Testing Statistical Laws in Complex Systems

Martin Gerlach[1] and Eduardo G. Altmann[2]
[1]*Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, USA*
[2]*School of Mathematics and Statistics, University of Sydney, 2006 NSW, Australia*

(Received 19 September 2018; revised manuscript received 19 December 2018; published 26 April 2019)

The availability of large datasets requires an improved view on statistical laws in complex systems, such as Zipf's law of word frequencies, the Gutenberg-Richter law of earthquake magnitudes, or scale-free degree distribution in networks. In this Letter, we discuss how the statistical analysis of these laws are affected by correlations present in the observations, the typical scenario for data from complex systems. We first show how standard maximum-likelihood recipes lead to false rejections of statistical laws in the presence of correlations. We then propose a conservative method (based on shuffling and undersampling the data) to test statistical laws and find that accounting for correlations leads to smaller rejection rates and larger confidence intervals on estimated parameters.

DOI: 10.1103/PhysRevLett.122.168301

Introduction.—Statistical regularities collected in the form of "universal laws" play a central role in complex systems [1–3]. Zipf's law of word frequencies [4], the Gutenberg-Richter law of earthquake magnitudes [5], scale-free degree distributions in networks [6], and inter-event time distributions between bursty events [7–10] are prominent examples that triggered entire research lines devoted to explaining the origin and to exploring the consequences of these laws.

Recently, the empirical support of such laws has been heavily questioned. The best known example is the case of scale-free degree distribution of networks; after the seminal work of Barabasi and Albert in 1999 [6], the early 2000s were marked by findings of power law distributions in various network datasets, while in the last five years the trend has reversed and it is now common to read that networks with power law degree distribution are rare [11,12] (see Ref. [13] for a journalistic account). This recent shift in conclusions, which appears in the analysis of Zipf's law in language [3,14,15] and also in other areas [16,17], is partially due to new (larger) datasets but mostly due to the improved statistical methods: least-squared fitting and visual inspection of double-logarithmic plots (used since Zipf) have been replaced by maximum likelihood methods made popular in the influential article by Clauset, Shalizi, and Newman [17], see Refs. [18–21] for variations. A point often ignored in the interpretations of the recent findings is that these methods rely on two hypotheses:

H1: The observations x are distributed as $p(x;\vec{\alpha})$, where $\vec{\alpha}$ are parameters, e.g., for a power law

$$p(x;\alpha) = Cx^{-\alpha}. \tag{1}$$

H2: The empirical observations x_i, $i = 1, \ldots, N$ are independent (e.g., of i or x_{i-1}).

While the statistical laws correspond to H1, the statistical tests rely also on H2 [implicitly assumed, e.g., when the log-likelihood is computed as $\sum_{i=1}^{N} \log p(x_i)$ [5,12,17,21,22]]. Complex systems are characterized by strong (temporal and spatial) interdependencies [23] and it is thus not clear whether the recent claims [11,12,16] of violation of the statistical laws arise from systematic deviations of the law itself (H1) or, instead, whether they are due to the well-known fact that observations are not independent (H2).

In this Letter we show that dependencies in the data (violation of H2) have a strong impact on the empirical analysis of statistical laws, leading to rejections even in processes that satisfy the law (H1), and to overconfident selection of models and parameters. We then propose an alternative method that distinguishes between H1 and H2, yielding an upper bound on the degree of correlations for which the statistical law is rejected.

General setting.—Let $\{x_i\} = x_1, x_2, \ldots, x_N$ be an ordered sequence obtained from a measurement process that asymptotically has a well-defined distribution $p(x) = (\#x_i = x/N)$ as $N \to \infty$. In observations of dynamical systems (or time series), x_i will typically depend on the observations at previous times so that for all times τ smaller than some (relaxation) time τ^* we find $p(x_i|x_{i-\tau}) \neq p(x)$. Violations of H2 happen also when data is not measured as a time series. In the case of Zipf's law of word frequencies, syntax restrict the valid sequences of word tokens, in violation of H2 (both in the rank-frequency and frequency distribution pictures [3,15]). In the case of networks, H2 can be violated because of the generative process or because of the sampling employed to *observe* the nodes and links (typically a subsample of an underlying network). In fact, it has been shown that the degree distribution of networks is sensitive to the sampling procedure [24–26].

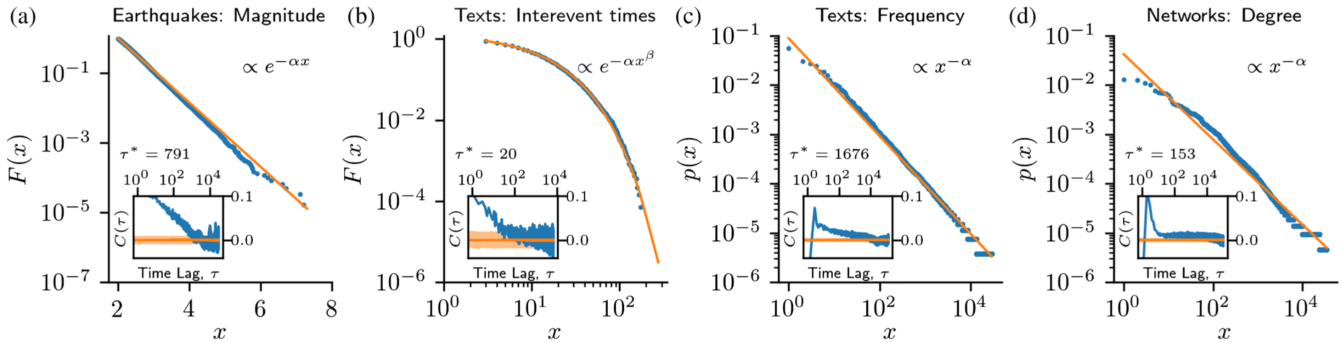

FIG. 1. Statistical laws and strong correlations occur simultaneously in complex systems. Main Panels: Distribution $p(x)$ [or its cumulative $F(x)$] of observable x for the data (blue dots) and maximum-likelihood fit of different statistical laws (orange, see Supplemental Material [30], Sec. I). Insets: autocorrelation $C(\tau)$ with time lag τ of the observable x for the original data (blue) and randomized data (orange, average and 1 or 99 percentiles over 1000 realizations); τ^* indicates the value at which the original and randomized $C(\tau)$ are statistically indistinguishable [31]. (a) Sequence of magnitudes x of earthquakes in Southern California from 1981–2010 [32] ($N = 59555$ with commonly used threshold $x \geq 2$ [33]). (b) Sequence of interevent times x (measured in words) of consecutive occurrences of the word "the" in the book *Moby Dick* obtained from Project Gutenberg [34] ($N = 14042$ with threshold $x \geq 3$). (c) Sequence of words (tokens) in the order they appear in the book *Ulysses* by James Joyce obtained from Project Gutenberg [34]; x the rank of the word (type) in terms of frequency in the whole book ($N = 264971$ word tokens, obtained removing punctuation and nonalphabetic characters). (d) Sequence of degrees of nodes from a network; x is the rank of the degree of the node; the sequence $\{x_i\}$ used to compute $C(\tau)$ was obtained applying an edge-sampling method to the complete network (see Supplemental Material [30], Sec. II); the network corresponds to the connections between autonomous systems of the Internet [35], $V = 34761$ vertices (nodes) and $E = 107720$ unique edges (in our case N is the number of half edges and thus $N = 2E$).

Moreover, the hypotheses H1 and H2 of the standard tests for power law distribution do not build a proper probabilistic network model [27], are thus not suitable to a rigorous statistical analysis [29], and the analysis of the degree distribution of networks requires further assumptions about the sampling or generative process.

More generally, strong correlations are ubiquitous in complex systems [23] and it is hard to imagine a case for which H2 holds. In Fig. 1 we show how previously proposed statistical laws and correlations appear together in paradigmatic complex systems: the Gutenberg-Richter law for earthquakes (exponential [5]), interevent times of words (stretched exponential [7–9]), Zipf's law for word frequencies (power law [4]), and scale-free distribution for the node degree in networks (power law [6]). While earthquake events and interevent times naturally occur as time series data, we mapped word frequencies in texts and the network data into ordered sequences $\{x_i\}$ based on a simple sampling process (see caption of Fig. 1) in order to illustrate and quantify the violation of H2 in an unified framework.

Constructed example.—We now show that the traditional methods [17] lead to a rejection of a power law distribution [Eq. (1)] even for data which are power law distributed for $N \to \infty$. This is done by building a Markov process [36,37] in which H1 is satisfied but H2 is violated [i.e., x_i depends on x_{i-1} and $p(x) = Cx^{\alpha}$ for $N \to \infty$, see the Supplemental Material [30], Sec. III].

In Fig. 2 we show that the violations of H2 have a strong influence on the analysis of statistical laws formulated in H1. In particular, the application of the traditional recipes [17] lead to the wrong conclusion that the data are not compatible with a power law distribution: the probability of rejecting the null hypothesis at a 5% significance level is much larger than 5% even for small sample sizes N [inset of Fig. 2(b)]. This corresponds to a type-I error because, by construction, the data satisfy H1. The origin of this failure thus originates from the fact that correlations lead to an effective reduction of the number of independent observations implying larger fluctuations which lead to larger deviations from the fitted model. Specifically, we recall that the test employed in Ref. [17] consists of comparing the Kolmogorov-Smirnov (KS) distance between the correlated data and the fitted curve, $\mathrm{KS}_{\mathrm{correlated}}$ (blue curve), and the KS distance between independent samples of the model (H1 + H2) and the fitted curve, $\mathrm{KS}_{\mathrm{model}}$ (orange curve). More precisely, the statistical law is rejected at 5% significance level if $\mathrm{KS}_{\mathrm{correlated}} > \mathrm{KS}_{\mathrm{model}}$ in 95% realizations (samplings) of the model. While in our artificial data $\mathrm{KS}_{\mathrm{correlated}} \propto 1/\sqrt{N}$ (as expected) and thus $\mathrm{KS}_{\mathrm{correlated}} \to 0$ for $N \to \infty$, this convergence is shifted from the convergence of $\mathrm{KS}_{\mathrm{model}}$ [Fig. 2(b)] due to the correlations. This shift leads to an increased rejection rate (≈ 1, p value ≈ 0).

Violations of H2 are important not only in the hypothesis-testing setting discussed above, they also lead to increased systematic and statistical errors (bias and fluctuations) in the fitting of the parameter $\hat{\alpha}$ [Fig. 2(c)] and, thus, in the selection between models [38,39].

Real data.—In order to confirm that the results discussed above are also relevant in real datasets—which have a fixed size N—we consider two types of undersampling of data to

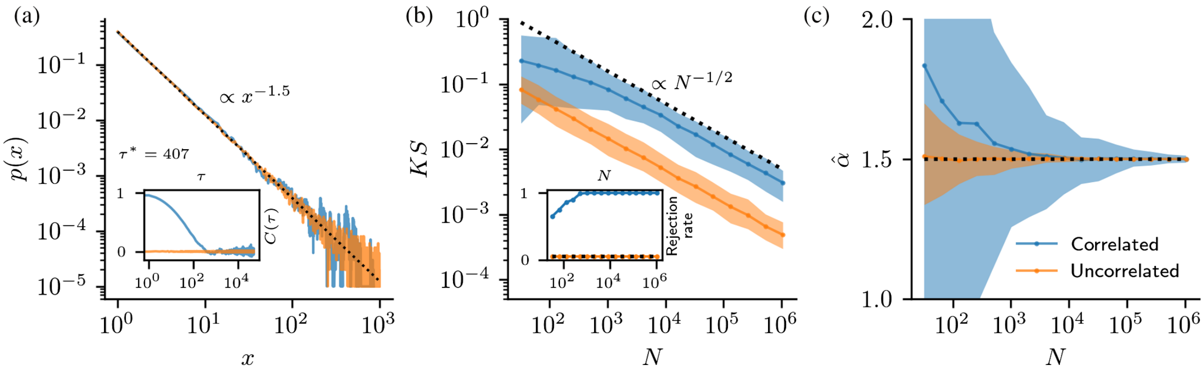

FIG. 2. Correlations impact the fitting of power law distributions using Maximum Likelihood methods. Two synthetic datasets following a power law with exponent $\alpha = 1.5$ for $x = 1, \ldots, 1000$ were generated: one using independent sampling (in orange or light gray) and one with correlations (in blue or dark gray); see the Supplemental Material [30], Sec. II, for different choices of the maximum cutoff leading to similar results. (a) Distribution $p(x)$ for a single realization with $N = 10^5$. Inset: Autocorrelation function $C(\tau)$. (b) Average and 95% confidence interval of the KS distance over 100 different realizations of the synthetic data. Inset: Rejection rate, i.e., fraction of realizations for which the power law is rejected on a 0.05-significance level according to method of Ref. [17] (dotted line) for datasets of varying length N. (c) Average and 95% confidence interval of the estimated power law exponent $\hat{\alpha}$ over 100 different realizations of the synthetic data.

sizes $n < N$: taking n points either randomly or preserving the structures or correlations by taking consecutive portions of the time series (the network and word-frequency databases are first mapped to a time series, as in Fig. 1). In order to distinguish between the effect of the shape of the distribution (H1) and correlations (H2) we compare the distribution of the n points with (i) the proposed statistical law and (ii) the empirical distribution (i.e., the one obtained for $n = N$). Our results (see the Supplemental Material [30], Sec. IV) confirm that correlated data show higher rejection rate and fluctuations of parameters.

Alternative approach.—In the vast literature of statistical methods for dependent data, two general approaches can be identified. The first approach is to incorporate the violation of independence in more sophisticated (parametric) models, e.g., in a time series one could consider Gaussian Markov processes [40]. This is of limited use in our case because statistical laws aim to provide a coarse-grained description (stylized facts) valid in many systems, instead of different detailed models of particular cases. The second (nonparametric) approach, which we pursue here, is to decorrelate or decluster the data, leading to a dataset with an "effective" sample size $N^* \leq N$ [41–43]. In practice, the analysis consists of multiple realizations of the following three steps. (i) Randomize (shuffle) the original sequence and select randomly n points, for different $n \in [1, N]$. (ii) Apply the traditional statistical analysis (i.e., the hypothesis test, model comparison, and fitting based on H1 + H2) to the randomized dataset obtained in (i), investigating their dependence on n. (iii) Estimate the correlation τ^*, defined as the time after which two observations (in the time series) are independent from each other. Out of the total N samples we thus estimate $N^* = N/\tau^*$ to be the number of independent samples and therefore we select the results from step (ii) for $n \approx N^*$.

The determination of τ^*—or the effective sample size N^*—in step (iii) requires knowledge or assumptions about how the data were generated. For the case of temporal sequences we propose to compute the autocorrelation and take as τ^* the lag for which it reaches an interval around zero (1 percentile of the random realizations, as in Fig. 1). In the constructed example (Fig. 2), we obtain $\tau^* = 407 \Rightarrow N^* = 245$, which leads to a rejection rate (at p value $= 0.05$) equal to 5% for all $n < N^*$. For the case of networks, the determination of the effective sample size N^* depends on the generative process and/or the sampling used to measure the data (here we assumed a specific edge sampling method, as described in Fig. 1.) In Fig. 3, we show evidence of the effectiveness of our approach through a systematic analysis of the p value distribution as a function of n for both the constructed and empirical datasets. This is further corroborated in artificial data (see the Supplemental Material [30], Sec. V) showing that (i) our method for the selection of τ^* is superior to the one proposed in Ref. [41] (sum of the autocorrelation function) and (ii) can be equally applied to data with other types of correlation: a Markov process with negative correlation and a Gaussian process with long-range correlations. In all cases our approach shows an uniform distribution of p values under the null hypothesis.

An important message of our analysis is that conclusions about the statistical law can be obtained even when the precise value of τ^* (or the effective sample size N^*) is unknown in step (iii). By shuffling and undersampling the sequence at different sizes n—steps (i) and (ii)—we can investigate how the results depend on n and obtain the range in τ^* for which the different conclusions hold. For instance, in the case of earthquakes [Fig. 3(c)] we see that the rejection increases dramatically around $n \approx 10^3$. We thus conclude that, in this dataset of size $N \approx 10^5$, we

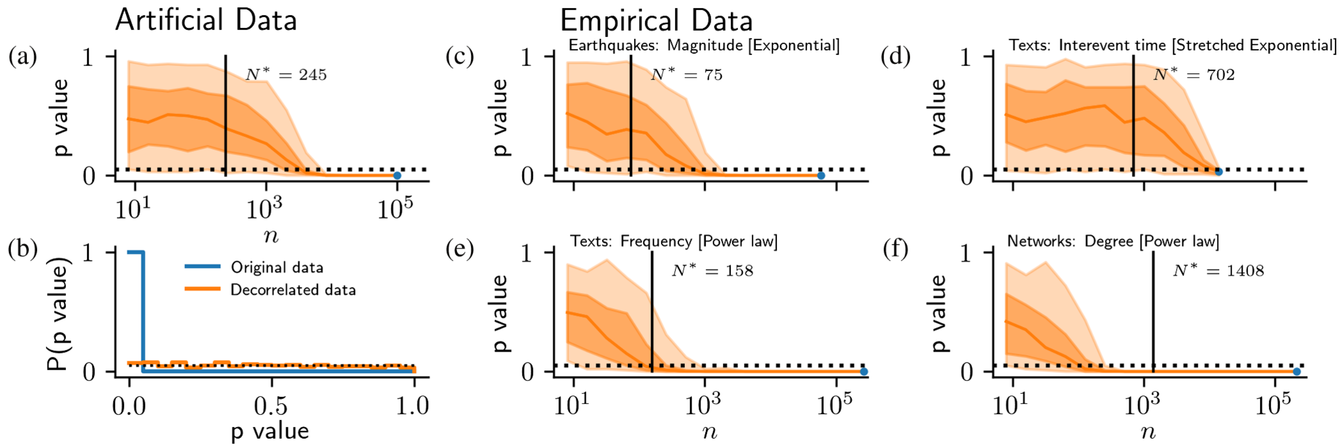

FIG. 3. Decorrelating data leads to different conclusions in hypothesis testing in artificial (a) and (b) and empirical (c)–(f) data. The distribution of p values from fitting correlated (of size N, in blue) and subsampled shuffled data (of size $n \leq N$, in orange). While the correlated data leads to a peaked distribution of small p values (i.e., rejection), the decorrelated data obtained from our approach leads to the expected flat distribution of p values. The effective sample size N^* (black vertical line) was obtained from τ^* reported in Fig. 1 as $N^* = N/\tau^*$. While all cases are rejected when fitting the full dataset, in three out of four cases we cannot reject the null hypothesis for decorrelated data (median p value ≥ 0.05 at $n = N^*$).

falsify the Gutenberg-Richter law if $\tau^* \leq N/n \approx 10^2$ observations ≈20 days [44]). The conservative estimate of τ^* in Fig. 1 was $\tau^* = 791 > 10^2$ and therefore we conclude that based on this data we cannot reject the Gutenberg-Richter law, contrary to the conclusion obtained assuming independent observations. We find similar results [Fig. 3(d)] for the stretched exponential distribution of inter-event times between words, while for Zipf's law [Fig. 3(c)] the outcome is uncertain, and the power law degree distributions in networks [Fig. 3(d)] is rejected even in the correlated case.

Discussion and Conclusion.—Statistical laws in complex systems are typically formulated (as in H1) without reference to the generative process of the data. Therefore, ideally, the empirical test of these laws should be designed to account for a large class of processes generating $\{x_i\}$. Traditional methods [17] based on the hypothesis of independent data (H2) are weak tests because they include a strong hypothesis that is easily violated, therefore favoring rejection. In fact, here we have shown how these methods (i) lead to wrong rejections of the laws because of correlated data, and (ii) are over-optimistic regarding uncertainties of the estimated parameters. Stronger tests of statistical laws should make weaker assumptions about the generative process so that rejections of the compound hypothesis provide much stronger evidence of the rejection of the law (H1). Here we proposed a methodology which allows us to identify the strongest assumption about correlations of the data τ^* for which the law can be rejected. Being conservative in the choice of τ^* (i.e., choosing large values for which we are confident that x_i and $x_{i+\tau^*}$ are uncorrelated) overcomes the main shortcoming of the traditional approach [17] and ensures that when we reject the law this is not happening due to correlations in the data (failing to reject the law is never a confirmation of its validity). In this sense, our approach is similar in spirit to the Bonferroni correction to account for multiple hypothesis testing [45] (both aim to avoid overconfident or spurious rejections of hypotheses). Our approach is even applicable in cases with no well-defined mixing time τ^* (e.g., long-range correlations) because it yields very large values of $\tau^* \approx N$ (no two points are independent).

Instead of directly testing whether the statistical law is valid (hypothesis testing), often the best we can do is to compare different alternatives (model comparison) [3,11,14,17,22,38,39]. Also in this case, violations of the hypothesis of independence are important and have been mostly ignored in the analysis of statistical laws in complex systems (see Refs. [5,12,25] for exceptions). As shown above, due to correlations (and violations of H2) actual data show much larger fluctuations than expected under the hypothesis of independent observations. By using a shuffled and undersampled dataset, we obtain larger uncertainties in the estimated parameters; we expect similar lack of certainty in the choice of best models. The need to account for violations of the independence assumption, shown in this Letter, applies much more broadly than the cases treated above. Correlations should be accounted for whenever testing statistical laws in complex systems, such as linguistic laws [3], scaling laws with system size—maximum likelihood methods based on H2 have been applied to biological allometric laws [46] and to city data [47]—and different distributions of interevent time (burstiness) [7–10].

The code and data shown in this Letter can be obtained by following the link in Ref. [48].

We thank F. Font-Clos and J. Moore for the careful reading of the manuscript.

[1] M. Mitzenmacher, Internet Math. **1**, 226 (2004).
[2] M. E. J. Newman, Contemp. Phys. **46**, 323 (2005).
[3] E. G. Altmann and M. Gerlach, *Creativity and Universality in Language*, edited by M. Degli Esposti, E. G. Altmann, and F. Pachet, Lecture Notes in Morphogenesis (Springer, New York, 2016).
[4] G. K. Zipf, *The Psycho-Biology of Language* (Routledge, London, 1936); *Human Behavior and the Principle of Least Effort* (Addison-Wesley Press, Oxford, 1949).
[5] V. F. Pisarenko and D. Sornette, Pure Appl. Geophys. **161**, 839 (2004).
[6] A. L. Barabási and R. Albert, Science **286**, 509, (1999).
[7] E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter, PLoS One **4**, e7678 (2009).
[8] A. Corral, R. Ferrer-i-Cancho, G. Boleda, and A. Diaz-Guilera, arXiv:0901.2924.
[9] A. Bunde, J. F. Eichner, J. W. Kantelhardt, and S. Havlin, Phys. Rev. Lett. **94**, 048701 (2005).
[10] A.-L. Barabási, Nature (London) **435**, 207 (2005).
[11] A. D. Broido and A. Clauset, Nat. Commun. **10**, 1017 (2019).
[12] R. Khanin and E. Wit, J. Comput. Biol. **13**, 810 (2006).
[13] E. Klarreich, Quanta Magazine, https://www.quantamagazine.org/scant-evidence-of-power-laws-found-in-real-world-networks-20180215/.
[14] M. Gerlach and E. G. Altmann, Phys. Rev. X **3**, 021006 (2013).
[15] I. Moreno-Sánchez, F. Font-Clos, and A. Corral, PLoS One **11**, 0147073 (2016).
[16] M. P. H. Stumpf and M. A. Porter, Science **335**, 665 (2012).
[17] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Rev. **51**, 661 (2009).
[18] M. L. Goldstein, S. A. Morris, and G. G. Yen, Eur. Phys. J. B **41**, 255 (2004).
[19] H. Bauke, Eur. Phys. J. B **58**, 167 (2007).
[20] A. Deluca and A. Corral, Acta Geophysica Polonica **61**, 1351 (2013).
[21] R. Hanel, B. Corominas-Murtra, B. Liu, and S. Thurner, PLoS One **13**, 0170920 (2017).
[22] M. P. H. Sumpf and P. J. Ingram, Europhys. Lett. **71**, 152 (2005).
[23] Z. Eisler, I. Bartos, and J. Kertész, Adv. Phys. **57**, 89 (2008).
[24] M. P. H. Stumpf, C. Wiuf, and R. M. May, Proc. Natl. Acad. Sci. U.S.A. **102**, 4221 (2005).
[25] S. H. Lee, P.-J. Kim, and H. Jeong, Phys. Rev. E **73**, 016102 (2006).
[26] M. P. H. Stumpf and C. Wiuf, Phys. Rev. E **72**, 036118 (2005).
[27] Consider x_is to be a sequence of degrees of a network sampled from H1 and H2. For large networks, only half of the realizations lead to graphical degree sequences [28].
[28] R. Arratia and T. M. Liggett, Ann. Appl. Probab. **15**, 652 (2005).
[29] H. Crane, *Probabilistic Foundations of Statistical Network Analysis* (Chapman and Hall/CRC, New York, 2018).
[30] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.122.168301 for additional text and figures supporting our results.
[31] For the data with assumed power law distributions, we calculate the autocorrelation for the variable log x. We determine τ^* as the minimum τ for which the lower bound $C(\tau)$ (1 percentile) of the original data is smaller or equal than the upper bound $C(\tau)$ (99 percentile) of the randomized data. We generate an ensemble of "original" time series of the same length N by selecting a random point as the first observation and using periodic boundary conditions.
[32] Southern California Earthquake Data Center, http://scedc.caltech.edu/research-tools/alt-2011-yang-hauksson-shearer.html.
[33] A. Corral, Phys. Rev. Lett. **92**, 108501 (2004).
[34] Project Gutenberg, http://www.gutenberg.org.
[35] KONECT Project: Internet topology, http://konect.cc/networks/topology.
[36] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods, Springer Texts in Statistics*, 2nd ed. (Springer, Berlin, 2005).
[37] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times* (AMS, Providence, Rhode Island, 2017).
[38] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2009).
[39] K. P. Burnham and D. R. Anderson, *Model selection and multimodal inference: A practical information-theoretic approach* (Spinger, New York, 2002).
[40] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, New York, 1992).
[41] T. Gasser, Biometrika **62**, 563 (1975).
[42] M. S. Weiss, J. Am. Stat. Assoc. **73**, 872 (1978).
[43] R. Chicheportiche and J.-P. Bouchaud, J. Stat. Mech. (2011) P09003.
[44] Assuming observations occur at an approximately constant rate over the 30 years of available data.
[45] J. P. Shaffer, Annu. Rev. Psychol. **46**, 561 (1995).
[46] P. S. Dodds, D. H. Rothman, and J. S. Weitz, J. Theor. Biol. **209**, 9 (2001).
[47] J. C. Leitao, J. M. Miotto, M. Gerlach, and E. G. Altmann, R. Soc. Open Sci. **3**, 150649 (2016).
[48] Codes and datasets are available at https://doi.org/10.5281/zenodo.2641375.