The distribution functions of vessel calls and port connectivity in the global cargo ship network

Michael T. Gastner* and César Ducruet†
*Institute of Technical Physics and Materials Science, Research Centre for Natural Sciences,
Hungarian Academy of Sciences, P.O. Box 49, H-1525 Budapest, Hungary, Email: mgastner@gmail.com
‡CNRS / University of Paris-I Panthéon Sorbonne, UMR 8504 Géographie-Cités / P.A.R.I.S.,
13 rue du Four, F-75006 Paris, France, Email: cdu@parisgeo.cnrs.fr

***Abstract*—Characterizing the connectivity of nodes in economic and technological networks is of key interest to assess their role and function. Here we study the distributions of vessel calls and the ports' degrees (i.e. the number of other ports that a port is directly linked with) in the network of worldwide cargo ship movements – the main transport network for world trade – for twenty different years between 1890 and 2008. We compare the Akaike information criterion and goodness-of-fit statistics of various common probabilistic models. Simple power laws, once believed to be a universal feature of degree distributions in complex networks, are inadequate to fit the data. Other subexponential distributions, such as lognormal or Weibull distributions, perform consistently better. Cargo ship traffic has thus for the entire study period been heavy-tailed with some ports being significantly busier than the average, but the distribution is not scale-free. Vuong's likelihood ratio test confirms that since 1975 a Weibull distribution can be regarded as a plausible null hypothesis. Lognormal distributions perform well for most years in Kolmogorov-Smirnov and Anderson-Darling tests for the call distribution. The Gini coefficient of the distribution has slightly, but statistically significantly, decreased over the study period, highlighting a tendency towards a more polycentric distribution in port traffic.**

I. Introduction

Recent years have seen intense research activity in the modelling and analysis of complex networks, mainly driven by the availability of new large-scale data bases for social, biological and technological networks (see for example [1] for a review). Maritime transport networks are one area where these new ideas and techniques have found fertile ground [2], [3], [4], [5]. In this study, we analyze a data base generated from Lloyd's Shipping Index, a weekly publication of cargo ship movements by Lloyd's List, over the period from 1890 to 2008. For twenty selected years, an entire volume of the Index, each containing data for one entire week, was extracted and the data transformed into a network where the nodes are ports and links are nonstop ship voyages. Because cargo shipping is the dominant transport mode for world trade [6], it is of great economic relevance to understand the importance of the nodes. Here we measure importance in two different ways:

- the number of vessel calls and
- the degree defined as the number of ports that the node is connected to by at least one arriving or departing ship.

The call and degree distributions are arguably the two most important summary statistics of the network. They may not allow a complete reconstruction of traffic on the links. However, unweighted and weighted degree distributions are an important feature of a network's topology and have often been used as circumstantial evidence for mechanistic models of the network's evolution [7], [8], [9]. The call distribution also plays a crucial role for predicting the full origin-destination matrix (i.e. the traffic between all pairs of ports) because it is an input in transport forecasting (e.g. in the gravity model or the intervening opportunities model [10]).

In the early phase of complex network science, many degree distributions of real-world networks were characterized as power laws [7]. In the parlance of statistics, an integer-valued power law is a probabilistic model that assigns the probability $\Pr(k)$ to the event that an arbitrary node has degree or weight k so that

$$\Pr(k;\tau) = \frac{k^{-\tau}}{\zeta(\tau)}, \qquad k = 1, 2, 3, \ldots \tag{1}$$

Here $\zeta(\tau)$ is the Riemann zeta function and $\tau > 1$ a fixed parameter that has to be fitted to the data. Because $\Pr(ak;\tau) = a^{-\tau}\Pr(k;\tau)$, the distribution of Eq. 1 is also called scale-free. The interest in power law distributions stems mainly from the fact that these are particularly heavy-tailed (i.e. the tail of the distribution decays more slowly than an exponential function). As a consequence, a power-law distribution has a large range in degrees: while most nodes have only a small degree, some nodes possess a much larger degree than the average. At first glance, such a heterogeneity in degrees is indeed found in many empirical networks [11].

Most of the time, a power-law degree distribution was inferred from straight-line fits to the log-log diagram of the degree frequency, but this is now generally viewed as an unsatisfactory approach [12]. Identifying a region where the data appear more or less linear is largely arbitrary because most distributions are too noisy and substantially curved on double-logarithmic scales. Straight-line fits based on standard least-squares algorithms can also lead to a bias in the estimated exponents. Most importantly, however, there is no a priori reason why $\Pr(k)$ has to be a power law. Many other common probability distributions are also heavy-tailed and may fit the data better. Recent studies in fact doubt that power laws are as ubiquitous as once believed [13], [14], [15].

This study proposes that statistical methods should be applied to the maritime network to understand which type of call or degree distribution best explains the observed data. We apply information-based model selection [16] and statistical tests to compare different candidate distributions. Another goal of the paper is to assess whether the call and degree distributions of the cargo ship network have significantly changed over time. One hypothesis from the geographic literature is that the call and degree distributions may structurally change alongside major technological transformations of the shipping industry and their consequences on port operations and maritime network configurations [17]. The studied period goes across different dominant ship technologies, such as sail, steam, combustion, specialized vessels (e.g. container, tanker), and mega-carriers. Such technological evolutions are believed to have been selective, as some ports were dropped from the network and replaced or superseded by new ones better adapted to changing standards, sometimes resulting in an increasing concentration of port activity favouring fewer and larger ports. Containerization is seen as a revolution in itself with profound impacts on network configuration and world trade [18], [19]. In this study we find indeed evidence that the call distributions have evolved so that the fraction of small ports has decreased. At the same time the Gini coefficient, a common measure for inequality of a distribution, has slightly decreased over the study period.

Before proceeding with the statistical analysis, we emphasize one caveat. The voyages reported by Lloyd's certainly form only a subset of the entire global traffic distribution and may possibly be biased, for example if certain ports, ships or routes are systematically underreported. The quality of reporting may also differ between different years. We currently have no sufficient knowledge whether such biases are present and thus cannot apply any corrections. Lloyd's Shipping Index, however, is the most complete and consistent data source available to study the development of the cargo ship network over the investigated time period. Therefore, we are confident that the trends reported below are genuinely representative of the network's evolution.

II. Probabilistic models

We investigate eight different models that have frequently been used to fit empirical degree distributions in complex networks (Table I). We restrict our study to discrete distributions

(a) whose support are all positive integers and
(b) that depend on maximally two parameters.

Restriction (a) reflects that degree or port calls only have integer values. One might argue that $k = 0$ should also be included and that the maximum degree should have an upper bound because the network is finite. However, from Lloyd's Shipping Index we cannot directly infer which ports were in principle open to traffic, but remained unused. Consequently, distributions with infinite support, but excluding $k = 0$ are more appropriate in the present context. The number of ports with $k > 0$ can differ slightly between the call and degree distributions: some ports have a positive number of calls but degree zero because in the raw data some vessels are reported to call at one, not two ports (namely origin and destination) in their latest known voyage. For the call distribution we kept these isolated ports, but removed them from the degree distribution since zero is unlikely to be their true degree. Restriction (b) avoids overfitting of the data, but still includes the "usual suspects" for degree distributions in socio-economic networks.

We include four one-parameter models: besides the power law of Eq. 1 (also known as zeta distribution), we assess the likelihood of Poisson, geometric and Yule-Simon distributions. The Poisson distribution describes the node degrees of large sparse Erdős-Rényi random graphs, a common null model in network studies. The tail of a Poisson distribution decays faster than exponentially so that degrees in Erdős-Rényi graphs are effectively limited to values near the mean degree. The geometric distribution decays exponentially, whereas the Yule-Simon distribution has a power law tail and only differs mildly from a strict power law for small k. Because the Yule-Simon distribution is the exact solution of popular "preferential attachment" models [20] (i.e. models where nodes are constantly added to the network and linked preferentially to nodes of high degree), we have included it in our list.

As a mixed case we introduce the exponentially truncated power law as one of our two-parameter models in Table I. The negative binomial is another two-parameter example that decays more slowly than an exponential if its parameter r exceeds 1, but with much less weight in the tail than a power law. All the distributions mentioned so far are discrete: they are defined for integer numbers k. In the case of the Poisson and negative binomial distributions, $k = 0$ is conventionally included in the distributions' support. In order to restore restriction (a) from above, we constrain these distributions to exclude $k = 0$, which explains why the equations in Table I differ from the ordinary textbook form.

Among continuous distributions, there are two further canonical candidates whose decay is between an exponential and a power law: the lognormal distribution and the Weibull distribution (also known as stretched exponential if $\beta < 1$ in the formula stated in Table I). We include these models in our study because previous studies have reported lognormal [21], [22], [23], [24] and Weibull distributions [25], [26], [27], [28] in real-world networks. To allow a direct comparison with the other models, we have to discretize the continuous lognormal and Weibull distribution, which can be accomplished in a variety of ways. Here we have chosen to integrate the continuous distributions between subsequent integers k and $k+1$. In terms of the cumulative distribution F, we can express the integral as the probability $F(k+1)-F(k)$. For the lognormal distribution, we assign this probability to the integer at the upper boundary $k+1$ and call this the "discrete lognormal" distribution. In the case of the Weibull distribution preliminary tests showed that the likelihoods are in general slightly larger if $F(k+1)-F(k)$ is assigned to k instead of $k+1$. Imposing the constraint (a), yields the expression for the "discrete Weibull" distribution in Table I.

TABLE I
THE INVESTIGATED PROBABILISTIC MODELS.

distribution	parameters	$\mathrm{Pr}(k)$, $k = 1, 2, 3, \ldots$
Poisson (POIS)	$\lambda > 0$	$\frac{\lambda^k}{(e^\lambda - 1)k!}$
geometric (GEOM)	$p \in (0, 1)$	$p(1-p)^{k-1}$
power law (ZETA)	$\tau > 1$	$[\zeta(\tau)k^\tau]^{-1}$
Yule-Simon (YULE)	$\rho > 0$	$\rho \mathrm{B}(k, \rho + 1)$
negative binomial (NEGB)	$p \in (0, 1)$, $r > 0$	$\frac{\Gamma(k+r)p^k(1-p)^r}{k!\Gamma(r)[1-(1-p)^r]}$
truncated power law (TPOW)	$q \in (0, 1)$, $\tau > 1$	$\frac{q^k}{\mathrm{Li}_\tau(q)k^\tau}$
discrete lognormal (DLGN)	$\mu \in \mathbb{R}$, $\sigma > 0$	$\begin{cases} \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(-\frac{\mu}{\sqrt{2}\sigma}\right) & \text{if } k = 1 \\ \frac{1}{2}\left[\mathrm{erf}\left(\frac{\ln k - \mu}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{\ln(k-1) - \mu}{\sqrt{2}\sigma}\right)\right] & \text{if } k = 2, 3, \ldots \end{cases}$
discrete Weibull (DWEI)	$q \in (0, 1)$, $\beta > 0$	$\frac{1}{q}\left[q^{(k^\beta)} - q^{((k+1)^\beta)}\right]$

One subtlety to note is that in this study we do not judge the fit of the distributions by the tails alone as it is often done elsewhere (e.g. [7], [12]). In the study of continuous phase transitions in physics it is justified to restrict attention to the tails because only these are important for determining "universal" power-law features. However, in the present context it is far-fetched to assume that the cargo ship network has anything to do with a physical phase transition. Instead we will assess the match of the distribution over the full set of positive integers $k = 1, 2, 3, \ldots$ with the same motivation as in the study of city size distributions in Ref. [29]. Although it is in principle possible to restrict the analysis to the tails by introducing a lower cutoff $k > 1$, this would introduce an additional parameter and ignore the bulk of the data which consists of ports with only few calls and low degree. On the contrary, we regard it as valuable information for practitioners to model low-traffic ports too, not only the small fraction of busy hubs that make up the distribution's tail.

III. AKAIKE INFORMATION CRITERION

Assuming that all calls and degrees are independent, the likelihood function for any of the models in Table I has the general form

$$L(\mathbf{v}) = \prod_{i=1}^{n} \mathrm{Pr}(k_i; \mathbf{v}), \tag{2}$$

where k_i is the number of calls (or the degree) at port i, n is the number of ports in the sample, and $\mathbf{v}$ is the set of the parameters in the second column of the table. The calls (or degrees) may in reality depend on each other so that $L(\mathbf{v})$ in Eq. 2 is more properly thought of as a composite likelihood. We can justify the use of a composite likelihood in our present context because the call (or degree) distribution $\mathrm{Pr}(k)$ that we would like to model is a marginal rather than the complete joint distribution of all calls (or degrees). Therefore, the full dependence structure is in statistical terms a "nuisance parameter" which neither matters to us nor is it clear how to specify the full likelihood. In such cases, composite likelihood methods have proved to be a well-behaved alternative [30], [31]. Another, more pragmatic, point of view is that there is no straightforward method to establish from our data how the k_i may depend on each other so that assuming independence is the most parsimonious choice.

For a specified model j, we determine the parameter $\hat{\mathbf{v}}_j$ that maximizes L and hence also the log-likelihood $\ln(L)$. A comparison between different models can then be performed by ranking their Akaike information criterion (AIC) [32],

$$\mathrm{AIC}_j = -2\ln(L(\hat{\mathbf{v}}_j)) + 2K_j, \tag{3}$$

where K_j is the number of parameters in the respective model. The AIC not only tells us which model is closest to the data in information content, properly taking into account that higher K_j generally allows better fits to the data, but weakens the explanatory power of the model. We can also make quantitative comparisons between different models based on the AIC differences. If $\mathrm{AIC}_{\mathrm{min}}$ is the minimum AIC over all models, then the difference

$$\Delta_j = \mathrm{AIC}_j - \mathrm{AIC}_{\mathrm{min}} \tag{4}$$

estimates the relative expected information gain between model j and the estimated best model. Because the likelihood of model j given the number of calls (or degrees) $k_1, k_2, \ldots$ is proportional to $\exp(-\Delta_j/2)$ [16], the relative likelihood is the so-called Akaike weight

$$w_j = \frac{\exp(-\Delta_j/2)}{\sum_i \exp(-\Delta_i/2)}, \tag{5}$$

where the summation in the denominator is over all models included in the comparison. Model selection by Akaike weights has become an increasingly popular tool to compare different hypothesized probability distributions [13], [21], [33], [34]. The Akaike weights for our data sets are summarized in Table II for the call distributions and in Table III for the degree distributions.

As a quick glance at the tables reveals, the maximum Akaike weights are achieved by the discrete lognormal and Weibull distributions and, in the case of the calls, in some years

TABLE II

AKAIKE WEIGHTS FOR THE DISTRIBUTION OF VESSEL CALLS. VALUES BELOW 10^{-200} ARE ROUNDED TO ZERO. THE LARGEST AKAIKE WEIGHT IN EACH YEAR IS HIGHLIGHTED IN BOLD TYPE.

Year	ports	POIS	GEOM	ZETA	YULE	NEGB	TPOW	DLGN	DWEI
1890	904	0	0	2.61×10^{-16}	1.60×10^{-10}	4.61×10^{-30}	3.64×10^{-1}	$\mathbf{6.10 \times 10^{-1}}$	2.58×10^{-2}
1910	1200	0	0	2.90×10^{-25}	4.56×10^{-17}	4.76×10^{-31}	$\mathbf{9.37 \times 10^{-1}}$	6.13×10^{-2}	2.12×10^{-3}
1915	992	0	0	1.51×10^{-23}	1.23×10^{-16}	3.55×10^{-23}	$\mathbf{7.70 \times 10^{-1}}$	2.26×10^{-1}	3.94×10^{-3}
1920	994	0	0	9.36×10^{-20}	2.07×10^{-13}	2.42×10^{-30}	$\mathbf{8.83 \times 10^{-1}}$	1.10×10^{-1}	6.22×10^{-3}
1925	1205	0	0	2.51×10^{-29}	4.17×10^{-20}	6.09×10^{-30}	6.97×10^{-2}	$\mathbf{8.27 \times 10^{-1}}$	1.03×10^{-1}
1930	1254	0	0	4.78×10^{-33}	3.16×10^{-23}	5.61×10^{-29}	9.64×10^{-3}	$\mathbf{9.64 \times 10^{-1}}$	2.66×10^{-2}
1935	1282	0	0	4.29×10^{-33}	3.15×10^{-22}	5.01×10^{-29}	1.03×10^{-2}	$\mathbf{5.24 \times 10^{-1}}$	4.66×10^{-1}
1940	1309	0	0	1.42×10^{-23}	2.62×10^{-15}	1.50×10^{-41}	$\mathbf{9.79 \times 10^{-1}}$	1.96×10^{-2}	1.20×10^{-3}
1946	1281	0	0	6.05×10^{-26}	7.30×10^{-17}	9.37×10^{-37}	$\mathbf{8.47 \times 10^{-1}}$	8.47×10^{-2}	6.80×10^{-2}
1951	1321	0	0	1.91×10^{-34}	1.06×10^{-23}	1.42×10^{-25}	3.21×10^{-1}	$\mathbf{5.67 \times 10^{-1}}$	1.13×10^{-1}
1960	1541	0	0	1.49×10^{-47}	1.51×10^{-34}	2.66×10^{-23}	8.99×10^{-2}	$\mathbf{8.78 \times 10^{-1}}$	3.22×10^{-2}
1965	1554	0	0	2.58×10^{-59}	1.07×10^{-44}	1.42×10^{-17}	3.52×10^{-3}	$\mathbf{6.40 \times 10^{-1}}$	3.56×10^{-1}
1970	1512	0	0	5.70×10^{-57}	2.29×10^{-43}	4.73×10^{-14}	$\mathbf{5.83 \times 10^{-1}}$	3.44×10^{-1}	7.25×10^{-2}
1975	1610	0	0	6.08×10^{-60}	1.08×10^{-44}	2.98×10^{-22}	1.15×10^{-4}	4.91×10^{-1}	$\mathbf{5.09 \times 10^{-1}}$
1980	1637	0	0	1.40×10^{-76}	1.40×10^{-59}	3.60×10^{-15}	2.23×10^{-6}	2.17×10^{-1}	$\mathbf{7.83 \times 10^{-1}}$
1985	1925	0	0	3.21×10^{-107}	1.22×10^{-84}	2.54×10^{-19}	8.01×10^{-14}	9.46×10^{-2}	$\mathbf{9.05 \times 10^{-1}}$
1990	1903	0	0	1.14×10^{-107}	2.37×10^{-85}	1.12×10^{-18}	6.86×10^{-14}	4.29×10^{-2}	$\mathbf{9.57 \times 10^{-1}}$
1995	1953	0	0	2.93×10^{-107}	6.45×10^{-85}	1.71×10^{-15}	8.34×10^{-11}	2.09×10^{-2}	$\mathbf{9.79 \times 10^{-1}}$
2000	2050	0	0	1.41×10^{-113}	1.36×10^{-89}	2.15×10^{-15}	4.87×10^{-11}	1.13×10^{-2}	$\mathbf{9.89 \times 10^{-1}}$
2008	2157	0	0	1.82×10^{-96}	3.97×10^{-74}	8.03×10^{-15}	5.93×10^{-5}	4.31×10^{-2}	$\mathbf{9.57 \times 10^{-1}}$

TABLE III

AKAIKE WEIGHTS FOR THE DEGREE DISTRIBUTION. VALUES BELOW 10^{-200} ARE ROUNDED TO ZERO. THE LARGEST AKAIKE WEIGHT IN EACH YEAR IS HIGHLIGHTED IN BOLD TYPE.

Year	ports	POIS	GEOM	ZETA	YULE	NEGB	TPOW	DLGN	DWEI
1890	895	0	3.05×10^{-132}	4.96×10^{-37}	1.22×10^{-25}	8.99×10^{-8}	3.62×10^{-4}	$\mathbf{6.21 \times 10^{-1}}$	3.79×10^{-1}
1910	1186	0	3.66×10^{-186}	1.97×10^{-69}	5.02×10^{-53}	1.20×10^{-9}	1.54×10^{-8}	$\mathbf{9.82 \times 10^{-1}}$	1.80×10^{-2}
1915	970	0	9.77×10^{-140}	4.07×10^{-62}	8.23×10^{-49}	1.48×10^{-4}	1.87×10^{-4}	$\mathbf{6.04 \times 10^{-1}}$	3.96×10^{-1}
1920	955	0	8.66×10^{-147}	4.53×10^{-60}	1.01×10^{-46}	5.13×10^{-7}	1.18×10^{-6}	$\mathbf{9.65 \times 10^{-1}}$	3.49×10^{-2}
1925	1170	0	1.79×10^{-188}	3.85×10^{-67}	3.23×10^{-52}	5.22×10^{-5}	3.74×10^{-4}	1.05×10^{-1}	$\mathbf{8.94 \times 10^{-1}}$
1930	1231	0	1.65×10^{-179}	6.11×10^{-89}	1.59×10^{-71}	9.58×10^{-9}	8.04×10^{-9}	$\mathbf{7.94 \times 10^{-1}}$	2.06×10^{-1}
1935	1259	0	3.08×10^{-181}	1.07×10^{-90}	2.22×10^{-72}	5.79×10^{-11}	5.18×10^{-11}	$\mathbf{9.92 \times 10^{-1}}$	8.34×10^{-3}
1940	1273	0	3.41×10^{-196}	1.25×10^{-85}	6.69×10^{-67}	2.44×10^{-13}	5.43×10^{-13}	$> \mathbf{9.99 \times 10^{-1}}$	4.92×10^{-4}
1946	1220	0	3.38×10^{-197}	5.90×10^{-57}	1.13×10^{-41}	1.13×10^{-6}	1.51×10^{-3}	$\mathbf{5.11 \times 10^{-1}}$	4.87×10^{-1}
1951	1294	0	3.85×10^{-180}	3.94×10^{-88}	2.04×10^{-69}	1.19×10^{-9}	1.29×10^{-9}	$\mathbf{9.94 \times 10^{-1}}$	6.48×10^{-3}
1960	1506	0	0	8.78×10^{-105}	1.09×10^{-83}	3.47×10^{-9}	3.26×10^{-9}	$\mathbf{9.57 \times 10^{-1}}$	4.29×10^{-2}
1965	1534	0	0	6.26×10^{-119}	3.03×10^{-97}	2.53×10^{-7}	1.12×10^{-7}	1.18×10^{-1}	$\mathbf{8.82 \times 10^{-1}}$
1970	1487	0	0	2.72×10^{-109}	2.20×10^{-89}	2.71×10^{-4}	1.74×10^{-4}	6.07×10^{-3}	$\mathbf{9.93 \times 10^{-1}}$
1975	1579	0	0	1.92×10^{-103}	3.00×10^{-82}	1.57×10^{-6}	2.00×10^{-6}	2.75×10^{-1}	$\mathbf{7.25 \times 10^{-1}}$
1980	1591	0	9.20×10^{-198}	3.34×10^{-124}	2.44×10^{-102}	4.49×10^{-5}	2.19×10^{-5}	1.25×10^{-4}	$> \mathbf{9.99 \times 10^{-1}}$
1985	1872	0	0	2.18×10^{-150}	1.57×10^{-123}	1.64×10^{-8}	5.10×10^{-9}	6.41×10^{-4}	$\mathbf{9.99 \times 10^{-1}}$
1990	1875	0	0	4.46×10^{-162}	1.44×10^{-136}	1.96×10^{-5}	7.40×10^{-6}	1.15×10^{-8}	$> \mathbf{9.99 \times 10^{-1}}$
1995	1897	0	4.51×10^{-191}	5.31×10^{-176}	3.15×10^{-149}	3.07×10^{-4}	8.99×10^{-5}	6.78×10^{-10}	$> \mathbf{9.99 \times 10^{-1}}$
2000	1969	0	0	2.60×10^{-148}	6.90×10^{-121}	1.84×10^{-3}	9.28×10^{-4}	1.17×10^{-7}	$\mathbf{9.97 \times 10^{-1}}$
2008	2007	0	8.85×10^{-183}	2.55×10^{-127}	1.78×10^{-99}	2.01×10^{-1}	1.61×10^{-1}	9.33×10^{-8}	$\mathbf{6.38 \times 10^{-1}}$

a truncated power law. These two-parameter models always perform better than even the best one-parameter model, which is in all cases the Yule-Simon distribution. The added term $2K_i$ in Eq. 3 for introducing a second parameter is therefore more than compensated by an increased likelihood for the best-performing models.

The effect of including a second parameter can be seen in Fig. 1(a) and (b) where we compare the observed call

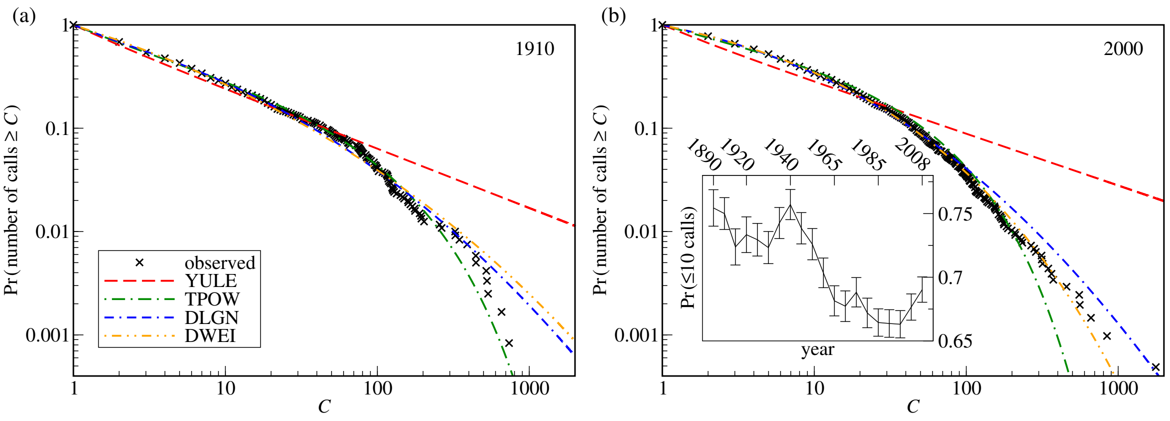

Fig. 1. Observed complementary cumulative call distribution function in (a) 1910, (b) 2000 together with the maximum-likelihood Yule-Simon, truncated power law, discrete lognormal and discrete Weibull distributions. The one-parameter Yule-Simon distribution fits the data far worse than any of the two-parameter alternatives. In 1910 the truncated power law has the highest likelihood among all models. In 2000 the Weibull distribution fits best (Table II). Inset in (b): The fraction of ports with no more than ten ports has dropped from around 75% to approximately 68% during the 1960s. This subtle, but statistically significant decrease is responsible for more curvature in later years on the left-hand side of the observed data. Error bars are jackknife estimates.

distributions in 1910 and 2000 with the maximum-likelihood estimates for the Yule-Simon, truncated power law, lognormal and Weibull distributions. Both observed distributions are substantially curved on a log-log scale and thus difficult to fit by an asymptotic power law such as the Yule-Simon distribution. All other plotted candidate distributions have the flexibility to follow the curvature more accurately. Among these, the truncated power law decays in the limit $k \to \infty$ most rapidly and the lognormal distribution most slowly in the tail.[1]

Comparing the observed call distributions in 1910 and 2000, the most obvious difference is that the initial decay on the left-hand side appears less curved in Fig. 1(a) than in (b). For this reason, the truncated power law that fits well in 1910 is no longer a suitable candidate in 2000. In general, we observed by visual inspection that in the small to medium port range the complementary cumulative call distribution tends to be more curved in later years. This trend is detected by the Akaike weights in Table II that have shifted over the years from the truncated power law to the Weibull distribution. In practice, this change in the distribution implies that there is now a smaller fraction of ports listed with maximally ten calls in one week. The inset in Fig. 1(b) confirms this trend, showing a statistically significant decrease between the years 1890 and 2008 from 75.4% to 69.0% of ports having no more than ten calls.[2]

The interpretation of the degree distribution is a little trickier. The Akaike weights in Table III seem to suggest a clear distinction: before the mid-1960s the most likely model is in all but one case a lognormal, afterwards always a Weibull distribution. However, Figure 2(a) shows that in 1910 the lognormal and Weibull distributions are visually more or less equally good fits. Only in 2000 (Fig. 2b) does the maximum-likelihood Weibull distribution fit clearly better in the tail than the lognormal. Unlike for the call distribution, we do not find a significant trend that the percentage of low-degree ports has decreased (inset in Fig. 2b). In order to shed light on the significance of the apparent trend in the Akaike weights for the degree distribution, we will in the next section compare the performance of the maximum-likelihood models with another statistical technique.

IV. Vuong's likelihood ratio tests

For models with an equal number of parameters, Akaike weights compare the models purely by the differences in their log-likelihood. Proponents of AIC-based model selection have argued that the Akaike weights are sufficient to judge the significance of the best model [16]. However, others argue that the log-likelihood alone does not in itself allow an assessment when we should reject the second-ranked model in favour of the model with the highest Akaike weight [35]. Likelihood ratio tests, on the other hand, can inform us how significant the difference in the log-likelihood is [12].

In this section we investigate the three two-parameter models that achieved maximal Akaike weight in at least one year for either the call or degree distribution: truncated power law, discrete lognormal and Weibull distributions. For these nonnested models we apply the likelihood ratio test devised by Vuong [36]. The test statistic for comparing models r and s is the ratio of their likelihoods from Eq. 2 or equivalently its logarithm

$$R = \ln\left(\prod_{i=1}^{n} \frac{p_r(k_i)}{p_s(k_i)}\right) = \sum_{i=1}^{n} \left(\ln p_r(k_i) - \ln p_s(k_i)\right), \qquad (6)$$

where $p_r(k_i)$ is the probability $\Pr(k_i, \hat{\mathbf{v}}_r)$ assigned to observing degree k_i in model r with the maximum-likelihood parameters $\hat{\mathbf{v}}_r$. If we assume that all observed k_i are independent (as

[1] Depending on the parameters, k may have to be larger than the maximum port size for the lognormal to exceed the Weibull distribution. For example in Fig. 1(a) we are not yet far enough in the asymptotic regime on the right-hand side for the Weibull distribution to fall below the lognormal.

[2] Because the number of ports, however, has more than doubled between 1890 and 2008 (see second column of Table II), the *absolute* number of ports with less than or equal to ten calls has of course still increased.

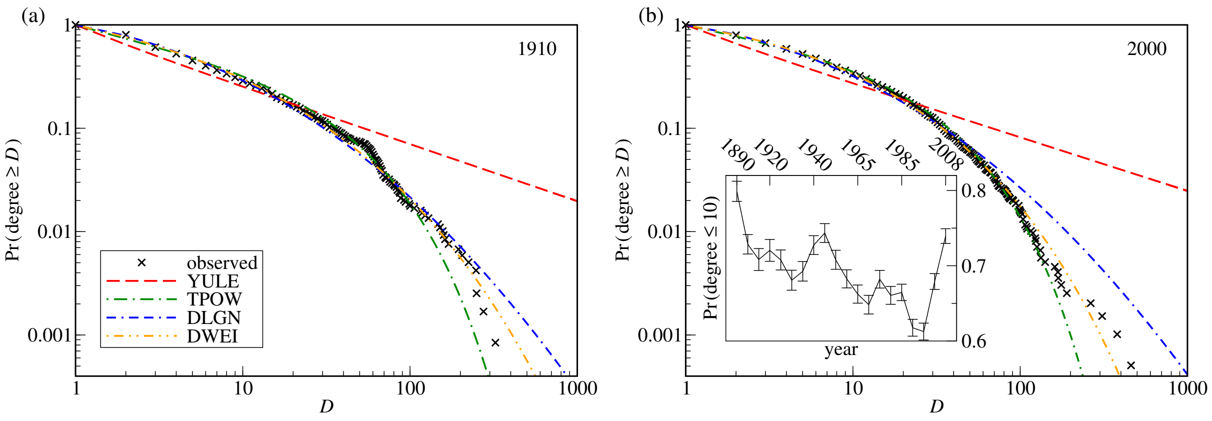

Fig. 2. Observed complementary cumulative degree distribution function in (a) 1910, (b) 2000 together with the maximum-likelihood Yule-Simon, truncated power law, discrete lognormal and discrete Weibull distributions. The highest Akaike weight is achieved by the (a) lognormal, (b) Weibull distribution. Inset in (b): The fraction of ports with degree ≤ 10 does not show a clear trend.

we discussed after Eq. 2), then all terms $\ln p_r(k_i) - \ln p_s(k_i)$ in the sum on the left-hand side of Eq. 6 are also independent. With the shorthand notation $l_i = \ln p_r(k_i) - \ln p_s(k_i)$, the variance of one term in the sum can be estimated as

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} l_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} l_i\right)^2 . \tag{7}$$

For sufficiently large n, the random variable R thus becomes normally distributed with an estimated variance $n\sigma^2$. We can then apply a conventional Z-test to determine whether the observed value R is significantly different from zero given the observed variance. The p-value can be expressed as

$$p = \operatorname{erfc}\left(\frac{|R|}{\sqrt{2n\sigma^2}}\right), \tag{8}$$

where erfc is the complementary error function.

This handwaving derivation, which follows essentially that of Ref. [12], is admittedly oversimplified. For both models r and s in Eq. 6 we have fitted the parameters $\hat{\mathbf{v}}_r$ and $\hat{\mathbf{v}}_s$ to the same data so that there are nontrivial correlations between R and σ^2. However, Ref. [36] proved that Eq. 8 still remains true. One noteworthy point about this equation is the appearance of the variance σ^2. The variance of the data is a crucial piece of evidence whether one of the two models in question is likely to be significantly better. The Akaike weights, by contrast, did not account for the variance.

We list the p-values for all pairwise comparisons between truncated power laws, lognormal and Weibull distributions in Table IV. We highlight in bold type all p-values less than 0.1 and, where the likelihood ratio test indicates a deviation from randomness at this 10% significance level, we list in parentheses the more likely model. As an overall pattern for the call distribution (left half of the table), the tests are for most years indifferent between the three candidate models. However, for the degrees (right half of the table) the test strongly rejects for most years the truncated power law, consistent with its small Akaike weights in Table III. Also in agreement with our earlier findings, the tests favour the Weibull distribution in some of the more recent years for both calls and degrees. There are examples where the Akaike weights suggest a high likelihood for one particular model, yet the likelihood-ratio test does not lend strong support to it. For example, in 2008 the Weibull distribution has an Akaike weight 0.957 for the calls, but, after factoring in the variance in the data, the likelihood ratio test does not reject the possibility that the data could be from a lognormal distribution (Akaike weight 0.043) or even a truncated power law despite its much lower Akaike weight (5.93×10^{-5}).

There is thus for most years no simple answer if the call distribution is better described by a Weibull or lognormal distribution. Over the range of observed calls (i.e. between 1 and approximately 2000 – the precise upper bound of course depends on the year in question) the maximum-likelihood distributions from the lognormal and Weibull family do in fact not differ very much as can be seen for example in Fig. 1(a). Likewise, for the degree distribution there is no clear support in favour of the lognormal hypothesis prior to 1960 (except in 1940) despite generally having the highest Akaike weight. Afterwards, there is increasing evidence in favour of the Weibull distribution which might have to do with an increasing number of ports in the sample that allows us to distinguish more clearly between the models.

One has to bear in mind that neither Akaike weight nor likelihood ratio test can tell us that a model is good in an absolute sense, only that it is more plausible than its competitors. In other words, if all our candidate models are bad, then “in the country of the blind, the one-eyed man is king.” We will now apply two classic goodness-of-fit tests that show that some of our candidate models are indeed a good match for the observations.

V. Kolmogorov-Smirnov and Anderson-Darling tests

The key idea behind both the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) test is to compare the difference

TABLE IV

The p-values for Vuong's likelihood ratio test of the maximum-likelihood distributions. We show all pairwise comparisons between truncated power law, discrete lognormal and discrete Weibull distributions. We highlight p-values smaller than 0.1 in bold type and add in parentheses the more likely distribution. Values below 5×10^{-3} are rounded to zero.

	calls			degrees		
Year	TPOW – DLGN	TPOW – DWEI	DLGN – DWEI	TPOW – DLGN	TPOW – DWEI	DLGN – DWEI
1890	0.83	0.49	0.20	**0.09** (DLGN)	**0.04** (DWEI)	0.67
1910	0.35	0.18	0.18	**0.01** (DLGN)	**0.01** (DWEI)	0.11
1915	0.66	0.22	**0.09** (DLGN)	0.16	**0.05** (DWEI)	0.84
1920	0.47	0.26	0.28	**0.03** (DLGN)	**0.01** (DWEI)	0.16
1925	0.52	0.94	0.34	0.35	**0.09** (DWEI)	0.19
1930	0.25	0.85	0.10	**0.02** (DLGN)	**0.00** (DWEI)	0.64
1935	0.37	0.48	0.94	**0.00** (DLGN)	**0.00** (DWEI)	0.13
1940	0.19	0.15	0.35	**0.00** (DLGN)	**0.00** (DWEI)	**0.02** (DLGN)
1946	0.59	0.66	0.93	0.30	0.18	0.97
1951	0.88	0.84	0.37	**0.01** (DLGN)	**0.00** (DWEI)	0.10
1960	0.64	0.86	0.10	**0.02** (DLGN)	**0.00** (DWEI)	0.33
1965	0.40	0.49	0.68	0.11	**0.00** (DWEI)	0.55
1970	0.92	0.73	0.39	0.64	0.10	**0.07** (DWEI)
1975	0.19	0.22	0.98	0.14	**0.02** (DWEI)	0.72
1980	0.13	**0.08** (DWEI)	0.30	0.84	**0.06** (DWEI)	**0.01** (DWEI)
1985	**0.01** (DLGN)	**0.00** (DWEI)	0.22	0.26	**0.00** (DWEI)	**0.08** (DWEI)
1990	**0.03** (DLGN)	**0.01** (DWEI)	0.18	0.56	0.10	**0.00** (DWEI)
1995	0.12	**0.03** (DWEI)	**0.09** (DWEI)	0.29	0.17	**0.00** (DWEI)
2000	0.16	**0.05** (DWEI)	**0.06** (DWEI)	0.39	0.32	**0.00** (DWEI)
2008	0.70	0.54	0.23	0.19	0.86	**0.00** (DWEI)

between the observed and hypothesized cumulative distribution functions. If the observed degree (or number of calls) are $k_1, k_2, \ldots, k_n$, then the observed cumulative distribution $F_{\text{obs}}(x)$ is the number of ports with $k_i \geq x$ divided by n. The KS statistic is defined by

$$D_{\text{KS}} = \max_{x=1,2,3,\ldots} |F_{\text{obs}}(x) - F_{\text{model}}(x)|, \tag{9}$$

where F_{model} is the cumulative distribution function of the model to be tested [37]. In words, D_{KS} is the maximum absolute difference between observed and hypothesized cumulative distribution function for any possible value x. While D_{KS} has a very intuitive interpretation, it also has one shortcoming when applied to heavy-tailed distributions: $|F_{\text{obs}}(x) - F_{\text{model}}(x)|$ is typically maximized where $F_{\text{model}}(x) \approx 0.5$ and therefore D_{KS} does not effectively sample the tail where $F_{\text{model}}(x)$ is close to 1. This phenomenon can be understood as follows. If the model were correct and the difference between F_{obs} and F_{model} only the consequence of random chance, the jackknife estimate of the standard deviation in the difference is $\sqrt{F_{\text{model}}(x)[1 - F_{\text{model}}(x)]/(n-1)}$, which has a maximum at $F_{\text{model}}(x) = 1/2$.

There is one obvious cure to this problem: we divide the difference to be maximized in Eq. 9 by the expected standard deviation,

$$D_{\text{AD}} = \max_{x=1,2,3,\ldots} \frac{|F_{\text{obs}}(x) - F_{\text{model}}(x)|}{\sqrt{F_{\text{model}}(x)[1 - F_{\text{model}}(x)]}}, \tag{10}$$

where we dropped the term $\sqrt{n-1}$ because it is independent of x. D_{AD} is called the Anderson-Darling statistic [38]. We have decided to carry out tests for both D_{KS} and D_{AD} because these two statistics measure different features of the distribution. A good model should be able to have small values of D_{KS} as well as D_{AD}.

We test the significance of the same three two-parameter models as in the likelihood ratio test (truncated power law, lognormal and Weibull distribution), but also include for comparison the Yule-Simon distribution which the Akaike weights identified as the best one-parameter model. We calculate p-values with Monte Carlo simulations based on the following algorithm.

First, we determine for a given model the maximum-likelihood parameters $\hat{\mathbf{v}}_{\text{obs}}$ that fit the Lloyd's Shipping Index data best. For the model distribution with parameters $\hat{\mathbf{v}}_{\text{obs}}$ we calculate the observed test statistics $D_{\text{KS, obs}}$ and $D_{\text{AD, obs}}$. Next we generate n random numbers drawn from the model distribution with parameters $\hat{\mathbf{v}}_{\text{obs}}$. We then pretend that we do not know $\hat{\mathbf{v}}_{\text{obs}}$ and determine the maximum-likelihood parameters $\hat{\mathbf{v}}_{\text{rnd}}$ that fit the random numbers best. In general, $\hat{\mathbf{v}}_{\text{rnd}}$ will differ slightly from $\hat{\mathbf{v}}_{\text{obs}}$. From the difference between the random numbers (now treated as surrogate observation) and the model with $\hat{\mathbf{v}}_{\text{rnd}}$ we calculate $D_{\text{KS, rnd}}$ and $D_{\text{AD, rnd}}$. We repeat drawing n random numbers 10^5 times and estimate the p-value p_{KS} for the KS test by the fraction of runs with $D_{\text{KS, rnd}} \geq D_{\text{KS, obs}}$. The same calculations are also carried out for the AD statistic. The repeated calculation of $\hat{\mathbf{v}}_{\text{rnd}}$ slows down the simulation, but is necessary to mimic the steps in the calculation of $D_{\text{KS, obs}}$ and $D_{\text{AD, obs}}$. Otherwise we obtain p-values with a strong downward bias that would lead us to accept the null hypothesis (i.e. that the real data follows the model distribution) more often than truly justified [12].

The p-values are listed in Table V. The highlighted entries in bold type are those cases where there is no reason to suspect at the 10% significance level that the model is wrong, neither in terms of the KS nor the AD statistic. It is striking that the Yule-Simon distribution fails as a null hypothesis for call and degree distributions in all years. The truncated power law is accepted only for the call distribution (left half of the table) and mostly in the early years of our data base. By contrast, the lognormal distribution is a suitable null hypothesis for the call distribution in all except one year (1995), and even then the null hypothesis would not be rejected at a 5% significance level. For the degree distribution (right half of the table), a lognormal null hypothesis is accepted in most, but not all years. Especially in the later years, the Weibull distribution shows better performance than the lognormal, confirming the trends we observed in the Akaike weights and the likelihood ratio tests. However, the KS and AD tests in earlier years only sporadically support a Weibull distribution.

VI. Discussion

The overall picture that emerges from the KS and AD tests is a surprisingly consistent performance of the lognormal model for the call distribution. The only rejection, namely by the KS test for the data of 1995, could plausibly be by random chance. After all, at the chosen 10% significance level it is likely that at least one false positive exists among the twenty years which we have tested. The lognormal hypothesis also gains support from a recent analysis of world container port throughput [39] that reported a good fit between the number of containers handled at 300 top ports and a lognormal distribution. There is also a simple mechanistic model that could explain how a lognormal distribution might come about: Gibrat's law of proportionate growth [40]. It is in principle possible to carry out further tests whether Lloyd's Shipping Index supports the key principle behind Gibrat's law, namely that the growth rates in calls are independent of the number of calls. Such a test will make more stringent demands on the data quality than what we currently have available. Right now such an effort would be hampered, for example, by the irregular time intervals between the samples. As more data becomes available, however, an analysis of port growth rates is clearly an intriguing research direction.

For the time being, we can instead view the call distribution from yet another angle. As an alternative to plotting the complementary distribution function directly as we did in Fig. 1, economists frequently employ Lorenz curves [41] to visualize inequality in distributions. Translated to our application, the Lorenz curve $y(x)$ shows the percentage of ship calls that were made at the x percent of lowest ranked ports (ranked by the number of calls). We plot the Lorenz curves for three representative years in Fig. 3(a). If all ports had an equal number of calls, the Lorenz curve would be the dashed diagonal line. One measure of inequality is the area between this diagonal and the actually observed Lorenz curve: the more unequal the distribution, the larger this area. Multiplied by two, this measure is known as Gini coefficient [42]. The coefficient itself as well as a jackknife estimate of its standard error can be conveniently calculated with ordinary least-squares regression [43]. The results for all years in our data base (Fig. 3b) reveal that the Gini coefficient for the calls has decreased from 0.80 in 1890 to 0.74 in 2008. Although this is a subtle decline, it is statistically significant: the values are more than five standard errors apart.

Why is the inequality declining? The maximum number of calls has increased (from 822 in 1890 to 2422 in 2008), which at first glance suggests rather an increasing inequality. The resolution to this apparent paradox lies in the inset of Fig. 1(b): the fraction of small ports with ≤ 10 calls has decreased. As the total number of ports has grown over the years, an overproportional number of new medium-sized ports were added to the network. Together with a flattening global hierarchy this has reduced the gap between core and periphery, thereby making the network more polycentric. This trend more than compensates the growth in maximum port size and has led to an overall decrease in the Gini coefficient.

There is more than just the Gini coefficient that we can infer from the Lorenz curve. One complementary measure is the Lorenz asymmetry coefficient (LAC) [44]. A Lorenz curve is defined to be symmetric if it has the same slope as the diagonal "line of equality" (i.e. a slope of 1) at the point where the curve and the antidiagonal line $y = 1 - x$ (i.e. the dotted line in Fig. 3a) intersect. One can show that for a continuous cumulative distribution F with mean μ the slope equals 1 at $x = F(\mu)$. At this point $y = \int_0^\mu (x/\mu)\, dF(x)$, so a criterion for symmetry is $\text{LAC} \equiv F(\mu) + \int_0^\mu (x/\mu)\, dF(x) = 1$.[3] If $\text{LAC} < 1$, the Lorenz curve is skewed such that it has slope 1 below the dotted antidiagonal. Conversely, if $\text{LAC} > 1$, the Lorenz curve is parallel with the line of equality above the antidiagonal symmetry axis.

The LAC is of interest because curves with the same Gini coefficient can have different asymmetries. If $\text{LAC} < 1$, the inequality in the distribution is caused by a large gap between a roughly equal number of small and large ports. By contrast, if $\text{LAC} > 1$, the inequality is due to a small number of very busy ports, whereas the majority of ports experiences approximately equally low traffic. The base case is a lognormal distribution where $\text{LAC} = 1$ regardless of the parameters μ and σ [44]. For the call distribution, we see in Fig. 3(b) that $\text{LAC} = 1$ is always included in the error bar which represents a jackknife estimate of the standard deviation. This observation gives additional credence to the lognormal distribution as a working hypothesis for ship calls.

[3] Strictly speaking, this is the definition only for a continuous distribution F. For discrete distributions the Lorenz curve is a polygon instead of a smooth curve so that there is typically no point where the slope is exactly 1. However, one can generalize the definition so that it still works for the discrete distributions obtained from finite samples, see Ref. [44] for details.

TABLE V

THE p-VALUES FOR THE KOLMOGOROV-SMIRNOV TEST p_{KS} AND THE ANDERSON-DARLING TEST p_{AD}. VALUES BELOW 5×10^{-3} ARE ROUNDED TO ZERO. WE HIGHLIGHT THOSE DISTRIBUTIONS IN BOLD TYPE WHERE THE NULL HYPOTHESIS (I.E. THAT THE DATA IS GENERATED BY THE MODEL) IS NOT REJECTED AT A 10% SIGNIFICANCE LEVEL.

Year	calls								degrees							
	YULE		TPOW		DLGN		DWEI		YULE		TPOW		DLGN		DWEI	
	p_{KS}	p_{AD}	p_{KS}	p_{AD}	p_{KS}	p_{AD}	p_{KS}	p_{AD}	p_{KS}	p_{AD}	p_{KS}	p_{AD}	p_{KS}	p_{AD}	p_{KS}	p_{AD}
1890	0.00	0.06	**0.22**	**0.37**	**0.98**	**0.78**	0.04	0.38	0.00	0.02	0.00	0.11	**0.93**	**0.63**	**0.34**	**0.47**
1910	0.00	0.04	**0.22**	**0.27**	**0.72**	**0.46**	0.01	0.18	0.00	0.00	0.00	0.03	0.07	0.17	0.00	0.10
1915	0.00	0.04	**0.24**	**0.21**	**0.81**	**0.69**	0.01	0.20	0.00	0.00	0.00	0.06	**0.50**	**0.43**	0.03	0.31
1920	0.00	0.05	**0.51**	**0.58**	**0.95**	**0.80**	0.04	0.37	0.00	0.00	0.00	0.03	**0.18**	**0.37**	0.01	0.11
1925	0.00	0.03	0.04	0.06	**0.96**	**0.82**	0.07	0.40	0.00	0.00	0.00	0.07	**0.78**	**0.44**	**0.29**	**0.55**
1930	0.00	0.03	0.02	0.03	**0.84**	**0.70**	0.01	0.26	0.00	0.00	0.00	0.02	**0.69**	**0.55**	0.02	0.28
1935	0.00	0.03	0.00	0.04	**0.91**	**0.77**	**0.42**	**0.49**	0.00	0.00	0.00	0.03	**0.35**	**0.51**	0.00	0.10
1940	0.00	0.04	**0.12**	**0.40**	**0.61**	**0.64**	0.03	0.17	0.00	0.00	0.00	0.02	**0.11**	**0.20**	0.00	0.06
1946	0.00	0.03	0.02	0.02	**0.31**	**0.22**	0.03	0.10	0.00	0.00	0.00	0.06	**0.30**	**0.30**	0.01	0.24
1951	0.00	0.03	0.03	0.06	**0.78**	**0.68**	0.08	0.33	0.00	0.00	0.00	0.03	**0.11**	**0.45**	0.00	0.08
1960	0.00	0.01	0.11	0.04	**0.55**	**0.65**	0.01	0.25	0.00	0.00	0.00	0.02	0.03	0.22	0.00	0.09
1965	0.00	0.00	0.02	0.02	**0.88**	**0.49**	0.08	0.48	0.00	0.00	0.00	0.03	**0.49**	**0.29**	0.06	0.42
1970	0.00	0.00	0.07	0.03	**0.84**	**0.54**	0.00	0.23	0.00	0.00	0.00	0.07	**0.23**	**0.25**	0.09	0.53
1975	0.00	0.00	0.00	0.02	**0.87**	**0.63**	**0.37**	**0.47**	0.00	0.00	0.00	0.04	**0.22**	**0.34**	0.02	0.31
1980	0.00	0.00	0.00	0.01	**0.60**	**0.52**	**0.20**	**0.75**	0.00	0.00	0.00	0.04	**0.23**	**0.14**	**0.21**	**0.75**
1985	0.00	0.00	0.00	0.00	**0.79**	**0.88**	**0.94**	**0.54**	0.00	0.00	0.00	0.01	**0.41**	**0.18**	**0.85**	**0.64**
1990	0.00	0.00	0.00	0.00	**0.26**	**0.46**	**0.26**	**0.24**	0.00	0.00	0.00	0.00	0.02	0.09	**0.34**	**0.58**
1995	0.00	0.00	0.00	0.00	0.06	0.25	0.02	0.40	0.00	0.00	0.01	0.01	0.01	0.06	**0.10**	**0.42**
2000	0.00	0.00	0.00	0.00	**0.60**	**0.25**	**0.56**	**0.16**	0.00	0.00	0.00	0.00	0.09	0.07	**0.16**	**0.34**
2008	0.00	0.00	0.04	0.00	**0.31**	**0.11**	0.04	0.10	0.00	0.00	0.05	0.00	0.03	0.06	0.00	0.18

While the lognormal model is hence a generally promising candidate for the call distribution, it is not equally good for the degree distribution because LAC < 1 in the later years (Fig. 3b). These numbers confirm the results from Tables III–V showing that the lognormal distribution is in those years not the best model for the degrees. There is, however, no immediate contradiction between this finding and a lognormal call distribution. The relationship between the call and degree distributions is complicated: regardless whether one or multiple voyages are made between the same two ports, it always only adds 1 to the ports' degrees. In other words, the call distribution is a measure of the weighted multigraph of voyages, whereas the degree distribution is derived from the unweighted network that forms a so-called simple graph. The collapse of multiple voyages into one unweighted link can conceivably change the distribution so that it appears to come from a completely different probabilistic law. In future analysis, we will explore with suitable null models (e.g. random graphs with a fixed weighted degree sequence [45]) how the degree distribution changes when a heavy-tailed (in particular lognormal) multigraph is mapped to a simple graph and if this may explain our observations for the cargo ship data.

VII. CONCLUSION

We have statistically analyzed call and degree distributions of the cargo ship network extracted from snapshots of Lloyd's Shipping Index in twenty different years between 1890 and 2008. We have applied information-based model selection and statistical hypothesis tests to quantify the empirical distribution in a mathematically principled manner. For the call distribution a lognormal null model passes the Anderson-Darling goodness-of-fit test in all years and the Kolmogorov-Smirnov test in 19 out of 20 years at a 10% significance level. In some early years the Akaike weight is higher for truncated power laws than a lognormal distribution; in later years a Weibull distribution is preferred. However, Vuong's likelihood ratio test does not reject the lognormal null hypothesis at a 5% significance level for any tested year, neither compared with the maximum-likelihood truncated power law nor Weibull distribution. When the empirical call distribution is replaced by the degree distribution, the lognormal model is plausible in early years, but in later years Weibull distributions fit the data better.

As in all model selection problems, one should bear in mind that reality is of course more complex than any of the candidate models. In our case, it might be possible to reduce the AIC further by allowing more than two parameters, but we feel that two parameters are a good compromise between simplicity of the model and goodness-of-fit. With additional data it might become possible to analyze the dynamics of port calls in more detail, especially to test if Gibrat's law applies in our case, which could explain a lognormal call distribution. It may also become feasible to test the assumption of inde-

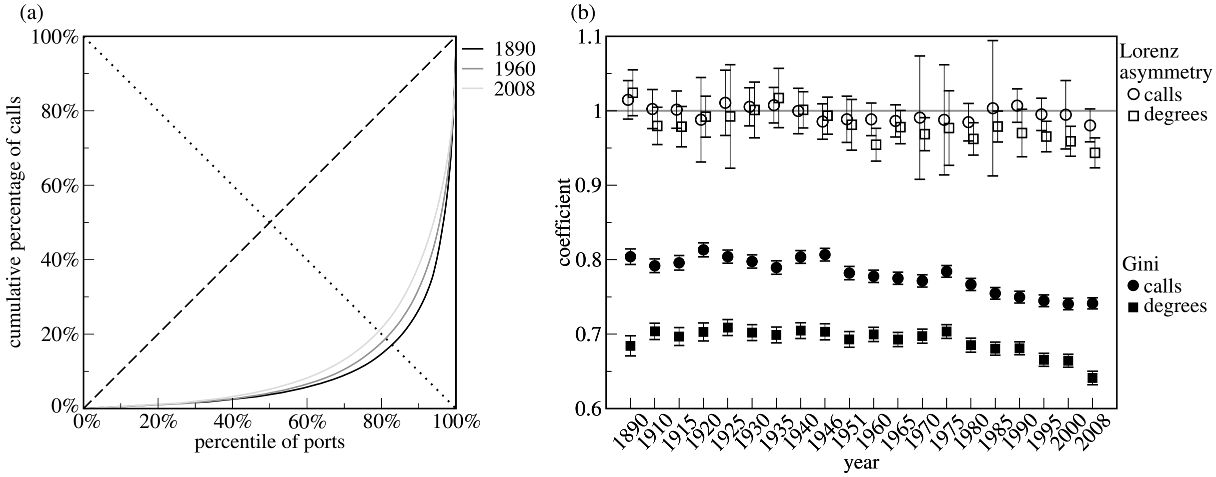

Fig. 3. (a) Lorenz curves for the call distributions in three exemplary years (black: 1890, dark grey: 1960, light grey: 2008). The Gini coefficient measures inequality as twice the area between the Lorenz curve and the line of equality (dashed diagonal). The Lorenz asymmetry coefficient (LAC) measures whether curves have a slope of 1 at the intersection with the dotted antidiagonal line (LAC < 1 if the slope equals 1 below the antidiagonal, LAC > 1 if the slope equals 1 above it). For the three curves in the figure the Gini coefficient has decreased over time, but they all have a slope close to 1 at the symmetry axis. (b) Gini coefficients and LACs for all twenty years for which we have data. Error bars represent jackknife estimates of the standard deviation. The Gini coefficient for the call and degree distribution both show a slightly decreasing tendency that is statistically significant, similarly the LAC for the degrees. The LAC for the call distribution, however, shows no significant deviation from 1 in any of the investigated years.

pendence between ports and apply full likelihood methods to the dynamics of the network [46].

Acknowledgment

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. [313847] "World Seastems". M. T. G. thanks the European Commission for financial support (project number FP7-PEOPLE-2012-IEF 6-4564/2013).

References

[1] M. E. J. Newman, *Networks: An introduction*. Oxford: Oxford University Press, 2010.

[2] D. Wei-Bing, G. Long, L. Wei, and C. Xu, "Worldwide marine transportation network: efficiency and container throughput," *Chinese Phys. Lett.*, vol. 26, p. 118901, 2009.

[3] Y. Hu and D. Zhu, "Empirical analysis of the worldwide maritime transportation network," *Physica A*, vol. 388, pp. 2061–2071, 2009.

[4] P. Kaluza, A. Kölzsch, M. T. Gastner, and B. Blasius, "The complex network of global cargo ship movements," *J. Roy. Soc. Interface*, vol. 7, pp. 1093–1103, 2010.

[5] C. Ducruet, "Network diversity and maritime flows," *J. Transp. Geogr.*, vol. 30, pp. 77–88, 2013.

[6] United Nations Conference on Trade and Development, *Review of Maritime Transport 2013*, available at unctad.org/en/publicationslibrary/rmt2013_en.pdf

[7] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[8] B. Tadić, "Temporal fractal structures: origin of power laws in the world-wide web," *Physica A*, vol. 314, pp. 278–283, 2002.

[9] C. Moore, G. Ghoshal, and M. E. J. Newman, "Exact solutions for models of evolving networks with addition and deletion of nodes," *Phys. Rev. E*, vol. 74, p. 036121, 2006.

[10] A. G. Wilson, "A statistical theory of spatial distribution models," *Transport. Res.*, vol. 1, pp. 253–269, 1967.

[11] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. Mod. Phys.*, vol. 74, pp. 47–97, 2002.

[12] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, pp. 661–703, 2009.

[13] A. M. Edwards *et al.*, "Revisiting Lévy flight search patterns of wandering albatrosses, bumblebees and deer," *Nature*, vol. 449, pp. 1044–1047, 2007.

[14] W. Willinger, D. Alderson, and J. C. Doyle, "Mathematics and the Internet: a source of enormous confusion and great potential," *Not. Am. Math. Soc.*, vol. 56, pp. 586–599, 2009.

[15] M. P. H. Stumpf and M. A. Porter, "Critical truths about power laws," *Science*, vol. 335, pp. 665–666, 2012.

[16] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference*. New York: Springer, 1998.

[17] R. J. McCalla, "From 'Anyport' to 'Superterminal'," in *Shipping and Ports in the Twenty-first Century*, D. Pinder and B. Slack, Eds. London: Routledge, 2004, pp. 123–142.

[18] K. P. B. Cullinane and M. Khanna, "Economies of scale in large containerships: optimal size and geographical implications," *J. Transp. Geogr.*, vol. 8, pp. 181–195, 1999.

[19] D. M. Bernhofen, Z. El-Sahli, and R. Kneller, "Estimating the effects of the container revolution on world trade," Lund University Working Paper, Department of Economics, School of Economics and Management, 2013.

[20] H. A. Simon, "On a class of skew distribution functions," *Biometrika*, vol. 42, pp. 425–440, 1955.

[21] M. P. H. Stumpf, P. J. Ingram, I. Nouvel, and C. Wiuf, "Statistical model selection methods applied to biological networks," in *Lecture notes in computer science*, vol. 3737, *Transactions on computational systems biology III*, C. Priami, E. Merelli, P. Gonzalez, and A. Omicini, Eds. Berlin: Springer, 2005, pp. 65–77.

[22] K. Bhattacharya, G. Mukherjee, J. Saramäki, K. Kaski, and S. S. Manna, "The international trade network: weighted network analysis and modelling," *J. Stat. Mech.*, vol. 2008, p. P02002, 2008.

[23] V. Gómez, A. Kaltenbrunner, and V. López, "Statistical analysis of the social network and discussion threads in Slashdot," *Proc. 17th Int. Conf. World Wide Web*, pp. 645–654, 2008.

[24] A. Todor, A. Dobra, and T. Kahveci, "Uncertain interactions affect degree distribution of biological networks," *IEEE Int. Conf. Bioinformatics and Biomedicine*, pp. 457–461, 2012.

[25] J. Laherrere and D. Sornette, "Stretched exponential distributions in nature and economy: fat tails with characteristic scales," *Eur. Phys. J. B*, vol. 2, pp. 525–539, 1998.

[26] A. Broido and kc claffy, "Internet topology: connectivity of IP graphs," *Proc. SPIE*, vol. 4526, pp. 172–187, 2001.

[27] Y. He, G. Siganos, M. Faloutsos, and S. Krishnamurthy, "A systematic framework for unearthing the missing links: measurements and impact,"

Proc. 4th USENIX Conf. Networked Systems Design & Implementation, p. 14, 2007.

[28] L. E. C. Rocha, F. Liljeros, and P. Holme, "Information dynamics shape the sexual networks of Internet-mediated prostitution," *Proc. Nat. Acad. Sci.*, vol. 107, pp. 5706–5711, 2010.

[29] J. Eeckhout, "Gibrats law for (all) cities," *Am. Econ. Rev.*, vol. 94, pp. 1429–1451, 2004.

[30] D. R. Cox and N. Reid, "A note on pseudolikelihood constructed from marginal densities," *Biometrika*, vol. 91, pp. 729–737, 2004.

[31] C. Varin, N. Reid, and D. Firth, "An overview of composite likelihood methods," *Stat. Sinica*, vol. 21, pp. 5–42, 2011.

[32] H. Akaike, "A new look at the statistical model identification," *IEEE T. Automat. Contr.*, vol. 19, pp. 716–723, 1974.

[33] D. T. Hamilton, M. S. Handcock, and M. Morris, "Degree distributions in sexual networks: a framework for evaluating evidence," *Sex. Transm. Dis.*, vol. 35, pp. 30–40, 2008.

[34] F. Prieto and J. M. Sarabia, "Fitting the degree distribution of real-world networks," *Int. J. Complex Systems in Science*, vol. 1, pp. 129–133, 2011.

[35] P. A. Stephens, S. W. Buskirk, G. D. Hayward, and C. Martínez del Rio, "Information theory and hypothesis testing: a call for pluralism," *J. Appl. Ecol.*, vol. 42, pp. 4–12, 2005.

[36] Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica*, vol. 57, pp. 307–333, 1989.

[37] A. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *G. Ist. Ital. Attuari*, vol. 4, pp. 83–91, 1933.

[38] T. W. Anderson and D. A. Darling, "Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes," *Ann. Math. Stat.*, vol. 23, pp. 193–212, 1952.

[39] D. Ding and C.-P. Teo, "World container port throughput follows lognormal distribution," *Maritime Policy & Management*, vol. 37, pp. 401–426, 2010.

[40] R. Gibrat, *Les inégalités économique*. Paris: Librairie du Recueil Sirey, 1931.

[41] M. O. Lorenz, "Methods of measuring the concentration of wealth," *Publ. Am. Stat. Assoc.*, vol. 9, pp. 209–219, 1905.

[42] C. Gini, *Variabilità e mutabilità.* Bologna: P. Cuppini, 1912.

[43] D. E. A. Giles, "Calculating a standard error for the Gini coefficient: some further results," *Oxford B. Econ. Stat.*, vol. 66, pp. 425–433, 2004.

[44] C. Damgaard and J. Weiner, "Describing inequality in plant size or fecundity," *Ecology*, vol. 91, pp. 1139–1142, 2000.

[45] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random Struct. Algor.*, vol. 6, pp. 161–180, 1995.

[46] C. Wiuf, M. Brameier, O. Hagberg, and M. P. H. Stumpf, "A likelihood approach to analysis of network data," *Proc. Nat. Acad. Sci*, vol. 103, pp. 7566–7570, 2006.