Markov chain analysis of an agent based growth model

Eric Thierry, Bruno Gaujal, Laszlo Gulyas, Yuri Mansury

Laboratoire de l'Informatique du Parallélisme

École Normale Supérieure de Lyon
Unité Mixte de Recherche CNRS-INRIA-ENS LYON-UCBL n° 5668

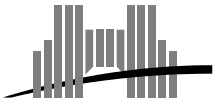
Markov chain analysis of an agent based growth model

Bruno Gaujal ,
Lazlo Gulyas ,
Yuri Surdati Mansuri ,
Eric Thierry

Mar 2007

Research Report N° 2007-15

École Normale Supérieure de Lyon
46 Allée d'Italie, 69364 Lyon Cedex 07, France
Téléphone : +33(0)4.72.72.80.37
Télécopieur : +33(0)4.72.72.80.80
Adresse électronique : `lip@ens-lyon.fr`



Markov chain analysis of an agent based growth model

Bruno Gaujal , Lazlo Gulyas , Yuri Surdati Mansuri , Eric Thierry

Mar 2007

Abstract

In this paper we investigate the asymptotic behavior of a discrete and probabilistic dynamical system which can be described as a growth model where autonomous agents aggregates.
The aim of this paper is to give a mathematical analysis of the dynamics. The analysis uses face homogeneous Markov chains and thanks to this study we validate a conjecture set by Laszlo Gulyas and Yuri Mansury concerning a growth model for cities where simulations had shown that the sizes of the cities asymptotically distribute as a Zipf's law. In light of our analysis, we discuss how the emergence of such a Zipf's law could be expected in Gulyas-Mansury' model and in its variants.

Keywords: discrete probabilistic dynamics, face homogeneous Markov chains, Zipf's law.

Résumé

Nous étudions le comportement asymptotique d'un modèle probabiliste discret qui peut être vu comme un modèle de croissance où des agents s'aggrègent.
L'objectif de ce travail est de donner une analyse mathématique de la dynamique. L'analyse s'intéresse à une chaîne de Markov à faces homogènes. Grâce à cette analyse, nous prouvons une conjecture énoncée par Laszlo Gulyas et Yuri Mansury concernant un modèle de croissance de villes dont les populations se distribuent asymptotiquement selon une loi de Zipf. A la lumière de cette analyse, nous montrons en quoi l'émergence d'une telle loi est naturelle dans le modèle de Gulyas et Mansury, ainsi que dans ses variantes.

Mots-clés: dynamique de modèle probabiliste discret, chaîne de Markov, loi de Zipf.

1 Introduction

Many growth phenomena in human or natural multi-site systems exhibit a quite remarkable property, known as Zipf's law, which says that if one calculates the logarithm of the rank and of the site sizes and plot the resulting data in a diagram you will get a remarkable log-linear pattern. In other words, Zipf's law states that the size v of the r'th largest site is a power of its rank: $v \sim Cr^{-\alpha}$, where α is a constant greater than or equal to 1 and C a positive constant (in Zipf's original work [18], α was 1).

For instance, it appears that Zipf's law governs many features of the Internet. It has been measured that distributions of the sizes of websites in term of the number of pages they include or the number of links given to other sites follow a Zipf's law [1]. This behavior of the World Wide Web has also been observed on the node degree distribution of the graph underlying the Internet backbone (the physical network) [13]. It has lead to the design of growth models in order to explain these features such as growth models by Huberman and Adamic [10] with intuitive and simple assumptions or new random graph growth models by Albert and Barabasi [2] which can be seen as agents aggregating with preferential attachments. In economics, some studies suggest that several types of ranking of companies follow Zipf's laws and that some other economic data is also concerned [3]. Another famous example is the distribution of city populations. After a seminal work of Zipf [19], city sizes growth has been investigated by many authors. One may mention the survey papers by Gabaix [5, 6, 7] which also provide their own explanations of the emergence of Zipf's laws. In another recent paper [15], Soo uses a new data set on 75 countries and shows that the Zipf's law applies (with different parameters according to the countries).

The reader interested by Zipf's laws is referred to the very comprehensive bibliography gathered by Li on `http://linkage.rockefeller.edu/wli/zipf/`.

However, in many cases, either this remarkable feature remains unexplained or rather complicated models are used to derive it.

During the 2003 Exystence Thematic Institute on Complex Systems [11], a simple probabilistic model solely based on individual decisions (autonomous agents) was presented by Laszlo Gulyas. It was introduced as a agent based model to study the growth of city sizes [9, 14]. This model consists in a finite number of sites, which can be seen as the cities, which are ranked by their size, then some agents arrive sequentially and join the cities with the following rule: when arriving, each agent has a random set of eligible cities to join, he then decides to join the largest city proposed to him (and thus increase by one its size). While this process is only based on individual decisions of the agents, Gulyas observed on simulations that the distribution of the city sizes asymptotically behaves like a Zipf's law. He conjectured that its parameter α was equal to 2.

Motivated by this conjecture, our work focusses on the mathematical analysis of a discrete probabilistic dynamics for which Gulyas' model is a special case. As a result, we prove that Gulyas' conjecture is true and that some variants of its model also follow Zipf's law.

As mentioned earlier, a large amount of work has been done to try explaining the recurrent apparition of Zipf's laws when studying city sizes growth. Note that our aim is not to compare all these previous models with Gulyas' new model or to confront its behavior with real data. Our work presents some mathematical tools that prove to be useful in the analysis of such discrete probabilistic models. In particular our proofs are based on a face homogeneous Markov chain description of the dynamics of the system.

In Section 2, we present a general growth model for which we provide an asymptotic

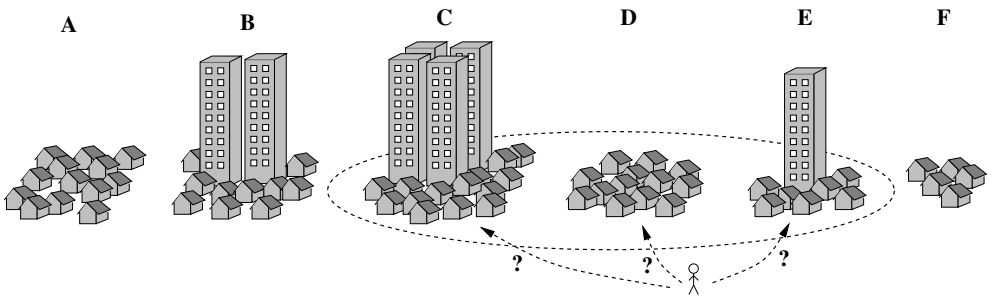


Figure 1: Once the set of possible choices is fixed ({**C,D,E**}), the new inhabitant joins the largest city ({**C**}).

analysis. In Section 3, we precisely describe Gulyas' model and we show how it falls into the framework presented before. Applying the preceding asymptotic analysis, we prove that his conjecture about the distribution of sizes is true. We then show that the general growth model apply to variants of Gulyas' model yielding some other Zipf's laws. Section 5 concludes the paper by summing up features that made the analysis work and by presenting some open problems.

2 The growth model and its asymptotic analysis

2.1 Presentation of the model

We consider the following growth model in discrete time, where agents sequentially aggregate with several sites. We will consider that these are new inhabitants joining cities. Let $C = \{c_1, \ldots, c_n\}$ be a finite set on n cities. The respective populations in the cities at time t are denoted $s_1(t), \ldots, s_n(t)$. The initial population is set to 0 (i.e. $s_i(0) = 0$ for all $1 \leq i \leq n$).

A set of probabilities $\{p_1, \ldots, p_n\}$ is also given ($p_1 + \cdots + p_n = 1$).

At time t, the cities can be ranked by their respective populations, let σ_t be one permutation of $\{1, \ldots, n\}$ such that $s_{\sigma_t(1)}(t) \geq \ldots \geq s_{\sigma_t(n)}(t)$.

At time $t+1$, a new person arrives. This person joins one of the cities with the following probabilities: he joins the i'th largest city $c_{\sigma_t(i)}$ with probability p_i. In other words, $s_{\sigma_t(i)}(t+1) = s_{\sigma_t(i)}(t) + 1$ with probability p_i.

Note that this may induce a new permutation σ_{t+1} if this arrival has altered the order of the sizes of the cities. Note also that the behavior of this system is not completely well defined when several cities have the same size. Indeed, in such cases, several permutations σ_t verify $s_{\sigma_t(1)}(t) \geq \ldots \geq s_{\sigma_t(n)}(t)$ and the growth of each city at time t depends on σ_t. However it should be clear that if we focuss on the sequence of city sizes $s_{\sigma_t(1)}(t) \geq \ldots \geq s_{\sigma_t(n)}(t)$ at time t, it does not depend on the choices of $\sigma_1, \ldots, \sigma_t$ made at all steps up to t.

Our aim is to study the evolution of the city sizes when t goes to infinity and under the assumption that the sequence $p_1, \ldots, p_n$ is strictly positive and **decreasing** ("people prefer larger cities").

For notation purposes, we also set $p_{n+1} = 0$ and we introduce $I(x) = \max\{i \mid 1/p_i \leq x\}$ that will be used later.

2.2 An equivalent model

We consider another representation of the model where each city is replaced by a token and its size is replaced by the position (or slot) of the token over the integer line. If several tokens have the same position, they are piled up on their slot. An example of the old and new representations is given in Figure 2.

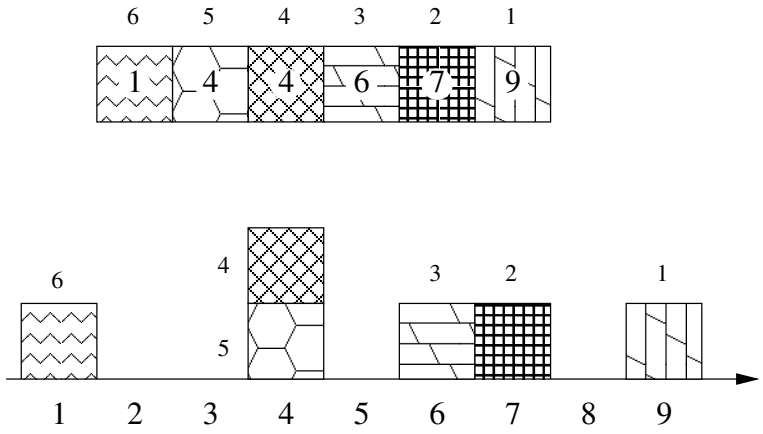


Figure 2: The old representation (size on each token) and the new one representation (each token at a position equal to its size).

Note that the selection process for the new model makes one token (the right most among all the picked tokens) move to the right by one slot. The models differ slightly is in the presence of ties. In the new model, if two tokens are in the same position, we decide to select the one on "top", rather than selecting arbitrarily. Similarly when one token moves to the right and joins other tokens at its new position, we choose to insert it at the bottom of the stack. This convention has one main feature which will be used in the following: the relative order between the tokens never changes: the tokens cannot overtake each other. Without loss of generality, we number them in decreasing order: the rightmost token is labeled by 1 and the last token (leftmost) is labeled by n (in a stack of tokens, the labels are decreasing from top to bottom). The position (or value) of token i is denoted by v_i.

It should be obvious that the probabilistic laws of both models are the same. In the remaining, we will rather use the token model where no overtaking can occur because it makes the model easier to describe, however, all the analysis below could have been made using the original model.

2.3 Markov chain analysis

The state of the system at step t $V(t) \stackrel{\text{def}}{=} (v_1(t), \cdots, v_n(t))$ obviously forms a Markov chain with a state space included in $\mathbb{N}^n$, since the future evolution only depends on the present state. However, its structure is rather complicated and cannot to be analyzed directly. It should be clear that all the tokens "tend" to move to the right during the evolution. However, to

compute the speed of this shift to the right, one needs to look first at second order quantities such as the relative positions of the tokens.

Let $g_i, 1 \leq i \leq n$, be the size of the gap between token i and token $i+1$. For all $1 \leq i \leq n-1$,

$$g_i = v_i - v_{i+1} \text{ and } g_n = v_n$$

We also denote by h_i the smallest label of all tokens piled above token i.

$$h_i = \min\{j \leq i, v_j = v_i\} = \min\{j \leq i, \prod_{k=i}^{j} g_k = 0\}.$$

It should also be obvious that $G = (g_1, \cdots, g_n)$ is a Markov chain living in $\mathbb{N}^n$, where the space can be divised in 2^n zones $Z_S = \{(g_1, \cdots, g_n) \in \mathbb{N}^n | \forall i \in S,\ g_i = 0\}$ for all $S \subseteq \{1, \ldots, n\}$ on which the chain is homogeneous. It is a face-homogeneous Markov chain [4].

The chains $V = (v_1, \cdots, v_n)$ and $G = (g_1, \cdots, g_n)$ have the following transition kernels, respectively denoted by O and P:

At step $t+1$, with probability p_i, $1 \leq i \leq n$,

$$\begin{aligned} (v_1, \cdots, v_n) &\xrightarrow{p_i} (v_1, \cdots, v_{h_i} + 1, \cdots, v_n) \\ (g_1, \cdots, g_n) &\xrightarrow{p_i} (g_1, \cdots, g_{h_i - 1} - 1, g_{h_i} + 1, \cdots, , g_n) \end{aligned}$$

The kernel P of the chain G is displayed in Figure 3 for $n = 2$.

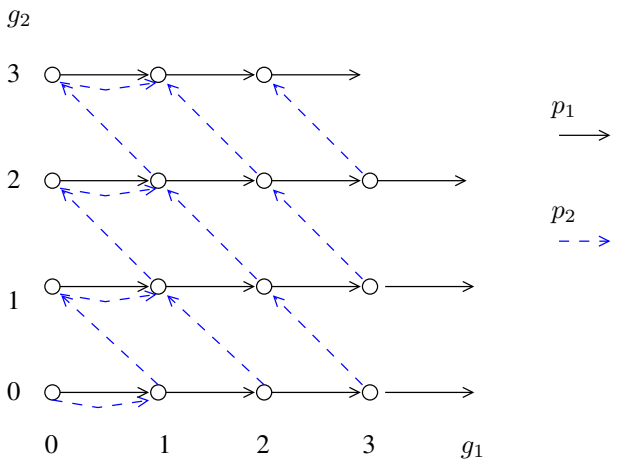


Figure 3: The kernel P for $n = 2$.

Lemma 1. *The chain $G = (g_1, \cdots, g_n)$ reaches states where $g_i > 0$ for all $1 \leq i \leq n$, in finite time, almost surely.*

Proof. The proof involves the construction of a new chain $B(t) = (b_1(t), \cdots, b_n(t))$ with a bounded state space. As shown below, this new chain will reach states with no null components with probability one and this will imply that G reaches states where $g_i > 0$ for all i, with probability one.

Here is the description of $B(t) = (b_1, \cdots, b_n)$. The state space is $\{0,1,2\}^n$. We define a function f_i similar to h_i but for the chain B.

$$f_i = \min\{j \le i, \prod_{k=i}^{j} b_k = 0\}.$$

The transition kernel W of B is the following. For each $1 \le i \le n$,

$$\begin{array}{lll} (b_1, \cdots, b_n) \xrightarrow{p_i} (b_1, \cdots, b_{f_i-1} - 1, b_{f_i} + 1, \cdots, b_n) & \text{if } f_i = i \text{ and } b_{f_i} < 2, \\ (b_1, \cdots, b_n) \xrightarrow{p_i} (b_1, \cdots, b_{f_i-1} - 1, b_{f_i}, \cdots, b_n) & \text{if } f_i = i \text{ and } b_{f_i} = 2, \\ (b_1, \cdots, b_n) \xrightarrow{p_i} (b_1, \cdots, b_{f_i-1} - 1, b_{f_i}, \cdots, b_n) & \text{if } f_i < i. \end{array} \tag{1}$$

The finite chain B is displayed in Figure 4. It should be clear that the chain $B(t)$ can reach state $(2, \cdots, 2)$ from any starting point, with positive probability.

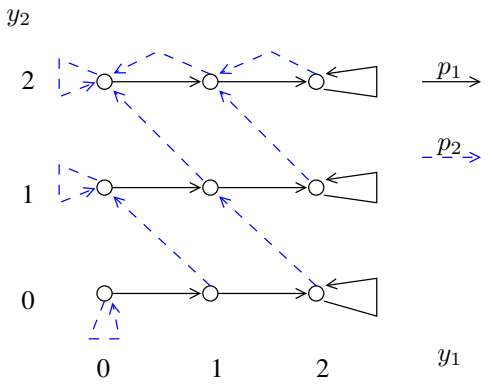


Figure 4: The kernel W when $n = 2$.

Now, we claim that if in the initial state, $B(0) \le G(0)$, component-wise, then at any time t, $B(t) \le_{st} G(t)$. This can be shown by induction on t. Assume that $B(t) \le_{st} G(t)$, then, by coupling the transitions in both chains and looking at the transition kernel, we will show that $B(t+1) \le_{st} G(t+1)$.

With probability p_i,

$$G(t+1) = (g_1(t), \cdots, g_{h_i-1}(t) - 1, g_{h_i}(t) + 1, \cdots, , g_n(t))$$

$$\begin{array}{rcll} B(t+1) & = & (b_1(t), \cdots, b_{f_i-1}(t) - 1, b_{f_i}(t) + 1, \cdots, , b_n(t)) & (2) \\ & or & (b_1(t), \cdots, b_{f_i-1}(t) - 1, b_{f_i}(t), \cdots, , b_n(t)). & (3) \end{array}$$

First, note that if $f_i = h_i$, then it is straightforward to check that in both cases, if $B(t) \le_{st} G(t)$, then $B(t+1) \le_{st} G(t+1)$.

Since $B(t) \le_{st} G(t)$, the only case remaining is $f_i < h_i \le i$. This implies that $b_{h_i-1}(t) = 0$ and the transition kernel of B uses case (3).

Since $G(t+1)$ remains non-negative by construction, this implies

$$\begin{array}{rcll} g_{h_i}(t+1) = g_{h_i}(t) + 1 & \ge & b_{h_i}(t) + 1 \ge b_{h_i}(t+1) & \\ g_{h_i-1}(t+1) = g_{h_i-1}(t) - 1 & \ge & 0 = b_{h_i-1}(t) = b_{h_i-1}(t+1) & \\ g_{f_i}(t+1) = g_{f_i}(t) & \ge & b_{f_i}(t) = b_{f_i}(t+1) & (\text{ if } f_i < h_i - 1) \\ g_{f_i-1}(t+1) = g_{f_i-1}(t) & \ge & b_{f_i-1}(t) \ge b_{f_i-1}(t+1). & \end{array}$$

All the other coordinates are unchanged in G and in B with probability p_i.

This coupling shows by induction that $G(0) \leq_{st} B(0)$ implies for all t, $G(t) \leq_{st} B(t)$.

Now,to finish the proof, it is enough to note that since $B(t)$ reaches states where $b_i > 0$ for all i in finite time , then so does $G(t)$ with probability one.

□

The second step of the analysis of the chain $G(t)$ is to show that there is a drift towards states where the gaps between the tokens grow.

Lemma 2. *After some finite time t_0, the chain $G(t)$ has all its components that remain positive almost surely. Furthermore, for all i,*

$$\lim_{t\to\infty} \frac{g_i(t)}{t} = p_i - p_{i+1}.$$

Proof. Let us first construct a Markov chain $S(t) = (s_1(t), \cdots, s_n t))$ with the following kernel Q.
If X has all its components positive , then $Q(X,Y) = Pr(G(1) = Y | G(0) = X) = P(X,Y)$.
If X has some null coordinates then using Lemma 1, the time

$$\nu(X) = \inf\{t \geq 0, s.t. G(t) \geq (1, \cdots, 1), G(0) = X\}$$

is finite almost surely and uniformly bounded. Indeed, for all X with null coordinates, let us denote by $X_2 \stackrel{\text{def}}{=} (\max(X_1, 2), \cdots (\max(X_n, 2))$ the state X with all its components truncated at 2. Then

$$\begin{array}{rcl} \nu(X) & \leq_{st} & \inf\{t \geq 0, s.t. B(t) \geq (1, \cdots, 1), B(0) = X_2\} \\ & \leq_{st} & \max\limits_{Z \in \{0,1,2\}^n} \inf\{t \geq 0, s.t. B(t) \geq (1, \cdots, 1), B(0) = Z\}, \end{array}$$

which does not depend on X. Then, we set $Q(X,Y) = Pr(G(\nu) = Y | G(0) = X) = P^{\nu(X)}(X,Y)$.

Now, let us consider the Lyapounov function $\min : \mathbb{R}^n \to \mathbb{R}$, which has bounded average increments for the kernel Q. For all $X \in \mathbb{N}^n$, if X has some null components

$$\mathbb{E}(\min(s_1(1), \cdots, s_n(1)) - \min(s_1(0), \cdots, s_n(0)) | (s_1(0), \cdots, s_n(0)) = X) \geq 1,$$

and if X has no null components

$$\mathbb{E}(\min(s_1(1), \cdots, s_n(1)) - \min(s_1(0), \cdots, s_n(0)) | (s_1(0), \cdots, s_n(0)) = X) \geq \min_{i=1}^{n}(p_i - p_{i+1}).$$

An adequate version of Foster's theorem, (see for example [12]), shows that the chain S is transient and $\min(s_1(t), \cdots, s_n(t))$ goes to infinity when t goes to infinity. This implies that there exists a finite time τ such that for all $t \geq \tau$, $s_i(t)$ remains strictly positive for all i. Also, the transition kernels of S and G coincide as soon as $X > 0$ for all components. Therefore, there exists two finite random variables n_0 and m_0 such that $S(t) = G(t + m_0)$ for all $t \geq n_0$. This means that after a finite time t_0, all the components of G remain positive. This ends the proof the first statement.

As for the second statement, consider $(\delta_j)_{j\in\mathbb{N}}$, an i.i.d. sequence such that $\delta_j = 1$ with probability p_i, $\delta_j = -1$ with probability p_{i+1}, and $\delta_j = 0$ with probability $1 - p_{i+1} - p_i$.

Since $g_i(t)$ is strictly positive almost surely after a finite time t_0 and since for all $t \geq t_0$, $g_i(t) - g_i(t_0) = \sum_{j=t_0}^{t} \delta_j$, the strong law of large numbers yields

$$\lim_{t\to\infty} \frac{g_i(t)}{t} = \lim_{t\to\infty} \frac{\sum_{j=1}^{t} \delta_j}{t} = p_i - p_{i+1} \quad a.s.$$

□

Using the same method as in the previous lemma, one can use the central limit theorem to get a more precise result on the behavior of $g_i(t)$ when t goes to infinity (in distribution).

$$\lim_{t\to\infty} \frac{|g_i(t) - t(p_i - p_{i+1})|}{\sigma^2\sqrt{t}} = \mathcal{N}(0,1)$$

Where $\mathcal{N}(0,1)$ is the normal distribution with mean 0 and variance 1 and $\sigma^2 = p_i + p_{i+1} - p_i^2 - p_{i+1}^2 + 2p_ip_{i+1}$.

Let $N_\ell(t)$ be the number of tokens whose positions are larger than ℓ after t steps.

$$N_\ell(t) = \max\{i, \sum_{j=i}^{n} g_j(t) \geq \ell\}.$$

Using the central limit theorem,

$$\sum_{j=i}^{n} g_j(t) = p_i t + \sqrt{t} S_i(t), \tag{4}$$

where $S_i(t)$ is a centered random variable such that $S_i(t)$ converges in distribution to a normal law with zero mean and bounded variance, since the $g_j(t)$ are eventually independent variables.

If one denotes $k := N_\ell(t)$, then by definition of k,

$$p_k t + \sqrt{t} S_k(t) \geq \ell > p_{k+1} t + \sqrt{t} S_{k+1}(t).$$

Dividing by t, one gets

$$p_k + \frac{S_k(t)}{\sqrt{t}} \geq \frac{\ell}{t} > p_{k+1} + \frac{S_{k+1}(t)}{\sqrt{t}}.$$

Now, $\frac{S_k(t)}{\sqrt{t}}$ converges to 0 in probability when t goes to infinity, uniformly in ℓ.

Finally we get the following result.

Theorem 3. *If $\ell = p_i t + O(t^{1/2})$ for some $i \in \{0, \ldots, N\}$, then there exists a discrete distribution $\mathcal{J}$ whose support is $\{i-1, i\}$ such that*

$$N_\ell(t) \xrightarrow[t\to\infty]{} \mathcal{J} \text{ in distribution.}$$

For all the other cases,

$$N_\ell(t) \xrightarrow[t\to\infty]{} I(\ell/t) \text{ in probability,} \tag{5}$$

where $I(x) := \max\{i \mid p_i \geq x\}$.

Remark 4. *Let p_i be a Zipf's law with parameter α. In the above discussion, Equation (4) shows that when t is large, the sizes of the cities follow a Zipf's law (with parameter α). As for Equation (5), it gives the related cumulative distribution function, which is Pareto (with parameter $1/\alpha$), i.e. at time t, the probability for city sizes to be larger than ℓ is approximatively $A\ell^{-1/\alpha}$ where $A = (Ct)^{1/\alpha}/n$ since $p_i \sim Ci^{-\alpha}$ and $I(\ell/t) = \max\{i \mid p_i \geq \ell/t\} \sim (\frac{\ell}{Ct})^{-1/\alpha}$. The same holds in the first case.*

Remark 5. *It should also be clear from all the reasoning that initializing the process with city sizes different from zero does not change the asymptotic behavior.*

Remark 6. *We have used the fact that the sequence $(p_i)_{1\leq i\leq n}$ is decreasing to apply Forster's theorem which ensures that from a certain time the gap between city sizes remains strictly positive and then the city of rank i grows almost independently from others at speed p_i.*

If the sequence $(p_i)_{1\leq i\leq n}$ is not decreasing, then the asymptotic behavior of the face-homogeneous Markov chain is more difficult to analyze. In general the cities group into clusters, each with the same growing speed. However, given the p_i's, it is an open question to know, for example, if all sites form a single cluster (which is the same as assessing the ergodicity of the corresponding face-homogeneous Markov Chain, see [4] for discussions about this).

3 Application to Gulyas' Model

The next model was proposed by Laszlo Gulyas as an alternative simple model for city growth. Given a set of n cities with initial populations equal to 0, new inhabitants arrive sequentially. When a new inhabitant arrives, a random sample of cities is drawn by first picking at random an integer k between 1 and n (with uniform distribution) and then a random subset of k cities from the n ones (with uniform distribution). The person joins the city with the largest size in the selected subset (if there are several ones, choose one of them arbitrarily) and it increments its size by 1.

Running some simulations, Laszlo Gulyas observed that the distribution of city sizes follows a Zipf's law. He conjectured that its parameter was 2.

We will first show that Gulyas' model is a special case of our general model presented in Section 2, and then apply our analysis to validate Gulyas' conjecture.

3.1 The probability to choose one city

With this model, the probability p_i that the i'th largest city is chosen and has its size incremented by 1 is equal to:

$$\frac{1}{n}\sum_{k=0}^{n-i}\frac{\binom{n-i}{k}}{\binom{n}{k+1}},$$

since this city can be chosen only if it belongs to a subset of cardinality $k+1$ with $1 \leq k+1 \leq n-i+1$ and once $k+1$ is fixed (with probability $1/n$), among the $\binom{n}{k+1}$ subsets of cardinality $k+1$ only the ones containing the i'th largest city and k cities taken from the $n-i$ smaller ones will lead to its choice (thus it occurs with probability $\binom{n-i}{k}/\binom{n}{k+1}$).

This sum has a closed and very simple expression that we may compute thanks to formal series. We will use the equality:

$$\sum_{k\geq 0}\binom{k+m}{k}x^k = \frac{1}{(1-x)^{m+1}}$$

We have:

$$\begin{aligned}
\frac{1}{n}\sum_{k=0}^{n-i}\frac{\binom{n-i}{k}}{\binom{n}{k+1}} &= \frac{1}{n}\sum_{k=0}^{n-i}\frac{(n-i)!}{n!}(k+1)\frac{(n-k-1)!}{(n-i-k)!}\\
&= \frac{1}{n}\frac{(n-i)!}{n!}\sum_{k=0}^{n-i}(k+1)\frac{(n-k-1)!}{(n-i-k)!}\\
&= \frac{(i-1)!}{n}\frac{(n-i)!}{n!}\sum_{k=0}^{n-i}(k+1)\frac{(n-k-1)!}{(n-i-k)!(i-1)!}\\
&= \frac{(i-1)!}{n}\frac{(n-i)!}{n!}\underbrace{\sum_{k=0}^{n-i}(k+1)\binom{n-k-1}{i-1}}_{(*)}
\end{aligned}$$

By denoting $[x^n]S(x)$ the coefficient of x^n in the series $S(x)$, the under-braced sum $(*)$ is equal to:

$$\begin{aligned}
\sum_{k=1}^{n-i+1}k\binom{n-k}{i-1} &= [x^n]\Big(\sum_{k\geq 1}kx^k\Big)\Big(\sum_{j\geq 0}\binom{j}{i-1}x^j\Big)\\
&= [x^n]\Big(\sum_{k\geq 1}kx^k\Big)\Big(\sum_{j\geq 0}\binom{j+i-1}{i-1}x^{j+i-1}\Big)\\
&= [x^n]x^i\underbrace{\Big(\sum_{k\geq 1}kx^{k-1}\Big)}_{\frac{1}{(1-x)^2}}\underbrace{\Big(\sum_{j\geq 0}\binom{j+i-1}{j}x^j\Big)}_{\frac{1}{(1-x)^i}}\\
&= [x^n]\frac{x^i}{(1-x)^{i+2}}\\
&= [x^{n-i}]\frac{1}{(1-x)^{i+2}}\\
&= \binom{n+1}{i+1}
\end{aligned}$$

By reintroducing this value in the initial sum, we obtain for the probability to increment the i^{th} greatest integer:

$$p_i = \frac{(i-1)!}{n}\frac{(n-i)!}{n!}\binom{n+1}{i+1} = \frac{n+1}{n}\frac{1}{i(i+1)}.$$

The sequence of p_i's is strictly positive and decreasing, thus we can apply all the results of Section 2. Since the p_i's follow a Zipf's law with parameter 2 and $I(x) = \max\{i \mid 1/p_i \leq x\} \sim ((n+1)x/n)^{1/2}$, with Remark 4, we prove Gulyas' conjecture:

Theorem 7. *In Gulyas' growth model, the city sizes asymptotically distribute as a Zipf's law with parameter 2. More precisely when t tends to infinity, the size of the i'th largest city $v_i(t)$ verifies $v_i(t) \sim \frac{n+1}{n}\frac{t}{i(i+1)}$ and the number $N_\ell(t)$ of cities of size larger or equal to ℓ verifies $N_\ell(t) \sim \frac{n+1}{nt}^{1/2}\ell^{-1/2}$.*

3.2 Simulation

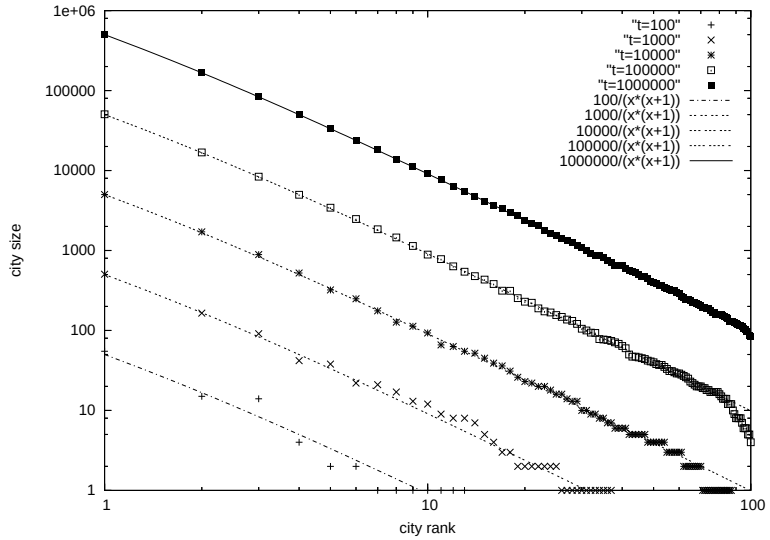


Figure 5: The sizes of cities depending on their rank at various times during the process, with log-scaled axes.

The Zipf's law in the previous model was observed by Gulyas thanks to simulation. For a bench of experimental results, the reader is referred to his work on such models [9, 14].

We have also run some simulations of the system, mainly to estimate the speed of convergence towards the asymptotic regime given above. Section 2 gives some clues about it, but there are several hidden functions or constants which asymptotically disappear but play a role in the convergence speed. For now, a precise quantification of this speed remains a difficult problem.

Figure 5 presents an example of the evolution of the system when $n = 100$ for a sequence of random draws. At time $t = 100, 1000, 10000, 100000$ and 1000000, we have plotted the distribution of the 100 city sizes, each city having as coordinates its rank and its size. We have also plotted the ideal curves where the points should be located under the asymptotic regime, namely $p_i t \approx t/(i * (i + 1))$ for the point of rank i. We have used log-scaled axes to put the data all together and to point out the power law behavior.

It can be clearly seen that the output data match the mathematical analysis: it converges towards the ideal curves which correspond to Zipf's laws with parameter 2.

The simulation uses aliasing techniques to select which city will be joined by the new

person in constant time [16].

4 Variations on this agent-based model

A natural way to generalize Gulyas' model is to keep the scheme that a newcomer chooses the largest city proposed to him but to change the probability distribution used to pick the random sample of cities.

Suppose that we do not wish that the city names play any role in the random sampling, i.e. we do not wish to impose before starting some correlated growth of cities. Then given the n initial cities $c_1, \ldots, c_n$, for all $1 \leq k \leq n$, all the subsets of $\{c_1, \ldots, c_n\}$ of cardinality k should be picked with the same probability. Then the system is fully defined by a set of probabilities $(\alpha_k)_{1\leq k\leq n}$, $\sum_{1\leq k\leq n} \alpha_k = 1$, where α_k is the probability to choose the cardinality k and thus the probability to pick a fixed subset of cardinality k is $\alpha_k/\binom{n}{k}$. Gulyas' original model is when you take $\alpha_k = 1/n$ for all $1 \leq k \leq n$.

Given the sequence $(\alpha_k)_{1\leq k\leq n}$, this model is clearly equivalent to the general model of Section 2 where the probability p_i to join the i'th largest city is equal to:

$$\sum_{k=0}^{n-i} \alpha_{k+1} \frac{\binom{n-i}{k}}{\binom{n}{k+1}}.$$

Moreover the sequence $(p_i)_{1\leq i\leq n}$ is decreasing if and only if there exists $2 \leq k_0 \leq n-1$ such that $\alpha_{k_0} > 0$. In that case, all the analysis of Section 2 applies. We have seen that asymptotically the size of the i'th largest city is approximatively $p_i t$, thus the emergence of a Zipf's law is directly linked to the form of the p_i's and not to intricate combinatorics due to cities overtaking each other.

Note that formal series may be also used here to compute the values of p_i. All the reasoning of Section 3.1 applies and following the same steps, it can be checked that :

$$p_i = \frac{(i-1)!(n-i)!}{n!}[x^{n-i}]\frac{S'(x)}{(1-x)^i}, \tag{6}$$

where $[x^{n-i}]$ is the coefficient of x^{n-i} in the series and $S(x) = \sum_{k=1}^{n} \alpha_k x^k$, $S'(x)$ being its derivative.

4.1 Non uniform choices for city sets

In the following, we analyse several cases. We have seen above that when the probability to choose a set of cities of size k is uniform ($\alpha_k = 1/n$) then, $p_i = C_n/i^2$, where C_n is a normalizing constant. This implies that the long-term city sizes follow Zipf's laws with parameter 2. In the following we consider other values for α_k. We focus on two cases. First let $\alpha_k = a_n/k$ (where a_n is a normalizing constant). This means that large sets have a smaller chance to be chosen which could actually be the case in real life where choices are often limited.

The other case is $\alpha_k = b_n\binom{n}{k}$ (again b_n is a normalizing constant). This corresponds to a uniform distribution over all subsets of cities, without taking the number of cities into account.

Case ($\alpha_k = a_n/k$) If $\alpha_k = a_n/k$, then by definition, $S(x) = a_n \sum_{k=1}^{n} x^k/k$ and

$$S'(x) = a_n \sum_{k=0}^{n-1} x^{k-1} = a_n \frac{1-x^n}{1-x}.$$

Therefore,

$$\frac{S'(x)}{(1-x)^i} = a_n \frac{1-x^n}{(1-x)^{i+1}}.$$

Now, looking at the coefficients,

$$\begin{aligned}
[x^{n-i}]\frac{S'(x)}{(1-x)^i} &= a_n[x^{n-i}]\frac{1-x^n}{(1-x)^{i+1}} \\
&= a_n[x^{n-i}]\left((1-x^n)\sum_{k=1}^{n}\binom{k+i}{i}x^k\right) \\
&= a_n\binom{n}{i}.
\end{aligned}$$

Using (6),

$$p_i = \frac{1}{\binom{n}{i}}\frac{1}{i}\binom{n}{i} = \frac{a_n}{i}. \tag{7}$$

Therefore, p_i follows a Zipf's law with parameter 1. Now, using Theorem 3, the city sizes also follow a Zipf's law with parameter 1.

Case ($\alpha_k = b_n\binom{n}{k}$) If $\alpha_k = b_n/k$, then by definition, $S(x) = b_n \sum_{k=1}^{n} \binom{n}{k} x^k$ and

$$S'(x) = b_n n(1+x)^{n-1}.$$

Now, looking at the coefficient of x^{n-i},

$$p_i = \frac{1}{\binom{n}{i}}\frac{1}{i}[x^{n-i}]\frac{S'(x)}{(1-x)^i} = \frac{b_n 2^n}{2^i}.$$

In this case, p_i does not follow a Zipf's law. It is exponentially distributed as well as the sizes of the cities.

These two cases can be seen as extreme. In general, playing with the probability α_k changes the laws of the cities sizes drastically. Actually, it enables one to model many different behaviors of the city sizes, from exponential to power laws with almost arbitrary parameters.

5 Conclusion

The main feature of this paper is to provide an example of face homogeneous Markov chain that may appear in the modelization of real discrete systems. The mathematical analysis matches simulation results and offers a powerful prediction tool. As shown in the preceding section, the general model we have presented includes some agent-based and multi-site growth models where each site has a value (which may be seen as its power of attraction) and where new agents preferentially choose the more attractive sites among the ones proposed to them

(and then increase the attraction of the chosen sites).

Several variations of this type of dynamics may appear in models of real systems.

For instance, if the sequence $(p_i)_{1\leq i\leq n}$ is not decreasing, the behavior of the asymptotic distribution of sizes is more difficult to describe and the mathematical analysis is much more difficult. Even worse, it might be undecidable to predict some features of the dynamics [8].

Of course, many other features could be added to the general model described in this paper. One may cite: some synchronicity of the arrivals, some departures from the sites also depending on the site ranks, non-uniform distribution when drawing subsets of the same cardinal e.g. to take into account the geography of the sites and the fact that newcomers will choose their destination among close sites which leads to a correlated evolution of their sizes ... Some of these variations may drastically change the long run behavior and thus the study of each of them constitutes a new interesting challenge.

References

[1] L. A. Adamic. The small world web. In *ECDL'99, Lecture Notes in Computer Science 1696*, pages 443–452. Springer, 1999. Berlin.

[2] R. Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Review of Modern Physics*, 74:47–94, 2002.

[3] R. L. Axtell. Zipf distribution of u.s. firm sizes. *Science*, 293:1818–1820, 2001.

[4] G. Fayolle, V. A. Malyshev, and M.V. Menchikov. *Constructive theory of countable Markov chains*. Cambridge University Press, 1995.

[5] X. Gabaix. Zipf's law and the growth of cities. *American Economic Review Papers and Proceedings*, 89(2):129–132, May 1999.

[6] X. Gabaix. Zipf's law for cities: An explanation. *Quaterly Journal of Economics*, 114(3):739–767, August 1999.

[7] X. Gabaix and Y. Ioannides. *Handbook of Urban and Regional Economics*, volume 4, chapter 53 - The Evolution of City Size Distribution, pages 2341–2378. V. Henderson and J.-F. Thisse, 2004.

[8] D. Gamarnik. On deciding stability of constrained homogeneous random walks and queueing systems. *Math. Oper. Res.*, 27(2):272–293, 2002.

[9] L. Gulyas and Y. Mansury. Patterns of firm agglomeration: an autonomous agent based model of city formation. In *International Conference on Complex Systems, ICCS'2002*, Nashua, 2002.

[10] B. Huberman and L. Adamic. Growth dynamics of the world wide web. *Nature*, 401:131, 1999.

[11] Exystence Thematic Institute. "discrete and computational aspects of complex systems". Lyon, France, June-July 2003. Exystence web page: `http://www.complexityscience.org`.

[12] J. Lamperti. Criteria for the recurrence or transience of stochastic process, I. *Journal of Mathematical Analysis and Applications*, 1:314–330, 1960.

[13] P. Faloutsos M. Faloutsos and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM'99*, pages 251–262, 1999.

[14] Y. Mansury and L. Gulyas. Explaining the zipf-law pattern of city formation: An agent based approach. In *Association of Collegiate Schools of Planning (ACSP) 2002 Conference*, November 2002. Washington D.C.

[15] Kwok Tong Soo. Zipf's law for cities: A cross country investigation,. Technical report, mimeo, London School of Economics, 2002.

[16] A.J. Walker. An efficient method for generating random variables with general distributions. *ACM Trans. Math. Software*, pages 253–256, 1974.

[17] D. Williams. *Probability and Martingales*. Cambridge University Press, 1991.

[18] G. K. Zipf. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, Cambridge MA, 1932.

[19] G. K. Zipf. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge MA, 1949.

A Comparison with Simon's model

In this section, we discuss the relations between our model of city growth and the well known Simon's model.

Let us first give a brief description of Simon's model, which is also agent-based.

Consider a system which current state at time t is made of n cities of respective sizes $s_1(t), \cdots, s_n(t)$. Each newcomer decides either to create a new city with a given probability α (this new city is then of size $s_{n+1}(t+1) = 1$), or to settle down in one of the existing cities with probability $1 - \alpha$. In the latter case, the new comer chooses its home city according to a Bernouilli process with probability $s_i(t)/S(t)$, where $S(t) = \sum_k s_k(t)$.

Simon has shown using a continuous approximation that this model follows a Zipf's law. More precisely that $\mathbb{E}s_{\sigma(i)}(t)/s_{\sigma(i+1)}(t)$ has a limit $\ell(\alpha)$, when t goes to infinity.

This model differs in many ways from ours. The first main difference is the fact that new cities are created during the evolution of the system. However, even when $\alpha = 0$ (in which case new cities are never created w.p. 1), both systems behave in a very different way.

In the following we will analyse the discrete long-run behavior of Simon's model when $\alpha = 0$, , which seems to have gone undetected so far.

Theorem 8. *Consider a system following Simon's dynamic with $\alpha = 0$ and n cities, all of initial size* 1*. Then,*

$$\lim_{t\to\infty} \frac{s_i(t)}{S(t)} = \ell_i \quad a.s.$$

where ℓ_i is a real random variable, uniformly distributed over $(0, 1)$*.*

Before proving this theorem, several comments can be made.

- The ratio between the sizes of any two cities converges. In particular, if $\sigma_i(t)$ is the i-th largest city at time t, , the ratio with the next largest city exists and can be written

$$\lim_{t\to\infty} \frac{s_{\sigma_i(t)}(t)}{s_{\sigma_{i+1}(t)}(t)} = \frac{\ell_{\sigma_i(\infty)}}{\ell_{\sigma_{i+1}(\infty)}}.$$

  It ranges from 0 to ∞ when $\ell_{\sigma(i)}$ and $\ell_{\sigma(i+1)}$ vary from 0 to 1. Its distribution is rather complex and depends on order statistics (same with the laws of $\ell_{\sigma_i(\infty)}$ and $\ell_{\sigma_{i+1}(\infty)}$)). Its derivation is not reported here.

- The ratios may depend on which cities are considered It can be shown that the ratios between the sizes of consecuitve cities are smaller for large cities than for small cities (again, this is not reported here). Thefore, the system does not follow a Zip'f law with a constant parameter.

- Furthermore, this system is not ergodic: the limits are random variables, so that one realisation of this system has a very different limit behavior than a second one. Moreover, since ℓ_i is uniformly distributed, the range of possible behaviors is maximal and equi-probable.

- the non-ergodicity of the system makes it very tricky to control or predict. A simulation will not provide any information on its long term behavior, and the ratio between the cities (Zipf's paramters) can only be computed afterwards. For example, in the simple case of two cities, one very long simulation may provide an asymptotic ratio of $\ell_{\sigma_1(\infty)} = 1$ (meaning that almost everyone settles in the largest city), while another very long simulation, starting with the same state may provide a ratio $\ell_{\sigma_1(\infty)} = 1/2$ (meaning that only half the population chooses the largest city). Both simulation are valid and equi-probable (as well as any simulation providing any other ratio between 1 and 1/2).

Proof. Let us first prove that the limit exists. This is based on Martingale theory.

Let us consider the random variable $X_i(t) = \frac{s_i(t)}{S(t)}$. Computing its conditional expectation gives

$$\begin{aligned}
\mathbb{E}(X_i(t+1)|s_1(t),\dots,s_n(t)) &= \left(\frac{s_i(t)}{S(t)}\right)\frac{s_i(t)+1}{S(t)+1} + \left(1-\frac{s_i(t)}{S(t)}\right)\frac{s_i(t)}{S(t)+1} \\
&= \frac{s_i(t)}{S(t)} \\
&= X_i(t).
\end{aligned}$$

Therefore, $X_i(t)$ is a martingale with respect to $s_1(t),\dots,s_n(t)$, bounded by 1. Using the Martingale Convergence Theorem [17], this means that there exists a random variable ℓ_i such that

$$\lim_{t\to\infty} X_i(t) = \ell_i \quad a.s.$$

Next, we show that ℓ_i is uniformly distributed in $[0,1]$. For that, consider the distribution of $s_1(t),\dots,s_n(t)$. We will show by induction that for all $x_1,\dots,x_n \geq 1$ such that $x_1 + \cdots + x_n = n + t$, $\mathbb{P}(s_1(t) = x_1,\dots,s_n(t) = x_n) = 1/K(t)$, where $K(t) = \binom{t+n-1}{t}$. At $t = 0$, $x_1 = \cdots = x_n = 1$ is the only case , $K(0) = 1$ and $\mathbb{P}(s_1(t) = x_1,\dots,s_n(t) = x_n) = 1$.

Assume that the equation is true at step t and consider the distribution at step $t+1$. If $x_i \geq 2$ for all i, then

$$\begin{aligned}
\mathbb{P}(s_1(t+1)=x_1,\ldots,s_n(t+1)=x_n) &= \sum_i \mathbb{P}(s_1(t)=x_1,\ldots,s_i(t)=x_i-1,\ldots,s_n(t)=x_n)\frac{s_i(t)}{S(t)} \\
&= \frac{1}{\binom{t+n-1}{t}}\sum_i \frac{x_i-1}{t+n} \\
&= \frac{1}{\binom{t+n-1}{t}}\frac{t+1}{t+n} \\
&= \frac{1}{\binom{t+n}{t+1}}.
\end{aligned}$$

On the other hand, if $x_i = 1$ for some i, then without any loss of generality, one may assume that $x_1 = 1, \ldots, x_k = 1, x_{k+1} \geq 2, \ldots, x_n \geq 2$.

$$\begin{aligned}
\mathbb{P}(s_1(t+1)=x_1,\ldots,s_n(t+1)=x_n) &= \sum_{i=k+1}^{n} \mathbb{P}(s_1(t)=x_1,\ldots,s_i(t)=x_i-1,\ldots,s_n(t)=x_n)\frac{s_i(t)}{S(t)} \\
&= \frac{1}{\binom{t+n-1}{t}}\sum_{i=k+1}^{n} \frac{x_i-1}{t+n} \\
&= \frac{1}{\binom{t+n-1}{t}}\frac{t+1}{t+n} \\
&= \frac{1}{\binom{t+n}{t+1}}.
\end{aligned}$$

To end the proof, if $\mathbb{P}(s_1(t)=x_1,\ldots,s_n(t)=x_n) = 1/K(t)$ for all t, this means that the distribution of the sizes of the cities is uniform. Taking the limit when t goes to infinity, the distribution remains uniform so that for each city, $s_i(t)/S(t)$ converges in distribution to the uniform distribution over $[0,1]$. Combining this with the almost sure limit ℓ_i shown above implies that ℓ_i is indeed random and uniform over $[0,1]$. □

Actually, a more general theorem can be shown.

Theorem 9. *Consider a system following Simon's dynamic with $\alpha = 0$ and n cities, with initial sizes $a_1, \ldots, a_n$. Then,*

$$\lim_{t\to\infty} \frac{s_i(t)}{S(t)} = \ell_i \quad a.s.$$

where ℓ_i is a real random variable, with a Dirichlet distribution.

The proof uses the fact that the probablity to go from state $a_1, \ldots, a_n$ to state $b_1, \ldots, b_n$ has a nice combinatorial structure.

Finally, unlike Simon's model, our model has an ergodic behavior. The ratios between the city sizes converge to a deterministic value which can be evaluated on all simulations and which can also be computed in many cases using the parameters of the model $(n, \alpha_1, \ldots, \alpha_n)$.