

**Wild@Ace**<sub>2004</sub>

**Industry and Labour Dynamics II**

**Proceedings of the wild@ace 2004 conference**

**12**

**Herding and clustering in Economics: the Yule-Zipf-Simon model**

**Domenico Costantini, Stefania Donadio, Ubaldo Garibaldi  
and Paolo Viarengo**

# Herding and clustering in Economics: the Yule-Zipf-Simon model

D. Costantini,\* S. Donadio,†U. Garibaldi‡and P.Viarengo§

10 February 2005

## Abstract

The clustering of agents in the market is a typical problem discussed by the new approaches to macroeconomic modelling, that describe macroscopic variables in terms of the behavior of a large collection of microeconomic entities. Clustering is often described by Ewens Sampling Formula (ESF), that admits a very nice interpretation in terms of rational *vs* herding behavior. Focusing on the evergreen problem of the size of firms, we discuss the incompatibility between empirical data and ESF. An alternative model is suggested, inspired to Simon's approaches to the firm size problem. It differs from the Ewens model both in destruction and in creation. In particular the probability of herding is independent on the size of the herd. This very simple assumption destroys the exchangeability of the random partitions, and forbids an analytical solution. Simple computational simulations look to confirm that actually the mean number of clusters of size  $i$  (the equilibrium distribution) follows the corresponding Yule distribution. Finally we introduce a Markov chain, that resembles the marginal dynamics of a cluster, which drives the cluster to the right-censored Yule distribution.

---

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

†INFN, Department of Physics, 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)

# 1 Introduction

The clustering of agents in the market is a typical problem dealt with by the new approaches to macroeconomic modelling, that describe macroscopic variables in terms of the behavior of a large collection of microeconomic entities. Clustering [1] has often been described by Ewens Sampling Formula (ESF) [6]. At variance to the usual complex derivations [13], we have suggested a finitary characterization of the ESF pointing to real economic processes [9], that admits a nice interpretation in terms of rational *vs* herding behavior. In this paper we apply the clustering point of view to the problem of the size of firms. Initially we discuss the compatibility between empirical data [2] and ESF, that is poor. Thus we suggest an alternative model, traced to Simon [15]. A microscopic explanation of the empirical power law should be constructed on some elementary units and their interaction as time goes by. In this frame we hope to find that the power law is obtained as the equilibrium distribution. What is to consider elementary is in some sense conventional. Usually one starts from firms (following Gibrat [10]), that can grow or diminish their size. More deeply we shall start from individual agents, that change job following some well-defined probabilistic rule, and the number and the size of the firms would result as consequences of the individual motions. The stochastic process we are looking for could be a homogeneous Markov chain, where time and state space can eventually achieve continuous limits. To start with a continuous stochastic process describing firms, even if results were satisfactory, would shadow the concrete (discrete) dynamics of agents that determines the final result. This is the case of most of the stochastic explanation quoted in [2]. Here we want to study microscopic explanations of these models, which can be compared with the so-called agent-based computational models [3]. Our ambition is to bridge the gap between agent-based computational models (where there is a lack of probabilistic insight) and stochastic processes (that appear “phenomenological” if they are non “agent-based”). We start from the general notions about state changes occurring to microscopic elements (units, agents). Then we recall the essential of Ewens model, that is the best known structure of Exchangeable Random Partitions [13]. As the line inspired by Simon starts from a pure creation process different from Ewens’ one, we analyze the implications of this variation on the deep properties of the dynamics of the system.

## 2 Destructions, Creations, Attitudes

A dynamical system is composed of  $n$  entities and  $d$  categories (cells), whose state is described by the non negative integer occupation number vector  $\mathbf{n} = (n_1, \dots, n_i, \dots, n_d)$ ,  $n_i \geq 0$ ,  $\sum_{i=1}^d n_i = n$ . The state of the system changes over time as units change cell. The probabilistic dynamics is modelled as an extraction of some units, which temporarily abandon the system, followed by a re-accommodation of the same units, usually into cells different from the original ones. In the time interval from  $t$  to  $t + 1$  the size of the population shrinks as long as selection occurs, and it returns gradually to the original size after all accommodations, such a way that  $\mathbf{n}(t)$  and  $\mathbf{n}(t + 1)$  have the same size  $n$ . The system could be represented by shoppers in an open air market. The simplest selection is the nominal one, where units are extracted individually, like in the pioneer work of the Ehrenfests [5]. A destruction occurs in a stall if that stall loses a shopper, who retires from the stall and sets off for a new stall. An alternative (collective) mechanism is that a stall closes (it is “killed”), so that all its units are compelled to move to new stalls. In this case the destruction probability is the probability for a stall to close. On the contrary, the creation is always intended as individual, that is agents accommodate one by one, and the creation probability for a stall is the probability to be chosen by a moving shopper. This choice may occur because the stall is already occupied (in this case the unit can be influenced by the presence of the “herd”), or even if the stall is empty (in this case the unit follows some “rational”, “fundamentalist” behavior). Thus if a currently empty stall can be chosen again there is no absorbing state, all possible vectors  $\mathbf{n}$  communicate, and under simple conditions  $\mathbf{n}$  can be represented as a homogeneous irreducible aperiodic Markov chain, and the equilibrium distribution  $\pi(\mathbf{n})$  exists. The sequence  $\mathbf{n}(0), \mathbf{n}(1), \dots, \mathbf{n}(t)$  is the joint description of the occupation numbers of the  $d$  categories as time goes by, and each stall has a precise identity that lasts over an infinite time.

If the number of stalls  $d$  tends to infinite (i.e. it is much greater than that of the shoppers), a consistent description requires that each active stall closes when it remains without shoppers, and the closure is for ever (with probability 1). This amounts to say that each precise stall has a vanishing weight for the fundamentalist attitude. Conversely at each accommodation a new stall (which we cannot say) can open, and we suppose that this stall is not a reincarnation of old ones. This happens when the entity chooses to be a

pioneer rather than to join the herd: the two strategies are “orthogonal”, as to follow the initial weights implies to choose a new stall, that is to be a pioneer (innovation). In order to label all stalls in a consistent way, we consider a finite number of sites  $g > n$ , with the proviso that each site hosts a stall at most or it is empty. If  $k$  denotes the actual number of clusters,  $g - k$  is the number of empty sites, so that there is always room for the accommodation of a new stall. Denoting now by  $\mathbf{s}$  the occupation number vector of the  $g$  sites, given that an empty site can be chosen in case of innovation, it is easy to see that  $\mathbf{s}$  can be represented as a homogeneous irreducible aperiodic Markov chain, and the equilibrium distribution  $\pi(\mathbf{s})$  exists. The sequence  $\mathbf{s}(0), \mathbf{s}(1), \dots, \mathbf{s}(t)$  is the joint description of the occupation numbers of the  $g$  sites as time goes by, and each site has a precise identity that lasts over an infinite time. Each occupied site describes an active stall, a cluster. The great difference from the previous case is that every time the occupation number of a site reaches 0, it denotes the death of the presently described cluster, and when it is occupied again it indicates the birth of a new cluster.

While in the finite case each stall has a precise identity that lasts over an infinite time, in the infinite case all present open stalls perish sooner or later, and they are substituted by new stalls, that grow and perish too, so that their identity is given by their birth-and -death date and the site where they lived. Then the framework is quite apt to describe a population where shoppers stand for workers and stalls stand for firms, considered as clusters of aggregated workers. Firms can be born, grow and perish as a result of the state of aggregation on the  $n$  units.

The statistical description of the non empty sites is given by the occupancy (or partition) vector  $\mathbf{z} = (z_1, \dots, z_n)$ , where  $z_i$  denotes the number of sites whose occupation number is  $i$ .  $\sum_{i=1}^n iz_i = n$  is a deterministic constraint, while  $\sum_{i=1}^n z_i = k$  is a random variable (the number of clusters,  $1 \leq k \leq n$ ). From the very definition of  $z_i$  it follows that for any probability distribution  $P(\cdot)$

$$E(z_i) = \sum_{j=1}^g P(s_j = i) \tag{1}$$

holds, and in the case that  $P(s_j = i) = P(i)$

$$E(z_i) = gP(i), \tag{2}$$

where  $P(i)$  is the marginal distribution of a fixed site. If  $E(z_i) = \sum_{\mathbf{z}} z_i Q(\mathbf{z})$ , where  $Q(\mathbf{z})$  denote the equilibrium distribution on  $\mathbf{z}$ , then (2) represents

the main link between the joint distribution on partitions  $Q(\mathbf{z})$  and the site marginal distribution  $P(i)$  at equilibrium (that are both functions of  $\pi(\mathbf{s})$ ).

### 3 The Ehrenfest-Brillouin model and the Ewens limit

To characterize a statistical model we need to define exactly both the destruction and the creation probabilistic mechanism. Recalling the essential of the Ehrenfest-Brillouin model [4], in the simplest case of unary changes (the size of the extracted sample is  $m = 1$ ), we pose that:

i) the selection of the moving agent is individual and random; hence the probability for a stall to loose a shopper (the destruction probability) is proportional to  $n_i$ ; ii) the probability of re-accommodation in the  $j$ th stall (the creation probability) is proportional to  $\theta_j + \nu_j$ , where  $\theta_j > 0$  is the initial weight of the stall and  $\nu_j = n_j - \delta_{j,i}$  is its current occupation number after destruction<sup>1</sup>. The two terms of the accommodation probability can

be interpreted as resulting from two attitudes:  $\frac{\theta_j}{\Sigma\theta_j}$  is the probability of choosing the  $j$ th stall following the initial weight distribution, while  $\frac{\nu_j}{\Sigma\nu_j}$  is the probability of the same choice following the current frequency distribution of the “herd”. In the finite case for each stall  $\theta_j$  is positive, that allows that the stall can be chosen also if it is actually empty. This essential feature makes the dynamics representable by a homogeneous Markov chain which is ergodic, and the equilibrium is given by the generalized Polya distribution [4].

The case in which  $d \rightarrow \infty$ ,  $\theta_j \rightarrow 0$ ,  $\theta = \Sigma\theta_j$  is the Ewens’ limit. Considering sites  $g > n$  sites, the destruction term does not change, while the site creation probability is proportional to  $\nu_j$  if the site is occupied, or to  $\frac{\theta}{g-k}$  if it is empty ( $k$  denotes the actual number of clusters, and  $g - k$  is the number of empty sites). The normalized creation probability can be written as  $\frac{\nu}{\nu + \theta} \frac{\nu_j}{\nu}$

---

<sup>1</sup>The generalization to  $1 < m \leq n$  is obtained performing  $m$  sequential destructions followed by  $m$  sequential creations, both conditioned to the current  $\nu_i$  and  $\nu_j$ . The frequency distribution  $\mathbf{m}$  of the moving units follows the Hypergeometric distribution  $H(\mathbf{m} | \mathbf{n})$ , and the frequency distribution  $\mathbf{m}'$  of the accommodated units follows the Polya distribution  $Po(\mathbf{m}' | \mathbf{n} - \mathbf{m} + \theta)$ .

for  $\nu_j > 0$ , and then  $\frac{\theta}{\nu + \theta} \frac{1}{g - k}$  for  $\nu_j = 0$ , , where  $\frac{\theta}{\nu + \theta} := u$  is the probability of innovation, and  $\frac{\nu}{\nu + \theta} = 1 - u$  is the probability of herding. This creation probability is very similar to Hoppe’s urn scheme[11]<sup>2</sup>. The dynamics of the site occupation number vector  $\mathbf{s}$  is still homogeneous Markov, and the equilibrium distribution of the occupancy (or partition) vector  $\mathbf{z}$  is the  $ESF(n, \theta)$ .

The number  $m \leq n$  of changes-per-step can be fixed at will. The rate of approach to equilibrium of the chain is an increasing function of  $m$ , while the equilibrium distribution is independent on  $m$ , both for the finite and the infinite case. Then the case  $m = 1$  is enough for most applications. In the extreme case  $m = n$  at each step the system is razed to the ground and reconstructed following the Polya (or Hoppe) urn scheme.

We stress some essential features of Ewens model. 1) The birth or death of a cluster is a consequence of the motion of the elementary units. A cluster dies if all its units leave it, a new cluster (a new firm) starts up if a moving agent chooses to be a pioneer. 2) The alternative “innovation *vs* herding” of the moving element has probability depending on the size of the herd . It is essential that the “innovation probability” is a function of two parameters, the weight  $\theta$  of the rational attitude and the size  $v$  of the herd, that is the actual number of fixed units (already accommodated in the sites just at the moment of the re-accommodation of the moving element). The (virtual) size of the population within each step follows the sequence  $(n, n - 1, \dots, n - m, n - m + 1, \dots, n - 1, n)$ . In the general case  $n - m \leq v \leq n - 1$ . The innovation probability at each accommodation is  $u = \frac{\theta}{\theta + v}$ . Hence in the extreme case  $m = n$ , when at each step all units leave the present clusters and then re-accommodate in sequence, the first rebirth occurs when  $v = 0$ , so that it is all committed to the rational (fundamentalist) attitude, while the last rebirth occurs when  $v = n - 1$ , and the weight of the herd is maximum. Only in the case of  $m = 1$  (unary changes), when all rebirths occur at  $v = n - 1$ , the dependence of  $u$  on  $\nu$  is hidden. 3) The equilibrium distribution is independent on  $m$ . 4) The marginal chain is just the projection of the joint dynamics on a fixed site. As we shall see in the following these pleasant properties fail in more general cases.

---

<sup>2</sup>The difference is discussed in [9]

## 4 Ewens marginal description

The marginal description of one fixed site is a reduced description of the system. The above-said dynamics is easily projected on one site exactly. Denoting by  $X_s = i$  the site occupation number after  $s$  steps, and posing  $\{w(i, j) := P(X_{s+1} = j | X_s = i), i, j = 0, 1, \dots, n\}$ , for unary changes the non-vanishing entries of the transition matrix of the marginal chain are:

$$\begin{aligned}
 w(i, i+1) &= \frac{n-i}{n} \frac{i}{\theta+n-1}, i = 1, \dots, n \\
 w(i, i-1) &= \frac{i}{n} \frac{\theta+n-i}{\theta+n-1}, i = 1, \dots, n \\
 w(i, i) &= 1 - w(i, i+1) - w(i, i-1), i = 1, \dots, n \\
 w(0, 1) &= \frac{1}{g - E_n(k)} \frac{\theta}{\theta+n-1} \\
 w(0, 0) &= 1 - w(0, 1)
 \end{aligned} \tag{3}$$

The first three rows in (3) deal with an occupied site, that is its occupation number is  $i > 0$ , and its initial weight is vanishing. All external sites are merged in a single one (the thermostat), whose initial weight is  $\theta$ , and  $n - i$  denotes its occupation number. The last two rows hold if the site is empty (and thus  $i = 0$ ) it can be reactivated, as it can be chosen in case of innovation, representing thus a newborn cluster.  $E_n(k)$  is the mean number of clusters.

Starting from a cluster whose size is  $i$  we must distinguish the history of the cluster (that terminates when the size reaches  $i = 0$ ) from that of the site, that sooner or later reactivates. The equilibrium distribution of the

chain (3) is  $\begin{cases} P(i) = \frac{\theta}{gi} \frac{\theta^{[n-i]}/(n-i)!}{\theta^{[n]}/n!}, i = 1, \dots, n \\ P(0) = 1 - \sum_{i=1}^n P(i) \end{cases}$

Hence by (1):

$$E(z_i) = \frac{\theta}{i} \frac{\theta^{[n-i]}/(n-i)!}{\theta^{[n]}/n!} \tag{4}$$

While  $P(i)$  is the marginal equilibrium probability that a fixed site contains  $i$  units,  $E(z_i)$  the mean number of cluster of size  $i$  in the complete description. In the case of empirical data like Axtell's ones [2] [3], the data

represent the frequency distribution  $\mathbf{z}=(z_1, z_2, \dots, z_n)$ , that from a Markovian point of view is a snapshot, taken from the moving sequence  $\{\mathbf{z}(t)\}$ , driven by the transition probability of the chain. We can compare the empirical distribution with  $E(z_1), \dots, E(z_1)$  which is the equilibrium expected distribution. It is proportional to  $P(i)$ , that is the probability that a random selected cluster has size  $i$ .  $P(i)$  and  $P(\mathbf{z})$  are linked by (1).  $P(i)$  is the equilibrium distribution of the size of a cluster, resulting from the marginal chain that drives the motion of a cluster around its state space.

## 5 Mean values of ESF and the size of firms

For  $n \gg 1$ , using Stirling's approximation, we have [9]

$$E_n(z_i) \approx \frac{\theta}{i} \left(1 - \frac{i}{n}\right)^{\theta-1} \quad (5)$$

Note that  $\sum_{i=1}^n E_n(z_i) = E_n(k)$  is the mean number of clusters, and

$$E_n(k) = \sum_{i=0}^{n-1} \frac{\theta}{\theta + i} \approx \theta \ln \frac{n-1+\theta}{\theta} + \gamma, \quad (6)$$

where  $\gamma$  is the Euler constant.  $E_n(z_i)$  is the mean number of clusters of size  $i$  when the size of the population is  $n$ . The meaning of  $E_n(k)$  in the firm size problem is the equilibrium time average number of firms, that increases logarithmically with total number of individual agents  $n$ .

For a comparison with empirical data,  $E_n(z_i)$  and  $E_n(k)$  are the theoretical quantities candidate to be (dis)proved. In particular  $\frac{E_n(z_i)}{E_n(k)}$  is the expected fraction of firms with size  $i$ . From the marginal point of view, it is the fraction of time spent by a firm in the size  $i$ .

A very simple case of (5) is given by  $\theta = 1$ , where  $E(z_i) = \frac{1}{i}$ , that is the mean size follows a power law. This looks promising, instead it is deceiving. Following Axtell [2], in a realistic case (USA) the rough number of firms is  $k = 5.5 \cdot 10^6$  and the number of agents is about  $n = 105 \cdot 10^6$ .

In the Ewens model, given  $n$  the sole parameter is  $\theta$ , whose best estimate is given by  $k$ . Inverting (6) we get  $\hat{\theta} = 1.24 \cdot 10^6$ , and in that region the normalized (5) is indistinguishable from  $\frac{\hat{\theta}}{i} e^{-i\frac{\hat{\theta}}{n}}$ . This can be represented by

the LogSeries distribution  $L(i) = -\frac{1}{\log(1-y)} \frac{y^i}{i}$ ,  $y = 1, 2, \dots$ , where the parameter is  $y = e^{-\frac{\theta}{n}}$ . In this limit the normalizing constant (6) becomes  $E_n(k) = \theta \ln \frac{n}{\theta}$ . It follows that we have a power law  $\approx \frac{\hat{\theta}}{i}$  for small sizes, but an exponential tail for large firms. This looks sufficient to exclude that a Ewens-like dynamics produces equilibrium probabilities whose tail follows a power law [7].

## 6 Yule-Zipf-Simon herding

The Ewens herding consists in the attitude choice between  $u = \frac{\theta}{\theta + \nu}$  (the novelty) and  $1 - u = \frac{\nu}{\theta + \nu}$  (joining an existing cluster). This probability is not conditioned either on the previous status of the unit, or in the state of the population except for its size  $\nu$ . All units are in the same position with respect to the choice. In the frame of clustering of workers, it amounts to consider for instance two possible states of the units, “self-employed” and “employee”. If a unit founds a new cluster, i.e. he behaves as a pioneer, he enters the state of “self-employed”, and an innovation adds to the firm system. If he joins an existing cluster, he enters the state of “employee”, and no innovation is introduced in the firm system. This choice is independent on the previous status of the unit. Conserving all the meanings, an alternative probabilistic model is that of considering  $u$  and  $1 - u$  as unconditioned from  $\nu$  too. Hence  $u$  is a property of the unit itself, not a balance between an internal strength and an external influence. This leads us into the realm of the Yule-Zipf-Simon herding. In Zipf’s original words [17], when applied to human speech, the Principle of Least Effort produces a vocabulary balance between the “Force of Unification” (the represents the “speaker’s economy”,  $1 - u$ ) and the “Force of Diversification” (“the auditor’s economy”,  $u$ ). A stochastic process built up with these hypotheses was given by Simon [15], inspired also by a previous work of Yule, whose asymptotic stationary solution for relative frequencies tends to the Yule distribution  $f(i) = \rho B(i, \rho + 1)$ ,  $i = 1, 2, \dots$ , where  $B(\cdot, \cdot)$  is the Euler beta function.

Let us consider the following “Zipf’s urn scheme”, whose interpretation

is the construction of a text adding a new word at each step:

$$\begin{cases} P(Y_{n+1} = i | \mathbf{z}(n)) = (1 - u) \frac{iz_i(n)}{n} \\ P(Y_{n+1} = 0 | \mathbf{z}(n)) = u \\ P(Y_1 = 0) = 1 \end{cases} \quad n > 0 \quad (7)$$

where  $n$  is the length of the text so far,  $z_i(n)$  is the number of vocables<sup>3</sup> (distinct words) that appeared  $i$  times,  $iz_i(n)$  is the total frequency of these words (the number of items). Then  $Y_{n+1} = i$  means that the  $(n + 1)$ th word is a vocable that appeared  $i$  times,  $Y_{n+1} = 0$  means that the  $(n + 1)$ th word is a vocable appeared 0 times, that is a new vocable. Its probability is  $u \in [0, 1]$  independent on  $n$ . Note that the first step must result a new vocable, and this must be added in (7). The extreme cases  $u = 0$  and  $u = 1$  are trivially deterministic.  $u = 0$  implies a single cluster, i.e.  $n$  repetitions of the first vocable, and  $P(z_n(n) = 1) = 1$ ; while  $u = 1$  generates  $n$  singletons, i.e. different vocables, and  $P(z_1(n) = n) = 1$ . The evolution of  $\mathbf{z}(n)$  under  $Y_{n+1}$  is the following: the emission of a word appeared  $i$  times transforms a cluster (a vocable) of size  $i$  into a cluster of size  $i + 1$ . Hence it destroys a cluster of size  $i$  (that is  $z_i(n + 1) = z_i(n) - 1$ ) and creates a cluster of size  $i + 1$  (that is  $z_{i+1}(n + 1) = z_{i+1}(n) + 1$ ). If  $Y_{n+1} = 0$  the sole effect is  $z_1(n + 1) = z_1(n) + 1$ .

This scheme is very similar to ‘‘Hoppe’s urn scheme’’ [11], the sole difference is that in Hoppe  $u = \frac{\theta}{\theta + n}$  depends on  $n$ . The difference is decisive in that, while in Hoppe the growth of partitions is exchangeable, in Zipf’s scheme exchangeability fails (see Appendix 1). This feature forbids a closed form for the resulting  $P(\mathbf{z})$ . In fact Simon provides a regression of  $z_{i+1}(n + 1)$ , that is:

$$\begin{cases} E\{z_i(n + 1)\} - z_i(n) = (1 - u) \left( \frac{(i-1)z_{i-1}(n)}{n} - \frac{iz_i(n)}{n} \right), \quad i = 2, \dots, n \\ E\{z_1(n + 1)\} - z_1(n) = u - (1 - u) \frac{z_1(n)}{n} \end{cases} \quad (8)$$

as for  $i = 2, \dots, n$  it is  $\Delta z_i(n) = z_i(n + 1) - z_i(n) = 1$  if  $Y_{n+1} = i - 1$  with probability  $(1 - u) \frac{(i-1)z_{i-1}(n)}{n}$ ; while  $\Delta z_i(n) = -1$  if  $Y_{n+1} = i$  with probability  $(1 - u) \frac{iz_i(n)}{n}$ , and for extreme values  $i = 1$  the meaning is apparent. Taking

---

<sup>3</sup>Suggested in [16]. *F.i.* the text ‘‘home sweet home’’ consists in two vocables and three words.

the means of all equations, that is posing  $E\{z_i(n)\} := Z_i(n)$ , (8) becomes

$$\Delta Z_i(n) = (1-u) \left( \frac{(i-1)Z_{i-1}(n)}{n} - \frac{iZ_i(n)}{n} \right), i > 1 \quad (9)$$

$$\Delta Z_1(n) = u - (1-u) \frac{Z_1(n)}{n} \quad (10)$$

that admits a steady growth for the (mean) number of clusters (vocables) of size  $i$  proportional to  $n$  :

$$\frac{Z_i(n+1)}{n+1} = \frac{Z_i(n)}{n} \quad (11)$$

that is  $Z_i(n+1) = (1 + \frac{1}{n})Z_i(n)$ , or

$$\Delta Z_i(n) = \frac{Z_i(n)}{n}$$

Pay attention that  $\frac{Z_i(n)}{n}$  is not a relative frequency, a misleading term present in Simon. Given that the mean number of vocables is  $k(n) = \sum_{i=1}^n Z_i(n) = 1 + u(n-1) \approx nu$ , i.e. there is a constant inflation of new words, (11) implies that  $\frac{Z_i(n)}{k(n)}$ , that is the relative frequency of vocables of size  $i$  is invariant once the steady growth has been reached. The other important relative frequency is the mass of words,  $\frac{iZ_i(n)}{n}$ , that sums to 1. Then imposing (11) into (9) we get

$$(1-u)(i-1)Z_{i-1}(n) - (1-u)iZ_i(n) - Z_i(n) = 0 \quad (12)$$

$$nu - (1-u)Z_1(n) - Z_1(n) = 0 \quad (13)$$

$$Z_i(n) = \frac{(1-u)(i-1)}{1+(1-u)i} Z_{i-1}(n) = \frac{i-1}{\rho+i} Z_{i-1}(n) \quad (14)$$

with  $\rho = \frac{1}{1-u}$ . The solution of the system of difference equations is obtained solving first (12), that yields  $Z_1^*(n) = \frac{nu}{2-u} = \frac{\rho}{\rho+1} nu$ , and

$$Z_i^*(n) = \frac{i-1}{\rho+i} \frac{i-2}{\rho+i-1} \cdots \frac{1}{\rho+2} Z_1^*(n) = \frac{\Gamma(i)\Gamma(\rho+2)}{\Gamma(\rho+i+1)} Z_1^*(n) = \quad (15)$$

$$= (\rho+1) \frac{\Gamma(i)\Gamma(\rho+1)}{\Gamma(\rho+i+1)} Z_1^*(n) = \rho B(i, \rho+1) nu \quad (16)$$

and the corrected normalized Yule distribution, that holds also for  $i = 1$ , is the following <sup>4</sup>:

$$f_i = \frac{Z_i^*(n)}{nu} = \rho B(i, \rho + 1), i = 1, 2, \dots \quad (17)$$

$$\sum_{i=1}^{\infty} f_i = 1, \rho > 0$$

$$E(i) = \sum_{i=1}^{\infty} i f_i = \frac{\rho}{\rho - 1}, \text{Var}(i) = \frac{E(i)^2}{\rho - 2} \quad (18)$$

The right conclusion of this section is the following: if at each step a unit is added to the system following (7), Simon's exact regression (the mean number of clusters of size  $i$ ) tends to  $Z_i^*(n)$ . Of course  $Z_i(n) = 0$  for  $i < n$ , and  $Z_i(n) \rightarrow Z_i^*(n)$  only for  $n \gg i$  (see Fig.1). In words: considering a fixed size  $i$ , if the population grows you can find a size  $n^*$  such that for  $n > n^*$  is  $Z_i(n) \sim n$ , that is the the steady growth of a pure growth process.

Finally we recall that  $\rho B(i, \rho + 1) = \rho \frac{\Gamma(i)\Gamma(\rho+1)}{\Gamma(i+\rho+1)} = \frac{\rho\Gamma(\rho+1)}{i(i+1)\dots(i+\rho)}$ , which for  $i \gg \rho$  has the form  $f_i \sim i^{-(\rho+1)}$ , that is the prescribed power tail. The case  $u = 0$  is trivial, and cannot be represented by any well-behaved distribution. Instead the limit  $u \rightarrow 0_+$  corresponds to  $\rho \rightarrow 1_+$ , where  $B(i, 2)$  is a well-behaved distribution, with  $E(i) = \infty$ . The relationship between the mutation rate  $u$  and the exponent is  $u(\rho) = \frac{\rho-1}{\rho}$ ,  $\rho > 1$ .  $E(i)$  is finite for  $\rho > 1$ ,  $\text{Var}(i)$  is finite for  $\rho > 2$ .

## 7 Birth and death Simon-Zipf's process

But Simon himself realizes [15] that the complete solution of the problem wants a birth-death process, where the length of the text (the size of the population) is fixed, and clusters are created and destructed with some probability law. He tackles the question in Section III of the paper, in a very confused way. The discussion about the point can be traced to Steindl [16]. In Simon's paper the equation (9) is interpreted in this way: the term  $\frac{1-u}{n}$

---

<sup>4</sup>Simon calls Yule distribution  $B(i, \rho + 1) \frac{Z_i^*(n)}{nu}$ ,  $i > 1$ , as there is a mistaken in his recursive form (15). Anyway the calculated values in Table 1 of the paper are correct

$((i - 1)Z_{i-1}(n) - iZ_i(n))$  is the increment of  $Z_i(n)$  due to a new arrival,  $\frac{-Z_i(n)}{n}$  is the decrement due to the probability of a destruction. If the two contributes balance, it means that  $Z_i^*(n)$  is the invariant distribution of a Markov chain. While in the original (9) the parameter  $n$  represents the size of the text, in the new interpretation it must be assumed as time. This is quite unclear.

Forty years after Simon's seminal work we can give some clear statement. The first statement deals with the lack of exchangeability of the creation process, that forbids close formulas for the sampling distribution, and makes all "proofs" not quite satisfactory. Further let us suppose that our system is "built up" to the size  $n$  following (7), so that the mean values of the cluster size follow exactly Simon's regression. Then we stop the growth, and we leave the system to evolve following unary changes, where a word is cancelled randomly (all words are on a par) and a new word is produced following (7): then the marginal transition probability is just (3), with  $\frac{\theta}{\theta+n-1} = u$ . Hence the equilibrium distribution is the  $ESF(\theta, n)$ , with  $\theta = \frac{u(n-1)}{1-u}$ <sup>5</sup>. Alternatively suppose to pass to  $n$ -ary changes: if at each step the system is razed to the ground and reconstructed following the Zipf scheme, the equilibrium distribution is just Simon's regression. Hence the equilibrium distribution depends strongly on the number of deaths-per-step. If we want to conserve a creation probability like (7), in order to escape the Ewens basin of attraction we must change the death model in a substantial way.

Indeed Simon's (verbal) suggestion is to kill a whole cluster, and then to put back the corresponding number of items into the population following Zipf's scheme. Suppose that destruction consists in eliminating a vocable (an old word, together with all items), while creation consists in adding a number of items equal to the size of the destructed cluster. Supposing first to eliminate a cluster, all clusters being on a par, and then to put back the corresponding number of items into the population following (7). At each step the size of the population first shrinks of a random size  $m$  equal to the size of the destructed cluster, and then returns to its natural size  $n$  via  $m$  accommodations. In words, at each step a firm dies, and all its workers either join existing firms or found new ones with individual probability  $1 - u$  or  $u$  respectively. Newborn firms accommodate in some empty site, and they

---

<sup>5</sup>We may ask which is the rate of approach to equilibrium. As in the Ehrenfest-Brillouin case,  $r = \frac{\theta}{n(\theta+n-1)}$  is the rate of approach of the mean, and  $r^{-1}$  is the associated number of steps. For Axtell's values  $r^{-1} \approx \frac{n(\theta+n)}{\theta} \approx \frac{105 \cdot 106}{1.24} 10^6 = 9 \cdot 10^9$ , that is the number of unary changes needed for the mean to achieve equilibrium. It is like 86 changes for unit.

are labelled by their site. The chain of the site process is ergodic, but an analytic solution is cumbersome being the transition matrix very complex, due to the lack of exchangeability of the creation process. A suggestion about the reasonability of this approach is present in Section II of Simon's quoted paper. Consider (8) and compare it with (9). If all  $z_i(n)$  are equal to their equilibrium mean  $Z_i(n)$  the increment of  $z_i(n)$  is given by (11). Instead if  $z_{i-1}(n) = Z_{i-1}(n)$  but  $z_i(n) = Z_i(n) + \epsilon(i, n)$ , then its mean increment  $E\{z_i(n+1)\} - z_i(n)$  differs from the mean increment by a restoring term  $-(1-u)\frac{i\epsilon(i,n)}{n}$ . In words, a fluctuation from the mean is smoothed by new arrivals on average.

Computational simulations look to confirm that actually the mean number of cluster of size  $i$  (the equilibrium distribution) is closer to the corresponding Yule distribution than Simon's regression (see Fig 2, 4)

## 8 Marginal description of the Simon-Zipf process.

If we consider a  $i$ -cluster belonging to a population whose dynamics is described by the previous mechanism, it is easy to see that the evolution is not a simple function of  $i$ ,  $n-i$  and  $u$ . In fact the death mechanism is deeply different from the Ewens case, as it amounts to "kill" a whole cluster, with equal probability for all the  $1 \leq k \leq n$  existing clusters. Two alternatives are possible: the former is to suppose that at each step a killing occurs; the latter is to pose the probability of killing each cluster equal to  $\frac{1}{n}$ , so that the probability of some killing is  $\frac{k}{n} \leq 1$ , with the proviso that if no killing occurs the system is left in the initial state<sup>6</sup>. With this assumption the probability of a decrease is identical to the probability of the death of the cluster, that is  $w_{i,0} = \frac{1}{n}$ . If death occurs in some other cluster, our  $i$ -cluster can increase up to the (random) size of the destructed cluster. In the case of  $n \gg 1$ , as the joining probability is proportional to  $\frac{i}{\nu} \simeq \frac{i}{n}$ , we can simplify the mechanism supposing that no more than one unit is allowed to eventually join our  $i$ -cluster, with probability given by the usual  $(1-u)\frac{i}{n}$ , that is  $w_{i,i+1} := P(i+1|i) = (1-u)\frac{i}{n}$ . The size is still with probability  $w_{i,i} = 1 - w_{i,0} - w_{i,i+1}$ . We introduce the "re-birthing term" for the label  $w_{0,1}$ , in order to avoid that 0 state is absorbing. Summarizing the transition

---

<sup>6</sup>The latter version can be interpreted supposing that each cluster is represented by its founder, like in the agent-based simulation of Axtell [3]. If at each step an unit is chosen, killing occurs when a founder is chosen

matrix, we get the following square matrix  $\{w_{i,j} : i, j = 0, \dots, n\}$

$1 - w_{0,1}$	$w_{0,1}$	0	0	...	0	0
$\frac{1}{n}$	$w_{1,1}$	$(1-u)\frac{1}{n}$	$(1-u)\frac{1}{n}$	...	0	0
$\frac{1}{n}$	0	$w_{2,2}$	$(1-u)\frac{2}{n}$	...	...	0
...	...	...	...	...	...	...
$\frac{1}{n}$	0	0	0	...	$w_{n-1,n-1}$	$(1-u)\frac{n-1}{n}$
$\frac{1}{n}$	0	0	0	...	0	$w_{n,n}$

We look for the invariant distribution  $P_i, i = 0, \dots, n$ , that must satisfy the Forward Chapman-Kolmogorov equations  $P_i = \sum_j P_j w_{j,i}, i = 0, 1, \dots, n$ . Then

$$P_0 = P_0 w_{0,0} + P_1 w_{1,0} + P_2 w_{2,0} + \dots = P_0 w_{0,0} + (1 - P_0)\varphi, \text{ where } \varphi = \frac{1}{n};$$

$$P_1 = P_0 w_{0,1} + P_1 w_{1,1}.$$

The generic term is  $P_i = P_{i-1} w_{i-1,i} + P_i w_{i,i}$ , that is  $P_i(1 - w_{i,i}) = P_{i-1} w_{i-1,i}$ . Substituting values,

$P_i(\frac{1}{n} + (1-u)\frac{i}{n}) = P_{i-1}\frac{(1-u)^{i-1}}{n}$  that is identical to (14). The explicit solution of  $P_i$  is  $P_i = \frac{\theta_{i-1}}{\varphi+\theta_i} \dots \frac{\theta_1}{\varphi+\theta_2} \frac{\theta_0}{\varphi+\theta_1} \frac{\varphi}{\varphi+\theta_0}$ , where  $\theta_i = w_{i,i+1}$ .

The meaning of  $P_0 = \frac{\varphi}{\varphi+\theta_0}$  is the probability (the fraction of time) that the label is idle. Conditioned to the time when the cluster is alive, or introducing  $\pi_i = P(i|i > 0) = \frac{P_i}{1-P_0} = P_i \frac{\varphi+\theta_0}{\theta_0}$ , then  $\pi_i = \frac{\theta_{i-1}}{\varphi+\theta_i} \dots \frac{\theta_1}{\varphi+\theta_2} \frac{\varphi}{\varphi+\theta_1}$  that does not depend on the ‘‘rebirthing term’’  $\theta_0$ . Note that  $\frac{\theta_{i-1}}{\varphi+\theta_i} = \frac{\frac{i-1}{n\rho}}{\frac{1}{n} + \frac{i}{n\rho}} = \frac{i-1}{\rho+i}$ , so that

$\pi_1 = \frac{1}{1+(1-u)} = \frac{\rho}{\rho+1}, \pi_2 = \frac{\rho}{\rho+1} \frac{1}{\rho+2}, \pi_3 = \frac{\rho}{\rho+1} \frac{1}{\rho+2} \frac{2}{\rho+3}, \dots$  It is apparent that  $\pi_i = \rho \frac{\Gamma(i)\Gamma(\rho+1)}{\Gamma(\rho+i+1)}, i = 1, \dots, n-1$ , that is exactly the value of Yule distribution.

The last term is different, as  $\theta_n = 0$ , and  $\pi_n = \frac{\theta_{n-1}}{\varphi+\theta_n} \pi_{n-1} = \frac{n-1}{\rho} \pi_{n-1}$ , that is different from the expected  $\frac{n-1}{\rho+n} \pi_{n-1}$ . This is not surprising as the domain of  $\pi_i$  is finite, while for the Yule distribution (17) it is unbounded. The equilibrium distribution of the marginal chain is then the right-censored Yule distribution

$$\begin{cases} \pi_i = f_i, & i = 1, \dots, n-1 \\ \pi_n = \sum_{i=n}^{\infty} f_i = \frac{\Gamma(n)\Gamma(\rho+1)}{\Gamma(\rho+n)}, \end{cases}$$

where the last term adds all the tail of the Yule distribution.

This chain has the same structure as the ‘‘problem of runs’’, because at each step the state, if it moves, either increases by one or it jumps down

to zero. The cluster suffers a “sudden death”, that is not a consequence of successive departures of all its units.

## 9 A formal birth-and-death marginal chain

From Yule distribution, the equilibrium mean numbers are  $Z_i = \frac{(1-u)(i-1)}{1+(1-u)^i} Z_{i-1} = \frac{i-1}{\rho+i} Z_{i-1}$ , that is

$$Z_i(1 + (1 - u)i) = Z_{i-1}(1 - u)(i - 1) \quad (19)$$

Interpreting  $\frac{Z_i}{nu}$  as the equilibrium distribution of the marginal process of a firm  $P_i$ , then  $\frac{P_i}{P_{i-1}} = \frac{Z_i}{Z_{i-1}}$  implies that

$$P_i(1 + (1 - u)i) = P_{i-1}(1 - u)(i - 1) \quad (20)$$

A finite probabilistic dynamics that is compatible with (20) is the following: at each step a firm can loose or gain a worker at most (unary moves). The increment-decrement probability is

$$\begin{cases} w_{i,i+1} = \psi_i = A \frac{(1-u)^i}{n} \\ w_{i,i-1} = \phi_i = A \frac{1+(1-u)^i}{n} \end{cases} \quad (21)$$

Then (20) can be interpreted as the balance equation of the chain, with the additional normalization constant  $A$ . This marginal chain produces the righth-truncated Yule distribution

$$\pi_i = f(i|X \leq n) = \frac{f_i}{F_n}, i = 1, \dots, n$$

as equilibrium probability, where  $F_n = \sum_{i=1}^n f_i$ . While in the previous case the marginal chain is obtained from of microscopic dynamics, in this case the destruction-creation mechanism does not look “agent-based” in a clear way. The factor  $\frac{1}{n}$  in  $w(i, i - 1)$  reflects the possibility that one of the  $Z_i$  firms may be destroyed, while the two factors  $\frac{(1-u)^i}{n}$  account for the addition of a moving unit. In fact in the model where a firm is destroyed (closed) with probability  $\frac{1}{n}$ ; the corresponding workers with probability  $u$  open new firms, and with probability  $1 - u$  join to existing firms proportional to their size, it is reasonable to put  $\begin{cases} P(\Delta Z_i = 1) = (1 - u) \frac{(i-1)Z_{i-1}}{n} \\ P(\Delta Z_i = -1) = \frac{Z_i}{n} + (1 - u) \frac{iZ_i}{n} \end{cases}$  if, given that  $n \rightarrow \infty$ ,

we suppose that at most one free worker joins to  $i$ -size or  $(i - 1)$ -size. If he joins a  $(i - 1)$ -cluster,  $Z_i \rightarrow Z_i + 1$ , if he joins a  $i$ -cluster,  $Z_i \rightarrow Z_i - 1$ . In the limit  $n \rightarrow \infty$  we suppose that these three events are disjoint. Then  $E(\Delta Z_i) = 0$  if (19) holds. In (21) the increasing term  $\psi_i$  is proportional to the size of the cluster and to the probability of a herd choice, while the decreasing term  $\phi(i)$  is still proportional to the size of the cluster and to the probability of a herd choice, with added a term that keeps into account the death probability.

The formal chain is smooth, and admits a continuum limit (see Appendix 2).

## 10 Provisional conclusions

The right-censored or truncated Yule distribution (which generate the full Yule distribution for  $n \rightarrow \infty$ , and eventually the Pareto distribution in the continuum limit) appear to be the goal of a probabilistic agent-based finitary approach to clustering whose sizes follow Zipf-like laws. Statistical models of (dis)aggregation must describe the motion of agents from cluster to cluster, and the equilibrium size distribution must result from this dynamics. Simon's Zipf accommodation process produces a cluster distribution which tends locally to the Yule distribution. Introducing dynamics along the same route, to escape Ewens basin of attraction we suppose that: 1) clusters are destroyed by a "sudden death" mechanism; 2) re-accommodations are such that the individual probability of herding does not depend on the size of the herd. Being the accommodation process not exchangeable, no exact calculation is possible, and we must be contented with Simon's recursive regression and simulations. Simulations of the process are shown, and equilibrium distributions closer to the Yule distribution than Simon's regression are obtained even if  $n$  is very small. A marginal chain that approximates the above-said dynamics can be solved exactly and it shows the right-censored Yule distribution as equilibrium distribution of the size of the cluster.

## 11 Appendix 1: Hoppe vs Zipf urn

If the population size is  $n$ , let introduce  $g > n$  sites. Suppose to start from the void state. Let us consider  $n$  random variables  $Y_1, \dots, Y_n$  whose range is  $(1, \dots, g)$ . Suppose that  $S_\nu = (\nu_1, \dots, \nu_g)$  is the current occupation vector of the first  $\nu$  variables, that is  $\nu_j = \#\{Y_i = j, i = 1, \dots, \nu\}$ , and  $k(\nu) = \#\{\nu_j : \nu_j > 0, j = 1, \dots, g\}$  is the number of non empty sites. Let the conditional predictive distribution of  $Y_{\nu+1}$  be the following:

$$P(Y_{\nu+1} = j | \nu_j, \nu) = \frac{\nu}{\theta + \nu} \frac{\nu_j}{\nu} + \delta(\nu_j) \frac{\theta}{\theta + \nu} \frac{1}{g - k(\nu)} \quad (22)$$

for  $\nu = 0, 1, \dots, n-1$ . This is the Hoppe urn [9], with the proviso that a not yet observed value accommodates in an empty site randomly. The first term with  $\nu_j > 0$  describes herding (i.e.  $\frac{\nu}{\theta + \nu}$ ), the second with  $\nu_j = 0$  innovation (i.e.  $\frac{\theta}{\theta + \nu}$ ).

Let us calculate the probability of the sequence  $t = Y_1 = i, Y_2 = j, Y_3 = i$ . Applying recursively the previous (22) it is:  $\frac{\theta}{\theta} \frac{1}{g} \cdot \frac{\theta}{\theta+1} \frac{1}{g-1} \cdot \frac{1}{\theta+2} = \frac{1}{g(g-1)} \frac{\theta^2}{\theta(\theta+1)(\theta+2)}$

Consider now a permutation of  $t$ , f.i.  $Y_1 = i, Y_2 = i, Y_3 = j$ . Its probability is  $\frac{\theta}{\theta} \frac{1}{g} \cdot \frac{1}{\theta+1} \cdot \frac{\theta}{\theta+2} \frac{1}{g-1} = \frac{1}{g(g-1)} \frac{\theta^2}{\theta(\theta+1)(\theta+2)}$ , that is the same. In fact it is a function of the occupation vector. Hence the probability of a site occupation vector  $\mathbf{n}$  is just the probability of a sequence times the number of distinct sequences to  $\mathbf{n}$ , that is  $\frac{n!}{n_1! \dots n_g!} P(t)$ , while the probability of a partition site  $\mathbf{z}$  is  $\frac{g!}{z_0! z_1! \dots z_n!} \frac{n!}{n_1! \dots n_g!} P(t)$ , that is the ESF [9].

Zipf creation process is very similar to (22), that is

$$P(Y_{\nu+1} = j | \nu_j, \nu) = \begin{cases} (1-u) \frac{\nu_j}{\nu} + \delta(\nu_j) \frac{u}{g-k(\nu)}, & \nu > 0 \\ \frac{1}{g}, & \nu = 0 \end{cases} \quad (23)$$

Lets us calculate the probability of the sequence  $t = Y_1 = i, Y_2 = j, Y_3 = i$ . Applying recursively the previous it is  $\frac{1}{g} \frac{u}{g-1} (1-u) \frac{1}{2} = \frac{1}{2} \frac{u(1-u)}{g(g-1)}$ , while  $P(Y_1 = i, Y_2 = i, Y_3 = j) = \frac{1}{g} \cdot (1-u) \cdot \frac{u}{g-1} = \frac{u(1-u)}{g(g-1)}$ , that is different. It is not a function of the occupation vector. Hence to calculate the probability of  $\mathbf{n}$  one must follow all possible sequences, and there is no close formula. In the Zipf process the sentence “home home sweet ” is twice more probable than “home sweet home”.

The text “home sweet home” contains one singleton and one couple. Exchangeability means that all distinct permutations of the text, that is “home

sweet home”, “home home sweet ”, “sweet home home” have the same probability.

## 12 Appendix 2. The continuum limit of the birth-and-death marginal chain.

Let us consider the mean and the variance of the increment of the cluster driven by (21), with  $A = 1$ :

$$E(i) = \psi(i) - \phi(i) = -\frac{1}{n}$$

$$Var(i) = \psi(i) + \phi(i) - (\psi(i) - \phi(i))^2 = \frac{1}{n} + \frac{2(1-u)i}{n} - \left(\frac{1}{n}\right)^2$$

introducing  $x = \frac{i}{n}$ , for large  $n$ ,  $E(i) = -\frac{1}{n}$ ,  $Var(i) \approx 2(1-u)x$ ,

and for the new variable  $x = \frac{i}{n}$ , that is the fraction of elements is:

$$E(x) = -\frac{1}{n^2}, Var(x) \approx \frac{1}{n^2}2(1-u)x$$

Rescaling then time to  $\tau = \frac{1}{n^2}$ , we have that the discrete Markov chain (21) converges to the continuum process  $X$  whose infinitesimal parameters are

$$\begin{aligned} \mu(x) &= -1 \\ \sigma^2(x) &= 2(1-u)x \end{aligned} \tag{24}$$

Solving the diffusion equation with infinitesimal parameters (24), following [12], we find the stationary distribution.

Now  $\frac{2\mu(x)}{\sigma^2(x)} = \frac{1}{(1-u)x} = \frac{\rho}{x}$ . Setting  $s(x) = \exp\{-\int \frac{2\mu(x)}{\sigma^2(x)} dx\} = \exp\{\int \frac{\rho}{x} dx\} = \exp\{\rho \ln x\} = x^\rho$ , the stationary solution has the form

$$f(x) = \frac{1}{s(x)\sigma^2(x)} \{C_1 \int s(x) dx + C_2\}; \text{ posing } C_1 = 0$$

$$f(x) = \frac{C_2 \rho}{2xx^\rho} = \frac{A}{x^{\rho+1}}, \text{ that is the Pareto Distribution.}$$

Computer simulations are in progress, and we mean to compare (24) against Gibrat’s law, whose meaning is not easy to understand.

## 13 Appendix 3. The continuum limit of the Ewens marginal chain.

From (3), posing  $w(i, i+1) = \theta(i)$ ,  $w(i, i-1) = \phi(i)$ , we have:

$$\begin{cases} \theta(i) = \frac{n-i}{n} \frac{i}{n-1+\theta} \approx \frac{i(n-i)}{n^2} \\ \phi(i) = \frac{i}{n} \frac{n-i+\theta}{n-1+\theta} \approx \frac{i(n-i)}{n^2} + \frac{i\theta}{n^2} \end{cases}, \text{ that is}$$

$$\begin{cases} E(i) = \theta(i) - \phi(i) \approx -\frac{i\theta}{n^2} \\ \sigma^2(i) = \theta(i) + \phi(i) - (\theta(i) - \phi(i))^2 \approx 2\frac{i(n-i)}{n^2} + \frac{i\theta}{n^2} \approx 2\frac{i(n-i)}{n^2} \end{cases}$$

Introducing  $x = \frac{i}{n}$ ,  $\begin{cases} E(i) \approx -\frac{\theta}{n}x \\ \sigma^2(i) \approx 2x(1-x) \end{cases}$  we have

$$\begin{cases} E(x) = E(i)/n \approx -\frac{\theta}{n^2}x \\ Var(x) \approx \frac{2x(1-x)}{n^2} \end{cases}$$

Rescaling the time interval to  $\tau = \frac{1}{n^2}$ , then the infinitesimal parameters

are  $\begin{cases} \mu(x) = -\theta x \\ \sigma^2(x) = 2x(1-x) \end{cases}$ .

Then following [12], we look for the stationary distribution.

$$\frac{2\mu(x)}{\sigma^2(x)} = -\frac{\theta}{1-x}, \int -\frac{\theta}{1-x} dx = \theta \ln(1-x), s(x) = Exp[-\theta \ln(1-x)] = (1-x)^{-\theta},$$

$$s(x)\sigma^2(x) = 2x(1-x)(1-x)^{-\theta} = 2x(1-x)^{-\theta+1}, \frac{1}{s(x)\sigma^2(x)} = kx^{-1}(1-x)^{\theta-1}$$

$P(x) = \theta x^{-1}(1-x)^{\theta-1}$ , that is the frequency spectrum [9], i.e. the continuum limit of (5).

## 14 Appendix 4. Simulations

The simulation of Fig.2 consists in a system of  $n = 100$  elements whose initial site state is  $(1, 1, \dots, 1, 0)$ . This corresponds to Axtell' s initial condition [3]. At each step 1) we eliminate a cluster, all clusters being on a par; the size of the population first shrinks of a random size  $m$  equal to the size of the destroyed cluster; 2) we put back the corresponding number of items into the population following (7), returning to the natural size  $n$  *via*  $m$  accommodations. In words, at each step a firm dies, and all its workers either join existing firms or found new ones with individual probability  $(1-u, u)$ . The number of steps is  $t = 2000$ , and we show the case  $u = .2$ . We show (black dots) the empirical distribution of the time average of the fraction of cluster of size  $i$ , while the line represents the Yule distribution with  $\rho = \frac{1}{1-u} = 1.25$ . We give a continuous representation of the Yule (which is discrete) for graphical opportunity. Gray dots represent the normalized Simon regression. In the following graphs (Fig.3) we show the occupation numbers of two sites as time goes by, wherefrom we extract the start-up, growth and death of some clusters. In Fig.4 and Fig.5 we show the same simulation except  $u = .5$ . It is apparent that low values of  $u$  (and  $\rho$ ) imply rapid growth of clusters, large attainable sizes and short mean life; while large values of  $u$  imply small

growth, small concentration and larger mean lives.

## References

- [1] Aoki M. (2000) “Cluster Size Distributions of Economic Agents of Many Types in a Market”, *Journal of Mathematical Analysis and Applications* **249**, 32-52.
- [2] Axtell R.L. (2001) “Zipf Distribution of U.S. Firm Sizes”, *Science*, **293**, 1818-1820.
- [3] Axtell R.L. (1999) “The emergence of firms in a population of agents: Local increasing returns . . . and Power Law Size Distribution”, Working paper n°3, Brookings Institutions.
- [4] Costantini D. and Garibaldi U. (2004) “The Ehrenfest model: from model to theory”, *Synthese*, 139 (1), 107-142.
- [5] Ehrenfest P. and Ehrenfest T. (1907) “Ueber zwei bekannte Einwände gegen Boltzmanns *H*-Theorem”, *Phys. Zeitschrift* **8**, 311-316.
- [6] Ewens W.J. (1972) “The sampling theory of selectively neutral alleles” *Theoretical Population Biology* **3** 87-112.
- [7] Fontanari J.F. and Perlovsky L.I. (2004) *Phys. Rev. E* **70**.
- [8] 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.
- [9] Garibaldi U., Costantini. and Viarengo P. (2004) “A finite characterization of Ewens sampling formula”, *Advances in Complex Systems*, **7** (2), 265-284.
- [10] Gibrat R. (1931) “Les inegalities economique:. . . la loi dell’effet proportionnel”, Libraire du Recueil Sirey, Paris.
- [11] Hoppe F.M. (1987) “The sampling theory of neutral alleles and an urn model in population genetics”, *J. Math. Biol.* **25**(2), 123-159.

- [12] Karlin S. and Taylor H.M. (1981) A second course in stochastic processes, Academic press.
- [13] Kingman J.F.C. (1978) “The representation of partition structures”, *Journal London Mathematical Society* **18** 374-380.
- [14] Kirman A. (1993) “Ants, Rationality and Recruitment”, *The Quarterly Journal of Economics*, **108**, 137-156.
- [15] Simon H.A. (1955) “On a class of skew distribution functions”, *Biometrika*, 425-440
- [16] Steindl J. (1965) “Random processes and the growth of firms”, Charles Griffin&Co.Ltd., London
- [17] Zipf G.K. (1949) “Human Behavior and the Principle of Least Effort”, Addison-Wesley, Reading, MA.

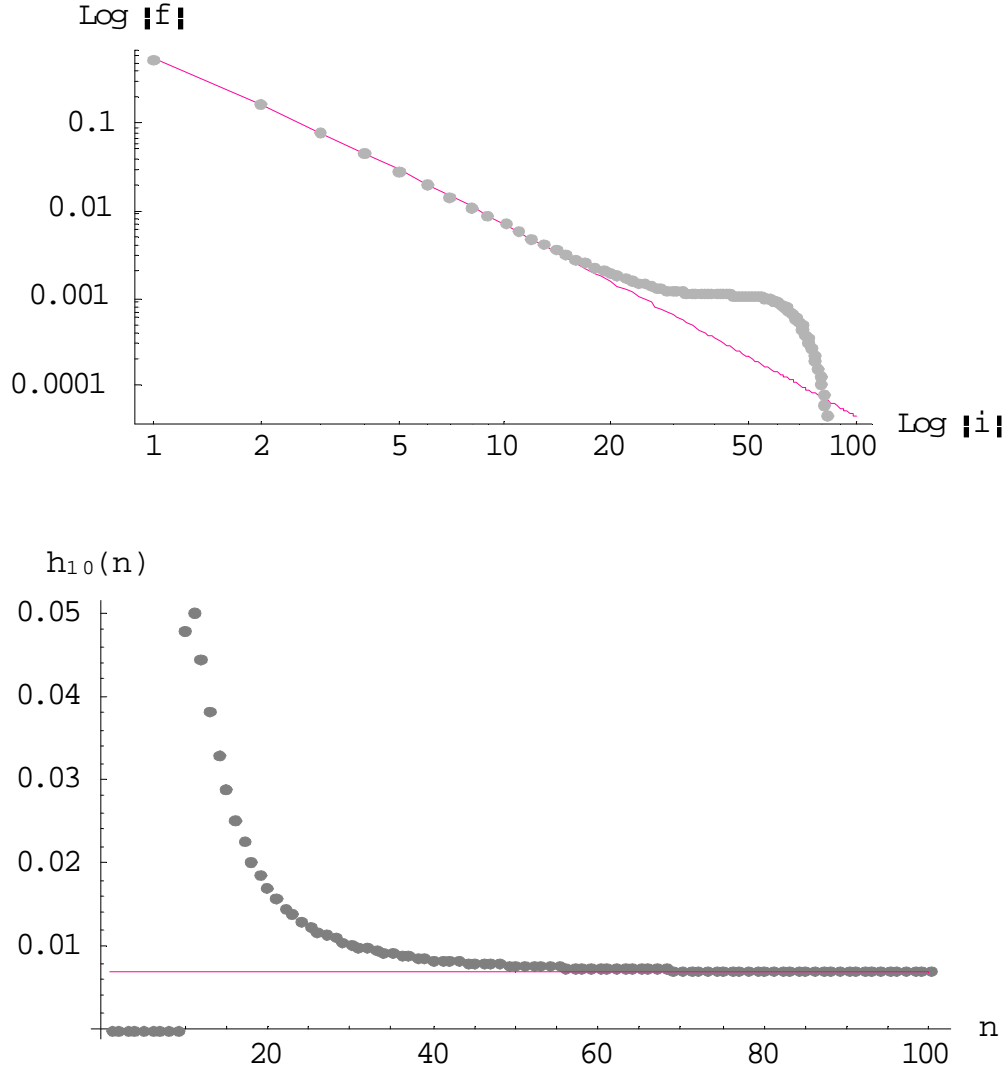


Figure 1: The exact calculation of  $\frac{Z_i(n)}{1+(n-1)u}$ , i.e. the normalized Simon regression (dotted), is compared with the Yule distribution (continuous for graphical opportunity) in the case of  $n = 100$  and  $u = .2$ . We see that the two curves almost coincide for  $i \leq 15$ . The scale of graph is log-log. The second graph shows the convergence of the term of the normalized Simon regression (dotted)  $h_{10}(n) := \frac{Z_i(n)}{1+(n-1)u}$  that describes the mean fraction of the clusters of size 10, with the same term of the Yule distribution as a function of the size of the population.

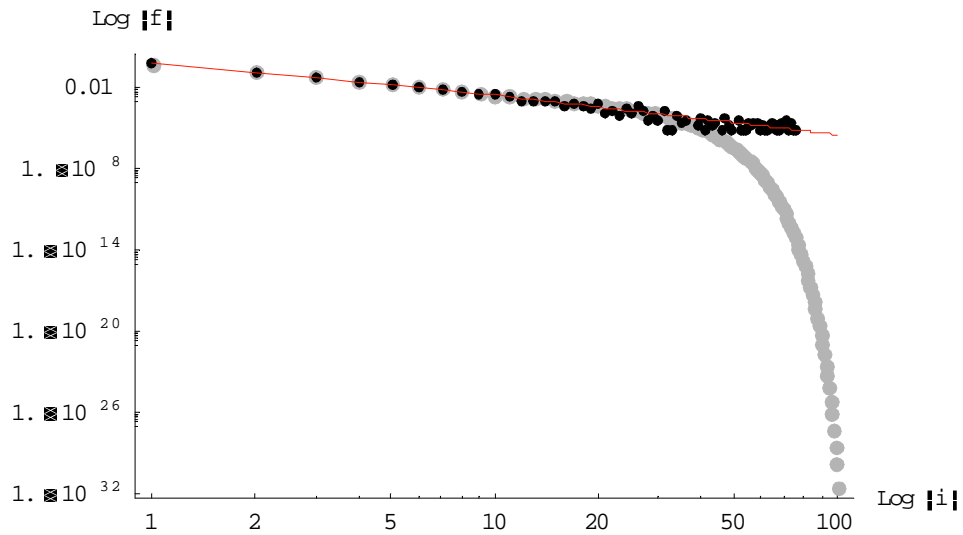


Figure 2: Equilibrium distributions,  $n = 100, u = .2, t = 2000$

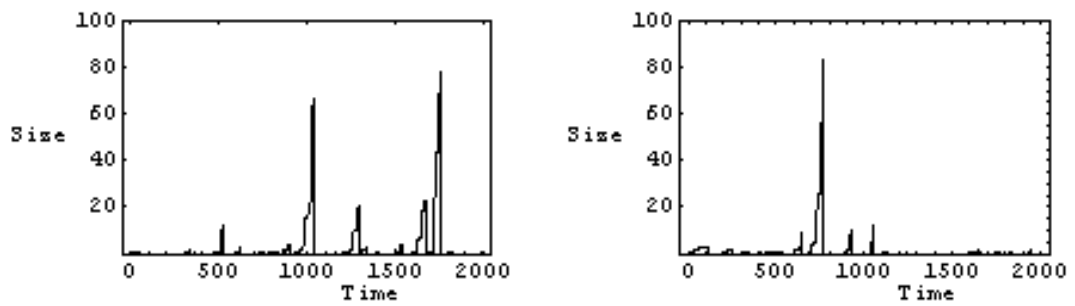


Figure 3: Cluster histories,  $n = 100, u = .2, t = 2000$

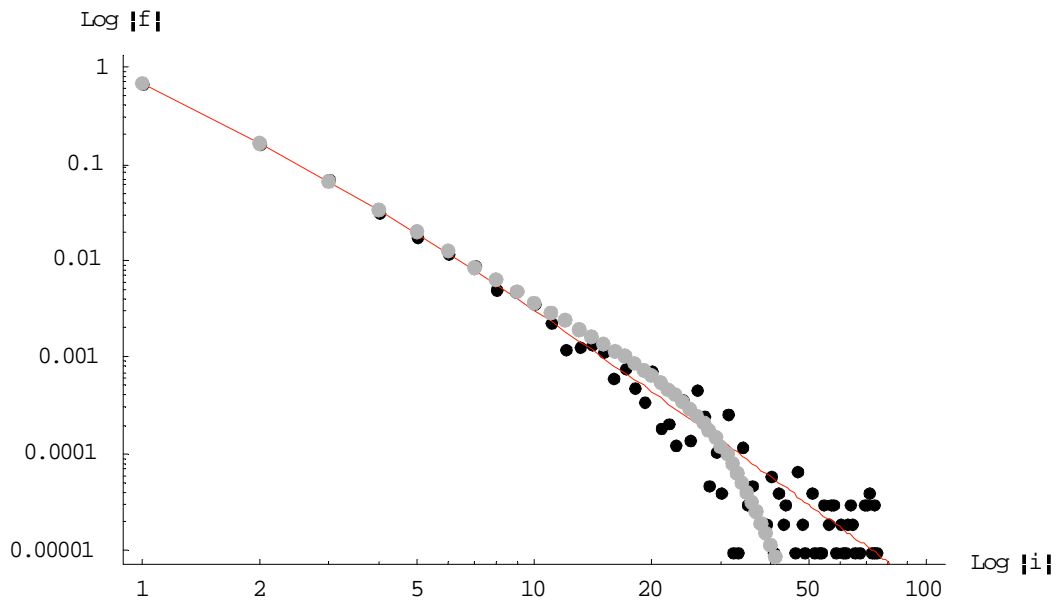


Figure 4: Equilibrium distributions,  $n = 100, u = .5, t = 2000$

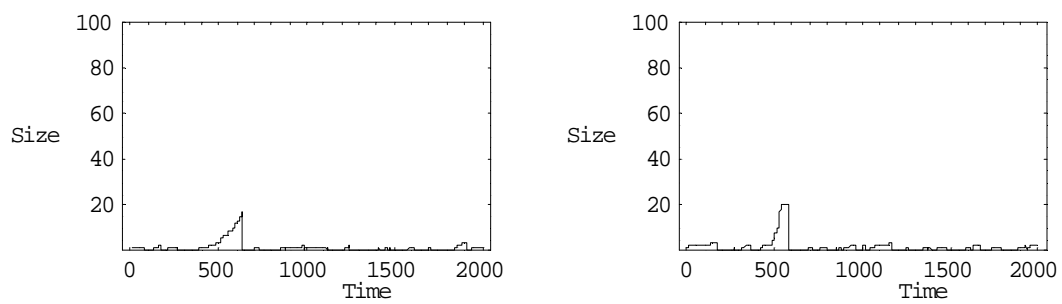


Figure 5: Cluster histories,  $n = 100, u = .5, t = 2000$