

A finitary characterization of Ewens sampling formula

D. Costantini*, U. Garibaldi† and P. Viarengo‡

September 2003

Abstract

The clustering of agents in the market is a typical problem dealt with by the new approaches to macroeconomic modeling, that describe macroscopic variables in terms of the behavior of a large collection of microeconomic entities. Clustering has a lot of economical interpretations [3], that are often described by Ewens Sampling Formula (ESF). This formula can be traced back to Fisher as “species sampling”, and its main use was restricted to Genetics for a long time. Contrary to the usual complex derivations [17], we suggest a finitary characterization of the ESF pointing to real economic processes. Our approach is finitary in the sense that we probabilize a system of n individuals considered as a closed system, a population, where individuals can change attributes as time moves on. The intuitive meaning of the probability is the fraction of time the system spends in the considered state of clustering. As ESF is an equilibrium distribution satisfying the detailed balance, some cumbersome properties are derived in a simple way.

1 Introduction

Thirty years ago, in the context of the researches devoted to population genetic, Ewens [12] introduced a distribution that later has been called the Ewens sampling formula (ESF)

$$P(\mathbf{z}) = \frac{n!}{\theta^{[n]}} \prod_{i=1}^n \left(\frac{\theta}{i}\right)^{z_i} \frac{1}{z_i!}, \quad (1)$$

$\theta > 0$, $\theta^{[n]} = \theta(\theta+1)\dots(\theta+n-1)$ is the Pochhammer symbol and $\mathbf{z} = (z_1, \dots, z_n)$, $\sum_{i=1}^n iz_i = n$, is the partition vector to be defined later. (1) has been applied to a wide variety of models for reproduction. A description of ESF, essentially due to Tavaré and Ewens itself, can be found in [19], together with

*Clinical Epidemiology, National Cancer Research Institute, Genoa, Italy

†IMEM-CNR, c/o Department of Physics, University of Genoa, Italy, via Dodecaneso 33, 16146, Genoa, Italy (*e-mail*: garibaldi@fisica.unige.it)

‡National Cancer Research Institute (IST), Genoa, Largo Benzi 10, 16132, Genoa, Italy and Department of Statistical Science, University of Bologna, Bologna, Italy (*e-mail*: paolo.viarengo@istge.it)

structural properties, characterizations, estimation, relations with other distributions, approximations and applications. Among these are quoted: Genetics, Bayesian statistics, permutations, Ecology, Physics, the spread of news and rumors, the law of succession, prime number, random sampling and combinatorial structures.

All the known characterizations of the ESF are either mathematically very sophisticated or appeal to the intuition *via* some urn scheme. They are essentially based upon populations being infinite both in the number of individuals and in that of categories.

In the recent years the (1) has been applied to Economics too (see for example [2][3]). Keeping in mind these new applications, we are suggesting a finitary characterization of the ESF pointing to real economic processes. One of the reasons to publish our characterization is that it has been used by Aoki [4]. This author in many occasion has quoted our finitary characterization that he knows through a mimeographed working paper. This finitary approach allows comprehensible demonstrations, some of them collected in the Appendix.

1.1 Some known characterizations of ESF

In [19] only two kind of characterizations are taken into account. After quoting the pioneer work of Antoniak [1], that obtains ESF is in the field of sampling from a Dirichlet process, the main infinitary characterization is due to Kingman [17], that considers random partitions, an unconventional notion in field other than Biology. The predictive approach starts with the prediction rule of Blackwell-McQueen [5], deepened by Donnelly [11] with respect to Biology, interpreted by Hoppe's urn model [15], finally extended by Hansen and Pitman [14], This approach can be traced to Johnson [20] and more recently to Zabell [24]

All these characterizations take for granted that the n individuals that appear in (1) are a sample from an infinite random population. This is essential from the Bayesian point of view, where the probability (1) is achieved as a mean of the likelihood over the initial (equilibrium) distribution on all possible infinite populations [17]. Also in predictive approach it is essential that the population is at equilibrium, so that time plays no role.

Our approach is finitary in the sense that we want to probabilize our system of n individuals considered as a closed system, that is a finite population whose individuals change attributes as time moves on. Whatever the initial state may be, the probabilistic dynamics is able to drive the system into equilibrium. The intuitive meaning of the probability in (1) is the fraction of time the system spends in the state \mathbf{z} . To summarize this point we say that our characterization does not arrive at the ESF as a distributions ensuing from a sampling procedure from some static superpopulation. We shall prove that the ESF is the equilibrium distribution of a Markov chain ruled by a transition probability build up by exchangeable and invariant creation and destruction terms[7][8][9].

Our method is finitary in a second sense: the "infinite allele model" is obtained as a limit of a finite model. In this way, that can be traced back to

Boltzmann, each assumption can be submitted to concrete inspection. This is of primary importance for applications. In fact a finitary characterization of a probability distribution can be concretely inspected. This allows to check whether the assumptions on which the statistical model is based hold good in the field of application.

2 The Ehrenfest-Brillouin model

Consider a dynamical system composed of n entities and g categories (cells), whose state is described by the non negative integer occupation number vector $\mathbf{n} = (n_1, \dots, n_i, \dots, n_g)$, $n_i \geq 0$, $\sum_{i=1}^g n_i = n$. The system is closed (it is a population), but individuals may change their category at discrete times, so that occupation numbers change step-by-step. The dynamical discrete evolution of the occupation number random variable is a realization of a homogeneous Markov chain, whose stochastic matrix has elements $w(\mathbf{n}, \mathbf{n}') := P(\mathbf{N}_{t+1} = \mathbf{n}' | \mathbf{N}_t = \mathbf{n})$ not depending on time explicitly. A unary moves make an entity to change its state from the cell i to the cell k . $\mathbf{n} = (n_1, \dots, n_i, \dots, n_k, \dots, n_g)$ denotes the initial state and $\mathbf{n}_i^k := (n_1, \dots, n_i - 1, \dots, n_k + 1, \dots, n_g)$ the final one in terms of the coordinates of the starting vector. This transition (a destruction followed by a creation) can be split into two distinct operations [9][13]. The resulting (death-and-birth) transition probability is:

$$w(\mathbf{n}, \mathbf{n}_i^k) = P(\mathbf{n}_i | \mathbf{n}) P(\mathbf{n}_i^k | \mathbf{n}_i) = \frac{n_i}{n} \frac{\alpha_k + n_k - \delta_{i,k}}{\alpha + n - 1} \quad (2)$$

with $\alpha = \sum_i \alpha_i$, and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_g)$ is a vector of parameters that represent initial weights, and $\delta_{i,k}$ is the Kronecker symbol. The meaning of α_i is to allow the accommodation on void cells. We shall consider the case where all $\alpha_i > 0$. It is apparent that, starting from a given \mathbf{n} , by repeated applications of (2), each possible vector of $S_g^n = \{(n_1, \dots, n_i, \dots, n_g) : n_i \geq 0, \sum n_i = n\}$ is reachable, $\#S_g^n = \binom{n+g-1}{n}$. Further all these states are persistent, as no absorbing state exists. It follows that the set of states is ergodic, the chain is irreducible and there exists an (unique) invariant measure $\pi(\mathbf{n})$ on the ergodic set [21]. In addition (2) does not exclude $\mathbf{n}_i^k = \mathbf{n}$, that is the case $j = k$. Hence the chain is aperiodic, and the invariant measure $\pi(\mathbf{n})$ is also the equilibrium distribution on the ergodic set [22], that is

$$\lim_{t \rightarrow \infty} P(\mathbf{N}_t = \mathbf{n} | \mathbf{N}_0 = \mathbf{n}') = \pi(\mathbf{n}) \text{ whatever } \mathbf{n}'$$

The problem of finding out the invariant distribution is solved by the detailed balance equations, that are satisfied by the generalized g -dimensional *Polya* distribution.

$$\pi(\mathbf{n}; \boldsymbol{\alpha}) = \frac{n!}{\alpha^{[n]}} \prod_{i=1}^g \frac{\alpha_i^{[n_i]}}{n_i!}, \quad \mathbf{n} : \sum_{i=1}^g n_i = n \quad (3)$$

Thus $\pi(\mathbf{n}; \boldsymbol{\alpha})$ is dynamically invariant, and by uniqueness theorem is equal to the equilibrium distribution $\pi(\mathbf{n})$.

As $\{\alpha_i\} > 0$ accommodations are positively correlated with the occupation numbers \mathbf{n} . The correlation is large for small α , while it tends to zero for $\alpha \rightarrow \infty$, where the creation probability is $p_k = \frac{\alpha_k}{\alpha}$, independent on \mathbf{n} .

Very simple cases of (3) are [7] the Bose-Einstein distribution, that is uniform on all \mathbf{n} , obtained for $\alpha_i = 1; \alpha = g$; the Maxwell-Boltzmann distribution, obtained for $\alpha_i = c; \alpha = gc; |c| \rightarrow \infty$. The different behavior of the microscopic entities is ascribed to the vector parameter $\boldsymbol{\alpha}$, that determines the type of inter-entity correlation at the moment of choosing the new category, and to $|\alpha|^{-1}$, that fixes its strength. The deviation from independence, known in Economics as “herd behavior”, depends on the ratio of the total initial weight α and the size of the population n .

Considering the mechanism of category change described by (2), the first part of the move is the drawing of an agent, and the first term $\frac{n_i}{n}$ gives the probability that the drawing occurs in the i th cell. This depends only on \mathbf{n} . The second part of the move is the accommodation of the moving agent into the final cell, and the term $\frac{\alpha_k + n_k}{\alpha + n - 1}$ gives the probability that the final cell is the k th cell. This second term can be re-written as $\frac{\alpha p_k + (n-1)f_k}{\alpha + n - 1}$, that is a mixture of the initial probability $\{p_k = \frac{\alpha_k}{\alpha}, \Sigma p_k = 1\}$ and of the current normalized occupation vector $\{f_k = \frac{n_k}{n-1}, \Sigma f_k = 1\}$. The mixture can be interpreted as a randomization of two attitudes [18]. Suppose that the choice is performed in two stages: first the choice of an attitude (theoretical *vs* empirical, with weights α and $n - 1$), followed by the choice of the cell given the distribution attached to chosen attitude.

In Economics this can be interpreted as following: if the agent chooses the theoretical distribution (he behaves as a “fundamentalist”), he is not influenced by his colleagues. In this case we have self-conversion [18]. If the agent chooses the empirical distribution (he behaves as a “chartist”), he chooses a colleague at random and he converts to its strategy [13].

2.1 An auxiliary urn process

Let us consider n random variables Y_1, \dots, Y_n whose range is $(1, \dots, g)$. Suppose that $S_m = (m_1, \dots, m_g)$ is the current occupation vector, that is $m_j = \#\{Y_i = j, i = 1, \dots, m\}$. Let the conditional predictive distribution of Y_m be the following:

$$P(Y_{m+1} = j | m_j, m) = \frac{m_j + \alpha_j}{m + \alpha} \quad (4)$$

for $m = 0, 1, \dots, n - 1$. This is a simple urn (sampling) process (a Polya one, with initial weights $\{\alpha_j\}$) whose n -predictive distribution $P(S_n = \mathbf{n})$ is the same that the equilibrium distribution of our Markov chain $\pi(\mathbf{n}) = \lim_{t \rightarrow \infty} P(X_t = \mathbf{n})$.

2.2 The approach to equilibrium of the mean occupation number

The most elementary transition probability (2) refers to a unary move. A more complicated transition (a m -ary move) [9] is obtained when m units

are drawn without replacement and then reallocated in the cells by recurring application of (4). Let the starting vector be $\mathbf{N}_t = \mathbf{n}$, and let $\mathbf{D}_{t+1} = \mathbf{d} = (d_1, \dots, d_g)$, $\sum_i d_i = m$ the frequency vector of the extracted units, classified by their categories. After the destruction, we reallocate the m units into the cells, and let $\mathbf{C}_{t+1} = \mathbf{c} = (c_1, \dots, c_g)$ $\sum_i c_i = m$ be the frequency vector of the “created” units. The resulting final occupation vector is:

$$\mathbf{N}_{t+1} = \mathbf{N}_t - \mathbf{D}_{t+1} + \mathbf{C}_{t+1} = \mathbf{N}_t + \mathbf{I}_{t+1} \quad (5)$$

where $\mathbf{I}_{t+1} = -\mathbf{D}_{t+1} + \mathbf{C}_{t+1}$ is the increment. Destruction is understood to precede creation. The probabilistic structure of the process is completely given, as $P(\mathbf{D}_{t+1}|\mathbf{N}_t)$ is a Hypergeometric distribution $H(m, \mathbf{n})$, while $P(\mathbf{C}_{t+1}|\mathbf{N}_t, \mathbf{D}_{t+1})$ is a Polya distribution $Po(m, \boldsymbol{\alpha} + \mathbf{n} - \mathbf{d})$.

To simplify we consider the marginal occupation number of the i th category, posing for simplicity $n_i = k$; we denote by K_t, D_{t+1}, C_{t+1} and I_{t+1} the related random variables

$$K_{t+1} = K_t - D_{t+1} + C_{t+1} = K_t + I_{t+1} \quad (6)$$

Suppose $K_t = k$, and collect the complementary $n - k$ statistical units belonging to the remaining $g - 1$ categories in a sole one. The destruction chooses m units from $(k, n - k)$ without replacement, so that it destroys d units from the i th category, and $m - d$ from the rest. The resulting occupation vector is $(k - d, n - k - m + d)$. The creation is modeled by m drawings from a *Polya* urn with initial composition $(\alpha_i + k - d, \alpha - \alpha_i + n - k - m + d)$, so that it creates c units in the category with initial weight α_i , and $m - c$ in the rest with initial weight $\alpha - \alpha_i$. Given the initial state $(k, n - k)$, we get $E(D_{t+1}|k) = m \frac{k}{n}$; adding to the evidence the destroyed state $(d, m - d)$, we have $E(C_{t+1}|k, d) = m \frac{\alpha_i + k - d}{\alpha + n - m}$. It results that $E(C_{t+1}|k) = E(E(C_{t+1}|k, d)) = m \frac{\alpha_i + k - E(D_{t+1}|k)}{\alpha + n - m}$, and finally

$$E(I_{t+1}|k_t) = -\frac{m\alpha}{n(\alpha + n - m)}\left(k_t - \frac{n\alpha_i}{\alpha}\right).$$

The mean increment is null if $\frac{n\alpha_i}{\alpha} = E(n_i) = \mu_i$, that we know to be the equilibrium value of the i th occupation number. It is apparent that

$$r = \frac{m\alpha}{n(\alpha + n - m)} \quad (7)$$

is the rate of approach to equilibrium, that depends on the size n , the total initial weight α and the number of changes m . Hence

$$E(I_{t+1}|k_t) = -r(k_t - \mu_i) \quad (8)$$

$$Cov(K_t, K_{t+1}) = (1 - r)\sigma_i^2$$

where μ and σ_i^2 are the mean and the variance of the marginal Polya distribution. The autocorrelation of each occupation number is universal, that is

$$Corr(K_t, K_{t+s}) = Corr(s) = (1 - r)^s$$

Being a Markov chain, (5) behaves like an $AR(1)$.

The rate is rapidly increasing with m , and it tends to 1 for $m = n$, where the destruction term is trivial and the creation term is just the auxiliary process of Section 2.2. In the domain $\alpha_i > 0$ the rate attains its maximum ($= \frac{m}{n}$) for $\alpha \rightarrow \infty$, while it tends to 0 for $\alpha \rightarrow 0$. Comparing this with the conclusions of the previous section, the process is fast in the independence limit, while it slows down for high “herd behavior”. In this case the representative point needs a lot of transitions in order to move around all the possible states.

3 Infinite number of categories

In the Ehrenfest-Brillouin model [13][10] the distinctive feature of the system are the number of statistical entities n , the number of categories g , and the set of initial weights of the categories $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_g)$. All these parameters are independent, so that they can vary independently, and represent very different scenarios. For instance it is well-known that in the continuum limit $n \rightarrow \infty$ (g and α fixed) the equilibrium distribution $Polya(n, \boldsymbol{\alpha})$ becomes $Dirichlet(\boldsymbol{\alpha})$. Now we want investigate what happen when the number of cell tends to infinity being fixed the size n of the system and the total weight of the initial distribution α . Posing $\alpha_i = \frac{\alpha}{g}$, $\lim_{g \rightarrow \infty} \alpha_i = 0$, and $\lim_{g \rightarrow \infty} \sum_{i=1}^g \alpha_i = \alpha = \theta < \infty$ (for historical reasons: [12]). Note that in the limit $g \rightarrow \infty$ with fixed n any vector $\mathbf{n} = (n_1, \dots, n_g)$ has infinite terms almost all null, and some caution is needful. We have shown in [13] that in the Ewens limit

$$P(\mathbf{n}_i^j | \mathbf{n}) = \begin{cases} \frac{n_i}{n} \frac{n_j - \delta_{i,j}}{\theta + n - 1} & \text{for } j \leq k \\ \frac{n_i}{n} \frac{\theta}{\theta + n - 1} & \text{for } j = k + 1 \end{cases} \quad (9)$$

where the indexes $1, \dots, k \leq n$ relabel the categories initially occupied, and $k + 1$ denote a “new” category $j \notin \{1, \dots, k\}$. Typical examples from Economics are n agents (firms)[4] and g sectors (or goods). The term $\frac{n_i}{n} \frac{n_j - \delta_{i,j}}{\theta + n - 1}$ describes the probability that an unity leaves the i th cluster (strategy) and joints the j th, while $\frac{n_i}{n} \frac{\theta}{\theta + n - 1}$ describes the probability of a change between an existing cluster and the foundation of new one. Summing over all possible starting clusters, we have that $\frac{\theta}{\theta + n - 1}$ is the absolute probability of the emergence of a new cluster at each step. This is a constant, that does not depend on the occupation vector. Due to the absence of the initial weight α_j in the creation term, any state with $n_j = 0$ is an absorbing state; that is, once n_j becomes 0 it cannot undergo a rebirth. There is no equilibrium distribution on \mathbf{n} . In facts all the present k cluster soon or later disappear, or better they go to the “equilibrium” value ($= 0$) with rate $\frac{\theta}{n(\theta + n - 1)}$. All the mass will be transferred to new clusters. With probability 1 a new born clusters cannot be a reincarnation of old ones.

In order to represent these features by a homogeneous Markov chains there are two routes. The first is that of abandoning once for all the labels of the categories to introduce partitions $\mathbf{z} = (z_1, \dots, z_n)$, $z_i = \#\{n_j = i, j = 1, \dots, g\}$, with $k = \sum_{i=1}^n z_i$ and $\sum_{i=1}^n i z_i = n$. The Markov chain induced

on partitions by (9) has indeed ESF as equilibrium distribution, but the demonstration is cumbersome (see Appendix). A second route is an improvement of a complex method suggested by Kelly [16], is the following “label process”.

3.1 The label process

If the population size is n , let introduce $g > n$ fixed labels. Suppose to start with $k \leq n$ distinct clusters. Label them with the first k labels, and $\mathbf{n} = (n_1, \dots, n_k, 0, \dots, 0)$. Define $A(\mathbf{n}) = \{i : n_i > 0\}$ the set of all active (occupied at the moment) labels, then $\#A(\mathbf{n}) = k$ is their number, and $g - k$ is the number of the inactive (unused at the moment) ones. Put $P(\mathbf{n}_i^j | \mathbf{n}) = \frac{n_i}{n} \frac{n_j - \delta_{i,j}}{\theta + n - 1}$ for $j \in A$ (that is $n_j > 0$) while we suppose that in case of innovation a new cluster will be random labeled by one of the $g - k$ available labels:

$$P(\mathbf{n}_i^j | \mathbf{n}) = \begin{cases} \frac{n_i}{n} \frac{n_j - \delta_{i,j}}{\theta + n - 1} & \text{for } n_j > 0 \\ \frac{n_i}{n} \frac{1}{g - k} \frac{1}{\theta + n - 1} & \text{for } n_j = 0 \end{cases} \quad (10)$$

The label process $\mathbf{L}_0, \mathbf{L}_1, \dots, \mathbf{L}_t, \dots$, where $\mathbf{L}_t = (n_1, \dots, n_k, \dots, n_g)$, with the constraint $\sum_1^g n_i = n$, describes the random occupation vector with respect to the g categories. This curious procedure has the merit of producing an ergodic set of label states. In fact a label j that is inactive (and thus $n_j = 0$) can be reactivated, as it can be chosen in case of innovation. The transition probability (10) induces an irreducible aperiodic Markov chain, whose equilibrium distribution can be derived by the detailed balance equations (see Appendix):

$$P(\mathbf{n}) = \frac{(g - k)!}{g!} \frac{n!}{\prod_{j \in A} n_j} \frac{\theta^k}{\theta^{[n]}} = \quad (11)$$

$$= \frac{(g - k)!}{g!} \frac{n!}{\theta^{[n]}} \prod_{i=1}^n \left(\frac{\theta}{i} \right)^{z_i} \quad (12)$$

where in (12) we introduce the partition vector $\mathbf{z} = (z_1, \dots, z_n)$, $z_i = \#\{n_j = i, j = 1, \dots, g\}$, with $k = \sum_{i=1}^n z_i$ and $\sum_{i=1}^n i z_i = n$. Now z_i is the number of active labels (representing clusters) with $n_j = i$, and k is the number of active labels (that is the number of clusters in \mathbf{n}). Note that $\prod_{j \in A(\mathbf{n})} n_j = \prod i^{z_i}$. The label associated with a cluster is not intended to reproduce any physical characteristic of the cluster- it simply labels it. As time goes by, the same label will be used for different clusters, but after the extinction of a cluster an interval will elapse before its label is used again (due to $g > n$). Following the “history” of a label, each passage from 1 to 0 indicates the death of a cluster, and each passage from 0 to 1 indicates a newborn one.

An auxiliary urn process whose n -predictive distribution is (12) is very similar to the so-called Hoppe urn [15](see Appendix). The label process has equilibrium distribution (11), and $E(n_j) = \frac{n}{g}$ for reasons of symmetry. As a consequence, all “true names” of the clusters are lost. This labeling process is analogous to that introduced by Hansen and Pitman in species

sampling [14]: “it is just a device to encode species ... in a sequence of random variables”. A very simple example is shown below (see Fig.1 to Fig.3.).

Now let us consider \mathbf{z} , the statistical description of clusters associated to a label state $\mathbf{L} = (n_1, \dots, n_k, \dots, n_g)$. It is apparent that the number of distinct occupation numbers of n objects in g categories with the same \mathbf{z} is $\frac{g!}{z_0!z_1!\dots z_n!}$, where z_0 is the number of empty categories. Hence the number of distinct label states with the same \mathbf{z} is $\frac{g!}{(g-k)!z_1!\dots z_n!}$.

Further (12) depends only on \mathbf{z} , so that finally:

$$P(\mathbf{z}) = \frac{g!}{(g-k)!z_1!\dots z_n!} P(\mathbf{n}) = \frac{n!}{\theta^{[n]}} \prod_{i=1}^n \left(\frac{\theta}{i}\right)^{z_i} \frac{1}{z_i!} \quad (13)$$

(13) is the ESF [12], that appears as the equilibrium distribution on partitions generated by the label process \mathbf{L} . We can pose $g = n + 1$, that is the minimum number that allows to label unambiguously our death-and birth process.

Example 1 Consider a population of 5 agents, initially separated (they are all singletons). We need at least 6 labels, and we use A, B, C, D, E for labeling the 5 initial clusters.

The six graphics of Fig.1 represent the size of the clusters labeled A, \dots, F for 300 steps. The parameter θ is very small, $\theta = 0.1$, so that herding dominates. In the first 30 steps the clusters B and C are in competition, from $t = 30$ to 180 the cluster B dominates, at $t = 180$ the cluster F begins to compete with B , that is canceled at $t = 200$, where after F dominates. The most probable partition is $(0, 0, 0, 0, 1)$, that is $z_5 = 1$. The theoretical mean life of a cluster is $E(k) \frac{(\theta+n-1)}{\theta} = 49$. The six graphics of Fig.2 represent the size of the clusters when the parameter θ is small, $\theta = 1$, so that herding and pioneering compete. The dynamics is faster, cluster mean life is shorter (the theoretical mean life is 11.4), all labels are really used. The time when $z_5 = 1$ is much shorter than before. This is an indication of the lowered strength of herding. The six graphics of Fig.3 represent the size of the clusters when the parameter θ is not small, $\theta = 5$, so that pioneering prevails. The dynamics is still faster, the theoretical mean life is 5.9. The system never experiences $z_5 = 1$. The time when $z_1 = 5$ is large. This is an indication of the strength of pioneer behavior.

3.2 The probabilistic dynamics of a cluster

It is noteworthy that the probabilistic dynamics of a cluster is independent of the rest of the system. In fact if a cluster has size i , then the probability of an increase or a decrease is:

$$\begin{cases} w(i, i+1) = \frac{n-i}{n} \frac{i}{\theta+n-1} & i = 1, \dots, g-1 \\ w(0, 1) = 0 \\ w(i, i-1) = \frac{i}{n} \frac{n-i+\theta}{\theta+n-1} & i = 1, \dots, g \end{cases} \quad (14)$$

This behavior is a direct consequence of Johnson sufficiency postulate [24]. Hence, given that $E(\Delta i|i) = w(i, i + 1) - w(i, i - 1)$,

$$E(\Delta i|i) = -\frac{i}{n} \frac{\theta}{\theta + n - 1} = -ri \quad (15)$$

where r is the rate of approach of the mean to equilibrium(7) for unary changes. Note that (15) is a particular case of (8) when the equilibrium mean is null. Then each particular existing cluster is driven toward its extinction by (15).

The matrix (14) drives the evolution of the represented cluster. All states except 0 are transient, and the state 0 is absorbing. In order to study the temporal behavior of a cluster, we introduce $w^{(s)}(i, j)$, that is the transition probability from i to j after s steps. Then $w^{(s)}(i, 0)$ is the probability that a cluster of size i is dead after s steps. Hence $w^{(s)}(i, 0) - w^{(s-1)}(i, 0)$ is the probability of dying at the sth step, and thus $\tau_i = \sum_{s=1}^{\infty} s(w^{(s)}(i, 0) - w^{(s-1)}(i, 0))$ is the expected duration time of a cluster of size i . Note that from (14) $w^{(s)}(i, 0) - w^{(s-1)}(i, 0) = \frac{w^{(s-1)}(i, 1)}{n}$, as to die at the sth step is the same as to get the size 1 at the $(s - 1)th$ step and then to die, with transition probability $w(1, 0) = \frac{1}{n}$ from 14. Hence

$$\tau_i = \sum_{s=1}^{\infty} s \frac{w^{(s-1)}(i, 1)}{n} \quad (16)$$

where we pose $w^{(0)}(i, 1) = \delta_{i,1}$. The expected duration time of a newborn cluster (of size 1, hence τ_1) is the mean life of a newborn cluster. It can be calculated exactly by 16, but a shorter formula exists. In fact considering that $\frac{\theta}{\theta+n-1} = u$ is the innovation rate, that is the mean number of new clusters at each step, and posing $E(k)$ as the stationary mean number of clusters, it is apparent that the meanlife of a cluster is just

$$\tau_1 = \frac{E(k)}{u} \quad (17)$$

where

$$E(k) = \sum_{i=0}^{n-1} \frac{\theta}{\theta + i} \quad (18)$$

is the mean number of clusters [19] as a function of θ and n .

Theoretical values in Fig.1,2 and 3 follow (17), while the empirical mean life is calculated as the arithmetic mean of the life durations of all the appeared clusters.

A recurrence relation exists for τ_i , so that it can be solved exactly avoiding the cumbersome (16). In fact if a cluster has size i (and its expected duration time is τ_i), at the following step we find three possibilities, and then three possible expected duration times, all increased by the duration of the step (= 1). Then

$$\tau_i = w(i, i + 1)\{\tau_{i+1} + 1\} + w(i, i)\{\tau_i + 1\} + w(i, i - 1)\{\tau_{i-1} + 1\},$$

that is

$$\tau_i = 1 + w(i, i+1)\tau_{i+1} + w(i, i)\tau_i + w(i, i-1)\tau_{i-1},$$

and substituting (14), reordering and introducing $\Delta_i = \tau_i - \tau_{i-1}, i \geq 2$

$$\Delta_i = \Delta_{i-1} \frac{n-i+1+\theta}{n-i+1} - \frac{n(n-1+\theta)}{(i-1)(n-i+1)}, \quad \Delta_1 = \frac{E(k)}{u}$$

In Fig.4 we represent a simulation, and in 4.1 we show the age of the oldest cluster of the population as a function of time. The size of the oldest cluster in the temporal window $130 < t < 300$ is shown in 4.2. The parameters are $n = 7, \theta = 1$. It is apparent that the oldest cluster is the largest too only in a shorter temporal window ($190 < t < 250$): the correlation between these two order statistics deserve more investigation.

4 Conclusion

Economics is the field of application we have looked at. As showed by Aoki [3] the system could be that of shoppers waiting at one stall in an open air market. The destruction probability is the probability with which a shopper waiting at a stall, leaves this stall for going to another one. On the contrary, the creation probability is the probability with which a shopper goes to a stall coming from a different one. If we suppose that these probabilities are exchangeable and invariant the equilibrium is given by the generalized Polya distribution. In the case in which the number of stalls is much larger than that of the shoppers, that is when Ewens' hypothesis hold and the greatest part of the stalls have no shoppers, the equilibrium distribution is near to the ESF appears. Being strictly finitary our derivation is based on assumptions whose validity can be easily checked. In this ease rests the greatest advantage of the characterization we have suggested. Note that these derivations have nothing to do with sampling, and the resulting meaning of ESF is the fraction of time that the state of systems is described by \mathbf{z} .

In Economics the interpretation is straightforward: you can think to a fixed number of agents (farms, costumers,...) that are wondering around a very large number of possibilities. The only important thing is the ratio between the "herd behavior" (proportional to n) and the pioneer behavior (proportional to θ), that controls the size and the number of existing clusters. In the Ewens limit, when an agent chooses "by himself", he is a pioneer, given the infinite possibilities of available choices. When he chooses the "herd", he joins clusters proportionally to their mass.

In Genetics the ESF comes out for the neutral infinite allele models. A treatise whose conclusions are similar to ours can be found in the Chapter 7 of the book of Kelly [16], where the ESF is derived as equilibrium distribution of a Markov process, that represent a limit case of the reproductive mechanism. Our treatise is confined to discrete times (it is a Markov chain), but about our chain we have a very satisfactory description of the approach to equilibrium.

Further from this point of view is simple to study the typical “life” of any particular cluster, that depends on its size, on n and θ . Computer simulations are easy to perform [13], and can be useful to study properties interesting to economics, like the relationships between size and age, duration time and size, and so on. The label process is conceived as the most natural environment from which the essential information about clusters can be extracted. This interpretation is also useful to derive some properties of the *ESF* in a very natural way.

Appendix

Partition Probabilistic Dynamics

The second route can be introduced taking just the statistical description of clusters $\mathbf{z} = (z_1, \dots, z_n)$ (the “partition”) as state variable of a homogeneous Markov chain. Let \mathbf{u} be the initial partition, \mathbf{v} the partition after the destruction and \mathbf{z} the final partition. If the destruction affects an unit belonging to a cluster of size i , then it destroys a cluster of size i and transform it in a cluster of size $i - 1$, hence

$$v_i = u_i - 1, \quad v_{i-1} = u_{i-1} + 1;$$

Afterward if the created entity increases the clusters of size j , as it joins to a cluster of size $j - 1$, then $z_j = v_j + 1$ and $z_{j-1} = v_{j-1} - 1$.

Hence destructions in i are proportional to the number of entities belonging to a cluster of size i , that is iu_i , while creations in j are proportional to the number of entities belonging to a cluster of size $j - 1$, that is $(j - 1)v_{j-1}$

$$P(\mathbf{z}|\mathbf{u}) = P\left(\mathbf{u}_i^j|\mathbf{u}\right) = \begin{cases} \frac{iu_i}{n} \frac{(j-1)v_{j-1}}{\theta+n-1} & \text{for } j > 1 \\ \frac{iu_i}{n} \frac{\theta}{\theta+n-1} & \text{for } j = 1 \end{cases} \quad (19)$$

as iu_i is the number of agents initially in some i -cluster, $(j - 1)v_{j-1}$ is the number of agents in some $(j - 1)$ -cluster after destruction. If the agent joins to some of them, the number of j -clusters increases by one. Once again the transition probability (19) induces an homogeneous Markov chain irreducible and aperiodic, whose equilibrium distribution (the *ESF*) can be derived by the detailed balance equations, also if this case is more cumbersome than the label cases [12]. The description \mathbf{z} being statistical, it gives less information than the description \mathbf{L} . If two agents exchange their places \mathbf{z} in insensitive to this move, while \mathbf{L} is not.

Equilibrium distribution of the label process

Suppose that $\pi(\mathbf{n}) \propto \frac{f(k)}{g(\mathbf{n})}$, where $k = \#A$, and $A = \{i : n_i > 0\}$. The transition probability $W(\mathbf{n}_i^j|\mathbf{n}) = \begin{cases} \frac{n_i}{n} \frac{n_j}{\theta+n-1} & \text{for } j \in A \\ \frac{n_i}{n} \frac{g-k}{\theta+n-1} & \text{for } j \in \bar{A} \end{cases}$ is such that: for $n_j > 0, k(\mathbf{n}_i^j) = k(\mathbf{n}) = k, W(\mathbf{n}_i^j|\mathbf{n}) = Bn_in_j; W(\mathbf{n}|\mathbf{n}_i^j) = B(n_j + 1)(n_i - 1)$, then

$$\frac{P(\mathbf{n}_i^j)}{P(\mathbf{n})} = \frac{W(\mathbf{n}_i^j|\mathbf{n})}{W(\mathbf{n}|\mathbf{n}_i^j)} = \frac{n_i}{n_i - 1} \frac{n_j}{n_j + 1}$$

that is satisfied by $g(\mathbf{n}) = \prod_{j \in A} n_j$ for $n_j = 0$, if $n_i > 1$ then $k(\mathbf{n}_i^j) = k(\mathbf{n}) + 1$, $W(\mathbf{n}_i^j|\mathbf{n}) = B n_i \frac{\theta}{g-k}$; $W(\mathbf{n}|\mathbf{n}_i^j) = B(n_i - 1)$, finally

$$\frac{P(\mathbf{n}_i^j)}{P(\mathbf{n})} = \frac{n_i}{n_i - 1} \frac{\theta}{g - k}$$

that is satisfied by $f(k) = (g - k)! \theta^k$. In fact $\frac{f(k+1)}{f(k)} = \frac{(g-1-k)! \theta^{k+1}}{(g-k)! \theta^k} = \frac{\theta}{g-k}$ if $n_i = 1$ then $k(\mathbf{n}_i^j) = k(\mathbf{n})$, $W(\mathbf{n}_i^j|\mathbf{n}) = B \frac{\theta}{g-k}$; $W(\mathbf{n}|\mathbf{n}_i^j) = B \frac{\theta}{g-k}$, and $\frac{P(\mathbf{n}_i^j)}{P(\mathbf{n})} = 1$

The auxiliary urn process of the label process

Let us consider n random variables Y_1, \dots, Y_n whose range is a set of $g > n$ labels $L = (l_1, \dots, l_g)$. Suppose that the each type of label has almost n occurrences, so that in principle is possible to label all the sample in the same way. To chose the type of the label to be attached, suppose that exists an urn U , that contains just g balls, each of them with printed a distinct name from (l_1, \dots, l_g) . When Y_1 is extracted, we random chose a ball from U (without replacement) and we label Y_1 with say l_{i_1} . Turning to Y_2 if $Y_2 = Y_1$ it will be labeled by l_{i_1} , otherways the make a second extraction from U , and a new label is given to Y_2 , say l_{i_2} and so on. Proceeding this way all repeated values of Y_i are labelled in the same way, while different values have different labels.

This process can be reduced to a simpler one if labels have no names at all, so that a new label is simply introduced whenever a new value appears. In this case the label i denotes the i th label that has been introduced. The model of this process is the so called Hoppe urn.

$S_m = (m_1, \dots, m_k)$ is the current occupation vector, that is $m_j = \#\{Y_i = j, i = 1, \dots, m\}$, and $k = \#\{m_j > 0\}$ is the number of present labels. If the conditional predictive distribution of Y_m is the following, for $m = 0, 1, \dots, n - 1$:

$$P(Y_{m+1} = j | m_j, m) = \begin{cases} \frac{m_j}{m+\theta} & j \leq k \\ \frac{\theta}{m+\theta}, & j = k + 1 \end{cases} \quad (20)$$

and hence $P(Y_1 = 1) = 1$ by definition.

The new urn (sampling) process (the Hoppe urn, that can be traced back to A.De Moivre, following [24]) consists in an urn that contains initially a white ball whose weight is θ , with the rule that, whenever the white ball is extracted it is painted with a colour not yet present in the sample. After this operation a new white ball (of the same weight θ) is reintroduced in the urn, together with a ball (of weight 1) of the just painted colour. If a colored ball is extracted, it is reintroduced in the urn, together with a ball (of weight

1) of the same colour, just like in the Polya scheme. The probability of a sequence is derived by (20) and results:

$$P(Y_1, \dots, Y_n) = \frac{\theta^k}{\theta^{[n]}} \prod (n_i - 1)! \quad (21)$$

where $n_j = \#\{Y_i = j, i = 1, \dots, n\}$, and $k = \#\{n_j > 0\}$. The n - predictive distribution is

$$\begin{aligned} P(S_n = \mathbf{n}) &= \binom{n-1}{n_1-1} \binom{n-n_1-1}{n_2-1} \cdots \binom{n_{k-1}+n_k-1}{n_{k-1}-1} \frac{\theta^k}{\theta^{[n]}} \prod (n_i - 1)! \\ &= \frac{n!}{n_k(n_k+n_{k-1}) \cdots (n_k+n_{k-1}+\dots+n_1)} \frac{\theta^k}{\theta^{[n]}} \end{aligned} \quad (22)$$

called by Donnelly [11] “the size-biased permutation of the *ESF*”. If we introduce the label urn U whose cardinality is g , and indicate with Y_1^*, \dots, Y_n^* the labels attached at each observation, we have that while $Y_1 = 1$ with certainty, $Y_1^* \in \{l_1, \dots, l_g\}$ equiprobably; and when the second label appears in the Hoppe scheme a new label will be extracted from U . The recursive predictive probability is

$$P(Y_{m+1}^* = j | \mathbf{m}) = \begin{cases} \frac{m_j}{\theta+m} & \text{for } m_j > 0 \\ \frac{1}{g-k(\mathbf{m})} \frac{\theta}{\theta+n} & \text{for } m_j = 0 \end{cases} \quad (24)$$

where $k = \#\{m_j > 0\}$. The probability of a sequence is derived by (24) and results

$$P(Y_1^*, \dots, Y_n^*) = \frac{P(Y_1, \dots, Y_n)}{g(g-1) \cdots (g-k+1)}$$

Hence a sequence Y_1, \dots, Y_n (that is not exchangeable¹) is partitioned into $\frac{g!}{(g-k)!}$ exchangeable sequences Y_1^*, \dots, Y_n^* .

The n -predictive distribution on the occupation numbers on the g states is

$$\begin{aligned} P(S_n^* = \mathbf{n}) &= \frac{n!}{n_1! \cdots n_g!} P(Y_1^*, \dots, Y_n^*) = \frac{n!}{n_1! \cdots n_g!} \frac{(g-k)!}{g!} \frac{\theta^k}{\theta^{[n]}} \prod (n_i - 1)! = \\ &= \frac{(g-k)!}{g!} \frac{n!}{\prod n_j} \frac{\theta^k}{\theta^{[n]}} \end{aligned}$$

that is just same that the equilibrium distribution of the label Markov chain (11), that is $\lim_{t \rightarrow \infty} P(L_t = \mathbf{n})$

The mean number of clusters $k = \#\{m_j > 0\}$ is easily obtained by (20), and results

¹(21) depends only on \mathbf{n} , but this not implies Y_1, \dots, Y_n to be exchangeable. In fact $Y_1 = 1, Y_2 = 2, Y_3 = 1$ and $Y_1 = 1, Y_2 = 1, Y_3 = 2$ are both admissible (equiprobable) with occupation vector $(m_1 = 2, m_2 = 1)$, while the permuted $Y_1 = 2, Y_2 = 1, Y_3 = 1$ is forbidden by construction. This is cleared by the multiplicity factor in $P(S_n = \mathbf{n})$ of equation (22). The fact that (21) is symmetric with regards to n_i , we have that $Y_1 = 1, Y_2 = 2, Y_3 = 1$ and $Y_1 = 1, Y_2 = 2, Y_3 = 2$ are equiprobable, but the number of sequences belonging to $(m_1 = 2, m_2 = 1)$ is greater than the number of sequences belonging to $(m_1 = 1, m_2 = 2)$

$$E(k) = \sum_{i=0}^{n-1} \frac{\theta}{\theta + i} \quad (25)$$

and this remains true also for the L-process.

Mean values of ESF from the label process

Suppose n entities classified into $g > n$ categories. Let us introduce the indicators

$$1_i(j) = \begin{cases} 1 & \text{if } n_j = i \\ 0 & \text{if } n_j \neq i. \end{cases} \quad j = 1, \dots, g.$$

In words the i -indicator of a category is one iff the category contains exactly i entities. The number of clusters of size i is then $z_i = \sum_{j=1}^g 1_i(j)$ and

$$E(z_i) = \sum_{j=1}^g P(n_j = i) = gP(n_j = i) \quad (26)$$

due to the equidistribution of the g categories.

$P(n_j = i)$ is the marginal of (11), and can be calculated by the auxiliary urn process of the label process.

Now $P(Y_1^* = \dots = Y_i^* = j, Y_{i+1}^* \neq j, \dots, Y_n^* \neq j) = \frac{\theta}{g} \frac{(i-1)! \theta^{[n-i]}}{\theta^{[n]}}$, and due to exchangeability of $\{Y_k^*\}$

$$P(n_j = i) = \frac{n!}{i!(n-i)!} P\{Y_k^*\} = \frac{\theta}{gi} \frac{\theta^{[n-i]}/(n-i)!}{\theta^{[n]}/n!} \quad (27)$$

where from $E(z_i)$ follows. As regards to second moments the demonstration is similar, considering that, denoting by R a sequence of $n - i - j$ values different from both a and b ,

$$P(Y_1^* = \dots = Y_i^* = a, Y_{i+1}^* = b, \dots, Y_{i+j}^* = b, R) = \frac{\theta^2}{g(g-1)} \frac{(i-1)!(j-1)! \theta^{[n-i-j]}}{\theta^{[n]}}$$

References

- [1] Antoniak CE. (1974) "Mixtures of Dirichlet processes with applications to bayesian nonparametric problems", *The Annals of Statistics* Vol.2 **6**, 1552-1174.
- [2] Aoki M. (1996) *New Approaches to Macroeconomic Modeling: Evolutionary Stochastic Dynamics, Multiple Equilibria, and Externalities as Field Effects*, Cambridge University Press, New York.
- [3] Aoki M. (2000) "Cluster Size Distributions of Economic Agents of Many Types in a Market", *Journal of Mathematical Analysis and Applications* **249**, 32-52.
- [4] Aoki M. (2002) *Modeling Aggregate Behavior and Fluctuations in Economics*, Cambridge University Press, Cambridge, UK.

- [5] Blackwell D. and MacQueen, J.B. (1973) “Ferguson distribution via Polya urn scheme”, *The Annals of Statistics*, Vol.1, 353-355.
- [6] Costantini D. and Garibaldi U. (1989) “Classical and quantum statistics as finite random processes”, *Found. of Phys.* **19** 743-754.
- [7] Costantini D. and Garibaldi U. (1997) “A probabilistic foundation of elementary particle statistics. Part I”, *Stud. Hist. Phil. Phys.* **28**, 483-506.
- [8] Costantini D. and Garibaldi U. (1998) “A probabilistic foundation of elementary particle statistics. Part II”, *Stud. Hist. Phil. Phys.* **29**, 37-59.
- [9] Costantini D. and Garibaldi U. (2000) “A purely probabilistic representation for the dynamics of a gas of particles”, *Found. of Phys.* **30**, 81-99.
- [10] Costantini D. and Garibaldi U. (2004) “The Ehrenfest fleas: from model to theory”, to appear in *Synthese*, April 2004 issue (139,1).
- [11] Donnelly P. (1986) “Partition structures, Polya urns, the Ewens Sampling Formula, and the ages of alleles”, *Theor. Popul. Biol.* **30**, 271-288.
- [12] Ewens W. J. (1972) “The sampling theory of selectively neutral alleles” *Theoretical Population Biology* **3** 87-112.
- [13] Garibaldi U., Penco M.A. and Viarengo P. (2003) “An exact physical approach for market participation models”, Cowan,R. and Jonard, N. Eds, “Heterogeneous agents, Interactions and Economic Performance”, Lecture Notes in Economics and Mathematical Systems, Volume 521, Springer.
- [14] Hansen B. and Pitman J. (2000) “Prediction rules for exchangeable sequences related to species sampling”, *Statistics and Probability Letters* **46**, 251-256.
- [15] Hoppe F.M. (1987) “The sampling theory of neutral alleles and an urn model in population genetics”, *J. Math. Biol.* **25**(2), 123-159.
- [16] Kelly F.P. (1979) *Reversibility and Stochastic Networks*, John Wiley&Sons.
- [17] Kingman J. F. C. (1978) ‘The representation of partition structures’, *Journal London Mathematical Society* **18** 374-380.
- [18] Kirman A. (1993) “Ants, Rationality and Recruitment”, *The Quarterly Journal of Economics*, **108**, 137-156.
- [19] Johnson N. L., S. Kotz, and N. Balakrishnam (1997) *Multivariate Discrete Distributions*, Wiley New York; Tavaré, Chapter 41.

- [20] Johnson W. E. (1932) *Probability: The Relations of Proposal to Supposal; Probability: Axioms; Probability: The Deductive and the Inductive Problems* *Mind* 41, 1-16, 281-296, 409-423.
- [21] Penrose O. (1970) *Foundations of Statistical Mechanics*, Pergamon Press, Oxford, pages 72-75.
- [22] Sinai Y.G. (1992) *Probability Theory. An introductory Course*. Springer-Verlag Berlin Heidelberg.
- [23] Watterson G. A. (1976) “The stationary distribution of the infinitely-many neutral alleles diffusion model”, *J. Appl. Prob.*, **13**, 639-651.
- [24] Zabell S. (1992) “Predicting the unpredicable” *Synthese* **90**, 205-232.

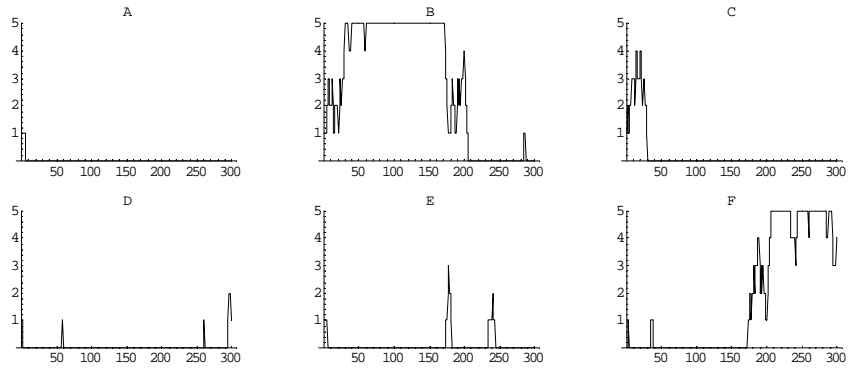


Fig1: $\theta = 0.1$

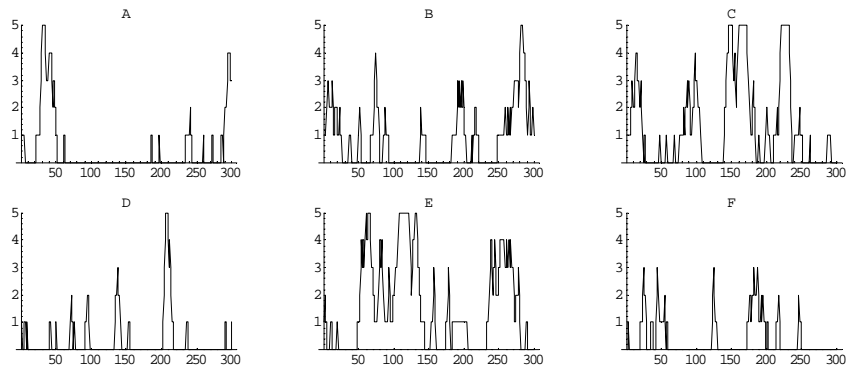


Fig2: $\theta = 1$

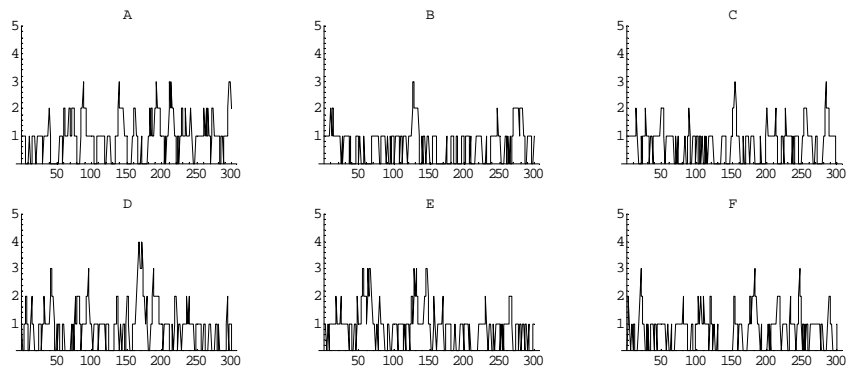


Fig3: $\theta = 5$

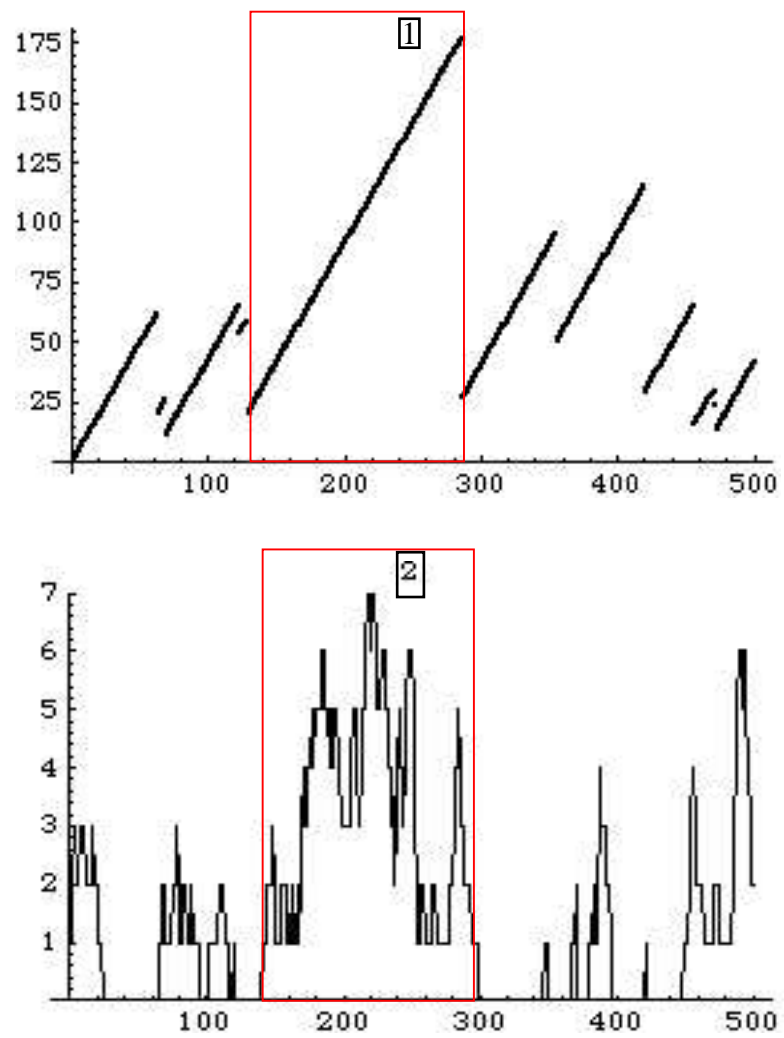


Fig. 4