Exact Solution for a Metapopulation Version of Schelling's Model

Richard Durrett and Yuan Zhang *

Department of Mathematics, Box 90320
Duke U., Durham, NC 27708-0320

July 16, 2014

Abstract

In 1971, Schelling introduced a model in which families move if they have too many neighbors of the opposite type. In this paper we will consider a metapopulation version of the model in which a city is divided into N neighborhoods each of which has L houses. There are ρNL red families and ρNL blue families. Families are happy if there are $\le \rho_c L$ families of the opposite type in their neighborhood, and unhappy otherwise. Each family moves to each vacant house at rates that depend on their happiness at their current location and that of their destination. Let $Tri(p_R, p_B)$ be a trinomial distribution with probability p_R and p_B of red and blue, and probability $1 - p_R - p_B$ of empty. Suppose first that $\rho_c > 0.25963$. In this case, if neighborhoods are large then there are critical values $\rho_b < \rho_d < \rho_c$. so that for $\rho < \rho_b$ the two types are distributed randomly in equilibrium, i.e, neighborhoods are $Tri(\rho, \rho)$. When $\rho > \rho_b$ a new segregated equilibrium $(1/2)Tri(\rho_1, \rho_2) + (1/2)Tri(\rho_2, \rho_1)$ appears with $\rho_1 > \rho_c > \rho_2$. When $\rho_b < \rho < \rho_d$ there is a bistability, but for $\rho > \rho_d$ the segregated state is the unique stationary distribution. When $\rho_c < 0.25963$, $Tri(\rho, \rho)$ may be the stationary distribution when ρ is close to $1/2$, and if so there is a region of bistability.

1 Model Description

In 1971, Schelling [1] introduced one of the first agent-based models in the social sciences. Families of two types inhabit cells in a finite square, with 25%–30% of the squares vacant. Each family had a neighborhood that consists of a 5×5 square centered at their location If the fraction of neighbors of the opposite type was too large then they move to the closest location that satisfies their constraints. Schelling simulated this and many other variants of this model (using dice and checkers) in order to argue that if people have a preference for living with those of their own color, the movements of individual families invariably led to complete segregation. [2]

*Both authors were partially supported by NSF grant DMS 1005470 from the probability program.

As Clark and Fossett [3] explain "The Schelling model was mostly of theoretical interest and was rarely cited until a significant debate about the extent and explanations of residential segregation in U.S. urban areas was engaged in the 1980s and 1990s. To that point, most social scientists offered an explanation that invoked housing discrimination, principally by whites." At this point Schelling's article has been cited more than 800 times. For a sampling of results from the social sciences literature see Fossett's lengthy survey [4], or other more recent treatments [5, 6, 7]. About ten years ago physicists discovered this model and analyzed it a number of variants using techniques of statistical mechanics, [8]–[14]. However to our knowledge the only rigorous work is [15] which studies the one-dimensional model in which the threshold for happiness is $\rho_c = 0.5$ and two unhappy families within distance w swap places at rate 1.

Here, we will consider a metapopulation version of Schelling's model in which there are N neighborhoods that have L houses, but we ignore spatial structure within the neighborhoods, and their physical locations. We do this to make the model analytically tractable, but these assumptions are reasonable from a modeling point of view. Many cities in the United States are divided into neighborhoods that have their own identities. In Durham, these neighborhoods have names like Duke Park, Trinity Park, Watts-Hillendale, Duke Forest, Hope Valley, Colony Park, etc. They are often separated by busy roads and have identities that are reinforced by email newsgroups that allow people to easily communicate with everyone in their neighborhood, so it is the overall composition of the neighborhood that is important not just the people who live next door. In addition, when a family decides to move they can easily relocate anywhere in the city.

Families, which we suppose are indivisible units, come in two types that we call red and blue. There are ρNL of each type, leaving $(1-2\rho)NL$ empty houses. This formulation was inspired by Grauwin et al. [16], who studied segregation in a model with one type of individual whose happiness is given by a piecewise linear unimodal function of the density of occupied sites in their neighborhood. To define the rules of movement, we introduce the threshold level ρ_c such that a neighborhood is happy for a certain type of agent if the fraction of agents of the opposite type is $\le \rho_c$. For each family and empty house, movements occur at rates that depend on the state of the source and destination houses:

from/to	Happy	Unhappy
Happy	$r/(NL)$	$\epsilon/(NL)$
Unhappy	$1/(NL)$	$q/(NL)$

where $q, r < 1$ and ϵ is small, e.g., 0.1 or smaller. Note that there are $O(NL)$ vacant houses so each family moves at a rate $O(1)$. In words, happy families, are very reluctant to move to a neighborhood in which they would be unhappy, while unhappy families move at rate 1 to neighborhoods that will make them happy. As we will see later, the equilibrium distribution does not depend on the values of q and r.

2 Convergence to a deterministic limit

To describe the dynamics more precisely, let $n_{i,j}$, $i, j \ge 0$, $i + j \le L$ be the number of neighborhoods with i red and j blue families, and let $n = NL$ be the total number of houses.

The configuration of the system at time t can be fully described by the numbers $u_{i,j} = n_{i,j}/N$, which is a probability measure on

$$\Omega_L = \{(i,j) \in \mathbb{Z}^2 : i,j \geq 0, i+j \leq L\}.$$

Thus we have a stochastic process ν_t^N taking values in $\mathcal{M}_1(\Omega_L)$, the space of probability measures on Ω_L, or if the reader prefers, a large vector of $u_{i,j}(t)$ of frequencies that change over time.

If one computes infinitesimal means and variances then it is natural to guess (and not hard to prove) that if we keep L fixed and let $N \to \infty$, then ν_t^N converges to a deterministic limit. Motivated by individual-based models in finance, Daniel Remenik [17] has proved a general result which takes care of our example. In order to state that result we use the same notation as in Remenik's work, despite the fact that it is somewhat difficult to parse.

To describe the jump rates, we need some notation. Let $\ell_c = [\rho_c L]$, where $[x]$ is the largest integer $\leq x$. In words, a family is happy if there are $\leq \ell_c$ families of the opposite type in their neighborhood. Let

$$\Delta(i_1, i_2) = \begin{cases} r & i_1 \leq l_c, i_2 \leq \ell_c \\ \epsilon & i_1 \leq l_c, i_2 > \ell_c \\ 1 & i_1 > \ell_c, i_2 \leq \ell_c \\ q & i_1 > l_c, i_2 > \ell_c \end{cases}$$

be the matrix of movement rates, which depends on the number of houses of the opposite type at the source i_1 and destination i_2. Let

$$\lambda(a_1, b_1; a_2, b_2) = \frac{1}{L}\left[a_1(L - a_2 - b_2)\Delta(b_1, b_2) + b_1(L - a_2 - b_2)\Delta(a_1, a_2)\right]$$

be N times the total rate of movement from one (a_1, b_1) neighborhood to one (a_2, b_2) neighborhood.

The distribution of the outcome of migration from $\omega_1 = (a_1, b_1)$ to $\omega_2 = (a_2, b_2)$, written as a measure on the new pair of states, is

$$\begin{aligned} b(\omega_1, \omega_2; d\omega_1' \otimes d\omega_2') = &\frac{a_1(L - a_2 - b_2)\Delta(b_1, b_2)}{L\lambda(a_1, b_1; a_2, b_2)} \delta_{(a_1 - 1, b_1; a_2 + 1, b_2)} \\ &+ \frac{b_1(L - a_2 - b_2)\Delta(a_1, a_2)}{L\lambda(a_1, b_1; a_2, b_2)} \delta_{(a_1, b_1 - 1; a_2, b_2 + 1)} \end{aligned}$$

where δ_x is a point-mass at x. In words, the new states will be $\omega_1' = (a_1 - 1, b_1)$ and $\omega_2' = (a_2 + 1, b_2)$ or $\omega_1' = (a_1, b_1 - 1)$ and $\omega_2' = (a_2, b_2 + 1)$ with the indicated probabilities.

Theorem 1. *The measure valued processes ν_t^N will converge weakly to the solution of the ODE:*

$$\begin{aligned} \frac{d\nu_t(i,j)}{dt} = &- \nu_t(i,j) \sum_{\omega \in \Omega} [\lambda(i,j;\omega) + \lambda(\omega; i, j)] \nu_t(\omega) \\ &+ \sum_{\omega, \omega' \in \Omega} \lambda(\omega; \omega') [b(\omega, \omega'; \{i,j\} \times \Omega) + b(\omega, \omega'; \Omega \times \{i,j\})] \, \nu_t(\omega) \nu_t(\omega'). \end{aligned} \tag{1}$$

The notation is a little cumbersome, but the result is not complicated. The first term comes from the fact that a migration from $(i,j) \to \omega$ or $\omega \to (i,j)$ destroys an (i,j) neighborhood, while the second reflects the fact that a migration $\omega \to \omega'$ may create an (i,j) neighborhood at the source or at the destination.

3 Special case $L = 2$

To illustrate the use of Theorem 1, we consider the case $L = 2$. When $L = 2$, a neighborhood with both types of families must be $(1,1)$, so the situation in which $\ell_c \geq 1$ is trivial because there are never any unhappy families. In the case $L = 2$ and $\ell_c = 0$, it is easy to find the equilibrium because there is detailed balance, i.e., the rate of each transition is exactly balanced by the one in the opposite direction.

rates	transitions
$r\nu_{1,0}^2 = 4r\nu_{0,0}\nu_{2,0}$	$(1,0)(1,0) \rightleftharpoons (0,0)(2,0)$
$r\nu_{0,1}^2 = 4r\nu_{0,0}\nu_{0,2}$	$(0,1)(0,1) \rightleftharpoons (0,0)(0,2)$
$2\nu_{0,0}\nu_{1,1} = \epsilon\nu_{1,0}\nu_{0,1}$	$(1,1)(0,0) \rightleftharpoons (1,0)(0,1)$
$\nu_{1,0}\nu_{1,1} = 2\epsilon\nu_{2,0}\nu_{0,1}$	$(1,1)(1,0) \rightleftharpoons (0,1)(2,0)$
$\nu_{0,1}\nu_{1,1} = 2\epsilon\nu_{0,2}\nu_{1,0}$	$(1,1)(0,1) \rightleftharpoons (1,0)(0,2)$

After a little algebra, we find that the fixed point must have the form:

$$\nu_{2,0} = \nu_{0,2} = x \quad \nu_{1,1} = 2\epsilon x \quad \nu_{1,0} = \nu_{0,1} = y \quad \nu_{00,} = y^2/4x.$$

At first, it may be surprising that the rate r has nothing to do with the fixed point, but if you look at the first two equations you see that the r appears on both sides. The parameter q does not appear either, but in this case it is for the trivial reason that transitions $(1,1)(1,0) \to (1,0)(1,1)$, which occur at rate q, do not change the state of the system.

Using now the fact that the equilibrium must preserve the red and blue densities, we can solve for x and y to conclude that the only fixed point must have the form above with

$$y = \frac{1 - \sqrt{8(1-\epsilon)\rho^2 - 4(1-\epsilon)\rho + 1}}{1-\epsilon}$$
$$x = \frac{(2-2\epsilon)\rho - 1 + \sqrt{8(1-\epsilon)\rho^2 - 4(1-\epsilon)\rho + 1}}{2(1+\epsilon)(1-\epsilon)}.$$

Since the formulas are somewhat complicated, Figure 1 shows how $\nu_{i,j}$ vary as a function of ρ.

Unfortunately, when $L \geq 3$, the Markov process on $\mathcal{M}_1(\Omega_L)$ is no longer be reversible. One can, of course, solve for the stationary distribution numerically. Figure 2 shows limit behavior of the system with $L = 20$, and $\rho_c = 0.3$, i.e., $\ell_c = 6$ for initial densities $\rho = 0.1$, 0.2, 0.25 and 0.35. In the first two cases, most of the families are happy. In the third situation, the threshold $\ell_c = 6$ while the average number of reds and blues per neighborhood is 5, but since fluctuations in the make of neighborhoods can lead to unhappiness, there is a tendency toward segregation. In the fourth case segregation is almost complete with most neighborhoods having 0 or 1 of the minority type.

4 Neighborhood-Environment Approach

Finding the stationary distribution requires solving $\sum_{i=0}^{L} L+1-i = (L+1)(L+2)/2$ equations, which is 231 when $L = 20$ and 5151 when $L = 100$. In this section, we will adopt a different approach, which allows us to explicitly compute the stationary distribution. We concentrate on the evolution of neighborhood 1 and consider neighborhoods 2–N to be its environment, which can be summarized by the following 4 parameters: (1) the average number of happy red and blue families per neighborhood, and (2) the average number of vacant sites happy for red or blue, again per neighborhood. If we let $n_{i,j}$ be the number of (i, j) neighborhoods then those parameters for red families can be written as:

$$h_R^1 = \sum_{j<l_c} n_{i,j} i, \qquad h_R^0 = \sum_{j<l_c} n_{i,j}(L-i-j) \tag{2}$$

Here, and in what follows, we will cut the number of formulas in half by not writing the analagous quantities for blues. From the four parameters h_R^1, h_0^R, h_1^B and h_0^B, we can calculate the rate at which reds arrive (superscript +) and leave (superscript −), sites in neighborhood 1 that are happy H and unhappy U for red. Letting N_R, N_B, and N_0 be the total number of red families, blue families, and empty sites,

$$\begin{aligned} H_R^+ &= [rh_R^1 + (N_R - h_R^1)]/NL \quad H_R^- = [rh_R^0 + (N_0 - h_R^0)]/NL \\ U_R^+ &= [\epsilon h_R^1 + q(N_R - h_R^1)]/NL \quad U_R^- = [h_R^0 + q(N_0 - h_R^0)]/NL \end{aligned}$$

From this, we see that the transition rates for neighborhood 1 are

	if $j_0 \le \ell_c$	if $j_0 > \ell_c$
$(i_0, j_0) \to (i_0, j_0+1)$	$(L-i_0-j_0)H_R^+$	$(L-i_0-j_0)U_R^+$
$(i_0, j_0) \to (i_0, j_0-1)$	$i_0 H_R^-$	$i_0 U_R^-$

If we specify the four parameters, then it is (almost) easy to compute the stationary distribution. Divide the state space $\Omega = \{(i, j) : i, j \ge 0, i + j \le L\}$ into four quadrants based on red and blue happiness. Writing 0 for H and 1 for U, we have

$$\begin{aligned} Q_{0,0} &= \{i \le \ell_c, j \le \ell_c\}, \qquad Q_{0,1} = \{i > \ell_c, j \le \ell_c\} \\ Q_{1,0} &= \{i \le \ell_c, j > \ell_c\}, \qquad Q_{1,1} = \{i > \ell_c, j > \ell_c\} \end{aligned}$$

If we let $Tri(p_R, p_B)$ be the trinomial distribution

$$\frac{L!}{i!j!(L-i-j)!} p_R^i p_B^j (1-p_R-p_B)^{L-i-j} \tag{3}$$

then inside $Q_{k,\ell}$, the detailed balance condition is satisfied by $Tri(p_R, p_B)$ where

$$p_R = \frac{\alpha_{k,\ell}}{1+\alpha_{k,\ell}+\beta_{k,\ell}} \quad \text{and} \quad p_B = \frac{\beta_{k,\ell}}{1+\alpha_{k,\ell}+\beta_{k,\ell}}$$

where the $\alpha_{k,l}$ and $\beta_{k,l}$ are as follows:

$$\alpha_{0,0} = \alpha_{0,1} = \frac{H_R^+}{H_R^-} \qquad \alpha_{1,0} = \alpha_{1,1} = \frac{U_R^+}{U_R^-}$$
$$\beta_{0,0} = \beta_{1,0} = \frac{H_B^+}{H_B^-} \qquad \beta_{0,1} = \beta_{1,1} = \frac{U_B^+}{U_B^-}.$$

Unfortunately, the Kolmogorov cycle condition, is not satisfied around loops that visit two or more quadrants, so there is no reversible stationary distribution.

5 Self-consistent distributions

The next step is to identify the stationary distribution of neighborhood 1, that are self-consistent. That is, when we calculate the expected values of h_R^1, h_R^0, h_B^1 and h_B^0 in equilibrium then they agree with the original parameters. The first step, accomplished in Section 1 of the supplementary materials is to show that such a measure cannot put mass on both $\Omega_{0,0}$ and $\Omega_{1,1}$.

Theorem 2. *Suppose there is no mass on $\Omega_{1,1}$. For $a \in (0, 1/2]$ let*

$$\rho_1(a,\rho) = \frac{-1 + (a+\rho)(1-\epsilon) + \sqrt{[1-(a+\rho)(1-\epsilon)]^2 + 4a(1-\epsilon^2)\rho}}{2a(1-\epsilon^2)}. \tag{4}$$

Let $\rho_1(0,\rho) = \rho/(1-\rho(1-\epsilon))$ and for $a \in [0, 1/2]$ let

$$\mu = (1-2a)\, Tri(\rho_0,\rho_0) + a\, Tri(\rho_1,\rho_2) + a\, Tri(\rho_2,\rho_1)$$

The distribution μ, it is self consistent if and only if it has the form above with parameters $\rho_1 > \rho_c$, $\rho_2 = \epsilon\rho_1 < \rho_c$ and $\rho_0 = \rho_1/[1+(1-\epsilon)\rho_1] < \rho_c$.

To clarify the last sentence: the definition of ρ_1 does not guarantee that the three conditions are satisfied for all values of $a \in [0, 1/2]$, so the inequalities are additional conditions. To explain the definition of $\rho_1(0,\rho)$, note that $\rho_1(a,\rho) \to \rho/(1-\rho(1-\epsilon)$ as $a \to 0$. A little algebra shows that

$$\rho_1(1/2,\rho) = \frac{2\rho(1-\epsilon)}{1-\epsilon^2}$$

Since families do not change type, we must have

$$\rho = (1-2a)\frac{\rho_1}{1+(1-\epsilon)\rho_1} + a\rho_1 + a\epsilon\rho_1$$

This equation shows that the mapping $a \to \rho_1(a,\rho)$ so it must be monotone, and in this case it is increasing.

Corollary 1. *If $\rho < \rho_c(1-\epsilon^2)/2(1-\epsilon)$ then $\rho_1(1/2,\rho) < \rho_c$ and hence $Tri(\rho,\rho)$ is the unique stationary distribution.*

The possible self-consistent stationary distributions are similar in the second case but the formulas are different.

Theorem 3. *Suppose there is no mass on $\Omega_{0,0}$ and for $a \in (0, 1/2]$ let*

$$\hat{\rho}_1 = \frac{\epsilon + (1-\epsilon)(a+\rho) - \sqrt{[\epsilon + (1-\epsilon)(a+\rho)]^2 - 4a(1-\epsilon^2)\rho}}{2a(1-\epsilon^2)}. \tag{5}$$

Let $\hat{\rho}_1(0,\rho) = \rho/(\epsilon + (1-\epsilon)\rho))$, and for $a \in [0, 1/2]$ let

$$\hat{\mu} = a\,Tri(\hat{\rho}_1, \hat{\rho}_2) + a\,Tri(\hat{\rho}_2, \hat{\rho}_1) + (1-2a)\,Tri(\hat{\rho}_3, \hat{\rho}_3)$$

The distribution $\hat{\mu}$ is self-consistent if and only if it has the form above with parameters $\hat{\rho}_1 > \rho_c$, $\hat{\rho}_2 = \epsilon\hat{\rho}_1 < \rho_c$ and $\hat{\rho}_3 = \epsilon\hat{\rho}_1/[1-(1-\epsilon)\hat{\rho}_1] > \rho_c$.

Again the formula for $\hat{\rho}_1(0,\rho)$ comes from taking the limit $a \to 0$. A little algebra shows

$$\rho_1(1/2, \rho) = \frac{2\rho(1-\epsilon)}{1-\epsilon^2}$$

i.e, the same formula as in the previous case, but this time $a \to \rho_1(a,\rho)$ is decreasing. In Section 3 of the supplementary materials we show that the situations in Theorems 2 and 3 correspond to $\rho < \rho_c$ and $\rho \geq \rho_c$.

6 Stability calculations

Since the measures in each quadrant are trinomial, the probabilities will decay exponentially away form the mean

$$\left(\frac{\alpha_{k,\ell}}{1+\alpha_{k,\ell}+\beta_{k,\ell}}, \frac{\beta_{k,\ell}}{1+\alpha_{k,\ell}+\beta_{k,\ell}}\right) L$$

Thus, unless one of the coordinates is close to ρ_c, the measure will be very small near the boundaries between the quadrants. This gives us a separation of time scales in the process.

Ansatz. Probability mass flows slowly, at rate $\exp(-cL)$, between quadrants, while equilibrium is restored in time $O(1)$, so the process is always in one of self-consistent stationary distributions.

We called our solution exact in the title of this paper, because we will not go through the pain of proving that here, and only give the answer that results if we assume this is correct.

Using "large deviations" for the trinomial distribution, which in this case is just using Stirling's formula, we conclude:

Theorem 4. *Suppose there is no mass on $\Omega_{1,1}$ and hence $\rho < \rho_c$. The flow into $Q_{0,0}$ from $Q_{0,1}$ and $Q_{1,0}$ is larger than the flow out if and only if*

$$\left(\frac{1-\epsilon\rho_1}{1-\rho_1}\right)^{\rho_c} < 1 + (1-\epsilon)\rho_1. \tag{6}$$

Suppose there is no mass on $\Omega_{0,0}$ and hence $\rho \geq \rho_c$. The flow out of $Q_{1,1}$ to $Q_{0,1}$ and $Q_{1,0}$ is larger than the flow in if and only if

$$\left(\frac{\hat{\rho}_1}{1-\hat{\rho}_1}\right)^{1-\rho_c} < (1-(1-\epsilon)\hat{\rho}_1)^{-1}. \tag{7}$$

7 Phase Transition

To illustrate the use of the results in the last two sections, we will now consider the special case $\rho_c = 0.2$ and $\epsilon = 0.1$. To follow the calculation it will be useful to refer to Figure 3. There the two curves are $\rho_1(0,\rho)$ and $\hat{\rho}_1(0,\rho)$, while the straight line is $\rho_1(1/2,\rho) = \hat{\rho}_1(1/2,\rho)$.

When $\rho_c = 0.2$ and $\epsilon = 0.1$, the inequality in (6) holds for $\rho_1 < 0.2183$. The upper bound on the interval of ρ_1's,

$$2\rho(0.9)/(0.99) = 0.2183 \quad \text{when} \quad \rho_b = 0.120065,$$

so we have uniqueness for $\rho < \rho_b$. The lower bound on the interval of ρ_1's,

$$\rho/(1 - 0.9\rho) = 0.2183 \quad \text{when} \quad \rho_d = 0.18245.$$

When $\rho_b < \rho < \rho_d$ there will be an $a_c \in (0, 1/2)$ so that the a in the mixture will decrease for $a < a_c$ and increase for $a > a_c$. so we have bistability. When $\rho = \rho_d$, $a_c = 0$ and the $Tri(\rho,\rho)$ fixed point loses its stability.

If $\rho_c > 0.25963$ then the inequality in (7) is always true. When $\rho_c = 0.2$, the two quantities are equal when $\rho_1 = 0.8724$. The upper bound on the interval of $\hat{\rho}_1$'s,

$$\rho/(0.1 + 0.9\rho) = 0.8724 \quad \text{when} \quad \hat{\rho}_b = 0.40607,$$

so bistability develops at this point. The lower bound on the interval of $\hat{\rho}_1$'s,

$$2\rho(0.9)/(0.99) = 0.8724 \quad \text{when} \quad \hat{\rho}_d = 0.47982.$$

At this point, the segregated fixed point loses its stability, and the answer again becomes $Tri(\rho,\rho)$.

8 Conclusions

Here, we have considered a metapopulation version of Schelling's model, which is arguably a better model for studying the dynamics of segregation in a city than a nearest neighborhood interaction on the two dimensional square lattice. Due to the simple two-level structure of the model, we are able to describe the phase transition in great detail. As ρ increases there is a discontinuous phase transition too a segregated state at ρ_d preceded by an interval (ρ_b, ρ_d) of bistability. Surprisingly the phase transition occurs at a value $\rho_d < \rho_c$, i.e., at a point where randomly distributed individuals are happy. This occurs because random fluctuations create segregated neighborhoods, which, as our stability analysis shows, are more stable than the random ones.

As ρ nears 1/2, there is another discontinuous transition at $\hat{\rho}_d$ which returns the equilibrium to the random state $Tri(\rho,\rho)$. This transition is preceded by an interval $(\hat{\rho}_b, \hat{\rho}_d)$ of bistability. To explain the return to $Tri(\rho,\rho)$ intuitively, we note that when families are distributed randomly, everyone is unhappy and moves at rate 1, maintaining the random distribution. In our concrete example, $\rho_c = 0.2$, $\epsilon = 0.1$, the faction of vacant houses at $\hat{\rho}_d$ only 4.036%, so it is very difficult to make segregated neighborhoods where one type is

happy. The stability analysis implies that these segregated neighborhoods are created at a slower rate than they are lost.

In Durham there are four or five dozen neighborhoods with roughly 100 houses in each. Fluctuations in the trinomial distributions are of order 10, so the phase transition will not be as sharp as in the $L \to \infty$ limit. However, our simulations show that even when $L = 20$, our predictions match the qualitative behavior of the model.

References

[1] Schelling, T.C. (1971) Dynamic models of segregation. *J. Mathematical Sociology.* 1, 143–186

[2] Schelling, T.C. (1978) *Micromotives and Macrobehavior.* Norton, New York

[3] Clark, W.A.V., and Fossett, M. (2008) Understanding the social context of Schelling's segregation model. *Proc. Natl. Acad. Sci.* 105, 4109–4114

[4] Fossett, M. (2006) Ethnic preferences, social science dynamics, and residential segregation: Theoretical explanations using simulation analysis. *J. Mathematical Sociology.* 30, 185–274

[5] Pancs, R., and Vriend, N.J. (2007) Schelling's spatial proximity model of segregation revisited. *Journal of Public Economics.* 91, 1–24

[6] Kandler, A., Perreault, C., and Steele, J. (2012) Cultural evolution in spatially structured populations: A review of alternative modeling frameworks. *Advances in Complex Systems.* 15, paper 1203001

[7] Hatna, E., and Benenson, I. (2009) The Schelling model of ethnic residential dyanmics: Beyon the integrated-segregation dichotomy of patterns. *Journal of Artificial Societies and Scoial Simulation.* 15 (1) 6

[8] Vinkovic, D., and Kirman, A. (2006) A physical analogue of the Schelling model. *Proc. Natl. Acad. Sci.* 103, 19261–19265

[9] Stauffer, D., and Solomon, S. (2007) Ising, Schelling and self-organizing segregation. *European Physical Journal B.* 57, 473–479

[10] Singh, A., Vainchtein, D., and Weiss, H. (2007) Schelling's segregation model: Parmaeters, Scaling, and Aggregation. arXiv:0711.2212

[11] Dall'Asta, L., Castellano, C., and Marsili, M. (2008) Statistical physics of the Schelling model of segregation. *Journal of Statistical Physics: Theory and Experiment.* Letter L07002

[12] Gauvin, L., Vannimenus, J., and Nadal, J-P. (2009) Phase diagram of a Schelling segregation model. *European Physical Journal B.* 70, 293–304

[13] Rogers, T., and McKane, A.J. (2011) A unified framework for Schelling's model of segregation. *Journal of Statistical Mechanics: Theory and Experiment.* Paper P07006

[14] Domic, N.G., Goles, E., and Rica, S. (2011) Dyanmics and complexity of the Schelling segregation model. *Physical Review E.* 83, paper 056111

[15] Brandt, C., Immorlica, N., Kamath, G., and Kleinberg, R. (2012) An analysis of one-dimensional Schelling segregation. arXiv:1203.6346

[16] Grauwin, S., Bertin, E., Lemoy, R., and Jensen, P. (2009) Competition between collective and individual dynamics. *Proc. Natl. Acad. Sci.* 106, 20622-20626

[17] Remenik, D. (2009) Limit theorems for individual-based models in economics and finance. *Stoch. Proc. Appl.* 119, 2401–2435

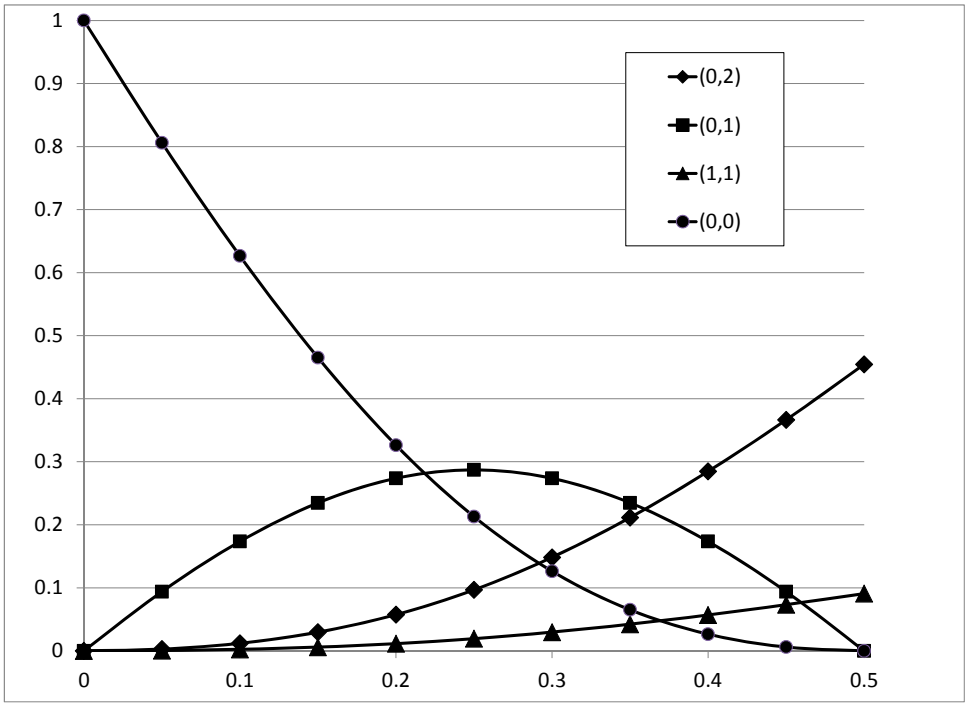

Figure 1: Equilibrium for the case $L = 2$ plotted against ρ.

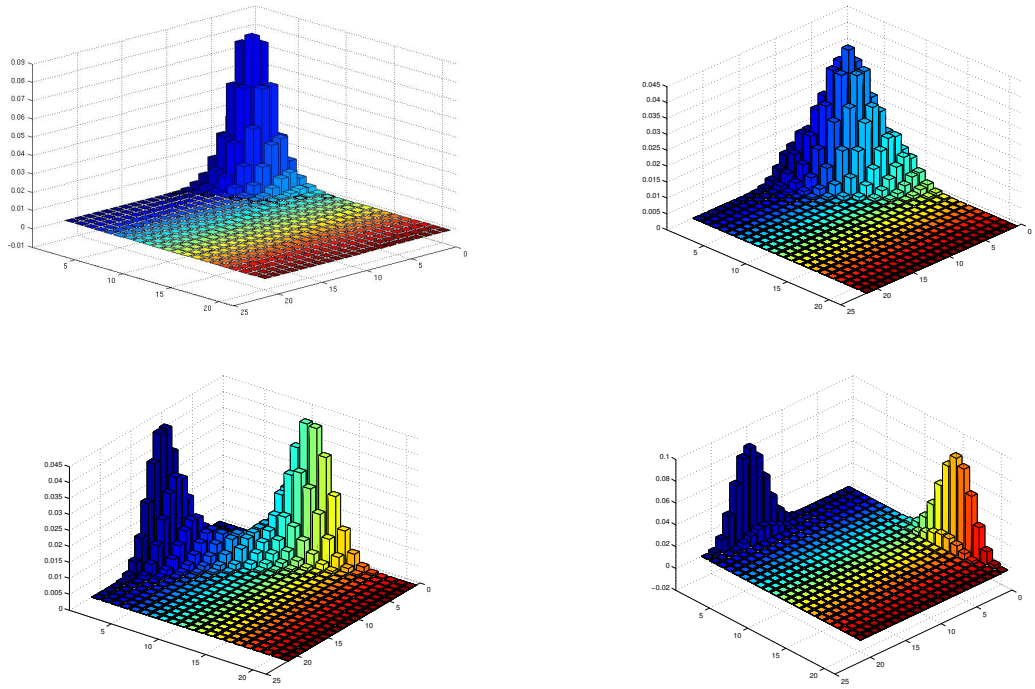

Figure 2: Limiting behavior of limit differential equation, with $\rho_c = 0.3, \epsilon = 0.01$, $\rho = 0.1$, 0.2, 0.25, and 0.35.

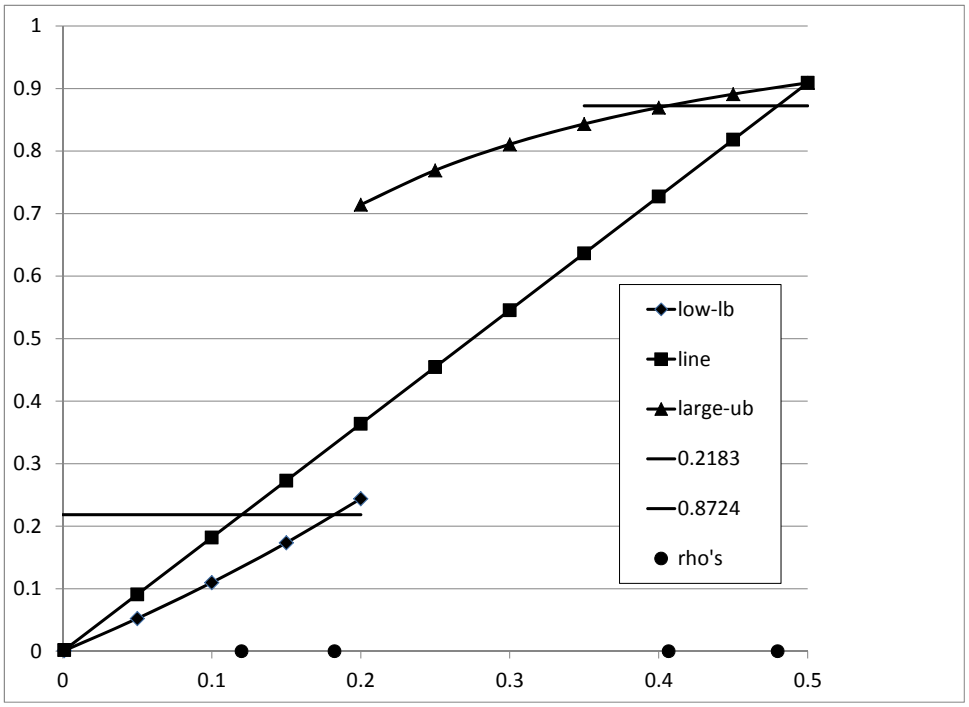

Figure 3: Picture to explain calculation of the phase transition when $\rho_c = 0.2$, $\epsilon = 0.1$. Dots on the axis are the locations of ρ_b, ρ_d, $\hat{\rho}_b$, $\hat{\rho}_d$.

Supplementary Materials for Durrett and Zhang

1 Proof of Theorem 2

The first step is to show

Lemma 1. *A measure of the form*

$$a\,Tri(\rho_0,\rho_0)+b\,Tri(\rho_1,\rho_2)+b\,Tri(\rho_2,\rho_1)+c\,Tri(\rho_3,\rho_3)$$

is self consistent only if $ac=0$, i.e. it cannot put positive mass on both $Q_{0,0}$ and $Q_{1,1}$.

Proof. Suppose $a,c\neq 0$. Then by self consistency, $\rho_0=\alpha_{0,0}/(1+2\alpha_{0,0})$ and $\hat{\rho}_3=\alpha_{1,1}/(1+2\alpha_{1,1})$. Since $\rho_0<\rho_c<\hat{\rho}_3$, we must have $\alpha_{0,0}<\alpha_{1,1}$. However, since $\epsilon<q,r<1$,

$$\alpha_{0,0}=\frac{rh_R^1+N_R-h_R^1}{rh_R^0+\epsilon(N_0-h_R^0)}>\frac{\epsilon h_R^1+q(N_R-h_R^1)}{h_R^0+q(N_0-h_R^0)}=\alpha_{1,1}$$

since the numerator of the first fraction is larger than the numerator of the second, and the denominator of the first fraction is smaller than the denominator of the second and we have a contradiction. □

Theorem 2 concerns the case in which there is no mass on $Q_{1,1}$ and the measure has the form

$$(1-2a)\,Tri(\rho_0,\rho_0)+a\,Tri(\rho_1,\rho_2)+a\,Tri(\rho_2,\rho_1)$$

with $\rho_0<l_c$, $\rho_2<l_c<\rho_1$. Our goal is to show that any self-consistent distribution of this form falls into the one-parameter family described in Theorem 2. The first step is recalling that under this case the environmental parameters are as follows:

$$\begin{aligned}
&h_R^1=h_B^1=(1-2a)\rho_0+a\rho_1\\
&N_R-h_R^1=N_B-h_B^1=a\rho_2\\
&h_R^0=h_B^0=(1-2a)(1-2\rho_0)+a(1-\rho_1-\rho_2)\\
&N_0-h_R^0=N_0-h_B^0=a(1-\rho_1-\rho_2).
\end{aligned}$$

Thus in $Q_{0,0}$:

$$\begin{aligned}
\alpha_{0,0}=\beta_{0,0}&=\frac{r[(1-2a)\rho_0+a\rho_1]+a\rho_2}{r[(1-2a)(1-2\rho_0)+a(1-\rho_1-\rho_2)]+\epsilon a(1-\rho_1-\rho_2)}\\
&=\frac{r(1-2a)\rho_0+a(r\rho_1+\rho_2)}{r(1-2a)(1-2\rho_0)+(r+\epsilon)a(1-\rho_1-\rho_2)}.
\end{aligned}$$

In $Q_{0,1}$, $\alpha_{0,1}=\alpha_{0,0}$ while

$$\begin{aligned}
\beta_{0,1}&=\frac{\epsilon[(1-2a)\rho_0+a\rho_1]+qa\rho_2}{[(1-2a)(1-2\rho_0)+a(1-\rho_1-\rho_2)]+qa(1-\rho_1-\rho_2)}\\
&=\frac{\epsilon(1-2a)\rho_0+a(\epsilon\rho_1+q\rho_2)}{(1-2a)(1-2\rho_0)+(1+q)a(1-\rho_1-\rho_2)}
\end{aligned}$$

since it is an unfriendly environment for blue individuals. Similarly, in $Q_{1,0}$, $\alpha_{1,0} = \beta_{0,1}$ and $\beta_{1,0} = \beta_{0,0}$. For self-consistency, the following equations have to be satisfied:

$$(i)\ \frac{\alpha_{0,0}}{1+\alpha_{0,0}+\beta_{0,0}} = \rho_0, \quad (ii)\ \frac{\alpha_{0,1}}{1+\alpha_{0,1}+\beta_{0,1}} = \rho_1, \quad (iii)\ \frac{\beta_{0,1}}{1+\alpha_{0,1}+\beta_{0,1}} = \rho_2.$$

To treat (i) we first note that if $\alpha_{0,0} = \beta_{0,0} = A/B$ where

$$\begin{aligned} A &= r(1-2a)\rho_0 + a(r\rho_1 + \rho_2) \\ B &= r(1-2a)(1-2\rho_0) + (r+\epsilon)a(1-\rho_1-\rho_2). \end{aligned}$$

With the notations above, one can easily see that

$$1 + \alpha_{0,0} + \beta_{0,0} = \frac{B+2A}{B}$$

and condition (i) is equivalent to $A = (B+2A)\rho_0$ or

$$r(1-2a)\rho_0 + a(r\rho_1+\rho_2) = [r(1-2a) + 2a(r\rho_1+\rho_2) + (r+\epsilon)a(1-\rho_1-\rho_2)]\rho_0. \tag{1}$$

Subtracting $r(1-2a)\rho_0$ and then dividing by a on both side of (1), we have

$$\rho_0(r+\epsilon)(1-\rho_1-\rho_2) = (r\rho_1+\rho_2)(1-2\rho_0). \tag{2}$$

This implies

$$1-\rho_1-\rho_2 = \frac{(r\rho_1+\rho_2)(1-2\rho_0)}{\rho_0(r+\epsilon)}. \tag{3}$$

Conditions (ii) and (iii) imply that $\alpha_{0,1}/\beta_{0,1} = \rho_1/\rho_2$, so we have

$$\frac{\rho_1}{\rho_2} = \frac{(1-2a)(1-2\rho_0) + (1+q)a(1-\rho_1-\rho_2)}{r(1-2a)(1-2\rho_0) + (r+\epsilon)a(1-\rho_1-\rho_2)} \times \frac{r(1-2a)\rho_0 + a(r\rho_1+\rho_2)}{\epsilon(1-2a)\rho_0 + a(\epsilon\rho_1 + q\rho_2)}. \tag{4}$$

Plugging (3) in to (4), we can simplify the equation and get

$$\frac{\rho_1}{\rho_2} = \frac{1}{r+\epsilon} \cdot \frac{(1-2a)\rho_0(r+\epsilon) + (1+q)a(r\rho_1+\rho_2)}{r(1-2a)\rho_0 + a(r\rho_1+\rho_2)} \times \frac{r(1-2a)\rho_0 + a(r\rho_1+\rho_2)}{\epsilon(1-2a)\rho_0 + a(\epsilon\rho_1+q\rho_2)}. \tag{5}$$

Canceling out $r(1-2a)\rho_0 + a(r\rho_1+\rho_2)$ and cross multiplying gives us

$$\rho_2[(1-2a)\rho_0(r+\epsilon) + (1+q)a(r\rho_1+\rho_2)] = \rho_1[\epsilon(1-2a)\rho_0(r+\epsilon) + a(r+\epsilon)(\epsilon\rho_1+q\rho_2)]. \tag{6}$$

For further simplification, note that we can rewrite equation (6) as

$$(1-2a)\rho_1(r+\epsilon)(\rho_2-\epsilon\rho_1) + a(1+q)\rho_2(r\rho_1+\rho_2) - a(r+\epsilon)\rho_1(\epsilon\rho_1+1\rho_2) = 0,$$

which is equivalent to

$$(1-2a)\rho_1(r+\epsilon)(\rho_2-\epsilon\rho_1) + a[(1+q)\rho_2 + (r+\epsilon)\rho_1](\rho_2-\epsilon\rho_1) = 0,$$

and

$$(\rho_2 - \epsilon\rho_1) \cdot [(1-2a)\rho_0(r+\epsilon) + a(r+\epsilon)\rho_1 + (1+q)a\rho_2] = 0. \tag{7}$$

Since $(1-2a)\rho_0(r+\epsilon) + a(\epsilon+r)\rho_1 + (1+q)a\rho_2 > 0$, (7) implies that

$$\rho_2 = \epsilon\rho_1. \tag{8}$$

Now plugging (8) back into (2), we have $\rho_0(1-(1+\epsilon)\rho_1) = \rho_1(1-2\rho_0)$ and

$$\rho_0 = \frac{\rho_1}{1+(1-\epsilon)\rho_1}. \tag{9}$$

To find ρ_1 note that $a\rho_1 + a\rho_2 + (1-2a)\rho_0 = \rho$ since the system preserves density, combine this with (8) and (9):

$$a(1+\epsilon)\rho_1 + (1-2a)\frac{\rho_1}{1+(1-\epsilon)\rho_1} = \rho.$$

Simplifying the equation above, we have:

$$a(1-\epsilon^2)\rho_1^2 + [1-(a+\rho)(1-\epsilon)]\rho_1 - \rho = 0.$$

Thus ρ_1 should be the positive solution of this quadratic equation:

$$\rho_1 = \frac{-1+(a+\rho)(1-\epsilon) + \sqrt{[1-(a+\rho)(1-\epsilon)]^2 + 4a(1-\epsilon^2)\rho}}{2a(1-\epsilon^2)} \tag{10}$$

and we have proved Theorem 2.

2 Proof of Theorem 3

We move now to the case when there is no mass on $Q_{0,0}$. The measure in this case can be written as:

$$a\,Tri(\hat{\rho}_1, \hat{\rho}_2) + a\,Tri(\hat{\rho}_2, \hat{\rho}_1) + (1-2a)\,Tri(\hat{\rho}_3, \hat{\rho}_3)$$

and the environmental parameters are now as follows:

$$\begin{aligned}
&h_R^1 = h_B^1 = a\hat{\rho}_1 \\
&N_R - h_R^1 = N_B - h_B^1 = (1-2a)\hat{\rho}_3 + a\hat{\rho}_2 \\
&h_R^0 = h_B^0 = a(1-\hat{\rho}_1 - \hat{\rho}_2) \\
&N_0 - h_R^0 = N_0 - h_B^0 = (1-2a)(1-2\hat{\rho}_3) + a(1-\hat{\rho}_1-\hat{\rho}_2).
\end{aligned}$$

As in case 1, in $Q_{1,1}$ we can compute the ratios α and β as follows:

$$\begin{aligned}
\alpha_{1,1} = \beta_{1,1} &= \frac{\epsilon a\hat{\rho}_1 + q[(1-2a)\hat{\rho}_3 + a\hat{\rho}_2]}{a(1-\hat{\rho}_1-\hat{\rho}_2) + q[(1-2a)(1-2\hat{\rho}_3) + a(1-\hat{\rho}_1-\hat{\rho}_2)]} \\
&= \frac{(1-2a)\hat{\rho}_3 + a(\epsilon\hat{\rho}_1 + q\hat{\rho}_2)}{q(1-2a)(1-2\hat{\rho}_3) + (1+q)a(1-\hat{\rho}_1-\hat{\rho}_2)}
\end{aligned}$$

while in $Q_{0,1}$, $\beta_{0,1} = \beta_{1,1}$ and

$$\begin{aligned}\alpha_{0,1} &= \frac{ra\hat{\rho}_1 + [(1-2a)\hat{\rho}_3 + a\hat{\rho}_2]}{ra(1-\hat{\rho}_1-\hat{\rho}_2) + \epsilon[(1-2a)(1-2\hat{\rho}_3) + a(1-\hat{\rho}_1-\hat{\rho}_2)]} \\ &= \frac{(1-2a)\hat{\rho}_3 + a(r\hat{\rho}_1 + \hat{\rho}_2)}{\epsilon(1-2a)(1-2\hat{\rho}_3) + (r+\epsilon)a(1-\hat{\rho}_1-\hat{\rho}_2)}.\end{aligned}$$

In case 2, a self-consistent distribution has to satisfy the following conditions:

$$(i)'\ \frac{\alpha_{1,1}}{1+\alpha_{1,1}+\beta_{1,1}} = \hat{\rho}_3; \quad (ii)'\ \frac{\alpha_{0,1}}{1+\alpha_{0,1}+\beta_{0,1}} = \hat{\rho}_1, \quad (iii)'\ \frac{\beta_{0,1}}{1+\alpha_{0,1}+\beta_{0,1}} = \hat{\rho}_2.$$

As before write $\alpha_{1,1} = \hat{A}/\hat{B}$ where

$$\begin{aligned}\hat{A} &= q(1-2a)\hat{\rho}_3 + a(\epsilon\hat{\rho}_1 + q\hat{\rho}_2) \\ \hat{B} &= q(1-2a)(1-2\hat{\rho}_3) + (1+q)a(1-\hat{\rho}_1-\hat{\rho}_2).\end{aligned}$$

Thus

$$\hat{B} + 2\hat{A} = q(1-2a) + (1+q)a(1-\hat{\rho}_1-\hat{\rho}_2) + 2a(\epsilon\hat{\rho}_1 + q\hat{\rho}_2)$$

and we need $\hat{A} = (\hat{B} + 2\hat{A})\hat{\rho}_3$ for condition $(i)'$, which can also be written as

$$q(1-2a)\hat{\rho}_3 + a(\epsilon\hat{\rho}_1 + q\hat{\rho}_2) = q(1-2a)\hat{\rho}_3 + (1+q)a(1-\hat{\rho}_1-\hat{\rho}_2)\hat{\rho}_3 + 2a(\epsilon\hat{\rho}_1 + q\hat{\rho}_2)\hat{\rho}_3.$$

Then again subtracting $q(1-2a)\hat{\rho}_3$ and dividing by a on both sides:

$$\epsilon\hat{\rho}_1 + \hat{\rho}_2 = (1+q)(1-\hat{\rho}_1-\hat{\rho}_2)\hat{\rho}_3 + 2(\epsilon\hat{\rho}_1 + q\hat{\rho}_2)\hat{\rho}_3$$

which can be simplified as

$$\begin{aligned}&(1+q)\hat{\rho}_3(1-\hat{\rho}_1-\hat{\rho}_2) = (1-2\hat{\rho}_3)(\epsilon\hat{\rho}_1 + q\hat{\rho}_2) \\ \Rightarrow &(1-\hat{\rho}_1-\hat{\rho}_2) = \frac{(1-2\hat{\rho}_3)(\epsilon\hat{\rho}_1 + q\hat{\rho}_2)}{(1+q)\hat{\rho}_3}.\end{aligned} \tag{11}$$

From conditions $(ii)'$ and $(iii)'$, $\alpha_{0,1}/\beta_{0,1} = \hat{\rho}_1/\hat{\rho}_2$. Thus

$$\frac{\hat{\rho}_1}{\hat{\rho}_2} = \frac{q(1-2a)(1-2\hat{\rho}_3) + (1+q)a(1-\hat{\rho}_1-\hat{\rho}_2)}{\epsilon(1-2a)(1-2\hat{\rho}_3) + (r+\epsilon)a(1-\hat{\rho}_1-\hat{\rho}_2)} \times \frac{(1-2a)\hat{\rho}_3 + a(r\hat{\rho}_1 + \hat{\rho}_2)}{q(1-2a)\hat{\rho}_3 + a(\epsilon\hat{\rho}_1 + q\hat{\rho}_2)}. \tag{12}$$

Then using exactly the same calculation as in the proof of Theorem 2 by plugging (11) into (12), we get

$$\frac{\hat{\rho}_1}{\hat{\rho}_2} = (1+q)\frac{q(1-2a)\hat{\rho}_3 + a(\epsilon\hat{\rho}_1 + q\hat{\rho}_2)}{\epsilon(1-2a)(1+q)\hat{\rho}_3 + (r+\epsilon)a(\epsilon\hat{\rho}_1 + q\hat{\rho}_2)} \times \frac{(1-2a)\hat{\rho}_3 + a(r\hat{\rho}_1 + \hat{\rho}_2)}{q(1-2a)\hat{\rho}_3 + a(\epsilon\hat{\rho}_1 + q\hat{\rho}_2)}$$

which implies:

$$\hat{\rho}_2[(1-2a)(1+q)\hat{\rho}_3 + (1+q)a(r\hat{\rho}_1 + \hat{\rho}_2)] = \hat{\rho}_1\left(\epsilon(1-2a)(1+q)\hat{\rho}_3 + (r+\epsilon)a(\epsilon\hat{\rho}_1 + q\hat{\rho}_2)\right) \tag{13}$$

after we cancel the term of $q(1-2a)\hat{\rho}_3 + a(\epsilon\hat{\rho}_1 + q\hat{\rho}_2)$. Simplifying (13) with exactly the same procedure as in the proof of Theorem 2, we have

$$(\hat{\rho}_2 - \epsilon\hat{\rho}_1) \cdot [(1-2a)(1+q)\hat{\rho}_3 + (r+\epsilon)a\hat{\rho}_1 + (1+q)a\hat{\rho}_2] = 0. \tag{14}$$

It is clear that the second term in the product on the left side of (14) is positive, which implies:

$$\hat{\rho}_2 = \epsilon\hat{\rho}_1. \tag{15}$$

Using this in (11) gives

$$2\hat{\rho}_3(1 - \hat{\rho}_1 - \epsilon\hat{\rho}_1) = (1 - 2\hat{\rho}_3)(2\epsilon\hat{\rho}_1)$$

which can be simplified to

$$\hat{\rho}_3 = \frac{\epsilon\hat{\rho}_1}{1 - \hat{\rho}_1(1-\epsilon)}. \tag{16}$$

Noting that $a(\hat{\rho}_1 + \hat{\rho}_2) + (1-2a)\hat{\rho}_3 = \rho$ and using (15) and (16), we have

$$a(1+\epsilon)\hat{\rho}_1 + (1-2a)\frac{\epsilon\hat{\rho}_1}{1-(1-\epsilon)\hat{\rho}_1} = \rho$$

and 0

$$a(1-\epsilon^2)\hat{\rho}_1^2 - [\epsilon + (1-\epsilon)(a+\rho)]\hat{\rho}_1 + \rho = 0. \tag{17}$$

The coefficient of $\hat{\rho}_1^2$ and the constant term are positive and the coefficient of $\hat{\rho}_1$ is negative, and we expect the roots to be real so the quadratic equation above has two positive solutions, say $0 < x_1 < x_2$. Suppose $\hat{\rho}_1$ equals to the bigger solution x_2. Then the smaller solution $x_1 < x_2 = \hat{\rho}_1 < 1$. Note that $a(1+\epsilon)\hat{\rho}_1 + (1-2a)\frac{\epsilon\hat{\rho}_1}{1-(1-\epsilon)\hat{\rho}_1} = \rho$, which implies

$$a(1+\epsilon)\hat{\rho}_1 \leq \rho.$$

Thus we have

$$x_1 x_2 < x_2 = \hat{\rho}_1 \leq \frac{\rho}{a(1+\epsilon)}.$$

However, from equation (17):

$$x_1 x_2 = \frac{\rho}{a(1-\epsilon^2)} = \frac{\rho}{a(1+\epsilon)}\frac{1}{1-\epsilon} > \frac{\rho}{a(1+\epsilon)}$$

and we get a contradiction. Thus $\hat{\rho}_1$ has to be the smaller solution x_1 of the equation above and

$$\hat{\rho}_1 = \frac{\epsilon + (1-\epsilon)(a+\rho) - \sqrt{[\epsilon + (1-\epsilon)(a+\rho)]^2 - 4a(1-\epsilon^2)\rho}}{2a(1-\epsilon^2)} \tag{18}$$

which completes the proof of Theorem 3.

3 Density of Self-Consistent Distributions

Our next step is to show that whenever a self-consistent distribution falls into form of Theorem 2 we must have the corresponding overall density

$$\rho = a\rho_1 + a\rho_2 + (1-2a)\rho_0$$

satisfies $\rho < \rho_c$. And similarly when it falls into case 2 we must have the density $\rho > \rho_c$.

For a self-consistent distribution in case 1, $\rho_2 = \epsilon\rho_1$ and $\rho_0 = \rho_1/[1+(1-\epsilon)\rho_1]$. Note that

$$\begin{aligned} 2\rho_0 - (1+\epsilon)\rho_1 &= \frac{2\rho_1}{1+(1-\epsilon)\rho_1} - (1+\epsilon)\rho_1 \\ &= \frac{2\rho_1 - (1+\epsilon)\rho_1 - (1+\epsilon)(1-\epsilon)\rho_1^2}{1+(1-\epsilon)\rho_1} = \frac{(1-\epsilon)\rho_1[1-(1+\epsilon)\rho_1]}{1+(1-\epsilon)\rho_1} \end{aligned}$$

since $(1+\epsilon)\rho_1 = \rho_1 + \rho_2 \le 1$, $2\rho_0 \ge (1+\epsilon)\rho_1 = \rho_1 + \rho_2$. Combine this with the fact that $\rho_0 < \rho_c$, we have $\rho = (1-2a)\rho_0 + a(\rho_1+\rho_2) \le \rho_0 < \rho_c$.

Similarly, for self-consistent distribution in Theorem 3, we have $\hat\rho_3 = \epsilon\hat\rho_1/[1-(1-\epsilon)\hat\rho_1]$, then

$$\begin{aligned} 2\hat\rho_3 - (1+\epsilon)\hat\rho_1 &= \frac{2\epsilon\hat\rho_1}{1-(1-\epsilon)\hat\rho_1} - (1+\epsilon)\hat\rho_1 \\ &= \frac{2\epsilon\hat\rho_1 - (1+\epsilon)\hat\rho_1 + (1+\epsilon)(1-\epsilon)\hat\rho_1^2}{1-(1-\epsilon)\hat\rho_1} = \frac{(1-\epsilon)\hat\rho_1((1+\epsilon)\hat\rho_1 - 1)}{1-(1-\epsilon)\hat\rho_1} < 0. \end{aligned}$$

Thus $(1+\epsilon)\hat\rho_1 \ge 2\hat\rho_3 > 2\rho_c$, and $\rho = (1-2a)\hat\rho_3 + a(\hat\rho_1+\hat\rho_2) \ge \rho_c$.

4 Stability Calculations

To have the formulas at hand we recall the statement.

Theorem 2. *Suppose there is no mass on $Q_{1,1}$. For $a \in (0, 1/2]$ let*

$$\rho_1(a,\rho) = \frac{-1+(a+\rho)(1-\epsilon) + \sqrt{[1-(a+\rho)(1-\epsilon)]^2 + 4a(1-\epsilon^2)\rho}}{2a(1-\epsilon^2)}. \tag{19}$$

Let $\rho_1(0,\rho) = \rho/(1-\rho(1-\epsilon))$ and for $a \in [0, 1/2]$ let

$$\mu = (1-2a)\,Tri(\rho_0,\rho_0) + a\,Tri(\rho_1,\rho_2) + a\,Tri(\rho_2,\rho_1)$$

The distribution μ, it is self consistent if and only if it has the form above with parameters $\rho_1 > \rho_c$, $\rho_2 = \epsilon\rho_1 < \rho_c$ and $\rho_0 = \rho_1/[1+(1-\epsilon)\rho_1] < \rho_c$.

Suppose $\rho < \rho_c$, The first task is to determine when the ρ_i given in Theorem 2 satisfy the desired inequalities. Let $b = 1-(a+\rho)(1-\epsilon)$. When a is small, $\sqrt{b^2 + 4a(1-\epsilon^2)\rho} \approx b + 2a(1-\epsilon^2)\rho/b$ so

$$\rho_1(a,\rho) \approx \frac{2a(1-\epsilon^2)\rho}{2a(1-\epsilon^2)b} \to \frac{\rho}{1-\rho(1-\epsilon)} \quad \text{as } a \to 0.$$

From this we see that when $a = 0$,

$$\rho_0 = \frac{\rho/(1-\rho(1-\epsilon))}{(1-\rho(1-\epsilon)+\rho(1-\epsilon))/(1-\rho(1-\epsilon))} = \rho.$$

When $a = 1/2$ the quantity under the square root is

$$C = [1-(1/2+\rho)(1-\epsilon)]^2 + 2(1-\epsilon^2)\rho.$$

We claim this is the same as

$$D = [1-(1/2-\rho)(1-\epsilon)]^2.$$

To check this note that

$$\begin{aligned} C - D &= -4\rho(1-\epsilon) + 2\rho(1-\epsilon)^2 + 2(1-\epsilon^2)\rho \\ &= \rho[-4+4\epsilon+2(1-2\epsilon+\epsilon^2+2(1-\epsilon^2)] = 0. \end{aligned}$$

Putting D under the square root

$$\rho_1(1/2,\rho) = \frac{2\rho(1-\epsilon)}{(1-\epsilon^2)}. \tag{20}$$

In order for the measure constructed above to be valid we must have

$$\rho_c > \rho_0 = \frac{\rho_1}{1+(1-\epsilon)\rho_1}. \tag{21}$$

which implies

$$\rho_1 < \frac{\rho_c}{1-(1-\epsilon)\rho_c} \tag{22}$$

When (21) fails mass will flow out of $Q_{0,0}$, so the solution will be

$$(1/2)Tri(\rho_1,\rho_2) + (1/2)Tri(\rho_2,\rho_1).$$

We will now investigate the stability of our proposed equilibria. Suppose we have a trinomial

$$\frac{L!}{i!j!(L-i-j)!}p_R^i p_B^j (1-p_R-p_B)^{L-i-j}.$$

Using Stirling's formula $n! \sim n^n e^{-n}\sqrt{2\pi n}$, dropping the square root terms, and noticing the e^{-n} terms cancel in a multinomial coefficient, this becomes

$$\frac{L^L}{i^i j^j (L-i-j)^{L-i-j}}p_R^i p_B^j (1-p_R-p_B)^{L-i-j}.$$

We are interested in what happens when $i = \rho_c L$. Dividing top and bottom by L^L and inserting the definitions

$$\begin{aligned} &= \rho_c^{-\rho_c L}(\theta)^{-\theta L}(1-\rho_c-\theta)^{1-\rho_c-\theta}p_R^{\rho_c L}p_B^{\theta L}(1-p_R-p_B)^{(1-\rho_c-\theta)L} \\ &= \left(\frac{p_R}{\rho}\right)^{\rho L}\left(\frac{p_B}{\theta}\right)^{\theta L}\left(\frac{1-p_R-p_B}{1-\rho-\theta}\right)^{(1-\rho-\theta)L} \end{aligned} \tag{23}$$

Taking logs and dividing by L we want to maximize:

$$\rho_c \log(p_R/\rho_c) + \theta \log(p_B/\theta) - (1-\rho_c-\theta)\log\left(\frac{1-p_R-p_B}{1-\rho_c-\theta}\right).$$

Taking the derivative with respect to θ

$$\frac{d}{d\theta} = \log(p_B/\theta) + \theta(-1/\theta) - \log\left(\frac{1-p_R-p_B}{1-\rho_c-\theta}\right) - (1-\rho_c-\theta)\cdot\frac{-1}{1-\rho_c-\theta}.$$

The derivative is 0 when

$$\frac{\theta}{p_B} = \frac{1-\rho_c-\theta}{1-p_R-P_B}, \tag{24}$$

i.e., the trials that do not result in R are allocated between B and 0 (i.e., neither R nor B) in proportion to their probabilities. Solving gives

$$(1-\rho_c-\theta)p_B = \theta(1-p_R-p_B) \quad \text{or} \quad \theta = \frac{(1-\rho_c)}{(1-p_R)}p_B.$$

Using (24) in (23), the maximum probability is

$$\left(\frac{p_R}{\rho_c}\right)^{\rho_c L}\left(\frac{1-p_R}{1-\rho_c}\right)^{(1-\rho_c)L}. \tag{25}$$

In $Q_{0,0}$ where $p_R = p_B = \rho_0 < \rho_c$ this is

$$\theta = \frac{(1-\rho_c)}{1-\rho_0}\rho_0 < \rho_0 < \rho_c.$$

In $Q_{0,1}$ where $p_R = \rho_1 > \rho_c$ and $p_B = \epsilon\rho_1 < \rho_c$, the maximizing θ is

$$\frac{(1-\rho_c)}{(1-\rho_1)}\epsilon\rho_1.$$

Using (22) this is

$$\leq (1-\rho_c)\frac{\epsilon\rho_c}{1-(1-\epsilon)\rho_c}\cdot\frac{1+\epsilon\rho_c}{1-(1-\epsilon)\rho_c} < \rho_c,$$

since $\rho_c \leq 1/2$.

Putting the information from the last paragraph into (25), and discarding the denominators we want to show

$$\rho_0^{\rho_c L}(1-\rho_0)^{(1-\rho_c)L} < \rho_1^{\rho_c L}(1-\rho_1)^{(1-\rho_c)L}.$$

Remembering $\rho_0 = \rho_1/(1+(1-\epsilon)\rho_1)$ and noting $1-\rho_0 = (1-\epsilon\rho_1)/(1+(1-\epsilon)\rho_1)$ this is equivalent to

$$\left(\frac{\rho_1}{1+(1-\epsilon)\rho_1}\right)^{\rho_c L}\left(\frac{1-\epsilon\rho_1}{1+(1-\epsilon)\rho_1}\right)^{(1-\rho_c)L} < \rho_1^{\rho_c L}(1-\rho_1)^{(1-\rho_c)L}.$$

Cancelling and rearranging we want

$$\left(\frac{1-\epsilon\rho_1}{1-\rho_1}\right)^{(1-\rho_c)} < 1+(1-\epsilon)\rho_1,$$

which proves the first part of Theorem 4.

Theorem 3. *Suppose there is no mass on $Q_{0,0}$ and for $a \in (0,1/2]$ let*

$$\hat{\rho}_1 = \frac{\epsilon+(1-\epsilon)(a+\rho)-\sqrt{[\epsilon+(1-\epsilon)(a+\rho)]^2-4a(1-\epsilon^2)\rho}}{2a(1-\epsilon^2)} \tag{26}$$

let $\hat{\rho}_1(0,\rho) = \rho/(\epsilon+(1-\epsilon)\rho))$, and for $a \in [0,1/2]$ let

$$\hat{\mu} = a\,Tri(\hat{\rho}_1,\hat{\rho}_2) + a\,Tri(\hat{\rho}_2,\hat{\rho}_1) + (1-2a)\,Tri(\hat{\rho}_3,\hat{\rho}_3)$$

The distribution $\hat{\mu}$ is self-consistent if and only if it has the form above with parameters $\hat{\rho}_1 > \rho_c$, $\hat{\rho}_2 = \epsilon\hat{\rho}_1 < \rho_c$ and $\hat{\rho}_3 = \epsilon\hat{\rho}_1/[1-(1-\epsilon)\hat{\rho}_1] > \rho_c$.

To explain the definition of $\hat{\rho}(0,\rho)$, let $b = \epsilon+(a+\rho)(1-\epsilon)$. When a is small,

$$\sqrt{b^2-4a(1-\epsilon^2)\rho} \approx b - 2a(1-\epsilon^2)\rho/b,$$

so we have

$$\hat{\rho}_1(a,\rho) \approx \frac{2a(1-\epsilon^2)\rho}{2a(1-\epsilon^2)b} \to \frac{\rho}{\epsilon+\rho(1-\epsilon)} \quad \text{as } a \to 0.$$

Note that when $a = 0$, we have

$$\hat{\rho}_3 = \frac{\epsilon\rho/(\epsilon+(1-\epsilon)\rho)}{\epsilon/(\epsilon+(1-\epsilon)\rho)} = \rho.$$

At the other extreme $a = 1/2$, the quantity under the square root is

$$\hat{C} = [\epsilon+(1-\epsilon)(1/2+\rho)]^2 - 2(1-\epsilon^2)\rho.$$

We claim that this is equal to

$$\hat{D} = [\epsilon+(1-\epsilon)(1/2-\rho)]^2.$$

To check this, note that

$$\begin{aligned}\hat{C}-\hat{D} &= 4\epsilon(1-\epsilon)\rho + (1-\epsilon)^2\cdot 2\rho - 2(1-\epsilon^2)\\ &= \rho[4\epsilon-4\epsilon^2+2-2\epsilon+2\epsilon^2-2+2\epsilon^2] = 0.\end{aligned}$$

Putting $\hat{D}$ under the square root,

$$\hat{\rho}_1(1/2,\rho) = \frac{2\rho(1-\epsilon)}{1-\epsilon^2},$$

which agrees with (20), but now the possible values of ρ_1 are $[\hat{\rho}_1(1/2,\rho),\hat{\rho}_1(0,\rho)]$.

To determine the rate of flow between $Q_{1,0}$ and $Q_{1,1}$, we use (25). We choose these quadrants so that again the boundary is at $i = \ell_c$. In $Q_{1,0}$ we have $p_R = \hat{\rho}_2$ and $p_B = \hat{\rho}_1$, so the maximum occurs at

$$\theta = \frac{(1-\rho_c)}{1-p_R} p_B = \frac{(1-\rho_c)}{1-\epsilon\hat{\rho}_1}\hat{\rho}_1.$$

In $Q_{1,1}$, we have $p_R = p_B = \hat{\rho}_3$, so the maximum occurs at

$$\theta = \frac{1-\rho_c}{1-\hat{\rho}_3}\hat{\rho}_3 > \hat{\rho}_3 > \rho_c.$$

Thus to show that there will be no mass on $Q_{1,1}$ we want to show

$$\hat{\rho}_2^{\rho_c L}\hat{\rho}_1^{(1-\rho_c)L} < \hat{\rho}_3^{\rho_c L}(1-\hat{\rho}_3)^{(1-\rho_c)L}.$$

Filling in the definitions we need

$$\epsilon^{\rho_c L}\hat{\rho}_1^{L} < \left(\frac{\epsilon\hat{\rho}_1}{1-(1-\epsilon)\hat{\rho}_1}\right)^{\rho_c L}\left(\frac{1-\hat{\rho}_1}{1-(1-\epsilon)\hat{\rho}_1}\right)^{(1-\rho_c)L}.$$

Cancelling, rearranging, and raising both sides to the $1/L$ power, we want

$$\left(\frac{\hat{\rho}_1}{1-\hat{\rho}_1}\right)^{(1-\rho_c)} < (1-(1-\epsilon)\hat{\rho}_1)^{-1}. \tag{27}$$

If $\hat{\rho}_1 \le 1/2$ the inequality is satisfied since

$$LHS \le 1 < RHS.$$

When $\epsilon = 0.1$ the maximum possible value is $\rho_{max} = 0.9/0.99$.

$$\rho_{max}/(1-\rho_{max}) = .9/.09 = 10$$
$$(1-.9\rho_{max})^{-1} = 5.5$$

When $\rho_c = 0.25963$ the two quantities are equal at ρ_{max}, so the inequality always holds when $\rho_c > 0.25963$ but when $\rho_c < 0.25963$ it fails for ρ close to $1/2$.