Back to the future: a simple solution to schelling segregation

Sylvain Barde

OFCE

Document de travail

BACK TO THE FUTURE: A SIMPLE SOLUTION TO SCHELLING SEGREGATION

N° 2011-05

MARCH 2011

Sylvain Barde

Affilié-OFCE,

School of Economics, Keynes College, University of Kent

Observatoire Français des Conjonctures Économiques

69, Quai d'Orsay 75340 Paris Cedex 07

Tel : 01 44 18 54 00 Fax : 01 45 56 06 15

E-mail: ofce@ofce.sciences-po.fr Web: http://www.ofce.sciences-po.fr

Back to the Future: A Simple Solution to Schelling Segregation*

Sylvain Barde[a,b]

[a] *School of Economics, Keynes College, University of Kent, Canterbury, CT2 7NP, UK. tel : +44 (0)1 227 824 092, email: s.barde@kent.ac.uk*
[b] *affiliate, Observatoire Français des Conjonctures Economiques*

February 2011

Abstract

The maximum entropy methodology is applied to the Schelling model of urban segregation in order to obtain a reliable prediction of the stable configuration of the system without resorting to numerical simulations. We show that this approach also provides an implicit equation describing the distribution of agents over a city which allows for directly assessing the effect of model parameters on the solution. Finally, we discuss the information theoretic motivation for applying this methodology to the Schelling model, and show that it effectively rests on the presence of a potential function, suggesting a broader applicability of the methodology.

JEL classification: C11, C63, D80, J15.

Keywords: Information theoretic measure, potential function, Schelling segregation model.

1 Introduction

The Maximum Entropy (MaxEnt) approach developed by Jaynes (1957a,b) is a generic methodology that can be applied to predicting probability distributions in situations where very little information is available. This is possible because the objective function used is the ignorance or uncertainty of an external observer as to the exact state of a system, measured by the Shannon entropy of a message revealing this unobserved state. This was initially integrated in economics through Theil (1967), but more recently Foley

*The author wishes to thank seminar participants at GREQAM for suggesting this application of the MaxEnt methodology and is grateful in particular to Jagjit Chadha for helpful suggestions, and to Sonia Moulet for allowing me to bounce my ideas off her head. Any remaining errors are the author's.

(1994) and Toda (2010) have used the approach to prove the existence of a statistical market equilibrium when agents have "offer sets" of transactions they are willing to accept and interact in a random fashion.

Barde (2011) further investigates the economic applications of this methodology using a simple allocation problem where the preferences of agents are unobservable. Information-theoretic analysis shows this approach provides the only consistent prediction the observer can make of the endowment distributions, as any other prediction violates the endowment constraint. A central economic aspect the fact that under the standard economic assumptions of monotonic and convex preferences, the allocation problem is in fact equivalent a congestion game and therefore possess a potential function.[1] This implies that if these assumptions are satisfied the allocation problem possesses the finite improvement property (FIP), and all myopic improvement paths end in an equilibrium. This allows the ignorant observer to be sure that the predicted aggregate endowment distribution corresponds to an underlying optimal equilibrium.

This paper serves a double purpose, centered around exploring the usefulness of this methodology in obtaining both analytical and empirical results for simulation-based models. The first is to provide an example of a practical application of the MaxEnt prediction approach in economics, by showing how it can be used to predict the outcome of simulations of a simple agent-based model. The second deeper purpose is to illustrate the key role of the potential function as the link between the optimal behaviour at the agent level and the aggregate description provided by the information-theoretic methodology.

These two objectives motivate the use of the model of urban segregation developed by Schelling (1969, 1971). As pointed out by Blume (1997) and Durlauf (1997) the Schelling model is the earliest and simplest example of an agent based model with local interaction. Furthermore, recent work on the Schelling model strongly suggests the existence a potential function. For example, in the physical analog to the Schelling model proposed by Vinkovic and Kirman (2006) particles on a lattice systematically rearrange themselves to

[1]This is defined by Monderer and Shapely (1996) as a unique function defined such that changes in the potential function across states of the system correspond to changes in the objective functions of individual agents in the system.

reduce the internal energy of their configuration. This provides a setting in which a single aggregate function, the overall energy of the system, provides the potential function. Further analysis of the model by Grauwin, Goffette-Nagot, and Jensen (2009) shows that it possess a potential function when bounded neighbourhoods are used. The simplicity of the framework and the presence of a potential function for the model make the Schelling model ideally suited as a test bed for the methodology.

The remainder of the paper is organised as follows. Section 2 presents the version of the Schelling model used in the paper, while section 3 presents the information theoretic methodology and the results obtained. Finally, section 4 discusses the findings and concludes.

2 The Schelling model of segregation

In the standard setting of the Schelling model two types of agents live in a city made up of discrete locations, and each type has a slight preference for living in a neighbourhood composed of agents of the same type. When agents are allowed to move, segregated neighborhoods will emerge from an integrated initial condition as agents relocate to unoccupied locations in the city that are more attractive. In practice this is shown by simulation, using a random initial distribution of agents and where opportunities to move arrive as a Poisson process for each location.

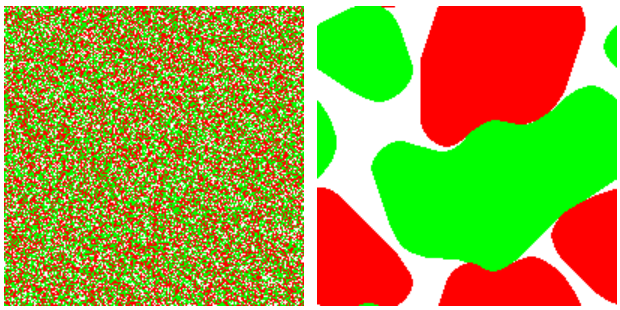
The attractiveness of a location to an agent is a function of the number of similar agents in the vicinity, usually determined by a convolution between the city and a neighbourhood of given width. If B is a $N \times N$ binary matrix which identifies the neighbours for all N locations, and p_j^c is the probability of a c-type agent living in location j this similarity is given by:

$$\left(B \times p^c\right)_i = \sum\nolimits_j B_{i,j} p_j^c \ . \tag{1}$$

Because movement opportunities arrive randomly, simulation is usually the method of

choice for investigating this model. Grauwin, Goffette-Nagot, and Jensen (2009) themselves point out that most analyses of this model rely on agent-based simulations and lack analytical solutions. A simulation is therefore provided as a point of reference for the prediction methodology presented below.

The parameters for the benchmark simulation are as follows: the city is 200 pixels across and each coloured pixel represents a location, so there are $N = 200^2 = 40000$ locations. There are 16000 red and green agents and 8000 free spaces. [2] The neighborhood agents consider when assessing the desirability of a given location is a 7×7 square area centered on that location. As a simplification here, the utility of an agent is directly given by the number of similar neighbors.[3] The random initial condition is represented in Figure 1 (a), while Figure 1 (b) represents the state of the city after 44841 individual moves have occurred. The final state in 1 (b), which exhibits the segregated outcomes typical of the Schelling model, is stable as no further utility-improving relocations exist.



(a) initial state (b) final state

Figure 1: **Initial and final state**

A breakdown of this process is shown in Figure 2. The sequence of images (a)→(i) shows the gradual emergence of the segregated equilibrium following migration away from the initial condition.

[2] As is the case with the work of Grauwin, Goffette-Nagot, and Jensen (2009), the space occupied by the city is toroidal, so that the top/bottom and left/right edges are in contact. This simplification allows the neighbourhood matrix B to be encoded as a Toeplitz matrix.

[3] This simplification does not change the general properties of the methodology presented here, as Grauwin, Goffette-Nagot, and Jensen (2009) show that the existence of the potential function does not depend on the specification of the utility function.

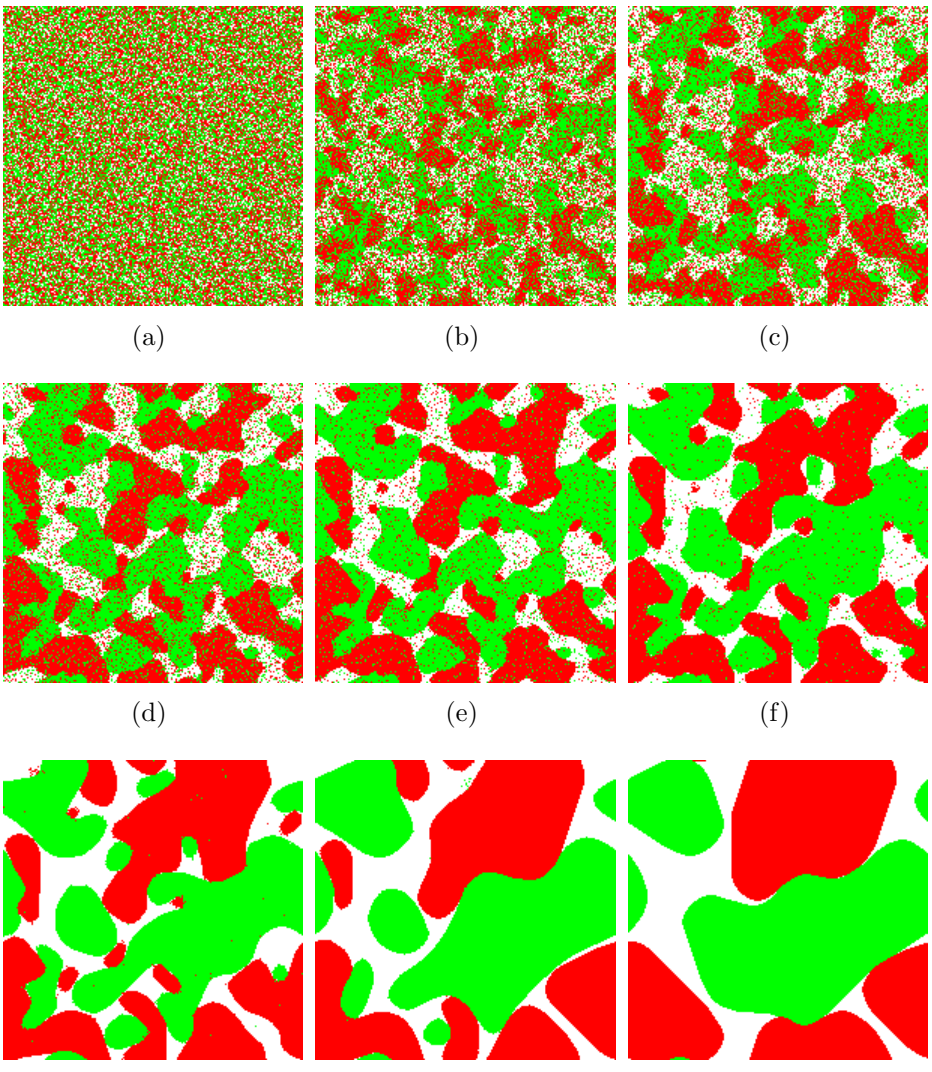


Figure 2: **Emergence in the Schelling model**

3 Information-theoretic prediction

The Schelling model provides a setting where an observer with knowledge of the the initial condition is ignorant of the final state of the system. As discussed in Barde (2011), such a setting motivates the use of MaxEnt as an information-theoretic prediction methodology. More specifically, running the sequence of images in Figure 2 backwards, (i)→(a), one has a situation where a well defined and coherent image gradually becomes more and more noisy and decays until most of the information content has disappeared.[4] An observer receiving the decayed image (a) will be practically ignorant of the original image (i), and the extent of this ignorance can be modeled using the Shannon (1948) entropy as an information measure.

The use of MaxEnt in addressing the problem of reconstructing a "clean" underlying image from initially noisy and distorted data has a long history in image processing and astronomy, where the noise process involved in measurement is similar to the (i)→(a) sequence of Figure 2.[5] In fact, the specific algorithm used to obtain the predictions is adapted from an application for astronomic data suggested in Cornwell and Evans (1985).

Within the setting described in section 2, p_i^c is the probability that the i^{th} location is occupied by and agent of the c^{th} colour, with $c \in \{R, G, W\}$. Given this, the information content of a message revealing the state of the i^{th} location is given by the Shannon (1948) information measure:

$$H(i) = -\sum\nolimits_c p_i^c \ln p_i^c$$

The image processing literature mentioned above prefers relative entropy, as this allows the integration of prior information in the form of a model m_i^c. In practice, one can see below that this results in subtracting the expected information content provided by the model from the overall information content of the message.

[4] This is in fact analogous to the framework of Foley (1994) and Toda (2010) where agent preferences are known and the ignorance relates to sequence of trades made by agents.

[5] Narayan and Nityananda (1986) and Skilling and Gull (1991) provide good introductions of the use of this methodology

$$H\left(i \| \, m_i\right) = -\sum\nolimits_c p_i^c \ln\left(\frac{p_i^c}{m_i^c}\right) = -\sum\nolimits_c p_i^c \ln p_i^c + \sum\nolimits_c p_i^c \ln m_i^c \tag{2}$$

As pointed out in Barde (2011), the attractiveness of the maximum entropy methodology as developed by Jaynes (1957a) is that the ignorance of the observer concerning the distribution of agents over the city can be constrained by integrating available information into the problem. Equation (2) shows that the first piece of information to be included is the underlying model, encoding the prior knowledge that the observer has of the location of agents in the city. The Schelling model, however, does not provide any prior information regarding the probability of a location being occupied by a particular type of agent, therefore the model m_i^c in expression (2) is not particularly useful. This is dealt with by following Skilling and Gull (1987) and considering the expected information content of a message revealing the state $\{c, d\}$ of two randomly picked locations $\{i, j\}$.[6] Using expression (3) enables the integration of a two-dimensional model $m_{i,j}^{c,d}$ which can contain knowledge of correlations across locations.[7] This is better suited to the prior information provided by the Schelling model, in which one expects neighbouring locations to have a relatively high probability of being occupied by similar agents.

$$E\left[H\left(i, j \| \, m_{i,j}\right)\right] = -\frac{2}{N}\sum\nolimits_i \sum\nolimits_c p_i^c \ln p_i^c + \frac{1}{N^2}\sum\nolimits_{i,j}\sum\nolimits_{c,d} p_i^c p_j^d \ln m_{i,j}^{c,d} \tag{3}$$

The second element known to the observer is that number of agents of each type stays constant, such that the average probability of a given colour must match the share of locations s^c that are of the c^{th} colour. This expression serves to normalise the probabilities.

$$\forall c, \quad \sum\nolimits_i \frac{p_i^c}{N} = s^c \tag{4}$$

Finally, the most important piece of information available to the observer is the initial condition of the system. Using the image-processing representation of a known state in

[6] The derivation of the double entropy specification is detailed in appendix A.

[7] This structure also allows correlations across agent types, for example if agents were to evaluate the attractiveness of a location not only by the number of similar agents but also by the number of agents of a different type. This is not the case here as in the basic Schelling model, agents only consider their own type in their location decision. In other words, the model structure (8) will impose $m_{i,j}^{c,d} = 0 \quad \forall d \neq c$.

Figure 2 (i) decaying to a noisy state in Figure 2 (a), this represents the information that has not been wiped out in the decayed image. Within the Schelling framework, this represents the key stable locations that are initially most attractive and are not modified as the segregated outcome emerges. This information is revealed by taking the convolutions of the initial state in order to determine the initial attractiveness (1) for each type of population. This is visible in Figure 3.

As is standard in the image processing literature, this information is integrated into the problem by constraining the noise level, measured by the chi-squared deviation between the initial data available and the prediction. While the historical literature on image processing suggests constraining it to the number of locations N, expression (5) below follows the suggestion of Skilling and Gull (1991) and constrains it to the number of locations N minus Γ^c, the number of good locations in the initial data,[8]

$$\forall c, \quad \frac{\left(\chi^2\right)^c}{N} = \frac{\left(N - \Gamma^c\right)}{N} \ . \tag{5}$$

The chi-squared deviation itself is given by the following expression, where d_i^c represents the initial attractiveness data obtained by taking the convolution of the initial condition and $(\sigma^c)^2$ the variance of this data,

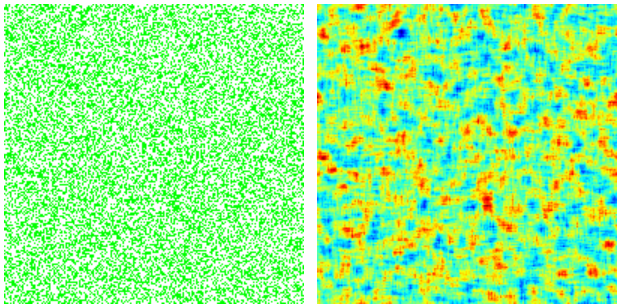
$$\left(\chi^2\right)^c = \sum\nolimits_i \frac{\left(\left(B * p^c\right)_i - d_i^c\right)^2}{\left(\sigma^c\right)^2} \ . \tag{6}$$

The the information theoretic problem is therefore to maximise the ignorance of an observer (3) subject to the known information provided by (4) and (5). The first order condition of the problem directly provides the best prediction of the probability distributions over the locations,

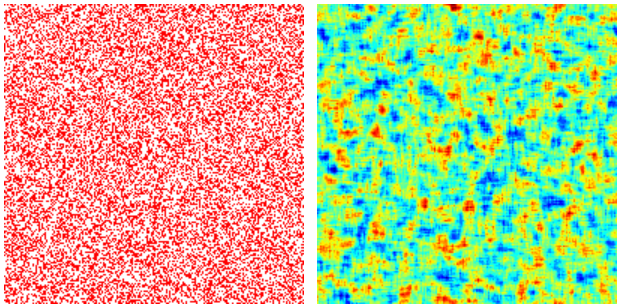
$$p_i^c = \frac{\mu_i^c e^{\alpha^c \frac{\partial\left(\chi^2\right)^c}{\partial p_i^c}}}{Z^c} \ . \tag{7}$$

The effective model μ_i^c and the normalisation parameter Z^c are given by:

[8]Equation (5) expresses this as a percentage of noisy locations. The calculation of Γ^c is explained in appendix B.



(a) initial green state (b) initial green convolution



(c) initial red state (d) initial red convolution

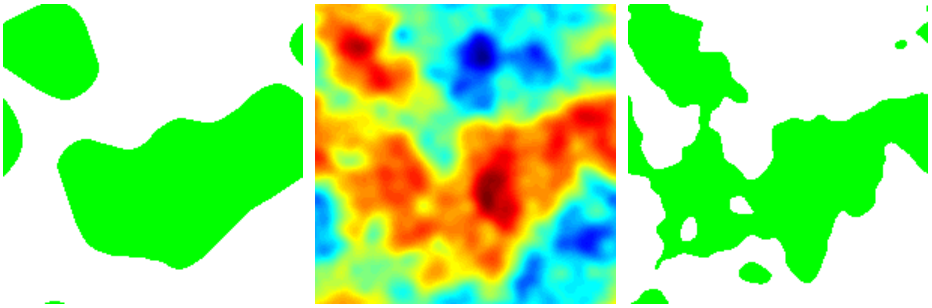
Figure 3: **Initial condition information**

$$\mu_i^c = \exp\left(\frac{1}{N}\sum\nolimits_j \sum\nolimits_{c,d} p_j^d \ln m_{i,j}^{c,d}\right) \quad \text{and} \quad Z^c = \frac{1}{Ns^c}\sum\nolimits_i \mu_i^c e^{\alpha^c \frac{\partial\left(\chi^2\right)^c}{\partial p_i^c}} \ .$$

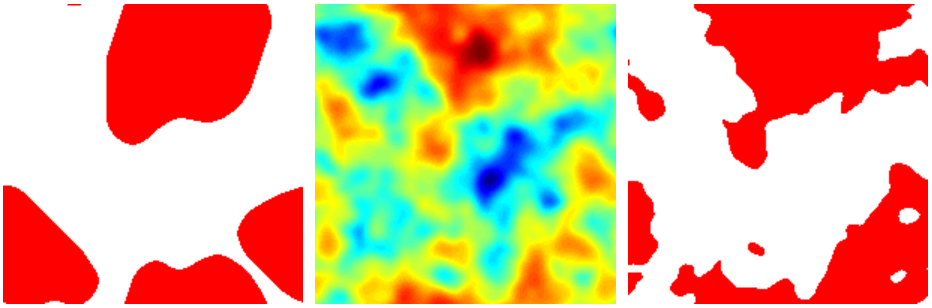
One can see that the effective model for a location μ_i^c is simply the geometric mean of the individual correlations $m_{i,j}$, weighted by the probability vector. As pointed out by Skilling and Gull (1987), this is effectively a convolution of the probability vector p^c with the logarithm of the $N \times N$ model matrix, similar to (1). As for Skilling and Gull (1991), however, the convolution used in the prediction algorithm is slightly different: instead of taking the geometric mean of the model weighted by the probabilities, the algorithm uses the geometric mean of the probabilities weighted by the normalised model,[9]

$$\mu^c = \exp\left(M^c \times \ln p^c\right) \ . \tag{8}$$

[9]This is done for computational reasons. Most of the entries in the model M are very small as one expects the correlations across locations to exist only over short distances. As a result they are truncated out of the matrix, which can be stored as a sparse matrix with many zero elements. Taking the logarithm of this $N \times N$ matrix is cumbersome, therefore in practice it is easier to take the logarithm of the $N \times 1$ vector of probabilities



(a) Final green state (b) Green probability density (c) Predicted green locations



(d) Final red state (e) Red probability density (f) Predicted red locations

Figure 4: **Predictions**

Figure 4 provides the result of the MaxEnt prediction, where (a) and (d) are the agent-specific results of the simulation in Figure 1. It is important to point out that expression (7) only provides an implicit solution for the probability distribution p_i^c as both the model term μ_i^c and noise term $(\chi^2)^c$ are themselves functions of p_i^c. The predicted distributions in Figure 4 (b) and (e) are therefore obtained using a gradient-based algorithm, outlined in appendix B. Figures 4 (c) and (f) are the locations where $p_i^G - p_i^R$ and $p_i^R - p_i^G$ are largest respectively, truncated to match the number of agents of each type. Crucially, even though the initial information available in Figure 3 is limited and very noisy, comparing Figures 4 (a),(c),(d) and (f) suggests that the MaxEnt approach nevertheless provides a reliable prediction for the final location of both types of population.

4 Discussion and Conclusion

The effectiveness of the MaxEnt approach in obtaining a prediction for the Schelling model provides a strong justification for its use. A further motivation is the fact that it also

provides an analytical expression for the predicted probability of a location being occupied by a given type of agent. While this provides only an implicit solution to the problem, it can nevertheless be used to examine directly the impact of changing the parameters of the model. Furthermore, by examining what it is that makes this methodology successful it is possible to make a broader point justifying its extension to other economic models.

The MaxEnt methodology provides an information-theoretically consistent reconstruction of a signal that has been distorted and/or contaminated by noise. In the traditional astronomy or image processing applications the typical sequence of events is one where the phenomenon observed is well defined but because of problems during measurement or transmission the information received subsequently is incomplete. Formally, however, the methodology does not rest on the direction of time but on the direction of increasing entropy. Its ability to successfully predict the outcomes of the Schelling model stems directly from the link between Shannon entropy and the concept of a potential function shown in Barde (2011). A core tenant of economic systems is that optimality increases as time goes by because agents only carry out welfare-increasing transactions. The presence of a potential function in a system is simply a strong statement of this fact, as Monderer and Shapely (1996) show that this implies the FIP, where transitions on the path bring systematic increments to the potential function as the system self-organises.

In a system with a potential function, reversing the sequence of steps on the improvement path, i.e. starting at the optimal final solution and finishing at the initial condition, provides a situation where each for step an agent shifts from an optimal to a random sub-optimal situation. This is analogous to the concept of signal decay mentioned above and is illustrated in the inverse sequence of the Schelling process in Figure 2, where one can effectively treat the final state (i) as a hypothetical data file and the initial condition (a) as the noisy/decayed version of (i). In this situation, the MaxEnt prediction of the end point of the improvement path provides is in fact a reconstruction of the origin point of the reversed improvement path.

As was discussed in Barde (2011) the concept of improvement paths is one that is fundamental to economics, as simple allocation problems have a potential function under

relatively innocuous assumptions on preferences. The broader suggestion stemming from this result is therefore that the methodology outlined here should consistently predict the aggregate properties of many economic systems.

References

Barde, S. (2011): "Ignorance is bliss: rationality, information and equilibrium," *School of Economics Discussion Paper*, 11(03).

Blume, L. E. (1997): *The Economic System as an Evolving Complex System II*chap. Population Games. Addison-Wesley.

Cornwell, T., and K. Evans (1985): "A simple maximum entropy deconvolution algorithm," *Astronomy and Astrophysics*, 143, 77–83.

Durlauf, S. (1997): *The Economic System as an Evolving Complex System II*chap. Statistical Mechanics Approaches to Socioeconomic Behavior. Addison-Wesley.

Foley, D. K. (1994): "A Statistical Equilibrium Theory of Markets," *Journal of Economic Theory*, 62, 321–345.

Grauwin, S., F. Goffette-Nagot, and P. Jensen (2009): "Dynamic models of residential segregation: Brief review, analytical resolution and study of the introduction of coordination," *GATE Working Papers*, 09-14.

Jaynes, E. T. (1957a): "Information Theory and Statistical Mechanics I," *The Physical Review*, 106, 620–630.

——— (1957b): "Information Theory and Statistical Mechanics II," *The Physical Review*, 108, 171–190.

Monderer, D., and L. S. Shapely (1996): "Potential Games," *Games and Economic Behaviour*, 14, 124–143.

Narayan, R., and R. Nityananda (1986): "Maximum entropy image restoration in astronomy," *Annual review of astronomy and astrophysics*, 24, 127–170.

Schelling, T. C. (1969): "Models of segregation," *American Economic Review*, 59, 488–493.

——— (1971): "Dynamic models of segregation," *Journal of Mathematical Sociology*, 1, 143–186.

Shannon, C. E. (1948): "A Mathematical Theory of Communication," *The Bell System Technical Journal*, 27, 379–423.

Skilling, J., and S. F. Gull (1987): *Maximum-Entropy and Bayesian spectral analysis and estimation problems*chap. Prior knowledge must be used. Kluwer.

——— (1991): "Bayesian Maximum-Entropy Image Reconstruction," *Spatial Statistics and Imaging*, 20, 341–367.

Theil, H. (1967): *Economics and Information Theory*. North-Holland.

Toda, A. A. (2010): "Existence of a statistical equilibrium for an economy with endogenous offer sets," *Economic Theory*, 45, 379–415.

Vinkovic, D., and A. Kirman (2006): "A physical analogue of the Schelling model," *Proceedings of the National Academy of Sciences,*, 103, 19261–19265.

A Information-theoretic framework

There are two main differences between the standard relative information content (2) typically used in the image processing literature and the specification (3) which is actually used. The first is the use of an expectation operator $E[...]$, which indicates that the message reveals the colour of a random location picked with probability $l = 1/N$, rather than that of a known location i, as is the case with specification (2). The rationale for this is that this allows the entropy measure to build in variation of probabilities across locations i as well as agent type c, which is required given that the constraints are all expressed in terms of summation across locations,

$$E\left[H\left(i\|\, m_i\right)\right] = -\sum\nolimits_i \sum\nolimits_c l p_i^c \ln\left(\frac{l p_i^c}{l m_i^c}\right) = -\frac{1}{N}\sum\nolimits_i \sum\nolimits_c p_i^c \ln\left(\frac{p_i^c}{m_i^c}\right) .$$

The second difference is the use of the double space entropy suggested by Skilling and Gull (1987) to integrate prior knowledge of *relative* rather than *absolute* positions of agents. Formally, the relative entropy is the same as (2), except that it encodes the information content of a message revealing the colour $\{c, d\}$ of a randomly chosen pair of locations $\{i, j\}$, relative to what would be expected given prior knowledge $m_{i,j}^{c,d}$:

$$E\left[H\left(i,j\|\, m_{i,j}\right)\right] = -\frac{1}{N^2}\sum\nolimits_{i,j} \sum\nolimits_{c,d} p_{i,j}^{c,d} \ln\left(\frac{p_{i,j}^{c,d}}{m_{i,j}^{c,d}}\right) .$$

$$E\left[H\left(i,j\|\, m_{i,j}\right)\right] = -\frac{1}{N^2}\sum\nolimits_{i,j} \sum\nolimits_{c,d} p_{i,j}^{c,d} \ln p_{i,j}^{c,d} + \frac{1}{N^2}\sum\nolimits_{i,j} \sum\nolimits_{c,d} p_{i,j}^{c,d} \ln m_{i,j}^{c,d}$$

Treating the joint probability as the product of the marginal probabilities $p_{i,j}^{c,d} = p_i^c p_j^d$, one obtains the following expression, which is the specification used in equation (3). Although the existence correlations in the model $m_{i,j}^{c,d}$ means that the probabilities are not in fact independent, this assumption allows the relative entropy to measure the extra information required to treat probabilities p_i^c and p_j^d as independent when they are in fact related by the model,

$$E\left[H\left(i,j\|\, m_{i,j}\right)\right] = -\frac{2}{N}\sum\nolimits_i \sum\nolimits_c p_i^c \ln p_i^c + \frac{1}{N^2}\sum\nolimits_{i,j}\sum\nolimits_{c,j} p_i^c p_j^d \ln m_{i,j}^{c,d} \ .$$

Given this specification for the relative entropy and the constraints (4) and (6), the lagrangian for the maximum entropy problem is:

$$\Lambda = E\left[H\left(i,j\|\, m_{i,j}\right)\right] - \sum\nolimits_c \beta^c \left(\sum\nolimits_i \frac{p_i^c}{N} - s^c\right) - \sum\nolimits_c \alpha^c \left(\frac{\left(\chi^2\right)^c}{N} - \frac{\left(N-\Gamma^c\right)}{N}\right) . \quad \text{(A-1)}$$

This leads to the following first order condition with respect to p_i^c:

$$\frac{\partial \Lambda}{\partial p_i^c} = -\ln\left(\frac{p_i^c}{\mu_i^c}\right) - 1 - \beta^c - \alpha^c \frac{\partial\left(\chi^2\right)_i^c}{\partial p_i^c} = 0 \ ,$$

$$p_i^c = \mu_i^c e^{-(1+\beta^c)} e^{-\alpha^c \frac{\partial\left(\chi^2\right)_i^c}{\partial p_i^c}} \ . \quad \text{(A-2)}$$

Replacing this into the normalisation constraint (4) allows the derivation of partition function, i.e. the $1+\beta^c$ exponential term:

$$\sum\nolimits_i \mu_i^c e^{-(1+\beta^c)} e^{-\alpha^c \frac{\partial\left(\chi^2\right)_i^c}{\partial p_i^c}} = N s^c \quad \Rightarrow \quad e^{1+\beta^c} = \frac{1}{N s^c}\sum\nolimits_i \mu_i^c e^{\alpha^c \frac{\partial\left(\chi^2\right)_i^c}{\partial p_i^c}} \ .$$

B Maximum entropy algorithm

The algorithm used to obtain the probability distribution (7) follows from Cornwell and Evans (1985) and integrates a chi-squared constraint in the spirit of Skilling and Gull (1991).[10] The initial probability and model vectors are given by the uniform distribution $p_i^c = m_i^c = s^c$. Prior to running the algorithm, the initial conditions are processed in order to extract the relevant data for calibrating the model constraints:

- The initial attractiveness data vector d^c is calculated as a convolution of initial state vector f^c, i.e. $d^c = B \times f^c$.

[10] The code for the Schelling simulation and the MaxEnt reconstruction algorithm is available from the author on request, as well as the initial condition matrix required for replicating the figures shown here.

- The mean $\overline{d^c}$ and standard deviation σ^c of the d^c data are calculated.
- The number of good locations Γ^c is determined as the number of locations with $d_i^c \gtrless \overline{d^c} \pm 2\sigma^c$.
- Because the good locations Γ^c are clustered, the number of distinct clusters and the mean radius b of a cluster are calculated. This is used to calibrate the model M^c, which is assumed to be a circulant matrix containing a gaussian convolution of radius b.

The iterative algorithm is based on the Newton-Raphson method, with the Jacobian vector and Hessian matrix of the lagrangian given by:

$$\begin{cases} \nabla\Lambda^c = \nabla H^c - \beta^c - \alpha^c \nabla \left(\chi^2\right)^c \\ \nabla\nabla\Lambda^c = \nabla\nabla H^c - \alpha^c \nabla\nabla \left(\chi^2\right)^c \end{cases} . \tag{A-3}$$

Each iteration starts with a calculation of the current entropy and chi-squared gradients ∇H^c and $\nabla(\chi^2)^c$. The step change in the probability vector at each iteration of the Newton-Raphson method is then:

$$\Delta p^c = -\left(\nabla\nabla\Lambda^c\right)^{-1} . \nabla\Lambda^c \ . \tag{A-4}$$

Given the large size of N, inverting the $N \times N$ Hessian matrix is computationally intensive. Cornwell and Evans (1985) show, however, that it is possible to use the structure of the Hessian to produce a simplified estimate. Given the specification of the information entropy, $\nabla\nabla H^c$ is simply a diagonal matrix with diagonal elements $\nabla\nabla H^c_{i,i} = 1/p^c_i$, and the only off-diagonal elements in the Hessian come from the chi-squared term. Cornwell & Evans suggest using assigning all the weight of the off diagonal terms to the diagonal term using q, the number of pixels in the neighbourhood. This leads to the following diagonal approximation to the Hessian, used to calculate (A-4):

$$\widehat{\nabla\nabla\Lambda}^c = \nabla\nabla H^c - \alpha^c \frac{2q}{\left(\sigma^c\right)^2} I \ .$$

The values of the α^c and β^c parameters need to be determined so that constraints (4) and (6) are satisfied for each step of the iteration. The β^c parameter is obtained simply by calculating $\ln Z^c$ for each iteration, while α^c is obtained by iterating the following expression, based on the deviation of the χ^2 term from its constrained level $N - \Gamma^c$:

$$\Delta\alpha^c = \alpha^c \frac{(\Delta\chi^2)^c - (N - \Gamma^c)}{\|\nabla (\chi^2)^c\|} \ . \tag{A-5}$$

$\Delta(\chi^2)^c$ is value of the chi-squared deviation calculated for the updated probability vector $p^c + \Delta p^c$. In between iterations of (A-5) the value of Δp^c is recalculated using the updated value of α^c in $A - 3$, keeping other terms in the Jacobian and Hessian constant.

Once the value of α^c satisfies the constraint (up to a tolerance ε), the current step vector Δp^c is accepted and the probability vector is updated, $p^c + \Delta p^c$. The model is also updated at this point using the following expression, obtained by deriving the model specification with respect to p_i^c:

$$\Delta\mu^c = [\mu^c] \, [p^c]^{-1} \, M^c \Delta p^c \ .$$

Finally, as in Cornwell and Evans (1985), the stopping condition for the iterations is that the norm of the Jacobian must fall below a tolerance ε relative to the norm of the unit vector:

$$\frac{\|\nabla \Lambda^c\|}{\|1\|} < \varepsilon \ .$$