# Growing inequality in systems showing Zipf's law

Journal of Physics: Complexity

PAPER

# Growing inequality in systems showing Zipf's law

Giordano De Marzo[1,2,3,6,*], Federico Attili[4] and Luciano Pietronero[1,2,5]

[1] Centro Ricerche Enrico Fermi, Piazza del Viminale, 1, I-00184 Rome, Italy
[2] Dipartimento di Fisica Università 'Sapienza', P.le A. Moro, 2, I-00185 Rome, Italy
[3] Sapienza School for Advanced Studies, 'Sapienza', P.le A. Moro, 2, I-00185 Rome, Italy
[4] Dipartimento di Scienze Economiche Università di Bologna, Piazza Scaravilli, 2, 40126 Bologna, Italy
[5] Istituto dei Sistemi Complessi (ISC)—CNR, UoS Sapienza, P.le A. Moro, 2, I-00185 Rome, Italy
[6] Complexity Science Hub Vienna, Josefstaedter Strasse 39, Vienna 1080, Austria
[*] Author to whom any correspondence should be addressed.

**E-mail:** giordano.demarzo@gmail.com

**Keywords:** Zipf's law, Gini index, power laws, inequality

OPEN ACCESS

RECEIVED
8 September 2022

REVISED
11 January 2023

ACCEPTED FOR PUBLICATION
2 March 2023

PUBLISHED
20 March 2023



## Abstract

A central problem in economics and statistics is the assessment of income or wealth inequality starting from empirical data. Here we focus on the behavior of Gini index, one of the most used inequality measures, in presence of Zipf's law, a situation which occurs in many complex financial and economical systems. First, we show that the application of asymptotic formulas to finite size systems always leads to an overestimation of inequality. We thus compute finite size corrections and we show that depending on Zipf's exponent two distinct regimes can be observed: low inequality, where Gini index is less than one and maximal inequality, where Gini index asymptotically tends to its maximal value one. In both cases, the inequality of an expanding system slowly increases just as effect of growth, with a scaling never faster than the inverse of the size. We test our computations on two real systems, US cities and the cryptocurrency market, observing in both cases an increase of inequality that is completely explained by Zipf's law and the systems expanding. This shows that in growing complex systems finite size effects must be considered in order to properly assess if inequality is increasing due to natural growth processes or if it is produced by a change in the economical structure of the systems. Finally we discuss how such effects must be carefully considered when analyzing survey data.

## 1. Introduction

Zipf's law is a characteristic feature of complex systems and is often considered as a footprint of complexity [1–3]. Such scaling law is observed in an astonishing number of natural and social systems characterized by emerging complex behaviors, such as solar flares, cosmic structures, earthquakes, language, urban systems and many more [2, 4–6]. Denoting by $S(k)$ the size of the $k$th largest element in the system, Zipf's law reads

$$S(k) = \frac{S(1)}{k^{\gamma}}. \tag{1}$$

Here $k$ is called rank, $\gamma$ is the Zipf exponent, while $S(1)$ is the size of the rank one object. For instance, having in mind an urban system, its elements would be the cities composing it and as size one can use different measures ranging from population to light emission [7]. Also most financial and economical systems are generally described by Zipf's scaling, which is found for instance in the distribution of returns, of stock prices, of the cryptocurrency market and of people' wealth [5, 8–12], this last being the first system where power law distributions were observed [13]. This explains a number of very counter-intuitive properties typical of financial and economics data. Examples are the uneven distribution of digits in stock prices [11], 80–20 rule or the occurrence of major financial crisis [10, 14].

Among the nontrivial effects of power law distributions and Zipf's law, the peculiar behavior of Gini index recently discussed in [15, 16] is of particular relevance. Gini index is a well know indicator used to quantify wealth inequality [17–19] and more generally to measure how resources such as money or

population are unevenly distributed in a system [20, 21]. Gini index and more generally inequality is strictly connected to Zipf's law and power laws, as noted for instance in [22] or [23]. These works interpret Zipf's law as the result of an optimization process, with the former suggesting a balance between the efficiency of the system and the inequality of its components. As a consequence, given the ubiquity of Zipf's law in economical systems, understanding how such a scaling law affect inequality measures is a crucial and non trivial point [24]. Indeed, as shown in [15, 16], a naive approach applied on finite mean but infinite variance distributions, results in a biased estimation of Gini index, which turns out to be the more under estimated the smaller is the sample considered. Since power law tails provides a non marginal contribution to inequality [25, 26], such result questions the reliability of wealth inequality studies and further stress the importance of studying the properties of Gini index computed on power law distributed data. However, despite such evidences and the vast literature on Gini index (see for instance [24, 27, 28]), only few steps have been moved in this direction.

Motivated by what just discussed, in the present paper we tackle the problem of deriving analytical expressions of Gini index for finite size systems showing Zipf's law. In particular we determine how Gini index varies as function of the system size $N$ and of Zipf's exponent $\gamma$. We find that for all values of the exponent, Gini index is increasing in $N$, but the convergence to the asymptotic value can be very slow. Such a result implies that a growing system characterized by a Pareto distribution is expected to increase its inequality just as effect of its expansion and it is particularly important since many complex systems shows such a behavior [1, 4]. Depending on the value of $\gamma$ different regimes can be found: for $\gamma < 1$ Gini index asymptotically tend to a finite value and the system is in a low inequality regime, while if $\gamma > 1$ Gini index tends to its maximal value and the system is in the maximal inequality regime. We provide as a practical example of the latter the cryptocurrency market, which in the last years has experienced a huge growth both in terms of the number of different cryptocurrencies and in its total market capitalization [9]. As expected we observe an increase of Gini index over time, which nicely follows the analytical expression we derived. Similar results are obtained considering the evolution of US cities, a system characterized by being in the low inequality regime. In both cases we conclude that the increase of inequality observed is not due to a change of the economical condition of the systems, but instead it is just the result of their expansion. This implies that finite size effects must be carefully taken into account when assessing the effects of policies aimed at reducing inequality. Finally, we shortly discuss how the approach here proposed can be used for estimating the inequality of a system showing Zipf's law when only a random sample of it is available, a situation typically occurring during surveys.

## 2. Results

The quantification of income or wealth inequality [20, 21] is a central problem in economic and statistics. An index of inequality provides a measure of how much richness is uniformly or unevenly distributed among individuals and is therefore minimized by a delta like distribution. Among the many proposed measures [18, 29], the most used one is Gini index [17]; given a set of $N$ incomes (or any other measure of size) $s_1, s_2, \ldots, s_N$ it is defined as

$$G = \frac{\sum_{i,j}^{N} |s_i - s_j|}{2N \sum_i s_i}. \tag{2}$$

When all the incomes are equal there is no inequality among the individuals and Gini index is null, while maximal inequality corresponds to $G = 1$. Particularly interesting is the case in which sizes satisfy Zipf's law equation (1), as it occurs for cities, wealth, stock prices and many other socio-economial systems. In this case also the probability distribution of the sizes is power law like $P(s) \sim s^{-\alpha}$ and its exponent $\alpha$ is related to Zipf's exponent $\gamma$ by the relation $\gamma = 1/(\alpha - 1)$ [4, 6]. We can recast equation (2) as

$$G = \frac{\sum_{i,j}^{N} |s_i - s_j| \int \mathrm{d}x \delta(x - s_i) \int \mathrm{d}y \delta(y - s_j)}{2N \sum_i^N s_i \int \mathrm{d}x \delta(x - s_i)} = N^2 \frac{\int \mathrm{d}x \int \mathrm{d}y |x - y| \frac{1}{N} \sum_i^N \delta(x - s_i) \frac{1}{N} \sum_j^N \delta(y - s_j)}{2N^2 \int \mathrm{d}x x \frac{1}{N} \sum_i^N \delta(x - s_i)}$$

where the integrals are done over $[s_{\min}, \infty]$, $s_{\min}$ being the lower cutoff of the power law distribution. By taking the limit $N \to \infty$ we express Gini index in terms of the inherent probability distribution

$$G = \frac{\int \mathrm{d}x \int \mathrm{d}y |x - y| P(x) P(y)}{2N \int \mathrm{d}x x P(x)} = \frac{\int \mathrm{d}x \int \mathrm{d}y |x - y| P(x) P(y)}{2\mu},$$

where $\mu$ is the mean value of the distribution. For $\alpha > 2$ the mean value is finite and the integrals can be solved, giving

$$G_\infty = \frac{1}{2\alpha - 3} = \frac{\gamma}{2-\gamma}. \tag{3}$$

However all real systems have a finite size and consequently it is generally not possible to compute the Gini index by directly applying equation (3). This can be easily understood by noticing that since the system shows Zipf's law, it holds

$$S(N) = \frac{S(1)}{N^\gamma} \;\rightarrow\; S(1) = S(N) \cdot N^\gamma.$$

We can freely assume $S(N)$ to remain constant as $N$ increases and to be equal to the lower cutoff of the distribution $s_{\min}$ (or, equivalently, we can express all sizes in the units of $S(N)$ since power laws are scale invariant). This implies that for finite $N$ the largest element in the system has a size of order $N^\gamma$, while the integral is computed up to infinity, producing an overestimation of inequality.

Equation (3) can be corrected for taking into account finite size effect. First we notice that equation (2) can be written as

$$G = \frac{1}{N}\left[N + 1 - 2\frac{\sum_k (N+1-k)T(k)}{\sum T(k)}\right] = \frac{1}{N}\left[N + 1 - 2\frac{\sum_k (N+1-k)S(N+1-k)}{\sum S(k)}\right],$$

where $S(k)$ is the sequence of sizes in descending order, while $T(k)$ is the reversed sequence. This gives

$$G = \frac{1}{N}\left[N + 1 - 2\frac{\sum_k^N kS(k)}{\sum_k^N S(k)}\right]. \tag{4}$$

In the case we are considering $S(k)$ satisfies Zipf's law and is given by equation (1), thus we obtain

$$G = \frac{1}{N}\left[N + 1 - 2\frac{\sum_k^N k^{1-\gamma}}{\sum_k^N k^{-\gamma}}\right] = \frac{1}{N}\left[N + 1 - 2\frac{H_N^{(\gamma-1)}}{H_N^{(\gamma)}}\right],$$

where $H_n^{(g)}$ is the generalized harmonic number of order $g$ of $n$. By using the Euler–Maclaurin formula we can approximate the harmonic numbers as

$$H_N^{(\gamma)} = \sum_{k=1}^{N} \frac{1}{k^\gamma} = \zeta(\gamma) + \frac{1}{1-\gamma}N^{1-\gamma} + \frac{1}{2}N^{-\gamma} + O(N^{-\gamma-1}).$$

Here $\zeta(x)$ is Riemann Zeta function and using this result we get

$$G \approx \frac{1}{N}\left[N + 1 - 2\frac{\zeta(\gamma-1) + \frac{1}{2-\gamma}N^{2-\gamma} + \frac{1}{2}N^{1-\gamma}}{\zeta(\gamma) + \frac{1}{1-\gamma}N^{1-\gamma}}\right]. \tag{5}$$

Depending on the value of $\gamma$, the fraction is dominated by the Zeta functions or by the terms in $N$, with $\gamma = 1$ dividing these two cases.

The case $\gamma = 1$, corresponding to the classical Zipf's exponent, must be considered separately. For this value of $\gamma$ the harmonic number of $N$ can be approximated as

$$H_N^{(\gamma)} \equiv H_N \approx \eta + \log N,$$

where $\eta \approx 0.58$ is the Euler–Mascheroni constant. This gives

$$G_{\gamma=1} \approx \frac{1}{N}\left[N + 1 - 2\frac{N}{\eta + \log N}\right],$$

that is

$$G_{\gamma=1} \approx 1 - \frac{2}{\log N} + \frac{2\eta}{(\log N)^2}. \tag{6}$$

This expression can thus be exploited in those systems showing an exact Zipf's law with unitary exponents.

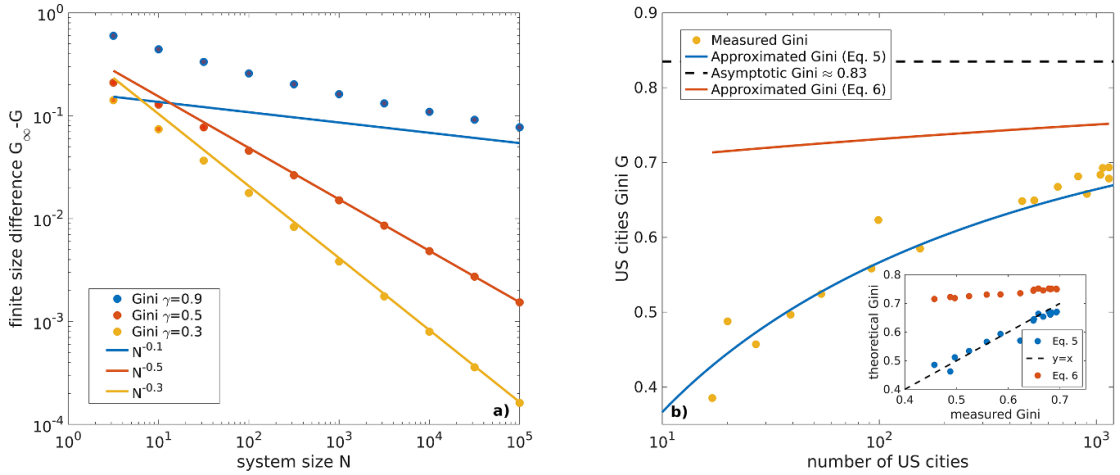


**Figure 1.** Low inequality regime. (a) Dots show the finite size difference between Gini index and its asymptotic value $\gamma/(2-\gamma)$ as function of $N$ for different values of Zipf exponent $\gamma$. The solid line represent the scaling $N^{1-\gamma}$ with the prefactor given by equation (7). As it is possible to see this scaling well fits the finite size difference, but for $\gamma$ close to one there are relevant discrepancies if $N$ is small. (b) Gini index of US cities from 1790 to 1950 as function of the number of cities $N$ (yellow dots), compared with the theoretical values given by equation (5) (blue line) and equation (7) (red line). The black dashed line shows the asymptotic value $G_\infty \approx 0.83$.

### 2.1. Low inequality regime: $\gamma < 1$

For $\gamma < 1$ we know that in the asymptotic limit $N \to \infty$ Gini index is given by equation (3) and is smaller than one. Since the system does not asymptotically tend to maximal inequality, we call $\gamma < 1$ low inequality regime. Let us focus on the last term of equation (5), we can write it as

$$\left[\zeta(\gamma-1)+\frac{1}{2-\gamma}N^{2-\gamma}+\frac{1}{2}N^{1-\gamma}\right]\frac{2(1-\gamma)}{N^{1-\gamma}\left[1+\zeta(\gamma)(1-\gamma)N^{\gamma-1}\right]}$$
$$\approx \frac{2(1-\gamma)}{N^{1-\gamma}}\left[\zeta(\gamma-1)+\frac{1}{2-\gamma}N^{2-\gamma}+\frac{1}{2}N^{1-\gamma}\right]\left[1-\zeta(\gamma)(1-\gamma)N^{\gamma-1}\right],$$

where we expanded for small $N^{\gamma-1}$. Retaining only the three largest contributions this gives

$$G \approx \frac{1}{N}\left\{N+1-\frac{2(1-\gamma)}{N^{1-\gamma}}\left[\frac{1}{2-\gamma}N^{2-\gamma}-\zeta(\gamma)\frac{1-\gamma}{(2-\gamma)}N+\frac{1}{2}N^{1-\gamma}\right]\right\}$$

that is

$$G \approx 1-2\frac{1-\gamma}{2-\gamma}+2\frac{\zeta(\gamma)(1-\gamma)^2}{(2-\gamma)}\frac{1}{N^{1-\gamma}}+\frac{\gamma}{N}=G_\infty+2\frac{\zeta(\gamma)(1-\gamma)^2}{(2-\gamma)}\frac{1}{N^{1-\gamma}}+\frac{\gamma}{N}. \tag{7}$$

Since Riemann Zeta function $\zeta(\gamma)$ is negative for $\gamma < 1$, this result implies that the asymptotic value $G_\infty$ given by equation (3) is approached by below with the scaling

$$G - G_\infty \sim -N^{\gamma-1} \tag{8}$$

and as a consequence a system showing Zipf's law increases its inequality just as a result of its expansion. Note that the condition $0 < \gamma < 1$ on Zipf's exponent implies that the exponent $\alpha$ of the inherent power law probability distribution must satisfy $\alpha > 2$. The scaling given by equation (8) coincide [30] with that derived in [16] under the hypothesis of infinite variance, that is $2 < \alpha < 3$. Our computation show that the same scaling is found also when the variance is finite.

A comparison between the approximation given by equation (7) and a direct computation of Gini index is reported in figure 1(a). More precisely, using synthetic data following Zipf's law, we show how the difference between the Gini index of the system and the asymptotic Gini index given by equation (3) scales with the system size $N$. Dots correspond to the exact Gini index while solid lines to the scaling $N^{\gamma-1}$ given by equation (7). Our approximation rapidly converges to the exact value for small values of $\gamma$, but for $\gamma$ close to one large discrepancies are observed up to large values of $N$. We also test the formulas we derived on a real system by studying how the inequality between US cities evolved from 1790 to 1950. This system is one of the first described in terms of Zipf's law, with studies dating back to George Kingsley Zipf himself [31]. Since historical GDP data are not available, we use population as a proxy of GDP and we exploit Gini index to

measure how unevenly distributed is richness in the US urban system. In this case $S(k)$ thus coincides with the population of the $k$th largest US city, while $N$ is the number of different urban areas. Note that we only considered cities in the power law tail of the distribution, see the section 4 for more details about the procedure and the data used. Figure 1(b) shows the evolution of Gini index as function of $N$ (yellow dots), as it is possible to see inequality has been increasing since the birth of the US. By using a maximum likelihood (ML) approach we estimated the average Zipf's exponent that turns out to be $\gamma \approx 0.91$, thus this system is in the low inequality regime. As expected the approximation of Gini index given by equation (7) is very unreliable, since Zipf's exponent is close to one and the sample size is small, while equation (5) perfectly describe the data. This implies that the increase of inequality observed can be totally explained as the result of the US urban system expanding and is not due to any specific economic reason. We also show in the inset a direct comparison between the measured Gini index and the theoretical predictions computed using the two different approximations. The results we obtained have strong implications, first they prove that using the asymptotic formula provided by equation (3) can be completely misleading, since the inequality observed in the system we considered is way below that predicted asymptotically. Second, they show that expanding systems with an underlying power law distribution present non negligible variation of inequality. This is particularly relevant since such variations must be taken into account when evaluating policies, whose beneficial impact over inequality could otherwise be obscured by the phenomenon we just discussed.

### 2.2. Maximal inequality regime: $\gamma > 1$

When $\gamma > 1$ the mean value of the distribution is diverging and so Gini index is not defined for an infinite system. However, as we are going to show, it can easily defined for a finite system and, by letting the size of such system go to infinity, it goes to one. We thus denote $\gamma > 1$ maximal inequality regime, since systems with such exponents expanding evolve toward a configuration where Gini index is maximal. Since $\gamma > 1$, the denominator appearing in equation (5) is dominated by the Zeta function and we can write the last term of this equation as

$$\left[\zeta(\gamma-1)+\frac{1}{2-\gamma}N^{2-\gamma}+\frac{1}{2}N^{1-\gamma}\right]\frac{2}{\zeta(\gamma)\left[1+\frac{1}{\zeta(\gamma)(1-\gamma)}N^{1-\gamma}\right]}$$
$$\approx \frac{2}{\zeta(\gamma)}\left[\zeta(\gamma-1)+\frac{1}{2-\gamma}N^{2-\gamma}+\frac{1}{2}N^{1-\gamma}\right]\left[1-\frac{1}{\zeta(\gamma)(1-\gamma)}N^{1-\gamma}\right].$$

Taking the two largest terms and plugging the result into equation (5) we obtain

$$G \approx \frac{1}{N}\left\{N+1-\frac{2}{\zeta(\gamma)}\left[\zeta(\gamma-1)+\frac{1}{2-\gamma}N^{2-\gamma}\right]\right\},$$

which gives

$$G \approx 1-\left[2\frac{\zeta(\gamma-1)}{\zeta(\gamma)}-1\right]N^{-1}-\frac{2}{\zeta(\gamma)(2-\gamma)}N^{1-\gamma} = G_\infty-\left[2\frac{\zeta(\gamma-1)}{\zeta(\gamma)}-1\right]N^{-1}-\frac{2}{\zeta(\gamma)(2-\gamma)}N^{1-\gamma}, \tag{9}$$

where $G_\infty$. Indeed as we mentioned above we see from this expression that Gini index tends to one in the limit of infinite size $N \to \infty$, thus even if the mean value is not defined in such a limit, Gini index is. Also note that depending on the value of $\gamma$ two different scalings toward $G_\infty$ are observed

$$G - G_\infty \sim \begin{cases} -N^{1-\gamma} \text{ for } 1 < \gamma < 2 \\ -N^{-1} \text{ for } \gamma > 2. \end{cases} \tag{10}$$

Also in this case we compare the theoretical prediction given by equation (9) with synthetic data following Zipf's law with exponent $\gamma > 1$. Results are reported in figure 2(a), where we plotted the difference between the Gini index of the system and the asymptotic Gini index $G_\infty = 1$ as function of the system size $N$. Dots correspond to the exact Gini index, solid lines to equation (9), while dashed lines are guides for eyes representing the scaling predicted by equation (10). As expected there are two distinct scaling regimes for $1 < \gamma < 2$ and $\gamma > 2$. Again we also test our analytical expressions on a real system, in this case we choose the cryptocurrency market. Such financial system has been found to follow Zipf's law with exponent $\gamma \approx 1.71$ and has been characterized by a very fast growth both in terms of different coins and total market cap [9]. In this case $S(k)$ is given by the market capitalization of the $k$th largest cryptocurrency, while $N$ is the number of different cryptocoins. We considered all cryptocurrencies over one billion dollars of market capitalization, see the Methods for more details about the procedure and the data used. Figure 2(b) shows the evolution of Gini index as function of $N$ (yellow dots), as it is possible to see inequality is converging to the asymptotic

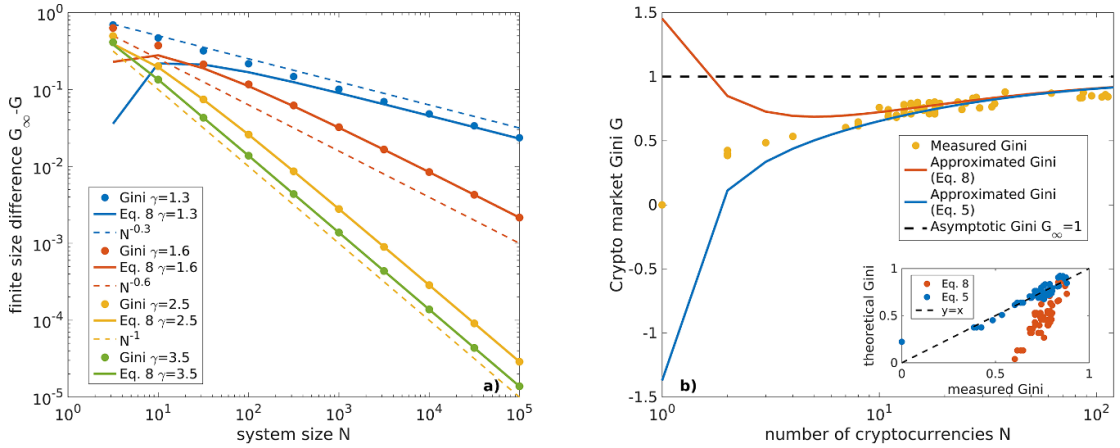


**Figure 2.** Maximal inequality regime. (a) Dots show the finite size difference between Gini index and its asymptotic value 1 as function of $N$ for different values of Zipf exponent $\gamma$. The solid line represent the approximation given by equation (9), while dashed lines are the scaling predicted by equation (10). As expected different scaling regimes are found for $1 < \gamma < 2$ and for $\gamma > 2$ and also in this case the convergence to the exact value is slow if Zipf's exponent is close to one. (b) Gini index of cryptocurrencies above one billion dollars of market cap from 2013 to 2021 as function of their number $N$ (yellow dots), compared with the theoretical values given by equation (5) (blue line) and equation (9) (red line). The black dashed line shows the asymptotic value $G_\infty = 1$.

value $G_\infty = 1$. Both equations (5) and (9) give a good prediction of the Gini index of the system already for $N \sim 10$. This is more clear in the inset where we report a direct comparison between the measured Gini index and the theoretical predictions computed using the two different approximations. As for the case of US cities, also for this system we can conclude that the increase of inequality we observed is the result of its expansion and does not reflect a change in the functioning of this financial system. This further confirms that is very important to disentangle the increase of inequality due to structural changes from that deriving from the expansion of the system.

## 3. Discussion

Assessing inequality is a central problem in economics and statistics that is often tackled using the well-known Gini index. In this context, many systems are characterized by being described in terms of Zipf's law, an ubiquitous scaling law that is by many considered one of the typical signs of complexity. Zipf's law relates the size of the elements $S(k)$ composing a system to their position in the ranking $k$ by a power law of exponent $\gamma$. In this paper we studied how the Gini index of a system showing Zipf's law depends on the exponent and on the size of the system $N$, given by the number of elements it contains. Indeed, since the long tail of a power law distribution is very hard to sample, using asymptotic expressions for Gini index in finite size systems may result in a severe miscalculation of inequality. This makes the assessment of finite size effects on inequality measures particularly relevant, since wealth, income, GDP and many other quantities typically considered in economics are characterized by a power law tail.

By means of analytical calculations and numerical simulations we thus compute finite size corrections to the asymptotic expressions and we show that depending on the value of Zipf's exponent $\gamma$ two different regimes can be identified: low inequality and maximal inequality. In the low inequality regime, observed for $\gamma < 1$, Gini index converges to a value smaller than one with the scaling $N^{\gamma-1}$. In the maximal inequality regime, corresponding to $\gamma > 1$, Gini index tends to its maximal value (one), but two different scalings are observed. For $1 < \gamma < 2$ the scaling is $N^{1-\gamma}$, while when $\gamma > 2$ it becomes $N^{-1}$ independently of Zipf's exponent. Note that for $\gamma > 1$ Gini index is not defined in infinite systems due to the divergence of the mean value, but our computation shows that considering a finite system and then taking the infinite size limit leads to a well-defined Gini index. In both cases the convergence to the asymptotic value is never faster than $1/N$ and for values $\gamma \approx 1$, exponent observed in many real systems, it becomes tremendously slow [32].

In order to test the validity of our approach, we applied it to two real systems, US cities and the cryptocurrency market. The former is characterized by $\gamma \approx 0.9$ and it is thus in the low inequality regime. As expected, by following its development from 1790 to 1950, we observe a growth of inequality that nicely obeys the theoretical prediction. Moreover, since the exponent is close to one, the system is far from reaching the asymptotic value of inequality. Conversely the cryptocurrency market in the maximal inequality regimes, since Zipf's exponent is $\gamma \approx 1.7$. In this case we observe a faster convergence to the asymptotic value 1 which is again well described by the formulas we derived. It is important to remark that in both situations, a direct

computation of Gini index not considering finite size effect, would result in the conclusion that both systems are changing their economical structure evolving toward a more uneven configuration. Instead we showed that such a growth is entirely explained by the growth of the systems and has nothing to do with a change on the economical conditions. Taking into account finite size effects is thus particularly important when assessing the effects of policies aimed at reducing inequality. Indeed the positive effects of such policies may be hidden by the growth of inequality spontaneously occurring as result of the system expanding. Furthermore, the results we discussed open questions about the claims of growing inequality in the world [33, 34]. Indeed, the world population has been rapidly increasing in the last century, as well as the number of billionaires. As a consequence the increase of inequality one observes could be largely caused by such a growth and not due economical factors.

We conclude by stressing that the approach here introduced can also be very useful when dealing with data coming from surveys. In such a situation one can only access a subsample of the system and thus a direct computation of Gini index on this subsample result in an underestimation of inequality. On the other side, also the parametric approach proposed in [15, 16] and based on ML techniques present some drawbacks. Indeed such an approach is based on asymptotic expressions and thus gives poor results in finite size systems with Zipf's exponent close to one leading to an overestimation of inequality. Much better results can be achieved by combining a ML approach with the finite size formulas we derived. The idea is to first computes the Zipf's exponent of the system by applying ML techniques to the available subsample [35] and then by leveraging on equation (5) (or the other approximations) it is possible to get the Gini index of the whole system, the only additional information needed being its size. This simple procedure allows to determine the inequality of a system characterized by long tails starting from survey data, while naive computations should be avoided since they give biased and unreliable results.

## 4. Methods

### 4.1. Fitting procedure

In order to compute the average Zipf's exponent of US cities we compute the exponent of the underlying probability distribution $P(S) \sim S^{-\alpha}$ for all the years in the time period under available. We do this using the Python power law package [36] which implements the technique described in [5]. In short this method exploit the maximum-likelihood fitting technique and the Kolmogorov–Smirnov statistic to asses both the exponent of the power law probability distribution $\alpha$ with its standard error $\sigma$ and the lower cutoff where the power-law behavior ceases to hold $s_{\min}$. In this way for each year $t$ we obtained the exponent $\alpha_t$ and its standard error $\sigma_t$. We then computed the mean power law exponent $\langle\alpha\rangle$ as an average weighted with the standard error

$$\langle\alpha\rangle = \frac{\sum_t \frac{\alpha_t}{\sigma_t}}{\sum_t \frac{1}{\sigma_t}}$$

and we obtained

$$\langle\alpha\rangle_{\text{cities}} = 2.10.$$

Starting from the power law exponent, Zipf's exponent $\gamma$ can be easily computed by the relation $\gamma = 1/(\alpha - 1)$ and so we ended up with the value

$$\gamma_{\text{cities}} = 0.91.$$

Note that for each year $t$ we computed Gini index only over those cities in the power law tail, i.e. those cities with size $S > s_{\min}^{(t)}$, where $s_{\min}^{(t)}$ is the lower cutoff returned by the fitting procedure for year $t$. As a consequence also the number of cities for year $t$ is computed as the number of cities with a population larger than $s_{\min}^{(t)}$.

### 4.2. Datasets

*4.2.1. Cryptocurrencies*

The dataset we exploited is the same studied in [9]. In particular historical market cap data have been downloaded from https://coinmarketcap.com/. The full dataset cover the period 28 April 2013–9 September 2021 and contains a total of 4588 cryptocurrencies (at the time of download). For each coin we only retained market capitalization data for each day in the period mentioned above, for a total of 3057 days. We then limit our analysis to cryptocurrencies above one billion dollar of market capitalization and we used the Zipf's exponent computed in [9].

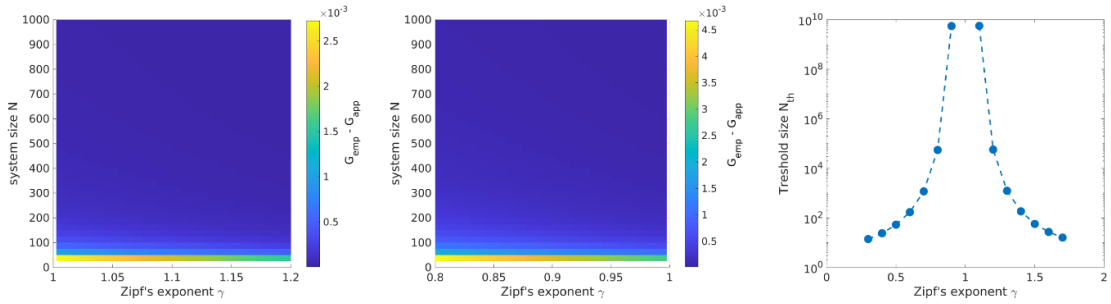


**Figure 3.** Goodness of approximations. (a) Difference between the empirical Gini index and that computed using equation (5) for different system sizes $N$ and Zipf's exponents $\gamma$ in the low inequality regime. As it is possible to see even for $\gamma \approx 1$ the approximation provides very reliable results. (b) As (a), but considering the maximal inequality regime. (c) Threshold size $N$th over which the approximations given by equations (7) and (9) can be considered reliable. One can notice that as the Zipf's exponents gets closer to one, both approximations only hold for very large systems.

*4.2.2. US cities*

The historical population of US cities has been gathered from [37]. The dataset provides United States historical city populations decennially between 1790 and 2010 and has been compiled mainly using US Census Bureau dataset of ∼7500 incorporated cities whose populations surpassed 2500 people at some point in their existence, for a total of 8911 cities. More information and the data can be found at https://github.com/cestastanford/historical-us-city-populations. For each year in the dataset we only retained those cities above the lower cutoff returned by the fitting procedure described above.

### 4.3. Goodness of the approximations proposed

As aforementioned, when Zipf's exponent becomes close to one, equations (5), (7) and (9) stop to be valid and instead the formula derived for $\gamma = 1$, given by equation (6), should be considered. Here we provide some results about the reliability of the different expressions in this scenario. First we test the goodness of the general expression equation (5), both in the low inequality and the maximal inequality regime. We report in figures 3(a) and (b) the difference between the empirically measured Gini index and that computed by means of the equation we just mentioned. As it is possible to see, even when $\gamma \approx 1$, there are only negligible differences between the two values, this meaning that equation (5) provides meaningful results also when Zipf's exponent is very close to one. This result implies that the approximations given by equations (7) and (9) break because of the expansion of the denominator appearing in equation (5). We can thus determine the threshold size $N$th over which these approximations are reliable by comparing the two terms in the denominator. In particular in the low inequality regime it must be

$$\frac{1}{1-\gamma} N^{1-\gamma} \gg |\zeta(\gamma)|,$$

while in the maximal inequality one the opposite must hold. By setting an arbitrary tolerance level at 0.1, the threshold size in the low and maximal inequality regimes can be derived imposing, respectively

$$\frac{1}{1-\gamma} N\text{th}^{1-\gamma} = 10|\zeta(\gamma)|$$

and

$$\frac{1}{\gamma-1} N\text{th}^{1-\gamma} = 0.1\zeta(\gamma).$$

We show in figure 3(c) the numerical solutions of these equations for different values of $\gamma$. As expected we observe that as $\gamma$ tends to one, larger and larger systems must be considered in order for the approximations to hold.

## Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

## ORCID iD

Giordano De Marzo https://orcid.org/0000-0002-3127-5336

## References

[1] Tria F, Loreto V, Servedio V D P and Strogatz S H 2014 *Sci. Rep.* **4** 1
[2] De Marzo G, Labini F S and Pietronero L 2021 *Astron. Astrophys.* **651** A114
[3] Corominas-Murtra B, Hanel R and Thurner S 2015 *Proc. Natl Acad. Sci.* **112** 5348
[4] De Marzo G, Gabrielli A, Zaccaria A and Pietronero L 2021 *Phys. Rev. Res.* **3** 013084
[5] Clauset A, Shalizi C R and Newman M E 2009 *SIAM Rev.* **51** 661
[6] Li W 2002 *Glottometrics* **5** 14
[7] Small C, Elvidge C D, Balk D and Montgomery M 2011 *Remote Sens. Environ.* **115** 269
[8] Clementi F and Gallegati M 2005 *Physica* A **350** 427
[9] Marzo G D, Pandolfelli F and Servedio V D 2022 *Sci. Rep.* **12** 1
[10] De Marzo G, Gabrielli A, Zaccaria A and Pietronero L 2022 *Phys. Rev. Res.* **4** 033079
[11] Pietronero L, Tosatti E, Tosatti V and Vespignani A 2001 *Physica* A **293** 297
[12] Gabaix X, Gopikrishnan P, Plerou V and Stanley H E 2003 *Nature* **423** 267
[13] Pareto V 1896 *Cours d'Economie Politique: Professe a l'Universite de Lausanne* vol 1 (Lausanne: F. Rouge)
[14] Taleb N N 2007 *The Black Swan: The Impact of the Highly Improbable* vol 2 (New York: Random House)
[15] Taleb N N 2015 arXiv:1510.04841
[16] Fontanari A, Taleb N N and Cirillo P 2018 *Physica* A **502** 256
[17] Gini C 1914 *Atti R. Ist. Veneto Sci. Lett. Arti* **73** 1203
[18] Allison P D 1978 *Am. Sociol. Rev.* **43** 865
[19] Ceriani L and Verme P 2012 *J. Econ. Inequal.* **10** 421
[20] Cowell F 2011 *Measuring Inequality* (Oxford: Oxford University Press)
[21] Atkinson A B 1975 *The Economics of Inequality* (Princeton, NJ: Citeseer)
[22] Chen Y 2012 *Physica* A **391** 767
[23] Wang Q A 2021 *Chaos Solitons Fractals* **153** 111489
[24] Davidson R 2009 *J. Econometrics* **150** 30
[25] Cowell F A and Flachaire E 2007 *J. Econometrics* **141** 1044
[26] Li Q, Li S and Wan H 2020 *China Econ. Rev.* **62** 101495
[27] Gastwirth J L 1972 *Rev. Econ. Stat.* **54** 306
[28] Farris F A 2010 *Am. Math. Mon.* **117** 851
[29] Eliazar I I and Sokolov I M 2010 *Physica* A **389** 117
[30] Pay attention to the different notation, in [16] the probability distribution is defined as $P(s) \sim s^{-\alpha'-1}$, thus $\alpha' = \alpha - 1$
[31] Zipf G K 1949 *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Cambridge, MA: Addison-Wesley)
[32] It can be shown that for $\gamma = 1$ the convergence is logarithmic
[33] Piketty T 2014 *Capital in the Twenty-First Century* (Cambridge, MA: Harvard University Press)
[34] Piketty T 2015 *The Economics of Inequality* (Cambridge, MA: Harvard University Press)
[35] A random sampling of a system showing Zipf's law results in a subsample also showing Zipf's law with the same exponent, see for instance [4]
[36] Alstott J, Bullmore E and Plenz D 2014 *PLoS One* **9** e85777
[37] Steiner E (US Census Bureau) 2018 Spatial history project (available at: https://github.com/cestastanford/historical-us-city-populations)