Data-driven stochastic simulation leading to the allometric scaling laws in complex systems

Yuh Kobayashi [*]
Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8502, Japan

Hideki Takayasu [†]
Sony Computer Science Laboratories, Tokyo 141-0022, Japan

Shlomo Havlin [‡]
Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel

Misako Takayasu [§]
Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8502, Japan

(Received 22 March 2022; revised 15 September 2022; accepted 19 September 2022; published 7 December 2022)

We propose a data-driven stochastic method that allows the simulation of a complex system's long-term evolution. Given a large amount of historical data on trajectories in a multi-dimensional phase space, our method simulates the time evolution of a system based on a random selection of partial trajectories in the data without detailed knowledge of the system dynamics. We apply this method to a large data set of time evolution of approximately one million business firms for a quarter century. Accordingly, from simulations starting from arbitrary initial conditions, we obtain a stationary distribution in three-dimensional log-size phase space, which satisfies the allometric scaling laws of three variables. Furthermore, universal distributions of fluctuation around the scaling relations are consistent with the empirical data.

DOI: 10.1103/PhysRevE.106.064304

I. INTRODUCTION

Studying the time evolution of complex systems without well-established first principles is a difficult task. For such systems, researchers usually adopt models that significantly simplify the phenomenon using strong assumptions on the system by inferring the first principles. Examples of well-known dynamical systems employed for this type of modeling include the Lotka-Volterra equation for predator-prey interaction in ecological communities and the SIR model for epidemics in biological and human populations [1,2]. Another strong assumption, which is often implicitly made when phase-space reconstruction [3,4] is applied to empirical data, is that the system dynamics are essentially deterministic. Stochastic and probabilistic models have also been used to analyze a variety of complex systems: financial markets are analyzed using the Ising model [5] and the spread of information and pathogens is modelled as random walks on networks [6]. Although these simplistic assumptions and models facilitate rigorous mathematical analysis and intuitive understanding of the model's inner workings, the applicabilities of the models tend to be unclear. Because radically different mechanisms can lead to similar probability distributions [7] and spatial patterns [8], the empirical verification of the proposed mechanism at microscopic scales is an essential part of assessing the veracity of a model. However, some models of this type can be solely tested by whether they agree with empirical data of macroscopic patterns, owing to the lack of data at microscopic levels, notably in ecological [9] and social [10] phenomena. Furthermore, a moderate modification to some well-known models sometimes leads to entirely different behaviors of their solutions (e.g., the varying number of coexisting species in ecological "niche" models [11,12] and the degree distribution of complex networks, which is sensitive to how the preferential attachment is formulated [13–15]). For complex systems without the established knowledge of underlying mechanisms, introducing a specific mathematical model might induce biases in modeling and predictions.

In contrast, prediction methods based on large-scale data sets such as machine learning have flourished in the last decades, largely owing to the fast advancements of computational power. Recent successes in the application of machine learning techniques in various fields has also inspired the interests by physicists, and led to numerous applications, for

[*]Also at Department of Mathematical Sciences, College of Science and Engineering, Aoyama Gakuin University, Sagamihara 252-5258, Japan; kobayashi@math.aoyama.ac.jp
[†]Also at Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8502, Japan.
[‡]Also at Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8502, Japan.
[§]Corresponding author: takayasu.m.aa@m.titech.ac.jp

example in condensed matter physics, from the automatic discovery of order parameter of matter to the acceleration of classical and quantum simulations of molecules such as proteins [16]. This type of approach has also been tried for the time evolution of complex systems, for example in weather forecasts [17] and projections of COVID-19 epidemic [18]. The same trend has led researchers to increasingly employ the method of analogs [19,20], which nonparametrically predicts the future state of a system directly from a large amount of past data in economic projections [21–23] as well as in weather forecasts in combination with numerical weather prediction [24,25]. It can be argued that the method of analogs is particularly well suited to the prediction of the future states of complex systems without well-established first principles [26], because it can predict the state of a system in the near future without detailed knowledge on its internal structures and interactions with other systems or the environment, which are often intractably entangled.

Here we further explore the potential of the method of analogs by repeatedly using this method to obtain a long-term simulation of a system's time evolution in a multidimensional phase space. The method of analogs is conventionally concerned with the time series data of a system—or an ensemble of systems—that are believed to be essentially deterministic (see Refs. [26–29] for considerations in relation to dynamical systems). However, we deem the set of analogs as approximating the true ensemble of *stochastic* time evolution, and attempt to simulate the evolution of system distributions in a phase space. Although the mathematical considerations of our approach were presented a few decades ago [30–33], this is arguably an approach that has not been tested and evaluated well with the performance on empirical data. Although a similar approach was considered in the context of a long-term climate simulation [34], it is challenging to apply this methodology to other complex systems with scale-free property and possibility of nonstationarity. The problem is even more difficult in the case that partial data are missing, which often occurs in empirically observed systems.

Accompanying the fast growth of computational power, allometric scaling with a fractional power-law exponent between system-size measures has been uncovered in large-scale data sets of complex systems such as business firms [35–40] and metropolises [41,42], in addition to the conventional example of animal bodies [43–46] and biodiversity in natural ecosystems [47,48]. Although plausible explanations of the power-law scaling have been proposed for some of the systems [49,50], origins of such scaling relations remain unclear for others, including business firms. We argue that simulating the time evolution of these systems are an important step toward understanding the systems. Therefore, in this study, we attempt to simulate the stochastic time evolution of business firms in a three-dimensional phase space of logarithms of system-size measures using our repetitive version of the method of analogs. Accordingly, we rely on a large-scale data set that describes the status of approximately one million firms during a quarter-century period. We determine that the stationary distribution of our simulation is surprisingly consistent with the empirical distribution of firms on the phase space when adjusting the effect of the nonstationary increase in the yearly data.

The remainder of this paper is organized as follows. In Sec. II A, we briefly introduce the systems studied here, namely business firms, and examine a few properties of the stochastic time evolution relative to previously reported allometric scalings. Next, we present a stochastic version of the method of analogs in Sec. II B and consider the assumptions that might affect the accuracy of our simulation method in Sec. II C. Subsequently, we investigate the transient system-size distributions of business firms in our simulations to ensure that the empirical system-size distributions are well approximated by the stationary distribution in our simulation (Sec. II D). We address the agreements and disparities between simulation results and empirical data relative to allometric scalings in Sec. II E. Then, we discuss our results to conclude the study in Sec. III. In Appendices A and B, we present further details on processing our empirical data and normalizing the number of trading partners, respectively.

II. RESULTS

A. System under study

As a real-world example of large-scale complex systems, we adopt business firms for our case study. A business firm is a typical complex system: it satisfies the definition set in Ref. [51], as it is "a system built from a large number of nonlinearly interacting constituents," i.e., individuals, "which exhibits collective behavior and, due to an exchange of energy or information," including goods, services, and money in this case, "with the environment, can easily modify its internal structure and patterns of activity." Firms are known to exhibit hierarchical structures [52,53], a range of nontrivial power-law scaling [37,38,40,54], and supposedly adaptive behaviors [55]. Business firms in a country, connected by an interfirm trading network, also collectively comprise a complex system characterized by nonlinear interactions such as nontrivial power laws in the money flow [56] and nonlinear preferential attachment in mergers and acquisitions [57]. Besides being a complex network as the "backbone" of a complex system [58], the interfirm trading network in Japan has been reported to exhibit several properties that have often been considered typical [51,59,60] for real-world networks, such as a heavy-tailed degree distribution approximated by a power law [61], a small-world property characterized by a short distance between two arbitrary nodes [62], and a modular structure with multiple communities [63].

The data employed in this study are provided by Teikoku Databank, Ltd., Japan (hereafter TDB) and describe the annual status of business $\approx 10^6$ firms incorporated in Japan during a 25-year period. In particular, they include the list of trading partners, along with the annual sales and number of employees. Therefore, we can consider the system to be a complex network with $\approx 10^6$ nodes of firms and up to 4×10^6 links of trading relations. See Appendix A for further details regarding the provided data and our data compilation in this study. It is important to note that the data are sometimes missing, although only partially. The processing of such missing data will be comprehensively discussed in Sec. II B.

Firms have been typically characterized by power-law distributions of system size. Whether it is measured by the

annual sales, number of employees, total assets or count of trading relationships with other business firms, a firm's size is usually distributed along several orders of magnitude and approximately follows a power-law distribution in exhaustive data sets [35,40,61,64–70]. Note that the number of trading relationships in this case amounts to the sum of in- and out-degrees of a node in the interfirm trading network. Therefore, a power-law distribution of the number of interfirm trading relationships implies that the network of trading and transactions between firms is scale-free. In the empirical data [40], the firm size distributions barely change in the timescale of one year and are very robust against fluctuations in the economic climate.

Less widely known is the nontrivial scaling relationships of the form, $x \propto y^{\gamma}$, between different measures of a firm's system size, such as the annual sales, numbers of employees, and number of business trading partners [35–40]. In a formal notation, the distribution of size by one measure (here y) normalized by $\langle y|x\rangle$, the average value of y conditional on the other size measure (x), is independent of the system size measured with x. In other words,

$$y/\langle y|x\rangle \perp\!\!\!\perp x,$$

where the symbol $\perp\!\!\!\perp$ represents the independence between two stochastic variables. Owing to the power-law scaling relationship,

$$\langle y|x\rangle \propto x^{\gamma},$$

the conditional probability of y can be described as

$$\mathrm{P}(y|x) = \frac{1}{x^{\gamma}}\widetilde{P}\left(\frac{y}{x^{\gamma}}\right), \tag{1}$$

using a single probability function $\widetilde{P}$. We refer to the random variable defined by this function as a universal probability function of fluctuation. Such functions also exhibit a power-law tail in empirical data [37,40].

A power-law scaling implies that the observed system is relatively densely distributed on a "scaling line" in the phase space of log-transformed size measures. In Fig. 1(a), the scaling line for the three-dimensional phase space of logarithms of the number of trading partners (k), number of employees (ℓ), and annual sales (s) in million yen is depicted as an orange line. Note that these three measures of size have been particularly well studied as the joint distributions of firm size measures fit into simple mathematical descriptions [40]. For example, their density functions have virtually no jump discontinuity, except at very low values where the ratio between consecutive integers is not approximated by 1.0. Power-law scaling relationships, $\ell \propto k^{1.0}$, $s \propto k^{1.2}$, and $s \propto \ell^{1.2}$, can be considered as two-dimensional projections of a single three-dimensional scaling. This suggests that a firm either returns toward the line or simply disappears after it deviates from the scaling. In our previous study [55], it was verified that the scaling relationships are maintained in both ways. After a large deviation from the scaling, firms usually return quickly onto the scaling line; however, they are also more likely to disappear owing to reasons such as being acquired and bankruptcy. A typical trajectory of surviving firms in the phase space is illustrated in Fig. 1(a).

In our previous study [55], we considered the average dynamical tendencies toward the scaling line, using the vector field of mean yearly displacement in the phase space. Note that the dimensions of the phase space are log-transformed system-size measures. After a normalization of k against the increase in data (see Appendix B for details), a location in the phase space at time t, $\mathbf{x}(t)$, is defined as

$$\mathbf{x}(t) \equiv [\ln\ \tilde{k}(t), \ln\ \ell(t), \ln\ s(t)]^{T},$$

where $\tilde{k}$ denotes the normalized number of trading partners defined by

$$\tilde{k}(t) \equiv \frac{N_{2016}}{N_t}k(t),$$

where N_t is the number of firms with one or more records of their trading partners in year t. Accordingly, a yearly displacement of a single firm at time t, $\mathbf{g}(t)$, is defined as

$$\begin{aligned}\mathbf{g}(t) &\equiv \mathbf{x}(t+1) - \mathbf{x}(t)\\ &= [\ln\ G_{\tilde{k}}(t), \ln\ G_{\ell}(t), \ln\ G_{s}(t)]^{T},\end{aligned}$$

where $G_x(t) = x(t+1)/x(t)$ for a system-size measure x and natural logarithms are adopted. The distribution of displacement substantially depends on the starting point of the displacement, $\mathbf{x}(t)$. The mean displacement conditional on the starting point, $\langle \mathbf{g}|\mathbf{x}\rangle$, can be estimated from a large data set of historical time evolution. Remarkably, the scaling line is approximately the attractor of this vector field, which implies that the scaling relationships are dynamically stable.

In the present study, we focus on the distribution, rather than the mean, of yearly displacements of firms, $\mathbf{g}(t)$. Examples of $\mathbf{g}(t)$ conditional on some values of $\mathbf{x}(t)$ are presented in Figs. 1(b)–1(d), and descriptive statistics for the distributions are tabulated in Table I. The three starting points of the displacements are illustrated in Fig. 1(a). The distribution of $\mathbf{g}(t)$ varies substantially with the starting point $\mathbf{x}(t)$. Yearly displacements starting from around the scaling line are rather symmetrically distributed [Fig. 1(b)]. Note that the data are abundant around this starting point, $(\tilde{k}, \ell, s) = (10, 10, 500)$, as firms are densely located near the scaling line. However, less data exist on the displacement starting from a state that is distant from the scaling line. In these zones, the displacement is usually biased towards positive or negative values in some directions, and exhibits relatively larger variance, as presented in Figs. 1(c) and 1(d) and Table I.

We consider the distribution of displacement as generally not consistent with a multivariate Gaussian distribution based on the following reasons. The non-Gaussian properties are evident for the point (c) and can be confirmed with the negative (for ℓ) and positive (for s) mean log-growth (Table I). Although there is none of such apparent peculiarity for the displacement distributions around the point (d), asymmetricity appears in the distribution of $\ln G_{\ell}$ and $\ln G_{s}$, where extreme values are expected to be larger in the negative (for ℓ) and positive (for s) directions than in the opposite directions. The ratio of the 95% interval length to the standard deviation of $\ln G_{\tilde{k}}$ and $\ln G_{\ell}$ is substantially lower than the theoretical value of 3.92 for the Gaussian distribution, which indicates that the distributions exhibit a slightly broader tail than the Gaussian. This is a natural result, as the empir-

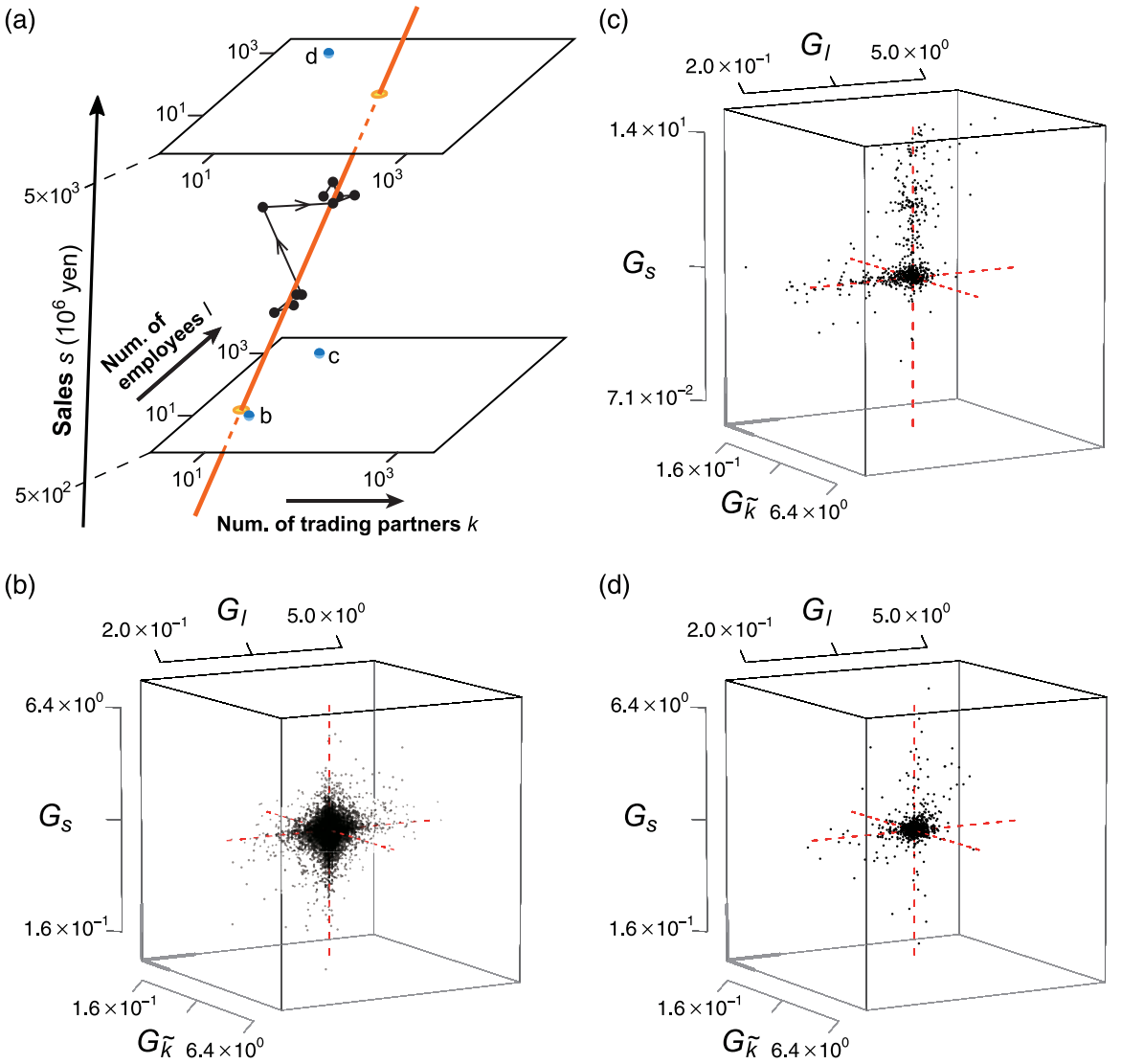

FIG. 1. Dynamical characteristics of three-dimensional scaling laws between measures of firm size. (a) The three-dimensional scaling relationship is illustrated with a typical trajectory of firms in the phase space. Note the use of logarithmic scales. The orange line represents the scaling line determined from three scaling relations, $\ell \propto k^{1.0}$, $s \propto k^{1.2}$, and $s \propto \ell^{1.2}$, where k, ℓ, and s denote the number of trading partners, number of employees, and annual sales in million yen, respectively. Firms are relatively densely distributed on the line. Black dots connected by solid lines represent yearly locations of a hypothetical firm in the three-dimensional phase space. When a firm deviates from the scaling line, it usually moves back toward the line in the following years. (b) Distribution of the 9513 samples for the one-year three-dimensional displacement starting from the neighborhood $(\tilde{k}, \ell, s) \approx (10, 10, 500)$. The data points are plotted in the logarithmic scale. Growth rate G_x for a size measure x is defined by $x(t+1)/x(t)$. Red-dashed lines mark the axes where the growth rates of two of the variables are equal to unity. (c), (d) Distribution of the 1000 samples of one-year displacement starting from the neighborhoods $(\tilde{k}, \ell, s) \approx (10, 1000, 500)$ and $(10, 1000, 5000)$, respectively. Data outside the range indicated by cuboid box are truncated and not shown in the figures. Note that the distribution of growth rates strongly depends on the location in the phase space and is evidently not consistent with a multivariate normal distribution.

ical distribution of yearly displacements in the logarithms of the single measures of the system size has been characterized with "tent-shaped" or slightly heavier-tailed than double exponential Laplace distribution for commercial and cultural organizations [35,54,70,71]. Note that these empirical distributions have been approximated differently in previous studies [70,72,73]. Because the distribution of displacements from around a point cannot be appropriately fitted by a Gaussian distribution, the system dynamic may not be approximated by a simple diffusion equation. This necessitates a nonparametric method for the simulation of the system's time evolution.

B. Simulation method

When transition data over a phase space are abundantly available and temporal correlations are negligible, a system can be modelled by phase-space dynamics that are either defined by a velocity field (when deterministic) or Markov transition probabilities (when stochastic). In such cases, these

TABLE I. Descriptive statistics for natural logarithms of growth rates around different locations in the phase space. SD stands for standard deviation, 0.025-q and 0.975-q for 2.5- and 97.5-percentiles, respectively, and 95%-I for 95%-interval length (i.e., the interval between the 2.5-percentile and 97.5-percentile).

Loc.[a]	Variable	Mean	SD	0.025-q	0.975-q	95%-I
(b)	$\ln G_{\tilde{k}}$	−0.011	0.167	−0.381	0.294	0.675
	$\ln G_{\ell}$	0.001	0.215	−0.511	0.435	0.946
	$\ln G_{s}$	−0.025	0.229	−0.553	0.405	0.958
(c)	$\ln G_{\tilde{k}}$	−0.003	0.275	−0.468	0.508	0.976
	$\ln G_{\ell}$	−0.185	0.605	−2.452	0.223	2.675
	$\ln G_{s}$	0.369	0.855	−0.202	2.841	3.043
(d)	$\ln G_{\tilde{k}}$	0.015	0.244	−0.409	0.467	0.876
	$\ln G_{\ell}$	−0.023	0.264	−0.478	0.283	0.761
	$\ln G_{s}$	0.043	0.252	−0.258	0.722	0.980

[a]Location in the phase space. See Fig. 1 for the exact locations that the alphabets indicate.

"equations of motion" can be determined with the method of analogs by obtaining the estimation of velocity or transition probability at every point in the phase space. This can be achieved by searching for the past data of the system that are close or "analogous" to each point in the space. Our previous method of obtaining mean flow diagrams in the phase space [55] was essentially a deterministic version of "phase-space reconstruction" by the method of analogs. Here we further develop our previous methodology to allow the simulation of the stochastic time evolution. Although the method of analogs was originally presented as a means for assessing the predictability of future atmospheric states [19,20], the method has been successfully applied to socioeconomic data such as GDP relative to international trades [21,22] and texts of patents [23].

To simulate the time evolution of business firms in the phase space of log-transformed system sizes by the method of analogs, we apply the following steps. First, we initialize the simulation by creating a predetermined number of simulated firms of a random size at the time step 0. Unless otherwise mentioned, this number is set to 10^6, which approximates the actual number of active firms in the data. Then the stochastic transitions are repeatedly applied to the firms and they normally make a firm evolve into a certain point in the phase space. They may sometimes trigger the disappearance or exit of a firm with a low probability ranging from less than 1% to 5% per year, depending on the point [55]. The average rate of disappearance in our data is 3.3% per year. When a simulated firm is determined to have disappeared and nonexistent, a new random firm is created at the following time step. These procedures are repeated until the predetermined number of time steps, here set to 1000, is reached.

The creation of a new firm is simulated by random draws from three independent discretized log-normal distributions that best fit to the empirical distributions of respective system-size measures (i.e., the normalized degree in the trading network, employee number, and annual sales with each log-transformed). We justify this adoption of empirical distributions as the proxy for the distributions of newly created firms by our lack of unbiased data of firms just at their birth. We fit a discretized log-normal distribution (with the values rounded up to integers) to the empirical distributions of the three system-size variables. We employ data from the year 2016, as it is the year with the largest amount of data for annual sales in the database; hence, these data are considered to be the most unbiased representation of all the active firms in the country (see Appendix A for details). The fitting is performed via the minimization of Kolmogorov-Smirnov statistics between the empirical and fitted distributions. The power-law tail of the empirical distribution is effectively excluded in this step. Consequently, we obtain the mean $\mu_{\tilde{k}} \approx 1.024$ and the standard deviation $\sigma_{\tilde{k}} \approx 1.176$ for the natural logarithm of $\tilde{k}$, $\mu_{\ell} \approx 1.414$ and $\sigma_{\ell} \approx 1.546$ for the natural logarithm of ℓ, and $\mu_{s} \approx 4.388$ and $\sigma_{s} \approx 1.649$ for the natural logarithm of s. We then employ these parameters to generate a simulated firm of random size by drawing independent samples from the log-normal distributions and rounding up the resulting numbers to integers.

Our simulation of stochastic transitions over the phase space is essentially the method of analogs [19,20] and conforms to the local bootstrap framework [32]. The local bootstrap was originally proposed to generate bootstrapped data for a time series, and utilizes the empirical data of transitions starting from a state that is sufficiently close to the present state of the simulated system. Here, we adopt this method according to the following considerations. First, the method should be nonparametric and without restrictive assumptions regarding the tail, as we are not provided with the approximate functional forms of the density of transition probability (see Fig. 1; also refer to the last paragraph of Sec. II A). Second, the method should be well-defined for a mixture of continuous and discretized data. Because our data are integer-valued, the values of size measures are discrete for small firms, although they can be considered approximately continuous for large firms. Finally, the method should be applicable to multivariate data. A brief review of bootstrap methods for time series data other than the one we employ is available in Ref. [74].

The procedures of our simulation method are defined in formal terms as follows. First, we determine the empirical data located at the neighborhood of the simulated firm. Let $\mathcal{C}$ denote the (multi)set of empirical data for the yearly time evolution of firms, called a catalog [28,29] (refer to Appendix A for details on the data compilation). We search the catalog, including 19 766 521 data of single- and double-year time evolution of firms, for the neighboring data $\mathcal{N}(\mathbf{x})$ around the location of the simulated firm in the phase space, $\mathbf{x}$. Neighbors are determined in terms of either the nearest neighbors or data within a certain distance, depending on the data density near $\mathbf{x}$. We first determine the set of data points whose distance to the simulated firm, c^*, is less than d^{thr} [here set to $(\ln\ 10)/16$], and if the number of such data points are less than $n^{\min}$ (here set to 50), we adopt the $n^{\min}$-nearest neighbors instead. In a mathematical notation,

$$\mathcal{N}(\mathbf{x}_{c^*}) \equiv \{c \in \mathcal{C} \mid \|\mathrm{x}_{c^*} - \mathrm{x}_c(0)\| \leqslant \max\{d^{\mathrm{thr}}, r_{n^{\min}}(\mathrm{x}_{c^*})\}\},$$

where

$$r_N(\mathbf{x}) = \min\{r \geqslant 0 \mid |\mathcal{N}_r(\mathbf{x})| \geqslant N\},$$
$$\mathcal{N}_r(\mathbf{x}; \mathcal{C}) = \{c \in \mathcal{C} \mid \|\mathbf{x} - \mathbf{x}_c(0)\| \leqslant r\},$$

and the norm symbols $\|\cdot\|$ and $|\cdot|$ denote the Euclidean norm and number of elements in a set, respectively. Note that this "adaptive" change in the number of resampled data according to the local data density is not unique to ours, but usual in meteorological predictions using the method of analogs (e.g., Ref. [28]). Next, we randomly select a single empirical datum of evolution from the neighborhood $\mathcal{N}(\mathbf{x}_{c^*})$ and allow the simulated firm to follow the evolution of the empirical firm. Letting c_J denote the Jth element of $\mathcal{N}(\mathbf{x}_{c^*})$, where J is the random variable that takes its values in the set $\{1, 2, ..., |\mathcal{N}(\mathbf{x}_{c^*})|\}$ with a uniform probability, the location of the simulated firm is obtained with the equation:

$$\mathbf{x}_{c^*}(\tau + i) = \mathbf{x}_{c^*}(\tau) + \mathbf{x}_{c_J}(i) - \mathbf{x}_{c_J}(0), \tag{2}$$

for $i = 1$ in most cases; however, $i = 2$ in special cases, as we mention below. Because the components of $\mathbf{x}$ in the catalog data are logarithms of integers for ℓ and s, the output of a simulation is generally real-valued instead of being integer-valued. Although this might seem erratic, the error is actually not severe as we discuss in Sec. II C. Moreover, the zones of our interest are primarily where the system size is substantially larger than 1 and approximately continuously distributed.

In Eq. (2), i can be equal to 2 and thus the simulation of a firm proceeds by two time steps if the vector $\mathbf{x}_{c_J}(1)$ contains a not-available (NA) data, i.e., the data are missing or unobserved in that year. Note that such partially missing data are usual in empirical data. In our case, for example, approximately 20% of firms in a year do not have the data of trading partners. Considering that collecting data of a small firm would be sometimes difficult, missing data should be more likely to occur for smaller firms. Therefore, simply excluding all the NA data might incur the biased representation of the true time evolution of smaller firms, for example, when data are missing only in a year, but available in the previous and following years. To avoid such inaccuracies in our simulations, we seek, when we identify missing data, to fully utilize the available data as follows. If every variable in the vector takes an NA value in the second step ($i = 2$), then it is highly likely that the firm in the data disappeared. There are 655 151 or 3.31% of such data in the catalog. In this case, the simulated firm is determined to be removed from our simulation and a new random firm is created in the next time step. In contrast, if a non-NA number is available for all the variables in the second year, the firm is considered to have existed through the period and the simulated firm is determined to survive. The number of such data is 460 078 or 2.33% of all data. A datum that has one or two available values of the three system-size measures even in the second year is excluded in the data compilation (see Appendix A) and does not exist in the catalog. The data of time evolution that starts from a partially defined coordinate on the phase space is also excluded in the compilation process.

Our modifications to the original local bootstrap methods are twofold. First, we adopt displacements instead of target states. Using our notation, one substitutes $\mathbf{x}_{c_J}(i)$ for $\mathbf{x}_{c^*}(\tau + i)$ in the original local bootstrap method, which is called the "locally constant operator" in the recent literature [28,29]. Nonetheless, this original method is not considered "physical" because the displacements do not converge to zero as the length of a timestep Δt tends to zero [28]. The negative effect of the locally constant operator to the accuracy makes our method, the "locally incremental operator" [28,29], more suited for our purpose under the assumption of continuity on the underlying stochastic process. Second, we adaptively change the radius of neighborhood spheres. Such adaptive changes based on the data density should be considered given our first modification, because firms can accidentally deviate from the zones where the empirical data exist. Accordingly, d^{thr} and n^{min} are "hyperparameters" in our simulations. Although we do not have a definitive basis for determining the value of d^{thr} and n^{min}, changes in d^{thr} between $(\ln 10)/16$ and $(\ln 10)/8$ and n^{min} between 2 and 200 do not make a discernible difference in our simulations.

In summary, the stochastic time evolution of a firm's system size in a stable environment is simulated by our method from the firm's birth to its demise. We first create 10^6 new small firms, unless otherwise mentioned. Although the majority of firms survive a time step and often change their size in either positive or negative directions, some of the firms disappear. When the number of firms decrease by such disappearance, the same number of new small firms are inserted in the simulation, such that the number of firms is constantly 10^6. When a firm grows by an acquisition, the growth is described in our database as a normal growth. However, the disappearance of firms by reasons including involvement in mergers or acquisitions, as well as bankruptcy and planned discontinuation of business, are all recorded in our database. In total, 72 100 events of mergers and acquisitions are recorded during the period of our catalog and they explain $\approx$11% of the disappearance of firms. Accordingly, both the disappearance of firms by mergers and the growth of firms by acquisitions should occur in our simulations by approximately the same average rate and balance in a long run, although they are not explicitly paired.

C. Assumptions and accuracy of simulation

Applying our simulation method to the data, we assume the stationarity of the system, homogeneity of system elements (business firms), and negligible error in observation. The first condition of stationarity posits that the system was already at the stationary state when it was observed. If this is not met, then a fraction of system elements could enter the zones of phase space where observed data are insufficient. Simulated evolution of such elements should heavily rely on the extrapolation of the existing data and thus lead to a substantial error. The second condition states that the systems do not have unobserved latent variables that affect their evolution in the phase space. For example, consider a situation in which the variance of one-year displacements differs substantially depending on the types of business that a firm is involved in. In this situation, which we assume is not the case, our method could overestimate the fluctuation exhibited by single firms in a long-term observation. The effect of substantial observation error in data is similar to that of the heterogeneity of system elements.

Although some characteristics of the outputs of our simulation may be inconsistent with the original data, we stress that the errors are not severe. First, the firm size often become

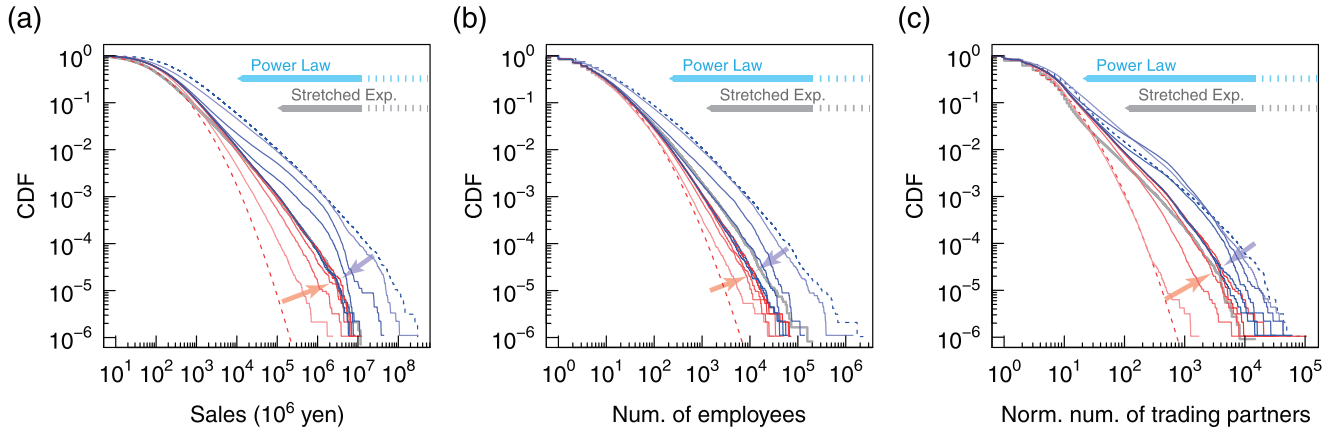

FIG. 2. Behavior of size distributions in the stochastic simulations presented as the change in cumulative distribution function (CDF) for (a) the annual sales, (b) employee count, and (c) normalized number of trading partners. Grey bold curves indicate the empirical distributions in 2016. Light blue and grey bars at the top right indicate the intervals for which a power-law and stretched exponential distributions fitted better to the empirical data in 2016. The ranges without empirical data are marked by broken bars. Red and blue dashed curves indicate the initial conditions of simulations for size distribution of 10^6 firms, while pale to dark solid curves represent the simulated distribution at the time step τ equal to 20, 100, 200, 500, and 1000. The initial distributions are log-normal for red, while the empirical samples in 2016 are raised to the power of 1.2 for blue. The simulated stationary distribution ($\tau > 500$) is remarkably consistent with the empirical distribution in 2016 for the annual sales. Although there are discrepancies between the simulated and empirical distributions for the other two size measures, the exponents of the power-law tails in the intermediate scales match well between the empirical data and simulation.

noninteger valued according to our simulation method (2). Although this may seem odd, it is necessary to avoid unfavorable biases from rounding the figure to integers when the firms are small. The size measure also sometimes becomes smaller than a unit, but it is typically larger than 0.5. This is because the applied displacement for a time step is mostly zero or in the positive direction, as the distribution of displacement is determined by "analogs." When a simulated firm has less than one employee, almost all possible displacements for the next step are collected from the empirical data of firms with only an individual employee that evolve to firms with one or more employees. Note that zero-valued data have been excluded at the data compilation step (refer to Appendix A). When the data should be presented as integers, we recommend rounding the numerical figures after all the simulations are conducted.

Another artifact can occur at the zones where very few data are available. Our method defines the transition probability from every point of the phase space including those without actual data nearby, by extrapolating the data at the edge of distribution. Therefore, even if the empirical system has a stable state in terms of the system-size distribution, a part of simulated firms may possibly leave the zones with empirical firms and proceed in the same direction indefinitely, thereby making the system unstable. Although this may in principle occur depending on the realization of the data, the probability that it occurs is low for an ensemble of systems that has reached the stationarity, as a system at peripheral zones almost certainly tends towards the central zones of the system distribution. Although this should be a serious problem for nonstationary processes, such a problem appears absent in our simulation results, as discussed in the next subsection.

D. Relaxation of system-size distribution

Having reviewed the basics of our simulation method, we explore the limit as $\tau \to \infty$ to characterize the stability of the system in our study. Because our simulation is free of memories, the stochastic system reaches a single stable distribution as $\tau \to \infty$, provided it indeed has such a distribution. Therefore, we simulate the collective system of firms until no change is visible in the system size distribution of simulated firms, i.e., $\tau = 1000$ in this case. The empirical distribution is compared to the limit distribution of simulations.

The empirical distribution of the size of firms has been characterized by power-law tails [35,40,61,64–70]. To characterize the empirical distributions of size measures in our data, we compare the fit of power-law and stretched exponential (Weibull) distributions as formulated in Ref. [75] to the empirical data within different intervals ranging from a threshold to infinity. The cumulative distribution function (CDF) $\mathrm{P}(X > x)$ of the two classes of distributions are parameterized as $(u/x)^b$ and $\exp[-(x/d)^c + (u/d)^c]$, where u is the lower threshold. We employ the Akaike information criterion (AIC) [76] for the model selection. The results are summarized at the top right in panels of Fig. 2. The empirical distributions of the most extreme values are rather consistent with a stretched exponential decay: data for $s \geqslant 10^5$ and $k \geqslant 100$ are better fitted by the stretched exponential distribution with $(c, d) \approx (0.127, 8.01 \times 10^{-3})$ and $(0.187, 6.43 \times 10^{-3})$, with the AIC values lower by 3.6 and 12.5 compared to the power-law fitting, respectively. The AIC rarely favors the stretched exponential distribution for ℓ, but data for $\ell \geqslant 1200$ are better fitted by the stretched exponential distribution with $(c, d) \approx (0.166, 6.76 \times 10^{-3})$, with the AIC value lower by 0.3. In contrast, data for $s \geqslant 10^4$, $\ell \geqslant 200$, and $k \geqslant 20$ are fitted by power-law distributions with $b \approx 1.00$, 1.30, and 1.34 with the

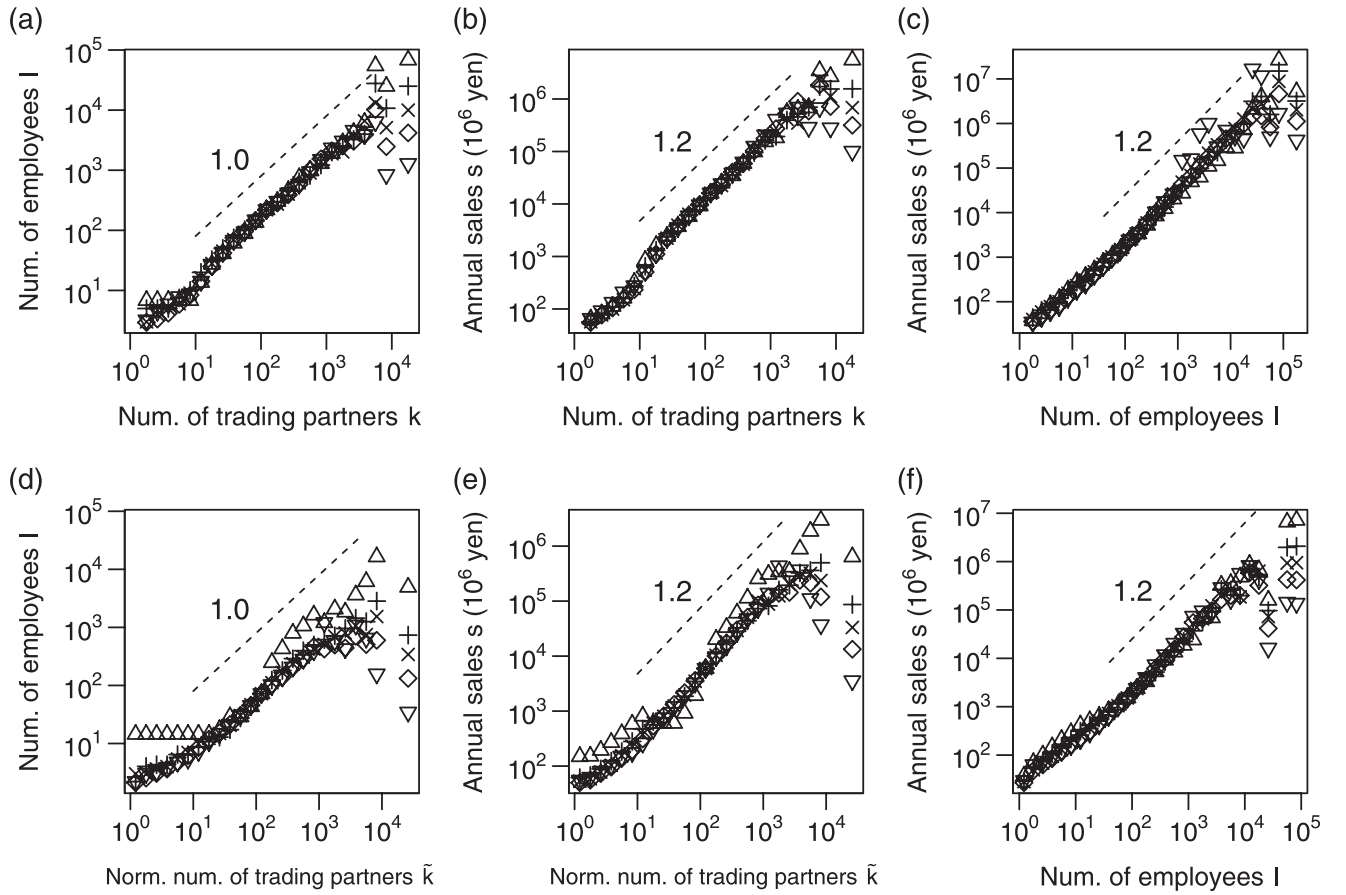

FIG. 3. Scaling relationships between pairs of system-size measures for the empirical data and simulated stationary state. Plots in the first row (a)–(c) are obtained from the empirical data in 2016, while those in the second row (d)–(f) are from the results of simulation at the time step $\tau = 1000$. Plots of 5-, 25-, 75-, and 95-percentiles of the y axis variable conditional on the x-axis variable are vertically shifted to collapse almost into the curve of the conditional median. The pairs are $(\tilde{k}, \ell)$ for the left column [panels (a) and (d)], $(\tilde{k}, s)$ for the middle [panels (b) and (e)], and (ℓ, s) for the right [panels (c) and (f)]. Dashed lines are placed at the same location in both panels of the same column. Gaps in the value of the conditional median are evident between the empirical data and simulation results for the pair $(\tilde{k}, \ell)$ in panels (a) and (d) and $(\tilde{k}, s)$ in panels (b) and (e), while no apparent discrepancy exists in scaling exponents between the data and simulation. The plots in panels in the upper row are from Ref. [40] for comparison. Refer to Ref. [37] for details on the plotting method.

AIC values lower by 42.2, 253.0, and 1635.2 compared to the stretched exponential fitting, respectively. Therefore, we conclude that a power law is, at least in an intermediate scale, a parsimonious description for the empirical distributions of measures of the system size, including the number of trading partners (i.e., the degree in the interfirm trading network). On this basis, we hereafter describe these distributions as heavy-tailed.

In Fig. 2, the time evolution of the distribution for each system size measures of firms is presented. The simulations are initialized with two contrasting distributions marked by blue and red dashed curves in the figures. The former comprises 10^6 random samples of small firms from the log-normal distribution assumed for newly created firms, while the latter comprises 10^6 empirical samples of firms in 2016 based on the database, but with the size arbitrarily enlarged by the power of 1.2. Although the system is initialized with such extremes, it almost collapses into a single distribution, 500 time steps after the initialization. Surprisingly, our simulation accurately predicts the empirical distribution of annual sales marked by the black bold curve in Fig. 2(a). Compared to the annual sales, the prediction for the number of employees or trading partners is apparently not as accurate. However, it still appears to predict the exponents of the power-law in the intermediate scale accurately, as the curve for the CDF that indicates the limit distribution is almost parallel to that for the empirical distribution in Figs. 2(b) and 2(c). In general, the empirical system size distribution is fairly well approximated by the stationary distribution of our simulation.

E. Scalings in stationary distribution

In addition to the system-size distribution measures by single-size measures, scaling relationships between the measures of size can be compared between our simulation and the empirical data. Because the allometric scaling relationships exert a strong limitation on the joint distribution of firm size measures, it is appropriate as a criterion of whether the simulation is consistent with the empirical data.

We demonstrate the allometric scaling property of the empirical and simulated firms, as presented in Fig. 3. Empirical data and results of simulations are presented in panels at the

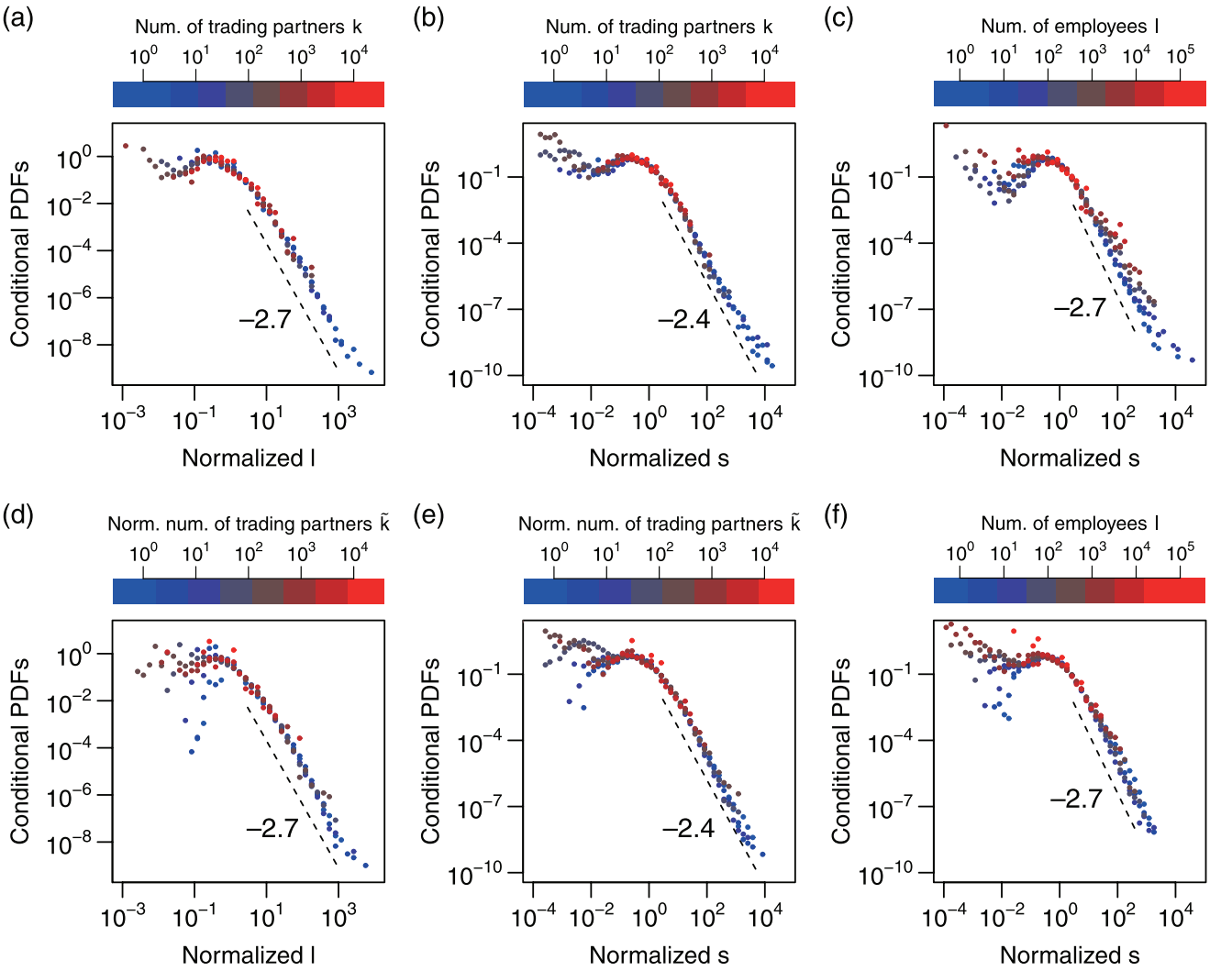

FIG. 4. Distributions of fluctuations from average relationships between pairs of system-size measures for the empirical and simulated firms. Plots in the first (a)–(c) and second (d)–(f) rows are obtained, respectively, from the empirical data in 2016 and the results of simulation at the time step $\tau = 1000$. Probability distribution functions (PDFs) of normalized size measures $(y/\langle y|x\rangle)$ conditional on eight intervals of x that are equal in length in a logarithmic scale are plotted. Color gradation from blue to red indicates the smallest to largest values of x. The x values indicated by the colors are presented in the respective legend above the main plots. The pairs of system-size measures are the same as in Fig. 3. Dashed lines are placed at the same location in both panels of the same column. There is no obvious discrepancy between the empirical data and simulation results except for small values of normalized system size. The plots in panels in the upper row are from Ref. [40] for comparison. Refer to Ref. [37] for details on the plotting method.

top and bottom, respectively. Data of the simulated firms are obtained from the final time step $\tau = 1000$ in the figures. The percentiles of the conditional distribution $\mathrm{P}(y|x)$ are plotted, where x and y denote the measures in the horizontal and vertical axes, respectively. Percentiles other than the median are vertically shifted in the plots, such that the fact can be visualized that the curves of every percentile collapse into a single curve, by the convention followed in Ref. [37]. Note that the power laws for each pair of system-size measures are formulated as $\ell \propto \tilde{k}^{1.0}$, $s \propto \tilde{k}^{1.2}$, and $s \propto \ell^{1.2}$. The vertical positions of the conditional medians differ between the simulation and empirical data, which is consistent with the slight disagreement between the two as already illustrated in Fig. 2. Nevertheless, the simulations agree with the empirical data regarding the values of power-law scaling exponents. This highly suggests that the conditional distribution exhibits a universal distribution of fluctuation, $\widetilde{P}$ [as in Eq. (1)] for every pair of $\tilde{k}$, ℓ, and s.

The existence of the universal function of fluctuation, $\widetilde{P}$, is directly verified by actually observing the distribution of $\mathrm{P}(\widetilde{y}|x)$, where $\widetilde{y} \equiv y/\langle y|x\rangle_{\mathrm{med}}$ and $\langle y|x\rangle_{\mathrm{med}}$ denote the conditional median. If allometric scaling (1) does hold, then the conditional distribution of $\widetilde{y}$ should be independent of the value of x. Such conditional densities in the empirical data and our simulation are plotted in Fig. 4 for each pair of $\tilde{k}$, ℓ, and s. It can be observed that the results of our simulation accurately approximate the empirical data. The existence of the power-law tails of the fluctuation and their exponents are particularly well approximated.

We conclude that the approximation of the empirical data by our simulation results is mostly accurate with a few minor inconsistencies. The number of employees is simulated to be

slightly larger and the number of trading partners is slightly smaller compared to the empirical system, leading to a rather major gap between average $\tilde{k}$–ℓ ratio.

The cause of such inconsistency could be explored by comparing the actual system against the three assumptions of our method outlined in the previous section. Among the three conditions, the incompatibility of our data to the assumption of the system's homogeneity might not be negligible. It is likely that the average $\tilde{k}$–ℓ ratio is different in different industrial sectors. For example, an intermediary firm engaging in the real estate trading might have a dozen trading partnerships with only several employees; however, a manufacturing firm would typically require more employees to have the same number of trading partners. Simulations based on the data of a specific industry would require more detailed consideration on the nonstationary growth and decline of firms in some industries, which is beyond the scope of our current study.

III. DISCUSSION

We have demonstrated that the multidimensional system-size distribution of business firms can be well approximated by the stationary state of a single stochastic process defined by a stochastic version of the method of analogs applied to a large-scale data set of the time evolution.

An immediate consequence of our results is a strong suggestion against the possibility that a majority of firms have long-range temporal correlations in their changes of size. If such long-range temporal correlations exist in the time series of the system size (which is characterized by a power-law tail of correlation function), then a simulation completely without memories such as ours should typically provide an inconsistent result with the empirical data. Our results are therefore consistent with the previous research, which found no evidence of long-range temporal correlations in firm size changes when considered collectively [77]. This fact validates our previous approach [55] that visualizes the mean flows of firms in the phase space, which implicitly assumes that a firm size change in a year is independent from that in other years. Nevertheless, it is not excluded that the size change of the firms in our study actually exhibit a weak or quickly diminishing temporal correlation, or even time evolution of the system size of a minor part of firms has a long-range memory. Mathematical studies of the local bootstrap method and its variants determined that these methods can accurately approximate the stationary distribution of a stochastic system even in the presence of short-term memories, although this does not hold if there are long-term memories [30–33].

Given that the dynamics of "average" firms can be simulated by our method without any memory effect, our simulation can serve as a null hypothesis that there are no strong temporal correlations in the system-size changes. When one can track the evolution of a system or a group of systems, if there is any inconsistency between the actual data of time evolution and the results of simulations, then one could state that some temporal correlations or hidden variables that are not included in the empirical data should exist. For the system in our study, considering that the dynamics of $\ell/\tilde{k}$ might differ according to the industry that the firm belongs to, we do not exclude the possibility that the industry classification is a hidden variable, which may reflect the major heterogeneity in the entire collection of firms in a country. Therefore, classifications of the industry that a firm is involved in can be considered as a hypothetical factor of the minor gap between our simulation and the empirical data in future studies. Another interesting question is centered on whether the dynamics of a firm depend on its age, to which we do not have any answer currently.

The system-size distribution in our simulations is not stable until $\tau \approx 500$, which implies that it takes approximately 500 years for the ensemble of firms to reach a stable state. This result is reminiscent of the hypothesis of self-organized criticality at work in a variety of complex systems [51], including human society [78], as it would suggest that the system is near the criticality between stationarity and nonstationarity. In contrast, one might question the result because the length of the relaxation time exceeds the timescale for the history of business firms by modern designs. We would like to note that some care should be taken to interpret the relaxation time. First, empirical size distributions of firms typically have a power-law tail with the exponent between -1.4 and -1 [35,40,61,64–66,68–70] and the exponent is larger than -1 in rare cases [67]. Therefore, the initial conditions in our simulations, a log-normal distribution and a heavy-tailed distribution with a power-law tail of exponent larger than -1 (in a CDF), are rather extreme to represent an empirical system-size distribution of all firms in a geographical area. The relaxation time would be shorter if we initialize the simulation with a more realistic size distribution. Second, we have not considered interactions between firms. When the ensemble of firms is in an extreme state, it is possible that strong interactions lead the distribution to a stationary state faster than in the absence of such interactions. Because our simulations are solely based on the empirical data near the stationarity, it should be noted that our method might fail when the model is interpolated for an ensemble far from the stationarity. Because the country under our study (Japan) experienced a rapid growth in population and economy during the post-war period between 1950 and 1980, we propose that the strength of interactions between firms should be elucidated using a large-scale data set of firms in that period.

Because our simulation is essentially a Markov process, we argue that models for the dynamical origin of allometric scalings in firms should exhibit the Markov property, at least approximately near their stationarity. We propose that a key step to such a model would be to determine mathematical formulations for the transition probabilities, which are denoted by $\mathbf{g}(t)$ conditional on $\mathbf{x}(t)$ (see also Fig. 1). Growth rate distributions of a single-size measure have been studied for decades [35,54,70,72,73,79]. However, there are insufficient studies on the conditional probability distribution of growth rate that depends on the location on a multidimensional phase space. Identifying statistical regularities for $\mathrm{P}(\mathbf{g}|\mathbf{x})$ in empirical data could be a promising research direction.

Although the simulations in our study are three-dimensional, two-dimensional simulations are certainly possible by ignoring a variable. We expect that the main consequence of variable omissions would be a faster convergence to the stationary distribution in simulation results in our case. As suggested by empirical examples of $\mathrm{P}(\mathbf{g}|\mathbf{x})$ at

different locations **x** (see Fig. 1 and Table I), the distribution of displacements can be heavier-tailed for atypical states located distantly from the "scaling line" compared to typical states on the "scaling line." This implies that a system has to be distant from typical states before a large yearly displacement occurs. Nevertheless, by neglecting a variable, we sometimes cannot distinguish an atypical state from typical ones. This in turn would allow a part of simulated systems to "fly" a long distance instantly, thus accelerating the convergence to the stationary distribution.

It is important to note that our proposed method is not specific to business firms and can be applied, rather universally, to any type of system with a large amount of observed time series. Although the method of analogs has been known to perform well with deterministic time series, our results are encouraging about using the method also for simulating time series of stochastic nature. Considering that the distinction between deterministic chaos and stochastic processes is sometimes not easily performed [80–86], the applicability of the method of analogs to both deterministic and stochastic processes is much advantageous as difficulties in recognizing or rejecting the determinism can be circumvented. The three assumptions of the method, which we mentioned earlier in Sec. II C, are those also referred to as the conditions for applying the method of analogs to deterministic time series [26,29]. A disagreement between the simulation results and empirical data is sufficient to reject at least one of these assumptions. Therefore, our approach may be beneficial in detecting long-range memories, nonstationarity, or large observation errors in general empirical time-series data, regardless of the level of determinism in the underlying processes.

ACKNOWLEDGMENTS

The authors thank Teikoku Databank, Ltd., Center for TDB Advanced Data Analysis and Modeling, for providing data and financial support. This work is partially supported by the Grant-in-Aid for Scientific Research (B), Grant No. 26310207 and JST, Strategic International Collaborative Research Program (SICORP) on the topic of "ICT for a Resilient Society" by Japan and Israel, and by Japan Ministry of Education, Culture, Sports, Science and Technology as "Exploratory Challenges on Post-K computer (Study on multilayered multiscale spacetime simulations for social and economical phenomena)." M.T. directed the project. Y.K., H.T., and S.H. developed the simulation method. Y.K. compiled the data and generated the plots and diagrams. All authors contributed in writing the paper.

APPENDIX A: DATA COMPILATION

We use the database of the summarized description of Japanese business firms provided by Teikoku Databank, Ltd., Japan (TDB), named COSMOS 2. A list of anonymized firms that were recognized as being active by TDB at the beginning of each year during the 1994–2018 period was available to us, along with their financial status (annual sales, capital, and profits), location, industrial classification, direct buying and selling trading partners, etc.

We first obtain a set of quantities that indicate the size of a firm in a year, namely the total number of selling and buying trading partners, the number of employees, and the annual sales in million yen, which are denoted by k, ℓ, and s, respectively. Note that the number of trading partners amounts to the sum of in- and out-degree of a firm in the trading network. These firm size measures are all integer-valued in the database. The three quantities have been intensively studied in previous research [37,40,55] and found to follow heavy-tailed distributions and relate to each other with "allometric" power-law scaling [37,40]. Records of quantities that are not available are left as such (NA). We exclude financial or banking firms and governmental entities based on their industrial classification in our data compilation process. These firms are also excluded from the count of trading partners. This is to filter the sales and trades that are referred to for the purpose of accounting, but inconsistent with the ordinary use of the words. Here we use the term "sales" for the amount of money that a firm receives in exchange for their goods and services in a year, and a "trade" for such an exchange between two firms. The sales figure published at the end of a 12-month fiscal year is adopted as the annual sales of the calendar year in which the fiscal year was closed. When the fiscal year was not 12-month long or there existed more than one end of a fiscal year in one year, sales figure for the year is left to be NA. In contrast, for the count of employees and trading partners and the industrial classification, we consider that the records represent the state of a firm in the previous calendar year of the timestamp of the data (i.e., January of each year). When zero values occur in the count of trading partners, we replace them with NAs. At the end of these processes, we obtain a triplet of matrices tabulating the size of all the firms during the studied period, measured by each of the three different size indices. Accordingly, the number of firms with full record of three system-size measures is 8.470×10^5 per year on average during the 25-year period between 1993 and 2017. The number of firms with partial record of the three measures of system size is 4.181×10^5 per year on average in the same period, of which 2.852×10^5 lack the data on trading partners. The number of unique firms recorded in this period is 2 515 679.

We then compile a "catalog" [28,29] that represents the randomized short-term evolution of firms (i.e., growth rates measured by k, ℓ, and s) in the period between 1993 and 2017. Because the evolution of firms indicated by k, ℓ, and s can be conceived as a realized stochastic process in three-dimensional phase space, we hereafter refer to the changes of the position of a firm in the phase space as transitions.

First, the number of trading partners is normalized according to the number of recorded firms in the trading network. This process makes the number of trading partners in a year comparable to that in another year. The details and rationales of the process is discussed in Appendix B. The normalized degree is denoted as $\tilde{k}$. NA values are kept as such during this process.

Next, we obtain the catalog, denoted by $\mathcal{C}$, the list of empirical transitions over the three-dimensional phase space. We search the triplet of matrices for pairs of a firm and a year for which $\tilde{k}$, ℓ, and s all have non-NA real values. We can then obtain a tentative list of single-year transitions over the phase space that start from a fully defined coordinate without NA

values. We neglect original records of the year and firm identity in this list to aggregate all the available data. The transition data that end with NA values indicate either the discontinued activity of a firm or an active firm that was just unobserved. We note that the frequency of single-year missing data of a firm is singularly high compared to the missing data of a firm for more than one consecutive year. To ensure that NA values at the end of a transition indicate the discontinuing activity of a firm, the time range of the evolution data is extended to accommodate three consecutive years if NA values exist at the end of a 1-year evolution. If some, but not all, of the data at the end of evolution are NA values after this 1-year extension of time range, then we simply ignore the data from the list of empirical transitions to ensure that the transition data end with either all NA or all non-NA values. We thus obtain a catalog for the transition of firms over the phase space during a single- or double-year range. This lengthening of transition data does not induce biased representation of the original data as all the data listed in the catalog are derived from mutually exclusive parts of the original matrices. It is known that the varying length of the transition data sometimes improves the accuracy of the class of methods we employ [87].

The final catalog contains a total of 19 766 521 entries of single- and double-year transitions. The numbers of single- and double-year transition data are 18 651 292 and 1 115 229, respectively. The double-year transitions end with either a fully defined coordinate (in 460 078 entries) or all-NA values (in 655 151 entries). The latter indicates the disappearance of a firm owing to various reasons including a merger by other firms. Although we do not have access to a data set that specifies the reasons of disappearance for each single firm, the recorded number of events of merger and acquisition in the 1994–2017 period, during which the acquired firms in the catalog should have disappeared, is 72 100. This amounts to 11.01% of all events of firm disappearance in the catalog.

For applications in the actual simulations, we log-transform the value of three firm size measures ($\tilde{k}$, ℓ, and s). One reason for this is that distributions of these firm size measures are typically heavy-tailed and the increasingly sparse data at the tail would severely affect the effectiveness of our simulation method. With the logarithms, one can suppress the appearance of heavy tails in the distribution. Moreover, the magnitude of relative fluctuation of firm sizes has been reported to be only weakly dependent on the size, particularly for large ones. Because this dependency is often characterized by a power law with the exponent typically ranging between 0 and -0.25 [54,70], the evolution of firm size can be regarded as rather multiplicative. Nevertheless, the method that we will adopt here is additive [see Eq. (2)], and this gap can be filled by adopting logarithms.

APPENDIX B: DEGREE NORMALIZATION

Here we discuss the details and rationale for the normalization of the number of trading partners. The number of trading partners (k) is the sum of the numbers of "selling" partners that a firm sells to and "buying" partners that a firm buys from in a year and amounts to the sum of in- and out-degrees in the interfirm trading network in terms of network science. In a region without severe economic or demographic fluctuations, the number of employees (ℓ) and annual sales (s) in local currency may not require further normalization. We believe that this was the case for Japan during the 1993–2017 period under this study, which is evidenced by the marked stability of the distributions of ℓ and s during the period [40]. However, the degree distribution should be substantially dependent on the size of the observed interfirm trading network, which is almost continuously increasing [Fig. 5(a)]. In Fig. 5(b), we plot percentiles of the degree distribution relative to the values in 1993. Fortunately, we observe that all the 75-, 95-, 99-, 99.9-, and 99.99-percentiles of the distribution of k are increasing at nearly the same rate. Our observation is further verified by the plots of the CDFs for the degree in each year, as observed in Fig. 5(c), which are almost parallel to each other with the exponent of the power law in the intermediate scale unchanged. Consequently, we can normalize k by a factor that only depends on the year and regardless of a firm's size.

The behavior of the aforementioned empirical degree distributions can be explained as follows. Assume that the firms in our database in each year are randomly sampled from the constant "true" population of all the active business firms in the region in an unbiased manner, and that the proportion of sampling continually increases with the years. Then, the series of seeming degree distributions should change in the same way as they do in the empirical data, because a fixed proportion of the true set of trading partners for a firm are sampled for every firm in a year. Simultaneously, the proportion of sampled firms should be proportional to the number of observed firms by definition. Therefore, it is hypothesized that we can normalize the total number of trading partners, $k_c(t)$, of a firm c in a year t, by the number of recorded firms in each year. To allow an intuitive interpretation, we represent the normalized number $\tilde{k}_c(t)$ with the 1-degree-equivalent unit in 2016:

$$\tilde{k}_c(t) \equiv \frac{N_{2016}}{N_t} k_c(t),$$

where the number of observed firms, N_t, in a year t is determined by the number of nonfinancial and nongovernmental firms with non-NA and nonzero degree values. It can be observed that the CDFs of normalized degree approximately collapse into a single function by the normalization [see Fig. 5(d)], which validates our normalization method.

The effects of this normalization can be observed from differences between the simulation results using catalogs with and without the degree normalization. The transient distributions of k in the simulation based on the catalog with the raw number of trading partners are plotted in Fig. 5(e). The stationary k-distribution without normalization substantially deviates from the empirical one to the right side, while the adoption of the catalog with the normalized degree ($\tilde{k}$) leads to a stationary distribution close to the empirical $\tilde{k}$-distribution [see Fig. 2(c)]. Therefore, we can conclude that the degree normalization exerts a substantial effect on the results of our simulations by the violation of the assumption of our simulation method, which states that the system under this study is stationary.

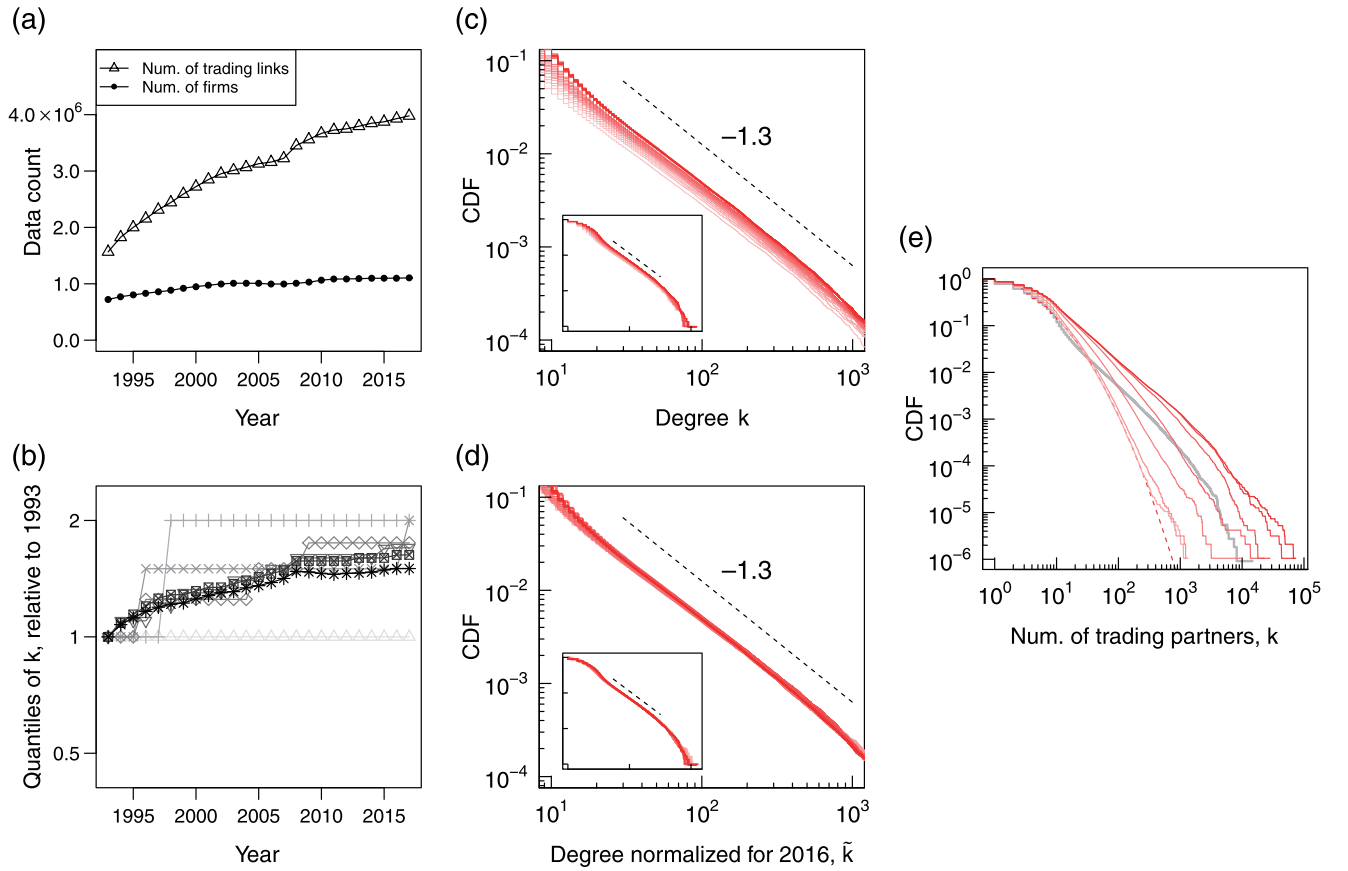

FIG. 5. Nonstationary increase in data and degree normalization for the apparently growing interfirm trading network and its effect on the simulation results. (a) The number of firms and trading links observed in the data for every year between 1993 and 2017. The number of trading links increased by more than two-fold during the period. (b) 5-, 25-, 50-, 75-, 99-, 99.9-, and 99.99-percentiles of the degree distribution (the distribution of the number of trading partners, k) in years between 1993 and 2017, relative to the 1993 data. Paler colors mark lower percentiles. The curves for higher percentiles approximately match each other. (c), (d) Original and normalized degree distributions for every year between 1993 and 2016. CDFs for the range of degree between 10^1 and 10^3 are shown in the main plots, while the whole range are shown in the inset plots. Older data are plotted with paler red. A tick indicates two orders of magnitude in the inset plots. (e) The behavior of simulated distribution of the number of trading partners without normalization, shown as the change of CDF. Grey bold curve indicates the empirical distribution in 2016. Red dashed curve indicates the log-normal distribution for the initial condition, while pale to dark solid curves represent the simulated distribution at the time step τ equal to 10, 20, 50, 100, 200, 500, and 1000. The simulated stationary distribution ($\tau > 500$) evidently disagrees with the empirical distribution in 2016.

[1] J. D. Murray, *Mathematical Biology I: An Introduction*, 3rd ed. (Springer-Verlag, New York, 2002).

[2] L. Edelstein-Keshet, *Mathematical Models in Biology* (Society for Industrial and Applied Mathematics, Philadelphia, 2005).

[3] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data*, Institute for Nonlinear Science (Springer, New York, 1996).

[4] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, 2nd ed. (Cambridge University Press, Cambridge, UK, 2003).

[5] H. C. Nguyen, R. Zecchina, and J. Berg, Adv. Phys. **66**, 197 (2017).

[6] N. Masuda, M. A. Porter, and R. Lambiotte, Phys. Rep. **716-717**, 1 (2017).

[7] S.-I. Kumamoto and T. Kamihigashi, Front. Phys. **6**, 20 (2018).

[8] J. D. Murray, *Mathematical Biology II: Spatial Models and Biomedical Applications*, 3rd ed. (Springer-Verlag, New York, 2003).

[9] S. P. Hubbell, in *The Unified Neutral Theory of Biodiversity and Biogeography*, edited by S. A. Levin and H. S. Horn (Princeton University Press, Princeton, NJ, 2001).

[10] W.-X. Zhou and D. Sornette, Eur. Phys. J. B **55**, 175 (2007).

[11] R. A. Armstrong and R. McGehee, Am. Nat. **115**, 151 (1980).

[12] W. Ebenhöh, Theoretical Population Biology **34**, 130 (1988).

[13] P. L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629 (2000).

[14] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. Lett. **85**, 4633 (2000).

[15] G. Bianconi and A.-L. Barabási, Phys. Rev. Lett. **86**, 5632 (2001).

[16] E. Bedolla, L. C. Padierna, and R. Castañeda-Priego, J. Phys.: Condens. Matter **33**, 053001 (2021).

[17] A. Chattopadhyay, E. Nabizadeh, and P. Hassanzadeh, J. Adv. Model. Earth Syst. **12** (2020).

[18] S. Arik, C.-L. Li, J. Yoon, R. Sinha, A. Epshteyn, L. Le, V. Menon, S. Singh, L. Zhang, M. Nikoltchev, Y. Sonthalia, H.

Nakhost, E. Kanal, and T. Pfister, in *Advances in Neural Information Processing Systems*, Vol. 33, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Red Hook, NY, 2020), pp. 18807–18818.
[19] E. N. Lorenz, Bull. Am. Meteorol. Soc. **50**, 345 (1969).
[20] E. N. Lorenz, J. Atmos. Sci. **26**, 636 (1969).
[21] M. Cristelli, A. Tacchella, and L. Pietronero, PLoS One **10**, e0117174 (2015).
[22] A. Tacchella, D. Mazzilli, and L. Pietronero, Nat. Phys. **14**, 861 (2018).
[23] A. Tacchella, A. Napoletano, and L. Pietronero, PLoS One **15**, e0230107 (2020).
[24] T. M. Hamill and J. S. Whitaker, Monthly Weather Rev. **134**, 3209 (2006).
[25] L. Delle Monache, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, Monthly Weather Rev. **141**, 3498 (2013).
[26] F. Cecconi, M. Cencini, M. Falcioni, and A. Vulpiani, Am. J. Phys. **80**, 1001 (2012).
[27] J. D. Farmer and J. J. Sidorowich, Phys. Rev. Lett. **59**, 845 (1987).
[28] R. Lguensat, P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet, Mon. Weather Rev. **145**, 4093 (2017).
[29] P. Platzer, P. Yiou, P. Naveau, P. Tandeo, Y. Zhen, P. Ailliot, and J.-F. Filipot, J. Atmos. Sci. **78**, 2117 (2021).
[30] M. H. Neumann, Ann. Statist. **26**, 2014 (1998).
[31] M. H. Neumann, Statistics **36**, 33 (2002).
[32] E. Paparoditis and D. N. Politis, J. Stat. Planning Infe. **108**, 301 (2002).
[33] V. Monbet and P.-F. Marteau, J. Stat. Plan. Infer. **136**, 3319 (2006).
[34] P. Yiou, Geosci. Model Dev. **7**, 531 (2014).
[35] K. Okuyama, M. Takayasu, and H. Takayasu, Physica A **269**, 125 (1999).
[36] Y. U. Saito, T. Watanabe, and M. Iwamura, Physica A **383**, 158 (2007).
[37] H. Watanabe, H. Takayasu, and M. Takayasu, Physica A **392**, 741 (2013).
[38] G. B. West, in *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies* (Penguin, New York, 2017), Chap. 9.
[39] C. Dang, Z. (Frank) Li, and C. Yang, J. Bank. Finance **86**, 159 (2018).
[40] Y. Kobayashi, H. Takayasu, S. Havlin, and M. Takayasu, Entropy **23**, 168 (2021).
[41] L. G. A. Alves, H. V. Ribeiro, E. K. Lenzi, and R. S. Mendes, PLoS One **8**, e69580 (2013).
[42] L. Alves, H. Ribeiro, E. Lenzi, and R. Mendes, Physica A **409**, 175 (2014).
[43] M. Kleiber, Physiol. Rev. **27**, 511 (1947).
[44] W. R. Stahl, Science **150**, 1039 (1965).
[45] K. Schmidt-Nielsen, *Scaling: Why is Animal Size So Important?* (Cambridge University Press, Cambridge, UK, 1984).
[46] V. M. Savage, J. F. Gillooly, W. H. Woodruff, G. B. West, A. P. Allen, B. J. Enquist, and J. H. Brown, Functional Ecol. **18**, 257 (2004).
[47] O. Arrhenius, J. Ecol. **9**, 95 (1921).
[48] M. L. Rosenzweig, *Species Diversity in Space and Time* (Cambridge University Press, Cambridge, UK, 1995).
[49] G. B. West, J. H. Brown, and B. J. Enquist, Science **276**, 122 (1997).
[50] K. Yakubo, Y. Saijo, and D. Korošak, Phys. Rev. E **90**, 022803 (2014).
[51] J. Kwapień and S. Drożdż, Phys. Rep. **515**, 115 (2012).
[52] H. A. Simon, *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*, 1st ed. (Macmillan, New York, 1947).
[53] A. D. J. Chandler, *The Visible Hand: The Managerial Revolution in American Business* (Belknap Press, Cambridge, MA, 1977).
[54] M. H. R. Stanley, L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M. A. Salinger, and H. E. Stanley, Nature (London) **379**, 804 (1996).
[55] Y. Kobayashi, H. Takayasu, S. Havlin, and M. Takayasu, New J. Phys. **21**, 043038 (2019).
[56] K. Tamura, W. Miura, M. Takayasu, H. Takayasu, S. Kitajima, and H. Goto, Int. J. Modern Phys.: Conf. Ser. **16**, 93 (2012).
[57] H. Goto, E. Viegas, H. J. Jensen, H. Takayasu, and M. Takayasu, Sci. Rep. **7**, 5064 (2017).
[58] V. Latora, V. Nicosia, and G. Russo, *Complex Networks* (Cambridge University Press, Cambridge, UK, 2017).
[59] R. Cohen and S. Havlin, *Complex Networks* (Cambridge University Press, Cambridge, UK, 2010).
[60] M. Newman, *Networks*, 2nd ed. (Oxford University Press, Oxford, UK, 2018).
[61] W. Miura, H. Takayasu, and M. Takayasu, Phys. Rev. Lett. **108**, 168701 (2012).
[62] M. Takayasu, S. Sameshima, T. Ohnishi, Y. Ikeda, H. Takayasu, and K. Watanabe, Annual Report of the Earth Simulator (Japan Agency for Marine-Earth Science and Technology) **6**, 263 (2008).
[63] T. Iino and H. Iyetomi, in *The Economics of Interfirm Networks*, edited by T. Watanabe, I. Uesugi, and A. Ono (Springer, Tokyo, 2015), Chap. 3, pp. 39–65.
[64] R. L. Axtell, Science **293**, 1818 (2001).
[65] E. Gaffeo, M. Gallegati, and A. Palestrini, Physica A **324**, 117 (2003).
[66] Y. Fujiwara, C. Di Guilmi, H. Aoyama, M. Gallegati, and W. Souma, Physica A **335**, 197 (2004).
[67] P. Cirillo and J. Hüsler, Physica A **388**, 1546 (2009).
[68] T. Ogwang, Empir. Econ. **41**, 473 (2011).
[69] S. Da Silva, R. Matsushita, R. Giglio, and G. Massena, Physica A **512**, 68 (2018).
[70] R. Pascoal, M. Augusto, and H. Rocha, Physica A **531**, 121797 (2019).
[71] S. Picoli and R. S. Mendes, Phys. Rev. E **77**, 036105 (2008).
[72] D. Fu, F. Pammolli, S. V. Buldyrev, M. Riccaboni, K. Matia, K. Yamasaki, and H. E. Stanley, Proc. Natl. Acad. Sci. USA **102**, 18801 (2005).
[73] M. Takayasu, H. Watanabe, and H. Takayasu, J. Stat. Phys. **155**, 47 (2014).
[74] R. Cerqueti, P. Falbo, and C. Pelizzari, Eur. J. Oper. Res. **256**, 163 (2017).
[75] Y. Malevergne*, V. Pisarenko, and D. Sornette, Quant. Financ. **5**, 379 (2005).
[76] H. Akaike, in *Selected Papers of Hirotugu Akaike* (Springer, New York, 1998), pp. 199–213.
[77] T. Mizuno, M. Katori, H. Takayasu, and M. Takayasu, in *Empirical Science of Financial Fluctuations*, edited by H. Takayasu (Springer, Tokyo, Japan, 2002), pp. 321–330.

[78] D. L. Turcotte, Rep. Prog. Phys. **62**, 1377 (1999).
[79] J. Sutton, Physica A **312**, 577 (2002).
[80] J. Hu, W.-w. Tung, J. Gao, and Y. Cao, Phys. Rev. E **72**, 056207 (2005).
[81] J. B. Gao, J. Hu, W. W. Tung, and Y. H. Cao, Phys. Rev. E **74**, 066204 (2006).
[82] S. Ramdani, F. Bouchara, and J.-F. Casties, Phys. Rev. E **76**, 036204 (2007).
[83] D. Naro, C. Rummel, K. Schindler, and R. G. Andrzejak, Phys. Rev. E **90**, 032913 (2014).
[84] A. Ji and P. Shang, Physica A **534**, 122038 (2019).
[85] W. Marszalek, M. Walczak, and J. Sadecki, IEEE Access **7**, 183245 (2019).
[86] J. Belaire-Franch, Physica A **555**, 124733 (2020).
[87] S. Datta and W. P. Mccormick, Can. J. Stat. **21**, 181 (1993).