Heliyon 7 (2021) e08422

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Estimating the epidemic growth dynamics within the first week

Vincenzo Fioriti [a], Marta Chinnici [a,*], Andrea Arbore [b], Nicola Sigismondi [b], Ivan Roselli [a]

[a] ENEA- C.R Casaccia, Via Anguillarese 301, Rome, 00123, Italy
[b] ICT-Technical Consultant, Rome, Italy

ARTICLE INFO

Keywords:
Complex network
Dynamical systems
Graph theory
Big data
Epidemic spreading
Infective diseases

ABSTRACT

Information about the early growth of infectious outbreaks is indispensable to estimate the epidemic spreading. A large number of mathematical tools have been developed to this end, facing as much large number of different dynamic evolutions, ranging from sub-linear to super-exponential growth. Of course, the crucial point is that we do not have enough data during the initial outbreak phase to make reliable inferences. Here we propose a straightforward methodology to estimate the epidemic growth dynamic from the cumulative infected data of just a week, provided a surveillance system is available over the whole territory. The methodology, based on the Newcomb-Benford Law, is applied to the Italian covid 19 case-study. Results show that it is possible to discriminate the epidemic dynamics using the first seven data points collected in fifty Italian cities. Moreover, the most probable approximating function of the growth within a six-week epidemic scenario is identified.

1. Introduction

During an outbreak, a major issue to stop or mitigate the virus diffusion is gathering information as soon as possible about the nature of the epidemic from a mathematical point of view. Pandemic outbreaks allow a reasonable amount of data to only ex-post the event. Therefore the analysis is severely limited. On the other hand, well-timed information about the epidemic growth is highly precious and justifies any effort in this direction. Panoply of tools is available, but their accuracy is subject to limitations due to the small number of data points available, which also restricts the choice of models for the epidemic curve, a time-series of the cumulative number of cases per day [1, 2]. These curves are produced by different dynamics, ranging from sub-linear to super-exponential, giving rise to a diversity of the early growth profiles with profound implications for estimating the disease transmission and the implementation of the countermeasure [3]. Therefore, at least of the essential outbreak characteristics, fast detection and estimation would be beneficial, but the mathematical tools able to deal with those very few data are rare.

Moreover, epidemic data gathered on the field are always polluted by human errors, different collection methods, limited territorial coverage, irregular or random sampling. Even the most recent work [4, 5, 6, 7, 8] have been implemented new sophisticated signal processing techniques such as the Graph spectral analysis, Compressive Sensing, the Signal on Graphs method are affected by these problems [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. However, the Newcomb-Benford law (NBL) seems to reveal the epidemic dynamics using only a week of infection data. NBL is the statistic of the first (actually also of the second, third and so on) digit for a set of numbers, discovered and rediscovered independently by S. Newcomb and F. Benford (NB), today more known as the Benford law [27, 28, 29]. Its wide popularity is due to the apparent ubiquity of the NB distribution and the extreme simplicity of the calculation procedure. In recent years, the NBL has been used to discover fiscal frauds and to confirm scientific data reliability, including epidemic data [30]. In our work, we are interested in studying the capabilities of the NBL to predict the outbreak growth dynamic using very few initial data. It is only necessary to collect the daily infected cumulative data over a week and in some of the most critical cities involved in the outbreak to form a unique sequence of these numbers and then calculate the first digit distribution. Given the limited amount of data points, the calculated and the actual Benford distribution will not coincide exactly, thus defining the accuracy of an appropriate Goodness-of-Fit (GoF) parameter is used. At this point, by trial-and-error or any numerical technique, an approximating function is chosen: if its first digit distribution is congruent with that of the Italian cities, we can consider the approximating function as an accurate approximation to the real cumulative curve (given the GoF is small enough). In the following sections, we will show how this is possible invoking the Theorems of Berger & Hill and the ergodicity of the epidemic SIS process during the initial expansion

* Corresponding author.
E-mail address: marta.chinnici@enea.it (M. Chinnici).

https://doi.org/10.1016/j.heliyon.2021.e08422
Received 15 June 2021; Received in revised form 22 October 2021; Accepted 15 November 2021

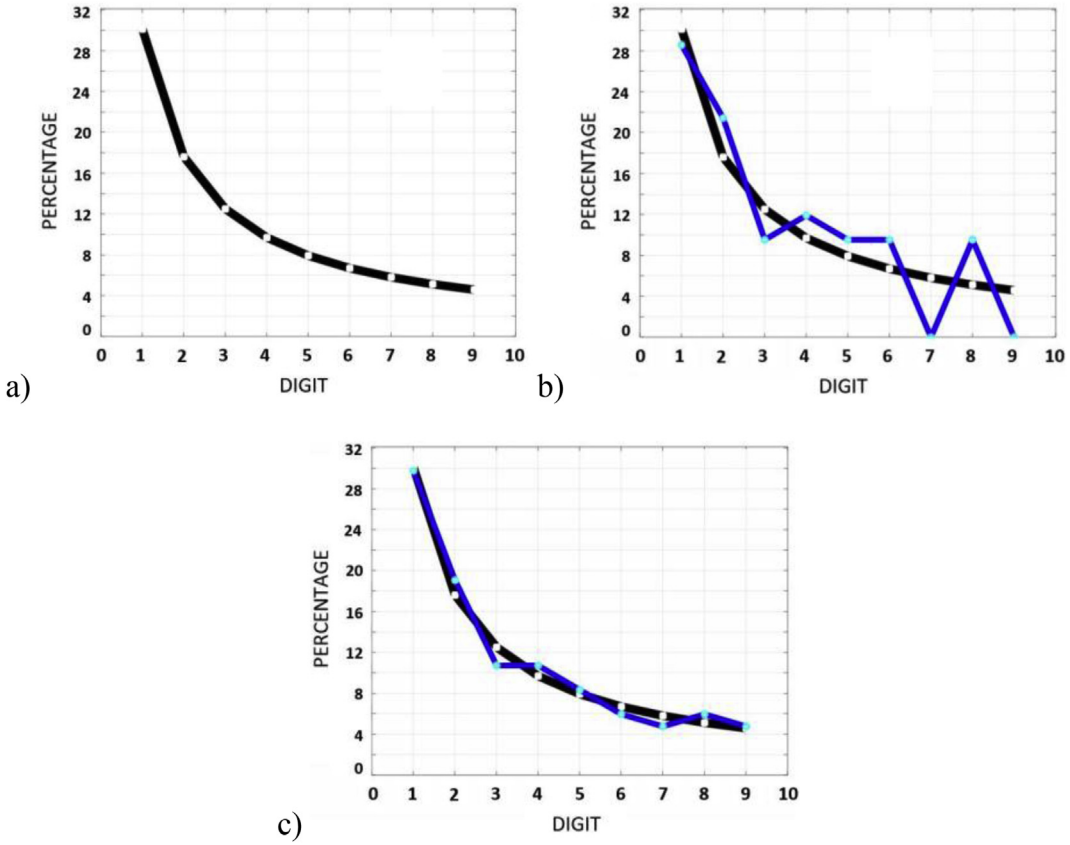

Figure 1. a) The Benford distribution: digits 1–9 and their percentages: 30.1, 17.6, 12.5, 9.69, 7.92, 6.69 5.80, 5.12, 4.58. First digit distribution with a limited amount of data: b) 2n first 42 data-points (blue) and c) 2n first 84 data-points (blue). The curve of Figure 1c is close to the real Benford distribution; in this case, gof = 0.99813, to be compared to the gof = 3.5803 for the previous one.

phase. In this work, the authors assume that the outbreak is uniformly distributed across an area but that the epidemic SIS process during the initial expansion phase is ergodic if [31]:

$$R^s{}_0 = 2(\beta N - \mu - \gamma) \,/\, \sigma^2 N^2 > 1$$

where μ is the birth rate, γ is the cure rate, and β is the contact rate [31]. This means, broadly speaking, that the ensemble statistics is equivalent to the time average. Therefore, it would be reasonable to use a very long sequence from only one city or very short sequences of many towns to get some mathematical insights. In particular, in this paper, the authors focus their interest on the second case. In other words, the local properties of an outbreak may be different, but the global behaviour is identifiable. It remains to define what it should be intended as "country" or "area". Since a national authority collects data, it seems natural to consider a country within its borders; that could be a bad idea if the state is so significant to generate too variable situations (meaning the economic development, social relations, languages, roads, etc.)

The paper is organised as follows: Section I – Introduction; Section II – Background: Newcomb – Benford Law; Section III – Application; Section IV –Conclusion.

2. Background: Newcomb – Benford Law

This section presents our proposal's essential points; we discuss the main statistical tool, namely the Newcomb-Benford Law. Since an extensive treatment can be found in the fundamental work of Hill and Berger [27, 28, 29], we will give just a brief introduction to the NBL from a practical point of view. S. Newcomb and F. Benford independently observed that the leading digits in many real-life numerical data sets such as macroeconomic, census, financial, fiscal data, were not distributed uniformly, as the common sense would suggest, instead they follow the logarithmic distribution of Figure 1a (see Figure 2).

More formally, the Benford distribution is a logarithmic curve:

$$Pr\,(D_1 = d_1) = log_{10}\,(1 + 1/d_1), \text{ for all } d_1 = 1, 2, \ldots, 9 \qquad (1)$$

where, D_1 is the first *significant* decimal digit. First significant digit means 4 in 0.467, 5 in 58.34, 3.7 in $3.7*10^2$, 8.21 in $8.21*10^{-4}$. When a set of numbers follows exactly the NBL, the digit "1" will appear about the 30% of the times, the digit "2" the 17%, "3" the 12% etc. Actually, the Generalized NB Law considers also the second, third etc. digit.

The compliance to the Benford distribution may be found also in some natural data-sets such as molecular weight tables, sport statistics, drainage areas of rivers that taken individually do not follow the NBL: what satisfies completely the NBL is the union of all those data-sets.

Some numerical sequences follow strictly the Benford first digit distribution. Let us consider the sequence of powers 2^n: {2 4 8 16 32 64 128 256 512 1024 2048 4096 8192 16384 32768 65536 131072 262144 524288 1048576 …}. The powers of 2 follow the Benford distribution (we say "is Benford", for short), but to verify it numerically, we would need of a large set of numbers. This constitutes one first difficulty, because it is not easy to determine the minimum cardinality of the set that guarantees *a priori* to reveal exactly the Benford distribution. Moreover, most of the time we do not have enough data to satisfy correctly the NBL, and, as a consequence, an error is introduced. In Figure 1b, c the effect of a limited data set is clearly illustrated: using only 42 data points we obtain a poor fit to the actual Benford distribution, although simply doubling the data-points reduces greatly the error. Hence, it is convenient to use a goodness-of-fit parameter to measure the error committed; here, we use a standard measure to this end:

$$GoF = \sqrt{\,(1/N_{data})*(\textstyle\sum_i (x_i - x_i^{real})^2)}\;\; i = 1, 2 \ldots N_{data} \qquad (2)$$

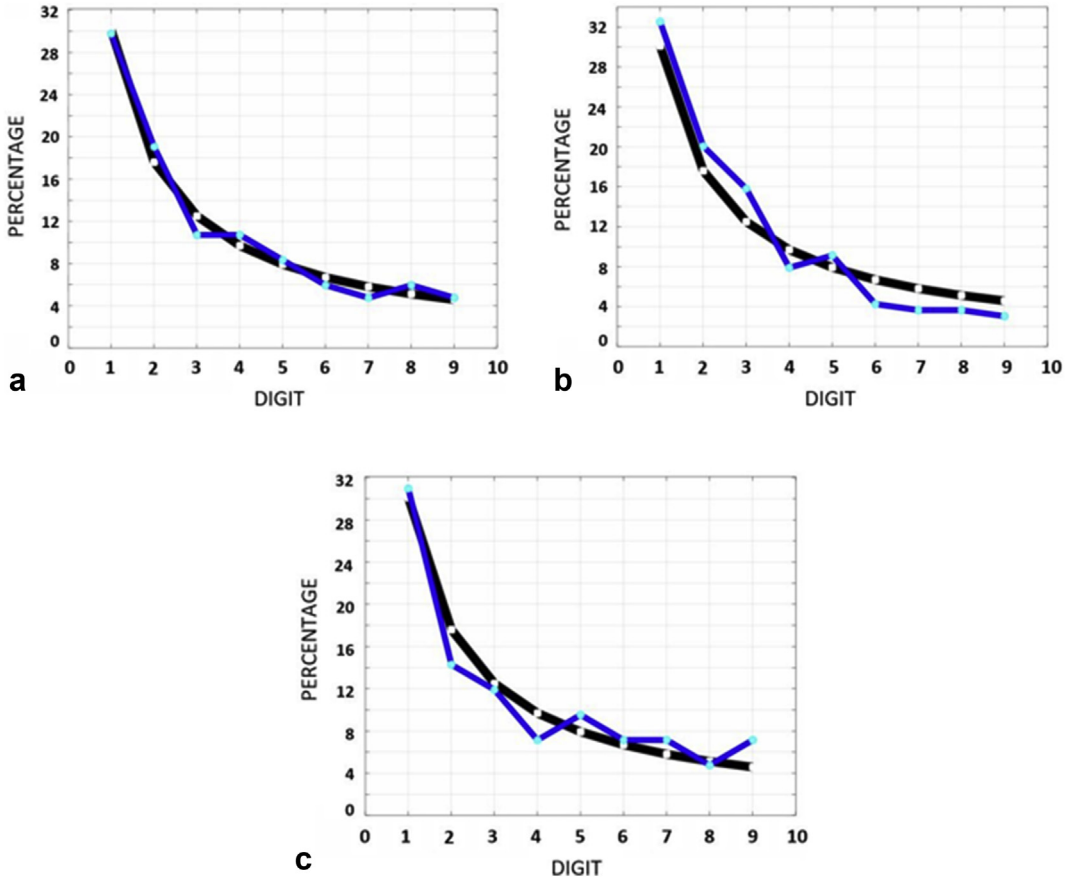

Figure 2. Visually, all the blue curves are similar to the Benford distribution (black). a) In blue: IT_real, the Italian cumulative data first digit distribution, GoF = 1.7313. b) Blue: the 50_cities first digit distribution, GoF = 1.8081. c) Blue: the logistic curve, GoF = 1.8206.

But, other statistical tests may be used as well. Of course, generally the first digit distribution of a data-set may or may not be a Benford (first digit) distribution; if this is the case, it will be clearly specified. In any case, the gof indicates the distance between the calculated firs digit distribution and the Benford distribution.

2.1. Theoretical justification and technical background

The main idea is to estimate an approximating function for the epidemic growth curve within a time horizon of Tf days, using only the first seven epidemic data points of fifty Italian cities, accounting for about 30% of the population, considered as a unique sequence formed of 50 × 7 data-points, called 50_cities sequence. We show that the first digit distribution of the 50_cities sequence converges to the first digit distribution of the cumulative daily infected Italian national sequence, formed summing the infected over all the national territory each day during the early Tf days of the epidemic ascending phase. Therefore, if the convergence exists, it is possible to know the compliance to Benford for the cumulative Italian curve in advance of Tf – 7 days. In turn, the level of compliance is used as a criterion to predict the accuracy of an approximating function to the actual epidemic national curve during the initial phase of Tf days (starting from the 21st February 2020).

Now we will sketch the theoretical justification of the above method. First of all, we have to ensure that the Italian national cumulative data first digit distribution is Benford during the initial spreading period. Based on the Berger & Hill theorems [27, 28], if the sequence is a power law, exponential or super-exponential, its first digit distribution is almost always Benford. Thus, we have only a sufficient condition. Below some of Berger & Hill's main results to support our method (formal demonstrations can be found in [28, 29, 30]).

Theorem. **(Berger & Hill 1).** *None of the classical probability distributions or random variables, such as the normal, uniform, exponential, beta, binomial, or gamma distributions are Benford. Specifically, no uniform distribution is even close to Benford, no matter how large its range or how it is centered. However, some distributions come close to being Benford, such as the Pareto and the Log-normal distribution.*

The next Theorem states that the solutions of an ordinary or differential equation system such as many of the classical epidemic models, under general conditions are Benford. Thus, the NBL is not restricted to discrete dynamics; on the other hand, general results for partial differential, delay or integral-differential equations are not known.

Theorem. **(Berger & Hill 2).** *Consider the dynamic system*:

$$x' = F(x),\ x(0) = x_0 \tag{3}$$

where, F: $\mathbb{R} \to \mathbb{R}$ *is continuously differentiable with* $F(0) = 0$, *and* $x_0 \in \mathbb{R}$. *Let* F: $\mathbb{R} \to \mathbb{R}$ be C^2 *with* $F(0) = 0$ and assume $F'(0) < 0$. *Then, for every* $x_0 \neq 0$ *sufficiently close to* 0, *the solution of the system is Benford.*

In the theory of differential dynamical systems related to Theorem 2 we have the so-called Shadowing Lemma. The Lemma describes the behavior of the pseudo-trajectories (or sequences) near a locally structurally stable hyperbolic invariant set [27].

Shadowing Lemma. Let T: $\mathbb{R} \to \mathbb{R}$ be a map, and β a real number with $|\beta| > 1$. If $\sup_{x \in \mathbb{R}} |T(x) - \beta x| < +\infty$ then there exists, for every $x \in \mathbb{R}$, one and only one point x° such that the sequence $(T^n(x) - \beta^n x^\circ)$ is bounded.

Table 1. Various approximating functions are ordered according to the GoF value (second column). In the last column is described the speed of growth within the first 42 days with respect to the real speed of growth of the case-study. The GoF of the actual Italian cumulative curve (IT_real) is 1.7313, very close to the 50_cities GoF; the best approximating function function is the cubic I(ti)n3. Samples is the number of data-points used to calculate the first digit distribution and therefore also the GoF; they are 42, except for the 50_cities GoF, whose distribution is calculated using not more than 294 data-points, the first 7 data from each city. In bold * are indicated the non-congruent dynamics: for example, 4*4n has a very large GoF, nonetheless belongs to the class of fast dynamics, instead of the slow one.

function	Benford goodness-of-fit	samples	growth to 42 days
$5*3^n$	0.9182	42	very fast
$12*(1.342)^n$	1.0269	42	fast
$3*5^n$	1.4893	42	very fast
$16*n^3$	1.4934	42	fast
$2*8^n$	1.5819	42	very fast
$2*9^n$	1.6308	42	very fast
IT_real	**1.7313 -**	42	-
50_cities	**1.8081** + 4.4 %	294	-
$\mathbf{1.675*n^3}$	**1.8110** + 4.6 %	42	**as *IT_real***
logistic	**1.8206** + 5.1 %	42	**as *IT_real***
$2*7^n$	2.0493 + 18.0%	42	**very fast ***
$16*n^4$	3.1133	42	**very fast ***
$8*2^n$	3.1237	42	**very fast ***
$16*n^2$	5.0668	42	slow
$3*6^n$	5.6255	42	**very fast ***
$16*n$	6.9965	42	linear
$4*4^n$	7.1002	42	**very fast ***
$16*\sqrt{n}$	10.783	42	Sub-linear
$50*\log_{10}(n)$	15.312	42	Sub-linear

This means that every hyperbolic set has the shadowing property, thus every pseudo-trajectory (or sequence) stays uniformly close to some true trajectory, i.e. a pseudo-trajectory is "shadowed" by a true one. Considering the epidemic curve as the solution of a dynamic system, the Shadowing Lemma allows to believe that very close to it a pseudo-sequence exists, is related to the dynamic system and is structurally stable. In other words, finding an approximating function to the cumulative epidemic Italian curve would not be a mere accident, at least locally. Moreover, small perturbations of the system initial conditions do not change the approximation: therefore, errors during the initial data collections do not alter the result.

Theorem. (Berger & Hill 3). *Let X be exponential with mean* 1, *that is* $F_X(t) = \max(0, 1 - e^{-t})$, $t \in \mathbb{R}$. Even *though X is not exactly Benford, it is close to being Benford for all* $t \in [1, 10)$.

Theorem. (Berger & Hill 4). *The sequences:* 2^n, 3^n *are Benford;* n, $n+1$, $n!$, $\sqrt{10^n}$, 10^n, $4*4^n$ *are not. In general,* $a*x^b$, *with* $a > 0$ *and* $b > 1$ *is Benford almost always, but not always, therefore* x^2 *is almost always Benford. Moreover, every mixture of* 2^n *with a random unbiased sequence, is Benford.*

Berger and Hill also state that apart from some particular cases, processes with linear growth are not Benford. This allows identifying the slow epidemic growths, which are a phenomenon more common than previously though [1]. By slow we mean a linear, sub-linear or a polynomial growth.

Theorem. (Berger & Hill 5). *If X and Y are Benford sequences, also their sum* $X + Y$ *is Benford. If the sequence Z is not Benford,* $X + Y + Z$ *is Benford.*

Hence, if the cumulative infected sequence of the Italian cities is Benford, also the national cumulative sequence is Benford as well, during the weeks of the increasing phase. Note that an epidemic cumulative sequence cannot be random, being non-decreasing, therefore by Theorems Berger & Hill 1, 3, 4, 5 the fast cumulative epidemic curves are all Benford, but could exist also fast non-Benford curves in particular circumstances. Thus, we cannot rule out the possibility of non-Benford growth curves to have a fast dynamics, thought this would be seldom the case, see Table 1.

Now we can discuss the convergence of the 50_*cities* distribution to the distribution of the cumulative daily infected of the Italian national sequence after T_f days. The 50_*cities* sequence is the union of the first 7 days data for each city, but to fix ideas, let us considers only three cities, A, B, C, whose sequences of seven elements are:

A = {1, 1, 34, 54, 60, 75, 94}, B = {3, 3, 3, 6, 6, 7, 14}, C = {5, 7, 10, 13, 15, 40, 55}

Thus the union U of A, B, C clearly reads:

$U\{A, B, C\}$ = {1, 1, 34, 54, 60, 75, 94, 3, 3, 3, 6, 6, 7, 14, 5, 7, 10, 13, 15, 40, 55 }

The cumulative number of cases per day is simply the total sum per day of the infected cases collected over all the N_{IT} towns and cities of Italy, during the first T_f days of the outbreak:

$$\sum_{k,h} (x^k(h))\ k = 1, 2 \ldots N_{IT}.\ h = 1, 2 \ldots T_f.$$

Indicating $B(\ldots)$ as the operator of the first digit distribution calculation, i.e. the Benford distribution, we show the following.

2.2. Remark on the convergence

$$B(U_{i,j}\{\ x^j(i)\ \}) \to B(\sum_{k,h} (x^j(h))) + O(m, N) \qquad (4)$$
$$N \to N_{IT}$$

where, $i = 1, 2, \ldots m.\ j = 1, 2, \ldots N.\ k = 1, 2, \ldots N_{IT}.\ h = 1, 2, \ldots T_f.$

In the trivial case, when the sequences x^j are (almost all) Benford with $m = T_f$ and $N = N_{IT}$ are the cities, the Berger & Hill (Theorem 5) guarantees that the summation of Benford sequence is Benford as well. If $m << T_f$, on the right side we still have a Benford distribution, but on the left side we have $m*N$ samples not selected randomly, as requested to form a Benford dataset, see [28] (see T. 6.20), [32]. Indeed, the cities can be chosen at random, but the m samples are always the first, therefore are not randomly chosen. Furthermore, consider that the sequences xj depend on a beta distribution, meaning that the real observed values during the data collection undergo a deterioration process described by a beta distribution [33] (this assumption may be relaxed, but for simplicity, we keep it). Ruankong and Sumetkijakan demonstrate that the truncate sequences union converges to Benford [34], provided the total number of sample $m*N$ is sufficiently large. Therefore, the intuitive reasoning that the ergodic nature of the epidemic should allow somehow to calculate the Benford distribution by a limited number of samples rests upon a solid basis. Of course, even if both sides of (4) converge to Benford, the error concerning the exact distribution of Figure 1a increases since we have very few samples.

Instead, if the sequences x^j are not Benford, generally, the convergence is not present.

2.3. Remark on the existence of an approximation

Given that the reproduction number is $R^s_0 > 1$ and that the N_{IT} local realizations of the epidemic process are almost all Benford, there is an approximation f to the solution T^n of the SIS system expressible as Benford summation sequence such that:

$$| T^n(h) - f(h) | \leq \beta^n h_0 \qquad (5)$$

with $\beta > 1$ *and* $h_0 \neq 0$.

By the Shadowing lemma there exists f close to $T^n(h)$ solution of the SIS model such that (5) is true. From the Theorem Berger & Hill 2 we know $T^n(h)$ is Benford, thus any good approximation f must be Benford too (and in our case this means also a low gof).

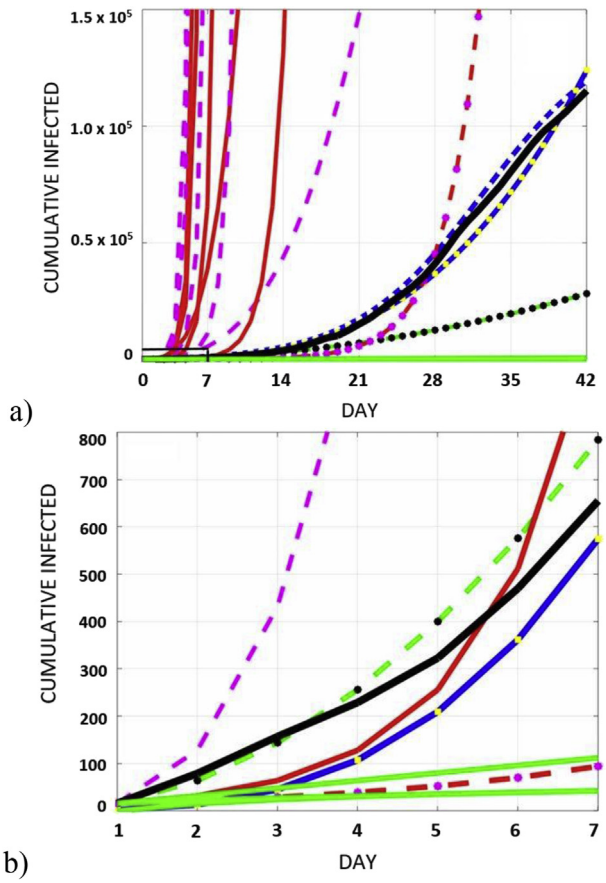

Figure 3. The first day is the 21th February 2020. a) 42 days scenario. From left to right in the color code the dynamics of Table 1; dotted magenta and red: very fast dynamics; black: the Italian cumulative curve and in blue the cubic and logistic approximations; green: slow dynamics. b) 7 days scenario. The cubic function (blue-yellow) approximate well the actual cumulative curve (black) also in this scenario, confirming the same result suggested in [35], while the logistic in this scenario performs poorly, and is not in the figure. Note the red dotted curve (12*1.342n), that seems slow, actually is a fast one; instead, the green black dotted curve (16n2) seems fast, but is very slow.

The SIS epidemic stochastic model during the initial phase of the outbreak is an *ergodic process* if $R^s{}_0 > 1$, as stated in [31], thus the local realization sequences of the cities have similar statistics (basically the time average is equal to the ensemble average). A natural candidate (but not the only one) for the approximation then could be:

$\sum_{k,h} x^j(h)$, for $N \leq N_{IT}$, $h = 1, 2, \ldots T_f$,

Since the ergodic properties prevent to consider $T_f \to \infty$ (provided N and T_f are not too small), while its "benfordness" can be readily calculated.

When $m = 7$, since it is not known an analytical method to determine *a priori* the values of m that guarantees a small *gof*, one can only calculate the first digit distribution $B(U_{i,j}\{\ x^j(i)\ \})$ and compare it to the Benford distribution of Figure 1a. If the *gof* stays high, the simplest heuristics would be to increase m, yet reducing the forecasting time-span of $T_f - 7$. Of course, the number of city-sequences N may well be expanded to all the cities N_{IT}: here we have restricted it to fifty cities only for demonstration purposes. Therefore, we do know that (4) is true, but cannot determine the minimum m necessary. Actually, we have chosen $m = 7$ because it is well below the prediction thresholds often suggested in the literature [2, 3]. Instead, to determine $T_f = 42$, we consider that during the initial phase of the outbreak if:

$$R^s{}_0 = 2(\beta N - \mu - \gamma) \,/\, \sigma^2 N^2 > 1 \tag{6}$$

where, μ is the birth rate, γ is the cure rate, and β is the contact rate, the overall epidemic process is ergodic [31]. Hence, all the local realizations have similar statistics, are non-random non decreasing sequences, and by the Shadowing lemma there exists an approximant function f to $\sum_{k,h}(x^j(h))$:

$$|\ \textstyle\sum_{k,h}(x^j(h) - f(x)\ | < \varepsilon \tag{7}$$

As a consequence, the first digit distribution of $f(x)$ approximates that of $\sum_{k,h}(x^j(h))$, but by the (4) also that of $B(U_{i,j}\{\ x^j(i)\ \})$. Therefore T_f can be determined heuristically or numerically as the value that get $B(f(x))$ closer to $B(U_{i,j}\{\ x^j(i)\ \})$ in terms of Benford Goodness-of-Fit.

From Table 1 it is readily seen that both the cubic and the logistic curve approximate the *IT_real GoF* very well for $T_f = 42$. In addition, $T_f = 42$ is very close to the inflection point of the real Italian cumulative curve, which indicates the end of the initial phase for the outbreak.

3. Application: the Italian case-study

Basically, we have three dynamics, very fast, fast, and slow (see Figure 3); we want to determine which one of them is prevailing by means of the Benford' GoF, and possibly to find an approximating function.

As said in the above Section, each city provides a sequence of 7 positive integers, and putting them together the fifty sequences make up a unique sequence of 50 × 7 integers called 50_*cities*. Note that sequences such as {1 2 2 2 2 2 2}, {3 1 1 1 1 1 1}, {1 1 1 1 1 1 1 1}, {1 0.3 … 0.01}, or {16 160 … 16000000}, are not taken into consideration; actually, 8 out of the 50 sequences have been discarded, thus the length of the 50_*cities* sequence is of 294 samples instead of 350.

In the Section above we have indicated that:

$$B(U_{i,j}\{\ x^j(i)\ \}) \to B(\textstyle\sum_{k,h}(x^j(h))) \tag{8}$$
$$N \to N_{IT}$$

where, $i = 1, 2, \ldots 7$. $j = 1, 2, \ldots 50$ $k = 1, 2, \ldots N_{IT}$ $h = 1, 2, \ldots 42$, $T_f = 42$ are now specified to the actual case-study scenario.

To classify the various possible approximant curves, we have calculated their Benford GoF, showed in the Table 1, together with the GoF of the real Italian epidemic data, the logistic curve, the cubic curve and of the 50_*cities*.

Keeping in mind that the *union* of the 50 sequences of seven data points from some of the most important Italian cities, converges to the Benford distribution of the *sum* of all the 6-week sequences for all the cities (5), Theorem Berger&Hill 3 [27] guarantee that *almost always* the exponential growth provides numbers according to the NB distribution. Therefore, as explained in other Sections, a good Benford gof obtained from the epidemic cumulative data of the 50_*cities*, should be able to identify the type of outbreak dynamic.

In Table 1 the 50 cities GoF is 1.8081, very close to the GoF for the real Italian data of the first six weeks, *GoF_IT* = 1.7313, while the cubic curve $GoF\_I(t_i)*n^3 = 1.8110$ has the minimal distance from the 50_cities GoF. Therefore n^3 is the most probable approximating function, whereas in Figure 3a the Italy' real epidemic curve is almost coincident with the cubic growth, confirming the results of [35]. Actually, the logistic curve GoF is in good agreement with the 50_*cities* GoF too, and fits very satisfactorily the Italy' real data *after* the fourth week (Figure 4b), but the cubic curve has an advantage in terms of Benford GoF and till the fourth week is also the best approximant. Therefore, Table 1 reveals that the cubic curve determines the initial stage of the growth.

Again a note of caution: a small GoF, say in the interval [0, 3] as in Table 1, is only a *sufficient* condition that guarantees a fast dynamic, but not a necessary one, meaning that a large *GoF* > 3 could represent a rapid growth too, as for the case of the function $B_0{}^*6^n$. Thus, most of the times,

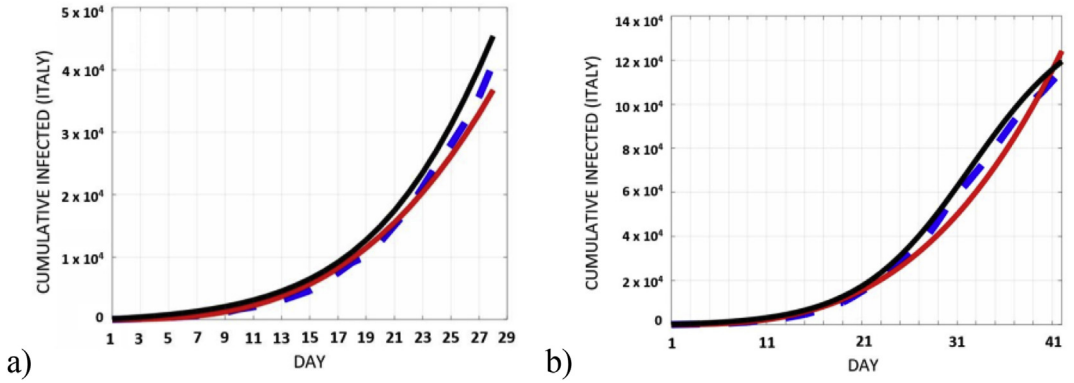

Figure 4. Dotted blue: the Italian cumulative curve, red: cubic approximant, black: logistic approximant. a) first four weeks, the cubic curve is very close to the real growth. b) first six weeks: the logistic curve now is a better approximant.

but not always, a large GoF indicates a linear or sub-linear dynamics; instead, a small GoF guarantees a rapid growth.

Summarizing, the first digit distribution of the 50_*cities* sequence converges to the first digit distribution of the cumulative infected curve that is Benford during the ascending epidemic phase. If the 50_*cities* GoF is small, on the basis of the Berger-Hill's Theorem we have a fast epidemic growth; instead, if the GoF is poor, the growth will be probably slow, although this cannot be assured formally. Moreover, in order to determine the form of the approximating function that shadows the real epidemic growth, one might extrapolate it analytically or heuristically. At this point, it suffices to choose the approximating curve whose GoF is close to that of the 50_*cities*. If the difference between the two GoF is small (in our case study less than 5%), we assume that the approximating function shadows correctly the real epidemic growth.

4. Conclusion

In this paper, we show how to estimate an approximating function of the epidemic curve within a time horizon of a five or six weeks, using only the first seven epidemic data points of fifty Italian cities, considered as a unique sequence. The level of compliance of the sequence to the Benford test is used as a criterion to predict the approximating function accuracy with respect to the real epidemic national curve. This procedure is made possible by the convergence of the unique sequence first digit distribution to the Benford distribution, since the national Italian cumulative data first digit distribution is Benford (almost always) when its sequence is a power, exponential or faster curve. Therefore, when a fast epidemic spread is taking place on the territory, its fingerprint will be the Benford distribution, otherwise the spreading will be, with high probability, slow (quasi-linear) or made of sporadic outbursts. Unfortunately, the Benford compliance of the fast growth is only a *sufficient* condition, not a necessary one. Our results make this point clear. On the other hand, early indications about the dynamic nature of the outbreak are made available through a simple statistical test applied to a handful of data. This method has been applied to the Italian covid 19 case study, where the most probable approximating function of the first five weeks has been identified clearly as a cubic curve.

Declarations

Author contribution statement

Vincenzo Fioriti: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Marta Chinnici: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Andrea Arbore, Nicola Sigismondi: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Ivan Roselli: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data included in article/supplementary material/referenced in article.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] C. Viboud, L. Simonsen, G. Chowell, A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks, Epidemics 15 (2016) 27–37.
[2] G. Texier, et al., Building test data from real outbreaks for evaluating detection algorithms, PLoS One (2017). September 1.
[3] Xia Jiang, G. Wallstrom, G.F. Cooper, M. Wagne, Bayesian prediction of an epidemic curve, J. Biomed. Inf. 42 (2009) (2009) 90–99.
[4] V. Fioriti, M. Chinnici, Identifying sparse and dense sub-graphs in large graphs with a fast algorithm, Europhys. Lett. 108 (2014) 50006.
[5] A. Arbore, V. Fioriti, Topological propagation of malware in networks: a survey, Int. J. Crit. Infrastruct. 9 (2) (2012).
[6] V. Fioriti, M. Chinnici, A. Arbore, Suboptimal Topological protection from Advanced Malware, Congresso Nazionale SIMAI, Politecnico di Torino, 2012.
[7] A. Arbore, V. Fioriti, M. Chinnici, The topological defense in SIS epidemic models, Chaos, Solit. Fractals 86 (2016) 16–22.
[8] M. Chinnici, V. Fioriti, Node seniority ranking in networks, Stud. Inf. Contr. 26 (4) (2017) 397–402.
[9] D. Shuman, et al., IEEE signal, Process. Mag. 84 (2013) 1053.
[10] S. Chen, R. Varma, A. Sandryhaila, J. Kovacevic, Discrete signal processing on

graphs: sampling theory, arXiv:1503.05432v1 [cs.IT] (3 Mar 2015).
[11] D.I. Shuman, S. Narang, P. Frossard, A. Ortega, P. Vandergheynst, The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains, IEEE Signal Process. Mag. 30 (3) (2013) 83–98.
[12] A. Sandryhaila, J.M.F. Moura, Discrete signal processing on graphs, IEEE Trans. Signal Process. 61 (7) (2013) 1644–1656.
[13] S. Hosseinalipour, J. Wang, Y. Tian, H. Dai, Infection analysis on irregular networks through graph signal processing, arXiv:1808.04879v1 [cs.SI] (2018), 14 Aug 2018.
[14] P.D. Lorenzo, S. Barbarossa, P. Banelli, S. Sardellitti, Adaptive least mean squares estimation of graph signals, IEEE Trans. Sign. Inform. Process. Over Network 2 (4) (2016) 555–568.
[15] P.D. Lorenzo, P. Banelli, E. Isufi, S. Barbarossa, G. Leus, Distributed recursive least squares strategies for adaptive reconstruction of graph signals, in: 2017 25th European Signal Processing Conference (EUSIPCO), 2017, pp. 2289–2293.
[16] M. Spelta, W. Martins, Online Temperature Estimation using Graph Signals, XXXVI Simposio Brasileiro de Telecomunicacoes e Processamento de Sinais, Campina Grande, 2018.
[17] V. Fioriti, M. Chinnici, J. Palomo, Predicting the sources of an outbreak with a spectral technique, Appl. Math. Sci. 8 (133-136) (2014) 6775–6782.
[18] P. Pinto, P. Thiran, M. Vetterli, Locating the source of diffusion in large scale network, Phys. Rev. Lett. 109 (2012) 68702–68709.
[19] V. Colizza, A. Barrat, M. Barthélemy, A. Vespignani, The role of the airline transportation network in the prediction and predictability of global epidemics, Proc. Natl. Acad. Sci. Unit. States Am. 103 (7) (2006).
[20] [a] G.C. Fanti, P. Kairouz, P. Viswanath, Hiding the rumor source, in: IEEE Transactions on Information Theory, 2017;
[b] X. Li, Y. Liu, C. Zhao, X. Zhang, D. Yi, Locating multiple sources of contagion in complex networks under the SIR model, Appl. Sci. 9 (20) (2019) 4472.
[21] M. Chinnici, V. Fioriti, A. Arbore, The network topology of connecting things: defense of IoT graph in the smart city, Lect. Notes Comput. Sci. 11539 (XV) (2019) 640.
[22] A. Vespignani, Modelling dynamical processes in complex socio-technical systems, Nat. Phys. (2012), 8 January 2012.
[23] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Phys. Rev. Lett. 86 (14) (2001) 3201–3203.
[24] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, C. Faloutsos, Epidemic thresholds in real networks, ACM Trans. Inf. Syst. Secur. 10 (4) (2008) 1–26.
[25] Y. Wang, D. Chakrabarti, C. Wang, C. Faloutsos, Epidemic Spreading in Real Networks: an Eigenvalue Viewpoint, in: SRDS Conference, 2003.
[26] S.J. Yan, A.A. Chughtai, C.R. Macintyre, Utility and potential of rapid epidemic intelligence from internet-based sources, Int. J. Infect. Dis. 63 (2017) 77–87.
[27] T. Hill, A statistical derivation of the significant digit law, Stat. Sci. 10 (4) (1995) 354–363.
[28] A. Berger, T. Hill, A basic theory of Benford's law, Probab. Surv. 8 (2011).
[29] T. Hill, The first digit phenomenon, Am. Sci. (1998) 86.
[30] R. Joannes-Boyau, T. Bodin, A. Scheffers, M. Sambridge, S.M. May, Using Benford's law to investigate Natural Hazard dataset homogeneity, Nature 2015 (July 2015).
[31] Q. Han, D. Jiang, C. Yuan, Extinction and Ergodic Property of Stochastic SIS Epidemic Model with Nonlinear Incidence Rate, Hindawi Publishing Corporation, 2013, p. 127321.
[32] Z. Cai, A.J. Hildebrand, J. Li, A local Benford Law for a class of arithmetic sequences, arXiv preprint arXiv:1808.01496 (2021).
[33] S. Pasquali, A. Pievatolo, A. Bodini, F. Ruggeri, A stochastic SIR model for the analysis of the COVID-19 Italian epidemic, arXiv preprint arXiv:2102.07566 (2021).
[34] P. Ruankong, S. Sumetkijakan, Chains of truncated beta distributions and Benford's law, Unif. Distrib. Theor. 14 (2019) 2.
[35] G. De Natale, et al., The COVID-19 infection in Italy: a statistical study of an abnormally severe disease, J. Clin. Med. (2020).