Zipf’s law for cosmic structures: How large are the greatest structures in the universe?

Giordano De Marzo1,2,3, Francesco Sylos Labini1, and Luciano Pietronero1,2,4

1 Centro Ricerche Enrico Fermi, Via Panisperna 89 A, 00184 Rome, Italy
e-mail: sylos@cref.it
2 Sapienza School for Advanced Studies, “Sapienza”, Ple A. Moro, 2, 00185 Rome, Italy
3 Istituto dei Sistemi Complessi (ISC) – CNR, UoS Sapienza, Ple A. Moro, 2, 00185 Rome, Italy
4 Dipartimento di Fisica Università “Sapienza”, Ple A. Moro, 2, 00185 Rome, Italy

Received 14 April 2021 / Accepted 12 May 2021

ABSTRACT

The statistical characterization of the distribution of visible matter in the universe is a central problem in modern cosmology. In this respect, a crucial question still lacking a definitive answer concerns how large the greatest structures in the universe are. This point is closely related to whether or not such a distribution can be approximated as being homogeneous on large enough scales. Here we assess this problem by considering the size distribution of superclusters of galaxies and by leveraging the properties of Zipf–Mandelbrot law, providing a novel approach which complements the standard analysis based on the correlation functions. We find that galaxy superclusters are well described by a pure Zipf’s law with no deviations and this implies that all the catalogs currently available are not sufficiently large to spot a truncation in the power-law behavior. This finding provides evidence that structures larger than the greatest superclusters already observed are expected to be found when deeper redshift surveys will be completed. As a consequence, the scale beyond which galaxy distribution crossovers toward homogeneity, if any, should increase accordingly.

Key words. cosmology: observations – large-scale structure of Universe – methods: statistical

1. Introduction

The distribution of matter in the universe is one of the most fascinating examples of complex structures (Pietronero & Labini 2005, Labini 2011) and it is surely the one concerning the largest objects ever observed. Indeed, such a distribution has been proven to have power-law correlations, corresponding to a fractal structure with a dimension of $D \approx 2$ up to several tens of megaparsecs (Pietronero & Labini 2005). Whether or not it presents a crossover toward homogeneity on larger scales is still a matter of considerable debate (Whibourn & Shanks 2014; Labini et al. 2014; Conde-Saaavedra et al. 2015; Alonso et al. 2015; Pandey & Sarkar 2015; Shirókov et al. 2016; Heinesen 2020; Teles et al. 2021). The fractal behavior of visible matter in the universe corresponds to the fact that galaxies are distributed in a hierarchical manner: They form small groups, which, in turn, aggregate into clusters of galaxies and then subsequently clusters are grouped into larger structures, that is to say superclusters and filaments (De Vaucouleurs 1953; Abell et al. 1989), which are the largest known structures in the universe.

Superclusters are linked by these filaments of galaxies and clusters forming the so-called cosmic web (Bond et al. 1996) or superclusters-void network (Einasto et al. 1980), corresponding to a complex distribution of matter characterized by large voids and connected over-densities (Tully et al. 2014; Pomarède et al. 2017, 2020; Colin et al. 2019).

After the surprising discovery of the cosmic web, much attention has been devoted to the quantitative statistical characterization of such large-scale structures, which represent a fundamental aspect of the observable universe and a key issue for any cosmological theory. Indeed, standard cosmological theories are based on the assumption of a uniform distribution of matter and so the identification of the “end of greatness” of galaxy structures represents a key issue in observational astrophysics. In the present paper, we show that there is no statistical evidence that the available data contain the largest existing structures. In addition, large-scale structures carry important information about the primordial universe and pose intriguing theoretical problems concerning their formation (Einasto et al. 2007; Oort 1983). For these reasons many authors have studied the features of the cosmic web, considering the distribution of voids (Hamaus et al. 2014; Pan et al. 2012; Einasto et al. 2011) and that of superclusters (Einasto et al. 1997), and they have tried to reproduce this complex structure with $N$-body numerical simulations of theoretical models (Springel et al. 2005).

In the present work, we analyze the large-scale structure of the universe considering the size distribution of galaxy superclusters and leveraging the properties of Zipf–Mandelbrot law (Mandelbrot 1953; Zipf 1949), which may provide interesting hints as to the greatest size of these objects. Previous analyses have considered Zipf’s scaling for voids (Gaite 2005; Tikhonov 2006), which were found to be well described by this statistical law, and the probability distribution of superclusters size (Chow-Martínez et al. 2014), which was found to be a power law. It is worth noting that cosmological $N$-body simulations do not show such a power-law behavior (Chow-Martínez et al. 2014) and consequently they miss this feature of the large-scale structure of the universe. On the other hand, to the best of our knowledge, the Zipf–Mandelbrot law for galaxy structures has never been considered. As we discuss below, the first problem to be considered is that of defining a reliable list of a supercluster’s size: We approached this by using different definitions and
trying to assess the robustness of the results on different observational catalogs. It is worth pointing out that for the present work no analysis on catalogs obtained by N-body simulation was performed. In particular, we tried to quantify the deviations from a pure Zipf’s law, which represent one of the key elements in this statistical characterization of structures. Indeed, it has been recently shown that such deviations allow one to characterize the upper cutoff of the power-law probability distribution of structures size (De Marzo et al. 2021). It is interesting to note that such a cutoff is indirectly related to the scale up to which the distribution of matter in the universe cannot be approximated as being statistically uniform, providing a complementary quantitative tool beyond that based on the correlation functions analysis.

The paper is organized as follows: In Sect. 2, we shortly recall the definition of Zipf’s law and Zipf–Mandelbrot law, pointing out how the level of deviations from Zipf’s law can be used to obtain a characterization of the upper cutoff of the power-law probability distribution. Then in Sect. 3, we present the supercluster catalogs used in the analysis and we compute Zipf–Mandelbrot parameters for all of them. Section 4 is devoted to the discussion of these results, which are also compared with those obtained for cluster catalogs. A measure of statistical fluctuations is introduced ad applied to the various samples. Finally, Sect. 5 is devoted to our final remarks and conclusions.

2. Zipf’s and Zipf–Mandelbrot’s laws

Zipf’s law (Zipf 1949) is a ubiquitous scaling law which can be considered as a sort of footprint of complexity. Indeed, this law has been observed in many complex socio-economic systems, such as cities, firms, and natural language, as well as in a large number of natural systems, such as earthquakes, solar flares, and lunar craters (Zipf 1949; Newman 2005; Cristelli et al. 2012).

Many attempts have been devoted to identify a mechanism capable of explaining all of these manifestations of Zipf’s scaling (Pietronero et al. 2001; Bak et al. 1987; Corominas-Murtra et al. 2015; Tria et al. 2014), but a universally accepted generating process is still missing.

Given a system composed of \( N \) objects and denoting by \( S(k) \), the size of the \( k \)th largest object, Zipf’s law reads

\[
S(k) = \frac{S(1)}{k^\gamma} \tag{1}
\]

where \( k \) is the rank and \( \gamma \) is called the Zipf’s exponent. In the case we are going to consider, \( S(1) \) is the size of the largest supercluster. \( S(2) \) is that of the second largest one, and so on. Zipf’s law is generally visualized by the so-called rank-size plot, obtained plotting the ordered sequence of the sizes as a function of their position in the sequence: If the rank-size plot is a straight line in the log–log scale, then the system is said to follow Zipf’s law.

Zipf’s law is intimately related to power-law probability distributions. We consider a set of objects whose sizes \( S \) are power-law distributed, that is

\[
P(S) = \frac{c}{S^\alpha}, \tag{2}
\]

where \( c \) is a normalization constant; the intrinsic upper and lower cutoffs of the distribution are called \( s_m \) and \( s_M \), respectively, and Eq. (2) is satisfied for \( s_m \leq S \leq s_M \). We note that such cutoffs are always present in real systems due to intrinsic limits on the size that an object can have: For instance, a supercluster cannot contain less than a galaxy or more than all of the existing galaxies. Generally a power-law behavior is found only in the tail of the distribution, so in correspondence of the largest objects. It can be proven (Cristelli et al. 2012; De Marzo et al. 2021) that if the underlying probability distribution is a power law, then Zipf’s law is valid asymptotically in the rank, that is

\[
S(k) \sim k^{-\gamma} \text{ for large values of } k \tag{3}
\]

and moreover the Zipf’s exponent \( \gamma \) is related to the exponent \( \alpha \) of the distribution \( P(S) \) by the following expression (Li 2002)

\[
\gamma = \frac{1}{\alpha - 1}. \tag{4}
\]

Real systems often show deviations from Zipf’s scaling law (Cristelli et al. 2012; De Marzo et al. 2021), which can be found at low ranks (for the largest objects) or at high ranks (in correspondence with the smallest elements). The latter may derive from the fact that the distribution has a power-law behavior only in the tail or from selection effects. For instance, we could observe only a fraction of small superclusters due to a selection effect in apparent luminosity.

On the other hand, deviations at low ranks can carry important information about the systems considered and they are usually described in terms of Zipf–Mandelbrot’s law (Mandelbrot 1953), which reads as follows:

\[
S(k) = \frac{S}{(k + Q)^\gamma}. \tag{5}
\]

Here \( S \) bar is defined by the relation \( S = S(1)(1+Q)^\gamma \). Zipf’s law is recovered for \( Q = 0 \), while strong deviations from the simple power-law behavior of Eq. (3) at low ranks are found for \( Q \gg 1 \). Deviations at low ranks are related to the level of sampling of the distribution and to the presence of cutoffs: Indeed, as is shown by De Marzo et al. (2021), the parameter \( Q \) satisfies

\[
Q = N \left( \frac{s_m}{s_M} \right)^{1/\gamma}, \tag{6}
\]

where \( N \) is the number of objects in the sample considered. The larger the extent of the distribution given by the ratio between the two cutoffs is, the smaller \( Q \) is; whereas the larger \( N \) is (i.e., the finer is the sampling), the larger \( Q \) is. We note that Zipf’s law may arise in the following two very different ways. First, the absence of deviations can be related to the fact that the sample that is available is too small to allow for the detection of the intrinsic upper cutoff of the distribution. This occurs, for instance, by analyzing the size distribution of earthquakes on small time scales (tens of years) (De Marzo et al. 2021). In this case, gradually enlarging the sample leads to a growth of deviations from a pure Zipf’s law. Second, some systems spontaneously evolve out of equilibrium toward Zipf’s law, as it occurs for cities (De Marzo et al. 2021). In this case, even considering the whole sample composed of all the urban settlements of a given country, deviations from Zipf’s law are absent. In both cases, whenever Zipf’s law is found in a sample consisting of only a fraction of the whole system, the observed maximum cannot be used as an estimate for the intrinsic upper cutoff since it is not possible to determine if the low level of sampling is an intrinsic property of the system or an effect produced by the limited system’s size. Indeed, as is shown in De Marzo et al. (2021), the intrinsic upper cutoff of the probability distribution \( s_M \) can be expressed as a function of the observed maximum \( S(1) \) as

\[
s_M = S(1) \left( 1 + \frac{Q}{Q} \right)^{-\gamma}. \tag{7}
\]
For $Q \to \infty$, the intrinsic upper cutoff coincides with the observed maximum (i.e., the largest observed object); whereas for $Q \to 0$, the upper cutoff diverges, meaning that the sample available is not sufficiently large to infer it.

3. Zipf’s law for superclusters of galaxies

Superclusters of galaxies are the largest objects in the observable universe and they can extend for hundreds of megaparsecs containing thousands of galaxies. In order to study the rank-size relation of superclusters, first of all we have to define a suitable measure of their size. Moreover, if we want to prove that Zipf’s law is a robust feature of the distribution of matter in the universe, such a scaling law should be found independently of the size definition adopted. For these reasons, following the hierarchy of cosmic structures, we considered three different definitions of size:

- the number of galaxies in the supercluster,
- the number of groups in the supercluster,
- the number of clusters in the supercluster.

The main problems of these catalogs are their completeness and the possible artifacts due to observational selection effects and clustering algorithms. Both of which are not simple to take into account. In order to verify that the properties we have detected are intrinsic properties of superclusters, we have analyzed several catalogs, which have been built by using different procedures. In the following subsections, we discuss the analyses we have performed in detail.

3.1. Number of galaxies

The rank-size plots of superclusters, which are ordered according to the number of galaxies they contain, are reported in Fig. 1a. Four different catalogs have been used: Einasto et al. (2006a,b), Einasto et al. (2007), and Liivamägi et al. (2012; which contains two distinct catalogs corresponding to different clustering procedures). We also reported the corresponding fits to Zipf–Mandelbrot law, so as to determine the parameter $Q$ and the extent of the deviations from Zipf’s law. For a detailed explanation of the fitting procedure and of the technique used for determining the uncertainty of the fit parameters, readers can refer to the Appendix A. As it is possible to see, superclusters follow Zipf’s law with negligible deviations at low ranks.

3.2. Number of groups

Groups are the first level of aggregation of galaxies, for instance the Milky Way is part of the Local Group, which contains approximately 70 galaxies. The catalogs we used for studying the rank-size plot of superclusters, which are ordered by the number of groups they contain, are the same we considered in the previous subsection. The rank-size plots and the corresponding fits to Zipf–Mandelbrot law are reported in Fig. 1b: Also, in this case, only small deviations from Zipf’s law can be detected.

3.3. Number of clusters

Finally, groups of galaxies merge into clusters, which are the level of aggregation just below superclusters. In order to study the adherence to Zipf’s law when superclusters are measured through the number of clusters they contain, we used five catalogs: Einasto et al. (2001), Einasto et al. (2006a,b), Einasto et al. (2007), and Chow-Martínez et al. (2014; which contains two distinct catalogs). The rank-size plots and the corresponding fits to Zipf–Mandelbrot law are reported in Fig. 1c and again the adherence to Zipf’s law is almost perfect.

4. Discussion

In the previous subsection we analyzed the rank-size distribution of superclusters of galaxies considering different definitions of their size and different catalogs. In all cases, negligible deviations from Zipf’s law have been observed. Results are summarized in Table 1, where we also report the minimal value of the deviation parameter $Q_{\text{min}}$ with 95% confidence computed using a parametric bootstrap method (see Appendix A for details). In all cases, the minimum value of $Q$, that is $Q_{\text{min}}$, is almost null, meaning that the system we considered is well described by a perfect Zipf’s law, regardless of the definition of size adopted or of the catalog analyzed. As discussed above, this implies that the available data are not sufficient to determine the intrinsic upper cutoff of the distribution of superclusters since, as shown in Eq. (7), this cutoff diverges for $Q \to 0$. In other words, the
conclusion is that the observed maximum is a bad estimator of the intrinsic one.

This suggests that cosmic structures that are much larger than those contained in the catalogs that we analyzed are likely to be found when larger portions of the universe will be available. Moreover, the absence of a cutoff implies that the distribution of matter in the universe cannot be approximated as homogeneous on the scales of the surveys considered. Indeed, even if a truncation point should exist, the distribution of superclusters is de facto scale invariant up to the largest scales currently observed. We discuss the implications of these results with respect to the standard analysis of homogeneity in the next section.

4.1. Inhomogeneity of the universe

Standard statistical methods based on the determination of the reduced two-point correlation function, or of its Fourier conjugate, are not suitable to test whether a certain distribution is homogeneous in a given sample. The reason is that these statistical tools compare the amplitude of fluctuations to the sample density, that is they assume the sample density to be an unbiased estimator of the density. Such a situation occurs only if the distribution is homogeneous and here the problem arises of testing whether or not a distribution is homogeneous without assuming a priori that this is verified inside a given sample. This point has been discussed at length, for example in Pietronero & Labini (2005), Labini (2011), and we refer the interested reader to those works for a more detailed discussion on the normalization problem.

There is an additional and more subtle problem which may affect any statistical determination in a finite sample, that is even those in which the sample density is an unbiased estimator of the intrinsic average density. Indeed, statistical analyses of finite sample distributions usually assume that fluctuations are self-averaging, meaning that they are statistically similar in different regions of the given sample volume. By determining the conditional density, that is the average galaxy density around a galaxy, Labini et al. (2009a,b) have tested whether this assumption is satisfied in several subsamples of the Sloan Digital Sky Survey (SDSS). The result being that the probability density function (PDF) of conditional fluctuations, filtered on large enough spatial scales (i.e., $r > 30$ Mpc $h^{-1}$), shows relevant systematic variations in different subvolumes of the survey. Instead for scales $r < 30$ Mpc $h^{-1}$, the PDF is statistically stable, and its first moment presents scaling behavior with a negative exponent around one. Thus, while up to 30 Mpc $h^{-1}$, galaxy structures have well-defined power-law correlations; on larger scales, it is not possible to consider whole sample average quantities as meaningful and useful statistical descriptors of intrinsic statistical properties. The conclusion of these studies was that this situation is due to the fact that galaxy structures correspond to density fluctuations which are too large in amplitude and too extended in space to be self-averaging on such large scales inside the sample volumes: The galaxy distribution is thus inhomogeneous up to the largest scales, that is $r \approx 100$ Mpc $h^{-1}$, probed by the SDSS samples. It is worth pointing out that the self-averaging property is a necessary, but not sufficient, condition for the distribution of matter to be homogeneous. Indeed, as aforementioned, even if up to 30 Mpc $h^{-1}$ galaxy structures are self-averaging, they still present power-law correlations and thus a fractal behavior, that is all but homogeneous.

As we are going to show in what follows, the lack of the self-averaging property is related to the absence of deviations from Zipf’s law for superclusters. Indeed, as we have pointed out, Zipf’s law corresponds to an under-sampled probability distribution and, consequently, splitting the sample worsens the level of sampling. This implies that when moving from subsample to subsample, strong fluctuations are expected. In order to show this, we introduced a measure of the large-scale statistical fluctuations of the sample, $R$. Given a sample divided in $M$ subsamples, this parameter is defined as

$$R(M) = \frac{\max_m \{s_m(1)\}}{\min_m \{s_m(1)\}}$$

where $s_m(1)$ is the size of the largest element contained in the $m$th subsample. For $R = 1$ all subsamples present similar large-scale structures, while for $R > 1$ fluctuations appear.

First of all, we consider the case $M = 2$, which corresponds to splitting the sample in two halves, and we suppose that the system obeys Zipf–Mandelbrot law as seen in Eq. (5). The size of the $k$th largest structure satisfies

$$S(k) = \frac{\hat{S}}{(Q + k)^\gamma}$$

and, as a consequence, it holds

$$\text{Prob} \left[ R = \frac{\hat{S}}{(1 + \frac{Q}{1 + k})^\gamma} \right] = \frac{1}{2}$$

$$\text{Prob} \left[ R = \left(1 + \frac{Q}{1 + k} \right)^\gamma \right] = \left(\frac{1}{2}\right)^2$$

$$\vdots$$

$$\text{Prob} \left[ R = \left(\frac{Q + Q}{1 + k} \right)^\gamma \right] = \left(\frac{1}{2}\right)^{k-1}.$$
The mean fluctuation parameter \( \langle R \rangle \) is then

\[
\langle R \rangle = \sum_{n=2}^{\infty} \left( \frac{n + Q}{1 + Q} \right)^{\gamma} \left( \frac{1}{2} \right)^{n-1} = 2 \sum_{n=2}^{\infty} \left( \frac{n + Q}{1 + Q} \right)^{\gamma} \left( \frac{1}{2} \right)^{n}.
\]

This quantity is plotted as a function of \( Q \) and for different values of \( \gamma \) in Fig. 2a. As can be seen, the larger \( Q \) is, the larger the level of sampling of the inherent distribution is, and the lower the level of fluctuations is. The values of \( \gamma \) measured in the catalogs are in the range \( \gamma \in [0.4, 0.8] \); then for \( Q = 0 \), that is, for a perfect Zipf’s law, we have \( 1.5 < R < 2.5 \). This implies that one subsample contains, on average, a structure that is between 50% and 150% larger than the largest structure of the other set. We then conclude that in absence of deviations from Zipf’s law, the two subsamples are consistently different when looking at the scale of superclusters.

Things worsen if we divide the sample in \( M > 2 \) groups, as is shown in Fig. 2b. In this case, we plotted the parameter \( R \) determined in the four catalogs we analyzed in Sect. 3.1 for increasing values of \( M \), that is when the sample is divided into more and more subsets. It is possible to see that the larger \( M \) is, the larger \( R \) is; in particular, for \( M = 16 \), the large-scale statistical fluctuation parameter \( R \) ranges from 6 (Liivamägi et al. 2012) to 16 (Einasto et al. 2007). We note that this implies that if we split the observed distribution of matter into 16 subsamples, one of them would contain, on average, a structure up to 16 times bigger than the largest supercluster contained in the other subsamples. This indicates that at the scale of superclusters, so at scales of about one hundred megaparsecs, the distribution of matter is not self-averaging and that consequently it cannot be approximated as being statistically uniform. In the same figure, we also plotted (black dotted line) the fluctuation parameter \( R \) obtained for a synthetic sample composed of \( N = 100 \) structures, which perfectly adhere to Zipf’s law with exponent \( \gamma = 0.6 \) and in absence of any statistical noise.

### 4.2. Clusters of galaxies

A natural issue to investigate is the behavior of the distribution of matter in the universe when looked at smaller scales. This can be done, for instance, by analyzing the rank-size distribution of clusters instead of that of superclusters. To this aim, we considered two large datasets: the first is described in Hao et al. (2010) and the second is described in Wen et al. (2012). The corresponding rank-size plots and the corresponding fits to Zipf–Mandelbrot law are reported in Fig. 3a. It is possible to see that there are strong deviations from Zipf’s law. The deviation parameters \( Q \) with their minimum value at 95% confidence and Zipf’s exponents \( \gamma \) are reported in Table 2. In both cases \( Q \) is not compatible with a null value, meaning that at the level of aggregation of clusters, the cutoff of the distribution is clearly visible and the distribution itself is completely sampled. Consequently we do not expect to observe clusters much richer than those already observed when larger portions of the universe will be mapped. Moreover, as discussed above, the presence of strong deviations from Zipf’s law also implies that at the scales of clusters, that is on scales of a few megaparsecs, the distribution of matter is self-averaging, confirming previous findings (Labini et al. 2009a,b). Indeed, as is shown in Fig. 3b, the inhomogeneity parameter \( R \), as seen in Eq. (8), is much smaller than that obtained in the case of superclusters, as can be seen in Fig. 2b, since its value is less than 2 when the sample is divided into \( M = 16 \) groups. We also plotted (black dotted line) the fluctuation parameter of a synthetic sample composed on \( N = 1000 \) elements described by a Zipf–Mandelbrot law with parameters \( \gamma = 0.27, Q = 5 \), and in the absence of any statistical noise. Again, we note that this does not imply that on small scales the distribution of matter is homogeneous, but only that it is self-averaging. Indeed, previous studies have shown that galaxy structures are fractal on the scale of galaxy clusters.

### 5. Conclusions

The determination of the properties of the distribution of matter in the universe is one of the most important and challenging
The black dotted line is the fluctuation parameter of a synthetic set of N Zipf–Mandelbrot law. In particular we focus on galaxy superclusters, the way; here we present a novel approach to the problem based on the structure rather than being spread in a statistically uniform manner in the universe is characterized by a complex hierarchy and properties of such a distribution using the tools of complex systems. The application of this Zipf’s law based analysis provides a novel and complementary vision with respect to those based on the study of the correlation functions or other statistical properties. Moreover it also demonstrates that Zipf’s law, which is often considered as a mere empirical law, can be used to gather fundamental information about systems characterized by a complex structure. The application of this Zipf’s law based analysis to the forthcoming galaxy surveys and to N-body simulated catalogs represent a natural and interesting extension of the study presented in this work. In particular, by following the dynamical evolution of the structures during a numerical simulation, it would be possible to determine if Zipf’s law is an intrinsic feature of superclusters of galaxies or a spurious manifestation of this scaling law (De Marzo et al. 2021).

### Table 2. Clusters of galaxies.

| Dataset      | Q  | Q_{min} | γ      |
|--------------|----|---------|--------|
| Hao 2010     | 13.28 | 8.96 | 0.30  |
| Wen 2012     | 6.13  | 2.39  | 0.25  |

**Notes.** Zipf–Mandelbrot parameters for the two catalogs of clusters we considered from Hao et al. (2010) and Wen et al. (2012).

problems in modern cosmology since the presence of large-scale structures of galaxies may be in contrast with the “end of greatness”. This distribution shows a power-law correlation, corresponding to a fractal behavior with the dimension \( D \approx 2 \) up to tens of megaparsecs; it is thus natural to assert statistical properties of such a distribution using the tools of complex systems. Many works have gone in this direction after the finding that matter in the universe is characterized by a complex hierarchical structure rather than being spread in a statistically uniform way; here we present a novel approach to the problem based on Zipf’s law. In particular we focus on galaxy superclusters, the largest structures in the universe, assessing their adherence to Zipf–Mandelbrot law.

Considering several catalogs and different definitions of the supercluster size, we found that superclusters show Zipf–Mandelbrot law with the deviation parameter \( Q \) close to zero, meaning that they almost perfectly adhere to Zipf’s law. This finding has several implications as follows.

- Despite the catalogs we considered are the largest currently available, they do not show the presence of an intrinsic upper cutoff of the probability distribution of superclusters, that is they consequently scale free up to the sizes of the largest structures observed.
- The absence of an intrinsic cutoff in the available samples implies that structures much larger than those currently observed are expected to be found as soon as deeper catalogs will be released.
- The probability distribution of superclusters is only poorly sampled and, as a consequence, different subsamples of the catalogs we considered show different large-scale properties. This implies that the distribution of matter is not self-averaging on the scales of tens of megaparsecs, confirming previous findings, and thus that it cannot be approximated as being homogeneous on such scales.

Exploiting the same methodology, we also analyzed the distribution of matter on smaller scales, considering the distribution of clusters rather that that of superclusters. In this case, strong deviations from Zipf’s law are observed and therefore, on scales of a few megaparsecs, the distribution of matter is self-averaging, confirming previous findings which have also shown this distribution to be fractal on such scales. Our study shows that the self-averaging property is deeply entangled with the level of sampling of the probability distribution of cosmic structures and with the presence of deviations from Zipf’s law. This approach thus provides a novel and complementary vision with respect to those based on the study of the correlation functions or other statistical properties. Moreover it also demonstrates that Zipf’s law, which is often considered as a mere empirical law, can be used to gather fundamental information about systems characterized by a complex structure. The application of this Zipf’s law based analysis to the forthcoming galaxy surveys and to N-body simulated catalogs represent a natural and interesting extension of the study presented in this work. In particular, by following the dynamical evolution of the structures during a numerical simulation, it would be possible to determine if Zipf’s law is an intrinsic feature of superclusters of galaxies or a spurious manifestation of this scaling law (De Marzo et al. 2021).

**Acknowledgements.** We thanks the referee Professor Bernard Jones for his helpful comments about the manuscript and for his suggestions about possible further studies.

**References**

Abell, G. O., Corwin, H. G., Jr., & Olowin, R. P. 1989, ApJS, 70, 1
Alonso, D., Salvador, A. I., Sánchez, F. J., et al. 2015, MNRAS, 449, 670
Bak, P., Tang, C., & Wiesenfeld, K. 1987, Phys. Rev. Lett., 59, 381
Bond, J. R., Kofman, L., & Pogosyan, D. 1996, Nature, 380, 603
Burroughs, S. M., & Tebbens, S. F. 2001, Pure Appl. Geophys., 158, 741
Chow-Martínez, M., Andernach, H., Caretta, C., & Trejo-Alonso, J. 2014, MNRAS, 445, 4073

Exploiting the same methodology, we also analyzed the distribution of matter on smaller scales, considering the distribution of clusters rather that that of superclusters. In this case, strong deviations from Zipf’s law are observed and therefore, on scales of a few megaparsecs, the distribution of matter is self-averaging, confirming previous findings which have also shown this distribution to be fractal on such scales. Our study shows that the self-averaging property is deeply entangled with the level of sampling of the probability distribution of cosmic structures and with the presence of deviations from Zipf’s law. This approach thus provides a novel and complementary vision with respect to those based on the study of the correlation functions or other statistical properties. Moreover it also demonstrates that Zipf’s law, which is often considered as a mere empirical law, can be used to gather fundamental information about systems characterized by a complex structure. The application of this Zipf’s law based analysis to the forthcoming galaxy surveys and to N-body simulated catalogs represent a natural and interesting extension of the study presented in this work. In particular, by following the dynamical evolution of the structures during a numerical simulation, it would be possible to determine if Zipf’s law is an intrinsic feature of superclusters of galaxies or a spurious manifestation of this scaling law (De Marzo et al. 2021).

**Acknowledgements.** We thanks the referee Professor Bernard Jones for his helpful comments about the manuscript and for his suggestions about possible further studies.

**References**

Abell, G. O., Corwin, H. G., Jr., & Olowin, R. P. 1989, ApJS, 70, 1
Alonso, D., Salvador, A. I., Sánchez, F. J., et al. 2015, MNRAS, 449, 670
Bak, P., Tang, C., & Wiesenfeld, K. 1987, Phys. Rev. Lett., 59, 381
Bond, J. R., Kofman, L., & Pogosyan, D. 1996, Nature, 380, 603
Burroughs, S. M., & Tebbens, S. F. 2001, Pure Appl. Geophys., 158, 741
Chow-Martínez, M., Andernach, H., Caretta, C., & Trejo-Alonso, J. 2014, MNRAS, 445, 4073
Appendix A

Derivation of the expressions for $Q$ and $s_M$

As in De Marzo et al. (2021), we consider a truncated power-law distribution of sizes, $P(S)$, that is

$$P(S) = \begin{cases} 0 & \text{for } S < s_m \\ \frac{c}{s^\gamma} & \text{for } s_m \leq S \leq s_M \\ 0 & \text{for } S > s_M \end{cases} \quad (A.1)$$

where $c$ is the normalization constant, and $s_m$ and $s_M$ are the lower and upper cutoff of the distribution. We recall that in the present case, $S$ coincides with the size of superclusters. These cutoffs are connected to $c$ by the normalization condition

$$c \int_{s_m}^{s_M} \frac{ds}{s^\gamma} = 1 \Rightarrow c = \frac{\alpha - 1}{s_m^{\alpha - 1} - s_M^{\alpha - 1}}. \quad (A.2)$$

Now, in considering the PDF $P(S)$ of a continuous variable $S$, the values of its cumulative distribution function (CDF) $C(S)$ are equiprobable. This is because the CDF is defined as $C(S) = \int_{S_m}^{S} ds' P(s')$ and so, in performing the change of variable from $S \rightarrow C = C(S)$ and denoting $f(C)$ as the PDF of $C$, we get $f(C) = \frac{ds}{s^\gamma} P(s) |_{S_m}^{S} = 1$ for $0 \leq C \leq 1$. As a consequence, given $N$ values of $S$ are independently extracted from $P(S)$, with a good approximation they can be considered as uniformly spaced in the corresponding variable $C$. Thus, the $k$th size ranked value $S(k)$ approximately corresponds to the CDF value $\frac{k-\frac{1}{2}}{N+1}$. In formulas

$$\int_{S_m}^{S(k)} P(S) dS = c \int_{s_m}^{S(k)} \frac{dx}{x^\gamma} \approx \frac{N + 1 - k}{N + 1},$$

which, together with Eq. (A.2), gives

$$S(k) \approx s_m + \frac{\gamma}{\alpha - 1} \cdot \left(\frac{k - \frac{1}{2}}{N + 1}\right)^{\frac{1}{\gamma}} = \frac{N + 1 - k}{N + 1} \cdot s_m + s_M \cdot \left(\frac{k - \frac{1}{2}}{N + 1}\right)^{\frac{1}{\gamma}}.$$ 

Making the assumption that $N + 1 \approx N$, $s_M \gg s_m$, and in introducing $\gamma = \frac{1}{\alpha - 1}$, we end up with the final rank-size formula

$$S(k) = \left[\frac{N s_m^{\frac{\gamma}{\alpha - 1}} s_M^{\frac{\gamma}{\alpha - 1}}}{N s_m + k s_M^{\frac{\gamma}{\alpha - 1}}}\right]^\frac{1}{\gamma} = \left[\frac{N^\gamma s_m}{k + N \left(\frac{s_m}{s_M}\right)^{\frac{\gamma}{\alpha - 1}}}\right]^\frac{1}{\gamma}. \quad (A.3)$$

By comparing this expression with the Zipf–Mandelbrot law, as seen in Eq. (5),

$$S(k) = \frac{S}{(k + Q)^\gamma}, \quad (A.4)$$

it follows that

$$\begin{cases} \gamma = \frac{1}{\alpha - 1} \\ \tilde{S} = N^{\gamma} s_m \\ Q = N \left(\frac{s_m}{s_M}\right)^{\gamma - 1} \end{cases}. \quad (A.5)$$

These expressions relate the number of values or objects in the system and the parameters of the PDF $P(S)$ on one side, and the Zipf–Mandelbrot parameters on the other. We note that $Q$ not only quantifies deviations from Zipf’s law, but it also quantifies the level of sampling of the inherent distribution. Indeed $Q$ is the larger the wider the statistical sample is, thus the larger the number $N$ of objects in the sample is and it is the smaller the wider the extension of the truncated power law is, given by the ratio between the upper cutoff and the lower one.

Finally, combining Eqs. (A.4) and (A.5), we obtain an expression for the upper cutoff $s_M$

$$S(1) = \frac{N^{\gamma} s_m}{Q + 1} = s_M \frac{N^{\gamma} s_m}{Q + 1} = s_M \left(\frac{Q}{Q + 1}\right)^\gamma,$$

which yields

$$s_M = S(1) \left(\frac{Q + 1}{Q}\right)^\gamma.$$

Fitting procedure and uncertainty assessment

We adopted a standard nonlinear least squares (NLS) fitting procedure to determine the parameters of Zipf–Mandelbrot law. This technique, if applied to the rank-size plot, or, equivalently, to the complementary cumulative distribution, has an accuracy which is comparable to maximum likelihood estimates (White et al. 2008), while being much simpler (Burroughs & Tebbens 2001) if the upper cutoff is unknown. In particular, we used Eq. (5) partially linearized through logarithms

$$\log S(k) = -\gamma \log (k + Q) + c,$$

where $Q$, $\gamma$, and $c$ are free parameters. The fitting algorithm returns 95% confidence bounds on these parameters; however, such uncertainty does not take into account statistical fluctuations that are encountered considering different samples generated by the same power-law distribution. For this reason we exploited a parametric bootstrap so as to obtain a more realistic measure of uncertainty. In particular we adopted the following procedure

- we computed the parameter $Q$ of the empirical data with the NLS technique;
- we used Eqs. (4) and (6) to determine the parameters of the underlying power-law distribution;
- we generated $M = 1000$ Monte Carlo samples with numerosity $N$ as the empirical sample using the power-law distribution obtained in the previous step;
- each synthetic sample $m$ was fitted with the NLS technique, so as to obtain the parameters $Q^m$;
- the empirical distribution of $Q$, $P(Q)$, was obtained by performing a histogram of the $M$ values $Q^m$;
- starting from the probability distribution $P(Q)$, the confidence bound for $Q$ was easily obtained using the cumulative distribution and determining the interval containing 95% of the probability.