Galaxy distribution and extreme-value statistics

T. ANTAL[1], F. SYLOS LABINI[2,3], N. L. VASILYEV[4] and Y. V. BARYSHEV[4]

[1] *Program for Evolutionary Dynamics, Harvard University - Cambridge, MA 02138, USA*
[2] *Museo Storico della Fisica e Centro Studi e Ricerche Enrico Fermi - Piazzale del Viminale 1,*
00184 Rome, Italy, EU
[3] *Istituto dei Sistemi Complessi CNR - Via dei Taurini 19, 00185 Rome, Italy, EU*
[4] *Institute of Astronomy, St. Petersburg State University - Staryj Peterhoff, 198504, St. Petersburg, Russia*

Abstract – We consider the conditional galaxy density around each galaxy, and study its fluctuations in the newest samples of the Sloan Digital Sky Survey Data Release 7. Over a large range of scales, both the average conditional density and its variance show a non-trivial scaling behavior, which resembles criticality. The density depends, for $10 \leqslant r \leqslant 80\,\mathrm{Mpc/h}$, only weakly (logarithmically) on the system size. Correspondingly, we find that the density fluctuations follow the Gumbel distribution of extreme-value statistics. This distribution is clearly distinguishable from a Gaussian distribution, which would arise for a homogeneous spatial galaxy configuration. We also point out similarities between the galaxy distribution and critical systems of statistical physics.

Introduction. – One of the cornerstones of modern cosmology is the mapping of three-dimensional galaxy distributions. In the last decade two extensive projects, the Sloan Digital Sky Survey (SDSS —[1]) and the Two-degree Field Galaxy Redshift Survey (2dFGRS —[2]), have provided redshifts of an unprecedented quality for more than one million galaxies. A common feature observed in these surveys [3,4] is that galaxies are organized in a complex pattern, characterized by large-scale structures: clusters, super-clusters, and filaments with large voids of extremely low local density [5]. Recent analyses of these catalogs have shown that galaxy structures display large amplitude density fluctuations at all scales limited only by sample sizes [6–10]. In addition, the conditional density [11] has been found to decay with distance as a power law function with an exponent close to one, up to ∼ 30 Mpc/h (see footnote[1]). At larger scales, the situation was unclear since in the 2dFGRS the relatively small solid angle prevents the proper characterization of correlations at larger scales [9,10]. Conversely, the SDSS samples (data release 6 —DR6) clearly show that conditional fluctuations are not self-averaging for $r > 30\,\mathrm{Mpc/h}$. In the latter case,

the sample volumes were found to be too small to obtain statistically stable result due to wild fluctuations [6,7]. Therefore, although there are unambiguous evidences for the inhomogeneity of the galaxy distribution at least up to scales of 100 Mpc/h [6,7,9,10], the scaling properties at scales larger than 30 Mpc/h were poorly understood.

The new galaxy samples from the data release 7 (DR7 —[12]) doubled in size since the DR6 sample. This new catalog is large enough to facilitate the study of fluctuations in the galaxy distribution. In particular, we calculate the galaxy density in a sphere of radius r around each galaxy, i.e., the conditional density. For uniformly positioned galaxies [11], the average conditional density is independent of the radius r, and the fluctuations over galaxies are Gaussian. Conversely, in DR7 we find that the average density depends logarithmically on r, while the fluctuations follow the Gumbel distribution of extreme-value statistics. This behavior has an analog in statistical physics, where logarithmically changing averages tend to correspond to Gumbel-type fluctuations [13].

The rest of the paper is organized as follows. We first discuss the quantities we consider in the measurements and briefly discuss the main properties of the Gumbel distribution. We then introduce the galaxy samples and our main results on the average, the variance and the

---

[1]We use $H_0 = 100\,\mathrm{h\,km/s/Mpc}$, with $0.4 \leq h \leq 0.7$, for Hubble's constant.

fluctuation distribution of the conditional density we measured in the data. Finally, we discuss the results and draw conclusions.

Statistical methods. – In this section we describe the estimators we use in the analysis and then discuss the properties of the Gumbel distribution. We also provide some physical examples where the Gumbel distribution was found fit to experimental data.

Estimators and their main properties. A particularly useful characterization of statistical properties of point distributions can be obtained by measuring conditional quantities [11]. In this paper we focus on such a quantity, namely we calculate the number $N(r)$ of galaxies contained in a sphere of radius r centered on a galaxy. Note that not all galaxies can be considered as sphere centers for a given radius r: a central galaxy has to be farther than distance r from any border of the sample, so that the sphere volume is fully contained inside the sample volume [7,10]. As r approaches the radius of the largest sphere fully contained in the sample volume, the statistics become poorer. To deal with these limitations for large values of r, two effects should be taken into account: i) the number of points $N_{tot}(r)$ satisfying the above condition is largely reduced and ii) most of the points are located in the same region of the sample. Any conclusion about statistical properties must consider a careful analysis of these limitations [7].

The Gumbel distribution. The Gumbel (also known as Fisher-Tippet-Gumbel) distribution is one of the three extreme-value distribution [14,15]. It describes the distribution of the largest values of a random variable from a density function with faster than algebraic (say exponential) decay. The Gumbel distribution's PDF is given by

$$P(w) = \frac{1}{\beta} \exp\left[-\frac{w-\alpha}{\beta} - \exp\left(-\frac{w-\alpha}{\beta}\right)\right]. \quad (1)$$

With the scaling variable

$$x = \frac{w-\alpha}{\beta} \quad (2)$$

the density function (eq. (1)) simplifies to the parameter-free Gumbel

$$P(x) = e^{-x-e^{-x}} \quad (3)$$

with (cumulative) distribution $e^{-e^{-x}}$. Note that this distribution corresponds to maxima values, while for minima values, x is used instead of $-x$ in the Gumbel distribution.

The mean and the standard deviation (variance) of the Gumbel distribution (eq. (1)) is

$$\mu = \alpha + \gamma\beta, \qquad \sigma^2 = (\beta\pi)^2/6, \quad (4)$$

where $\gamma = 0.5772\ldots$ is the Euler constant. For the scaled Gumbel (eq. (3)) the first two cumulants of eq. (4) simplify to γ and $\pi^2/6$.

Gumbel in critical systems. Away from criticality, any global (spatially averaged) observable of a macroscopic system has Gaussian fluctuations, in agreement with the central-limit theorem (CLT). At criticality, however, the correlation length tends to infinity, and the CLT no longer applies. Indeed, fluctuations of global quantities in critical systems usually have non-Gaussian fluctuations. The type of fluctuations is characteristic to the universality class of the system's critical behavior [16,17].

To fit experimental data, the generalized Gumbel PDF $P(x) = C(e^{-x-e^{-x}})^a$ has often been used, where $a > 0$ is a real parameter, and $C = a^a/\Gamma(a)$ is a normalization constant. For integer values of a, this distribution corresponds to the a-th maximal value of a random variable. The $a = 1$ case corresponds to the Gumbel distribution. Experimental examples for Gumbel or generalized Gumbel distributions include power consumption of a turbulent flow [18], roughness of voltage fluctuations in a resistor (original Gumbel $a = 1$ case) [19], plasma density fluctuations in a tokamak [20], orientation fluctuations in a liquid crystal [21], and other systems cited in [13]. The Gumbel distribution describing fluctuations of a global observable was first obtained analytically in [19] for the roughness fluctuations of $1/f$ noise. Its relations to extreme-value statistics have been clarified [22,23], generalizations have appeared [24], and related finite-size corrections have been understood [25].

In a recent paper Bramwell [13] conjectured that only three types of distributions appear to describe fluctuations of global observables at criticality. In particular, when the global observable depends logarithmically on the system size, the corresponding distribution should be a (generalized) Gumbel. For example the mean roughness of $1/f$ signals depends on the logarithm of the observation time (system size), and the corresponding PDF is indeed the Gumbel distribution [19].

The data. – We have constructed several sub-samples of the main-galaxy (MG) sample of the spectroscopic catalog SDSS-DR7[2]. We have constrained the flags indicating the type of object to select only the galaxies from the MG sample. We then consider galaxies in the redshift range $10^{-4} \leqslant z \leqslant 0.3$ with redshift confidence $z_{conf} \geqslant 0.35$ and with flags indicating no significant redshift determination errors. In addition we apply the apparent magnitude filtering condition $m_r < 17.77$ [26]. The angular region we consider is limited, in the SDSS internal angular coordinates, by $-33.5° \leqslant \eta \leqslant 36.0°$ and $-48.0° \leqslant \lambda \leqslant 51.5°$: the resulting solid angle is $\Omega = 1.85$ steradians. We do not use corrections for the redshift completeness mask or for fiber collision effects. Fiber collisions in general do not present a problem for measurements of large-scale galaxy correlations [26]. Completeness varies most near the current survey edges, which are excluded in our samples. In addition the completeness mask could be the main source of systematic effects on small scale only, while we are

---

[2]http://www.sdss.org/dr7

interested on the correlation properties on relatively large separations [8].

To construct volume-limited (VL) samples we computed the metric distances R using the standard cosmological parameters, *i.e.*, $\Omega_M = 0.3$ and $\Omega_\Lambda = 0.7$. We computed absolute magnitudes M_r using Petrosian apparent magnitudes in the m_r filter corrected for Galactic absorption. We considered the sample limited by $R \in [70, 450]\,\text{Mpc/h}$ and $M_r \in [-21.8, -20.8]$ containing $M = 93821$ galaxies. In this sample there are about $1/5$ of the whole galaxies in DR7; it has relatively large spatial extensions and small spread in galaxy luminosity. Note that in other samples limited at scales smaller than $\sim 400\,\text{Mpc/h}$ we found similar results.

We have checked that our main results in this VL sample do not depend significantly on K-corrections and/or evolutionary corrections as those used by [27]. In this paper we use standard K-correction from the VAGC data[3] (see discussion in [7] for more details).

Results. – In this section we present our findings from the analysis of the galaxy data.

We have computed the number of galaxies $N_i(r)$ within radius r around each galaxy i satisfying the boundary condition previously mentioned. By normalizing it by the volume $V_r = 4\pi r^3/3$ of the sphere, we obtain the conditional galaxy density $n_i(r) = N_i(r)/V_r$ around each galaxy i. This quantity is our main interest in this paper. The variable $n_i(r)$ differs for each galaxy, hence we consider this local density $n_i(r)$ as a random variable, and study its statistical properties. For example the conditional average density within radius r is defined as

$$\overline{n(r)} = \frac{1}{N_{\text{tot}}(r)} \sum_{i=1}^{N_{\text{tot}}(r)} n_i(r), \qquad (5)$$

where $\overline{n(r)}$ is "conditioned" on the presence of the central galaxy. Here $N_{\text{tot}}(r)$ is the total number of galaxies in the survey —counting only those ones which are farther from the sample borders than distance r [7]. The simplest quantity to characterize density fluctuations is the variance, or mean square deviation at scale r; its estimator is defined as

$$\sigma^2(r) \equiv \text{var}\,[n(r)] = \frac{1}{N_{\text{tot}}(r)} \sum_{i=1}^{N_{\text{tot}}(r)} n_i^2(r) - \overline{n(r)}^2. \qquad (6)$$

In the following subsections we are going to study the whole distribution of $n_i(r)$ as well.

Self-averaging properties. Conditional fluctuations have been found to be not self-averaging in several SDSS-DR6 samples, *i.e.,* there were systematic differences between statistical properties measured in different parts of a given sample [6,9]. It was concluded that this behavior is due to galaxy density fluctuations which are
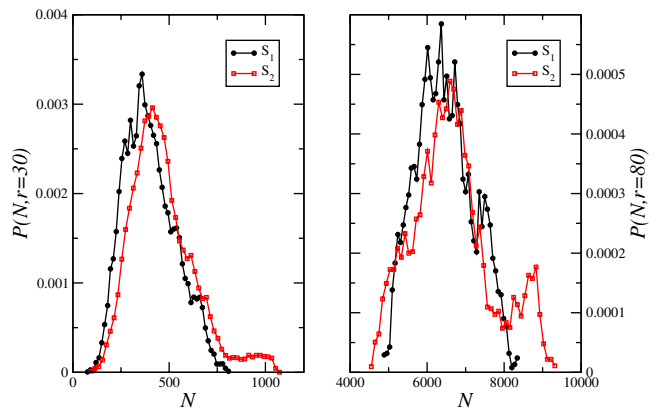
Fig. 1: (Colour on-line) PDFs of the conditional density in spheres of radius $r = 30\,\text{Mpc/h}$ (left) and $r = 80\,\text{Mpc/h}$ (right), in two distinct regions: a nearby (S_1) and a far away (S_2) one. Notice that the two PDFs statistically give the same signal.

too large in amplitude and too extended in space to be self-averaging inside the considered volumes. The lack of self averaging prevents one to extract a statistically meaningful information from whole sample average quantities, as for example the conditional average density. We repeated the stability test of statistical quantities within the new SDSS-DR7 sample, since it almost doubled in size compared to SDSS-DR6. To this aim we cut the sample volume into two regions, a nearby and a faraway one as in [6,7], and we determine the PDFs $P(n(r)) \equiv P(n; r)$ of the conditional density separately in both regions, and at two different r scales. We conclude from fig. 1 that the PDF is statistically stable and does not show systematic dependence on system size, as opposed to the case of the SDSS-DR6 data on scales $r > 30\,\text{Mpc/h}$ [6,7]. Hence in this new sample, conditional statistical quantities computed over the whole sample volume are useful and meaningful indicators.

Scaling at small scales. At small length scales ($r < 20\,\text{Mpc/h}$) the exponent for the conditional average density is close to minus one (see fig. 2). This result is in agreement with ones obtained by the same method in a number of different samples (see [6,7,9,10] and references therein). This scaling can be interpreted as a signature of fractality of the galaxy distribution in this range of scale. In addition, this implies that the distribution is not uniform at these scales, and thus the standard two-point correlation function is substantially biased.

Scaling at large scales. We first computed the average conditional density (eq. (5)) at large scales ($r > 10\,\text{Mpc/h}$). For a uniform point distribution this quantity is constant, *i.e.*, independent of the radius r [11]. Conversely, in our data we find a pronounced r dependence, as can be seen in fig. 2. Our best fit is

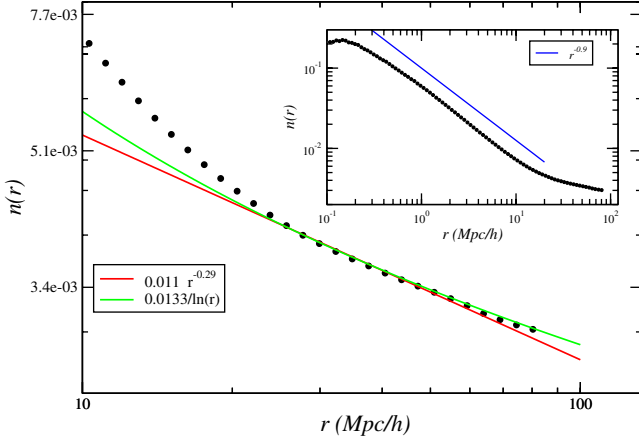$$\overline{n(r)} \approx \frac{0.0133}{\log r}, \qquad (7)$$

Fig. 2: (Colour on-line) Conditional average density $\overline{n}(r)$ of galaxies as a function of radius. In the inset panel the same is shown in the full range of scales. Note the change of slope at $\approx 20\,\mathrm{Mpc/h}$ and also the lack of flattening up to $\approx 80\,\mathrm{Mpc/h}$. Our conjecture is that we have a logarithmic correction to the constant behavior, although we cannot exclude the possibility that it is power law with an exponent ≈ -0.3.

that is the average density depends only weakly (logarithmically) on r. Alternatively, an almost indistinguishable power law fit is provided by

$$\overline{n(r)} \approx 0.011 \times r^{-0.29}. \tag{8}$$

We emphasize our preference for the logarithmic fit, where the only fitting parameter is the amplitude. Note that the logarithmic fit should in principle be written as $A/\log\left(r/r_0\right)$, where A is the amplitude and r_0 is a reference length scale. We avoid introducing parameter r_0, since the fit with eq. (7) is already quite good. As can be seen in fig. 2, we find a change of slope in the conditional average density in terms of the radius r at about $\approx 20\,\mathrm{Mpc/h}$. At this point the decay of the density changes from an inverse linear decay to a slow logarithmic one. Moreover, the density $\overline{n(r)}$ does not saturate to a constant up to $\sim 80\,\mathrm{Mpc/h}$, *i.e.*, up to the largest scales probed in this sample. Note that up to $r = 80\,\mathrm{Mpc/h}$ the number of points $N_{\mathrm{tot}}(r)$ is larger than 10^4, making this statistics very robust.

This result is in agreement with a study of the SDSS-DR4 samples [28], where, in the average conditional density, a similar change of slope was observed at about the same scale $r \approx 20\,\mathrm{Mpc/h}$, together with quite large sample to sample fluctuations. Indeed, some evidences were subsequently found to support that the galaxy distribution is still characterized by rather large fluctuations up to $100\,\mathrm{Mpc/h}$, making it incompatible with uniformity [6–10]. Similarly, in the Luminous Red Galaxy (LRG) sample of SDSS, Hogg *et al.* [29] also found a slope change in the average conditional density. On the other hand, we do not observe a transition to uniformity at about $70\,\mathrm{Mpc/h}$, which they reported. Note also that a study
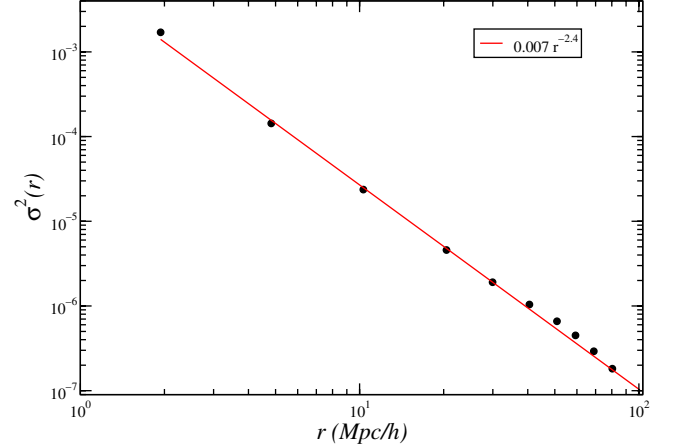

Fig. 3: (Colour on-line) Variance σ^2 of the conditional density $n_i(r)$ as a function of the radius. Conversely, the corresponding variance of a Poisson point process would display a $1/r^3$ decay.
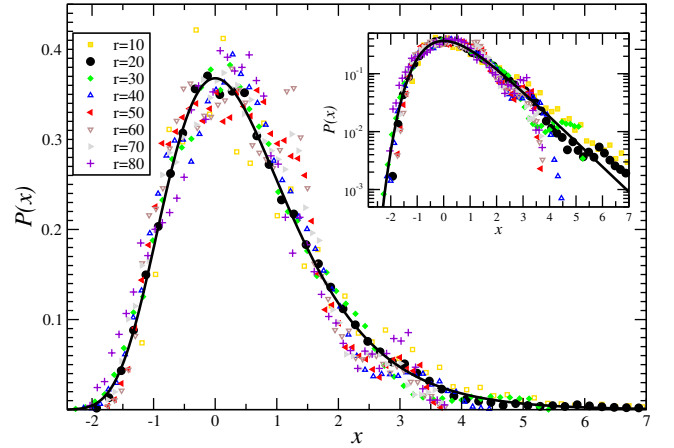

Fig. 4: (Colour on-line) Data curves of different r scaled together by fitting parameters α and β for each curves. The solid line is the parameter-free Gumbel distribution, eq. (3). In the inset panel the same is shown on log-linear scale to emphasize the tails of the distribution.

of the self-averaging properties of fluctuations in the LRG sample is still lacking.

Compared to the average density, it is harder to find the correct fit for the variance $\sigma^2(r)$ of the conditional density (eq. (6)). Our best fit is (see fig. 3)

$$\sigma^2(r) \approx 0.007 \times r^{-2.4}. \tag{9}$$

Given the scaling behavior of the conditional density and variance, we conclude that galaxy structures are characterized by non-trivial correlations for scales up to $r \approx 80\,\mathrm{Mpc/h}$.

To probe the whole distribution of the conditional density $n_i(r)$, we fitted the Gumbel distribution (eq. (1)) via its two parameters α and β. One of our best fits is obtained for $r = 20\,\mathrm{Mpc/h}$, see fig. 4. The data, moreover, convincingly collapses to the parameter-free Gumbel distribution (eq. (3)) for all values of r for

$10 \leqslant r \leqslant 80 \, \text{Mpc/h}$, with the use of the scaling variable x from eq. (2) (see fig. 4). Note that for a Poisson point process (uncorrelated random points) the number $N(r)$ (and consequently also the density) fluctuations are distributed exactly according to a Poisson distribution, which in turn converges to a Gaussian distribution for large average number of points $\overline{N(r)}$ per sphere. In our samples, $\overline{N(r)}$ is always larger than 20 galaxies, where the Poisson and the Gaussian PDFs differ less than the uncertainty in our data. Note also that due to the central-limit theorem, all homogeneous point distributions (not only the Poisson process) lead to Gaussian fluctuations. Hence the appearance of the Gumbel distribution is a clear sign of inhomogeneity and large-scale structures in our samples.

The fitting parameters in eq. (1) varied with the radius r approximately as

$$\alpha \approx \frac{0.007}{r^{0.21}}, \qquad \beta \approx \frac{0.035}{r}, \qquad (10)$$

although a logarithmic fit $\alpha \approx 0.0115/\log r$ cannot be excluded either. With the fitted values of α and β we recover the (directly measured) average conditional density of galaxies through eq. (4). On the other hand, we have a discrepancy when comparing the directly measured σ^2 to that obtained from the Gumbel fits through eq. (4). The reason for this discrepancy is that the uncertainty in the tail of the PDF $P(n, r)$ is amplified when we directly calculate the second moment.

Data collapse without fitting. It is possible to obtain a scaling of the data without any fitting procedure. We can compute the average, μ, and the standard deviation, σ^2, of the data and use the scaled variable

$$y = \frac{N - \mu}{\sigma}. \qquad (11)$$

The density functions for different values of r scale to the single curve

$$\Phi(y) = a e^{-(ay+\gamma) - e^{-(ax+\gamma)}} \qquad (12)$$

with $a = \pi/\sqrt{6}$. (This function, of course, has mean zero, and standard deviation one.) This type of fitting-free data collapse has been used extensively in statistical physics [16,19]. As shown in fig. 5 we find a satisfactory agreement with eq. (12). Note also that Gaussian fluctuations can be clearly excluded. Compared to the fitting results of fig. 4, the agreement in fig. 5 is better around the tails of the distribution, but it gets worse around the maximum. The reason for this latter mismatch is again due to the uncertainty in the second moment.

Discussion. – Given the observed scaling and data collapse in the spatial galaxy data, is there any supporting evidence for the appearance of the Gumbel distribution? Due to the scaling and data collapse we argue that the large-scale galaxy distribution shows similarities with critical systems. Here the galaxy density around each
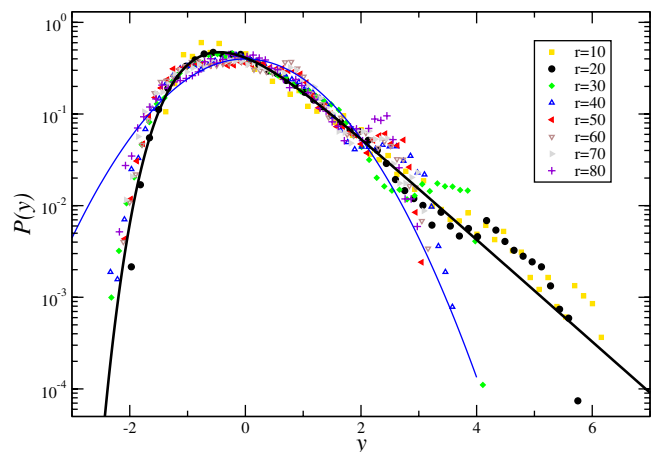

Fig. 5: (Colour on-line) The fitting-free data collapse (eqs. (11), (12)) based on the first two moments of the distribution. Note again the satisfactory data collapse on all scales. The black line is the Gumbel distribution of eq. (12), while the blue line is the corresponding normal Gauss distribution with zero mean and unit variance.

galaxy is analogous to a random variable describing a spatially averaged quantity in a volume. The average conditional galaxy density depends on the volume size ($\sim r^3$) only logarithmically $\overline{n(r)} \sim 1/\log r$ from eq. (7). According to the conjecture of Bramwell for critical systems [13], if a spatially averaged quantity depends only weakly (say logarithmically) on the system size, the distribution of this quantity follows the Gumbel distribution. This is indeed what we see in the galaxy data. Hence our two observations about the average density and the density distribution are compatible with the behavior of critical systems in statistical physics.

We note that standard models of galaxy formation predict homogeneous mass distribution beyond $\approx 10 \, \text{Mpc/h}$ [6,7,30]. To explain our findings about non-Gaussian fluctuations up to much larger scales presents a challenge for future theoretical galaxy formation models (see [6,7,9,10,30] for more details).

In summary, we have established scaling and data collapse over a wide range of radius (volume) in galaxy data. Scaling in the data indicates criticality. The average galaxy density depends only logarithmically on the radius, which suggests a Gumbel scaling function [13]. The scaled data is indeed remarkably close to the Gumbel distribution, which is one of the three extreme-value distributions. How this distribution arises through galaxy formation, or what the extreme quantity is in the galaxy data, are challenging questions needed to be addressed in the future.

$$* * *$$

REFERENCES

[1] York D. *et al.*, *Astron. J.*, **120** (2000) 1579.
[2] Colless M. *et al.*, *Mon. Not. R. Astron. Soc.*, **328** (2001) 1039.
[3] Busswell G. S. *et al.*, *Mon. Not. R. Astron. Soc.*, **354** (2004) 991.
[4] Gott J. R. III, Jurić M., Schlegel D., Hoyle F., Vogeley M., Tegmark M., Bahcall N. and Brinkmann J., *Astrophys. J.*, **624** (2005) 463.
[5] Hoyle F. and Vogeley M., *Astrophys. J.*, **607** (2004) 751.
[6] Sylos Labini F., Vasilyev N. L., Baryshev Yu. V. and Pietronero L., *EPL*, **86** (2009) 49001.
[7] Sylos Labini F., Vasilyev N.L. and Baryshev Yu. V., to be published in *Astron. Astrophys.* (2009) DOI: 10.1051/0004-6361/200811565.
[8] Sylos Labini F., Vasilyev N. L., Baryshev Yu. V. and López-Corredoira M., *Astron. Astrophys.*, **505** (2009) 981.
[9] Sylos Labini F., Vasilyev N. L. and Baryshev Yu. V., *EPL*, **85** (2009) 29002.
[10] Sylos Labini F., Vasilyev N. L. and Baryshev Yu. V., *Astron. Astrophys.*, **496** (2009) 7.
[11] Gabrielli A., Sylos Labini F., Joyce M. and Pietronero L., *Statistical Physics for Cosmic Structures* (Springer Verlag, Berlin) 2004.
[12] Abazajian K. *et al.*, *Astrophys. J. Suppl.*, **182** (2009) 543.
[13] Bramwell S. T., *Nat. Phys.*, **5** (2009) 444.
[14] Fisher R. A. and Tippett L. H. C., *Cambridge Philos. Soc.*, **28** (1928) 180.
[15] Gumbel E. J., *Statistics of Extremes* (Columbia University Press) 1958.
[16] Foltin G., Oerding K., Rácz Z., Workman R. L. and Zia R. K. P., *Phys. Rev. E*, **50** (1994) R639.
[17] Rácz Z., in *Slow Relaxations and Nonequilibrium Dynamics in Condensed Matter*, edited by Barrat J.-L. *et al.* (Springer-Verlag) 2002, cond-mat/0210435.
[18] Bramwell S. T., *Nature*, **396** (1998) 552.
[19] Antal T., Droz M., Györgyi G. and Rácz Z., *Phys. Rev. Lett.*, **87** (2001) 240601.
[20] Van Milligen B. P. *et al.*, *Phys. Plasmas*, **12** (2005) 052507.
[21] Joubaud S. *et al.*, *Phys. Rev. Lett.*, **100** (2008) 180601.
[22] Bertin E., *Phys. Rev. Lett.*, **95** (2005) 170601.
[23] Bertin E. and Clusel M., *J. Phys. A*, **39** (2006) 7607.
[24] Antal T., Droz M., Györgyi G. and Rácz Z., *Phys. Rev. E*, **65** (2002) 046140.
[25] Györgyi G., Moloney N. R., Ozogány K. and Rácz Z., *Phys. Rev. Lett.*, **100** (2008) 210601.
[26] Strauss M. A. *et al.*, *Astron. J.*, **124** (2000) 1810.
[27] Blanton M. R. *et al.*, *Astrophys. J.*, **592** (2003) 819.
[28] Sylos Labini F., Vasilyev N. L. and Baryshev Y., *Astron. Astrophys.*, **465** (2007) 23.
[29] Hogg D. W., Eisenstein D. J. and Blanton M. R. *et al.*, *Astrophys. J.*, **624** (2005) 54.
[30] Sylos Labini F. and Vasilyev N. L., *Astron. Astrophys.*, **477** (2008) 381.