University of Kent

School of Economics Discussion Papers

Back to the future: economic rationality and maximum entropy prediction

Sylvain Barde

January 2012

KDPE 1202

Back to the future: economic rationality and maximum entropy prediction*†

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*

December 2011

Abstract

An information-theoretic thought experiment is developed to clarify why the maximum entropy methodology is appropriate for predicting the equilibrium state of economic systems. As a first step, object allocation problems, modeled as knapsack problems, are shown to be equivalent to congestion games under weak assumptions. This proves the existence of finite improvement paths linking initial conditions and Nash equilibria. The existence of these improvement paths is precisely what enables the use of maximum entropy to make predictions concerning the equilibrium state. Finally an illustration of this predictive power is provided through an application to the Schelling model of segregation.

JEL classification: C02, C11, C63, D80.

Keywords: Information entropy, knapsack problem, potential function, Schelling segregation.

1 Introduction

A central fact of economics, pointed out by Hayek (1945), is that an observer's knowledge of the state of the economy is vanishingly small. Even when some economic

** This paper integrates and extends two previous papers released under the title 'Ignorance is bliss: rationality, information and equilibrium' and 'Back to the Future: A Simple Solution to Schelling Segregation', respectively School of Economics Discussion Papers 11/03 and 11/04.*

†The author wishes to acknowledge the suggestions received at the ESHIA 2010 and 2011 conferences, as well as thank the GREQAM seminar participants for suggesting the Schelling application for the MaxEnt methodology. Particular thanks goes to Jagjit Chadha, Mishael Milaković, Alexis Akira Toda and Alan Kirman for their helpful advice, and to Sonia Moulet for tirelessly providing a sounding board for the author's ideas. Any errors are the author's.

data is available, it does not completely describe the state of the economy. Given such a setting, Jaynes (1957a,b) shows that the correct measure of the observer's ignorance is the information-theoretic Shannon (1948) entropy. This measure in turn provides the objective function of the maximum entropy (MaxEnt) methodology which can be used to make predictions about the aggregate behaviour of a system in situations where little detailed information is available.

Within economics MaxEnt has been used by Foley (1994) and Toda (2010) to prove the existence of a statistical market equilibrium when agents have offer sets of transactions they are willing to accept and meet in a random fashion. While in their framework the ignorance relates to the sequence of transactions carried out by agents, this will be extended here to include a much more fundamental form of uncertainty, where a social planner is unable to even observe the preference rankings of agents. Applied investigations using the MaxEnt methodology include Castaldi and Milaković (2007), which investigates the distribution of wealth using information on turnover in portfolios and Alfarano and Milaković (2008) which similarly explores the origin of the Laplace distribution of firm growth rates.

A related use of information entropy in economics is the rational inattention literature developed by Sims (2003, 2006). In this setting, as explained by Tutino (2011), while information is freely available, agents are limited in their ability to process it, typically through an upper bound on the bits per unit of time that can be processed. Within this literature, MaxEnt can be seen as the limit case where the processing capacity is zero, or equivalently the communication channel is closed.

Jaynes' original motivation is that the methodology is "maximally noncomittal with regard to missing information" (Jaynes, 1957a, p. 623), providing a generalisation of Laplace's principle of insufficient reason. Foley (1994) and Toda (2010) provide a related combinatorial argument, which is that MaxEnt provides "the transaction distribution that can be realised in the largest number of ways" (Foley, 1994, p. 322). Finally, the most rigorous treatment, by Shore and Johnson (1980), proves

that it is the only method of inference satisfying three key axioms: the prediction unique; it is independent of the coordinate system used; and it does not depend on whether information about independent systems or sub-systems is accounted for jointly or separately.

Regardless of these justifications, MaxEnt potentially suffers from one key problem when directly transposed to an economic setting. If the preferences and behaviour of agents are unobserved, then the predictions obtained with MaxEnt do not depend on them, as by construction they are independent from any missing information. This feature is most visible in the kinetic models used in econophysics, as picked up by Gallegati et al. (2006), in which the aggregate distributions result from assuming agents trade randomly chosen quantities, in an analogy to molecules in a gas model.[1] Chakrabarti and Chakrabarti (2009) attempt to address this issue by providing a kinetic model where agents trade to maximise Cobb-Douglas utility, but in order to conserve random interaction, they have to assume that the elasticities with respect to commodities are randomly drawn from a uniform distribution at each point in time, which is not consistent with the stable preferences used in standard economic theory.

The first purpose of this paper is therefore to provide a stronger motivation for the use of MaxEnt in economics. This is achieved by using a thought experiment in which the problem of resource allocation is presented as a variant of the knapsack problem. This is a well known combinatorial optimisation problem where one has a set of objects with given values and weights and the objective is to pick the combination of objects with the highest value without exceeding fixed a weight limit, i.e. the capacity of the knapsack. In contrast to the extensive literature on object allocation mechanisms and mechanism design initiated by Hurwicz (1973), Harris and Raviv (1981) or Myerson (1981), the aim is not to provide a practical solution to, or an optimal design for, this allocation problem, but rather to clarify

[1] See Chatterjee et al. (2005) for a good illustration of how kinetic models can be used to describe the key features of wealth distributions.

the sequence of steps required to solve the problem in theory in order to derive its key properties.

The key finding of this thought experiment is that under standard assumptions on preferences the knapsack allocation problem is equivalent to a congestion game. In such games, identified by Rosenthal (1973), a single potential function encodes changes in payoffs when agents switch strategies and attains an extremum for Nash equilibria. This implies the existence of a set of finite improvement paths linking any initial state to a Nash equilibrium. It is shown that these two properties provide a strong rationale for the use of MaxEnt in economics, as it becomes possible to treat the reversed improvement path like a noise process, which can then be described by information-theoretic methods.

Following this, the second purpose of the paper it to illustrate how MaxEnt can be used to predict the outcome of a simple agent-based model, namely the model of urban segregation developed by Schelling (1969, 1971). The simplicity of the framework, and crucially the presence of a potential function for the model make it ideally suited as a test bed for the methodology. Indeed, in the physical analog to the Schelling model proposed by Vinkovic and Kirman (2006), particles on a lattice systematically rearrange themselves to reduce the internal energy of their configuration, and the overall energy of the system provides the potential function. Very recent analysis of the model by Grauwin et al. (2011) confirms that it possess a potential function when bounded neighbourhoods are used.

The rest of the paper is structured as follows. Section 2 presents the the knapsack framework used to model the allocation problem facing a social planner and shows that under reasonable assumptions on preferences this is equivalent to a congestion game. The implications for predicting aggregate distributions using the MaxEnt methodology are clarified in section 3, and an application on to the Schelling model of segregation is shown in section 4. Section 5 discusses these findings and concludes.

2 Finite improvement paths in object allocation

2.1 Object allocation as a simple knapsack problem

The allocation problem facing a social planner is modeled using the multichoice multi-dimensional variant of the knapsack problem (MMKP). Compared to a standard knapsack problem the MMKP enlarges both the number of choices and constraints, thus making the choice framework more general. In this variant several groups of objects are available, with each object providing a specific value and requiring a particular subset of several distinct sets of resources. The objective is to pick a single object from each of the groups, maximising their aggregate value while ensuring the multi-dimensional resource constraint is met. For instance, in the allocation problem each agent is faced with a group of bundles and the optimisation requires picking a single bundle for each agent. The MMKP has already been used in the operational research literature to model practical allocation problems, for example allocating nurses with different skills and time preferences to different types of shifts (Dowsland and Thompson, 2000), or allocating distinct computing resources such as memory and CPU cycles to several networked users with different session preferences (Khan et al., 2002).

There are N agents in the economy, labeled $i \in \{1, 2, ..., N\}$, and the social planner has to allocate Q different units amongst those agents. Although this does not influence the general problem, it will be convenient in the discussion to distinguish K types of commodities, labeled $k \in \{1, 2, ..., K\}$ for which $q_k \in \mathbf{N}$ units are available, in which case $Q = \sum_k q_k$. The allocation problem can be solved, in principle, with the following four steps.

- Step 1: The social planner labels all the possible bundles that can be built with the Q units available and lists them in a $2^Q \times Q$ binary identifier table B, shown in table 1. The binary string formed by each row provides a unique

identifier for the bundle as well as the bundle's composition.

Table 1: Binary bundle identifiers

B	$j=1$	$j=2$	$j=3$	$j=4$	...	$j=Q$
$b=1$	0	0	0	0	...	0
$b=2$	1	0	0	0	...	0
$b=3$	0	1	0	0	...	0
...	...	...	...	...		...
$b=2^Q$	1	1	1	1	...	1

- Step 2: The social planner sends the B-table to the N agents who, assuming completeness, rank the 2^Q bundles according to their preferences. The rankings are returned to the social planner who then builds a $2^Q \times N$ ranking table U, shown in table 2. Under the usual assumptions of transitivity and monotonicity, all agents will rank the full bundle highest and the empty bundle lowest.

Table 2: Bundle preference ranking

U	$i=1$	$i=2$	$i=3$	$i=4$	...	$i=N$
$b=1$	1	1	1	1	...	1
$b=2$	...	...	...	...	...	...
...	...	...	...	...		...
$b=2^Q$	2^Q	2^Q	2^Q	2^Q	...	2^Q

- Step 3: The social planner must pick a bundle for each agent, using a $2^Q \times N$ choice matrix X, where the choice variables are $X_{b,i} \in \{0,1\}$. Importantly, each agent only receives a single bundle, i.e. $\sum_{b=1}^{2^Q} X_{b,i} = 1 \quad \forall i \in N$.[2] The goal of the social planner is to maximise the sum of the ranks over agents while remaining within the resource constraint. Formally, this can be expressed as the following MMKP:

[2]One can see that even if the agent is allocated two bundles a and b from B, then $a+b$ is also a bundle in B.

$$\begin{aligned} &\max \operatorname{tr}(UX') \\ &\text{s.t. } : B'X1_N = 1_Q \end{aligned} \tag{1}$$

Here 1_N and 1_Q are the N and Q-length unit vectors respectively. Choosing an objective function for the MMKP is directly related to the problem of choosing a social welfare function. The standard approach of knapsack problems is to maximise the sum of the values of the objects, therefore maximising the sum of the individual rankings is equivalent to a standard Benthamite social welfare function.[3] The constraint ensures that the sum of the binary identifiers for each selected bundle equals the unit vector, i.e. each unit in Q is selected only once. Expressed in scalar notation, this corresponds to the standard MMKP as presented by Hifi et al. (2004); Sbihi (2007). The only differences compared to the more general framework in the operational research literature is that the resource requirement per bundle in B is the same for all i agents and the available capacity is restricted to one for all dimensions in Q:

$$\begin{aligned} &\max \sum_{i=1}^{N} \sum_{b=1}^{2^Q} U_{b,i} X_{b,i} \\ &\text{s.t. } : \sum_{i=1}^{N} \sum_{b=1}^{2^Q} B_{b,j} X_{b,i} = 1 \quad \forall j \in Q \end{aligned} \tag{2}$$

- Step 4: Once the optimal choice table X^* is obtained, the social planner can build a $Q \times N$ allocation table $A^* = B'X^*$, shown in table 3. This table uniquely assigns every unit in Q to an agent in N, and can therefore be used for the purpose of selecting goods one by one and dispatching them to their allocated owner.

In theory all four steps of the MMKP are feasible and A^* exists. The problem is not tractable in practice, however, and one of the main advantages of the framework

[3] Given that a utility function is never uniquely defined, it is possible to change the social welfare function within the linear sum framework of the MMKP by applying monotonic transformations to the rankings expressed by the agents in table U.

Table 3: Allocation table

A^*	$i=1$	$i=2$	$i=3$	$i=4$	...	$i=N$
$j=1$	0	0	1	0	...	0
$j=2$	1	0	0	0	...	0
$j=3$	0	1	0	0	...	0
...	...	...	...	...		...
$j=Q$	0	0	0	1	...	0

is that it neatly separates the types of hurdles facing a social planner. The first the choice of the correct social welfare function, followed, as pointed out by Hayek (1945), by a high and potentially unfeasible informational requirement (Step 2) and by a large computationally complex combinatorial optimisation (Step 3).[4] As a result, although it exists, the optimal allocation A^* is unknown to the social planner.

2.2 Knapsacks, congestion games and improvement paths

As explained in section 2.1, the operational research literature has used the MMKP to model resource allocation on a network. Similar network allocation frameworks also serve as illustrations of congestion games, for example the road congestion setting presented by Rosenthal (1973), where road users attempt to select routes so as to minimise the congestion they experience. We now show that the two are equivalent under standard assumptions on preferences, which has important implications in terms of convergence to a decentralised allocation.

The allocation of Q goods over N agents with preferences given by U is modeled as a road congestion game where $q_k \in \mathbf{N}$ users of K different types have to choose a route i in an N-edge multigraph between start point s and finish point f, in order to maximise their payoff $V^k_{b,i}$. Elaborating on Rosenthal (1973), one could imagine that the K different types represent different categories of vehicles, such as cars, trucks, etc. who each generate different congestion costs. This choice of graph as,

[4]The knapsack problem is known to be NP-complete, in other words solutions to the problem can be verified efficiently (in polynomial time), but there is no known algorithm for calculating the solutions efficiently in the first place.

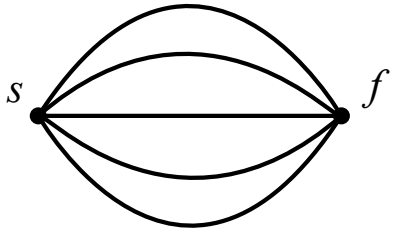


Figure 1: Multigraph congestion game

illustrated by Figure 1, implies that distinct routes follow separate edges i.e. there are no externalities between routes, where the benefit of a user choosing a route might depend on the number of users choosing another route.

As is standard in congestion games, the payoffs $V_{b,i}^k$ for choosing an edge are a function of the number of users already on the edge. These are derived from the ranking information returned in table 2. Specifically, let us define $\Delta_k U_{b,i} = U_{b,i} - U_{b\{-k\},i}$ as the change in ranks at the margin, following the addition of the last k-type good to bundle b. In terms of notation, $b\{-k\}$ is the bundle obtained by removing a k-type good from bundle b. Similarly, in the following, $b\{+k\}$ will refer to the bundle obtained by adding a k-type good to bundle b.[5] This allows the derivation of a $2^Q \times N \times K$ payoff array V which is set as $V_{b,i}^k = \Delta_k U_{b,i}$.

The following two assumptions on the rankings in the table 2 are required in order to show equivalence between the MMKP and the congestion game frameworks:

Monotonicity: $\Delta_k U_{b,i} > 0 \ \forall k, i, b$.

Concavity: Given two bundles a and b, $\forall i \in N$ if $U_{b,i} > U_{a,i}$ then $\Delta_k U_{b,i} < \Delta_k U_{a,i}$.

Monotonicity, which was already implicity assumed in the description of the rankings table 2 ensures that all the congestion game payoffs in V are strictly positive. Concavity intuitively means that the bundle rankings exhibits decreasing marginal

[5] $\Delta_k U_{b,i}$ is of course undefined for the empty bundle and whenever bundle b contains no k-type units.

values, as adding extra units of k-type goods to a bundle, keeping the rest of the bundle constant, will bring successively smaller increases in the ranking. Together with monotonicity, this is required in order to ensure that the payoff of choosing a particular edge is decreasing with the number of users on that edge, as in the basic congestion game framework of Rosenthal (1973).[6] We now prove that if the rankings table U satisfies these two assumptions, the MMKP and congestion game formulations are equivalent.

Proposition: *If the ranking table U displays monotonicity and concavity, the optimal solution to the MMKP problem is a Nash equilibrium for the corresponding multigraph congestion game based on payoff table V.*

Proof: By contradiction. Let X^* be the decision table that satisfies the MMKP (1) and $A^* = B'X^*$ the corresponding allocation of the Q users over the N edges of the multigraph. Let us assume that A^* is not a Nash equilibrium for the Q users. If b and a are the bundles allocated to edges i and x respectively by X^*, this implies:

$$\exists \quad i, x, k: \quad V^k_{a\{+k\},x} > V^k_{b,i}$$

Let Y be the decision table resulting from the switch of the k-type agent from edge i to x. Y is identical to X^*, except for edges i and x, which receive bundles $b\{-k\}$ and $a\{+k\}$ respectively. Using the definition of $V^k_{b,i}$:

$$\begin{aligned}
&\Delta_k U_{a\{+k\},x} > \Delta_k U_{b,i} \\
&U_{a\{+k\},x} + U_{b\{-k\},i} > U_{a,x} + U_{b,i} \\
&\mathrm{tr}\,(UY') > \mathrm{tr}\left(UX^{*\prime}\right)
\end{aligned}$$

X^* does not satisfy the MMKP, which is a contradiction. ■

[6] In the standard framework of Rosenthal (1973), congestion costs on an edge are increasing with the numbers of users on the edge, and the aim of the network users is to choose the edge with the lowest cost. This is equivalent to the framework used here, where the benefit of using an edge falls with the numbers of users on the edge, and users choose the edge with the highest benefit.

Corollary: *If the ranking table U displays monotonicity and concavity, the objective function of the MMKP is an exact potential function for the corresponding multigraph congestion game based on payoff table V.*

Proof: Immediate from the previous proof and the definition of the payoffs $V^k_{b,i}$. The change in payoff to a k-type user for switching from edge i to x is $V^k_{a\{+k\},x} - V^k_{b,i}$. Given $V^k_{b,i} = \Delta_k U_{b,i}$ and one thus has:

$$V^k_{a\{+k\},x} - V^k_{b,i} = \Delta_k U_{a\{+k\},x} - \Delta_k U_{b,i} = \mathrm{tr}\,(UY') - \mathrm{tr}\left(UX^{*\prime}\right)$$

The objective function of the MMKP is an exact potential for the congestion game based on the corresponding $V^k_{b,i}$ payoffs. ■

As was shown by Monderer and Shapely (1996), the equivalence of the MMKP and congestion game framework and the existence of a potential implies the existence of the finite improvement property (FIP). Starting from any initial state a simple best response path will lead to a Nash equilibrium in a finite number of steps.[7] In particular, agent pairs meeting randomly and trading goods with low marginal utility $\Delta_k U_{b,i}$ against goods with higher marginal utility would satisfy the requirement, as the potential function $\mathrm{tr}(UX')$ would increase following such trades.[8] The central implication of this result is that if the rankings expressed by the N agents are monotonic and concave, then even though the MMKP cannot be solved centrally, the social planner can be confident that the system will eventually reach a decentralised allocation. In itself, this is no surprise and only duplicates the findings of the objet alllocation literature mentioned in section 1. Nevertheless, it is shown below that the existence of the FIP has implications for the use of the MaxEnt methodology in economic systems.

[7] As pointed out by Rosenthal (1973), the equilibrium obtained in such a manner will not necessarily be the one that maximises the potential.

[8] Clearly, the switching process used in the proof is simplistic: one does not expect goods to choose their owners in order to maximise a payoff! A trade, however, can be broken down into a sequence of such switches: a k-type good moves from agent i to agent x, immediately followed by a another commodity switching from x to a.

3 Information-theoretic consistent prediction

The FIP implies that any initial state I is linked to a Nash equilibrium F by a finite sequence of intermediate states $I \to F$, where each transition is the result of agents making welfare increasing trades. For each commodity k we define $h_k^I(\varepsilon)$ and $h_k^F(\varepsilon)$ as share of the N agents with endowment level $\varepsilon \in \{0, 1, 2, ..., q_k\}$ in the initial and final states, respectively. h_k^I and h_k^F refer to the overall frequency distributions.

Assuming that only the initial endowment distribution h_k^I is known and that the intermediate states in the path $I \to F$ are not observable, a reasonable criterion for the social planner to use in predicting the equilibrium distribution h_k^F is to maximise the posterior probability given the knowledge of the initial frequency, i.e. $\arg\max_{h_k^F} p\left(h_k^F | h_k^I\right)$. Bayes' rule states that this can be expressed as the product of a prior on h_k^F and the likelihood $p(h_k^I | h_k^F)$, normalised by the evidence $p(h_k^I)$:

$$p\left(h_k^F | h_k^I\right) = p\left(h_k^F\right) \frac{p\left(h_k^I | h_k^F\right)}{p\left(h_k^I\right)} \tag{3}$$

The likelihood $p(h_k^I | h_k^F)$ can be rewritten using the log-likelihood $\ell(h_k^I | h_k^F)$:

$$p\left(h_k^I | h_k^F\right) = \frac{\exp\left(-\ell\left(h_k^I | h_k^F\right)\right)}{Z_\ell}$$

At this point one might think that finding the maximum of (3) is a matter of maximising the likelihood, effectively treating the prior $p(h_k^F)$ as constant. However, in this case, because h_k^F and h_k^I are simply histograms on the same discrete support $\{0, 1, 2, ..., q_k\}$, there are as many 'parameters' to determine in h_k^F as there are 'data points' in h_k^I. As a result the problem is likely to be degenerate and the choice of prior will have an effect on the prediction. In such cases, Jaynes (1957a) and Shore and Johnson (1980) advocate the use of an entropic prior of the form:

$$p\left(h_k^F\right) = \frac{\exp\left(\alpha S\left(h_k^F | h_k^I\right)\right)}{Z_S} \tag{4}$$

Here α is a regularisation parameter and $S\left(h_k^F|h_k^I\right)$ is the information entropy of the distribution. Many applications, including those mentioned in section 1 directly use the Shannon (1948) entropy.

$$S\left(h_k^F\right) = -\sum_{\varepsilon=0}^{q_k} h_k^F(\varepsilon) \ln h_k^F(\varepsilon) \tag{5}$$

The justification for this, epitomised in Foley (1994), is that $S\left(h_k^F\right)$ is in fact the logarithm of the multiplicity, i.e the number of ways a distribution can be realised. This implies that the prior probability (4) of a given distribution h_k^F is simply proportional to its multiplicity.

The existence of the FIP for well-behaved preferences provides an additional information-theoretic justification for the following relative entropy measure, which is more general and is used in particular in Bayesian image reconstruction. This measure is equal to minus the Kullback-Leibler (KL) divergence from h_k^I to h_k^F, and measures the similarity between two distributions h_k^I to h_k^F.[9] It reaches a global maximum of zero for $h_k^I = h_k^F$ and is strictly negative for $h_k^I \neq h_k^F$.

$$S\left(h_k^F|h_k^I\right) = -\sum_{\varepsilon=0}^{q_k} h_k^F(\varepsilon) \ln \frac{h_k^F(\varepsilon)}{h_k^I(\varepsilon)} \tag{6}$$

As stated above, changes in endowments on the improvement path $I \to F$ result from agents systematically trading to improve their welfare. If, however, the sequence of states forming the improvement path is view in reverse, $F \to I$, changes in endowments reflect a systematic sequence of errors as agents transition from optimal to sub-optimal states. The key consequence is that the known state I can be considered to be the result of a particular noise process applied to the unobserved state F. Therefore, the observer's absolute uncertainty as to the distribution of

[9]Formally, the KL divergence measures how many bits of information are gained by learning that the true distribution is h_k^F rather than h_k^I. As explained by Cover and Thomas (1991) it is often used as a measure of the distance between two distributions, and its additive inverse is therefore a measure of similarity. The reader is referred to Skilling and Gull (1991) for a discussion of this entropy measure in an image reconstruction context.

endowments h_k^F that occurs in the final state F, measured by (5), must be corrected by the knowledge of h_k^I, given that in theoretical terms it can be considered to be a noisy version of h_k^F itself.[10]

The posterior probability (3) can now be expressed as:

$$p\left(h_k^F|h_k^I\right) = \frac{\exp\left(\alpha S\left(h_k^F|h_k^I\right) - \ell\left(h_k^I|h_k^F\right)\right)}{p\left(h_k^I\right) Z_S Z_\ell} \tag{7}$$

One can see from (7) that finding the distribution h_k^F with the highest a posterior probability is therefore equivalent to solving a maximum entropy program with respect to h_k^F:

$$\arg\max_{h_k^F} p\left(h_k^F|h_k^I\right) \quad \Leftrightarrow \quad \arg\max_{h_k^F} \left(\alpha S\left(h_k^F|h_k^I\right) - \ell\left(h_k^I|h_k^F\right)\right)$$

The general prediction is given by the expression below, where α plays the role of a Lagrange multiplier.

$$h_k^F\left(\varepsilon\right) = h_k^I\left(\varepsilon\right) \exp\left(-\frac{1}{\alpha}\frac{\partial \ell\left(h_k^I|h_k^F\right)}{\partial h_k^F\left(\varepsilon\right)}\right) \tag{8}$$

As an illustration, let us suppose that the noise process $F \to I$ is such that agents can access all endowment levels in the initial state I with equal probability, regardless of their endowment level in the final state F. This effectively implies assuming ergodicity, with a uniform distribution $h_k^I(\varepsilon)$ over the support $\{0, 1, 2, ..., q_k\}$, frequencies given by $h_k^I(\varepsilon) = (q_k + 1)^{-1}$ and a likelihood given by $(q_k + 1)^{-N}$. However, in a pure allocation problem such as the MMKP outlined above, the amount of objects to be allocated q_k is constant, therefore $\ell(h_k^I|h_k^F)$ must reflect the fact that q_k is constrained to be equal to the aggregate endowments in the final state:

$$N\sum_{\varepsilon=0}^{q_k} \varepsilon h_k^F\left(\varepsilon\right) = q_k \tag{9}$$

[10] One can see that the two forms of entropy (5) and (6) are equivalent in a maximisation problem if h_k^I is a uniform distribution. This case is examined later.

The maximum entropy program for the prediction is therefore:

$$\arg\max_{h_k^F} \left(\alpha S\left(h_k^F | h_k^I\right) - N \ln\left(q_k + 1\right)\right) \tag{10}$$

Maximising (10) with respect to h_k^F, using (9) for the value of q_k, one obtains the following predicted distribution:

$$h_k^F\left(\varepsilon\right) = \frac{e^{-\lambda\varepsilon}}{q_k + 1}, \quad \text{with} \quad \lambda = \frac{N^2}{\alpha\left(q_k + 1\right)}$$

This replicates the MaxEnt prediction made in Foley (1994) of exponential endowment distributions in a pure exchange economy. However, it is important to point out that using relative (6) rather than the Shannon (5) entropy allows additional flexibility through the integration of the information provided by the initial state I. In the ergodic case presented above, the noise process $F \to I$ results in all endowment levels in the initial state I being equally accessible to any agent. As a result, I contains no information about F and using relative entropy (6) produces the same result as using (5). However, if it is not the case that the system is ergodic, i.e. the noise process obtained by reversing agent trades does not lead to a uniform h_k^I, this should be reflected in the prediction (8). For instance, let us suppose that the initial state I is arbitrarily close to a Nash equilibrium, such that resulting improvement path $I \to F$ is very short. With such a short path, h_k^I will contain very little noise compared to h_k^F, and one would expect intuitively that the best prediction for h_k^F displays a strong around h_k^I, which is what would be expected from (8). The relationship between the width of the distribution in the denominator of the relative entropy term (6) and the implicit length of the improvement path is investigated in a companion paper, Barde (2012) which applies the MaxEnt methodology to two well-known agent based models. The central finding is that the width of the model term around the initial condition does indeed control the time-horizon of the prediction.

4 Application to the Schelling model of segregation

4.1 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. This is because the attractiveness of a location to an agent is a function of the number of similar agents in the vicinity, which usually determined by counting the number of similar agents within a neighbourhood of given width. If B is a $N \times N$ binary matrix which identifies the neighbours for all N locations, and l^c is the binary vector for the location of c-type agents, this similarity for each location i is given by:

$$(B \times l^c)_i = \sum\nolimits_j B_{i,j} l^c_j \tag{11}$$

As is the case with the work of Grauwin et al. (2011), it is assumed that 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 circulant matrix, which greatly facilitates the analysis. A further assumption used here is that the utility of an agent is directly proportional to the number of similar neighbors. This is contrast to original Schelling (1969, 1971) model, where utility is a unimodal function of similarity, initially increasing with similarity, peaking for a balanced neighbourhood composed of 50% of agents of each type, then declining as similarity increases further. Indeed, Grauwin et al. (2011) show that in the case of continuous neighborhoods such as the (11), the existence of a potential function - critical to the argument in section 4.2 - requires utility functions that are linear

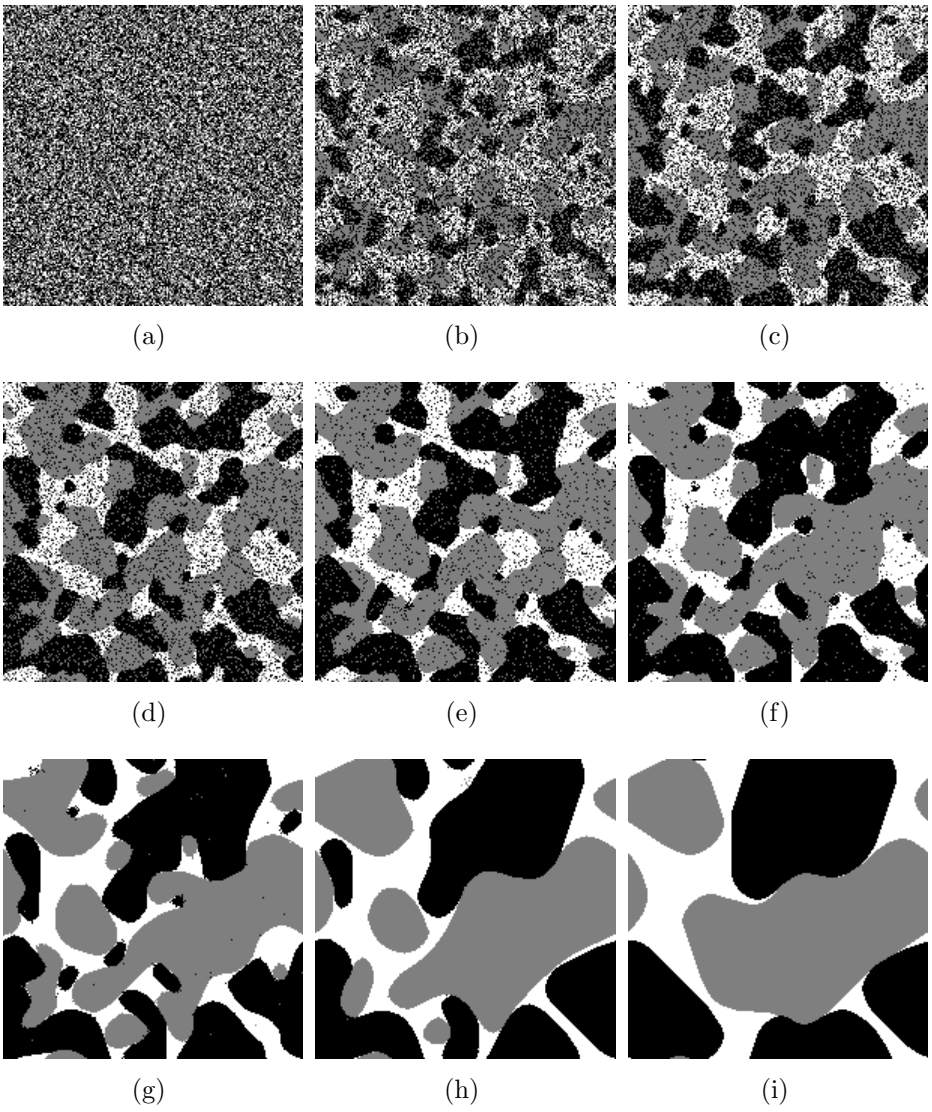


Figure 2: **Emergence of segregation in the Schelling model**

functions of the number of similar agents.

Because it is assumed that relocation opportunities arrive as a Poisson process, simulation is usually the method of choice for investigating this model. Grauwin et al. (2011) themselves point out that most analyses of this model rely on agent-based simulations and lack analytical solutions. Such a simulation is therefore provided as a point of reference for the MaxEnt prediction methodology. [11]

[11]The parameters for the benchmark simulation are as follows: the city is 200 pixels across and each pixel represents a location, so there are $N = 200^2 = 40000$ locations. There are $N^R = N^G = 16000$ red and green agents and $N^W = 8000$ free spaces. The continuous neighborhood agents consider when assessing the desirability of a given location is a 7×7 square area centered on that

The sequence of images in Figure 2 (a)→(i) provides a time-lapse of the simulation process. The random initial state is given in Figure 2 (a) , while Figure 2 (i) represents the state of the city after 44841 individual moves have occurred. The final state in 2 (i), which exhibits the segregated outcomes typical of the Schelling model, is stable as no further utility-improving relocations exist.

4.2 MaxEnt prediction

The Schelling model provides a simple yet effective setting for showcasing the use of MaxEnt as an information-theoretic prediction methodology. In particular, the sequence of images in Figure 2 provides an illustration of the FIP resulting from the presence of a potential function. As stated in section 4.2, viewing the finite improvement sequence in reverse, from (i)→(a), provides 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. Thus, predicting 2 (i) from 2 (a) is equivalent to retrieving a clean image 2 (i) from a noisy one 2 (a). The MaxEnt methodology has a long history in addressing such problems in image processing and astronomy, where the noise process involved in measurement is similar to the (i)→(a) sequence of Figure 2. As a result, the specific algorithm used to obtain the MaxEnt predictions of the Schelling model is a modified version of the image reconstruction algorithm of Skilling and Gull (1991).

Within the setting described above in 4.1, let p_i^c be the probability that the i^{th} location is occupied by an agent of the c^{th} colour, with $c \in \{R, G, W\}$ and $\sum_c p_i^c = 1$. Given this, relative entropy (6) measures the expected information content of a message revealing the final state of a randomly picked location, relative to prior information on location of agents:

$$S\left(p_i|\, m_i\right) = -\frac{1}{N}\sum\nolimits_i \sum\nolimits_c p_i^c \ln\left(\frac{p_i^c}{m_i^c}\right) \tag{12}$$

location.

As outlined in Section 4.2, relative entropy (12) encodes prior information through the underlying model m_i^c. In the setup of the Schelling model, however, agent satisfaction does not depend on absolute location, but on location relative to other agents. As a result, there is no prior information regarding the probability of a single location being occupied by a particular type of agent, and m_i^c in expression (12) 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 $\{i, j\}$.[12] Using expression (13) enables the integration of a two-dimensional model $m_{i,j}^{c,d}$ which can contain knowledge of correlations across locations.[13] This is better suited to the prior information provided in the Schelling model, in which one expects neighbouring locations to have a relatively high probability of being occupied by similar agents.

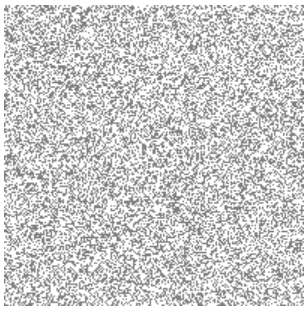
$$S\left(p_i, p_j \middle| m_{i,j}\right) = \frac{1}{N}\left(-2\sum\nolimits_i \sum\nolimits_c p_i^c \ln p_i^c + \frac{1}{N}\sum\nolimits_{i,j}\sum\nolimits_{c,d} p_i^c p_j^d \ln m_{i,j}^{c,d}\right) \tag{13}$$

The second important piece of information required for the MaxEnt prediction is the initial condition of the system, which provides the data entering the likelihood in (7). With the reversed FIP, where Figure 2 (i) decays 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 setting, this intuitively 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 (11) for each type of population, shown in Figure 3.

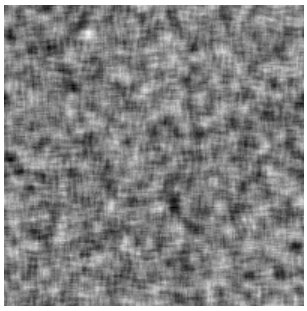
As a further simplification we assume, following the standard image recon-

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

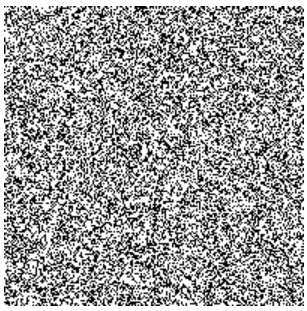
[13] This structure also allows correlations across agent types, for example if agents were to evaluate the attractiveness of a location not only by the number of similar agents but also by the number of agents of a different type. This is not the case here as in the basic Schelling model, agents only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



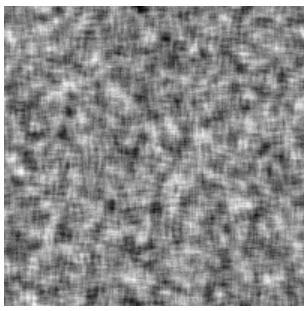
(a) initial green state



(b) initial green convolution



(c) initial red state



(d) initial red convolution

Figure 3: **Initial condition information**

struction literature, that the divergence between the prediction p^c and the initial attractiveness data d^c is normally distributed. This implies that the likelihood $p(d^c|p^c) = \exp(-\ell(d^c|p^c))/Z_\ell$ is gaussian, and therefore the log-likelihood $\ell(d^c|p^c)$ is directly related to the chi-squared deviation between the initial data available and the prediction, where $(\sigma^c)^2$ is the variance of the d^c data.

$$\ell\left(d^c|p^c\right) = \sum\nolimits_i \frac{\left((B * p^c)_i - d_i^c\right)^2}{\left(\sigma^c\right)^2} = \frac{\chi^2\left(p^c\right)}{2} \tag{14}$$

The the information theoretic problem is therefore to maximise the ignorance of an observer (13) subject to the information provided by the likelihood (14), normalising to ensure that the predicted number of agents of each colour equals the initial amount N^c. As pointed out by Skilling and Gull (1991), the value of the implicit Lagrange parameter α is used to constrain the noise level measured by (14) to be equal to number of degrees of freedom controlled by noise, i.e. the overall number

of locations N minus the number of good locations Γ^c in the initial data.[14]

The first order condition of the problem directly provides the best prediction for the distribution of agents over the locations:

$$p_i^c = \frac{\mu_i^c}{Z^c} \exp\left(-\frac{1}{2\alpha^c}\frac{\partial \chi^2\left(p^c\right)}{\partial p_i^c}\right) \tag{15}$$

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

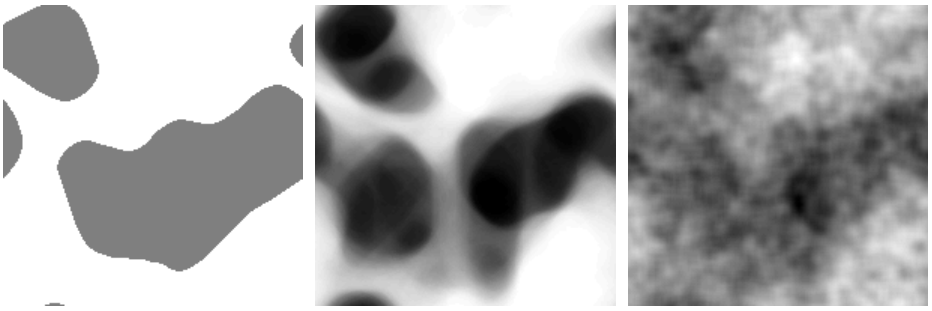
$$\mu_i^c = \exp\left(\frac{1}{2N}\sum\nolimits_j \sum\nolimits_{c,d} p_j^d \ln m_{i,j}^{c,d}\right) \quad \text{and} \quad Z^c = \frac{1}{N^c}\sum\nolimits_i \mu_i^c \exp\left(-\frac{1}{2\alpha^c}\frac{\partial \chi^2\left(p^c\right)}{\partial p_i^c}\right)$$

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 (11). [15] It is important to point out that expression (15) only provides an implicit solution for the probability distribution p_i^c as both the model term μ_i^c and noise term $\chi^2(p^c)$ are themselves functions of p_i^c. The predicted distributions are therefore obtained using a gradient-based algorithm, outlined in appendix B.

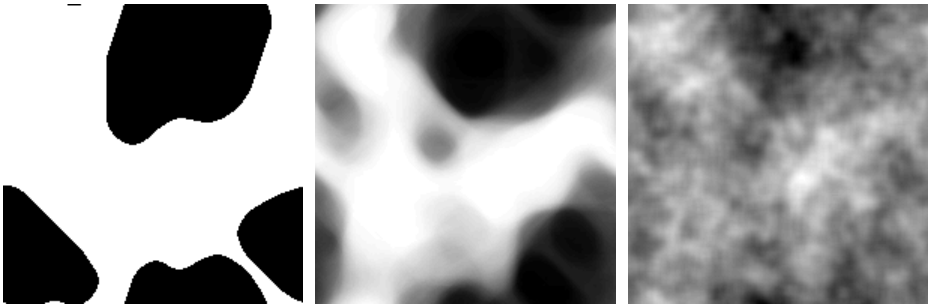
Figures 4 (c) and (f) provide the MaxEnt prediction (15) given the information from the initial condition in Figure 3. As a point of comparison, Figures 4 (a) and (d) are the colour-specific results of the simulation shown in Figure 2, and Figures 4 (b), (e) provide the colour-specific frequencies obtained by running 1000 Monte-Carlo (MC) iterations of the Schelling model on the same initial condition. Intuitively, these indicate the percentage of simulations that result in a particular location being occupied by a green or red agent.

[14] The relation between α and the number of noisy degrees of freedom as well as the calculation of Γ^c are explained in appendix B.

[15] In practice the convolution used in the prediction algorithm is different: instead of calculating $p_j^d \ln m_{i,j}^{c,d}$ the algorithm uses $m_{i,j}^{c,d} \ln p_j^d$. 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 efficiently as a sparse matrix with many zero elements. Taking the logarithm of this sparse matrix thus becomes problematic, therefore in practice it is easier to take the logarithm of the p^c.



(a) Final green state (b) Monte Carlo green frequency (c) MaxEnt Green probability density



(d) Final red state (e) Monte Carlo red frequency (f) MaxEnt Red probability density

Figure 4: **MaxEnt predictions vs. Monte Carlo frequencies**

The central finding is that although the initial information in Figure 3 is very noisy and therefore seems of limited use, comparing Figures 4 (b), (c) and (e), (f) respectively suggests that the MaxEnt methodology can extract it effectively and thus provide a reliable prediction for the expected outcome of the model. This is confirmed by the large and highly significant Spearman rank correlation in Table 4. Two measures of relative mean square error (MSE) are also provided: the first measures the sum of squared deviations between the MaxEnt prediction and the MC frequency relative to the variance of the MC frequency and the second provides the same measure using standardised MaxEnt and MC distributions. The standardised MSE supports the good predictive power of the MaxEnt procedure, but the regular MSE measure suggests that the MaxEnt prediction is only slightly better than the expectation N^C/N. What this difference indicates is that although MaxEnt correctly predicts the sign of the deviation from N^C/N, the predictions (15) are much

Table 4: Goodness of fit tests, Monte Carlo vs. MaxEnt

	Spearman's ρ	p-value	MSE/σ^2	Standardised MSE/σ^2
Green	0.910	0	0.9879	0.2292
Red	0.8190	0	0.9888	0.4546

flatter than the empirical MC frequencies. This is to be expected given that the reconstructed image (15) only controls a very small number of degrees of freedom Γ^c, the rest being controlled by the noise level, as measured by the chi-squared deviation (14).

5 Discussion and Conclusion

The justification provided for using MaxEnt in economics formally rests on the structural similarity between simple object allocation, as modeled by the MMKP, and congestion games frameworks. In fact, the only requirement for the optimal MMKP allocation to also be a Nash equilibrium in a corresponding congestion game is concavity in the bundle rankings, allowing the MMKP objective function to become an exact potential for the game. Assuming this is the case, the system displays the FIP, i.e. from any initial state there exists a finite path to a Nash equilibrium under even the simplest adjustment dynamics. It is the presence of the FIP and the corresponding improvement path that then provides the key motivation for the MaxEnt methodology in such a system, as well as a clarification of the link with the kinetic models mentioned in the introduction. This is because the sequence of states forming the improvement path can be interpreted in two ways, depending the direction in which it is viewed.

If the improvement path is viewed forward, starting at the initial state and finishing at the equilibrium, the picture one has is of a system that gradually self-organises as agents perform welfare-increasing trades. This is analogous to the definition of

biological entities as entropy-reducing systems, following in particular the initial insight of Schrödinger (1967). The information entropy measures the ignorance of the observer as to sequence of trades that occur, in particular the underlying preferences of agents. If, however, the improvement path is played backwards, i.e. starting at the equilibrium and traveling back towards the initial condition, then a physical analogy is more appropriate. In this setting agents systematically make welfare decreasing trades, in other words systematic errors. With this reverse view, the initially ordered state gradually decays through contamination by noise. This corresponds to the direct physical interpretation of entropy increases: An initially ordered system, say an ice cube in a glass of water, which gradually decays into a disordered thermal equilibrium. In this case the information entropy measures ignorance as to the amount and type of noise that has been introduced.

The central finding is therefore that if a system possesses the FIP, then predicting its equilibrium from a known initial state is formally equivalent to reconstructing an unobserved clean signal out of an observed noisy one. Not only does this provide a stronger justification for the use of MaxEnt and information-theoretic methods in economics, but it also suggests that image reconstruction algorithms, designed specifically for the purpose of removing noise from a signal, could become useful tools in economic prediction. This is illustrated by the application to the Schelling model, where such an algorithm is shown to be able to predict the emergence and location of segregated neighbourhoods with a good level of accuracy.

References

Alfarano, S., Milaković, M., 2008. Does classical competition explain the statistical features of firm growth? Economics Letters 101, 272–274.

Barde, S., 2012. Of ants and voters: Maximum entropy prediction of agent-based models with recruitment. Revue de l'OFCEForthcoming.

Castaldi, C., Milaković, M., 2007. Turnover activity in wealth portfolios. Journal of Economic Behavior and Organization 63, 537–552.

Chakrabarti, A. S., Chakrabarti, B. K., 2009. Microeconomics of the ideal gas like market models. Physica A 388, 4151–4158.

Chatterjee, A., Yarlagadd, S., Chakrabarti, B. K., 2005. Econophysics of wealth distributions. Springer.

Cover, T. M., Thomas, J. A., 1991. Elements of Information Theory. John Wiley & Sons.

Dowsland, K. A., Thompson, J. M., 2000. Solving a nurse scheduling problem with knapsacks, networks and tabu search. Journal of the Operational Research Society 51, 825–833.

Foley, D. K., 1994. A statistical equilibrium theory of markets. Journal of Economic Theory 62, 321–345.

Gallegati, M., Keen, S., Lux, T., Ormerod, P., 2006. Worrying trends in econophysics. Physica A 370, 1–6.

Grauwin, S., Goffette-Nagot, F., Jensen, P., 2011. Dynamic models of residential segregation: An analytical solution. Journal of Public EconomicsForthcoming.

Harris, M., Raviv, A., 1981. Allocation mechanisms and the design of auctions. Econometrica 49, 1477–1499.

Hayek, F., 1945. The use of knowledge in society. American Economic Review 35, 519–530.

Hifi, M., Michrafy, M., Sbihi, A., 2004. Heuristic algorithms for the multiple-choice multidimensional knapsack problem. Journal of the Operational Research Society 55, 1323–1332.

Hurwicz, L., 1973. The design of mechanisms for resource allocation. American Economic Review 63, 1–30.

Jaynes, E. T., 1957a. Information theory and statistical mechanics i. The Physical Review 106, 620–630.

Jaynes, E. T., 1957b. Information theory and statistical mechanics ii. The Physical Review 108, 171–190.

Khan, S., Li, K. F., Manning, E. G., Akbar, M., 2002. Solving the knapsack problem for adaptive media systems. Studia Informatica 2, 161–182.

Monderer, D., Shapely, L. S., 1996. Potential games. Games and Economic Behaviour 14, 124–143.

Myerson, R. B., 1981. Optimal auction design. Mathematics Of Operations Research 6, 58–73.

Rosenthal, R. W., 1973. A class of games possesing pure-strategy nash equilibria. International Journal of Game Theory 2, 65–67.

Sbihi, A., 2007. A best first search exact algorithm for the multiple-choice multidimensional knapsack problem. Journal of Combinatorial Optimisation 13, 337–351.

Schelling, T. C., 1969. Models of segregation. American Economic Review 59, 488–493.

Schelling, T. C., 1971. Dynamic models of segregation. Journal of Mathematical Sociology 1, 143–186.

Schrödinger, E., 1967. What Is Life? with Mind and Matter and Autobiographical Sketches. Cambridge University Press, Ch. What Is Life?

Shannon, C. E., 1948. A mathematical theory of communication. The Bell System Technical Journal 27, 379–423.

Shore, J. E., Johnson, R., 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Transactions on Information Theory 26, 26–37.

Sims, C. A., 2003. Implications of rational inattention. Journal of Monetary Economics 50, 665–690.

Sims, C. A., 2006. Rational inattention: Beyond the linear-quadratic case. American Economic Review 96, 158–163.

Skilling, J., Gull, S. F., 1987. Maximum-Entropy and Bayesian spectral analysis and estimation problems. Kluwer, Ch. Prior knowledge must be used.

Skilling, J., Gull, S. F., 1991. Bayesian maximum-entropy image reconstruction. Spatial Statistics and Imaging 20, 341–367.

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

Tutino, A., 2011. Rationally inattentive macroeconomic wedges. Journal of Economic Dynamics and Control 35, 344–362.

Vinkovic, D., Kirman, A., 2006. A physical analogue of the Schelling model. Proceedings of the National Academy of Sciences 103, 19261–19265.

A Information-theoretic framework

The key difference between the standard relative information content (12) typically used in the image processing literature and the specification (13) used here 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 (12), except that it encodes the information content of a message revealing the colours $\{c, d\}$ of a randomly chosen pair of locations $\{i, j\}$, relative to what would be expected given prior knowledge of correlations $m_{i,j}^{c,d}$:

$$S\left(p_i, p_j \middle| m_{i,j}\right) = -\frac{1}{N^2} \sum\nolimits_{i,j} \sum\nolimits_{c,d} p_{i,j}^{c,d} \ln \left(\frac{p_{i,j}^{c,d}}{m_{i,j}^{c,d}} \right)$$

$$S\left(p_i, p_j \middle| m_{i,j}\right) = -\frac{1}{N^2} \sum\nolimits_{i,j} \sum\nolimits_{c,d} p_{i,j}^{c,d} \ln p_{i,j}^{c,d} + \frac{1}{N^2} \sum\nolimits_{i,j} \sum\nolimits_{c,d} p_{i,j}^{c,d} \ln m_{i,j}^{c,d}$$

Treating the joint probability as the product of the marginal probabilities $p_{i,j}^{c,d} = p_i^c p_j^d$, and recognising that $\sum_i \sum_c p_i^c \ln p_i^c = \sum_j \sum_d p_j^d \ln p_j^d$, one obtains the specification used in equation (13). 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 underlying model.

Given the specifications for the entropy (13) and the the likelihood (14), maximising the posterior distribution involves solving the following maximum entropy problem. It is assumed that the α^c parameter integrates the multiplicative $2/N$ term in (13).

$$\arg\max_{p_i^c} \left(\alpha^c S\left(p_i, p_j \middle| m_{i,j}\right) - \frac{\chi^2\left(p^c\right)}{2} \right) \tag{A-1}$$

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

$$\alpha^c \left(-\ln p_i^c - 1 + \ln \mu_i^c\right) - \frac{1}{2}\frac{\partial \chi^2\left(p^c\right)}{\partial p_i^c} = 0$$

$$p_i^c \propto \mu_i^c \exp\left(-\frac{1}{2\alpha^c}\frac{\partial \chi^2\left(p^c\right)}{\partial p_i^c}\right) \tag{A-2}$$

Because N^c, the total number of agents of a particular colour, is given in the initial condition and does not change over time, it is possible to derivate a partition function Z^c which serves to normalise the distribution over locations:

$$\sum\nolimits_i p_i^c = N^c \quad \Rightarrow \quad Z^c = \frac{1}{N^c}\sum\nolimits_i \mu_i^c \exp\left(-\frac{1}{2\alpha^c}\frac{\partial \chi^2\left(p^c\right)}{\partial p_i^c}\right)$$

B Maximum entropy algorithm

The algorithm used to obtain the probability distribution (15) follows from Skilling and Gull (1991).[16] 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 l_0^c, i.e. $d^c = B \times l_0^c$.
- The initially most attractive locations G are determined as those where $d_i^c \gtrless \overline{d^c} \pm 2\sigma^c$. Because these good locations are clustered, the number of distinct clusters Γ^c is obtained by convolving the initial attractive locations G with B a second time to identify those which most attractive because located closest to each other. This provides $\Gamma^R = 14$ and $\Gamma^G = 11$.

[16] 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.

- Finally the expected radius of a cluster $b = \sqrt{G/(\Gamma^c * \pi)}$ is calculated. This is used to calibrate the model M^c, which is assumed to be a circulant matrix containing a gaussian convolution of standard deviation b.

B.1 Newton method iteration

The iterative algorithm is uses the Newton method. Referring to Q^c as the argument of the maximisation in (A-1), the Jacobian vector and Hessian matrix are given by:

$$\begin{cases} \nabla Q^c = \alpha^c \nabla S\left(p_i, p_j \middle| m_{i,j}\right) - \nabla \ell\left(d^c | p^c\right) \\ \nabla\nabla Q^c = \alpha^c \nabla\nabla S\left(p_i, p_j \middle| m_{i,j}\right) - \nabla\nabla \ell\left(d^c | p^c\right) \end{cases} \tag{A-3}$$

The step change in the probability vector at each iteration can be calculated using the standard Newton method:

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

Given that the Hessian matrix $\nabla\nabla Q^c$ is symmetric by construction, calculation of the iteration step (A-4) can be carried out efficiently by using the Preconditioned Conjugate Gradient method (PCG) to solve $-\nabla\nabla Q^c . \Delta p^c = \nabla Q^c$ without inverting the Hessian $\nabla\nabla Q^c$. Once this is done, the prediction is updated: $p^c + \Delta p^c$. The model is also updated at this point using $\Delta\mu^c = \left[\mu^c\right]\left[p^c\right]^{-1} M^c \Delta p^c$.

B.2 Control and termination

Two related control issues must be solved as the Newton iterations proceed. First of all, the value of the α^c parameter has to be determined and adjusted, and secondly the iteration must be terminated at some point. The main advantage of the Skilling and Gull (1991) approach is that it is the optimal value of α^c which both controls the iteration process and provides this termination condition. By integrating α^c into the hypothesis space of the Bayesian problem, they show that the most probable $\hat{\alpha^c}$

satisfies:

$$-2\alpha^c S\left(\hat{p}_i, \hat{p}_j \middle| m_{i,j}\right) = \mathrm{tr}\left(\left(\alpha^c I + L\right)^{-1} L\right) \quad \text{where} \quad \alpha^c I + L = [p^c]^{\frac{1}{2}} \nabla\nabla Q^c [p^c]^{\frac{1}{2}} \tag{A-5}$$

If λ_i are the eigenvalues of L, then $\mathrm{tr}\left(\left(\alpha^c I + L\right)^{-1} L\right) = \sum_i \lambda_i/(\alpha^c + \lambda_i)$. The trace term is therefore a measure of the number of good observations in the data, i.e. the number of dimensions for which $\lambda_i >> \alpha^c$, and the role of α is to identify the number of good observations and hence the amount of noise, as $-2\alpha^c S + \chi^2 = N$. If r is a $N \times 1$ vector of $N(0,1)$ errors, then the trace term can be estimated by:

$$\mathrm{tr}\left(\left(\alpha^c I + L\right)^{-1} L\right) = \left\langle r' [p^c]^{-\frac{1}{2}} \left(\nabla\nabla Q^c\right)^{-1} . [p^c]^{-\frac{1}{2}} Lr \right\rangle \tag{A-6}$$

This implies that the $\mathrm{tr}\left(\left(\alpha^c I + L\right)^{-1} L\right)$ term can be calculated by using PCG to solve $\nabla\nabla Q^c . Y = [p^c]^{-\frac{1}{2}} Lr$, then calculating $r' [p^c]^{-\frac{1}{2}} Y$. Given the similarity of (A-6) and the step-size problem (A-4), this is carried out in parallel to the main iteration at very little extra cost. This provides control by providing a target value $\tilde{\alpha^c} = -\mathrm{trace}/(2S)$ towards which the α^c parameter can be adjusted at each iteration.

In the original Skilling and Gull (1991) algorithm, equation (A-5) also provides the following termination condition for the algorithm, which is satisfied when $\Omega \approx 1$.

$$\Omega = -\frac{\mathrm{tr}\left(\left(\alpha^c I + L\right)^{-1} L\right)}{2\alpha^c S\left(p_i, p_j \middle| m_{i,j}\right)} \tag{A-7}$$

Given that the number of distinct good locations Γ^c is known in advance, (A-5) and (A-7) are modified to take this into account, by rescaling α^c with a free parameter θ, shown below. This parameter ensures that when the $\Omega \approx 1$ termination condition is reached $\hat{\alpha}^c = \tilde{\alpha}^c$. More importantly, it also ensures $2(\hat{\alpha^c}\theta)S = \mathrm{tr}((\hat{\alpha^c} I + L)^{-1} L) = \Gamma^c$ and $\chi^2(\hat{p^c}) = N - \Gamma^c$.

$$\theta = -\frac{2\alpha^c S\left(p_i, p_j \middle| m_{i,j}\right)}{\Gamma^c}$$

$$\tilde{\alpha}^c = -\theta \frac{\operatorname{tr}\left(\left(\alpha^c I + L\right)^{-1} L\right)}{2S\left(p_i, p_j \middle| m_{i,j}\right)}$$

$$\Omega = -\theta \frac{\operatorname{tr}\left(\left(\alpha^c I + L\right)^{-1} L\right)}{2\alpha^c S\left(p_i, p_j \middle| m_{i,j}\right)}$$

C Colour figures

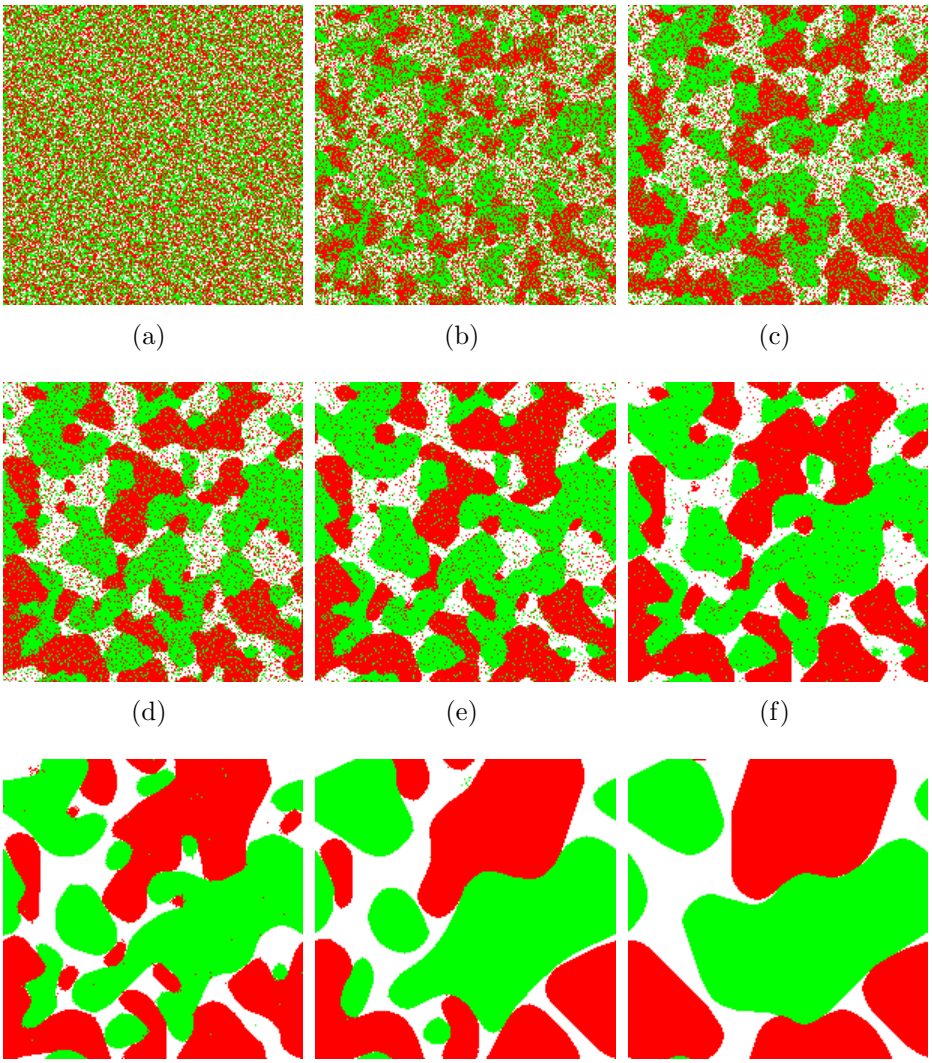
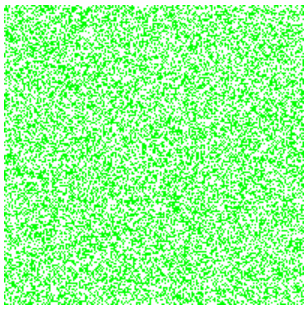
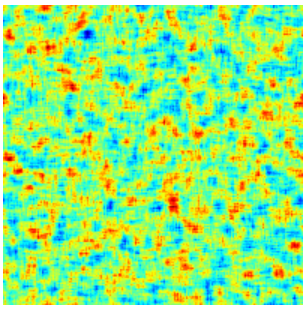


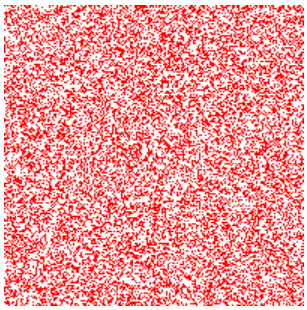
Figure 2: **Emergence of segregation in the Schelling model**



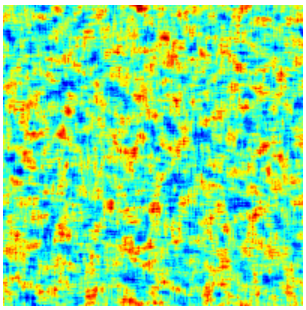
(a) initial green state



(b) initial green convolution



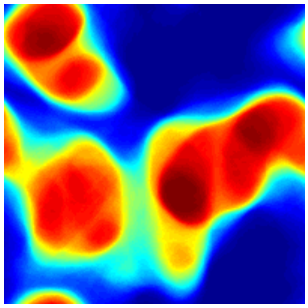
(c) initial red state



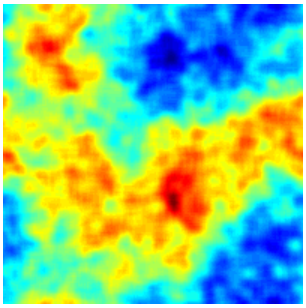
(d) initial red convolution

Figure 3: **Initial condition information**

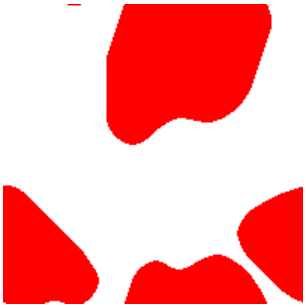
(a) Final green state



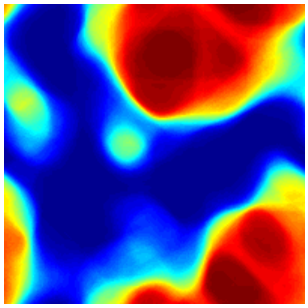
(b) Monte Carlo green frequency



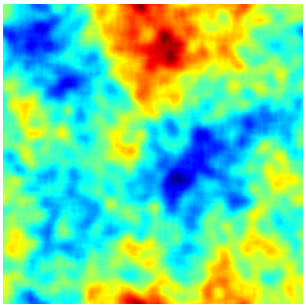
(c) MaxEnt Green probability density



(d) Final red state



(e) Monte Carlo red frequency



(f) MaxEnt Red probability density

Figure 4: **MaxEnt predictions vs. Monte Carlo frequencies**