Unravelling the size distribution of social groups with information theory in complex networks

A. Hernando\textsuperscript{1,a}, D. Villuendas\textsuperscript{2}, C. Vesperinas\textsuperscript{3}, M. Abad\textsuperscript{1}, and A. Plastino\textsuperscript{4,b}

\textsuperscript{1} Departament ECM, Facultat de Física, Universitat de Barcelona. Diagonal 647, 08028 Barcelona, Spain
\textsuperscript{2} Departament FEN, Facultat de Física, Universitat de Barcelona, Diagonal 647, 08028 Barcelona, Spain
\textsuperscript{3} Sogeti España, WTCAP 2, Plaça de la Pau s/n, 08940 Cornellà, Spain
\textsuperscript{4} National University La Plata, IFCP-CCT-CONICET, C.C. 727, 1900 La Plata, Argentina

Received 12 September 2009 / Received in final form 17 February 2010
Published online 2 July 2010 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2010

Abstract. The minimization of Fisher’s information (MFI) approach of Frieden et al. [Phys. Rev. E 60, 48 (1999)] is applied to the study of size distributions in social groups on the basis of a recently established analogy between scale invariant systems and classical gases [Phys. A 389, 490 (2010)]. Going beyond the ideal gas scenario is seen to be tantamount to simulating the interactions taking place, for a competitive cluster growth process, in a scale-free ideal network – a non-correlated network with a connection-degree’s distribution that mimics the scale-free ideal gas density distribution. We use a scaling rule that allows one to classify the final cluster-size distributions using only one parameter that we call the competitiveness, which can be seen as a measure of the strength of the interactions. We find that both empirical city-size distributions and electoral results can be thus reproduced and classified according to this competitiveness-parameter, that also allow us to infer the maximum number of stable social relationships that one person can maintain, known as the Dunbar number, together with its standard deviation. We discuss the importance of this number in connection with the empirical phenomenon known as “six-degrees of separation”. Finally, we show that scaled city-size distributions of large countries follow, in general, the same universal distribution.

1 Introduction

Regularities reflected in either scaling properties \cite{1} or power laws \cite{2-6} appear in different scenarios related to social groups. One of the most intriguing is Zipf’s law \cite{7}, a power law with exponent $-2$ for the density distribution function that is observed in describing urban agglomerations \cite{8,9} and firm sizes all over the world \cite{10}. This fact has received a remarkable degree of attention in the literature. The above mentioned regularities have been detected in other contexts as well, ranging from percolation theory and nuclear multi-fragmentation \cite{11-13} to the abundances of genes in various organisms and tissues \cite{14}, the frequency of words in natural languages \cite{7,15}, the scientific collaboration networks \cite{16}, the total number of cites of physics journals \cite{17}, the Internet traffic \cite{18,19} or the Linux packages links \cite{20}. More recently, Filho et al. \cite{21} found another special regularity in the density distribution function of the number of votes in the Brazilian elections, a power law with exponent $-1$. This law has also been found in reference \cite{22}, using an information-theoretic methodology \cite{23}, for both the city-size distribution of the province of Huelva (Spain) and the results of the 2008 Spanish General Elections. These findings allow one to conjecture that this behavior reflects a second class of universality.

What all these disparate systems have in common is the lack of a characteristic size, length or frequency for the observable under scrutiny, which makes them scale-invariant. In reference \cite{22} we have introduced an information-theoretic technique based upon the minimization of Fisher’s information measure \cite{23} (abbreviated as MFI) that allows for the formulation of a “thermodynamics” for scale-invariant systems. The methodology establishes an analogy between such systems and physical gases which, in turn, shows that the two special power laws mentioned in the preceding paragraph lead to a set of relationships formally identical to those pertaining to the equilibrium states of a scale invariant non-interacting system, the scale-free ideal gas (SFIG). The difference between the two distributions is thereby attributed to different boundary conditions on the SFIG.

However, there are many social systems that can not be included into any of these two universality classes and exhibit different kinds of behavior \cite{22}. In order to deal with them, during the last years researchers have worked out different mathematical models and thus addressed urban
The goals and motivation of this work are thus focused on gaining insight into such size-distributions in the case of systems that can not be described by recourse to the two power laws described above, i.e., by a non-interacting scenario. If an analogy with real gases is worked out when interactions are duly taken into account, a microscopic description is needed in order to obtain the pair correlation function [28]. This is achieved using numerical simulations as in molecular dynamics. A similar path will be followed here by recourse to the Fisher-derivation analogy of [22], where we showed that the dynamical process behind the scale-free behavior is that of a proportional growth process (PGP), which was validated numerically by using geometrical Brownian walkers. As shown in reference [29], several models are able to reproduce such a process, entailing that different choices are available in order to introduce the effect of interactions between the elements of the system.

City-size distributions and electoral results display a similar scale-free behavior, and both of them have the same constituents: groups of people. Although the resources of these groups or the interests of the individuals composing them may be different in each case, a naive approach is to assume that people are connected to other people, hence giving rise to a network where groups of interest develop. Diffusion processes as the “competitive cluster growth process” (CCG) – which is able to describe a PGP – have been successfully used before for dealing with electoral results and the spread of opinions [21,26] within the framework of Network theory [18,19]. This encourages us to employ CCGs to develop the microscopic description of the associated systems and thus go beyond the SFIG stage.

This work is organized as follows: in Section 2 we describe the application of the MFI approach [23] to obtain from first principles the degree distribution of the complex network where the CCG process takes place. This allows one to, in turn, microscopically simulate growth processes in a social group. In Section 3 we study the size distributions obtained using this methodology. We find a scale transformation that allows for systematically classifying the deviations from the SFIG that we encounter in the cluster-size distributions. This classification is effected by a single parameter, which we call the competitiveness. We also apply this criterion to classify the city-size distributions of the provinces of Spain and some electoral results. Moreover, we show that empirical assessments as the average path length and Dunbar’s number are well reproduced by our approach. Using such a scale transformation we demonstrate that most distributions of city population in large countries exhibit the same shape. Finally, in Section 4 we draw some conclusions.

2 Theoretical method

The basic elements of networks are “nodes” that are connected to other nodes by “edges”. The degree $k$ of each node is defined as the number of connections it possesses. The degree distribution (DD) $F(k)$ and the way the nodes are connected define the statistical properties of the network. Scale-free complex networks (networks with a power-law DD) display many interesting properties that have been found in techno-sociological systems such as the Internet (World Wide Web [30], e-mail networks [31] and also instant-message-sending networks [32], for example).

The competitive cluster growth process within a complex network falls into the category of a PGP, which are known to correctly describe scale-invariant behavior [29]. It is expected that, under certain conditions, the competition will make the final size of a cluster to be correlated with the sizes of the other clusters. Such a process illustrates the kind of interaction that we aim to describe. However, it has been shown that the final distribution of cluster-sizes strongly depends on the structure of the network [33–36], a fact that introduces some degree of arbitrariness into the model.

In order to reduce this arbitrariness in the network-choice and to also reduce the complexity of the problem at hand to a minimum we adopt two basic assumptions:

- that the network can be described at the macroscopic level as a scale-invariant system – which is reasonable due to the availability of abundant empirical evidence – and
- for simplicity’s sake, that neither the degree of each node depends on the degree of other nodes nor that other kind of correlations are present.

2.1 The scale free ideal network

How can we build a network based upon just those two assumptions? The configuration model [37,38] is a classical one of network-randomization, used to eliminate correlations that are not strictly dependent on the degree-distribution. We believe this to be the ideal tool for building up our network, although we have still to generate a way to predict DD-forms. Taking into account our assumptions, (i) the network can be described as a scale invariant system of $N$ nodes, with the number of connections $k$ being the “coordinate” that locates each node in the pertinent phase space, and (ii) the degree of each node does not depend on the degree of other nodes. In such circumstances we can legitimately describe the network as a SFIG in equilibrium [22], a scenario to be denoted as the scale free ideal network (SFN), in which one can derive a DD from first principles via minimization of Fisher’s information measure [17], a process that we pass now to recapitulate.
2.1.1 Minimum Fisher Information approach (MFI)

The Fisher information measure $I$ for a system of $N$ elements, described by the coordinate $k$ and the physical parameters $\theta$ has the form [39–41]

$$I(F) = c_k \int dk F(k|\theta) \left| \frac{\partial \ln F(k|\theta)}{\partial \theta} \right|^2,$$

(1)

where $F(k|\theta)$ is the density distribution in phase space and the constant $c_k$ accounts for proper dimensionality. According to MFI tenets [23], the equilibrium state of the system minimizes $I$ subject to prior conditions, such as the normalization of $F$ to a certain value $N$, namely $\langle 1 \rangle = N$. The MFI is then cast as a variation problem of the form [23]

$$\delta \left\{ I(F) - \mu(1) \right\} = 0,$$

(2)

where $\mu$ is the normalization-associated Lagrange multiplier.

2.1.2 Application of the MFI to a scale-free network

In the case of the derivation of the DD of our complex network, we consider $N$ nodes with a minimum connection-degree $= 1$ and a maximum connection-degree $k_M$. With the change of variable $u = \ln k$, the scale transformation $k' = k/\Theta_k$ transforms $u$ into $u' = u - \Theta_k$, where $\Theta_k = \ln \theta_k$. The distribution of physical elements is then described by the mono-parametric translation families $F^\prime(k|\theta) = f(u|\Theta_k) = f(u')$. Taking into account the fact that the Jacobian of the transformation is $dk = e^u du$, the information measure $I$ can be obtained in the continuous limit as [17]

$$I = c_u \int_0^{\ln k_M} du e^u f(u) \left| \frac{\partial \ln f(u)}{\partial u} \right|^2,$$

(3)

and, with the normalization constraint

$$\int_0^{\ln k_M} du e^u f(u) = N,$$

(4)

the variation problem reads now [17]

$$\delta \left\{ c_u \int_0^{\ln k_M} du e^u f \left| \frac{\partial \ln f}{\partial u} \right|^2 + \mu \int_0^{\ln k_M} du e^u f \right\} = 0.$$  

(5)

Introducing $f(u) = e^{-u} \Psi^2(u)$, and varying with respect to $\Psi$ leads to the Schrödinger-like equation [23]

$$-4 \frac{\partial^2 \Psi}{\partial u^2} + 1 + \mu' \Psi(u) = 0,$$

(6)

where $\mu' = \mu/c_u$. The general solution to this equation is $\Psi(u) = e^{-\alpha u/2}$ with $\alpha = \sqrt{1 + \mu'}$. Equilibrium corresponds to the ground state solution $\alpha = 0$ [23], which yields for the connection-degree $k$ the same density distribution as that of the SFIG in the thermodynamic limit (since $\ln k_M$ represents the volume in phase space [22], this condition is reached when $N/\ln k_M$ remains finite in the limit $N, \ln k_M \to \infty$), that is,

$$F(k)dk = \frac{N}{\ln k_M} dk, \text{ with } k < k_M.$$

(7)

2.1.3 Building the SFIN via the configuration model

Once we know, for a given total number of nodes $N$, the associated distribution for the connection-degree of the SFIN, we proceed by assigning to the $i$-th node a number of potential edges $k_i$ randomly obtained from $F(k)$. Accordingly, the $N$ nodes become now randomly connected among themselves by their assigned edges, with two restrictions: a node cannot be connected to itself nor twice to the same node. The ensuing process ends up when no more connections can be established. In this way, this configuration model predicts that all properties of the network are depend exclusively upon the DD, in this case via the value of the maximum degree $k_M$ and the number of nodes $N$. However we will show below that the values of $N$ and $k_M$ can be arbitrarily chosen in order to classify empirical distributions.

2.2 The competitive cluster growth process

Once we have built a SFIN with $N$ nodes and maximum connection-degree $k_M$, we apply a CCG process to it, as discussed in [33–36]. For starters, we fix the value of the total number of clusters $n_c$ that will grow in the network. Next, $n_c$ nodes of the network are randomly selected as cluster “seeds”. In the first iteration the first neighbors of the seeds are incorporated to the cluster in random order, unless they are seeds of other clusters. At subsequent iterations, the first neighbors of all nodes added to a cluster at the precedent step are, in turn, added to this cluster unless they already belong to a different cluster. The order of the clusters is chosen randomly at each iteration, adding all the new nodes at the same time. The process ends when all the nodes belong to some cluster. We display in Figure 1 the final result of a competitive growth process for $n_c = 100$ clusters in a network of 5000 nodes.

The available empirical data regarding city populations usually account for the population of the lowest-order administrative unit (municipalities or communes). In some cases several population nuclei are included into the same unit. Something similar happens in the case of electoral elections whenever independent groups form a coalition and join forces to increase their chances. Since these alliances can affect to the form of the final size-distribution, we have introduced in the model a new parameter $n_t$ which is the number of initial seeds that will actually considered as members of the same cluster. In this way, a total number of $n_t n_c$ seeds are now randomly selected at the beginning of the process, so that each of...
3 Present results

3.1 Study and classification of cluster-size distributions

We have studied the size distribution of SFIN-clusters with $N$ and $k_M$ ranging from $N=5000$ to $500000$ and $k_M=50$ to $500$. As observed in other opinion-dynamics’ models [43], finite-size effects are relevant, in our case for small values of $N$ and $k_M$. Since we are interested on macroscopic phenomena we focus our research on situations in which the values of $k_M$ and $N$ are large, and also for $n_c$ and $n_t$. Even if the maximum value of $N$ used here (500000) is not as large as in other studies, we have found that this value suffices for reducing finite-size effects and reproduce empirical size-distributions, allowing us to perform our calculations in an ordinary personal computer (PC).

3.1.1 Regime of low cluster-density: recovering the SFIG

When the density of seeds, defined as $\rho_s = n_s/n_N$, is much lower than unity, the probability density distribution of sizes $p(x)$ looks similar to that for the SFIG at equilibrium. Indeed, is in the continuous limit we have

$$p(x)dx = \begin{cases} \frac{1}{\Omega} \frac{1}{x} & \text{if } x_1 \leq x \leq x_M, \quad (8) \\ 0 & \text{otherwise} \end{cases}$$

where $\Omega = \ln(x_M/x_1)$ is the “volume” in the concomitant size-phase space. The maximum $x_M$ and minimum $x_1$ sizes generally depend on $n_s$, $n_c$, and $N$. Since finite-size effects make it difficult to estimate $x_1$ and $x_M$, we have found it useful to evaluate the volume as $\Omega = 2 \ln(x_{3/4}/x_{1/4})$, where $x_{1/4}$ and $x_{3/4}$ stand for the first and third quartiles of the distribution.

It is convenient for a scale-invariant system to introduce a new variable $x' = x/\theta$ without changing the physics, with $\theta$ a parameter to be later defined. Furthermore, we can re-scale the volume to $\Omega' = C\Omega$ according to

$$x' = \left(\frac{x}{\theta}\right)^C,$$

which leads to the scaled distribution

$$p(x')dx' = \begin{cases} \frac{1}{\Omega'} \frac{1}{x'} & \text{if } \left(\frac{x_1}{\theta}\right)^C \leq x' \leq \left(\frac{x_M}{\theta}\right)^C, \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Note that these changes do not affect the properties of the distribution, which remains that of a SFIG. It is also useful to employ the reduced units defined by $\theta = x_{1/2}$ and $\Omega' = 2$, where $x_{1/2}$ is the median of the distribution. In this particular case, for the new variable $y$ defined by the transformation

$$y = \left(\frac{x}{x_{1/2}}\right)^{\ln(\frac{1}{\ln2})},$$

the density distribution takes the form

$$p(y)dy = \begin{cases} \frac{1}{2} \frac{1}{\sqrt{y}} & \text{if } e^{-1} \leq y \leq e, \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

while the cumulative distribution reads

$$P(y) = \int_{e^{-1}}^{y} dy' P(y') = \frac{\ln y + 1}{2}. \quad (13)$$
In order to represent the size-distribution and the empirical data we will use here the rank-plot (or Zipf plot). For convenience we define a “normalized” rank-parameter $r$ that ranges within the interval [0,1]. This normalized rank-size distribution is derived from the cumulative distribution by extracting $y$ from the equation $r = 1 - P(y)$, which in the case of the distribution (13) gets cast as

$$y = e^{1-2r}. \quad (14)$$

Note that the density distribution (12) and associated rank-sizes (14) do no longer depend on $n_i$, $n_c$, $N$, or $k_M$, since $x_M$ and $x_1$ do not enter the definition of the maximum and minimum sizes when expressed in such units.

### 3.1.2 Regime of high cluster-density: classification by competitiveness

When we increase the number of clusters, the competition for space grows and the size of a cluster depends now on the size of the neighboring clusters. The size distribution exhibits important deviations from the SFIG, but the change to reduced units makes it still possible to compare between distributions obtained with different values of $n_i$, $n_c$, $N$ and $k_M$. These comparisons have led us to find a classification of the distributions using a parameter $\lambda$ – which we will call *competitiveness* – that we pass now to discuss.

The configuration model predicts that for the given degree distribution, the mean number of $j$-th neighbors of a node is

$$z_j = \left( \frac{z_2}{z_1} \right)^{j-1} z_1, \quad (15)$$

where $z_1$ and $z_2$ are the mean number of first and second neighbors, respectively. Consequently, the mean size $\langle x \rangle_s$ of the cluster generated for each individual seed is, at the end of the process,

$$\langle x \rangle_s = \sum_{j=1}^{j_f} z_j = \left[ \sum_{j=1}^{j_f} \left( \frac{z_2}{z_1} \right)^{j-1} \right] z_1 = \lambda^{-1} z_1 \quad (16)$$

where $j_f$ is the mean number of total iterations used in the process, and $\lambda^{-1}$ is a new parameter defined by (16) for future convenience. Since all nodes of the network belong, at the end of the process, to a certain cluster, the mean size times the total number of seeds – which is actually $n_c n_i$ when we take $n_i$ seeds to be members of the same cluster – must be equal to the total number of nodes, i.e.,

$$n_c n_i \lambda^{-1} z_1 = N. \quad (17)$$

For a scale-free ideal network, $z_1 = \langle k \rangle = (k_M - 1) / \ln k_M$, which gives for large $k_M$

$$\lambda = \frac{n_c n_i}{N} \frac{k_M}{\ln k_M} = \rho_s(k). \quad (18)$$

We interpret $\lambda$ – which is shown to be the density of seeds times the mean number of connections – as a quantifier of the strength of the competition between clusters since accounts for the dependence of the size of one cluster on the size of the other clusters: for low competition (low value of $\lambda$) the clusters grow freely with low interference and so behave as a SFIG. Instead, for high competition (high value of $\lambda$) the clusters interact with each other, which entails deviations from the SFIG. We use the competitiveness to classify the family of distributions obtained via our simulations and have studied distributions with values ranging from $\lambda \to 0$ – where the SFIG emerges naturally – up to $\lambda \sim 10$. Anyhow, we have found no evidence of an upper bound for $\lambda$. Since $\lambda$ has been obtained from the predictions of the configuration model, we expect better results for large values of $N$ and $k_M$, where finite size effects are reduced. We depict in Figure 2 the rank-size $y(\lambda, r)$, in semi-log scale, for different values of the competitiveness $\lambda$. The ensuing curves have been obtained by generating several large networks and working with a large number of competitive processes within them, so as to reduce numerical fluctuations.

### 3.2 Study and classification of empirical distributions by competitiveness

We have found that the change of variables performed above (to a reduced unit $y$) can be applied to empirical data so as to compare distributions both for city sizes.
and for electoral results in different countries or states, allowing for an assessment of the effects of societal features, such as policies and economic data. Furthermore, a comparison between these distributions and those obtained via our simulations can be performed to assign a competitiveness value to the empirical distributions. This assignment is effected by minimizing the distance between the data and our computed curve \( y(\lambda, r) \), using a Kolmogorov-Smirnov test [44] (some examples are displayed in Figures 3–10 for electoral results and city populations). Since our simulations nicely fit the data, we are compelled to conclude that in general, the scaled distributions of city populations and electoral results can be classified according to the values of \( \lambda \).

### 3.2.1 City size distributions

We have performed an exhaustive city-population study for the provinces of Spain [45], fitting each pertinent distribution to \( y(\lambda, r) \) and finding a competitiveness-distribution with a median of \( \lambda_{1/2} = 0.65 \). We depict in Figures 3 and 4 the scaled rank-size distributions for some of these provinces, together with the accompanying \( \lambda \)-family of distributions, which nicely fit the associated data. The rank-size distribution of the capital cities has a competitiveness of 0.71 (Fig. 3f), which does not significantly differ from the median value. We contend that the fact that to this last distribution one assigns such a competitiveness-degree is a signature of the scale invariant nature of the social system; the whole country can be thought of as a single network of (only) capital cities, which displays similar statistical properties as those of the complete network, which includes all cities.

We have detected some singular exceptions in the fitting of these curves, as illustrated in Figure 5 for the provinces of Guadalajara and Málaga. We understand that these deviations from the \( \lambda \)-family of distributions reflect either local effects in policies or in social, economical or geographical factors, as some studies have found [46]. In the case of Guadalajara, the uppermost cities in the plot – those that deviate from the best fit – are located in the neighborhood of Madrid, the capital of Spain. The capital of Spain thus affects the population distribution in its neighborhood.

The empirical maximum degree \( \hat{k}_M \), which defines the volume of the SFIN in phase space, has been estimated for each province by solving the equation

\[
\frac{\hat{k}_M}{\ln \hat{k}_M} = \lambda \frac{\hat{N}}{n_v n_s}.
\]
All empirical data are plotted using red dots, and simulation (b) Málaga, where the above deviations are also present. All empirical data are plotted using red dots, and simulation with black lines.

Fig. 5. (Color online) Examples of deviations from the λ-family of distributions. (a) Scaled rank-plot of the city population of Guadalajara province, compared with the best fit. Deviations are seen in the case of the cities at the graph’s top (see text). (b) Málaga, where the above deviations are also present.

Fig. 6. (Color online) Plot of the competitiveness λ of the Spanish provinces versus the empirical values of maximum degree \( k_M \). The medians of both parameters are represented by lines.

Here \( \bar{N} \) is the total population, \( \bar{n}_c \), the number of cities and \( \bar{n}_i \), the population of the smallest city used to estimate the minimum cluster size. We have found that the empirical distribution of the maximum degree exhibits a large tail, whose median is 166. The first and third quartiles are 37 and 319 respectively, hence

\[
\langle k_M \rangle = 166^{+153}_{-129}.
\]  

(20)

We display in Figure 6 the competitiveness of the Spanish provinces versus the empirical values for \( k_M \). The medians of both parameters are also shown. The maximum number of connections is an observable that has been evaluated by the size of their neocortex [48,49]. Dunbar’s number lies between 100 and 230, but a commonly detected value is \( k_D = 150 \), which fits quite well our results. As far as we know, the present work is the first in which Dunbar’s number is computed using a mathematical model based on first principles (our variational MFI).

If the value of \( k_M \) is not fixed but follows itself a distribution law, this may give rise to differences between the form of the empirically observed DD and our theoretical DD (7), obtained from first principles via the MFI variational approach. We can simulate the empirical form by (i) fitting the \( k_M \) - distribution to a log-normal distribution with a logarithmic mean \( \mu = \ln k_D \) and standard deviation \( \sigma = 0.9 \) and (ii) trying now to ascertain how its random nature changes the distribution-form. To this end we generate random values for \( k_i \) according to the distribution (7), but using as \( k_M \), in each case, the exponential of a normal random number with mean \( \mu \) and standard deviation \( \sigma \). In this way we encounter the special DD depicted in Figure 7 (red), that decays smoothly instead of exhibiting a sudden cutoff at a fixed value of \( k_D \). Interestingly enough, this special DD is reproduced by a scale-invariant distribution of the form

\[
F(k)dk \propto \frac{1}{2} \text{erfc} \left[ \frac{\ln k - \mu}{\sqrt{2\sigma}} \right] \frac{dx}{x},
\]  

(21)

where \( \text{erfc}(x) \) is the complementary error function. Equation (7) is recovered in the limit \( \sigma \to 0 \). An exceedingly interesting result can now be obtained. We compare the analytic distribution (21) with the empirical DD found in the largest empirical study thus far performed on a social network, that of Leskovec and Horvitz [32]. We are speaking of the Microsoft Messenger instant-messaging system [32]. This communication channel has two main advantages over other channels: (i) it is instantaneous and (ii) it is free. The Messenger-DD found by these authors is quite different from that of other social networks, which typically are power laws with an exponent between 2 and 3 [16,18,19,30,31]. Instead, its form is quite similar to that of our special DD, fitted by (21). This is illustrated by Figure 7. We get for the Messenger distribution \( \mu = \ln 55 \) and \( \sigma = 0.65 \). This result indicates that just an underestimation of high degree-values is the main difference between...
works with maximum degree computed (blue dots) and extrapolated (black line) – for net-
distribution (see text). Bottom panel, average path length – a BA network, which is not able to reproduce the empirical
tants and nodes. We also show the result of the process in
scaled, which implies a one to one relation between inhabi-
tness of (14). We consider the case of the province of Teruel,
out the need of scaling it via the change to the variable
number is able to reproduce a city-size distribution with-
with a maximum connection-degree equal to Dunbar's
destimation is to be expected.

This result encourages one to check whether a SFIN
with a maximum connection-degree equal to Dunbar’s
number is able to reproduce a city-size distribution with-
out the need of scaling it via the change to the variable
of (14). We consider the case of the province of Teruel,
where a total population of 109 810 inhabitants (exclud-
ing the capital city, since we have already seen that the
capital belongs to a larger national network) is distributed
among 235 cities. We can model it with a network of (i)
$N = 100 000$ nodes, (ii) maximum degree $k_M = k_D = 150$, and (iii) by “growing” $n_c = 250$ clusters with an initial size
of $n_i = 1$ nodes. We depict the rank size distribution in
Figure 8a, where it can be clearly seen that the simulation
nicely fits the data. The city-size distribution is also compared with the cluster growth process obtained in a
Barabasi-Albert network (BA) with the same number of
nodes and clusters. In this graph we see that not all kinds
of networks will be able to reproduce the empirical distrib-
ution, even if we employ a similar number of nodes and clusters.

The average path length (APL) of a network is de-

dined, for all possible pairs of nodes, as the average
umber of steps along the shortest path. It is one of the most
important quantities characterizing a network’s topol-
ogy [18,19]. The configuration model predicts a depend-
ence of the APL on $N$ and $k_M$ of the type

$$\text{APL} \sim \frac{\ln(N/k_M^{*})}{\ln(k_M^{*}/z)} + 1 \simeq \frac{\ln(N \ln(k_M)/k_M)}{\ln(k_M/2)} + 1 \quad (22)$$
i.e., a linear dependence on $\ln N$ ($\text{APL} = a \ln N + b$, where $a$ and $b$ are coefficients that only depends on $k_M$). We have numerically computed the APL of a SFIN with $k_M = 150$
as a function of $N$ up to $N = 100 000$ nodes. One easily
sees the expected dependence on $\ln N$, as illustrated by
Figure 8b. The extrapolation ($a = 0.279$ and $b = 0.826$)
gives $\text{APL} = 4.04$ for a SFIN of $100 000$ nodes, $\text{APL} = 5.73$ for $45 000 000$ nodes (population of Spain), $\text{APL} = 6.26$
for $300 000 000$ nodes (population of the USA [50]), and $\text{APL} = 7.12$ for $6 500 000 000$ nodes (World population).

These values are in accordance with the empirical mea-
sure of Travers and Milgram, known as the “six degrees
of separation” [51–53], and with the more recent results
of Dodds et al., who found an APL between 5 and 7 [54].
It is also the case of Leskovec and Horvitz, who found be-
tween Microsoft Messenger users an average path length of
6.6 degrees [32]. Note that the only parameter that char-
acterizes our DD (obtained from first principles) is the
maximum number of contacts $k_M$. Since in the configura-
tion model all the network’s observables result from this
DD – the coefficients $a$ and $b$ for the APL only depend
on $k_M$ –, one might be tempted to conjecture that the
“six degrees of separation” measure is a consequence of
Dunbar’s number.

3.2.2 Electoral results

We have carried out a similar competitiveness study for
the results of General Elections in different countries. The
cluster size-distribution is here compared with the dis-
tribution of votes and of the results allocated to each
party. We have computed the $\lambda$ value in the cases of
UK’05 [55], USA’04 [56], Italy’08 [57], and Spain’08 [58],
finding $\lambda = 6.5, 4.6, 2.7$, and 0.98, respectively (Fig. 9), all
values being larger than the average found for city pop-
ulations. In general, a high value of the competitiveness
increases, in the results’ tabulation, the difference in the
number of votes between two consecutive parties.

The estimates of the maximum degree are $k_M \sim$
600 000, 3 000 000, 19 000, and 35 000, respectively, i.e.,
many orders of magnitude larger than Dunbar’s number.
Thus, the volume in phase space of the SFIN that de-
scries the election process is larger than that for the city
populations. This is the effect of the creation and develop-
ment of temporary connections. A politician, journalist,
or blog writer can be easily connected during the electoral campaign to thousands of people via mass media, such as television, newspapers, or the Internet. In accordance with Dodds’ results, the world becomes smaller – more connected – when individual incentives exist [54], in this case to obtain good electoral results. These findings lead to interesting conclusions. In the USA’s case we find larger hubs than in the UK: 3 000 000 connections against 600 000, but since the total population is \(N_p = 300\,000\,000\) against 61 000 000 [59], the relative value is similar for each country, \(\bar{k}_M/N_p \sim 0.01\). This value indicates that the USA and the UK have similar social networks in electoral campaigns, but scaled. Since there are more parties competing in elections in UK’s case, the distribution of the results naturally displays a higher competitiveness than in the USA’s one.

### 3.2.3 The universal distribution

Studying the city population of different countries around the world [60] we have found that, in general, for countries with a population over 5 000 000 the main portion of the scaled distributions turns out to be quite similar, thus evidencing some degree of universality, as illustrated in Figure 10 for USA and Germany. Even the distribution of the size of companies in these countries follows this behavior, as depicted in the same figure for USA firms [50]. This universal distribution can be reproduced by our simulation. Note that the competitiveness has a local dependence, and thus data of a country are in fact several sets of data (for many states or provinces of that country), which have different values of the competitiveness. We have simulated this universal distribution by mixing data generated with different values of competitiveness, between 0.4 and 1, obtaining the curve \(y_0(r)\), which fits the empirical distributions as can be seen in Figure 10.

### 4 Conclusions

In this effort we have studied size distributions in social groups by recourse to an information theoretic approach based upon the minimization of Fisher’s information (MFI), being guided in such an endeavor by an analogy between scale invariant systems and classical gases. Going beyond the ideal gas scenario is tantamount to simulating the interactions taking place, for a competitive cluster growth process, in a scale-free ideal network, a non-correlated network with a connection-degree’s distribution that mimics the scale-free ideal gas density distribution.

We employed a scaling rule that allows one to classify the final cluster-size distributions using only one parameter that we called the competitiveness \(\lambda\), which can be seen as a measure of the strength of the interactions and encountered that both empirical city-size distributions and electoral results can be nicely reproduced and classified according to \(\lambda\). In our simulations this parameter is related to the density of clusters and to the maximum degree of the network via equation (18).

With our approach we can easily compute the empirical average of the maximum connection-degree, finding...
that it reproduces Dunbar’s number $k_D$ and its variance [48,49]. Furthermore, when we introduce $k_D$ and its variance into the Fisher-related theoretical DD (7) we are able to reproduce the form of the distribution found in the up-to-date larger empirical study performed on a social network [32]. The importance of $k_D$ becomes explicit when we use it to reproduce the rank-size distribution of the Teruel province using real values (without any sort of scaling) for the number of cities and total population. Our simulations also predict the empirical estimate of the average path-length when we use Dunbar’s number for the maximum connection-degree of the scale-free ideal network (SFIN).

Some studies have found correlations between city-size distribution and regional policies [46]. We believe that the use of the $\lambda$ parameter for such studies would add a very useful tool in order to classify the ensuing distributions. As seen in the case of electoral results, a high value of $\lambda$ in order to attain better chances in the final tallies, whereas a big party would choose a high value in order to increase relative differences with other parties. For city sizes, a low value of the competitiveness may work against supersaturated cities’ influence, whereas a high value may promote the importance of a capital city. In general, all empirical distributions agree quite well with those obtained with our simulation, but we also found some singular exceptions. We expect these to be related to the already mentioned regional policies, and to historical and geographical factors. Thus, our model could help to identify such possible causes. Exhaustive studies of data around the world are necessary to build a bridge between the three variables of equation (18), $\rho_s$, $k_M$, and $\lambda$, and the social and economic polices of a region.

Summing up, our results show that scale invariant gas- thermodynamics à la Fisher yields a useful framework for dealing with scale invariant phenomena. Its application to social sciences here seems to provide some insight into the way humans build up a society. This work only represents a step, and it is expected that subsequent studies will enhance the predictive power of the associated theory.

We would like to thank Manuel Barranco for useful discussions, and to Albert Díaz, Carles Panadés, Joan Manel Hernández, and Antoni García for their helpful comments and remarks. This work has been partially performed under grant FIS2008-00421/FIS from DGI, Spain (FEDER).

References

1. M. Batty, Science 319, 769 (2008)
2. A. Blank, S. Solomon, Physica A 287, 279 (2000)
3. X. Gabaix, Y.M. Ioannides, Handbook of Regional and Urban Economics (North-Holland, Amsterdam, 2004), Vol 4
4. W.J. Reed, J. Reg. Sci. 42, 1 (2002)
5. M.E.J. Newman, Contemp. Phys. 46, 323 (2005)
6. V. Pareto, Cours d’Economie Politique (Droz, Geneva, 1896)
7. G.K. Zipf, Human Behavior and the Principle of Least Effort (Addison-Wesley, Cambridge, MA, 1949)
8. L.C. Malacarne, R.S. Mendes, E.K. Lenzi, Phys. Rev. E 65, 017106 (2001)
9. M. Marsili, Y.-C. Zhang, Phys. Rev. Lett. 80, 2741 (1998)
10. R.L. Axtell, Science 293, 1818 (2001)
11. K. Paech, W. Bauer, S. Pratt, Phys. Rev. C 76, 054603 (2007)
12. X. Campi, H. Krivine, Phys. Rev. C 72, 057602 (2005)
13. Y.G. Ma et al., Phys. Rev. C 71, 054606 (2005)
14. C. Furusawa, K. Kaneko, Phys. Rev. Lett. 90, 088102 (2003)
15. I. Kanter, D.A. Kessler, Phys. Rev. Lett. 74, 4559 (1995)
16. M.E.J. Newman, Phys. Rev. E 64, 016131 (2001)
17. A. Hernando, D. Pugidomènech, D. Villuendas, C. Vesperinas, A. Plastino, Phys. Lett. A 374, 18 (2009)
18. R. Albert, A.L. Barabási, Rev. Mod. Phys. 2074, 2047 (2002)
19. M.E.J. Newman, A.L. Barabási, D.J. Watts, The Structure and Dynamics of Complex Networks (Princeton University Press, Princeton, 2006)
20. T. Maillard, D. Sornette, S. Spaeth, G. von Krogh, Phys. Rev. Lett. 101, 218701 (2008)
21. R.N. Costa Filho, M.P. Almeida, J.S. Andrade, J.E. Moreira, Phys. Rev. E 60, 1067 (1999)
22. A. Hernando, C. Vesperinas, A. Plastino, Phys. A 389, 490 (2010)
23. R. Frieden, A. Plastino, A.R. Plastino, B.H. Soffer, Phys. Rev. E 60, 48 (1999)
24. UrbanSim: http://www.urbansim.org, SLEUTH: http://www.ncgia.ucsb.edu/projects/gig/, DUEM: http://www.casa.ucl.ac.uk/software/duem.asp
25. C. Castellano, S. Fortunato, V. Loreto, Rev. Mod. Phys. 81, 591 (2009)
26. S. Fortunato, C. Castellano, Phys. Rev. Lett. 99, 138701 (2007)
27. M. Batty, Cities and Complexity: Understanding Cities Through Cellular Automata, Agent-Based Models, and Fractals (MIT Press, Cambridge, MA, 2005)
28. H. Gould, J. Tobochnik, An Introduction to Computer Simulation Methods: Applications to Physical Systems, 2nd edn. (Addison-Wesley, 1996)
29. W.J. Reed, B.D. Hughes, Phys. Rev. E 66, 067103 (2002)
30. R. Albert, H. Jeong, A.L. Barabási, Nature 401, 130 (1999)
31. R. Guimerà, R.L. Axtell, Science 293, 1818 (2001)
32. J. Leskovec, E. Horvitz (2008), e-print arXiv:0803.0939v1
33. K. Christensen, H. Flyvbjerg, Z. Olami, Phys. Rev. Lett. 71, 2737 (1993)
34. A. Mezhllumian, S.A. Molchanov, J. Stat. Phys. 71, 393 (1993)
35. S. Zapperi, K.B. Lauritsen, H.E. Stanley, Phys. Rev. Lett. 75, 4071 (1995)
36. A.A. Moreira, D.R. Paula, R.N.C. Filho, J.S. Andrade, Phys. Rev. E 73, 065101(R) (2006)
37. M. Molloy, B. Reed, Random Struct. Algorithms 6, 161 (1995)
38. M.E.J. Newman, S.H. Strogatz, D.J. Watts, Phys. Rev. E 64, 026118 (2001)
39. B.R. Frieden, B.H. Soffer, Phys. Rev. E 52, 2274 (1995)
40. B.R. Frieden, Physics from Fisher Information, 2nd edn. (Cambridge Univ. Press, Cambridge, 1998)
41. B.R. Frieden, Science from Fisher Information (Cambridge Univ. Press, Cambridge, 2004)
42. M. Batty, Nature 444, 592 (2006)
43. R. Toral, C.J. Tessone, Commun. Comput. Phys. 2, 177 (2007)
44. A. Clauset, C.R. Shalizi, M.E.J. Newman (2009), e-print arXiv:0706.1062v2
45. National Statistics Institute of Spain website, Government of Spain, (www.ine.es)
46. L. Valitova, V. Tambovtsev, Regional policy priorities in Russia: empirical evidence, RECEP Reports (2005), Vol. 5, p. 9
47. M. Gladwell, The Tipping Point – How Little Things Make a Big Difference (Little, Brown and Company, 2000)
48. R.I.M. Dunbar, J. Hum. Evo. 20, 469 (1992)
49. R.I.M. Dunbar, Beh. Brain Sci. 16, 681 (1993)
50. Census bureau website, Government of USA, www.census.gov
51. J. Travers, S. Milgram, Sociometry 32, 425 (1969)
52. D.J. Watts, Six Degrees: The Science of a Connected Age (Norton, New York, 2003)
53. D.J. Watts, Small Worlds: The Dynamics of Networks Between Order and Randomness (Princeton University Press, Princeton, 1999)
54. P.S. Dodds, R. Muhamad, D.J. Watts, Science 301, 827 (2003)
55. Electoral Commission, Government of UK, http://www.electoralcommission.org.uk
56. National Archives and Records Administration, Government of USA. www.archives.gov
57. Ministero dell’Interno – Elezioni Politiche, Government of Italy. politiche.interno.it
58. Ministerio del Interior, Elecciones, Government of Spain. www.elecciones.mir.es
59. UK Statistics Authority, www.statistics.gov.uk
60. Wolfram Mathematica CityData Source, based on a wide range of sources. www.wolfram.com