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

August 2, 2014

Classification. Physical Sciences: Applied Mathematics

Corresponding Author:
Richard Durrett
Department of Mathematics, Box 90320
Duke U., Durham, NC 27708-0320
Phone: (919) 660-6970
Email: rtd@math.duke.edu

Keywords: Schelling's model, segregation, metapopulation, bistability, large deviations

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 for some $\rho < 1/2$. Families are happy if there are $\leq \rho_c L$ 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. Our main result is that 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. When $\rho > \rho_b$ a new segregated equilibrium appears; for $\rho_b < \rho < \rho_d$ there is a bistability, but when increases past ρ_d the random state is no longer stable. When ρ_c is small enough, the random state will again be the stationary distribution when ρ is close to $1/2$. If so, this is preceded by a region of bistability.

Significance Statement

Over forty years ago, Schelling introduced one of the first agent-based models in the social sciences. The model showed that even if people only have a mild preference for living with neighbors of the same color, complete segregation will occur. This model has been much discussed by social scientists and analyzed by physicists using analogies with spin-1 Ising models and other systems. Here, we study the metapopulation version of the model, which mimics the division of a city into neighbor hoods and we present the first analysis which gives detailed information about the structure of equilibria, and explicit formulas for their densities.

1 Introduction

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 has a neighborhood that consists of a 5×5 square centered at their location If the fraction of neighbors of the opposite type is 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]

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 a number of variants of it 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. Because of this 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 $\leq \rho_c$. 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. Since there are $O(NL)$ vacant houses,

dividing the rates by NL makes each family moves at a rate $O(1)$. Since ϵ is small, 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 the rates q and r for transitions that do not change a families happiness. We do not have an intuitive explanation for this result.

2 Convergence to a deterministic limit

To describe the dynamics more precisely, let $n_{i,j}(t)$, be the number of neighborhoods with i red and j blue families for $(i,j) \in \Omega = \{(i,j) : i,j \geq 0, i+j \leq L\}$. The configuration of the system at time t can be fully described by the numbers $\nu_t^N(i,j) = n_{i,j}(t)/N$. 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. To describe the limit, we need some notation. Let $\lambda(a_1,b_1;a_2,b_2)$ be N times the total rate of movement from one (a_1,b_1) neighborhood to one (a_2,b_2) neighborhood. Let $b(\omega_1,\omega_2;\omega_1',\omega_2')$ be N times the rate at which a movement from one $\omega_1 = (a_1,b_1)$ neighborhood to one $\omega_2 = (a_2,b_2)$ neighborhood turns the pair ω_1,ω_2 into ω_1',ω_2'. The exact formulas for these quantities are not important, so they are hidden away in Section 1 of the Supplementary Materials.

Remth **Theorem 1.** *As $N \to \infty$, the $\nu_t^N(i,j)$ converge in probability to the solution of the ODE:*

$$\begin{aligned}\frac{d\nu_t(i,j)}{dt} =& -\nu_t(i,j)\sum_{\omega\in\Omega}[\lambda(i,j;\omega)+\lambda(\omega;i,j)]\nu_t(\omega) \\ &+ \sum_{\omega_1,\omega_2\in\Omega}\left[b(\omega_1,\omega_2;(i,j),\omega') + b(\omega_1,\omega_2;\omega',(i,j))\right]\nu_t(\omega_1)\nu_t(\omega_2).\end{aligned} \tag{1}$$ Remeq

We do not sum over ω' since its value is determined by ω_1 and ω_2. 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 $L = 2$, and let $\ell_c = [\rho_c L]$ be the largest number of neighbors of the opposite type which allows a family to be happy. Here, $[x]$ is the largest integer $\leq x$. 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.

$$
\begin{array}{ll}
r\nu_{1,0}^2 = 4r\nu_{0,0}\nu_{2,0} & (1,0)(1,0) \rightleftharpoons (0,0)(2,0) \\
r\nu_{0,1}^2 = 4r\nu_{0,0}\nu_{0,2} & (0,1)(0,1) \rightleftharpoons (0,0)(0,2) \\
2\nu_{0,0}\nu_{1,1} = \epsilon\nu_{1,0}\nu_{0,1} & (1,1)(0,0) \rightleftharpoons (1,0)(0,1) \\
\nu_{1,0}\nu_{1,1} = 2\epsilon\nu_{2,0}\nu_{0,1} & (1,1)(1,0) \rightleftharpoons (0,1)(2,0) \\
\nu_{0,1}\nu_{1,1} = 2\epsilon\nu_{0,2}\nu_{1,0} & (1,0)(1,1) \rightleftharpoons (1,0)(0,2)
\end{array}
$$

After a little algebra, see Section 2 of the Supplementary Materials, we find that this holds if and only if:

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

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 when $L = 2$ 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

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

The argument given here shows that this is the only fixed point that satisfies detailed balance. We prove in Section 2 of the Supplementary Materials that it is the only fixed point. Since the formulas, which result from solving a quadratic equation, are somewhat complicated, Figure 1 shows how the equilibrium probabilities $\nu_{i,j}$ vary as a function of ρ.

Unfortunately, when $L \geq 3$, there is no stationary distribution that satisfies detailed balance. 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 up 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 roughly $L^2/2$ equations. To be precise, 231 equations 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, h_R^1 and h_B^1 and (2) the average number of vacant sites happy for red or blue, h_r^0 and h_B^0, again per neighborhood.

If we specify these four parameters, then it is (almost) easy to compute the stationary distribution. Divide the state space Ω into four quadrants based on red and blue happiness. Writing 0 for H and 1 for U, we have

$j > \ell_c$	$Q_{1,0}$	$Q_{1,1}$
$j \le \ell_c$	$Q_{0,0}$	$Q_{0,1}$
	$i \le \ell_c$	$i > \ell_c$

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

$$\frac{L!}{i!j!(L-i-j)!} p_R^i p_B^j (1-p_R-p_B)^{L-i-j} \tag{2}$$ tri

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

$$p_R = \frac{\alpha_{k,\ell}}{1+\alpha_{k,\ell}+\beta_{k,\ell}}, \qquad p_B = \frac{\beta_{k,\ell}}{1+\alpha_{k,\ell}+\beta_{k,\ell}},$$

and formulas for $\alpha_{k,\ell}$ and $\beta_{k,\ell}$ are given in Section 3 of the Supplementary Materials.

Unfortunately, the Kolmogorov cycle condition, is not satisfied around loops that visit two or more quadrants, so there is no stationary distribution that satisfies detailed balance. A second problem is that a distribution satisfying detailed balance inside each quadrant but not across the boundary, may not be close to the true stationary distribution.

5 Outline of our solution

Our analysis of Schelling's model is carried out in three steps. The first is to identify the stationary distribution of the neighborhood-environment chain that are **self-consistent**. That is, if we cut the connections between the quadrants, so the trinomials introduced above are stationary distributions, and then we calculate the expected values of h_R^1, h_R^0, h_B^1 and h_B^0 in equilibrium, they agree with the original parameters. At this point, we can only do this under

Assumption 1. *Stationary distributions are symmetric under interchange of red and blue.*

The answer given in Section 6 is a one-parameter family of stationary distributions indexed by $a \in [0, 1/2]$. There, and in the next two steps, we have a pair of results, one for $\rho < \rho_c$ and one for $\rho > \rho_c$.

In Section 7 we investigate the flow of probability between quadrants when transitions between the quadrants are restored. The key idea is that the measures in each quadrant are trinomial, so the probabilities will decay exponentially away from the mean $(p_R L, p_B L)$. This implies that the flow between quadrants occurs at rate $\exp(-cL)$, which is much smaller than the time, $O(1)$, it takes the probability distributions to reach equilibrium. In words, the process comes to equilibrium on a fast time scale while the parameters change on a much slower one. We will prove this separation of time scales in a version of the paper for a mathematical audience. Here, we will only give the answers that result under

Assumption 2. *The process is always in one of self-consistent stationary distributions, but the value of a changes over time.*

In Section 7, we use the stability results to show that the only possible stable equilibria the end points:

$a = 0$ which represents a random distribution,

$a = 1/2$ which represents a segregated state.

We will call these measures μ_r and μ_s when $\rho < \rho_c$, $\hat{\mu}_r$ and $\hat{\mu}_s$ when $\rho > \rho_c$.

6 Self-consistent stationary distributions

The results given here are proved in Sections 4 and 5 of the Supplementary Materials. The formulas, which again come from solving a quadratic equation, are ugly but they are explicit and easily evaluated.

Theorem 2A. *Suppose $\rho < \rho_c$. For $a \in (0, 1/2]$ let*

$$\rho_1(a,\rho) = \frac{-1 + (a+\rho)(1-\epsilon) + \sqrt{[1-(a+\rho)(1-\epsilon)]^2 + 4a(1-\epsilon^2)\rho}}{2a(1-\epsilon^2)}. \tag{3}$$ rho1a

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

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

A symmetric distribution μ 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. As shown in Section 7 of the Supplementary Materials $a \to \rho_1(a,\rho)$ is increasing, so the range of possible values of ρ_1 for a fixed value of ρ is

$$[\rho_1(0,\rho), \rho_1(1/2,\rho)] = \left[\frac{\rho}{1-\rho(1-\epsilon)}, \frac{2\rho}{1+\epsilon}\right] \tag{4}$$ intlo

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

Theorem 2B. *Suppose $\rho \geq \rho_c$. For $a \in (0, 1/2]$ let*

$$\hat{\rho}_1(a,\rho) = \frac{\epsilon + (1-\epsilon)(a+\rho) - \sqrt{[\epsilon + (1-\epsilon)(a+\rho)]^2 - 4a(1-\epsilon^2)\rho}}{2a(1-\epsilon^2)}. \tag{5}$$ barrho1a

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

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

A symmetric 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 $\hat{\rho}_3 = \epsilon\hat{\rho}_1/[1 - (1-\epsilon)\hat{\rho}_1] > \rho_c$.

This time $a \to \rho_1(a,\rho)$ is decreasing, so the range of possible values of ρ_1 for a fixed value of ρ is

$$[\hat{\rho}_1(1/2,\rho), \hat{\rho}_1(0,\rho)] = \left[\frac{2\rho}{1+\epsilon}, \frac{\rho}{\epsilon+(1-\epsilon)\rho}\right] \tag{6}$$ inthi

i.e., the old upper bound on the range of ρ_1 in (4) has become the lower bound. See Figure 3 for a picture. There the two curves are $\rho_1(0,\rho)$ for $\rho < \rho_c$ and $\hat{\rho}_1(0,\rho)$ for $\rho \geq \rho_c$, while the straight line is $\rho_1(1/2,\rho) = \hat{\rho}_1(1/2,\rho) = 2\rho/(1+\epsilon)($.

7 Stability calculations

The results in this section are proved in Sections 7 and 8 of the Supplementary Materials. Using "large deviations" for the trinomial distribution, which in this case is just calculating probabilities using Stirling's formula, we conclude:

Theorem 3A. *Suppose $\rho < \rho_c$ and recall μ_a has no mass on $Q_{1,1}$. The flow into $Q_{0,0}$ from $Q_{0,1}$ and $Q_{1,0}$ is larger than the flow out if and only if*

$$\left(\frac{1-\epsilon\rho_1}{1-\rho_1}\right)^{1-\rho_c} < 1 + (1-\epsilon)\rho_1. \tag{7}$$ fsp

Theorem 3B. *Suppose $\rho \geq \rho_c$ and recall $\hat{\mu}_a$ has no mass on $Q_{0,0}$. 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*

$$\left(\frac{\hat{\rho}_1}{1-\hat{\rho}_1}\right)^{1-\rho_c} < (1-(1-\epsilon)\hat{\rho}_1)^{-1}. \tag{8}$$ flp

8 Phase Transition

Combining Theorems 2A and 3A, we can determine the behavior of the process for $\rho < \rho_c$. The set of possible values for $\rho_1(a,\rho)$ for a fixed ρ is the interval $[\rho_1(0,\rho), \rho_1(1/2,\rho)]$ given in (4). Since $0 \leq \rho < \rho_c$, we are looking for a solution to

$$\left(\frac{1-\epsilon x_0}{1-x_0}\right)^{1-\rho_c} = 1 + (1-\epsilon)x_0.$$

with $x_0 \in [0, 2\rho_c/(1+\epsilon))$. In Section 9 of the supplementary materials we show that x_0 exists and is unique. Here, we will concentrate on what happens in the example $\rho_c = 0.2$, $\epsilon = 0.1$. When $\rho_c = 0.2$ the intervals is $[0, 0.4/1.1]$ and we have $x_0 = 0.2183$.

Let ρ_b be chosen so that $x_0 = \rho_1(1/2,\rho_b)$ and ρ_d be chosen so that $x_0 = \rho_1(0,\rho_d)$. See Figure 3 for a picture. When a solution x_0 exists in the desired interval

$$\rho_b = \frac{(1+\epsilon)x_0}{2} \quad \text{and} \quad \rho_d = \frac{x_0}{1+x_0(1-\epsilon)}.$$

In our special case, $\rho_b = 0.1201$ and $\rho_d = 0.1825$.

Theorem 4A. *The stable stationary distribution for $\rho < \rho_c$ are*

$$\begin{aligned} \mu_r \quad & \text{for } 0 \geq \rho < \rho_b, \\ \mu_r \text{ and } \mu_s \quad & \text{for } \rho_b < \rho < \rho_d, \\ \mu_s \quad & \text{for } \rho_d < \rho < \rho_c. \end{aligned}$$

Why is this true? When $\rho < \rho_b$, $\rho_0 > \rho_1(1/2, \rho)$, so the flow into $Q_{0,0}$ is always larger than the flow out, so μ_r is the stationary distribution. When $\rho_b < \rho < \rho_d$ there will be an $a_c \in (0, 1/2)$ so that $\rho_1(a_c, \rho) = \rho_0$. The flow into $Q_{0,0}$ is larger than the flow out when $a < a_c$ and the a in the mixture will decrease, while for $a > a_c$ the flow out of $Q_{0,0}$ will be larger than the flow in and a will increase. Thus we have bistability. When $\rho_d < \rho < \rho_c$, $\rho < \rho_1(0, \rho)$, the flow out of $Q_{0,0}$ is always larger than the flow in, and the segregated fixed point with $a = 1/2$ is the stationary distribution. □

Using Theorems 2B and 3B, we can determine the behavior of the process for $\rho \geq \rho_c$. The set of possible values for $\hat{\rho}_1(a, \rho)$ for a fixed ρ is the interval $[\hat{\rho}_1(1/2, \rho), \hat{\rho}_1(0, \rho)]$ given in (6). Since $\rho_c \leq \rho \leq 0.5$, we are looking for a solution to

$$\left(\frac{\hat{x}_0}{1 - \hat{x}_0}\right)^{1-\rho_c} = (1 - (1 - \epsilon)\hat{x}_0)^{-1}.$$

with in $\hat{x}_0 \in [2\rho_c/(1 + \epsilon), 1/1 + \epsilon]$. In Section 10 of the supplementary materials we show that there is a solution in the desired interval if and only if $\epsilon^{\rho_c} \geq (\epsilon + 1)/2$, and it is unique. The condition comes from having $=$ in (8) at the right end point $1/(1 + \epsilon)$. When $\epsilon = 0.1$ the condition is $\rho_c < 0.25964$

When $\rho_c = 0.2$ and $\epsilon = 0.1$, this interval is $[0.4/1.1, 1/1.1]$, and $\hat{x}_0 = 0.8724$. Let $\hat{\rho}_b$ be chosen so that $\hat{x}_0 = \hat{\rho}_1(0, \hat{\rho}_b)$ and $\hat{\rho}_d$ be chosen so that $\hat{x}_0 = \hat{\rho}_1(1/2, \rho_d)$. When a solution $\hat{x}_0$ exists in the desired interval, we have

$$\rho_d = \frac{\epsilon \hat{x}_0}{1 - \hat{x}_0(1 - \epsilon)} \quad \text{and} \quad \rho_d = \frac{(1 + \epsilon)x_0}{2}.$$

In our example, $\hat{\rho}_b = 0.4061$ and $\hat{\rho}_d = 0.4798$.

Theorem 4B. *The stable stationary distribution for $\rho \geq \rho_c$ are*

$$\begin{aligned} \hat{\mu}_s \quad & \text{for } \rho_c \leq \rho < \hat{\rho}_b, \\ \hat{\mu}_r \text{ and } \hat{\mu}_s \quad & \text{for } \hat{\rho}_b < \rho < \hat{\rho}_d, \\ \hat{\mu}_r \quad & \text{for } \hat{\rho}_d < \rho < 0.5. \end{aligned}$$

The reasoning behind this result is the same as for Theorem 4A. To show that these result can be used to explicitly describe the phase transition, Figure 4 shows how the four critical values depend on ρ_c when $\epsilon = 0.1$.

9 Discussion

Here, we have considered a metapopulation version of Schelling's model, which we believe is 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 structure of the model, we are able to describe the phase transition in great detail. For $\rho < \rho_b$ a random distribution of families $\mu_r = Tri(\rho, \rho)$ is the unique stationary distribution. As ρ increases there is a discontinuous phase transition to a segregated state μ_s at ρ_d preceded by an interval (ρ_b, ρ_d) in which both μ_r and μ_s are stable. Surprisingly the phase transition occurs to a segregated state occurs at a value $\rho_d < \rho_c$, i.e., at a point where in a random distribution most families are happy. This shift in hbehavior occurs because random fluctuations create segregated neighborhoods, which, as the analysis in Section 7 shows, are more stable than the random ones.

If ρ_c is small enough then as ρ nears $1/2$, there is another discontinuous transition at $\hat{\rho}_d$ which returns the equilibrium to the random state $\hat{\mu}_r = Tri(\rho, \rho)$, and this is preceded by an interval $(\hat{\rho}_b, \hat{\rho}_d)$ of bistability. To explain this, 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 fraction of vacant houses at $\hat{\rho}_d = 0.4798$ only 4.04%, so it is very difficult to make segregated neighborhoods where one type is happy. The stability analysis in Section 7 implies that these segregated neighborhoods are created at a slower rate than they are lost, so the random state prevails.

The results in this paper has been derived under two assumptions (i) stationary distributions are invariant under interchange of red and blue, and (ii) the process is always in one of a one-parameter family of self-consistent stationary distributions indexed by $a \in [0, 1/2]$, but the value of a changes over time. We are confident that (ii) can be proved rigorously. Removing (i) will be more difficult, since when symmetry is dropped there is a two parameter family of self-consistent distributions. A more interesting problem, which is important for applications to real cities, is to allow the initial densities of reds and blues and their threshold for happiness to differ. While our solution is not yet complete, we believe it is is an important first step in obtaining a detailed understanding of the equilibrium behavior of Schelling's model in a situation that is relevant for applications.

Acknowledgements

Both authors were partially supported by grants DMS 10-05470 and DMS13-05997 from the probability program at NSF. They would like to thanks David Aldous, Nicholas Lanchier, and Simon Levin for helpful comments.

References

Sch1 [1] Schelling, T.C. (1971) Dynamic models of segregation. *J. Mathematical Sociology.* 1, 143–186

Sch2 [2] Schelling, T.C. (1978) *Micromotives and Macrobehavior.* Norton, New York

CF [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

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

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

KPS [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

HatBen [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

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

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

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

DACM [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

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

RogMcK [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

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

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

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

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

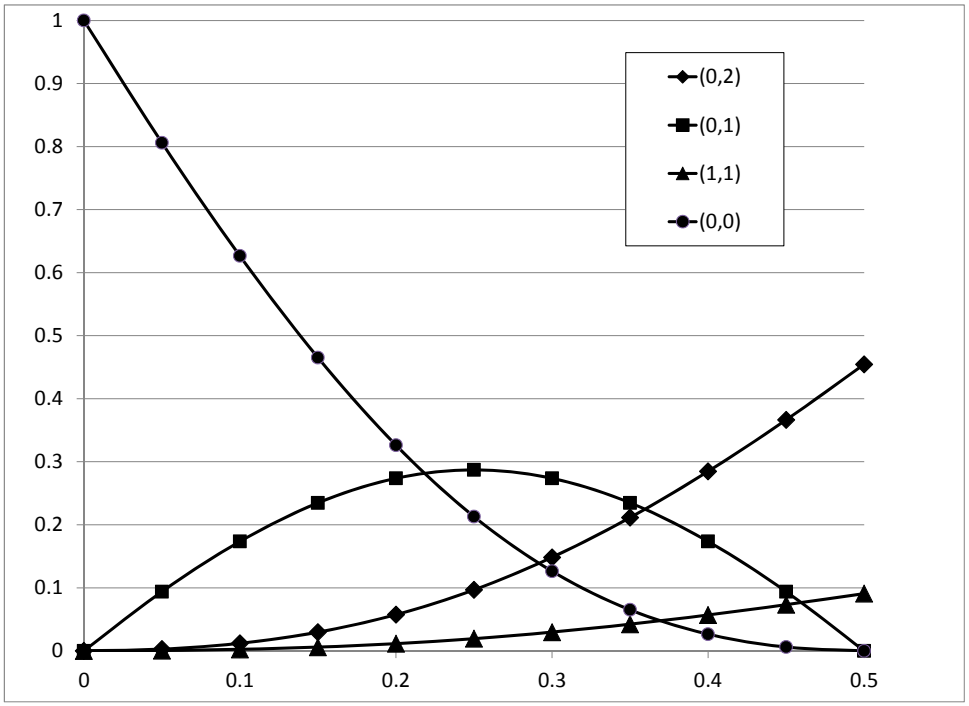

Figure 1: Equilibrium for the case $L = 2$ plotted against ρ.

fig:Figure1

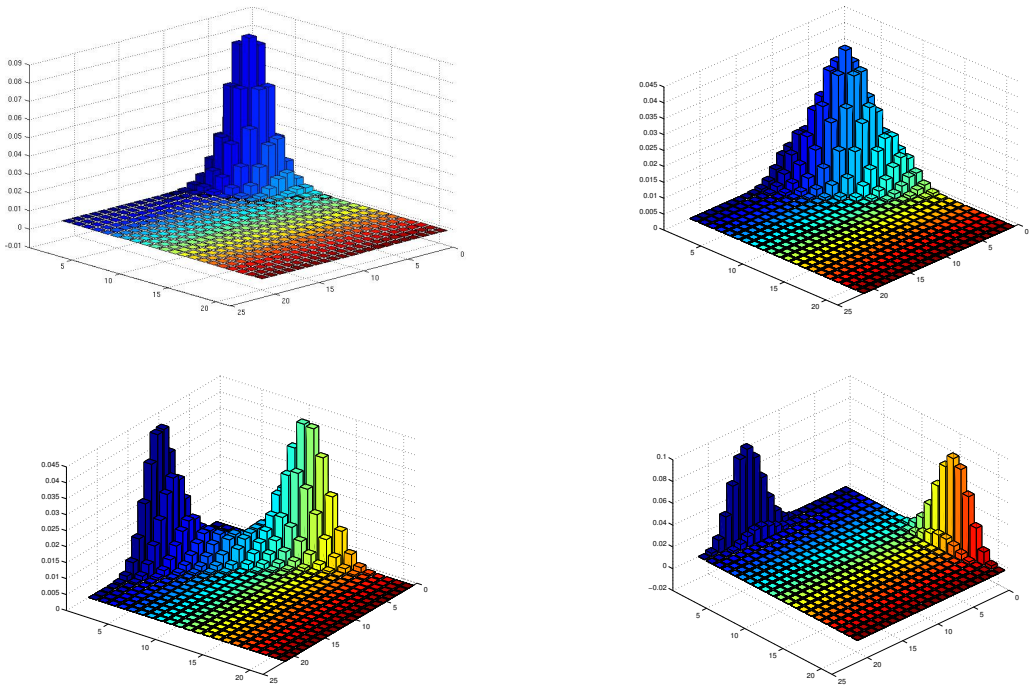

Figure 2: Limiting behavior of limit differential equation, with $\rho_c = 0.3, \epsilon = 0.01$, $\rho = 0.1$, 0.2, 0.25, and 0.35. fig:lim

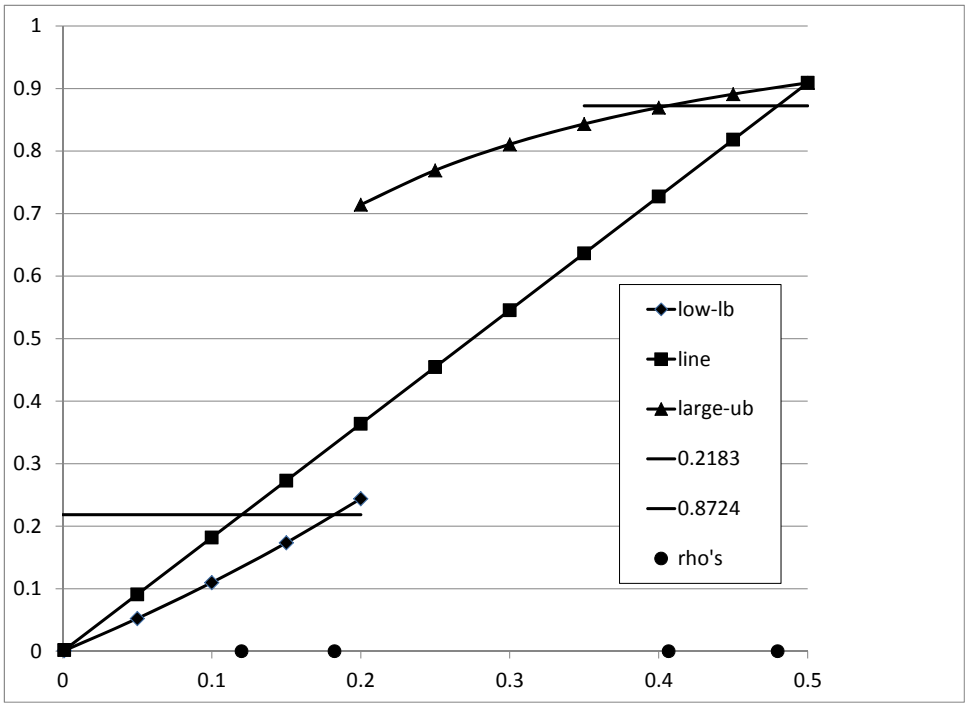

Figure 3: Picture to explain calculation of the phase transition when $\rho_c = 0.2$, $\epsilon = 0.1$. Dots on the axis are the locations of ρ_b, ρ_d, $\hat{\rho}_b$, $\hat{\rho}_d$. fig:Figure3

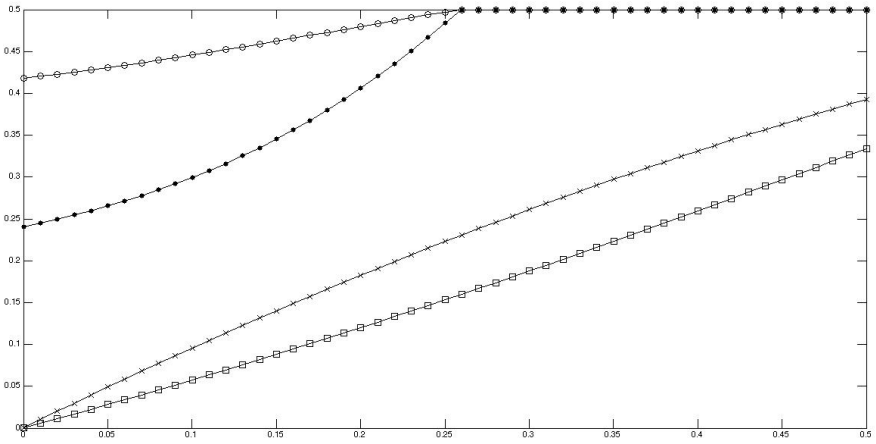

Figure 4: $\rho_b < \rho_d < \hat{\rho}_b < \hat{\rho}_d$ as a function of ρ_c when $\epsilon = 0.1$

fig:Phases

Supplementary Materials for Durrett and Zhang

1 Formulas for Theorem 1

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

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

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. Let $\omega_i = (a_i, b_i)$, $\omega_i' = (a_i', b_i')$

$$b(\omega_1, \omega_2; \omega_1', \omega_2') = \begin{cases} (a_1/L)(L - a_2 - b_2)\Delta(b_1, b_2) & \text{if } \omega_1' = (a_1 - 1, b_1)\omega_2' = (a_2 + 1, b_2) \\ b_1((L - a_2 - b_2)/L)\Delta(b_1, b_2) & \text{if } \omega_1' = (a_1, b_1 - 1)\omega_2' = (a_2, b_2 + 1) \end{cases}$$

and 0 otherwise.

2 Neighborhoods of size 2

To apply Theorem 1, we need to compute

$\lambda(\omega_1, \omega_2) =$

ω_1	$\omega_2 = (0,0)$	$(1,0)$	$(0,1)$
$(1,0)$	r^*	$r/2$	$\epsilon/2$
$(0,1)$	r^*	$\epsilon/2$	$r/2$
$(2,0)$	$2r$	r^*	ϵ
$(0,2)$	$2r$	ϵ	r^*
$(1,1)$	2	(1+q)/2	(1+q)/2

and $b(\omega_1, \omega_2; \cdot) =$

ω_1	$\omega_2 = (0,0)$	$(1,0)$	$(0,1)$
$(1,0)$	$r\mathbf{1}_{(0,0;1,0)}{}^*$	$r\mathbf{1}_{(0,0;2,0)}/2$	$\epsilon\mathbf{1}_{(0,0;1,1)}/2$
$(0,1)$	$r\mathbf{1}_{(0,0;0,1)}{}^*$	$\epsilon\mathbf{1}_{(0,0;0,2)}/2$	$r\mathbf{1}_{(0,0;1,1)}/2$
$(2,0)$	$2r\mathbf{1}_{(1,0;1,0)}$	$r\mathbf{1}_{(1,0;2,0)}{}^*$	$\epsilon\mathbf{1}_{(1,0;1,1)}$
$(0,2)$	$2r\mathbf{1}_{(0,1;0,1)}$	$\epsilon\mathbf{1}_{(0,1;0,2)}$	$r\mathbf{1}_{(0,1;1,1)}{}^*$
$(1,1)$	$\mathbf{1}_{(0,1;1,0)}$	$\mathbf{1}_{(0,1;2,0)}/2$	$\mathbf{1}_{(0,1;1,1)}/2$
	$+\mathbf{1}_{(1,0;0,1)}$	$+q\mathbf{1}_{(1,0;1,1)}/2$	$+q\mathbf{1}_{(1,0;0,2)}/2$

The entries marked with *'s do not change the $n_{a,b}$. Using the first table to see how things are destroyed (terms in square brackets below) and the second table to see how things are created – referring back to the first table for the rates, the limiting ODE is:

$$\begin{aligned}
\frac{d\nu_{0,0}(t)}{dt} &= -\nu_{0,0}[2r\nu_{2,0} + 2\nu_{1,1} + 2r\nu_{0,2}] + \epsilon\nu_{1,0}\nu_{0,1} + [r\nu_{0,1}^2 + r\nu_{1,0}^2]/2 \\
\frac{d\nu_{1,0}(t)}{dt} &= -\nu_{1,0}[r\nu_{1,0} + \epsilon\nu_{0,1} + \nu_{1,1}(1+q)/2 + \epsilon\nu_{0,2}] \\
&\quad + 4r\nu_{0,0}\nu_{2,0} + 2\nu_{1,1}\nu_{0,0} + \epsilon\nu_{2,0}\nu_{0,1} + \nu_{1,1}\nu_{0,1}/2 + \nu_{1,1}\nu_{1,0}q/2 \\
\frac{d\nu_{0,1}(t)}{dt} &= -\nu_{0,1}[r\nu_{0,1} + \epsilon\nu_{1,0} + \nu_{1,1}(1+q)/2 + \epsilon\nu_{2,0}] \\
&\quad + 4r\nu_{0,0}\nu_{0,2} + 2\nu_{1,1}\nu_{0,0} + \epsilon\nu_{0,2}\nu_{1,0} + \nu_{1,1}\nu_{1,0}/2 + \nu_{1,1}\nu_{0,1}q/2 \\
\frac{d\nu_{2,0}(t)}{dt} &= -\nu_{2,0}[2r\nu_{0,0} + \epsilon\nu_{0,1}] + [r\nu_{1,0}^2 + \nu_{1,0}\nu_{1,1}]/2 \\
\frac{d\nu_{0,2}(t)}{dt} &= -\nu_{0,2}[2r\nu_{0,0} + \epsilon\nu_{1,0}] + [r\nu_{0,1}^2 + \nu_{0,1}\nu_{1,1}]/2 \\
\frac{d\nu_{1,1}(t)}{dt} &= -\nu_{1,1}[2\nu_{0,0} + (\nu_{1,0} + \nu_{0,1})(1+q)/2] \\
&\quad + \epsilon\nu_{0,1}\nu_{1,0} + \epsilon\nu_{0,2}\nu_{1,0} + \epsilon\nu_{2,0}\nu_{0,1} + \nu_{1,1}(\nu_{1,0} + \nu_{0,1})q/2.
\end{aligned}$$

Inspired by the notion of detailed balance we make the following definitions;

$$\begin{array}{ll}
x_1 = r(\nu_{1,0}^2 - 4\nu_{0,0}\nu_{2,0}) & (1,0)(1,0) \rightleftharpoons (0,0)(2,0) \\
x_2 = r(\nu_{0,1}^2 - 4\nu_{0,0}\nu_{0,2}) & (0,1)(0,1) \rightleftharpoons (0,0)(0,2) \\
x_3 = 2\nu_{0,0}\nu_{1,1} - \epsilon\nu_{1,0}\nu_{0,1} & (1,1)(0,0) \rightleftharpoons (1,0)(0,1) \\
x_4 = \nu_{1,0}\nu_{1,1} - 2\epsilon\nu_{2,0}\nu_{0,1} & (1,1)(1,0) \rightleftharpoons (0,1)(2,0) \\
x_5 = \nu_{0,1}\nu_{1,1} - 2\epsilon\nu_{0,2}\nu_{1,0} & (1,0)(1,1) \rightleftharpoons (1,0)(0,2)
\end{array}$$

Introducing these variable into the differential equations. We have that in order to have a fixed point, the following equations has to be satisfied

$$\begin{cases}
\;\; x_1/2 + x_2/2 - x_3 = 0 \\
-x_1 + x_3 - x_4/2 + x_5/2 = 0 \\
-x_2 + x_3 + x_4/2 - x_5/2 = 0 \\
\;\; x_1/2 + x_4/2 = 0 \\
\;\; x_2/2 + x_5/2 = 0 \\
-x_3 - x_4 - x_5 = 0
\end{cases}. \tag{1}$$

Equations in (1) immediately implies that

$$x_1 = -x_4, \quad x_2 = -x_5, \quad x_1 + x_2 = x_3 = 0.$$

Plugging the formula for x_i's into the equations above, we have from $x_1 + x_2 = 0$

$$\nu_{1,0}^2 - 4\nu_{0,0}\nu_{2,0} + \nu_{0,1}^2 - 4\nu_{0,0}\nu_{0,2} = 0$$

which implies

$$\nu_{0,0} = \frac{\nu_{1,0}^2 + \nu_{0,1}^2}{4(\nu_{0,2} + \nu_{2,0})}.$$

And from $x_3 = 0$, we have

$$\nu_{0,0} = \frac{\epsilon \nu_{1,0}\nu_{0,1}}{2\nu_{1,1}}.$$

Combining the two equations above gives us

$$2\epsilon\nu_{1,0}\nu_{0,1}(\nu_{0,2} + \nu_{2,0}) = \nu_{1,1}(\nu_{0,1}^2 + \nu_{1,0}^2). \tag{2}$$

Moreover, from the equation that $x_4 + x_5 = -x_1 - x_2 = 0$, we have

$$2\epsilon(\nu_{2,0}\nu_{0,1} + \nu_{0,2}\nu_{1,0}) = \nu_{1,1}(\nu_{1,0} + \nu_{0,1}). \tag{3}$$

$\nu_{1,0} * (3) - (2)$ gives us:

$$2\epsilon\nu_{0,2}\nu_{1,0}(\nu_{1,0} - \nu_{0,1}) = \nu_{1,1}\nu_{0,1}(\nu_{1,0} - \nu_{0,1})$$

which implies

$$(2\epsilon\nu_{0,2}\nu_{1,0} - \nu_{1,1}\nu_{0,1})(\nu_{1,0} - \nu_{0,1}) = 0 \tag{4}$$

Equation (4) show that we must have at least one of the following two cases:

Case I: $2\epsilon\nu_{0,2}\nu_{1,0} - \nu_{1,1}\nu_{0,1} = 0$. By definition, we have $x_5 = 0$, and thus $x_2 = -x_5 = 0$, $x_1 = -x_2 = 0$ and $x_4 = -x_1 = 0$. Which implies that the fixed point of the ODE satisfies detailed balance, and thus must be the reversible measure.

Case II: $\nu_{1,0} - \nu_{0,1} = 0$. Under this case, note that

$$\nu_{1,0} + \nu_{1,1} + 2\nu_{2,0} = \nu_{0,1} + \nu_{1,1} + 2\nu_{0,2} = \rho.$$

We have $\nu_{2,0} = \nu_{0,2}$ and the fixed point is thus symmetric. This implies

$$x_1 = r(\nu_{1,0}^2 - 4\nu_{0,0}\nu_{2,0}) = x_1 = r(\nu_{0,1}^2 - 4\nu_{0,0}\nu_{0,2}) = x_2.$$

Thus $x_1 = x_2 = (x_1 + x_2)/2 = 0$, $x_4 = -x_1 = 0$ and $x_5 = -x_2 = 0$. So case II is actually the same as case I.

3 Rates for the neighborhood-environment chain

If we let $n_{i,j}$ be the number of (i, j) neighborhoods then those parameters for red families can be written as:

$$h_R^1 = \sum_{j<l_c} n_{i,j} i, \qquad h_R^0 = \sum_{j<l_c} n_{i,j}(L - i - j) \tag{5}$$

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,

$$H_R^+ = [rh_R^1 + (N_R - h_R^1)]/NL \quad H_R^- = [rh_R^0 + (N_0 - h_R^0)]/NL$$
$$U_R^+ = [\epsilon h_R^1 + q(N_R - h_R^1)]/NL \quad U_R^- = [h_R^0 + q(N_0 - h_R^0)]/NL$$

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

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

Using this it is easy to verify that inside $Q_{k,\ell}$, the detailed balance condition is satisfied by $Tri(p_R, p_B)$ where

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

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

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

4 Proof of Theorem 2A

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 2A 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 $\rho_0 < l_c$, $\rho_2 < l_c < \rho_1$. In the next section we will show that there is no mass on $Q_{1,1}$ if and only if $\rho < \rho_c$. Our goal is to show that any self-consistent distribution of this form falls into the one-parameter family described in Theorem 2A. 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

$$r(1-2a)\rho_0 + a(r\rho_1+\rho_2) = [r(1-2a) + 2a(r\rho_1+\rho_2) + (r+\epsilon)a(1-\rho_1-\rho_2)]\rho_0. \qquad (6)$$

Subtracting $r(1-2a)\rho_0$ and then dividing by a on both side of (6), we have

$$\rho_0(r+\epsilon)(1-\rho_1-\rho_2) = (r\rho_1+\rho_2)(1-2\rho_0). \tag{7}$$

This implies

$$1-\rho_1-\rho_2 = \frac{(r\rho_1+\rho_2)(1-2\rho_0)}{\rho_0(r+\epsilon)}. \tag{8}$$

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

$$\frac{\rho_1}{\rho_2} = \frac{(1-2a)(1-2\rho_0)+(1+q)a(1-\rho_1-\rho_2)}{r(1-2a)(1-2\rho_0)+(r+\epsilon)a(1-\rho_1-\rho_2)} \times \frac{r(1-2a)\rho_0+a(r\rho_1+\rho_2)}{\epsilon(1-2a)\rho_0+a(\epsilon\rho_1+q\rho_2)}. \tag{9}$$

Plugging (8) in to (9), we can simplify the equation and get

$$\frac{\rho_1}{\rho_2} = \frac{1}{r+\epsilon} \cdot \frac{(1-2a)\rho_0(r+\epsilon)+(1+q)a(r\rho_1+\rho_2)}{r(1-2a)\rho_0+a(r\rho_1+\rho_2)} \times \frac{r(1-2a)\rho_0+a(r\rho_1+\rho_2)}{\epsilon(1-2a)\rho_0+a(\epsilon\rho_1+q\rho_2)}. \tag{10}$$

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

$$\rho_2[(1-2a)\rho_0(r+\epsilon)+(1+q)a(r\rho_1+\rho_2)] = \rho_1[\epsilon(1-2a)\rho_0(r+\epsilon)+a(r+\epsilon)(\epsilon\rho_1+q\rho_2)]. \tag{11}$$

For further simplification, note that we can rewrite equation (11) 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

$$(\rho_2-\epsilon\rho_1)\cdot[(1-2a)\rho_0(r+\epsilon)+a(r+\epsilon)\rho_1+(1+q)a\rho_2]=0. \tag{12}$$

Since $(1-2a)\rho_0(r+\epsilon)+a(\epsilon+r)\rho_1+(1+q)a\rho_2>0$, (12) implies that

$$\rho_2 = \epsilon\rho_1. \tag{13}$$

Now plugging (13) back into (7), we have $\rho_0(1-(1+\epsilon)\rho_1) = \rho_1(1-2\rho_0)$ and

$$\rho_0 = \frac{\rho_1}{1+(1-\epsilon)\rho_1}. \tag{14}$$

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 (13) and (14):

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

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:

$$\rho_1 = \frac{-1+(a+\rho)(1-\epsilon)+\sqrt{[1-(a+\rho)(1-\epsilon)]^2+4a(1-\epsilon^2)\rho}}{2a(1-\epsilon^2)} \tag{15}$$

and we have proved Theorem 2A.

5 Proof of Theorem 2B

We move now to the case when there is no mass on $Q_{0,0}$. Again we will prove in the next section that this corresponds to $\rho \geq \rho_c$. 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

$$\begin{aligned}(1+q)\hat{\rho}_3(1-\hat{\rho}_1-\hat{\rho}_2) &= (1-2\hat{\rho}_3)(\epsilon\hat{\rho}_1+q\hat{\rho}_2)\\ \Rightarrow(1-\hat{\rho}_1-\hat{\rho}_2) &= \frac{(1-2\hat{\rho}_3)(\epsilon\hat{\rho}_1+q\hat{\rho}_2)}{(1+q)\hat{\rho}_3}.\end{aligned} \tag{16}$$

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

$$\frac{\hat{\rho}_1}{\hat{\rho}_2} = \frac{q(1-2a)(1-2\hat{\rho}_3)+(1+q)a(1-\hat{\rho}_1-\hat{\rho}_2)}{\epsilon(1-2a)(1-2\hat{\rho}_3)+(r+\epsilon)a(1-\hat{\rho}_1-\hat{\rho}_2)} \times \frac{(1-2a)\hat{\rho}_3+a(r\hat{\rho}_1+\hat{\rho}_2)}{q(1-2a)\hat{\rho}_3+a(\epsilon\hat{\rho}_1+q\hat{\rho}_2)}. \tag{17}$$

Then using exactly the same calculation as in the proof of Theorem 2 by plugging (16) into (17), 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:

$$\hat{\rho}_2[(1-2a)(1+q)\hat{\rho}_3+(1+q)a(r\hat{\rho}_1+\hat{\rho}_2)] = \hat{\rho}_1\left(\epsilon(1-2a)(1+q)\hat{\rho}_3+(r+\epsilon)a(\epsilon\hat{\rho}_1+q\hat{\rho}_2)\right) \tag{18}$$

after we cancel the term of $q(1-2a)\hat{\rho}_3 + a(\epsilon\hat{\rho}_1+q\hat{\rho}_2)$. Simplifying (18) with exactly the same procedure as in the proof of Theorem 2, we have

$$(\hat{\rho}_2-\epsilon\hat{\rho}_1)\cdot[(1-2a)(1+q)\hat{\rho}_3+(r+\epsilon)a\hat{\rho}_1+(1+q)a\hat{\rho}_2] = 0. \tag{19}$$

It is clear that the second term in the product on the left side of (19) is positive, which implies:

$$\hat{\rho}_2 = \epsilon\hat{\rho}_1. \tag{20}$$

Using this in (16) 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

$$\hat{\rho}_3 = \frac{\epsilon\hat{\rho}_1}{1-\hat{\rho}_1(1-\epsilon)}. \tag{21}$$

Noting that $a(\hat{\rho}_1+\hat{\rho}_2)+(1-2a)\hat{\rho}_3 = \rho$ and using (20) and (21), we have

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

and 0

$$a(1-\epsilon^2)\hat{\rho}_1^2-[\epsilon+(1-\epsilon)(a+\rho)]\hat{\rho}_1+\rho = 0. \tag{22}$$

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

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

Thus we have

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

However, from equation (22):

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

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

$$\hat{\rho}_1 = \frac{\epsilon + (1-\epsilon)(a+\rho) - \sqrt{[\epsilon + (1-\epsilon)(a+\rho)]^2 - 4a(1-\epsilon^2)\rho}}{2a(1-\epsilon^2)} \tag{23}$$

which completes the proof of Theorem 2B.

6 Density of Self-Consistent Distributions

Our next step is to show that whenever a self-consistent distribution falls into form of Theorem 2A 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 $(1+\epsilon)\rho_1 = \rho_1 + \rho_2 \leq 1$, $2\rho_0 \geq (1+\epsilon)\rho_1 = \rho_1 + \rho_2$. Combine this with the fact that $\rho_0 < \rho_c$, we have $\rho = (1-2a)\rho_0 + a(\rho_1 + \rho_2) \leq \rho_0 < \rho_c$.

Similarly, for self-consistent distribution in Theorem 2B, 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 $(1+\epsilon)\hat{\rho}_1 \geq 2\hat{\rho}_3 > 2\rho_c$, and $\rho = (1-2a)\hat{\rho}_3 + a(\hat{\rho}_1 + \hat{\rho}_2) \geq \rho_c$.

7 Proof of Theorem 3A

To have the formulas at hand we $\rho_1(0,\rho) = \rho/(1-\rho(1-\epsilon))$ while for $a \in (0, 1/2]$

$$\rho_1(a,\rho) = \frac{-1 + (a+\rho)(1-\epsilon) + \sqrt{[1-(a+\rho)(1-\epsilon)]^2 + 4a(1-\epsilon^2)\rho}}{2a(1-\epsilon^2)}. \tag{24}$$

$\mu_a = (1-2a)\text{Tri}(\rho_0,\rho_0) + a\text{Tri}(\rho_1,\rho_2) + a\text{Tri}(\rho_2,\rho_1)$, where

$$\rho_2 = \epsilon\rho_1 < \rho_c \quad \text{and} \quad \rho_0 = \rho_1/[1+(1-\epsilon)\rho_1] < \rho_c.$$

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

$$\rho_1(1/2,\rho) = \frac{2\rho(1-\epsilon)}{(1-\epsilon^2)}. \tag{25}$$

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, so in this case it is increasing.

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

$$\begin{aligned} &= \rho_c^{-\rho_c L}(\theta)^{-\theta L}(1-\rho_c-\theta)^{1-\rho_c-\theta} p_R^{\rho_c L} p_B^{\theta L}(1-p_R-p_B)^{(1-\rho_c-\theta)L} \\ &= \left(\frac{p_R}{\rho}\right)^{\rho L}\left(\frac{p_B}{\theta}\right)^{\theta L}\left(\frac{1-p_R-p_B}{1-\rho-\theta}\right)^{(1-\rho-\theta)L} \end{aligned} \tag{26}$$

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

$$\frac{\theta}{p_B} = \frac{1-\rho_c-\theta}{1-p_R-P_B}, \tag{27}$$

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 (27) in (26), the maximum probability is

$$\left(\frac{p_R}{\rho_c}\right)^{\rho_c L}\left(\frac{1-p_R}{1-\rho_c}\right)^{(1-\rho_c)L}. \tag{28}$$

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 (**??**) 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 (28), 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 of Theorem 3A.

8 Proof of Theorem 3B

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

$$\hat{\rho}_1 = \frac{\epsilon+(1-\epsilon)(a+\rho) - \sqrt{[\epsilon+(1-\epsilon)(a+\rho)]^2 - 4a(1-\epsilon^2)\rho}}{2a(1-\epsilon^2)} \tag{29}$$

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

$$\hat{\rho}_2 = \epsilon\hat{\rho}_1 < \rho_c \quad \text{and} \quad \hat{\rho}_3 = \frac{\epsilon\hat{\rho}_1}{1-(1-\epsilon)\hat{\rho}_1} > \rho_c.$$

The first step is to determine when the $\hat{\rho}_i$ satisfy the desired inequalities. 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 (25), 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 (28). 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

$$\left(\frac{\hat{\rho}_1}{1-\hat{\rho}_1}\right)^{(1-\rho_c)} < (1-(1-\epsilon)\hat{\rho}_1)^{-1}. \tag{30}$$

9 Existence and Uniqueness of x_0 when $\rho < \rho_c$

The first step in describing the phase transition for $\rho < \rho_c$ taken in Section 8 of the paper is to solve

$$\left(\frac{1-\epsilon x_0}{1-x_0}\right)^{(1-\rho_c)} = 1 + (1-\epsilon)x_0,$$

in $(0, 2\rho_c)/(1+\epsilon)$. When $x_0 = 0$ both sides are $= 1$. To look at the behavior near 0 we take log's:

$$(1-\rho_c)\log\left(1 + \frac{(1-\epsilon)x}{1-x}\right) \sim (1-\rho_c)(1-\epsilon)x$$
$$\log(1+(1-\epsilon)x) \sim (1-\epsilon)x$$

so near 0, the $LHS < RHS$.

The RHS is linear. Our next step is to check that the LHS is convex. To do this we rewrite it as

$$\left(\epsilon + \frac{1-\epsilon}{1-x}\right)^{1-\rho_c}$$

and differentiate twice

$$\begin{aligned}\frac{d}{dx} =&(1-\rho_c)\left(\epsilon+\frac{1-\epsilon}{1-x}\right)^{-\rho_c}\cdot\frac{1-\epsilon}{(1-x)^2}\\ \frac{d^2}{dx^2} =&-\rho_c(1-\rho_c)\left(\epsilon+\frac{1-\epsilon}{1-x}\right)^{-\rho_c-1}\cdot\frac{(1-\epsilon)^2}{(1-x)^4}\\ &+(1-\rho_c)\left(\epsilon+\frac{1-\epsilon}{1-x}\right)^{-\rho_c}\cdot\frac{2(1-\epsilon)}{(1-x)^3}\end{aligned}$$

Rearranging we have

$$\frac{d^2}{dx^2}=(1-\rho_c)\left(\epsilon+\frac{1-\epsilon}{1-x}\right)^{-\rho_c-1}\cdot\frac{1-\epsilon}{(1-x)^4}\cdot\left[-\rho_c(1-\epsilon)+\left(\epsilon+\frac{1-\epsilon}{1-x}\right)2(1-x)\right]$$

The quantity in square brackets is $-\rho_c(1-\epsilon)+2(1-x)\epsilon+2(1-\epsilon)>0$ since $\rho_c<1/2$.

Since $LHS=RHS$ at $x=0$ and $LHS<RHS$ for small $x>0$, we now know that there is at most one positive solution. To show that there is a solution in $(0,2\rho_c/(1+\epsilon))$ we evaluate the two functions at the right end point

$$\begin{aligned}RHS&=1+2\rho_c\frac{1-\epsilon}{1+\epsilon}\\ LHS&=\left(\frac{1+\epsilon-2\epsilon\rho_c}{1+\epsilon-2\rho_c}\right)^{1-\rho_c}\end{aligned}$$

To prove that $LHS>RHS$ for any $\rho_c\in[0,1/2)$, we consider the following function

$$g_{\rho_c}(\epsilon)=\left(\frac{1+\epsilon-2\epsilon\rho_c}{1+\epsilon-2\rho_c}\right)^{1-\rho_c}-2\rho_c\frac{1-\epsilon}{1+\epsilon}\tag{31}$$

one can easily see that it suffices to show $g_{\rho_c}(\epsilon)>1$ for all $\epsilon\in[0,1)$. To prove this, we establish:

Lemma 2. *For any $\rho_c\in[0,1/2)$, $g'_{\rho_c}(\epsilon)<0$ for all $\epsilon\in[0,1)$.*

To explain our interest in this lemma, note that it implies for any $\epsilon\in[0,1)$, $\rho_c\in[0,1/2)$,

$$\left(\frac{1+\epsilon-2\epsilon\rho_c}{1+\epsilon-2\rho_c}\right)^{1-\rho_c}-\rho_c\frac{1-\epsilon}{1+\epsilon}=g_{\rho_c}(\epsilon)>g_{\rho_c}(1)=1.$$

Thus, it is enough to prove the Lemma.

Proof. First, one can rewrite function $g_{\rho_c}(\epsilon)$ as

$$g_{\rho_c}(\epsilon)=\left[(1-2\rho_c)+\frac{4\rho_c(1-\rho_c)}{(1-2\rho_c)+\epsilon}\right]^{1-\rho_c}-2\rho_c\left(-1+\frac{2}{1+\epsilon}\right).$$

Taking the derivative we have

$$g'_{\rho_c}(\epsilon)=(1-\rho_c)\left[(1-2\rho_c)+\frac{4\rho_c(1-\rho_c)}{(1-2\rho_c)+\epsilon}\right]^{-\rho_c}\frac{-4\rho_c(1-\rho_c)}{[(1-2\rho_c)+\epsilon]^2}+\frac{4\rho_c}{(1+\epsilon)^2}$$

Returning the expression in square brackets to its original form $[1+(1-2\rho_c\epsilon)]/[(1-2\rho_c)+\epsilon]$ the above becomes

$$= 4\rho_c \left(\frac{1}{(1+\epsilon)^2} - (1-\rho_c)^2[\epsilon + (1-2\rho_c)]^{-2+\rho_c}[1+(1-2\rho_c)\epsilon]^{-\rho_c} \right).$$

Thus to show $g'_{\rho_c}(\epsilon) \leq 0$, we only need to prove that

$$(1-\rho_c)^2[\epsilon + (1-2\rho_c)]^{-2+\rho_c}[1+(1-2\rho_c)\epsilon]^{-\rho_c} \geq \frac{1}{(1+\epsilon)^2}$$

Taking the square root of each side and rearranging, this is equivalent to

$$(1+\epsilon)(1-\rho_c) \geq [\epsilon + (1-2\rho_c)]^{1-\rho_c/2}[1+(1-2\rho_c)\epsilon]^{\rho_c/2}. \tag{32}$$

Again for the RHS of (32), let $a = [\epsilon+(1-2\rho_c)]^{1-\rho_c/2}$, $b = [1+(1-2\rho_c)\epsilon]^{\rho_c/2}$, $p = 1/(1-\rho_c/2)$ and $q = 2/\rho_c$. Noting that $p^{-1} + q^{-1} = 1$. By Young's inequality,

$$\begin{aligned} RHS = ab &\leq \frac{a^p}{p} + \frac{b^q}{q} \\ &= [\epsilon + (1-2\rho_c)](1-\rho_c/2) + [1+(1-2\rho_c)\epsilon]\rho_c/2 \\ &= (1-\rho_c^2)\epsilon + (1-\rho_c)^2. \end{aligned} \tag{33}$$

Using theis identity in (32)

$$\begin{aligned} (1+\epsilon)(1-\rho_c) - (1-\rho_c^2)\epsilon - (1-\rho_c)^2 &= (1-\rho_c)[(1+\epsilon) - \epsilon(1+\rho_c) - (1-\rho_c)] \\ &= \rho_c(1-\rho_c)(1-\epsilon) > 0. \end{aligned} \tag{34}$$

Combining (33) and (34) we now get (32), and complete the proof of this lemma. □

10 Existence and Uniqueness of $\hat{x}_0$ when $\rho \geq \rho_c$

According to Theorem 3B, in order to compare the flows in/out region $Q_{1,1}$, it suffices to consider the sign of the function as follows:

$$f_{\rho_c,\epsilon}(x) = (1-\epsilon)x + \left(\frac{1-x}{x}\right)^{1-\rho_c} - 1. \tag{35}$$

I.e. $f_{\rho_c,\epsilon}(x) > 0$ if and only if inequality (8) in Theorem 3B holds. We have the following theorem

Theorem 1. *Let $A(\rho_c, \epsilon) = \epsilon^{1-\rho_c} + \epsilon^{-\rho_c} - 2$. For any $0 \leq \rho_c \leq 1/2$, $0 \leq \epsilon < 1$, we have:*

- *When $A(\rho_c, \epsilon) > 0$, $f_{\rho_c,\epsilon}(x) > 0$ for all $x \in (0, 1/(1+\epsilon)]$.*
- *When $A(\rho_c, \epsilon) < 0$, there is a $\hat{x}_0$, such that $f_{\rho_c,\epsilon}(x) > 0$ on $x \in (0, \hat{x}_0)$, $f_{\rho_c,\epsilon}(\hat{x}_0) = 0$ and $f_{\rho_c,\epsilon}(x) < 0$ on $x \in (\hat{x}_0, 1/(1+\epsilon)]$.*

Proof. First we show that for any $0 \leq \rho_c \leq 1/2$, $0 \leq \epsilon < 1$, $f'_{\rho_c,\epsilon}(x) < 0$ for all $x \in (0,1)$, i.e $f_{\rho_c,\epsilon}$ is a strictly decreasing function. We have

$$f'_{\rho_c,\epsilon}(x) = (1-\epsilon) - (1-\rho_c)\left(\frac{1-x}{x}\right)^{-\rho_c}\frac{1}{x^2} = (1-\epsilon) - (1-\rho_c)(1-x)^{-\rho_c}x^{-2+\rho_c} \tag{36}$$

Noting that $x^{-1} > 1$ and $\epsilon \geq 0$,

$$(1-\epsilon) - (1-\rho_c)(1-x)^{-\rho_c}x^{-2+\rho_c} < 1 - (1-\rho_c)(1-x)^{-\rho_c}x^{-1+\rho_c}$$

for all $x \in (0,1)$. Thus, to prove $f'_{\rho_c,\epsilon}(x) > 0$, it suffices to show that

$$(1-\rho_c)(1-x)^{-\rho_c}x^{-1+\rho_c} \geq 1 \tag{37}$$

for all $\rho_c \in [0, 1/2]$ and $x \in (0,1)$. First, one can immediately rewrite (37) as

$$(1-\rho_c) \geq (1-x)^{\rho_c}x^{1-\rho_c}. \tag{38}$$

Let $a = (1-x)^{\rho_c}$, $b = x^{1-\rho_c}$, $p = 1/\rho_c$ and $q = 1/(1-\rho_c)$. Noting that $p^{-1} + q^{-1} = 1$, apply Young's inequality to the RHS of (38):

$$\begin{aligned}(1-x)^{\rho_c}x^{1-\rho_c} = ab &\leq \frac{a^p}{p} + \frac{b^q}{q} \\ &= \rho_c(1-x) + (1-\rho_c)x = \rho_c + x - 2\rho_c x.\end{aligned} \tag{39}$$

Then note that

$$\begin{aligned}(1-\rho_c) - (\rho_c + x - 2\rho_c x) &= 1 - 2\rho_c - x + 2\rho_c x \\ &= (1-2\rho_c)(1-x) \geq 0.\end{aligned} \tag{40}$$

Combining (39) and (40) gives us the desired inequality (38) and (37). Thus $f'_{\rho_c,\epsilon}(x) < 0$ for all $x \in (0,1)$. The fact that $f_{\rho_c,\epsilon}$ is a strictly decreasing function immediately implies that

- When $f_{\rho_c,\epsilon}(1/(1+\epsilon)) > 0$, $f_{\rho_c,\epsilon}(x) > 0$ for all $x \in (0, 1/(1+\epsilon)]$.
- When $f_{\rho_c,\epsilon}(1/(1+\epsilon)) < 0$, there is a $\hat{x}_0$, such that $f_{\rho_c,\epsilon}(x) > 0$ on $x \in (0, \hat{x}_0)$, $f_{\rho_c,\epsilon}(\hat{x}_0) = 0$ and $f_{\rho_c,\epsilon}(x) < 0$ on $x \in (\hat{x}_0, 1/(1+\epsilon)]$.

Then note that

$$f_{\rho_c,\epsilon}(1/(1+\epsilon)) = \frac{1-\epsilon}{1+\epsilon} + \epsilon^{1-\rho_c} - 1 = \frac{\epsilon}{1+\epsilon}A(\rho_c,\epsilon).$$

Thus $A(\rho_c,\epsilon) > 0$ if and only if $f_{\rho_c,\epsilon}(1/(1+\epsilon)) > 0$, and the proof is complete.

□