PHYSICAL REVIEW E **90**, 022806 (2014)

Complex architecture of primes and natural numbers

Guillermo García-Pérez, M. Ángeles Serrano, and Marián Boguñá
Departament de Física Fonamental, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain
(Received 3 April 2014; published 12 August 2014)

Natural numbers can be divided in two nonoverlapping infinite sets, primes and composites, with composites factorizing into primes. Despite their apparent simplicity, the elucidation of the architecture of natural numbers with primes as building blocks remains elusive. Here, we propose a new approach to decoding the architecture of natural numbers based on complex networks and stochastic processes theory. We introduce a parameter-free non-Markovian dynamical model that naturally generates random primes and their relation with composite numbers with remarkable accuracy. Our model satisfies the prime number theorem as an emerging property and a refined version of Cramér's conjecture about the statistics of gaps between consecutive primes that seems closer to reality than the original Cramér's version. Regarding composites, the model helps us to derive the prime factors counting function, giving the probability of distinct prime factors for any integer. Probabilistic models like ours can help to get deeper insights about primes and the complex architecture of natural numbers.

DOI: 10.1103/PhysRevE.90.022806 PACS number(s): 89.75.Da, 02.10.De

I. INTRODUCTION

Prime numbers have fascinated and puzzled philosophers, mathematicians, physicists, and computer scientists alike for the past two and a half thousand years. A prime is a natural number that has no divisors other than 1 and itself; every natural number greater than 1 that is not a prime is called a composite. Despite the apparent simplicity of these definitions, the hidden structure in the sequence of primes and their relation with the set of natural numbers are not yet completely understood [1]. There is no practical closed formula that sets apart all of the prime numbers from composites [2], and many questions about primes and their distribution among the set of natural numbers still remain open. Indeed, most of the knowledge about the sequence of primes stands on unproved theorems and conjectures.

The mystery of primes is not a mere conundrum of pure mathematics. Unexpected connections can be discovered between primes and different topics in physics. For instance, the Riemann ζ function $\zeta(s)$—a sum over all integers equivalent to a product over all primes—has been considered as a partition function [3–5], such that its sequence of nontrivial zeros—encoding information about the sequence of primes—is similar to the distribution of eigenvalues of random Hermitian matrices used in classically chaotic quantum systems to describe the energy levels in the nuclei of heavy elements [6]. This idea traces back to the Hilbert-Pólya conjecture [7], which states that the zeros of the $\zeta(s)$ function might be the eigenvalues of some Hermitian operator on a Hilbert space. Indeed, the Riemann ζ function plays an integral role not only in quantum mechanics but in different branches of physics, from classical mechanics to statistical physics [8]. The interpretation of prime numbers or the Riemann ζ zeros as energy eigenvalues of particles appears also in statistical mechanics, as illustrated for instance by the Riemann gas concept as a toy model for certain aspects of string theory [9]. Recently, interesting connections have also been found between primes and self-organized criticality [10], or primes and quantum computation [11,12] (see Ref. [13] for an extensive bibliographical survey between the connection of number theory and physics). The importance of primes transcend theoretical aspects, and practical applications include public key cryptography algorithms [14] and pseudorandom number generators [15].

One of the most promising approaches to solve the enigmas of number theory is the use of probability theory and stochastic processes. Akin to chaotic dynamical systems, prime numbers, albeit purely deterministic, appear to be scattered throughout natural numbers in a nonhomogeneous random fashion. Indeed, for $n \gg 1$ the probability that a randomly chosen number in a "small" neighborhood of n is prime is given by [16]

$$P_n \sim \frac{1}{\ln n}. \tag{1}$$

This is equivalent to the well-known prime number theorem [17], which states that the prime counting function $\pi(N)$—counting the number of primes up to N—approaches $N/\ln N$ in the limit of $N \to \infty$, i.e.,

$$\pi(N) \sim \int_2^N \frac{dx}{\ln x} \equiv \mathrm{Li}(N) \sim \frac{N}{\ln N}, \tag{2}$$

where $\mathrm{Li}(N)$ is the offset logarithmic integral function. Taking advantage of this apparent randomness, Cramér formulated a simple model [18,19] where each integer n is declared as a "prime" with independent probability given by Eq. (1). The model—which generates sequences of random primes that are, obviously, in agreement with the prime number theorem—allowed him to "prove," in a probabilistic sense, his famous conjecture about gaps between consecutive primes [19].

Cramér's probabilistic model plays, still today, a fundamental role when formulating conjectures concerning primes. However, it presents three major drawbacks: (1) It does not "explain" the prime number theorem; instead, it is an input of the model. (2) Random primes in the model are totally uncorrelated, whereas there are both short- and long-range correlations in the sequence of real primes. (3) Finally, it says nothing about the relation between prime and composite numbers. In this paper, we combine a complex network approach with the theory of stochastic processes to introduce a parameter-free non-Markovian dynamical model that naturally generates random primes as well as the relation between primes and composite numbers with remarkable accuracy. Our

model is in agreement with Eqs. (1) and (2) and satisfies a modified version of Cramér's conjecture about the statistics of gaps between consecutive primes that seems closer to reality than the original Cramér's version. Regarding composites, the model helps us to derive the prime factors counting function, giving the probability of distinct prime factors for any integer.

II. BIPARTITE NETWORK OF NATURAL NUMBERS

Primes are the building blocks of natural numbers. The fundamental theorem of arithmetic states that any natural number $n > 1$ can be factorized uniquely as

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k} \cdots, \tag{3}$$

where p_i is the ith prime and α_i are nonnegative integers. From a complex networks perspective, natural numbers can be thought of as a weighted bipartite network with two types of nodes, primes and composites. A composite n is linked to primes p_i with weights α_i according to the factorization in Eq. (3), as shown in Fig. 1.

For a given network size N, the probability that a randomly chosen prime inside the network is connected to k_p different composites, that is, the degree distribution $P(k_p)$ for prime numbers, can be exactly determined in terms of the prime counting function as (see Appendix A for details)

$$P(k_p) = \frac{\pi\left(\frac{N}{k_p+1}\right) - \pi\left(\frac{N}{k_p+2}\right)}{\pi(N)}, \tag{4}$$

with $k_p = 0, 1, \cdots, \lfloor \frac{N}{2} \rfloor$, where $\lfloor x \rfloor$ stands for the floor function. Using the prime number theorem, Eq. (2), it is easy to see that in the limit $N/k_p \gg 1$ this distribution behaves as $P(k_p) \sim k_p^{-2}$. Quite surprisingly, we obtain a scale-free network with an exponent -2, very similar to many real complex networks, like the Internet [20], and similar to the degree distribution of the causal graph of the de Sitter space-time [21]. As we shall show, this is a consequence of an effective preferential attachment rule induced by the growth mechanism.

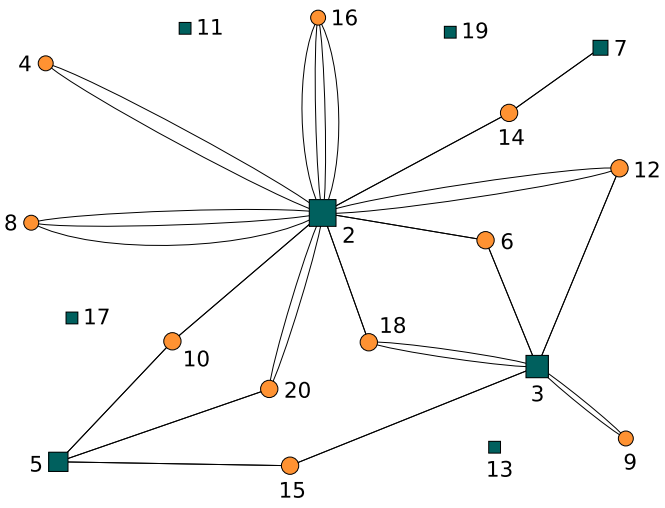

FIG. 1. (Color online) Example of the bipartite network of natural numbers grown up to size 20. Orange circles represent composite numbers and green squares prime numbers. The degree of a prime, k_p, is the number of distinct composites to which it is connected, whereas its strength, s_p, is the sum of its weighted connections. Similarly, the degree of a composite, k_c, is its number of distinct prime factors and its strength, s_c, the total number of prime factors.

The result in Eq. (4) allows us to derive a simple but yet interesting identity relating $\pi(n)$ and the number of distinct prime factors of any integer n, $\omega(n)$. We name $\omega(n)$ the prime factors counting function. We start from the trivial identity $[N - 1 - \pi(N)]\langle k_c \rangle = \pi(N)\langle k_p \rangle$, where k_c is the degree of a composite (or its number of distinct prime factors). Plugging Eq. (4) into this identity, we obtain

$$\sum_{n=2}^{N} \omega(n) = \sum_{i=1}^{\lfloor N/2 \rfloor} \pi\left(\frac{N}{i}\right). \tag{5}$$

Replacing the sum by an integral, we can approximate this expression as

$$\sum_{n=2}^{N} \omega(n) \approx N \int_2^N \frac{\pi(x)dx}{x^2} \sim N \ln\ln N + \mathcal{O}(N). \tag{6}$$

The final asymptotic behavior is directly related to the Hardy-Ramanujan theorem [22], which now becomes a simple consequence of the prime number theorem. Function $\omega(n)$ can be easily computed from Eq. (5) as

$$\omega(n) = \sum_{i=1}^{\lfloor n/2 \rfloor} \left[\pi\left(\frac{n}{i}\right) - \pi\left(\frac{n-1}{i}\right)\right]. \tag{7}$$

Notice that if n is a composite number, then $\omega(n)$ is, in our network representation, its degree. Therefore, the degree distribution of composite numbers is given by $P(k_c) = [\sum_{n=2}^{N} \delta_{\omega(n),k_c} - \delta_{k_c,1}\pi(N)]/[N - 1 - \pi(N)]$. Besides, Eq. (7) naturally leads to a set of arithmetic functions giving the sum of the prime factors of n raised to any exponent (see Appendix C).

Equations (4) and (7) are a remarkable result. Beyond potential applications to find better estimates of function $\omega(n)$, they state that the local properties of the network of natural numbers are fully determined by the prime counting function $\pi(N)$ alone. We then expect that any model producing random versions of the network that is able to reproduce well the prime counting function, $\pi(N)$, will also reproduce well the large scale of the real network topology.

III. MODELING THE EVOLUTION AND STRUCTURE OF NATURAL NUMBERS

The order relation implicit in the natural numbers allows us to consider the bipartite network representation of natural numbers as a growing system. In the growing process, natural numbers join the network sequentially and try to connect to already existing primes. Those new numbers that succeed in this process are said to be composites, otherwise, they become prime numbers. In this paper, we show that a very simple connection rule based upon a soft version of Eq. (3) generates networks with the same architecture as that of the real network of natural numbers. Taking advantage of the apparent randomness of prime numbers, we develop a stochastic model that generates growing bipartite natural number networks connecting random primes with composites.

The growth process only imposes two basic facts trivially implied by the fundamental theorem of arithmetic, that is, that the product of the prime factors of a natural number n must be n, and that n can have no more than one prime factor larger than $\sqrt{n}$. The model starts by assuming that number 2 is a prime and adds natural numbers $n \geqslant 3$ sequentially. It proceeds as follows:

(1) Each new number n that joins the network tries to connect to already existing random primes $p_i \leqslant \sqrt{n}$ with independent probabilities $1/p_i$ one by one, starting from the smallest prime, until the first connection is established.

(2) If number n first connects to an existing prime p smaller or equal to $\sqrt{n}$, it keeps trying to connect sequentially to existing primes in the range $[R_m, R_M]$, with $R_m = p$ and $R_M = \sqrt{n'}$, and $n' = \frac{n}{p}$. Each time n connects to a new random prime p' the range is redefined with $R_{m,\text{new}} = p'$ and $n'_{\text{new}} = \frac{n'_{\text{old}}}{p'}$. If $p' > R_{M,\text{new}}$ or n does not get new connections in the evaluation range, n is connected to the prime closest to R_M^2 and a new node $n+1$ is added to the system.

(3) If number n does not connect to any existing prime smaller or equal to $\sqrt{n}$, it is declared as a prime and a new number $n+1$ is added to the system.

The intuition behind the second step in our model is as follows. In the case of the real primes, a composite number n must have at least a prime factor smaller or equal to $\sqrt{n}$. Let p be the smallest prime factor of n. Then, n/p is also an integer number that is either a prime or, else, it can be expressed as a product of prime factors. However, in the latter case the smallest prime factor of n/p cannot be smaller than p because this would contradict the assumption that p is the smallest prime factor of n. Then, the smallest prime factor of n/p, let it be p', must lie in the closed interval $[p, \sqrt{n/p}]$. The same logic can now be applied to the prime factors of the ratio $n/(pp')$ until n is fully factorized. Our model tries to mimic in a stochastic manner this factorization property of composite numbers, with the difference that, in our case, n/p may not be an integer. Thus, at the end of a stochastic realization of our model, every number n is either declared as a prime or it is a composite such that the product of its prime factors is approximately n.

It is worth noticing the following properties of the model. (i) The model has no tunable parameters. (ii) It is a generative model, in the sense that the model generates simultaneously the number of primes and how primes and composites are connected. (iii) The model is able to generate multiple connections between composite and a prime numbers with no extra mechanism. (iv) The model is non-Markovian because the probability of a number being prime depends on the whole history of the stochastic process. At this respect, it is important to notice that all results in this paper are considered to be averages over all histories of the stochastic process. We also notice that the first step of the algorithm is similar to the random sieve proposed by Hawkins [23–26], the main difference being that the random sieve does not provide connections between composite and prime numbers.

A. The prime counting function

The analytical treatment of the model is quite involved due to its non-Markovian character (see Appendix D). However, it is possible to work out a relatively simple mean-field approximation. For instance, the probability that number n is a prime according to the model, P_n, satisfies the following recurrence relation:

$$P_n = e^{\sum_{i=2}^{\lfloor\sqrt{n}\rfloor} \ln[1-\frac{P_i}{i}]} \approx e^{-\int^{\sqrt{n}} \frac{P_x}{x} dx}, \tag{8}$$

where in the last term we have considered n as a continuous variable and approximated $\ln[1-\frac{P_i}{i}]$ by $-\frac{P_i}{i}$. It is easy to see that Eq. (8) is equivalent to the following nonlinear and nonlocal differential equation

$$\frac{dP_n}{dn} = -\frac{P_n P_{\sqrt{n}}}{2n}. \tag{9}$$

Although the full analytical solution of this equation is difficult to obtain, it is quite easy to check that, asymptotically, P_n behaves as $P_n \sim 1/\ln n$ and, thus, our model satisfies the prime number theorem as an emerging property. Figure 2 shows a comparison between the real $\pi(N)$, the one generated by our model $\Pi(N)$, and Eq. (2). As expected, $\lim_{N\to\infty} \pi(N)/\Pi(N) = 1$. However, for finite sizes the relative error of our model with respect to the real $\pi(N)$ is one order of magnitude smaller than the one given by Eq. (2).

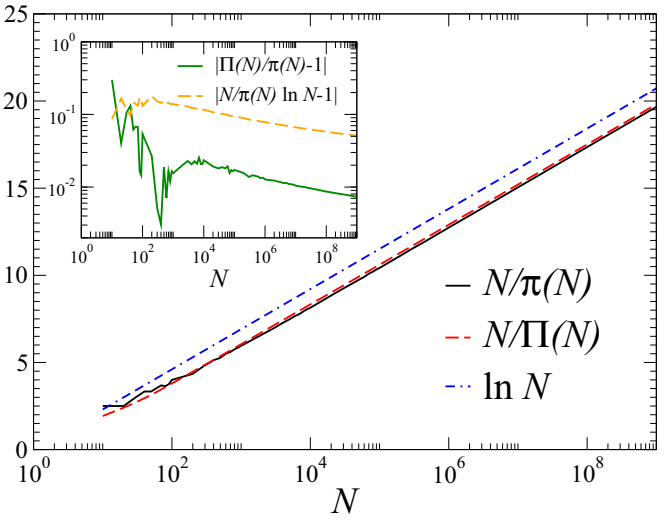

FIG. 2. (Color online) Comparison of the prime counting function $\pi(N)$, the prime number theorem Eq. (2), and the prime counting function of our random model $\Pi(N)$, averaged over 1000 realizations. The inset shows the corresponding relative errors. The relative error of the random model is one order of magnitude smaller than the one of Eq. (2).

B. Network properties

One of the strengths of our model lays in its ability to reproduce, not only the sequence of primes, but also the connections of each composite number. To check to what extent our model fulfills the fundamental theorem of arithmetic, we measure the relative error between a composite and its factorization according to the model, $\epsilon(N)$, as follows. Let c_i be the ith composite in a network of size N and let $\bar{c}_i$ be its stochastic factorization, then we define $x_i \equiv \bar{c}_i/c_i$. The relative error is then $\epsilon(N) \equiv 1 - \langle x \rangle = 1 - [N - 1 - \Pi(N)]^{-1} \sum_i \bar{c}_i/c_i$, where $\langle\cdot\rangle$ means the population average. In Fig. 3, we show

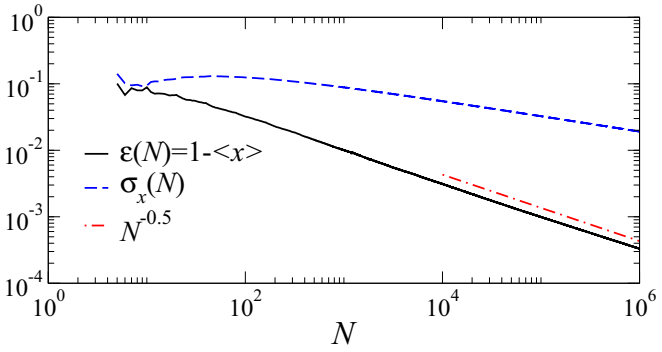

FIG. 3. (Color online) Average relative error $\epsilon(N) = 1 - \langle x \rangle$ between a composite number and its factorization in the network as a function of the system size N and the standard deviation of the ratio x, $\sigma_x(N)$.

$\epsilon(N)$ as a function of the system size N averaged over 1000 network realizations. As it can be seen, this error decreases as a power law of the size of the system $\epsilon(N) \sim N^{-\alpha}$ with $\alpha \approx 0.5$. We also show the standard deviation of x_i, which also approaches zero in the large system size limit. These two results indicate that the model fulfills the fundamental theorem of arithmetic for relatively small numbers with high accuracy.

The model also does an excellent job at reproducing well the large-scale topology of the real network. The left column in Fig. 4 shows the complementary cumulative degree distributions of primes and composites as compared to the real ones for the network grown up to $N = 10^6$. In both cases the agreement is excellent. The right column in Fig. 4 shows the strength distributions for primes and composites, that is, the equivalent to the left column measures when multiple links between primes and composites are considered (see Fig. 1). Again, the agreement between the model and the real network is excellent. This result is particularly interesting as it shows that our model

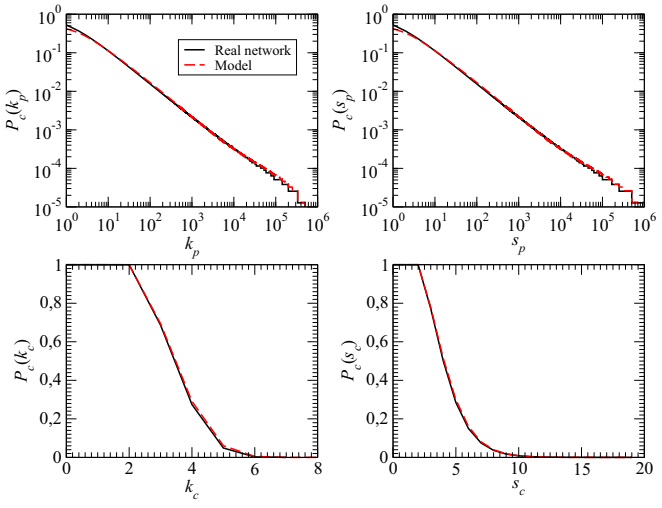

FIG. 4. (Color online) Comparison between the complementary cumulative distribution functions of the real bipartite network of natural numbers of size $N = 10^6$ and the network generated by our model averaged over 1000 realizations. The left column shows the unweighted properties and the right column the weighted ones. The legend explaining line types applies to the four plots.

is able to capture statistical properties of the multiplicities of composites' factorizations, i.e., the αs in Eq. (3). In particular, it recovers that $P_c(s_p)$ behaves asymptotically as s_p^{-2}, as expected from the almost linear correlation between strength and degree. Other topological properties are explored in Appendices B and E. For instance, it is possible to show that the model satisfies the Erdös-Kac theorem [27], which states that $[\omega(n) - \ln\ln n]/\sqrt{\ln\ln n}$ is, *de facto*, a random variable that follows the standard normal distribution.

C. The Cramér's conjecture revisited

Cramér's conjecture provides an absolute upper bound on the gaps between consecutive primes. Using his model, Cramér was able to prove that [19]

$$\limsup_{i\to\infty} \frac{p_{i+1} - p_i}{\ln^2 p_i} = 1 \tag{10}$$

and conjectured that the same relation also holds for real primes. Here, we study the statistics of prime gaps in our model and refine Cramér's conjecture for real primes. We start by noticing that in our model, all numbers between two perfect squares have the same probability of being primes and, more importantly, they are conditionally independent given their common history. Therefore, as a first approximation, we consider that every number in the interval $[m^2,(m+1)^2)$; $m = 2,3,\cdots$ has an independent probability $P_n = 1/\ln n$ of being a prime, where $n = m^2$. Under this assumption, the probability that a given gap G within the interval is smaller than g is $\text{Prob}\{G < g\} = 1 - (1 - P_n)^{g-1}$ [28]. If we assume that there are $N_G = 2\sqrt{n}P_n$ gaps within the interval, the probability that the largest gap G_m within the interval is smaller than g_m is

$$\text{Prob}\{G_m < g_m\} = [1 - (1 - P_n)^{g_m - 1}]^{N_G}. \tag{11}$$

The average largest gap can be evaluated from this expression, yielding

$$\langle G_m \rangle = \left(\frac{1}{P_n} - \frac{1}{2}\right) H_{N_G} + \mathcal{O}(P_n) \sim \frac{1}{2}\ln^2 n, \tag{12}$$

where $H_{N_G} = \sum_{k=1}^{N_G} k^{-1}$ is the N_Gth harmonic number (interestingly, a similar approach has been recently proposed in Ref. [29]). We can now define the normalized largest gap as $\overline{G}_m \equiv G_m/\langle G_m \rangle$, which distribution function satisfies

$$\text{Prob}\{\overline{G}_m < \overline{g}_m\} \sim e^{-N_G^{1-\overline{g}_m}}. \tag{13}$$

In the limit $n \to \infty$, $N_G \to \infty$ and this distribution becomes a step function (although very slowly). Thus, the largest gap stops being a random variable to become a deterministic quantity equal to $\ln^2 n/2$. Notice that this bound is twice as small as the bound given by Cramér's conjecture, apparently suggesting that it could be false for real primes.

To check our prediction, we compute the gaps between real primes up to 10^{11}. We divide this set in intervals between perfect squares and for each such interval we evaluate the largest gap. The top plot in Fig. 5 shows the series of largest gaps and the inset shows the normalized largest gaps obtained by using Eq. (12). As it can be seen, after normalization, the series becomes a stationary one but its average is not 1, as we would expect from our model, but $2c \approx 0.88$, with c a constant below $1/2$. As we see, our model suffers from

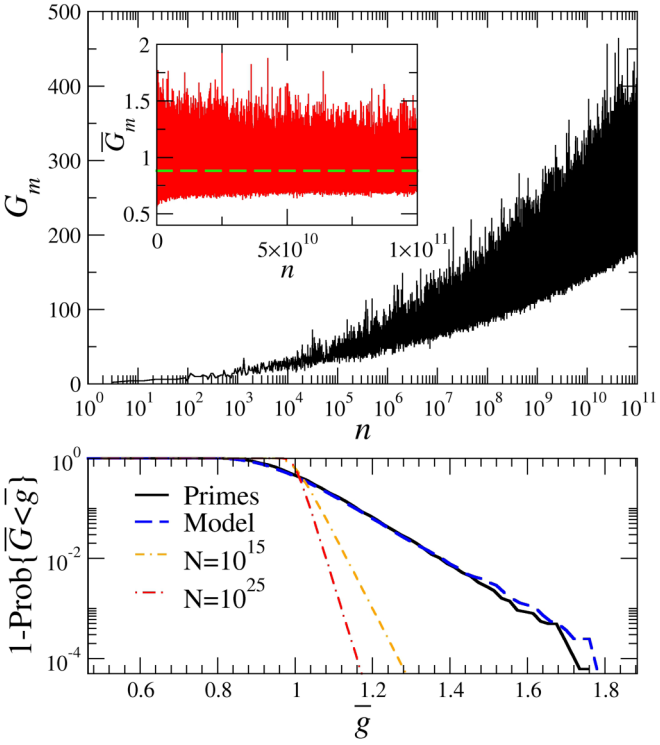

FIG. 5. (Color online) Gaps between primes. Top. Series of largest gaps between real primes in intervals between perfect squares. Top inset: The same series normalized by using Eq. (12). In both plots, primes are considered up to 10^{11}. Bottom: Complementary cumulative distribution function of the normalized largest gaps for real primes and the model in the range $[9\times10^{10},10^{11}]$. To make evident the slow convergence of the distribution, we also show extrapolations from Eq. (13) for $N=10^{15}$ and $N=10^{25}$.

the same problems affecting Cramér's model in what respects short-range correlations induced by small primes. For instance, the probability of n being a prime if $n-1$ is a prime is zero for real primes (except for 2 and 3), whereas our model would predict a nonzero probability; in addition, the probabilistic prediction that the number of primes in a short interval of length y about x is given by $y/\ln x$ was proved false by Maier [30,31]. Some other deviations from real primes on a very large scale have also been reported [32,33]. In the case of Cramér's model, it is possible to make heuristic corrections allowing one to reach right answers on several properties of real primes, like the number of twin primes below N [34]. In general, these corrections have only a numerical effect on the studied property since the bear model already predicts the right asymptotic behavior as a function of N. The same type of heuristics can be, in principle, applied to our model and we expect them to account for the observed discrepancy. For instance, a simple modification assumes that the probability of n being a prime is zero if the previous number is a prime, whereas it is $(\ln n-1)^{-1}$ otherwise. This simple modification preserves the prime number theorem and leads to a better estimate of constant $2c\approx0.92$.

Even more interesting is the analysis of the fluctuations of the normalized largest gaps around their average. A preliminary analysis of their distribution suggests that largest gaps of real primes behave as in the model after a global rescaling. Thus, to have a coherent comparison between the model and real primes, we divide the series shown in the inset of Fig. 5 by $2c$ so that its average is equal to 1, like in the model. We then evaluate the complementary cumulative distribution function for all largest gaps in the range $[9\times10^{10},10^{11}]$ and compare it with the one obtained from numerical simulations of our model; see bottom plot in Fig. 5. Interestingly, both distributions are nearly indistinguishable. This implies that fluctuations of largest gaps for real primes are governed asymptotically by the distribution Eq. (13). From this equation, we can evaluate the expected number of gaps up to N that are above a certain fraction α of the average largest gap, with $\alpha\geqslant1$, that is,

$$\#\text{ gaps with }\overline{G}_m>\alpha\approx\sum_{n=1}^{\sqrt{N}}\left(\frac{\ln n}{n}\right)^{\alpha-1}. \tag{14}$$

This quantity diverges when $1\leqslant\alpha<2$ as $\mathcal{O}(N^{1-\alpha/2}\ln^{\alpha-1}N)$ and as $\mathcal{O}(\ln^2 N)$ for $\alpha=2$. Putting all the pieces together, we refine Cramér's conjecture as follows. For all real prime gaps $G_i\equiv p_{i+1}-p_i$, with $p_i<N$ and $N\to\infty$, we have

$$\begin{aligned} G_i &< \alpha c\ln^2 p_i, \quad \text{for all but } \mathcal{O}(N^{1-\frac{\alpha}{2}}\ln^{\alpha-1}N) \text{ gaps},\\ G_i &< 2c\ln^2 p_i, \quad \text{for all but } \mathcal{O}(\ln^2 N) \text{ gaps}. \end{aligned} \tag{15}$$

For any $\alpha>2$, the number of gaps above this threshold is $\mathcal{O}(1)$. Notice, however, that this asymptotic behavior is only reached for extremely large values of N. For not so large values it is better to replace $\ln^2 p_i$ in Eq. (15) by $2[\ln p_i-1/2][\ln(2\sqrt{p_i}/\ln p_i)+\gamma]$, with γ the Euler-Mascheroni constant, as derived from Eq. (12). We check these predictions for all gaps up to 10^{11} in Fig. 6. We measure empirically the number of gaps that, up to a given size N, satisfy $G_i>2\alpha c[\ln p_i-1/2][\ln(2\sqrt{p_i}/\ln p_i)+\gamma]$ and compare them with the predictions in Eq. (15). Aside from statistical errors, our predictions agree well with the empirical measures.

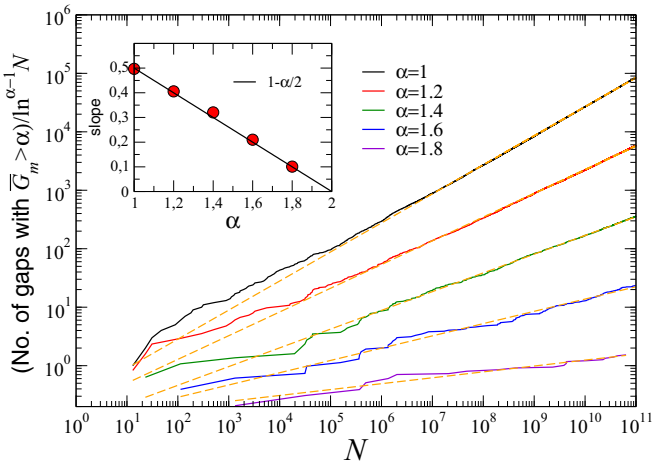

FIG. 6. (Color online) Number of gaps with $\overline{G}_m>\alpha$ for different values of α as a function of N rescaled by the factor $\ln^{\alpha-1}N$. According our estimates, this should behave as a power law of the form $N^{1-\alpha/2}$. Dashed lines are power law fits, which exponents are shown in the inset plot and compared to the theoretical prediction $1-\alpha/2$.

IV. CONCLUSIONS

Probabilistic approaches to understand usual patterns of primes as well as their extreme statistics brought a new perspective to the study of prime numbers. The big first step by Cramér was significantly developed afterwards bringing this kind of approach to maturity. With our work, we introduce a new dimension that allows us to understand primes and their statistical properties not in isolation but as building blocks of natural numbers. We have introduced a parameter-free non-Markovian stochastic model based on a bipartite complex network representation that naturally generates random primes as well as the relation between primes and composite numbers with remarkable accuracy. Our model satisfies the Erdös-Kac theorem, as well as the prime number theorem and a refined version of Cramér's conjecture about the statistics of gaps between consecutive primes that seems closer to reality than the original Cramér's version. Even though we are still unable to fully understand the finer details about primes and the complex architecture of natural numbers, probabilistic models like ours provide valuable tools helping to elaborate conjectures about primes and, perhaps, also to prove results on number theory. Beyond the implications in mathematics, our stochastic model generates the sequence of random primes and some of their statistical correlations as an emergent property, which allows probabilistic computations of number theoretical approaches to open problems in physics involving the Riemann ζ function, which plays an integral role in different branches from quantum mechanics to condensed matter.

ACKNOWLEDGMENTS

We thank Dmitri Krioukov for useful comments and suggestions. We acknowledge support from the James S. McDonnell Foundation 21st Century Science Initiative in Studying Complex Systems Scholar Award; the ICREA Academia prize, funded by the *Generalitat de Catalunya*; MICINN Projects No. FIS2010-21781-C02-02 and No. BFU2010-21847-C02-02; *Generalitat de Catalunya* Grant No. 2014SGR608; and the Ramón y Cajal program of the Spanish Ministry of Science.

APPENDIX A: BIPARTITE NETWORK REPRESENTATION OF NATURAL NUMBERS

In this section we derive the expressions that characterize the bipartite network representation of natural numbers presented in the paper.

1. Degree distribution

The degree distribution for primes in the network can be derived reasoning as follows: a prime number $p > N/2$ has degree $k_p(p) = 0$ since its product by any other prime number is greater than N and, hence, it cannot belong to the network (the subscript in k_p is used to denote the degree of primes; we use k_c to refer to the degree of composites). Identically, if $N/3 < p \leqslant N/2$, p has a multiple that belongs to the network ($2p \leqslant N$). In general,

$$p \in \left(\frac{N}{n+1}, \frac{N}{n}\right] \Leftrightarrow k_p(p) = n - 1, \tag{A1}$$

since $mp \leqslant N, m = 2, \ldots, n$ but $(n+1)p > N$. This directly leads to the expression for $P(k_p)$:

$$\begin{aligned} P(k_p) &= \frac{\#\{p : p \text{ prime} : k_p(p) = k_p\}}{\#\{p : p \text{ prime} \leqslant N\}} \\ &= \frac{\#\left\{p : p \text{ prime} \in \left(\frac{N}{k_p+2}, \frac{N}{k_p+1}\right]\right\}}{\#\{p : p \text{ prime} \leqslant N\}} \\ &= \frac{\pi\left(\frac{N}{k_p+1}\right) - \pi\left(\frac{N}{k_p+2}\right)}{\pi(N)}. \end{aligned} \tag{A2}$$

This expression is, interestingly, similar to a probability measure with multifractal properties used in Ref. [35]. We can derive an approximation for Eq. (A2) using the fact that, according to the prime number theorem,

$$\lim_{x\to\infty} \frac{\pi(x)}{x/\ln(x)} = 1. \tag{A3}$$

We first evaluate the complementary cumulative distribution function $P_c(k_p) = \sum_{k \geqslant k_p} P(k)$, which reads

$$P_c(k_p) = \frac{\pi\left(\frac{N}{k_p+1}\right)}{\pi(N)}. \tag{A4}$$

Using the prime number theorem, in the limit $N/k_p \gg 1$ this function behaves as

$$P_c(k_p) \approx \frac{1}{k_p\left(1 - \frac{\ln k_p}{\ln N}\right)} \sim \frac{1}{k_p}, \tag{A5}$$

from where it follows that the degree distribution behaves nearly as a power law:

$$P(k_p) \sim k_p^{-2}. \tag{A6}$$

Another useful relation is

$$k_p(p) = \left\lfloor \frac{N}{p} \right\rfloor - 1, \tag{A7}$$

which can be proved considering Eq. (A1):

$$\begin{aligned} p \in \left(\frac{N}{n+1}, \frac{N}{n}\right] \Leftrightarrow \frac{N}{p} \in [n, n+1) \Leftrightarrow \left\lfloor \frac{N}{p} \right\rfloor = n \Leftrightarrow k_p(p) \\ = \left\lfloor \frac{N}{p} \right\rfloor - 1. \end{aligned}$$

2. Strength of a prime number

The expression for the strength of a prime number p in the network of size N is

$$s_p(p) = \sum_{n=1}^{\lfloor \log_p N \rfloor} \left\lfloor \frac{N}{p^n} \right\rfloor - 1. \tag{A8}$$

The explanation of this formula is rather straightforward. The prime p inside the bipartite network is connected to $\lfloor N/p \rfloor - 1$ composites [Eq. (A7)]. Nevertheless, $\lfloor N/p^2 \rfloor$ of these composites can be divided by p twice. In general, there are $\lfloor N/p^n \rfloor$ composites which can be divided by p n times. Since the strength of the prime p is defined as the sum of the weights of all its connections, we can simply sum all these

terms as

$$s_p(p) = k_p(p) + \sum_{n=2}^{\infty} \left\lfloor \frac{N}{p^n} \right\rfloor = \sum_{n=1}^{\infty} \left\lfloor \frac{N}{p^n} \right\rfloor - 1.$$

An upper limit for the sum can be found by taking into account the fact that, if $p^n > N \Rightarrow N/p^n < 1$ and, hence, such term does not contribute to the sum. Let us then find the values of n that need to be considered:

$$\left\lfloor \frac{N}{p^n} \right\rfloor > 0 \Leftrightarrow \frac{N}{p^n} \geqslant 1 \Leftrightarrow p^n \leqslant N \Leftrightarrow n \leqslant \log_p N.$$

This allows us to write the upper limit in Eq. (A8), since the last term to be added is the one for $n = \lfloor \log_p N \rfloor$.

3. Strength distribution

A reasonable approximation of the strength as a function of the degree k_p is given by

$$s_p(k_p) \sim \frac{N(k_p+1)}{N-(k_p+1)} - 1, \tag{A9}$$

which shows that weights do not play an important role in our representation since, for small values of k_p, Eq. (A9) exhibits a linear behavior $[s_p(k_p) \sim k_p]$. This result is a consequence of the fact that only primes less or equal to $\sqrt{N}$ have connections with weight greater than 1, which implies that the fraction of nodes for which this is possible, $1/\sqrt{N}$, tends to zero in the thermodynamic limit. Equation (A9) can be derived by approximating Eq. (A8) as

$$s_p(p) = \sum_{n=1}^{\lfloor \log_p N \rfloor} \left\lfloor \frac{N}{p^n} \right\rfloor - 1 \sim \sum_{n=1}^{\infty} \frac{N}{p^n} - 1 = \frac{N}{p-1} - 1. \tag{A10}$$

We can finally use Eq. (A7) to give an approximate value of $p(k_p)$, i.e., a prime with degree k_p,

$$k_p(p) = \left\lfloor \frac{N}{p} \right\rfloor - 1 \Rightarrow p \sim \frac{N}{k_p+1}. \tag{A11}$$

The substitution of Eq. (A11) into Eq. (A10) yields Eq. (A9).

The cumulative strength distribution can also be derived as follows. From Eq. (A10), we see that any prime p such that

$$p \gtrsim \frac{N}{s_p+1} + 1$$

must have strength less or equal to s_p. We can therefore approximate $P_c(s_p) = \text{Prob}\{S > s_p\} = 1 - \text{Prob}\{S \leqslant s_p\}$, where S stands for the strength of a randomly chosen prime, as

$$\begin{aligned} P_c(s_p) &\sim 1 - \frac{\pi(N) - \pi\left(\frac{N}{s_p+1}+1\right)}{\pi(N)} = \frac{\pi\left(\frac{N}{s_p+1}+1\right)}{\pi(N)} \\ &\sim \frac{\pi\left(\frac{N}{s_p+1}\right)}{\pi(N)} \sim \frac{N}{(s_p+1)\ln\left(\frac{N}{s_p+1}\right)} \frac{\ln N}{N} \\ &= \frac{1}{1-\frac{\ln(s_p+1)}{\ln N}} \frac{1}{s_p+1} \sim s_p^{-1}, \end{aligned}$$

so we see that, indeed, $P(s_p) \sim s_p^{-2}$.

APPENDIX B: ONE-MODE PROJECTION

Given a bipartite network, we can build a new graph composed exclusively of nodes belonging to one of its classes by performing the so called one-mode projection. Since no pair of these nodes can be initially connected by the definition of bipartite network, linking must be ruled by some other

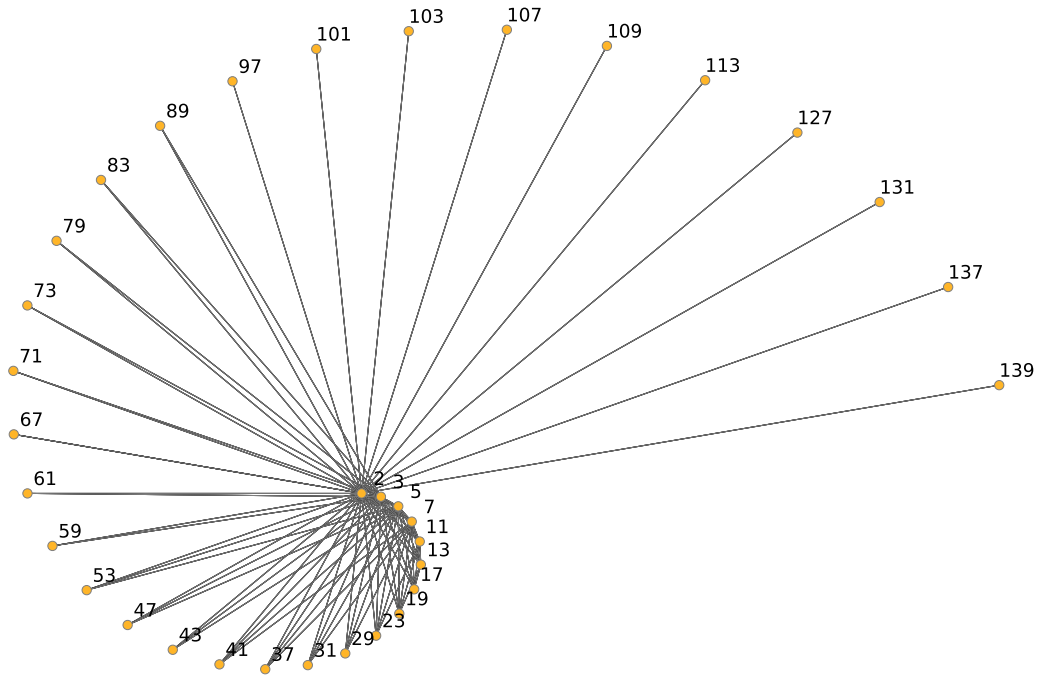

FIG. 7. (Color online) One-mode projection for $N = 289$. Primes greater than $N/2$ do not appear in the picture since they are all unconnected. Self-loops are not depicted either. Primes {2,3,5,7,11,13,17} form a clique, and there are no links between nodes not belonging to it.

criteria in the new graph. The most usual one is to establish a connection between two nodes with a weight equal to the number of common nodes to which they were both connected in the original network. Hence, whenever two nodes had no common neighbors in the bipartite network, they are left unconnected.

In order to go deeper into the study of the statistical properties of prime numbers, we have performed a one-mode projection onto that class in the bipartite network discussed so far following the latter criteria (see Fig. 7) and, in addition, allowing self-loops to exist in the resulting graph (whenever a perfect power of a prime exists in the bipartite network, we regard that prime as connected to itself, thus forming a self-loop).

As can be seen in Fig. 7, and as the results presented in this section imply, this graph has a structure made of a maximally connected core containing all the primes less or equal to $\sqrt{N}$ that is surrounded by nodes connected to some but not all of the inner nodes. In addition, the inner two prime numbers are, the strongest the connection among them. This suggests that this network could exhibit a self-similar behavior; i.e., it could be statistically invariant under a network renormalization procedure. This interesting property would allow us to predict some of its statistical properties on any scale.

1. Degree distribution

The degree k of a prime number p in the one-mode projection of a bipartite network of size N is given by

$$k(p) = \pi\left(\frac{N}{p}\right). \tag{B1}$$

This expression is justified as follows: p can be connected to any prime number p' as long as $pp' \leqslant N$. As a consequence, in order to obtain the number of primes p' to which p can be connected, we must count the number of primes $p' \leqslant N/p$, which is precisely the result in Eq. (B1). Notice that, if $p \leqslant \sqrt{N}$, p is counted as well; hence, this expression takes self-loops into account.

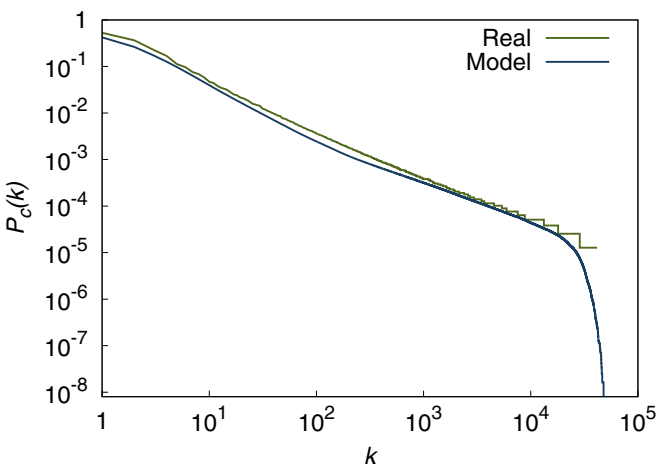

FIG. 8. (Color online) Complementary cumulative degree distribution $P_c(k)$ of the one-mode projection graph for both the real and the stochastic model networks.

Using p_k to denote the kth prime, the degree distribution $P(k)$ is exactly determined by

$$P(k) = \frac{\pi\left(\frac{N}{p_k}\right) - \pi\left(\frac{N}{p_{k+1}}\right)}{\pi(N)}, \quad p_0 \equiv 1. \tag{B2}$$

This result starts with the observation that if a prime p has degree k, it must be connected to the first k prime numbers $p_1, p_2, \ldots, p_k$. Hence, $pp_k \leqslant N$ but $pp_{k+1} > N$. In order to count how many primes are subject to these conditions, we must count the number of primes in the interval $p \in (N/p_{k+1}, N/p_k]$, which can be written in terms of the prime counting function as $\pi(N/p_k) - \pi(N/p_{k+1})$. Dividing that quantity by the amount of primes in the graph $\pi(N)$ yields Eq. (B2). We must take into account that, in the particular case of $k = 0$, we are considering the primes p for which $pp_1 > N$ and $p \leqslant N$, i.e. the primes $p \in (N/p_1, N]$. Defining $p_0 \equiv 1$, the latter equation is extended to that case.

In Fig. 8 we compare Eq. (B2) with its stochastic homologous.

2. Weight of a connection and strength of a prime

The weight of the connection ω_{ij} between two primes p_i and p_j is

$$\omega_{ij} = \left\lfloor \frac{N}{p_i p_j} \right\rfloor. \tag{B3}$$

This quantity is defined as the number of composites in the bipartite network to which both primes are connected. Such composites must be divisible by both p_i and p_j, i.e., by $p_i p_j$. Since there are $\lfloor N/p_i p_j \rfloor$ such numbers among the first N natural numbers, Eq. (B3) effectively gives ω_{ij}.

The strength of a prime number is straightforward to obtain from the latter result. By its definition, the only thing to do is adding the weights of the connections to all the other prime numbers in the network, from p_1 to $p_{\pi(\frac{N}{p})}$ (notice that if $p_i > N/p$ the weight of the connection is equal to zero). This leads to

$$s(p) = \sum_{i=1}^{\pi\left(\frac{N}{p}\right)} \left\lfloor \frac{N}{pp_i} \right\rfloor. \tag{B4}$$

Equation (B4) also adds the weight of the self-loop of p if existing (if $p^2 \leqslant N$).

3. Clustering coefficient

We can derive an expression for the clustering coefficient $C(p)$ of a prime number inside this graph. This quantity is a real number $C(p) \in [0,1]$ representing the fraction of possible links between the neighbors of p that actually exist. This coefficient affects many processes in networks such as percolation, dynamic processes, etc., and it is closely related to the small-world property as well as to hidden geometries. In our case, if $p \geqslant \sqrt{N}$, it can only be connected to primes $p_i \leqslant \sqrt{N}$. As the product of two numbers below $\sqrt{N}$ cannot be greater than N, all the primes $p_i \leqslant \sqrt{N}$ are connected to each other. Consequently, the clustering coefficient is

$C(p) = 1$ for any $p \geqslant \sqrt{N}$. However, when $p \leqslant \sqrt{N}$, the expression for $C(p)$ is given by

$$C(p) = \frac{[\pi(p)-1]\{2[\pi(\frac{N}{p})-1]-\pi(p)\}+2\{\sum_{j=\pi(p)+1}^{\pi(\sqrt{N})}[\pi(\frac{N}{p_j})-j]+\pi(\sqrt{N})-1\}}{\pi(\frac{N}{p})[\pi(\frac{N}{p})-1]}. \quad \text{(B5)}$$

To derive Eq. (B5) we need to count the number of connections between the primes to which p is connected. Let us compute several quantities separately.

(1) The number of neighbors of the prime p that we need to consider is not $k(p)$ as given by Eq. (B1), but $k \equiv k(p) - 1 = \pi(\frac{N}{p}) - 1$; since $p \leqslant \sqrt{N}$, we must correct the fact that Eq. (B1) is counting the self-loop of prime p. The number of possible links among these nodes is, allowing the possibility for self-loops to exist,

$$L_{\max} = \frac{1}{2}k(k+1) = \frac{1}{2}\pi\left(\frac{N}{p}\right)\left[\pi\left(\frac{N}{p}\right)-1\right]. \quad \text{(B6)}$$

(2) The number of self-loops existing among the neighbors of p, $L_{\rm sl}$, can be derived easily; a self-loop exists if and only if the corresponding prime is less or equal to $\sqrt{N}$. In addition, p is connected to all such primes, so

$$L_{\rm sl} = \pi(\sqrt{N}) - 1. \quad \text{(B7)}$$

The minus one term corrects the overcount due to the self-loop of prime p. This result allows us to simply count the number of links among the neighbors of p regardless of self-loops. This calculation is conveniently separated into two more parts.

(3) Links concerning primes less than p: let p_i denote any prime less than p [so $i = 1, \ldots, \pi(p) - 1$]. Then, if for some prime p' it is true that $pp' \leqslant N$, it must be true that $p_i p' < N$. In other words, all the p_i are connected to all the primes to which p is connected. Therefore, we need to count the number of different connections that $\pi(p) - 1$ elements can form with k elements (regardless of self loops, as explained above). We can proceed in the following manner: the first of the p_i, p_1, is connected to $k - 1$ elements. The second prime, p_2, forms $k - 2$ new bonds, since the connection to p_1 is not counted again. The elements in this succession can be written as $k - j$, which allows us to write the corresponding series as

$$L_{p_i<p} = \sum_{j=1}^{\pi(p)-1}(k-j) = [\pi(p)-1]k - \sum_{j=1}^{\pi(p)-1} j$$

$$= [\pi(p)-1]k - \frac{[\pi(p)-1]\pi(p)}{2}.$$

Making now use of the expression for k derived previously yields

$$L_{p_i<p} = \frac{1}{2}[\pi(p)-1]\left\{2\left[\pi\left(\frac{N}{p}\right)-1\right]-\pi(p)\right\}. \quad \text{(B8)}$$

(4) Links not concerning primes less than p: consider any pair of primes p_i and p_j such that $p_j > p_i > p$. Then, if $p_i p_j \leqslant N$, the chained inequalities $pp_i < pp_j < N$ must hold as well. This means that any link between p_i and p_j (both greater than p) is a link among neighbors of p; in particular, those that we have not counted yet. An easy way to count such links is to count, for every p_i, the number of p_j such that $p_i p_j \leqslant N$. For any given p_i we see that the value for p_j is bounded by $p_i < p_j \leqslant N/p_i$, so there are $\pi(N/p_i) - \pi(p_i) = \pi(N/p_i) - i$ links to be counted for prime p_i. The only thing left to do is adding the terms for all the p_i. Note, however, that the upper bound for i is given by $i \leqslant \pi(\sqrt{N})$ (if both p_i and p_j are greater than $\sqrt{N}$, their product cannot belong to the bipartite network). Finally, we can write

$$L_{p_i>p} = \sum_{j=\pi(p)+1}^{\pi(\sqrt{N})}\left[\pi\left(\frac{N}{p_j}\right)-j\right]. \quad \text{(B9)}$$

Equation (B5) is obtained directly by adding Eqs. (B7)–(B9) and dividing the result by Eq. (B6).

$$C(p) = \frac{L_{\rm sl} + L_{p_i<p} + L_{p_i>p}}{L_{\max}}. \quad \text{(B10)}$$

We have obtained a numerical relation between the clustering coefficient C and the degree k as well, which is plotted in Fig. 9 with the corresponding measurement on the stochastic model.

APPENDIX C: ARITHMETIC FUNCTIONS

The perspective of number theory that we have presented in this work provides us with a new approach to some arithmetic functions as well. In this section, we present a few results derived from our network representation of natural numbers concerning several of them. We have been able to derive exact and approximated expressions for the prime factors counting function $\omega(n)$ (indeed, we have informally obtained its normal order in accordance with the Hardy-Ramanujan theorem), the sum of the prime divisors of a number n raised to the rth power [which we denote by $\tau_r(n)$] and, indirectly, the sum of divisors of n to the rth power, $\sigma_r(n)$.

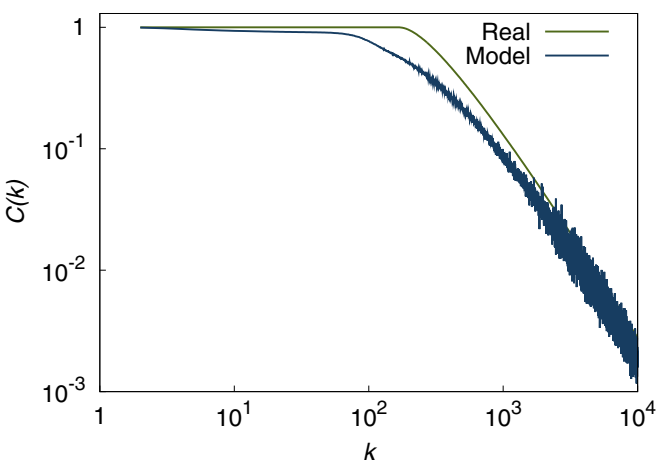

FIG. 9. (Color online) Clustering coefficient as a function of the degree $C(k)$.

1. Prime factors counting function $\omega(n)$

In the bipartite network that we have studied, every link connects a prime and a composite. Therefore, counting all the distinct links in the graph (i.e., with no multiplicities) yields the sum of the distinct prime divisors of all the composites up to N,

$$\sum_{n\text{composite}\leqslant N}\omega(n)=\pi(N)\sum_{k_p}k_pP(k_p). \tag{C1}$$

Since $\omega(p)=1$ for any prime, we can extend the latter sum to all $n\in[2,N]$ simply as

$$\sum_{n=2}^{N}\omega(n)=\pi(N)\left[1+\sum_{k_p}k_pP(k_p)\right]. \tag{C2}$$

Expanding the sum over k_p gives

$$\begin{aligned}\pi(N)\sum_{k_p}k_pP(k_p)&=\pi\left(\frac{N}{2}\right)-\pi\left(\frac{N}{3}\right)\\&\quad+2\left[\pi\left(\frac{N}{3}\right)-\pi\left(\frac{N}{4}\right)\right]+\cdots\\&=\sum_{k\geqslant 2}\pi\left(\frac{N}{k}\right).\end{aligned} \tag{C3}$$

We can find an upper limit for the sum in the latter expression considering that $\pi(N/k)>0\Leftrightarrow N/k\geqslant 2$, so only the terms with $k\leqslant\lfloor N/2\rfloor$ need to be added. We are finally led to the interesting identity

$$\sum_{n=2}^{N}\omega(n)=\sum_{k=1}^{\lfloor N/2\rfloor}\pi\left(\frac{N}{k}\right). \tag{C4}$$

The arithmetic function $\omega(n)$ is given in terms of Eq. (C4) as the difference between two consecutive sums, i.e., between the sums up to n and $n-1$,

$$\omega(n)=\sum_{k=1}^{\lfloor n/2\rfloor}\left[\pi\left(\frac{n}{k}\right)-\pi\left(\frac{n-1}{k}\right)\right]. \tag{C5}$$

2. Sum of prime factors of n raised to the rth power $\tau_r(n)$

A further analysis of Eq. (C5) reveals that $\phi(k;n)\equiv\pi(\frac{n}{k})-\pi(\frac{n-1}{k})$ gives

$$\phi(k;n)=\begin{cases}1 & \text{if } \frac{n}{k} \text{ is prime}\\ 0 & \text{otherwise}\end{cases} \tag{C6}$$

This result allows us to write an expression for $\tau_r(n)$, which we define as the sum of prime divisors of n raised to the rth power,

$$\tau_r(n)=\sum_{k=1}^{\lfloor n/2\rfloor}\left(\frac{n}{k}\right)^r\phi(k;n), \tag{C7}$$

so $\omega(n)=\tau_0(n)$. However, we need to prove Eq. (C6).

First notice that $\phi(k;n)$ is equal to the number of primes in the interval $p\in(\frac{n-1}{k},\frac{n}{k}]$. Let us thus count the number of integers in the interval. Suppose that $\frac{n}{k}\notin\mathbb{N}$. Then,

$$\frac{n}{k}=\frac{qk+r}{k}>q, \tag{C8}$$

with $q=\lfloor\frac{n}{k}\rfloor\in\mathbb{N}$ and $r\geqslant 1\in\mathbb{N}$. In addition, we have

$$\frac{n-1}{k}=\frac{qk+r-1}{k}\geqslant q. \tag{C9}$$

Even though $\frac{n-1}{k}$ can be an integer (if $r=1$), it does not belong to the interval $(\frac{n-1}{k},\frac{n}{k}]$, so every number in the interval is greater than q. We thus conclude that if $\frac{n}{k}\notin\mathbb{N}$ there are no integers (and therefore no primes) in the interval $[\frac{n}{k}\notin\mathbb{N}\Rightarrow\phi(k;n)=0]$.

On the other hand, if $\frac{n}{k}\in\mathbb{N}$, Eq. (C8) reads

$$\frac{n}{k}=q, \tag{C10}$$

while Eq. (C9) becomes

$$\frac{n-1}{k}=\frac{(q-1)k+k-1}{k}\geqslant q-1. \tag{C11}$$

In this case, we see that every number in the interval $x\in(\frac{n-1}{k},\frac{n}{k}]$ lies between $q-1<x<q$ (and, hence, they cannot be integers) except for $x=\frac{n}{k}\in\mathbb{N}$. We can thus conclude that, if $\frac{n}{k}$ is prime, $\pi(\frac{n}{k})=\pi(\frac{n-1}{k})+1$ and, therefore, $\phi(k;n)=1$. Notice, however, that even though $\frac{n}{k}\in\mathbb{N}$, if it is not prime, $\pi(\frac{n}{k})=\pi(\frac{n-1}{k})\Leftrightarrow\phi(k;n)=0$.

3. Approximation of $\tau_r(n)$

We can derive an approximation of Eq. (C5) exchanging the sum for an integral and making use of the prime number theorem [from Eq. (A3), we see that $\pi(x)\sim x/\ln x$],

$$\begin{aligned}\omega(n)=\tau_0(n)&=\sum_{k=1}^{\lfloor n/2\rfloor}\left[\pi\left(\frac{n}{k}\right)-\pi\left(\frac{n-1}{k}\right)\right]\\&\sim\int_1^{n/2}\left[\frac{n}{k\ln\frac{n}{k}}-\frac{n-1}{k\ln\frac{n-1}{k}}\right]dk\\&\sim\int_1^{n/2}\frac{dk}{k\ln\frac{n}{k}}=\int_2^{n}\frac{dp}{p\ln p}=\ln\ln n-\ln\ln 2.\end{aligned} \tag{C12}$$

The latter expression yields, according to the Hardy-Ramanujan theorem, the normal order of $\omega(n)$.

By the same line of reasoning, we can approximate any of the $\tau_r(n)$ for $r>0$,

$$\begin{aligned}\tau_r(n)&\sim\int_1^{n/2}\left(\frac{n}{k}\right)^r\frac{dk}{k\ln\frac{n}{k}}=\int_2^n\frac{p^{r-1}}{\ln p}dp\\&=\int_{2^r}^{n^r}\frac{dq}{q}=\text{li}(n^r)-\text{li}(2^r).\end{aligned} \tag{C13}$$

Eq. (C13) yields a very interesting result; in the particular case of $r=1$, we see that $\tau_1(n)\sim\text{Li}(n)\sim\pi(n)$; i.e., the sum of the distinct prime factors of n is close to the the number of primes up to n.

4. Sum of divisors of n raised to the rth power $\sigma_r(n)$

The proof of Eq. (C6) can be used to find an expression for $\sigma_r(n)$, defined as the sum of the divisors of n raised to the rth power. Indeed, using Eqs. (C8) and (C9) we see that, if $\frac{n}{k} \notin \mathbb{N} \Rightarrow \lfloor \frac{n}{k} \rfloor = q = \lfloor \frac{n-1}{k} \rfloor$. On the other hand, if $\frac{n}{k} \in \mathbb{N} \Rightarrow \lfloor \frac{n}{k} \rfloor = q$ but $\lfloor \frac{n-1}{k} \rfloor = q - 1$. If we define $\psi(k;n) \equiv \lfloor \frac{n}{k} \rfloor - \lfloor \frac{n-1}{k} \rfloor$, we can write

$$\psi(k;n) = \begin{cases} 1 & \text{if } \frac{n}{k} \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}. \tag{C14}$$

As a consequence, we can easily sum all the divisors of n raised to any power r simply as

$$\sigma_r(n) = \sum_{k=1}^{n} \left(\frac{n}{k}\right)^r \psi(k;n) = \sum_{k=1}^{n} k^r \psi(k;n). \tag{C15}$$

The reason why the two sums in Eq. (C15) are equivalent is that, if $\frac{n}{k} \in \mathbb{N}$, both $\frac{n}{k}$ and k divide n.

APPENDIX D: THE PRIME COUNTING FUNCTION IN THE STOCHASTIC MODEL

In this section we derive an expression for P_N, the probability that N is prime in a network chosen at random from the set of all networks of size greater or equal to N generated by our model. Since this probability cannot depend on numbers that join the network after N, we only need to study these networks up to N in the calculation. We can describe the state of a particular realization up to N using the set of dichotomous random variables $(n_2, \ldots, n_N)$, where

$$n_k = \begin{cases} 1 & \text{if } k \text{ is prime} \\ 0 & \text{otherwise} \end{cases} \quad \text{with } k = 2, \ldots, N. \tag{D1}$$

This allows us to write P_N as

$$P_N = \langle n_N \rangle = \sum_{n_2=0}^{1} \cdots \sum_{n_N=0}^{1} n_N \rho(n_2, \ldots, n_N), \tag{D2}$$

where $\langle \cdot \rangle$ denotes the statistical average and $\rho(n_2, \ldots, n_N)$ is the joint probability of the particular sequence $(n_2, \ldots, n_N)$. It is convenient to define its characteristic function

$$\hat{\rho}(z_2, \ldots, z_N) \equiv \sum_{n_2=0}^{1} \cdots \sum_{n_N=0}^{1} z_2^{n_2} \ldots z_N^{n_N} \rho(n_2, \ldots, n_N). \tag{D3}$$

P_N can be derived from this expression as

$$P_N = \left. \frac{\partial \hat{\rho}}{\partial z_N} \right|_{z_2 = z_3 = \cdots = z_N = 1}. \tag{D4}$$

The set of random variables $(n_2, \ldots, n_N)$ defines a sequence of causal variables, in the sense that n_i only depends on n_j with $j < i$. This implies that $\rho(n_2, \ldots, n_N)$ satisfies the following Chapman-Kolmogorov equation:

$$\rho(n_2, \ldots, n_N) = \rho(n_2, \ldots, n_{N-1}) \, \text{Prob}\{n_N | n_2, \ldots, n_{N-1}\}, \tag{D5}$$

with $N \geqslant 3$ and the initial condition $\rho(n_2 = 1) = 1$. The conditional probability that N is prime given the sequence $(n_2, \ldots, n_{N-1})$ is the probability that N does not connect to any of the existing primes below $\sqrt{N}$, that is

$$\text{Prob}\{n_N = 1 | n_2, \ldots, n_{N-1}\} = \prod_{i=2}^{\lfloor \sqrt{N} \rfloor} \left(1 - \frac{1}{i}\right)^{n_i}, \tag{D6}$$

and $\text{Prob}\{n_N = 0 | n_2, \ldots, n_{N-1}\} = 1 - \text{Prob}\{n_N = 1 | n_2, \ldots, n_{N-1}\}$. Plugging this expression into Eq. (D5) and then into Eq. (D3) leads to the following recurrence relation

$$\begin{aligned} &\hat{\rho}(z_2, \ldots, z_N) \\ &\quad = \hat{\rho}(z_2, \ldots, z_{N-1}) + (z_N - 1) \\ &\qquad \times \hat{\rho}(z_2 \alpha_2, \ldots, z_{\lfloor \sqrt{N} \rfloor} \alpha_{\lfloor \sqrt{N} \rfloor}, z_{\lfloor \sqrt{N} \rfloor + 1}, \ldots, z_{N-1}), \end{aligned} \tag{D7}$$

where we have defined the compact notation $\alpha_i \equiv 1 - 1/i$. Finally, by making use of Eq. (D4), we obtain

$$P_N = \hat{\rho}(\alpha_2, \ldots, \alpha_{\lfloor \sqrt{N} \rfloor}). \tag{D8}$$

From Eq. (D7) it is clear that the random variables $(n_2, \ldots, n_{N-1})$ are not statistically independent. This implies that the exact solution of the problem can only be obtained by solving Eq. (D7) and plugging the solution into Eq. (D8), a task that is, currently, beyond our mathematical skills. Nevertheless, it is possible to derive a very accurate mean-field approximation. We start by expanding $\hat{\rho}(z_2, \ldots, z_{\lfloor \sqrt{N} \rfloor})$ around $z_1 = z_2 = \cdots = z_{\lfloor \sqrt{N} \rfloor} = 1$ as

$$\begin{aligned} P_N = 1 &+ \sum_{i=2}^{\lfloor \sqrt{N} \rfloor} \left. \frac{\partial \hat{\rho}}{\partial z_i} \right|_{z_i = 1} \beta_i + \frac{1}{2!} \sum_{i=2}^{\lfloor \sqrt{N} \rfloor} \sum_{j=2}^{\lfloor \sqrt{N} \rfloor} \left. \frac{\partial^2 \hat{\rho}}{\partial z_i \partial z_j} \right|_{z_i = z_j = 1} \\ &\times \beta_i \beta_j + \frac{1}{3!} \sum_{i=2}^{\lfloor \sqrt{N} \rfloor} \sum_{j=2}^{\lfloor \sqrt{N} \rfloor} \sum_{k=2}^{\lfloor \sqrt{N} \rfloor} \left. \frac{\partial^3 \hat{\rho}}{\partial z_i \partial z_j \partial z_k} \right|_{z_i = z_j = z_k = 1} \\ &\times \beta_i \beta_j \beta_k + \cdots, \end{aligned} \tag{D9}$$

where we have used the convenient notation $\beta_i \equiv \alpha_i - 1 = -1/i$. All terms in the latter expansion that involve a derivative of order higher than one with respect to any of the z_i are null, since $n_i(n_i - 1) = 0$ (n_i is either 0 or 1). Using this fact and the properties of generating functions, we can rewrite Eq. (D9) as

$$\begin{aligned} P_N = 1 &+ \sum_i \langle n_i \rangle \beta_i + \sum_{i<j} \langle n_i n_j \rangle \beta_i \beta_j \\ &+ \sum_{i<j<k} \langle n_i n_j n_k \rangle \beta_i \beta_j \beta_k + \cdots + \langle n_2 n_3 \cdots n_{\lfloor \sqrt{N} \rfloor} \rangle \\ &\times \beta_2 \beta_3 \cdots \beta_{\lfloor \sqrt{N} \rfloor}. \end{aligned} \tag{D10}$$

Despite the fact that random variables n_i are not statistically independent, in most of the cases they are conditionally independent. For instance, let us first consider the term $\langle n_i n_j \rangle$ for $i > j$. If $j > \sqrt{i}$ then the only correlation between n_i and n_j is given through their common history, that is, the sequence of primes up to $\sqrt{j}$ and, therefore, they are conditionally independent. In the opposite case, n_i is correlated to n_j. However, notice that (i) n_j is only one out of $\sqrt{i}$ variables that have a direct influence on n_i. (ii) The common history between

n_i and n_j is even smaller than before, and (iii) the number of correlated terms for a given N is $\sum_{j=3}^{\lfloor\sqrt{N}\rfloor}\sqrt{j}\sim N^{3/4}$, whereas the total number of terms scales as N^2. Given these considerations, it is quite reasonable to factorize $\langle n_i n_j\rangle \approx \langle n_i\rangle\langle n_j\rangle = P_i P_j$. A similar analysis can be performed for higher-order correlation functions. Under this approximation, Eq. (D10) can be written as

$$P_N \approx 1+\sum_i \langle n_i\rangle\beta_i+\sum_{i<j}\langle n_i\rangle\langle n_j\rangle\beta_i\beta_j + \sum_{i<j<k}\langle n_i\rangle\langle n_j\rangle\langle n_k\rangle\beta_i\beta_j\beta_k+\cdots + \langle n_2\rangle\cdots\langle n_{\lfloor\sqrt{N}\rfloor}\rangle\beta_2\cdots\beta_{\lfloor\sqrt{N}\rfloor}. \quad \text{(D11)}$$

The latter sum can be expressed as

$$P_N \approx \sum_{m_2=0}^{1}\cdots\sum_{m_{\lfloor\sqrt{N}\rfloor}=0}^{1}(\langle n_i\rangle\beta_i)^{m_i} = \prod_{i=2}^{\lfloor\sqrt{N}\rfloor}(1+\langle n_i\rangle\beta_i) = \prod_{i=2}^{\lfloor\sqrt{N}\rfloor}\left(1-\frac{P_i}{i}\right). \quad \text{(D12)}$$

Finally, we can write

$$P_N \approx e^{\sum_{i=2}^{\lfloor\sqrt{N}\rfloor}\ln(1-\frac{P_i}{i})}. \quad \text{(D13)}$$

In the limit $N\to\infty$, the sum in the exponent of the exponential function is dominated by the upper limit and, therefore, it can be approximated as

$$P_N \approx e^{-\sum_{i=2}^{\lfloor\sqrt{N}\rfloor}\frac{P_i}{i}}. \quad \text{(D14)}$$

APPENDIX E: THE ERDÖS-KAC THEOREM IN THE STOCHASTIC MODEL

The Erdös-Kac theorem states that the quantity $[\omega(N)-\ln\ln N]/\sqrt{\ln\ln N}$ behaves as a random variable that follows a standard normal distribution. This is known as the fundamental theorem of probabilistic number theory. In our model, this quantity is, indeed, a random variable. In this section, we develop an approximation for the probability that number N in our model has ω distinct prime factors, $P(\omega|N)$. To do so, we first define the set of dichotomous random variables $(m_2,m_3,\ldots,m_{\lfloor\sqrt{N}\rfloor})$ as follows:

$$m_k = \begin{cases}1 & \text{if } k \text{ is a prime factor of } N\\ 0 & \text{otherwise}\end{cases} \quad \text{with } k=2,\ldots,\lfloor\sqrt{N}\rfloor. \quad \text{(E1)}$$

In terms of these variables, we can write

$$P(\omega|N) = \sum_{n_2=0}^{1}\cdots\sum_{n_{\lfloor\sqrt{N}\rfloor}=0}^{1}\rho(n_2,\ldots,n_{\lfloor\sqrt{N}\rfloor})\sum_{m_2=0}^{1}\cdots\sum_{m_{\lfloor\sqrt{N}\rfloor}=0}^{1}\text{Prob}\{m_2,\ldots,m_{\lfloor\sqrt{N}\rfloor}|n_2,\ldots,n_{\lfloor\sqrt{N}\rfloor}\}\delta_{\omega,1+\sum_i m_i}, \quad \text{(E2)}$$

where $\delta_{\cdot,\cdot}$ is the Kronecker δ function. The conditional probability of variables m_i satisfies

$$\text{Prob}\{m_2,\ldots,m_{\lfloor\sqrt{N}\rfloor}|n_2,\ldots,n_{\lfloor\sqrt{N}\rfloor}\} = \text{Prob}\{m_2|n_2\}\text{Prob}\{m_3|n_3,m_2\}\text{Prob}\{m_4|n_4,m_2,m_3\}\cdots, \quad \text{(E3)}$$

with

$$\text{Prob}\{m_j|n_j,m_2,m_3,\ldots,m_{j-1}\} = \delta_{m_j,1}\frac{n_j}{j}\theta\left(\sqrt{\frac{N}{\prod_{i=1}^{j-1}i^{n_i m_i}}}-j\right) + \delta_{m_j,0}\left[1-\frac{n_j}{j}\theta\left(\sqrt{\frac{N}{\prod_{i=1}^{j-1}i^{n_i m_i}}}-j\right)\right]. \quad \text{(E4)}$$

In the latter expression, $\theta(x)$ is the Heaviside step function. Notice that this step function accounts for the fact that j cannot be a prime factor of N if there already exist smaller prime factors such that j is above the square root of the ratio between N and the product of all prime factors smaller than j. Dropping this restriction would correspond to evaluate the distribution of a random variable $\hat{\omega}$ that is an upper bound of ω. However, in the limit $N\to\infty$, since the probability of j being a prime factor decreases as $1/j$, most of the prime factors of N are small numbers for which the argument of the Heaviside function in Eq. (E4) is always positive. We then expect that, in such limit, $\hat{\omega}\to\omega$ and so we can safely drop the Heaviside function in Eq. (E4). Under this approximation, the generating function of $P(\omega|N)$ can be written as

$$\hat{P}(z|N)\equiv\sum_{\omega=1}^{\infty}z^{\omega}P(\omega|N) = z\sum_{n_2=0}^{1}\cdots\sum_{n_{\lfloor\sqrt{N}\rfloor}=0}^{1}\rho(n_2,\ldots,n_{\lfloor\sqrt{N}\rfloor}) \times\prod_{j=2}^{\lfloor\sqrt{N}\rfloor}\left[1+(z-1)\frac{n_j}{j}\right], \quad \text{(E5)}$$

and using the same mean-field approximation that we used in the previous section, we can write

$$\hat{P}(z|N) = z\prod_{j=2}^{\lfloor\sqrt{N}\rfloor}\left[1+(z-1)\frac{P_j}{j}\right]\approx z e^{(z-1)\sum_{j=2}^{\lfloor\sqrt{N}\rfloor}\frac{P_j}{j}}. \quad \text{(E6)}$$

We now use Eq. (D14) to obtain

$$\hat{P}(z|N)\approx z e^{-(z-1)\ln P_N}, \quad \text{(E7)}$$

or, equivalently,

$$P(\omega|N) = \frac{P_N}{(\omega-1)!}[-\ln P_N]^{\omega-1}. \quad \text{(E8)}$$

This is nothing but a Poisson distribution of average $-\ln P_N\sim\ln\ln N$ and standard deviation $\sqrt{\ln\ln N}$, which, for large N, converges to a Gaussian distribution.

[1] Leonhard Euler about the structure of primes: *"Mathematicians have tried in vain to this day to discover some order in the sequence of prime numbers, and we have reason to believe that it is a mystery into which the human mind will never penetrate."*—Leonhard Euler, 1751.

[2] J.-P. Delahaye, *Merveilleux Nombres Premiers* (Belin-Pour La Science, Paris, 2013), p. 336.

[3] B. L. Julia, *Number Theory and Physics*, edited by J. M. Luck, P. Moussa, and M. Waldschmidt (Springer-Verlag, Berlin, 1990).

[4] J. Bost and A. Connes, Selecta Math. **1**, 411 (1995).

[5] D. I. Fivel, arXiv:hep-th/9409150.

[6] P. J. Forrester and A. M. Odlyzko, Phys. Rev. E **54**, R4493 (1996).

[7] J. V. Armitage, *Number Theory and Dynamical Systems*, edited by M. M. Dodson and J. A. G. Vickers, LMS Lecture Notes, series 134 (Cambridge University Press, Cambridge, 1989).

[8] D. Schumayer and D. A. W. Hutchinson, Rev. Modern Phys. **83**, 307 (2011).

[9] D. Spector, J. Math. Phys. **39**, 1919 (1998).

[10] B. Luque, O. Miramontes, and L. Lacasa, Phys. Rev. Lett. **101**, 158702 (2008).

[11] P. W. Shor, SIAM J. Comput. **26**, 1484 (1997).

[12] J. I. Latorre and G. Sierra, Quant. Informat. Computat. **14**, 577 (2014).

[13] M. R. Watkins, *Number Theory and Physics Archive* (2013), http://empslocal.ex.ac.uk/people/staff/mrwatkin/zeta/physics.htm.

[14] R. Rivest, A. Shamir, and L. Adleman, Commun. ACM **21**, 120 (1978).

[15] L. Blum, M. Blum, and M. Shub, SIAM J. Comput. **15**, 364 (1986).

[16] *"As a boy I considered the problem of how many primes there are up to a given point. From my computations, I determined that the density of primes around n is about* $1/\ln n$*"*—Carl Friedrich Gauss, 1849.

[17] T. M. Apostol, *Introduction to Analytic Number Theory* (Springer-Verlag, New York, 1976).

[18] H. Cramér, Skand. Math. Kongr. **8**, 107 (1935).

[19] H. Cramér, Acta Arith. **2**, 23 (1936).

[20] M. Faloutsos, P. Faloutsos, and C. Faloutsos, in *SIGCOMM* (1999).

[21] D. Krioukov, M. Kitsak, R. S. Sinkovits, D. Rideout, D. Meyer, and M. Boguñá, Sci. Rep. **2**, 793 (2012).

[22] G. H. Hardy and S. Ramanujan, Quart. J. Math. **48**, 76 (1917).

[23] D. Hawkins, Math. Mag. **31**, 1 (1957).

[24] D. Hawkins, Number Theory **6**, 192 (1974).

[25] H. M. Bui and J. P. Keating, J. Number Theory **119**, 284 (2006).

[26] J. Lorch and G. Okten, Math. Mag. **80**, 112 (2007).

[27] P. Erdös and M. Kac, Am. J. Math. **62**, 738 (1940).

[28] We are implicitly assuming that the length of the interval is much larger than the typical gap and, therefore, that the number of gaps within the interval is large.

[29] M. Wolf, Phys. Rev. E **89**, 022922 (2014).

[30] H. Maier, Mich. Math. J. **32**, 221 (1985).

[31] A. Granville, in *Proceedings of the International Congress of Mathematicians (Zürich, 1994), Vols. 1 and 2* (Birkhäuser, Basel, 1995), pp. 388–399.

[32] A. M. Odlyzko and H. J. J. te Riele, J. Reine Angew. Math. **357**, 138 (1985).

[33] M. Rubinstein and P. Sarnak, Exp. Math. **3**, 173 (1994).

[34] G. H. Hardy and J. E. Littlewood, Acta Math. **44**, 1 (1923).

[35] M. Wolf, Physica A **160**, 24 (1989).