I. INTRODUCTION

Power law distributions have fascinated generations of researches for decades. Cluster sizes at percolation [1]. Avalanche sizes in self-organized criticality [2]. The degree distribution of scale-free networks [3]. However, when faced with empirical data, calling a power law distribution is on the eye of the beholder [4–8]. The power law claims are challenged by other distributions. Most commonly the lognormal and stretched exponential distributions [4, 9, 10].

The correct distribution is important if scale invariance is all what matters. That property is unique to power law distributions. There may be other questions where having a heavy tail is sufficient. The distribution being power law or lognormal a matter of degree. One such case is the sum of independent and identically distributed random variables. The power law claims are challenged by other distributions. Most commonly the lognormal and stretched exponential distributions [4, 9, 10].

The correct distribution is important if scale invariance is all what matters. That property is unique to power law distributions. There may be other questions where having a heavy tail is sufficient. The distribution being power law or lognormal a matter of degree. One such case is the sum of independent and identically distributed random variables. The power law claims are challenged by other distributions. Most commonly the lognormal and stretched exponential distributions [4, 9, 10].

The work on stable distributions have established the types of distributions that are fully invariant to summation [11]. In that class we find the normal distribution. For several distributions, bounded or with superexponential tails, the distribution of $S_n = \sum_{i=1}^n X_i$. In this context we encounter a different type of invariance, the distribution invariance to summation.

The shape of empirical distributions with heavy tails is a recurrent matter of debate. There are claims of a power laws and the associated scale invariance. There are plenty of challengers as well, the lognormal and stretched exponential among others. Here I point out that, with regard to summation invariance, all what matters is they are subexponential distributions. I provide numerical examples highlighting the key properties of subexponential distributions. The summation invariance and the black swan dominance: the sum is dominated by the maximum. Finally, I illustrate the use of these properties to tackle problems in random networks, infectious dynamics and project delays.

II. THE CENTRAL LIMIT THEOREM

The sum of random variables

$$S_n = X_1 + \cdots + X_n, \quad (1)$$

is found in many problems. The random variables $X_i$ may be daily changes of a stock price or activity delays in a construction project. The standard text book procedure is to invoke the central limit theorem. If the $X_i$ are independent random variables from the same probability density function $f(x)$ then the probability density...
The idyllic picture in Fig. 1 does not hold in general. The central limit theorem is central for a reason.

The work on subexponential distribution shifts the attention from the mode to the tail. We are interested in positive defined variables and therefore the right tail. We define the tail distribution of \( f(x) \) and \( g_n(x) \) as

\[
\bar{F}(x) = \text{Prob}\{X > x\} = \int_x^\infty f(\xi)d\xi, \quad (4)
\]

\[
\bar{G}_n(x) = \text{Prob}\{S_n > x\} = \int_x^\infty g_n(\xi)d\xi. \quad (5)
\]

The tail of \( \bar{G}_2(x) \) has the lower bound

\[
\bar{G}_2(x) \geq 2\bar{F}(x) \quad (6)
\]

for \( x \to \infty \) \cite{12}. The tail of two identically independent random variables is at least as heavy as \( 2\bar{F}(x) \). For example, for an exponential distribution \( f(x) = \exp(-x/\mu)/\mu \), we get \( x \geq 2\mu \), which is certainly true for \( x \to \infty \).

Subexponentials are distributions with a slow decaying tail that cannot be bounded by any exponential. There is no \( \mu > 0 \) such that \( \bar{F}(x) < \exp(-x/\mu) \) for \( x \to \infty \). For subexponential distributions the lower bound in Eq. (6) becomes an approximation \cite{12}

\[
\bar{G}_2(x) \approx 2\bar{F}(x) \quad (7)
\]

for \( x \to \infty \). The distribution tail of two identically independent random variables with a subexponential distribution is as heavy as \( 2\bar{F}(x) \). The property (7) is used as an alternative definition of subexponentiality.

Equation (7) can be extended to any \( \bar{G}_n(x) \) with \( n \geq 2 \). That was proven by Chistyakov in 1964 \cite{13}. If \( f(x) \) is a subexponential distribution, as defined by Eq. (7), then

\[
\bar{G}_n(x) \approx n\bar{F}(x) \quad (8)
\]

for \( x \to \infty \). Upon summation the tail gets heavier by a factor of \( n \), but the shape remains the same as that of \( \bar{F}(x) \). I will call this the subexponential limit theorem.
A. Lognormal distribution

The lognormal probability density function is

\[ f(x) = \frac{1}{x \sqrt{2\pi \sigma^2}} \exp\left( -\frac{\ln x - \mu^2}{2 \sigma^2} \right), \quad (9) \]

where \( \mu \) and \( \sigma^2 \) are the location and variance of \( \ln x \). The tail of the lognormal density is of the order of \( \exp(-\ln x^2/2\sigma^2)/x \) and decays slower than any exponential. The larger is \( \sigma \) the heavier is the tail. The sum of random variables with a lognormal is not approximated by the normal density of the central limit theorem (Fig. 3a, b, solid circles vs dashed line). In contrast, the subexponential limit \( n_f(x) \) is a good approximation for the tail of \( g_n(x) \) (Fig. 3a, b, solid circles vs solid line). The agreement expands over a wider range the higher is the degree of subexponentiality. Higher \( \sigma \) in this case.

B. Power law distributions

A Pareto or power law distribution has the probability density function

\[ f(x) = \begin{cases} \frac{1}{\alpha} \left( \frac{\mu}{x} \right)^{1+\alpha}, & \text{if } x < \mu \\ 0, & \text{otherwise} \end{cases} \]

where \( \alpha \) is the tail exponent and \( \mu \) is scale parameter. The tail of the power law density is slower than any exponential. The tail is heavier the smaller is \( \alpha \). The power law distribution is scale invariant. If we apply the scale transformation \( x \to ax \) the distribution is transformed as \( f(x) \to f(ax)d(ax) = f(x)dx/a^\alpha \). The shape of the distribution remains invariant. Scale invariance is specific to power law densities. Since the power law distribution is subexponential the tail distribution is also invariant to summation. Figures 3c, d show that indeed, \( g_n(x) \approx n_f(x) \) for large \( x \). Here again the agreement extends to a wider range the higher is the degree of subexponentiality. Lower \( \alpha \) in this case.
V. MAXIMUM LIMIT THEOREM

The subexponential limit theorem has a follow up associated with the maximum. Let

\[ M_n = \max(X_1, X_2, \ldots, X_n), \] (11)

be the maximum of independent and identically distributed random variables, with tail distribution \( H_n(x) = \text{Prob.}\{M_n \geq x\} \). For \( x \to \infty \) we have \( \bar{F}(x) \to 0 \) and therefore (Ref. [12, Definition 3.1])

\[ H_n(x) = 1 - (1 - \bar{F}(x))^n \approx n\bar{F}(x). \] (12)

Note this limit is valid for any \( \bar{F}(x) \), independently of the subexponential property. Now, from Eqs. (12) and (8) it follows that

\[ G_n(x) \approx H_n(x). \] (13)

The tail distribution of the sum is approximated by the tail distribution of the maximum. I will call this result the maximum limit theorem.

Figure 3 shows that the tail of \( h_n(x) \) is a good approximation for the tail of \( g_n(x) \), the probability density function of the sum \( S_n = X_1 + \cdots + X_n \) (circles vs triangles).

VI. BLACK SWAN DOMINANCE

For the ancient Romans \textit{niger cyanus} was quintessential \textit{rara in terris} [14]. Today black swan is a common term for rare events of extreme magnitude. Black swan events are associated with the tail statistics. The heavier the tail of the distribution the higher the chance to observe a large event, a black swan. There is debate on whether black swans are extreme events in the tail of the distribution or unforeseen events happening at a higher frequency than expected from the reference statistics [13].

I propose the maximum limit theorem as a working definition of black swans. If \( \text{Prob.}\{X_1 + X_2 > x\} \approx \text{Prob.}\{\max(X_1, X_2) > x\} \) then the only way \( X_1 + X_2 \) is large is that \( \max(X_1, X_2) \gg \min(X_1, X_2) \). Using Eq. (14) we can extend that argument. The only way \( \sum X_i \) is large is that \( \max X_i \gg \max_{i \neq k} X_i \), where \( k = \arg\max X_i \) is the largest event index. One event, the black swan, dominates all. That is the definition of black swan dominance.

The black swan is defined in relative terms. We do not specify a threshold or scale. We demand that \( \max X_i \gg \max_{i \neq k} X_i \). Therefore a black swan is context dependent. We can have a construction project where all activities are not delayed, except one delayed by 30 days. Since 30 \( \gg 1 \) days we call that a black swan. In another construction project we can have 10 critical activities delayed by about 30 days, resulting in an aggregated delay of 300 days. Since 30 \( \ll 300 \) we don’t call any of these 30 days delays a black swan. The stress is in the dominance. One event of magnitude much larger than the others.

VII. APPLICATIONS

A. Aggregate connectivity

Suppose we have a network \( N \) composed of \( N \) nodes and \( M \) links between pairs of nodes. The degree of a node is the number of links containing that node or, equivalently, the number of neighbors or the number of interacting partners. Let \( p_k \) be the degree distribution across nodes. We want to know the aggregate number of links between a community \( C \) of \( n \) nodes and the complement \( N \setminus C \). The \( n \) nodes may represent a community in a social network or a pathway in a biological network, for example.

We can approximate the aggregate connectivity by the sum in Eq. (14), where \( X_i \) are the node degrees. This approximation holds if (i) there is a small overlap between the direct neighbors of the \( n \) nodes in \( N \setminus C \) and (ii) the number of links between the \( n \) nodes is much smaller than \( S_n \). Most real networks, including social and biological networks are characterized by heavy tailed degree distributions [16, 17]. If \( p_k \) has a subexponential tail then the probability density function of \( S_n \) has a tail approximated by \( np_k \). Bottom line, the aggregation of nodes makes the tail heavier by a factor of \( n \). The probability to find extreme hubs with large aggregate connectivity increases by a factor of \( n \).

B. Giant component of random networks

Consider a random network with given degree distribution \( p_k \) and excess degree distribution \( q_k = k p_k / \langle k \rangle \). \( p_k \) characterizes the degree of a vertex selected at random and \( q_k \) the degree distribution of a vertex at the end of link selected at random. Molloy & Read [18] demonstrated a transition from disconnected clusters to a giant component when

\[ \theta = \sum_k (k - 1) q_k = \frac{\langle k(k-1) \rangle}{\langle k \rangle} = 1. \] (14)

This result can be recapitulated using the subexponential limit theorem in Eq. (8).

The cavity method provides estimates for networks with a tree like structure [19–21]. In a nutshell, we focus on a node \( i \) at the end of a link selected at random. The excess degree is the set of links emanating from node \( i \) excluding the link we came from. Node \( i \) will have excess degree \( k - 1 \) with probability \( q_k \). At this point we want to calculate the probability distribution \( Q_k(s) \) that we can reach \( s \) other nodes following the excess links of node \( i \), given that node \( i \) has degree \( k \). If the network has a tree like structure then how many nodes are reached from node \( i \) can be written as the sum of \( k - 1 \) neighbors plus how many nodes can be reached from each neighbor

\[ S_k = k - 1 + X_1 + X_2 \cdots + X_{k-1}. \] (15)
We note that $S_k$ is a random variable extracted from $Q_k(s)$ and $X_i$ is a random variable extracted from $(Q(s) = \sum_k q_k Q_k(s)$.

If $(Q(s)$ is a subexponential distribution then $X_1 + X_2 \cdots + X_{k-1}$ is approximately distributed as $(k - 1)(Q(s)$ for $s \to \infty$. In this case

$$Q_k(s) = (k - 1)(Q(s - k + 1)$$

Taking the expectation over $q_k$ in both sides and the limit $s \to \infty$ we obtain

$$\langle Q(s) \rangle = \theta \langle Q(s) \rangle$$

This identity holds for $\theta = 1$. Therefore, $(Q(s)$ is subexponential only when $\theta = 1$. This is the critical point first obtained by Molloy & Reed [18] and later extended to percolation on random networks [19, 20]. For $\theta < 1$, the less connected phase, $Q(s)$ is bounded by an exponential decay. For $\theta > 1$, the more connected phase, the deviation from subexponentiality can only be explained by the presence of a giant component.

This example illustrates the use of the subexponential property in the context of networks with a tree-like structure. We did not invoke scale invariance or a power law distribution of components size. We just demanded the subexponential property, the invariance with respect to summation of the distribution tail.

C. Spreading dynamics

The spreading of an infectious agent (disease, computer virus, rumor, etc) in static or temporal networks can be modeled as a branching process [22]. The average number of new infectious at time $t$ is approximated by

$$n(t) = R_0 \sum_{n=1}^{D} R^{n-1} g_n(t),$$

where $R_0$ is the expected reproductive number of the starting node (patient zero), $R$ is the expected reproductive number of any other node in the transmission tree, $D$ is the maximum number of generations (the diameter for static networks) and $g_n(t)$ is the distribution of the sum of $n$ generation times.

The generation time is the time elapsed from acquiring to transmitting the disease. Let $f(t)$ be the probability density function of generation times. If the infection spreads at a constant rate $\mu$, a Poisson process, then $f(t)$ is an exponential distribution and $g_n(t)$ is given by the gamma distribution in Eq. [16]. Substituting Eq. [16] into [18] we obtain

$$n(t) = R_0 \exp \left( -\frac{t}{\mu} \right) R_0 \sum_{n=1}^{D} \frac{1}{(n-1)!} \left( \frac{Rt}{\mu} \right)^{n-1}.$$  

When $t \ll \mu D / R$ the sum approximates the Taylor expansion of the exponential and $n(t) \approx R_0 \exp(-1 - R(t/\mu)$.

An exponential decay provided $R < 1$. When $t > \mu D / R$ the sum is dominated by the $r = D$ term and $n(t) \propto t^{D-1} \exp(-t/\mu)$ [23]. Again an exponential decay for large $t$.

In contrast, if the generation time distribution is subexponential, then according to the subexponential limit theorem, Eq. [8], $g_n(t) \approx n f(t)$ in the limit $t \to \infty$. Substituting this result in Eq. [13] we obtain

$$n(t) \approx f(t) R_0 \sum_{n=1}^{D} n R^{n-1},$$

The long time limit of infection dynamics with a subexponential distribution of generation times is characterized by the same subexponential tail. In particular, this holds true when $f(t)$ has a power law tail, as previously reported [24].

The time between sexual intercourses follows a power law decay [22]. Therefore, the subexponential decay is relevant for the infection dynamics of sexually transmitted diseases. The distribution of the inter-event time between submission of two consecutive emails by an email user has a fat tail [25]. Interestingly, whether the tail is a power law or lognormal was subject to debate [5]. The marriage of these two distributions into the subexponential class closes the debate. The subexponential decay is relevant for the infection dynamics of email worms and other infectious agents transmitted via email. Finally subexponential distributions of inter-event times are found in many other systems [26].

D. Activity networks dynamics

Human projects are organized as activity networks, where nodes represent activities and arcs represent logical constraints: predecessor must finish before the successor starts. In the planning phase a project schedule is generated based on estimates of activity durations. A critical path is established, containing a sequence of activities from beginning to end that are executed without spare time between them. For paths outside the critical path there is some spare time (float). The total float between two activities, denoted by $T_{ij}$, is the spare time along the path of minimum spare time starting at $i$ and ending at $j$, with a negative sign (consumes delay). If there is no path then $T_{ij} = -\infty$, indicating that no delay will be transmitted from $j$ to $i$.

Once the project starts there are delays in the activity completions. If the delay at an activity exceeds the spare time between the activity an a successor then a delay is transmitted to the successor, initiating a delay cascade. There are two sources of delay. Exogenous delays increasing activity duration independently of what happened to the predecessors, with magnitude $z$. Endogenous delays carried from predecessors, with magnitude $h$. The total delay an activity will experience is $S_z = z + h$. If the distribution of delays is subexponential then, invoking the
maximum limit theorem, \( S_N \approx \max(z, h) \). In this case the delay \( d_i \) in the completion of activity \( i \) is approximated by

\[
d_i \approx \max_{j=1}^{N}(h_j + T_{ij}),
\]

(21)

where \( N \) is the number of activities and \( h_i \) the exogenous delay at activity \( i \).

Equation (21) has no reference to delay interactions at intermediate activities in the paths from \( j \) to \( i \). We simply aggregate the contribution of each exogenous delay independently and then take the max. When the exogenous delay follow a subexponential distribution, their contribution to delays at downstream activities can be calculated independently of each other. Just take the max at the end.

VIII. CONCLUSIONS

Researchers dealing with heavy tailed statistics should pay attention to the mathematical literature on subexponential distributions. This class includes, among others, power law, lognormal and stretched exponential distributions. From the practical point the subexponential distributions have two key properties: the distribution tail is invariant to summation and for large values the sum is approximated by the max. Here I have shown some examples were subexponentiality is the only thing that matters. These results are not limited to power law distributions. Debates on whether some empirical data is better described by a power law, lognormal or stretched exponential distribution may be irrelevant for certain problems.

ACKNOWLEDGEMENTS

This work was partly supported by European Union’s Horizon 2020 and the Cyprus Research & Innovation Foundation under the SEED program (grant agreement 0719(B)/0124). Nodes & Links provided support in the form of salary for AV, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.