Dynamics of book sales: Endogenous versus exogenous shocks in complex networks

F. Deschâtres[1] and D. Sornette[2,3,*]
[1]*Ecole Normale Supérieure, rue d'Ulm, Paris, France*
[2]*Institute of Geophysics and Planetary Physics and Department of Earth and Space Sciences, University of California, Los Angeles, California 90095, USA*
[3]*Laboratoire de Physique de la Matière Condensée, CNRS UMR 6622 and Université de Nice-Sophia Antipolis, 06108 Nice Cedex 2, France*
(Received 28 December 2004; published 18 July 2005)

We present an extensive study of the foreshock and aftershock signatures accompanying peaks of book sales. The time series of book sales are derived from the ranking system of Amazon.com. We present two independent ways of classifying peaks, one based on the acceleration pattern of sales and the other based on the exponent of the relaxation. They are found to be consistent and reveal the coexistence of two types of sales peaks: exogenous peaks occur abruptly and are followed by a power law relaxation, while endogenous sale peaks occur after a progressively accelerating power law growth followed by an approximately symmetrical power law relaxation which is slower than for exogenous peaks. We develop a simple epidemic model of buyers connected within a network of acquaintances which propagates rumors and opinions on books. The comparison between the predictions of the model and the empirical data confirms the validity of the model and suggests in addition that social networks have evolved to converge very close to criticality (here in the sense of critical branching processes of opinion spreading). We test in detail the evidence for a power law distribution of book sales and confirm a previous indirect study suggesting that the fraction of books (density distribution) $P(S)$ of sales S is a power law $P(S) \sim 1/S^{1+\mu}$ with $\mu \approx 2$.

DOI: 10.1103/PhysRevE.72.016112 PACS number(s): 89.75.Hc, 64.60.Ak, 87.23.Ge

I. INTRODUCTION

In the context of linear response theory, the fluctuation dissipation theorem provides an explicit relationship between microscopic dynamics at equilibrium and the macroscopic response that is observed in a dynamic measurement. It relates equilibrium fluctuations to close-to-equilibrium observables. In out-of-equilibrium systems, the existence of a relationship between the response function to external kicks and spontaneous internal fluctuations is not settled [1]. In many complex systems, this question amounts to distinguishing between endogeneity and exogeneity and is important for understanding the relative effects of self-organization versus external impacts. This is difficult in most physical systems because externally imposed perturbations may lie outside the complex attractor which itself may exhibit bifurcations. Therefore, observable perturbations are often misclassified. It is thus interesting to study other systems, in which the dividing line between endogenous and exogenous shocks may be clearer in the hope that it would lead to insight about complex physical systems. The systems for which the endogenous-exogenous question is relevant span, beyond the physical sciences, the biological to the social sciences [2].

Here, we study a real-world example of how the response function to external kicks is related to internal fluctuations. We study the precursory and recovery signatures accompanying shocks in complex networks, that we test on the database of book ranks provided by Amazon.com. We find clear distinguishing signatures classifying two types of sales peaks. Exogenous peaks occur abruptly and are followed by a power law relaxation, while endogenous sales peaks occur after a progressively accelerating power law growth followed by an approximately symmetrical power law relaxation which is slower than for exogenous peaks. These results are rationalized quantitatively by a simple model of epidemic propagation of interactions with long memory within a network of acquaintances. The slow relaxation of sales implies that the sales dynamics is dominated by cascades rather than by the direct effects of news and advertisements, indicating that the social network is close to critical. We perform also a direct measurement on ranks that give important constraints on the conversion that transforms sales into ranks. A short version of this work is [3].

The structure of the paper is as follows. We first present our database (Junglescan.com) and explain how we use it to estimate the sales from the ranks. To do so, we need to make some assumptions about the mechanisms underlying the sales dynamics. We discuss these points on Sec. II. In Sec. V, we revisit this question and derive a way to measure directly this conversion from ranks to sales. The epidemic branching process used in our study leads to an effective linear coarse-grained description of the complex nonlinear dynamics. We present this model in Sec. III and derive some predictions about the behavior of the system before and after sales peaks. Then the data are analyzed in Sec. IV. We provide two independent classifications of exogenous and endogenous shocks. The obtained classifications are robust. The last Sec. VI concludes with suggestions for future research.

*Electronic address: sornette@moho.ess.ucla.edu

II. CONSTRUCTION OF THE DATA: CONVERTING RANKS INTO SALES

A. The database

The study was performed using data from Amazon.com. Amazon was founded in 1994 as an online bookseller, and since has expanded its business into areas such as clothing, gourmet food, sports equipment and jewelry. With an expected $6 billion net sales in 2004, the American giant of internet trading is by far the largest e-retailer.

Electronic data make it possible to deal with huge amount of information impossible to gather otherwise. With the birth of the Internet, the prospect of understanding quantitatively social phenomena is emerging [4]. It is perhaps, the first time that an organization can collect so much information [5]. What people buy can help give an image of the state of the society. And Amazon can provide almost all of what can be bought. It operates on a global scale. That is why Amazon is an outstanding means to probe the society. As Andreas S. Weigend, former chief scientist of Amazon.com (2002–2004) wrote on his webpage: "Amazon.com might be the world's largest laboratory to study human behavior and decision making." Amazon does not release actual sales data. However, it provides on its website the rank for each of its products based on their past sales.

In a first step, Y. Ageon in our group developed an application using "WebBrowser control" to capture automatically the ranks of books on Amazon.com at regular time intervals. WebBrowser control [6] allows a user to browse sites on the Internet's World Wide Web. The application first opens Amazon's page with the URL: http://207.171.185.16/exec/obidos/ASIN/XXXXXX where XXXXXX corresponds to the ISBN or ASIN code of the targeted product (book, DVD, CD, Game, etc.). The rank is then found on the page using an algorithm which processes character string. In this way, we constructed a database of the ranks of tens of books over several months with a sampling periodicity of one hour. This allowed us to check the quality of the time series of the ranks of thousands of books which have been recorded by **JungleScan** (http://www.junglescan.com) over several years. JungleScan scans books on Amazon typically every six hours for those updated hourly by Amazon. For all books we have checked, we found identical rank values in **JungleScan** and in our directly constructed database, showing that we could trust the data from **JungleScan** for our study.

B. Amazon ranking schemes

In order for this data to be useful, we need to convert the ranks into meaningful "physical" units, such as sales or sale rates. The problem is that Amazon.com does not divulge the exact formula for this conversion (otherwise, its secretive strategy on its sale figures would be useless). In our study, we use the time series of book ranks up to April 2004. Until October 2004, the following ranking system held, which is the relevant one for our study [7].

1. The official statement

Amazon gives some hints about its ranking system. The official statement of Amazon is the following.

What sales rank means

As an added service for customers, authors, publishers, artists, labels, and studios, we show how items in our catalog are selling. This bestseller list is much like The New York Times Bestsellers List, except it lists millions of items. The lower the number, the higher the sales for that particular title.

The calculation is based on Amazon.com sales and is updated regularly. The top 10 000 best sellers are updated each hour to reflect sales in the preceding 24 hours [8]. The next 100 000 ranks are updated daily. The rest of the list is updated monthly, based on *several different factors*.

2. Different time scales

Amazon can use different time scales as follows:

A short time scale: one day. That is the time scale we are interested in since our purpose is to study sales dynamics.

A long time scale. For instance, if Harry Potter sales were to crash overnight, its rank will not fall so sharply. It means that the whole history of a book can outweigh its instantaneous sales. It is worth understanding when the short time scale is outbalanced.

3. Evidence that Amazon uses different ranking schemes

Books with ranks in the range below 100 000 are reranked according to the sales during the last 24 hours. For these books, Amazon uses a short time scale to rank them. We will see that there are exceptions. For books with sales ranks over 100 000, Amazon does not explain the different factors used to update the ranks.

Figure 1 (top panel) shows an example in which the rank increased with large fluctuations up to 100 000 (the sales were steadily decreasing), then jumped to a few thousands and then followed a very smooth slowly increasing trajectory. The absence of fluctuations in the second rightmost part of the graph implies that Amazon reranked the books using its sales data using a long time scale. Figure 1 (bottom panel) shows a related effect. First, the sales fell off resulting in the rank increasing to 100 000. At this rank level, Amazon switched to a long time scale ranking scheme, leading to a reassessment of the rank in the range of 10 000. At some later time, some sales occurred which led Amazon to switch back to a 24 hours time scale ranking scheme, resulting in the rank increasing dramatically to a level around 10^5 but slightly below (so that the 24 hours ranking system is active).

To sum up, only books with a low rank (typically less than ten thousands) are ranked at a time scale from a few hours to a day and are devoid of the pathological behavior shown in Fig. 1. We will thus restrain our study to the books with ranks below 10 000.

C. Evidence that the rank-sales relation is close to a power law

We present the findings of Rosenthal [9]. For over six years, he has followed closely the sales ranks of his own books as both an author and a publisher. He has also used data points from other authors and publishers. Note that his analysis is unauthorized and in no way sponsored by Ama-

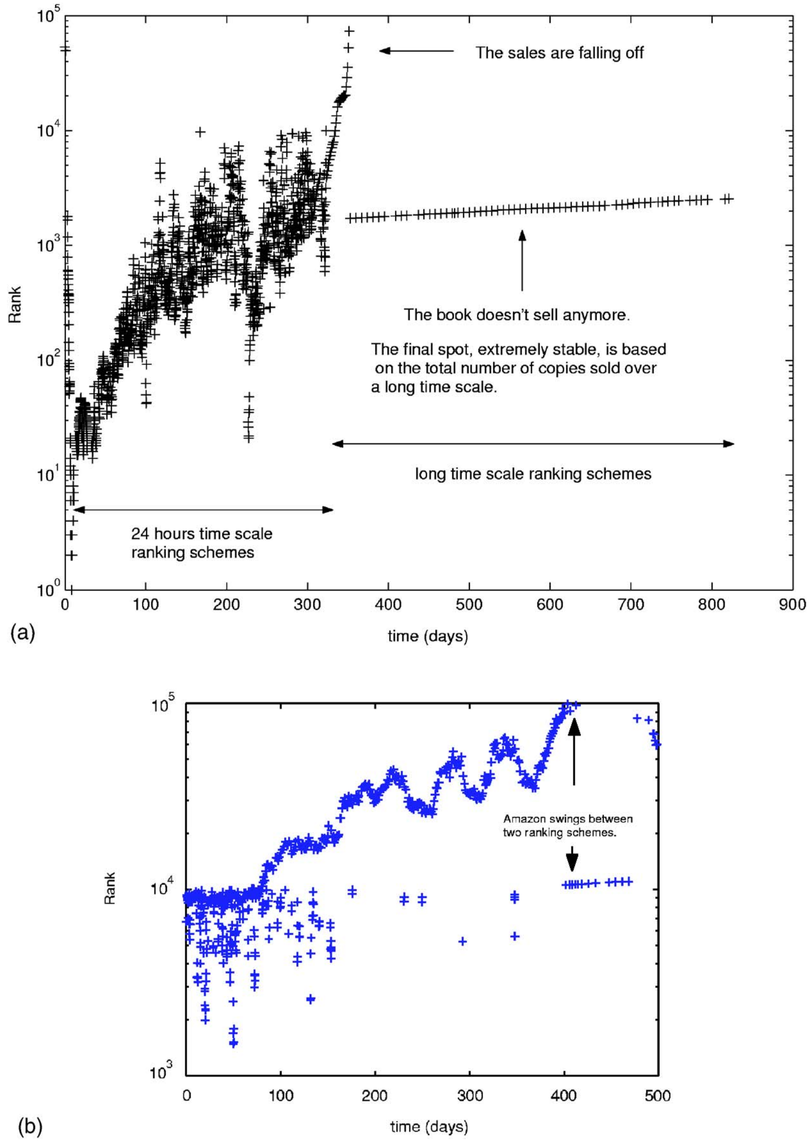

FIG. 1. (Color online) Two time series of ranks showing that Amazon switches between at least two ranking schemes, as explained in the text. The first book (top) is "See No Evil" by R. Baer. The second book (bottom) is "F'd Companies" by P. Kaplan.

zon, which keeps the sales-ranks conversion secret. His analysis allows us to get an approximate information on how to convert ranks into sales.

Table I shows the estimates of Rosenthal, which are converted into a rank-ordering or Zipf plot in Fig. 2. The power law dependence $S(R) \sim 1/R^{1/\mu}$ with exponent $\mu = 2.0 \pm 0.1$, as shown by the straight line with slope −0.5 in Fig. 2 translates into a standard power law of the complementary cumulative distribution of sales

$$R \sim 1/S^{\mu}, \quad \text{with } \mu = 2.0 \pm 0.1. \tag{1}$$

The bend for large ranks (small sales) is describing the bulk of the distribution of sales. The tail of the distribution of sales, i.e., for large sales, is described by the part of the figure corresponding to small ranks for which there does not appear to be a change of regime from a power law behavior, except for the fact that sales for the first ranks seem to fluctuation from book to book much more significantly that would be expected from a pure power law behavior. The current blockbuster with rank 1 may sell from hundreds of books per day to as much at tens of thousands of books per day. Such variations may perhaps be explained from the property that, while the typical fluctuations of the sales of the first rank is of the order of a few 100% [10], the distribution of the sales of the first rank is also a power law distribution of the form (1) with the same exponent, which means that it

TABLE I. Estimate of Rosenthal [9] on the rank-sale relationship. The symbol X reflects the unknown large fluctuations in the sales of the top book.

Rank	Copies/day
1	X
10	100
100	30
1000	10
10 000	2 (11 copies every 5 days)
100 000	1 copy a week
1000 000	around 15 total, depends on pub. date

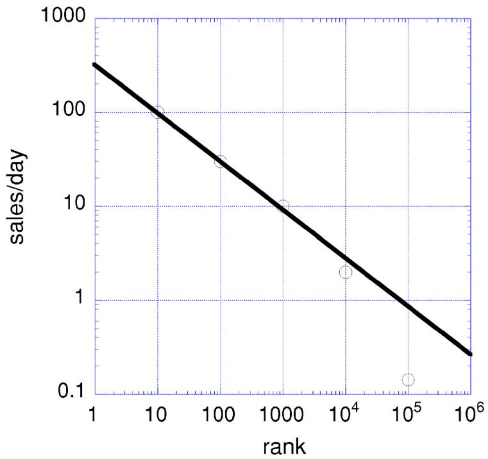

FIG. 2. (Color online) Rank ordering (Zipf) plot of the sales S per day, as a function of rank R. For R in the range from 10 to 10 000, the sales as a function of rank can be fitted with a good approximation by the power law $S(R) \sim 1/R^{1/\mu}$ with exponent $\mu = 2.0 \pm 0.1$, as shown by the straight line with slope -0.5. This translates into $R \sim 1/S^{\mu}$, which is proportional to the complementary cumulative distribution of sales. Note that the bend at large ranks (small sales) can be considered to be a finite size effect. The tail of the distribution of sales for large sales is described by the low ranks for which the distribution does not appear to bend but actually exhibits huge fluctuations as shown in Fig. 3 borrowed from [9].

is not impossible to have very large variations of the sales of rank 1, much larger than their typical values. These fluctuations are rendered in Fig. 3 in the left part for small ranks. More data would be needed to determine if the variations of the sales of the first ranks are explained by the power law distribution (1) or may perhaps reveal an amplification mechanism putting blockbusters apart. It is also possible that the increased fluctuations for the smallest ranks reflect the effect of the network structure of acquaintances. In contrast, the fluctuations of sales from rank 10 are typically of the order of 30% in agreement with expectations derived from the power law (1).

The correspondence between ranks and sales suggested in Figs. 2 and 3 and captured by the phenomenological formula (1) allows us to convert the time series of ranks for all studied books into time series of sales. Notwithstanding the possible uncertainties and errors in the calibration of the conversion from ranks to sales, as discussed above, one should consider this conversion as basically a convenient way to give a quantitative interpretation to the rank time series.

In summary, we will assume that the pdf $p(S)$ of book sales is a power law of the form

$$p(S) = \frac{C}{S^{1+\mu}}, \tag{2}$$

with $\mu = 2$. In Sec. V, we revisit this assumption and test it further from constraints on the averages of rank increments.

III. DESCRIPTION OF THE MODEL

A. Motivation: exogenous versus endogenous shocks

Consider the two sales time series shown in Fig. 4, exemplifying two characteristic patterns.

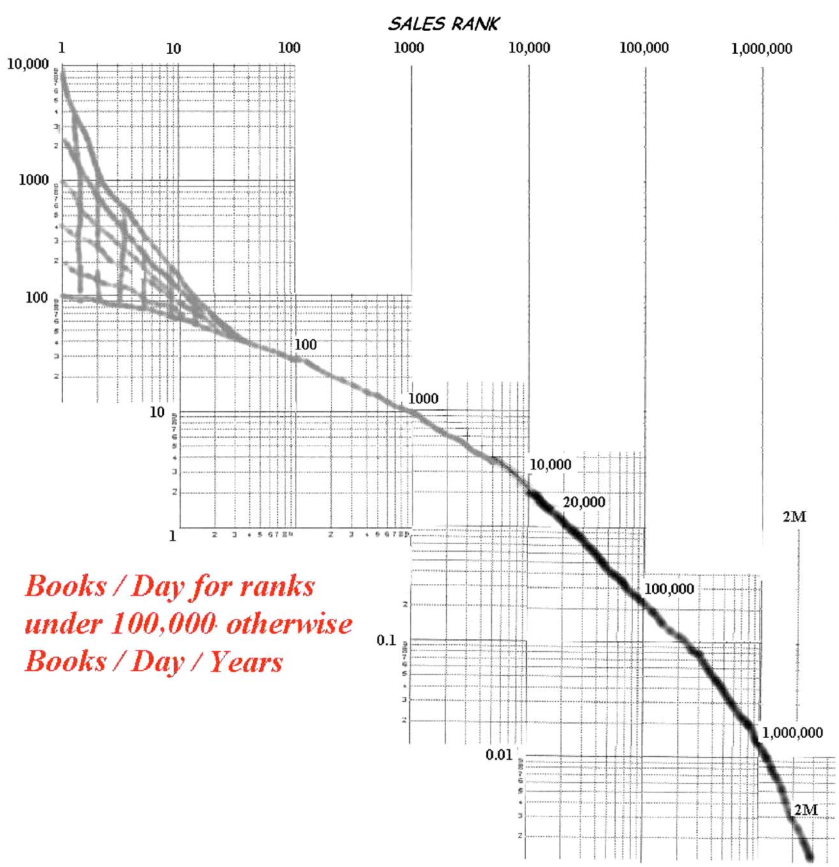

FIG. 3. (Color online) Rank ordering (Zipf plot) as in Fig. 2 presented in [9], extended to the smallest as well as largest ranks. Reproduced with kind permission from M. Rosenthal.

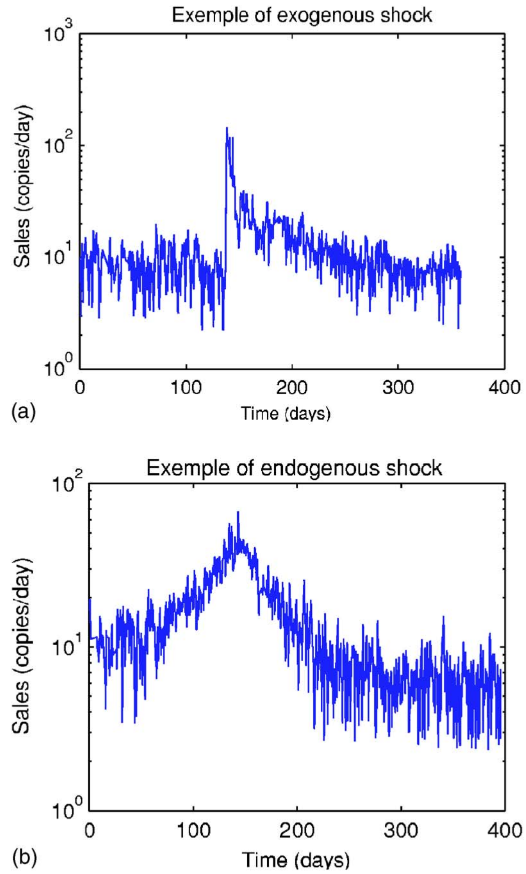

FIG. 4. (Color online) Time evolution over a year of the sales per day of two books: Book A (top) is "Strong Women Stay Young" by Dr. M. Nelson and book B (bottom) is "Heaven and Earth (Three Sisters Island Trilogy)" by N. Roberts. The difference in the patterns is striking, book A (B) exhibiting an exogenous (endogenous) peak.

Some books become best-sellers overnight, thanks to rocketing sales. Book A ("Strong Women Stay Young" by Dr. M. Nelson) jumped on June 5, 2002 from rank in the 2000s to rank 6 in less than 12 hours. On June 4, 2002, the New York Times published an article crediting the "groundbreaking research done by Dr. Miriam Nelson" and advising the female reader, interested in having a youthful postmenopausal body, to buy the book and consult it directly [11]. This case is the archetype of what we will refer to as an "exogenous" shock.

Some books become best-sellers after a long and steady increase in their sales. Book B ["Heaven and Earth (Three Sisters Trilogy)" by N. Roberts] culminated at the end of June 2002 after a slow and continuous growth, with no such newspaper article, followed by a similar almost symmetrical decrease, the entire process taking about 4 months.

B. Epidemic branching process with long-range memory

Such social epidemic process can be captured by the following simple model [12]. The model is based on the idea that the instantaneous sales flux of a given book results from a combination of external forces such as news, advertisement, selling campaign, and of social influences in which each past reader may impregnate other potential readers in her network of acquaintances with the desire to buy the book. This impact of a reader onto other readers is not instantaneous as people react at a variety of time scales. The time delays capture the time interval between social encounters, the maturation of the decision process which can be influenced by mood, sentiments, and many other factors and the availability and capacity to implement the decision. We postulate that this latency can be described by a memory kernel $\phi(t-t_i)$ giving the probability that a buy initiated at time t_i leads to another buy at a later time t by another person in direct contact with the first buying individual. We consider the memory function $\phi(t-t_i)$ as a fundamental macroscopic description of how long it takes for a human to be triggered into action, following the interaction with an already active human. $\phi(t)$ is normalized such that $\int_0^\infty \phi(t)=1$. Starting from an initial buyer (the "mother" buyer) who notices the book (either from exogenous news or by chance), she may trigger buying by first-generation "daughters," who themselves propagate the buying drive to their own friends, who become second-generation buyers, and so on. We describe the sum of all buys by a conditional Poisson branching process with intensity:

$$\lambda(t) = \eta(t) + \sum_{i|t_i \leq t} \mu_i \phi(t - t_i), \tag{3}$$

where $\eta(t)$ is the rate of buys initiated spontaneously (for instance by listening to a media coverage of a book or serendipity) without influence from other previous buyers and the mark μ_i is the number of potential buyers influenced by the buyer i who bought earlier at time t_i.

Our model is based on the key idea that the epidemic mechanism is basically the same for all books. Of course, the underlying networks of connected potential buyers are *a priori* not the same for different books. This can be accounted for by different values of the "branching ratio" as defined below.

While this version of the epidemic model of sales treats each book independently, in reality, we should consider correlations between sales of different books which may be related by a common growth of interest (see for instance the case of books on financial markets whose sales grow concomitantly during stock market bubbles [13]). In addition, sales will exhibit some correlation at special epochs, such as Christmas. At this period of the year, all books that have a gift appeal will sell more copies than they would have sold otherwise. In contrast, if a book does not have any gift appeal, its sales ranks will fall between Thanksgiving and Christmas, even if its actual sales remain steady. Likewise, university students buy a huge number of textbooks and other required reading titles through Amazon during September and from mid-January through mid-February, which will depress the ranks of books that do not fall into this category. They even vary in a regular manner during the course of the week. Some titles are primarily purchased by people at work or homemakers when their children are at school, while books with strong Associates support do relatively well on weekends. We will not take into account such effects.

We do not specifically describe the lifetime of a book and treat the innovations η's as stationary. Assuming sales as stationary allows us to define the probability that the sales reach a given value. This approximation should not be too bad over the time scales of months of our study but fails over longer time scales.

C. Mean field solution

Equation (3) can be written

$$\lambda(t) = \eta(t) + \int^{t} \int N[d\tau \times d\tau]\phi(t-\tau), \quad (4)$$

where $N[d\tau\times d\mu]$ is the standard notation for the number of events that occurred between τ and $\tau+d\tau$ with mark between μ and $\mu+d\mu$. In the physicist's notation $\int^t \int N[d\tau\times d\mu] = \int^t \int \delta(\tau-t_i)\delta(\mu-\mu_i)d\tau d\mu$. The lower bound of the integral over time in (4) is for instance the edition time of the considered book. Taking the ensemble average of (3) gives

$$S(t) \equiv \langle\lambda(t)\rangle = \eta(t) + n\int_{-\infty}^{t} d\tau\, \phi(t-\tau)S(\tau), \quad (5)$$

where the so-called "branching ratio" $n=\langle\mu\rangle$ is the average number of buys triggered by any "mother" within her acquaintance network. We have use the fact that $\langle \int N[d\tau \times d\mu]\rangle = n\langle\lambda(\tau)\rangle$. The branching ratio n depends on the network topology as well on the social behavior of influences. We consider only the subcritical regime $n<1$ in order to ensure stationarity. The linear structure of equation (5) does not mean that the dynamic is linear. It is an effective coarse-grained description of the complex nonlinear dynamics.

In order to solve $S(t)$, it is convenient to introduce the Green function or "dressed propagator" $\kappa(t)$ defined as the solution of (5) for the case where the source term $\eta(t)$ is a delta function centered at the origin of time,

$$\kappa(t) = \delta(t) + n\int_{-\infty}^{t} d\tau\, \phi(t-\tau)\kappa(\tau), \quad (6)$$

and by definition of $\kappa(t)$,

$$S(t) = \int_{-\infty}^{t} d\tau\, \eta(\tau)\kappa(t-\tau). \quad (7)$$

The cumulative effect of all the possible branching paths gives rise to the net sales flux $\kappa(t)$ triggered by the initial event at time $t=0$. The response function $\kappa(t)$ can easily be obtained by taking the Laplace transform of (6):

$$\hat{\kappa}(\beta) = \frac{1}{1-n\hat{\phi}(\beta)}. \quad (8)$$

Setting $\beta=0$, we get

$$\int_0^{\infty} d\tau\, \kappa(\tau) = \frac{1}{1-n}, \quad (9)$$

which means that $1/(1-n)$ is the average number of buyers influenced by one buyer through any possible lineage. This result can be recovered directly from the following argument: if n is the average number of buyers influenced directly by one buyer, the total number of buyers influenced through any possible lineage is $\Sigma_{k=0}^{\infty} n^k = 1/(1-n)$. The term $[\kappa(t)/(1-n)]dt$ is the probability that a purchase triggered by a buy at $t=0$ occurs at time t within dt.

We consider the case where the "bare propagator" is $\phi(t) \sim 1/t^{1+\theta}$ with $0<\theta<1$ corresponding to a long memory process. It leads to

$$\kappa(t) \sim 1/t^{1-\theta}, \quad \text{for } t < t^{\star}, \quad (10)$$

$$\kappa(t) \sim 1/t^{1+\theta}, \quad \text{for } t > t^{\star}, \quad (11)$$

with

$$t^{\star} \propto 1/(1-n)^{1/\theta}. \quad (12)$$

D. Prediction of the model

1. Distribution of sales

Starting from the evidence that the distribution $p(S)$ of sales is a power law (2), the linear expression (7) implies that the source terms $\eta(\tau)$ in (7) are also distributed as a power law with the same exponent μ. Actually, the generalized central limit theorem applied to (7) implies that the pdf $p(S)$ is a stable Lévy law L_μ with index μ as soon as the source terms $\eta(\tau)$ in (7) are independently and identically distributed as a power law with an exponent $\mu \leqslant 2$. By the generalized central limit theorem, the characteristic function of S is given by

$$\langle e^{iuS(t)}\rangle = \exp\left[-D|u|^{\mu}\int_0^{\infty} |\kappa(\tau)|^{\mu} d\tau\right], \quad (13)$$

where D is a measure of the magnitude of the sources $\eta(\tau)$. This translates into

$$P(S) \approx L_{\mu}\left(\frac{S}{\left[D\int_0^{\infty}|\kappa(\tau)|^{\mu}d\tau\right]^{1/\mu}}\right). \quad (14)$$

In reality, there is probably a dependence between the source terms $\eta(\tau)$. However, as long as their correlation function decays faster than $1/t$, this has only the effect of changing the coefficient D. We do not explore the situation in which the $\eta(\tau)$'s could have longer range correlations.

2. Dynamical properties

(a) Exogenous shock. An external shock occurring at $t=0$ can be modeled as a jump $S_0\delta(\tau)$. The response of the system for $t>0$ is then

$$S(t) = S_0\kappa(t) + \int_{-\infty}^{t} \eta(\tau)\kappa(t-\tau).$$

The expectation of the response to an exogenous shock is thus

$$E[S(t)] = S_0\kappa(t) + \frac{1}{1-n}\langle\eta\rangle, \tag{15}$$

where $\langle\eta\rangle$ is the average source level. Expression (15) simply expresses that the recovery of the system to an external shock is entirely controlled by its relaxation kernel.

For an external shock which is strong enough, for $t>0$:

$$E_{exo}(S(t)) \approx S_0\kappa(t). \tag{16}$$

It exemplifies that $\kappa(t)$ is the Green function of the coarse-grained equation of motions of the system.

(b) Endogenous shock. Consider a realization which exhibits a large sales burst $S(t=0)=S_0$ without any large external shock. In this case, a large endogenous shock requires a special set of realizations of the noise $\{\eta(t)\}$ [12]. We can write $\eta=\tilde{\eta}+\langle\eta\rangle$. By construction $\tilde{\eta}$ has a zero mean. Using this, we get

$$S(t) = \int_{-\infty}^{0} d\tau\,\tilde{\eta}(\tau)\kappa(t-\tau) + \int_{-\infty}^{0} d\tau\langle\eta\rangle\kappa(t-\tau) + \int_{0}^{t} d\tau\,\eta(\tau)\kappa(t-\tau).$$

The expectation of S is

$$E[S(t)] = \int_{-\infty}^{0} d\tau\,E[\tilde{\eta}(\tau)]\kappa(t-\tau) + \frac{\langle\eta\rangle}{1-n}. \tag{17}$$

As for an exogenous shock, the constant $\langle\eta\rangle/(1-n)$, equal to the unconditional average $\langle S\rangle$, can be neglected.

For $\tau<0$, the expectation of $\tilde{\eta}(\tau)$ is not zero, because the value $S(0)=S_0$ is specified. In contrast, for $\tau>0$, $E[\eta(\tau)]=\langle\eta\rangle$ since the conditioning does not operate after the shock. Consider the process $W(t)\equiv\int_{-\infty}^{t}d\tau\,\tilde{\eta}(\tau)$. A standard result is that for $t<0$:

$$E[W(t)|S(0)=S_0] = (S_0 - E[S])\frac{\mathrm{cov}[W(t),S_0]}{\mathrm{var}[S_0]} \simeq (S_0 - E[S])\int_{-\infty}^{t} d\tau\,\kappa(-\tau).$$

This expression predicts that the expected path of the continuous innovation flow prior to the endogenous shock (i.e., for $t<0$) grows like $\Delta W(t)\sim\kappa(-t)\Delta t$ upon the approach to the time $t=0$ of the large endogenous shock. In other words, conditioned on the observation of a large endogenous shock, there are specific sets of the innovation flow that led to it. These conditional innovation flows have an expectation $E[\eta(t<0)]-\langle\eta\rangle\simeq S_0\kappa(-t)$ (We assume $S_0\gg E[S]$). We thus obtain from (17) for $t>0$ and $t<0$:

$$E_{endo}[S(t)] \propto S_0\int_{-\infty}^{\min[t,0]} d\tau\,\kappa(t-\tau)\kappa(\tau). \tag{18}$$

(c) Distinguishing both shocks. The model predicts two different relaxations for exogenous and endogenous shocks according to expressions (16) and (18). Assuming that we are close to the critical point $n\approx 1$, we can use $\kappa(\tau)\sim 1/t^{1-\theta}$.

TABLE II. Aftershock and foreshock signatures for endogenous and exogenous shocks occurring at time $t=0$.

	Endo	Exo
Aftershock	$S(t)\propto 1/t^{1-2\theta}$	$S(t)\propto 1/t^{1-\theta}$
Foreshock	$S(t)\propto 1/\lvert t\rvert^{1-2\theta}$	Abrupt peak

Table II gathers the aftershock and foreshock signatures.

The prediction that the relaxation following an exogenous shock should happen faster (exponent $1-\theta$) than for an endogenous shock (exponent $1-2\theta$) agrees with the intuition that an endogenous shock should have impregnated the network much more and thus have a longer lived influence. This result is a nontrivial consequence of our model. If we perform the same calculation [12] for an exponential decaying memory kernel $\phi(t)$, the functional form of the recovery does not allow one to distinguish between an endogenous and an exogenous shock. For the memory kernel $\phi(t)$ decaying faster than an exponential, the endogenous relaxation turns out to be faster than the exogenous one [12].

IV. EMPIRICAL DETERMINATION OF THE EXPONENTS OF SALES DYNAMICS

A. Selection of peaks and fitting the power law

We qualify a peak as a local maximum over a 3-month time window which is at least $k=2.5$ time larger than the average of the time series over the same 3-month time window. The threshold value k was determined after looking at several examples of time series. The results do not change significantly by varying k within large bounds (see below).

We considered only the sales maxima corresponding to ranks reaching the top 50.

We selected those time series which had at least 15 days after the peak, so that we can analyze the recovery signature following shocks.

We fit the sales dynamics by a power law,

$$S(t) \sim \frac{A}{(t-t_c)^p},$$

with A, p and t_c unknown. The "critical time" t_c is expected to be close to the time t_0 of the peak. We know from previous experience in critical phenomena that the determination of the exponent p can be quite sensitive to the fitted value of t_c.

Moreover, we do not know *a priori* which window size should be taken. If the signal were a pure power law, it would not matter. But, here the signal is approximately a power law over a finite range. A big time window means more data and so more accurate results but we have to end the window before a possible change of regime due to (i) $n<1$ implying the existence of the finite crossover time $t^\star$ given by (12), (ii) $S(t)$ tending to background noise around to its average value, and (iii) the impact of other shocks that may interfere.

In order to determine A, p and t_c for a given window size, we use a least square method, i.e., we seek to minimize the following quantity:

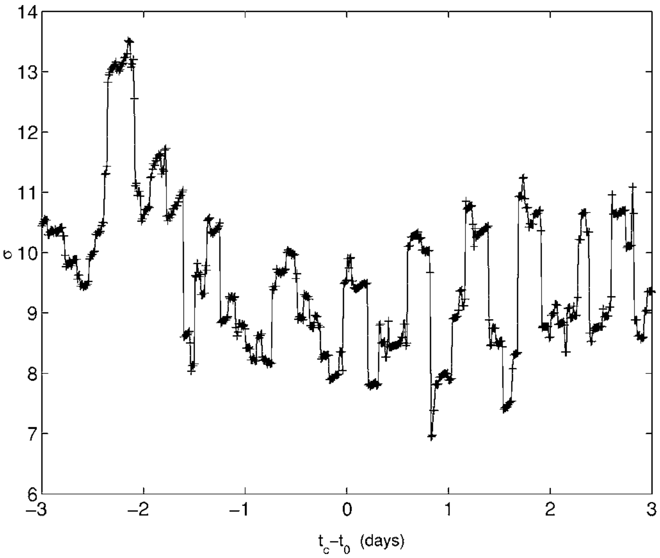

FIG. 5. $\sigma(t_c)$ defined in (19) as a function of t_c.

$$\sigma_{logA,p,t_c} = \sum_{t_i} F_{logA,p,t_c}(t_i)^2,$$

with

$$F_{logA,p,t_c}(t_i) = \log S(t_i) - \log A + p \log(t_i - t_c).$$

As σ is quadratic with respect to $\log A$ and p, for these two parameters, the minimization of σ is straightforward and can be done analytically. Setting the partial derivatives $\partial\sigma/\partial \log A$ and $\partial\sigma/\partial p$ equal to zero, we just need to solve a linear system of two equations with two unknowns, which gives $\log A(t_c)$ and p_{t_c} as a function of the still unknown t_c. Now, we just need to minimize

$$\sigma_{t_c} = \sigma_{logA(t_c),p(t_c),t_c}, \tag{19}$$

a function of one variable. Figure 5 shows an example of $\sigma(t_c)$. To minimize such irregular function, we scan over all the value of t_c. We typically use a total interval of one week around the time of the peak.

We perform a fit for different time windows ranging from a lower bound I_{min} to an upper bound I_{max}. I_{min} has a fixed value, set to 15 days. In contrast, I_{max} depends on the time series. I_{max} is calculated as the time at which the minimum of the sales occurs, over a time window running from 25 days to 6 months. I_{max} is fixed in this way to prevent a rise in sales from being taken into account in the fit.

Once the fit has been performed for the different time windows, we look at the correlation coefficients of the fits and choose the window which leads to the best correlation coefficient. This method turns out to give a robust estimate of the power law decay of the relaxation of sales.

B. Distribution of exponents

Out of some 14 000 books available on Junglescan on April 2004, our algorithm detects 1013 peaks which obey the constraints of reaching the top 50, with sufficient data before and after the peak and obeying the condition of not being contaminated by a closeby peak, as specified above. Among these 1013 peaks, we select those followed by a relaxation which can be well approximated by a power law, with the criterion that the correlation coefficient r of the corresponding fit is larger than 0.95. This leads us to keep 138 peaks. Making this selection does not change qualitatively our results but improve somewhat the quantitative findings. We have played with different values of the correlation coefficient between 0.8 and close to 1 and find the same results with larger error bars for lower correlation coefficients.

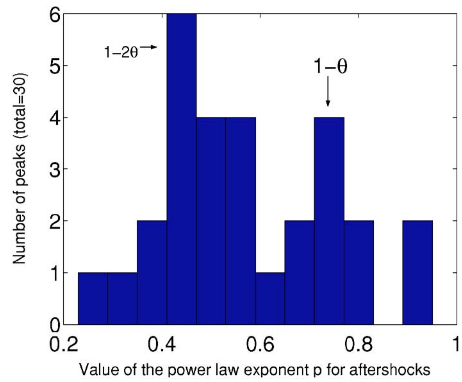

FIG. 6. (Color online) Histogram of the estimated power law exponents p of the relaxations of the sales. One can clearly identify two clusters: the endogenous cluster with exponent $1-2\theta$ close to 0.4 and the exogenous cluster with exponent $1-\theta$ close to 0.7, compatible with the estimation $\theta \approx 0.3$. The peaks shown here are those used in Fig. 7 (see Sec. IV C).

Figure 6 shows the distribution of power law exponents for the decrease of sales after peaks. We find two clusters corresponding to peaks respectively with an exponent $0.2 < p < 0.6$ and with an exponent $0.6 < p < 1$. This suggests that these two clusters can be seen as the endogenous cluster ($1-2\theta \approx 0.4$) and the exogenous one ($1-\theta \approx 0.7$). This provides a first estimate for the exponent $\theta \approx 0.3$.

According to the epidemic model proposed here, the small values of the exponents (close to $1-\theta$ and $1-2\theta$) for the exogenous and endogenous relaxations, respectively, imply that the sales dynamics is dominated by cascades involving high-order generations rather than by interactions stopping after first-generation buy triggering. Indeed, if buys were initiated mostly by the direct effects of news and advertisements without amplification by triggering cascades in the acquaintance network, the cascade model would predict an exponent $1+\theta$ given by the "bare" memory kernel $\phi(t)$. The values smaller than 1 for the two exponents for exogenous and endogenous shocks imply accordingly that the average number n (the average branching ratio in the language of branching models) of impregnated buyers per initial buyer in the social epidemic model is on average very close to its critical value 1, because the renormalization from $\phi(t)$ to $\kappa(t)$ given by (10) only operates close to the criticality characterized by the occurrence of large cascades of buys. Reciprocally, a value of the exponent p larger than 1 would suggest that the associated social network is far from critical.

C. Stacking the peaks

According to our model and as summarized in Table II, the peaks belonging to the cluster with high p ($p \approx 0.7$)

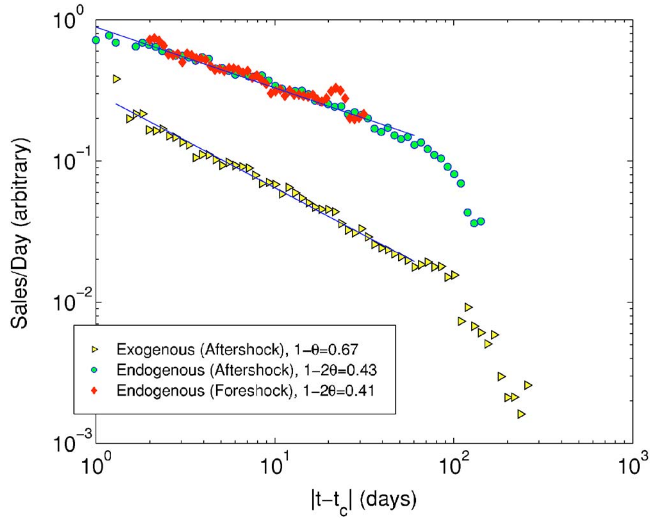

FIG. 7. (Color online) Relaxation after the peak (for both endogenous and exogenous cases) and precursory acceleration (for the endogenous case). We average (log average) over the peaks classified as endogenous and exogenous according to their precursory growth. The same sample of books was used for both Fig. 6 and Fig. 7.

should be in the exogenous class, and therefore should be reached by the occurrence of abrupt jumps without detectable precursory growth. Alternatively, the peaks belonging to the cluster with $p \approx 0.4$ should be in the endogenous class, and therefore should be associated with a progressive precursory power law growth $1/(t_c - t)^p$ with exponent $p = 1 - 2\theta$.

To check this prediction, the following algorithm categorizes the growth of sales before each of the peaks according to its acceleration pattern. We differentiate between peaks that have an increase in sales by a factor of at least k_{exo} prior to the peak and peaks that have an increase in sales by a factor of less than k_{endo} at the same time. More precisely, we compared the value of the sales at the time of the peak (D day) and the average value of the sales from day D-4 to D-1. We tried several values for these two coefficients k_{exo} and k_{endo}.

We find that the bigger k_{exo}, the largest is the exponent of the average relaxation for books that have an increase in sales by a factor more than k_{exo}. Conversely, we find that the smaller k_{endo}, the smaller is the exponent of the average relaxation for books that have an increase in sales by a factor less than k_{exo}. Both results agree with the intuition that, for selective criteria, we only keep peaks that are very easy to classify and thus reduce the probability to make a misclassification.

Finally, we set $k_{exo} = 30$, $k_{endo} = 2$. It means that peaks for which the acceleration factor were between 2 and 30, were not considered for the subsequent analysis leading to Fig. 7. Out of the 138 peaks, 30 remains.

Figure 7 shows the average relaxation and precursory acceleration. For shocks classified as exogenous according to their acceleration pattern, we find a relaxation governed by an exponent $1 - \theta \approx 0.7$. For shocks classified as endogenous, both aftershocks and foreshocks are controlled by an exponent $1 - 2\theta \approx 0.4$. These results match what has been predicted by the model (see Table II).

Aftershock. The best fit of the least square method with a power law gives a slope $1 - \theta \approx 0.7$ for exogenous shocks and a slope $1 - 2\theta \approx 0.4$ for endogenous shocks. One can observe a crossover for $t - t_c \approx 60 - 80$ days, in both cases. It is tempting to interpret this crossover as the change of regime predicted by the model for $t \approx t^\star$ (see Sec. III C). Indeed for $t > t^\star$, we should expect a power law with exponent $1 + \theta$, both for exogenous and endogenous shocks. Unfortunately, the crossover does not extend sufficiently far to allow us to constrain the exponent of the second regime.

Foreshock. The best fit of the least square method with a power law gives a slope $1 - \theta \approx 0.4$ for the endogenous foreshocks. The time on the x axis has been reversed to compare the precursory acceleration with the aftershock relaxation. The superposition of the two top curves for the precursory and relaxation behavior of the endogenous peaks confirms the symmetric behavior predicted by the model (see Table II).

D. Detailed analysis of exogenous peaks

Table III lists the 10 peaks that have an exponent larger than 0.65 among the thirty peaks used to make Fig. 6. Figure 9 shows the evolution of the sales for these 10 peaks, from ten days before to 70 days after the peaks. Figure 8 shows the time evolution of the sales of the book "Get with the program" written by the personal trainer of Ophra Winfrey and its remarkable succession of exogenous peaks associated with regular appearance of the book in Oprah's TV show.

In Fig. 9, eight peaks are reached by a fast acceleration of the sales, as expected from our model and the classification. However for two books, "Star Wars" and Stephen King's novel, the situation is different. Their acceleration patterns are classified as endogenous (i.e., slow acceleration growth) by our algorithm. Yet, they exhibit a fast relaxation (large exponent $p = 1 - \theta$) which means that they are classified as exogenous according to the first criterion, based on the exponent of the relaxation after the peak. This mismatch between the results of the two classifications can be explained as follows.

Consider the example of the book entitled "Stars Wars." Its success was triggered by the release of the movie. But, as the advertisement campaign lasted several months, rather than a sharp acceleration of sales, we observe a slow acceleration growth. For such huge selling campaign, modeling the source term $\eta(t)$ responsible for the shock by a Dirac function is a very poor approximation. Here, the time duration of the advertisement campaign cannot be neglected.

For Stephen King's novel, the situation looks quite the same. On September 2003, the Board of Directors of the National Book Foundation announced that its 2003 Medal for Distinguished Contribution to American Letters would be conferred to Stephen King. This is America's most prestigious literary prize. The ceremony took place two months later, on 11/19/2003, shortly after the observed peak. As for "Star Wars," the time extension of the exogenous impact of news cannot be neglected.

This finding suggests that modeling external influence—news, advertisement—by a Dirac function $\delta(t)$ is not ad-

TABLE III. List of the 10 peaks that have an exponent p more than 0.65 (see Fig. 6). The last column suggests a possible cause of the exogenous shock, as far as we have been able to tell.

Title	Author	Time	Rank	Reason
Get with the Program	Bob Greene	3/16/2002	1st	The book was seen on the Oprah Winfrey Show (B. Greene is O. Winfrey's trainer).
		10/19/2002	1st	
		1/4/2003	1st	
Get with the Program Daily Journal	Bob Greene	10/19/2002	2nd	The book was seen on the Oprah Winfrey Show (B. Greene is O. Winfrey's trainer).
Adios Muchachos	Daniel Chavarría	7/5/2002	6th	On 7/4/2002, the book was on the first page of the Art Section of the New York Times
Sacred Contracts: Awakening Your Divine Potential	Caroline Myss	8/16/2002	1st	?
Micawber's Museum of Art	John Lithgow	9/30/2002	3rd	?
Refrigerator Rights: Creating Connections and Restoring Relationships	Dr. Will Miller	1/26/2003	12nd	The author has appeared on a variety of national television programs
Stars Wars: Episode II, Attack of the Clones	RA Salvatore	5/20/2002	13rd	The movie was released on 5/16/2002.
Wolves of the Calla (The Dark Tower, Book 5)	Stephen King	11/11/2003	3rd	On 11/19, the 2003 Medal for Distinguished Contribution to American Letters was conferred upon Stephen King.

equate in some cases. The external shock may have a significant duration T. We then expect the sales to grow slowly and exhibit a plateau over a time scale proportional to T and then to crossover to the exogenous $1/t^{1-\theta}$ decay rate for times $t > T$. Figure 10 shows that this is indeed the case for the two books "Star Wars" and "Wolves of the Calla" which were exceptions to our classification in terms of their acceleration pattern: the long durations of the peaks are consistent with the fact that the external news impacted over an extended period of time, leading to our misclassification as endogenous with respect to their acceleration before the peak. The time scale of about 20 days of the plateau is consistent with the known duration of the external news.

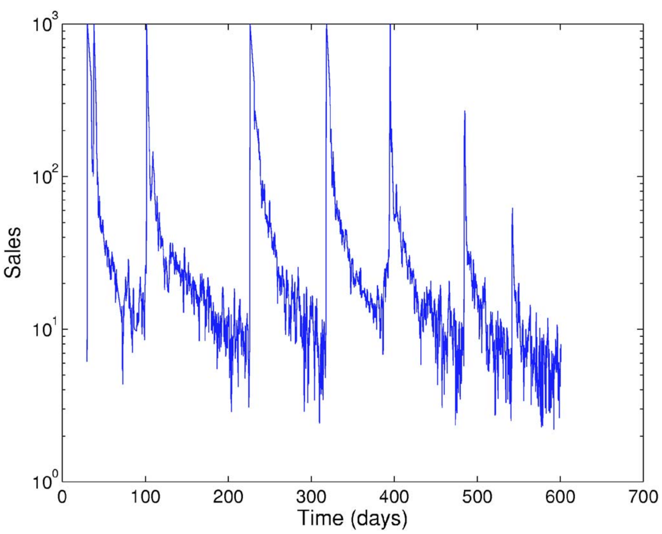

FIG. 8. (Color online) Time evolution of the book "Get with the Program" (see Table III). Each time the book is presented in the Oprah Winfrey Show, the sales jump overnight and then relax according to the exogenous response function $\kappa(t) \sim 1/t^{1-\theta}$.

These two examples show the need to refine our analysis to allow for a more general description of external news. In particular, the epidemic model can in principle be used to invert the amplitude of the flow of news $\eta(\tau)$ from the time series of the sales $S(t)$. But to be effective and reliable, this would require better data, for instance directly working on Amazon sales rather than on reconstructed sales from ranks.

E. Robustness of the results upon variation of the conditions of peak selection

Previously, we kept 30 peaks corresponding to those which give a fit with a power law with a correlation coefficient larger than 0.95 and which have an increase before the peak of a factor less than $k_{endo}=2$ or of a factor more than $k_{exo}=30$. It is interesting to discuss the results for less restrictive conditions.

Figure 11 is similar to Fig. 6 and shows the distribution of power law exponents for the decrease of sales after peaks for those peaks whose relaxation can be fitted by a power law

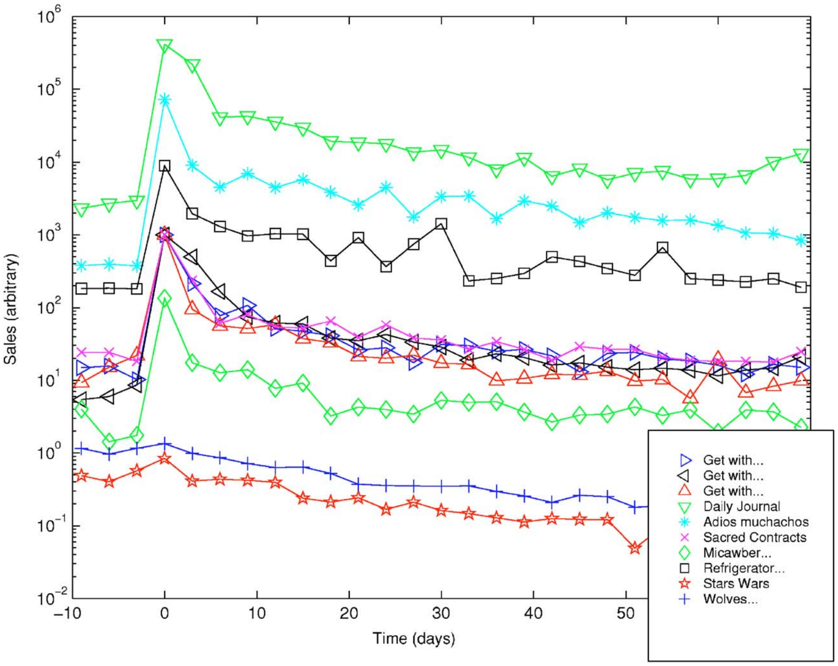

FIG. 9. (Color online) Time evolution of the sales of the 10 books of Table III. For each time series, $t=0$ is the time of the peak. The sales jumped just before the peak except for "Stars Wars" and "Wolves of the Calla."

with a correlation coefficient larger than 0.9. This less restrictive condition on the quality of the power law fit selects 388 peaks out of the initial 1013 peaks of our prefiltered data set, i.e., close to three times more peaks. We again find two rather clearly defined clusters, which can be associated with the endogenous and the exogenous classes as previously.

Among the 388 peaks that have a correlation coefficient more than 0.9, we kept the 270 peaks which have an increase before the peak less than $k_{endo}=8.5$ or more than $k_{exo}=12.5$. These two coefficients were selected empirically, such that the algorithm make a distinction between slow and fast acceleration similar to what common sense would suggest. The $388-270=118$ rejected peaks correspond to blurred situations where the acceleration of the sales is neither fast nor progressive. Figure 12 is similar to Fig. 7 but for these 270 peaks. We again obtain two different power law slopes, a faster decrease and larger exponent for the class classified as exogenous with respect to its fast foreshock acceleration than for the class classified as endogenous with respect to its progressive foreshock acceleration. For $(k_{endo},k_{exo})=(8.5,12.5)$, we obtain $p=0.54$ for the exogenous class and $p=0.4$ for the endogenous class. To test the sensitivity with respect to the coefficients k_{endo} and k_{exo}, we report other values. For $(k_{endo},k_{exo})=(5,20)$ [respectively (2,30)] which are not shown, we obtain a mean exponent for the exogenous class equal to 0.55 (0.57) and 0.39 (0.39) for the endogenous class. While the results are qualitatively consistent with those obtained with more stringent conditions, we see that the expo-

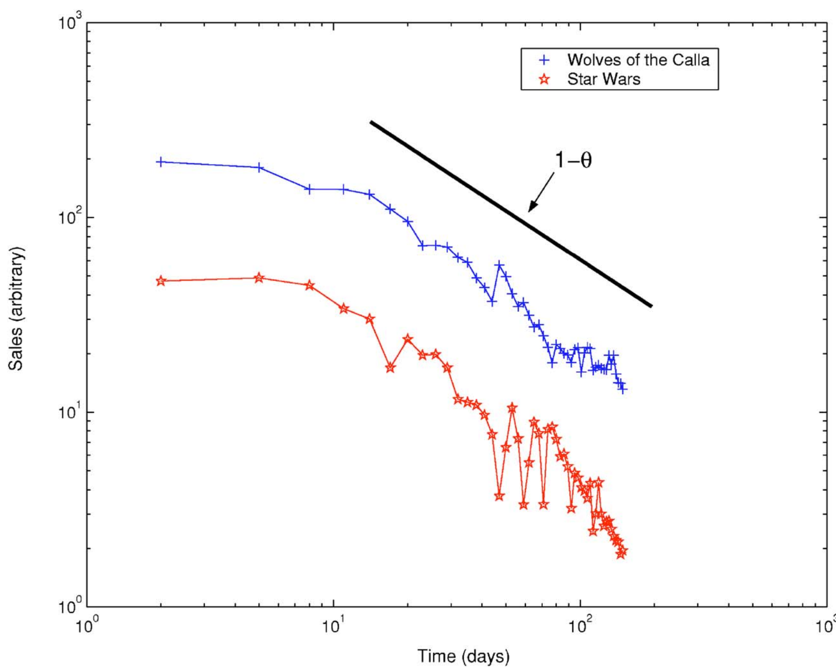

FIG. 10. (Color online) Relaxation of the sales after the peak for "Star Wars" and "Wolves of the Calla." The expected power law behavior for exogenous shocks can only be observed for $t>10-20$ days. The small value of the slope for $t<10-20$ days can be explained by considering the time duration of the external shock triggered by the media as explained in the text.

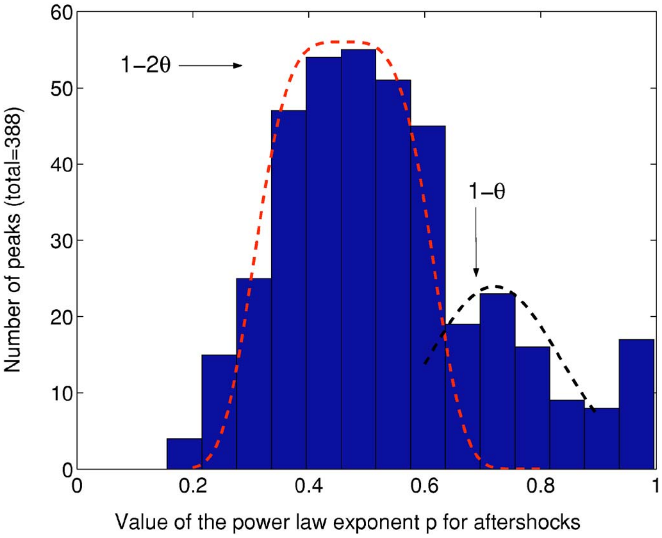

FIG. 11. (Color online) Histogram of the estimates of the power law exponents p of the relaxations of the sales using a selection criterion different from that of Fig. 6: among the 1013 peaks of our prefiltered database, we kept those which have a correlation coefficient larger than 0.9, giving 388 peaks.

nent for the exogenous class is a bit too small. This may be due to the duration of the news which are not instantaneous as discussed above and to the existence of other factors and disturbances not described here.

V. FURTHER TESTS AND CONSTRAINTS ON THE RANK-SALE CONVERSION (2)

In this section, we compare the prediction of the model with the data on changes of ranks to test the validity of the power law distribution (2) that we used to convert ranks into sales.

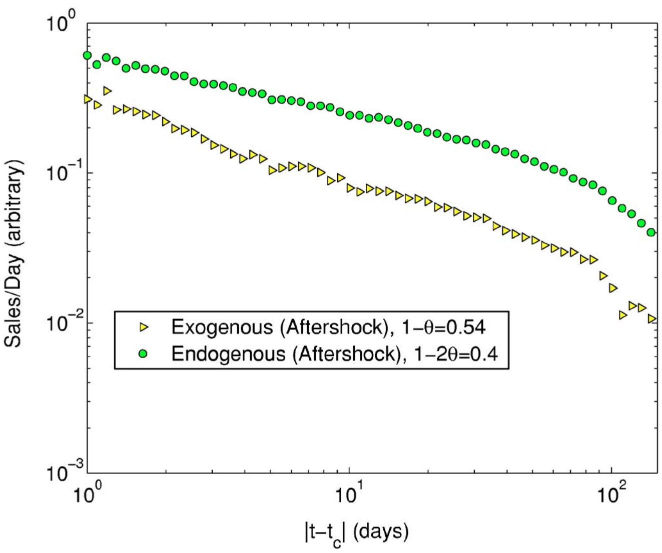

FIG. 12. (Color online) Relaxation of sales after the peak following the same methodology as for Fig. 7 but with the same set as in Fig. 11 with the clustering procedure explained in the text.

Let us denote $\Delta R(t) \equiv R(t+\Delta t) - R(t)$ the variation of rank over the time interval Δt (which will be taken fixed and equal to 1 day). Let us call $\Delta S(t) \equiv S(t+\Delta t) - S(t)$, the variation of sales over the same time interval. It is clear that variations of ranks must be interpreted relative to past ranks. A change of a few ranks when a book is selling at rank 10 is not the same as when it has rank 10 000. This motivates us to study the average of the conditional rank variation defined by

$$\langle \Delta R(R) \rangle \equiv E[\Delta R(t) | R(t)]. \tag{20}$$

$\Delta R(t) | R(t)$ is defined as the rank variation from time t to $t+1$ for a book which was at rank $R(t)$ at time t. Similarly, the average of the conditional sale variation

$$\langle \Delta S(S) \rangle \equiv E[\Delta S(t) | S(t)] \tag{21}$$

is the expectation of the variation of the sales conditioned on its value before.

First, we will derive $\langle \Delta S(S) \rangle$ from our model. Then, using the postulated rank-sales conversion given by (2) will provide a prediction for $\langle \Delta R(R) \rangle$. We will also measure $\langle \Delta R(R) \rangle$ directly without using the conversion rank-sales, and compare to test the ranks-sales conversion.

A. Prediction of the model of Sec. III

1. Derivation of $\langle \Delta S(S) \rangle$

Let us define

$$\tilde{S}(t) \equiv \int_{-\infty}^{t} d\tau \, \tilde{\eta}(\tau) \kappa(t-\tau), \tag{22}$$

with $\tilde{\eta} \equiv \eta - \langle \eta \rangle$. Then, $S(t) = \tilde{S}(t) + \langle \eta \rangle / (1-n)$ and $\Delta S(t) = \Delta \tilde{S}(t)$. Let us expand $\Delta S(t)$:

$$\Delta S(t) = \Delta \tilde{S}(t) = \int_{-\infty}^{t} d\tau \, \tilde{\eta}(\tau)(\kappa(t-\tau+\Delta t) - \kappa(t-\tau)) + \int_{t}^{t+\Delta t} d\tau \, \tilde{\eta}(\tau) \kappa(t+\Delta t - \tau).$$

Conditioned on the value of $S(t)$, the expectation of the second term of the right-hand side (RHS) is zero because the conditioning does not affect times posterior to t. This implies

$$\langle \Delta S(S) \rangle = \int_{-\infty}^{t} d\tau \, E[\tilde{\eta}(\tau) | S(t)](\kappa(t-\tau+\Delta t) - \kappa(t-\tau)). \tag{23}$$

Following the same reasoning as in Sec. III D 2, we have $\forall S(t)$, $E[\tilde{\eta}(\tau) | S(t)] \propto \tilde{S}(t) \kappa(t-\tau)$. Replacing this expression in (23) obtains

$$\langle \Delta S(S) \rangle = -\alpha \tilde{S} \tag{24}$$

$$= -\alpha(S - \langle S \rangle), \tag{25}$$

with

$$\alpha \simeq -\int_{-\infty}^{t} d\tau(\kappa(t-\tau+\Delta t)-\kappa(t-\tau)), \tag{26}$$

which is positive because κ is a decreasing function.

2. Derivation of $\langle \Delta R(R)\rangle$ from $\langle \Delta S(S)\rangle$ assuming (2) to be valid

Expression (2) implies that the function $R(S)$ is a power law of exponent $-\mu$. Thus, taking the logarithmic derivative expressed as a finite difference gives

$$\frac{\Delta R}{R} = -\mu\frac{\Delta S}{S}. \tag{27}$$

From (25) and (27), we get

$$\frac{\langle \Delta R(R)\rangle}{R} = \mu\alpha\left(1-\frac{\langle S\rangle}{S}\right). \tag{28}$$

S can be expressed as a function of R as follows. Starting from (2), we express the constant C as $C=\mu S_{min}^{\mu}$ from the normalization of $p(S)$ in the interval from a minimum sale S_{min} to infinity. Then, $\langle S\rangle=\int p(S)\ S\,dS=[\mu/(1-\mu)]S_{min}$. We have assumed $\mu>1$, which implies that a majority of the sales are coming from books with large ranks (i.e., low sales) and not from the few blockbusters in the top ranks. This leads to

$$R(S) = NP_{>}(S) = N\left(\frac{\mu-1}{\mu}\right)^{\mu}\left(\frac{\langle S\rangle}{S}\right)^{\mu}, \tag{29}$$

where N is the "total" number of books. It is not very clear which value should be taken for N. Amazon sells and ranks several millions of books. But, should our power-law conversion be true, it will not probably be valid for the largest ranks in the million range. We can probably expect to have $N\approx 10^4$ because of Amazon's change of ranking scheme around $R=10^4$, which creates a naturally clustered population of the first 10 000 ranked books. Or perhaps, it could be 10^5.

Putting together (28) and (29) leads to

$$\langle \Delta R(R)\rangle = \alpha\mu R\left[1-\frac{\mu}{\mu-1}\left(\frac{R}{N}\right)^{1/\mu}\right]. \tag{30}$$

Our epidemic model of book buys together with the assumption of a power law distribution of sales with exponent μ, expression (30) predicts that the average variation of ranks is proportional to the rank itself for small R and to $-R^{1+1/\mu}$ for large rank. The nonmonotonicity of $\langle \Delta R(R)\rangle$, i.e., $\langle \Delta R(R)\rangle>0$ for small R and $\langle \Delta R(R)\rangle<0$ for large R simply reflects that best-sellers tend to lose ranks because being a best-seller requires to have continuously sources $\eta(t)$'s of buyers above the average, which can only be achieved for a relatively short time. Conversely, poorly ranked books can only improve their ranking on average.

B. Empirical test

Figure 13 shows the overall behavior of $\langle \Delta R(R)\rangle$. For our purpose, we will ignore ranks $R>10^4$ because, for such ranks, the Amazon ranking scheme is not appropriate as already discussed in Sec. II B and illustrated by the artifacts for $R\approx 10^4$ and $R\approx 10^5$. These spurious peaks reflect the shift of Amazon between different ranking schemes.

The part of Fig. 13 for $R<10^4$ is magnified in Fig. 14. The predicted nonmonotonous behavior is observed but expression (30) does not fit the experimental data for the small ranks. The log-log plot of Fig. 14 clearly shows that $\langle \Delta R\rangle$ is given by

$$\langle \Delta R\rangle \sim R^{\beta}, \quad \text{with } \beta = 1.5 \pm 0.05, \tag{31}$$

over approximately three decades. Thus, $\langle \Delta R\rangle$ is not proportional to R since β is significantly larger than 1. This depar-

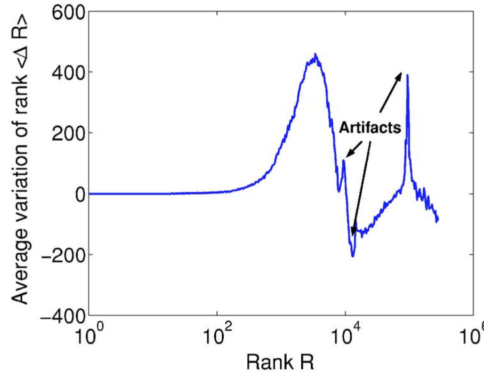

FIG. 13. (Color online) Average variation $\langle \Delta R\rangle$ of the rank as a function of the rank itself. This measure was performed over a sample of 14 000 books. One can note spurious behaviors for ranks close to 10^4 and 10^5. This can be rationalized by a shift in ranking schemes for these value (see Sec. II B 1). The data can only be exploited for $R<10^4$ to avoid these artifacts. The most striking feature is the nonmonotonous behavior of $\langle \Delta R\rangle$ for $R<10^4$. This enables us to reject an exponential conversion (see Sec. V B 2).

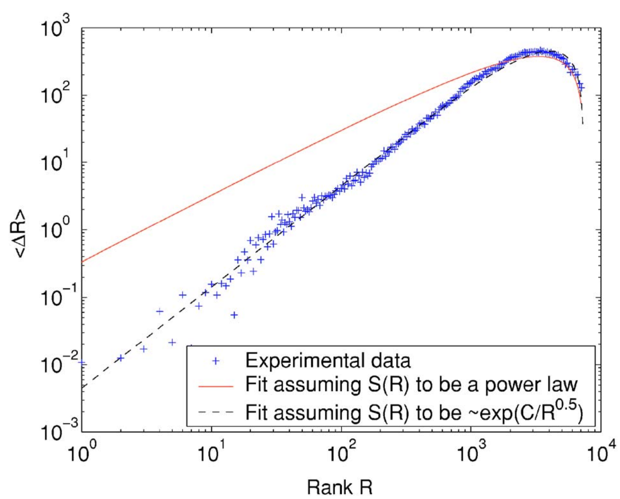

FIG. 14. (Color online) Average variation of the rank $\langle \Delta R\rangle$ as a function of the rank itself. We only take into account those time series that do not exhibit abrupt peaks (i.e., we discard exogenous shocks). *Plus*: experimental data; *solid line*: fit assuming $S(R)$ to be a power law; *dashed line*: fit assuming $S(R)$ to be $S(R) \propto \exp(C/\sqrt{R})$.

ture from linearity can either signal a problem with the prediction (25) of the model or can be due to a breakdown of the power law assumption (2).

1. First option: breakdown of $E[\tilde{\eta}(\tau)|S(t)] \propto \tilde{S}(t)$

Suppose instead of (25) that

$$\langle \Delta S(S) \rangle \sim -\tilde{S}^x, \tag{32}$$

where x may be different from 1. Then, following step by step the derivation leading to (30), we obtain

$$\langle \Delta R(R) \rangle \sim R^{(1-x)/\mu} R, \tag{33}$$

where we have used the power law conversion (1) between sales and ranks derived from (2). When (25) holds, $x=1$ and expression (33) recovers the linear dependence of $\langle \Delta R(R) \rangle$ with R, for not too large R, quantified by the first term in the RHS of (30). Now, expression (33) is compatible with the data (31) together with the measurement $\mu \simeq 2$ only if we take $x=0$, which implies that $\langle \Delta S(S) \rangle$ does not depend in a first approximation on $\tilde{S}$. This in turn implies that the result $E[\tilde{\eta}(\tau)|S(t)] \propto \tilde{S}(t)\kappa(t-\tau)$ obtained above following the reasoning of Sec. III D 2 does not hold and must be replaced by the statement that $E[\tilde{\eta}(\tau)|S(t)]$ is independent of $\tilde{S}(t)$. Note that the dependence of $E[\tilde{\eta}(\tau)|S(t)]$ on $\kappa(t-\tau)$ is not necessarily linked to the validity of $E[\tilde{\eta}(\tau)|S(t)] \propto \tilde{S}(t)$ and thus the proportionality $E[\tilde{\eta}(\tau)|S(t)] \propto \kappa(t-\tau)$ can still hold, ensuring the validity of the exponent $1-2\theta$ for the foreshock and aftershock sales associated with an endogenous peak (see Table II).

A possible interpretation of the breakdown of $E[\tilde{\eta}(\tau)|S(t)] \propto \tilde{S}(t)$ is that the amplitude of the sales have a large stochasticity from book to book and the dependence of precursory innovations on the future amplitude of the sales's peak is lost. In sum, if we accept the conclusion drawn in insight that the sales' innovations are weakly dependent or are actually independent of the amplitude of the sales' peak, then the observed law (31) can be seen as a dramatic confirmation of the power law (2) through the rank-to-sale power law conversion (1) in the range of ranks up to a few thousands. However, as a note of caution, we cannot exclude the possibility that $x \neq 0$ and $\mu \neq 2$ as long as $(1-x)/\mu = 1/2$ holds: in such a case, expression (33) shows that we would recover the observed power law relationship (31) from the analogous derivation that led to (30).

2. Second option: deviations from the power law conversion

Let us consider a general function $S(R)$ relating ranks to sales. Equation (27) becomes $\Delta S = S'(R)\Delta R$. Using the empirical evidence $\langle \Delta R \rangle \propto R^{\beta}$ and assuming still the validity of $\langle \Delta S \rangle \propto -S$ for small R, we get $S(R) = S_{min}\exp(C/R^{\beta-1})$, where S_{min} and C are two constants. Assuming that this expression for small R can be used for all ranks of interest, we get

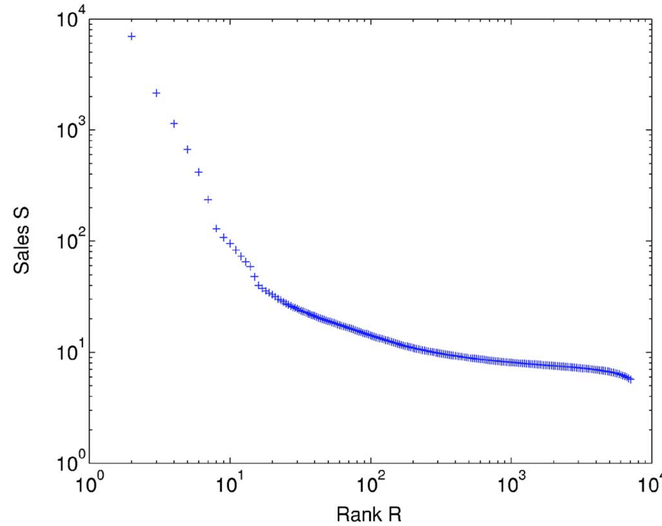

FIG. 15. (Color online) Rank-to-sales conversion function $S(R)$ estimated using expression (36) for arbitrary values of α, $\langle S \rangle$, S_{max}. Consequently, the units on the y axis are arbitrary units.

$$\langle \Delta R(R) \rangle = \frac{\alpha R^{\beta}}{C(\beta-1)} \left[1 - \frac{\langle S \rangle}{S_{min}} \exp\left(-\frac{C}{R^{\beta-1}} \right) \right]. \tag{34}$$

The two constants S_{min} and C can be adjusted to provide a rather good fit by (34) to the data, as shown in Fig. 14. By construction, $\langle \Delta R \rangle$ has the correct behavior for small R. All the difficulty is to describe how $\langle \Delta R \rangle$ reaches its maximum and then decreases. Typically, the fit gives $C=165$, which leads to the unrealistic value $S(1)/S_{min} = \exp(165) \approx 10^{70}$. Our mistake was to derive $S(R)$ for small R in order to use it to calculate the behavior of $\langle \Delta R \rangle$ for any R.

A better approach is to invert the empirical function $\langle \Delta R(R) \rangle$ to get $S(R)$, without making assumptions about it. For a given function $S(R)$, Eq. (28) becomes

$$S'(R)\langle \Delta R(R) \rangle = -\alpha(S(R) - \langle S \rangle). \tag{35}$$

Thus we can express $S(R)$ as a function of $\langle \Delta R \rangle$ and obtain

$$S(R) = \langle S \rangle + (S_{\max} - \langle S \rangle)\exp\left(-\alpha \int_1^R \frac{1}{\langle \Delta R(R') \rangle} dR' \right), \tag{36}$$

where S_{max} is a constant equal to $S(R=1)$.

The problem in using (36) is that we do not know the value of the constant α, which is rather crucial as it appears in the exponential. Nevertheless, whatever the value of α, we observe roughly the same behavior for $S(R)$. Figure 15 shows $S(R)$ obtained by inserting the empirical dependence $\langle \Delta R(R) \rangle$ as a function of R in expression (36) for fixed values of α, S_{max} and $\langle S \rangle$. The "fast" decrease for small R (typically $R < 10-20$) followed by a "slow" one for large R's is typical of most reasonable parameters. This shows that the conversion can be roughly seen as a power law (an approximate straight line in a log-log plot) in the range $R \in [20, 10^4]$. But we are not able to determine the value of the exponent from this approach because we do not know α.

In summary, the direct determination of the function $S(R)$ requires some additional assumption as shown in the various attempts developed above. Nevertheless, we have seen that a power law would explain the observed nonmonotonous behavior for $\langle \Delta R(R) \rangle$. Notice that if we assume $S(R)$ to be exponential: $S(R)=S_0 \exp(-R/R_\star)$, the relation (35) implies:

$$\langle \Delta R(R) \rangle = \alpha R_\star \left(1 - \frac{\langle S \rangle}{S_0} \exp\left(\frac{-R}{R_\star}\right)\right), \qquad (37)$$

which is obviously monotonous and can thus be rejected by comparison with the data.

VI. CONCLUSION

In our study of the ranks of books sold by Amazon.com, we have shown that sales shocks can be classified into two categories: endogenous and exogenous. We have used two independent ways of classifying peaks, one based on the acceleration pattern of sales (see Sec. IV B) and the other based on the exponent of the relaxation (see Sec. IV C). We have developed an epidemic model of influences between buyers connected within a network of acquaintances. The comparison between the predictions of the model and the empirical data suggests that social networks have evolved to converge very close to criticality (here in the sense of critical branching processes). It means that tiny perturbations, which in any other state would be felt only locally, can propagate almost without any bound. Studies of critical phenomena shows that very different systems can exhibit fundamental similarities, this is generally referred as universality.

While we have emphasized the distinction between exogenous and endogenous peaks to set the fundamentals for a general study, we also find closely repeating peaks as well as peaks that may not be pure members of a single class. In a sense, there are no real "endogenous peaks," one could argue, because there is always a source or a string of news impacting on the network of buyers. We have thus distinguished between two extremes, the very large news impact and the structureless flow of small news amplified by the cascade effect within the network. One can imagine and actually observe a continuum between these two extremes, with feedbacks between the development of endogenous peaks and the attraction of interest of the media as a consequence, feeding back and providing a kind of exogenous boost, and so on. Our framework allows us to generalize beyond these two classes and to predict the sales dynamics as a function of an arbitrary set of external sources.

If Amazon.com would release its data, we suggest future promising directions of inquiries.

Before anything else, the same study should be done again. We expect more accurate results as we work with the real sales and not only with an estimate derived from the ranks.

Secondly, a direct access to sales should make it possible to inverse $S(t)$ to get access to the news innovations $\eta(t)$ describing the sources of spontaneous buys. We currently describe $\eta(t)$ as a white noise distributed according to a power law. This zero order approximation could be improved and lead to a better understanding of the statistical properties of the news $\eta(t)$. One can even imagine to reconstruct the specific history of news that were amplified by the epidemic process to obtain a given sales history for a given book.

Different kinds of books should involve different social networks. Take the example of the book "Divine Secrets of the Ya-Ya Sisterhood" by R. Wells. It became a bestseller two years after publication, with no major advertising campaign [14]. Following the reading of this originally small budget book, "Women began forming Ya-Ya Sisterhood groups of their own […]. The word about Ya-Ya was spreading […] from reading group to reading group, from Ya-Ya group to Ya-Ya group [14]." By looking at different classes of books, we would expect to highlight different network characteristics [15,16].

Amazon has the addresses of its customers. This can be used to study the geographical spread of books.

ACKNOWLEDGMENTS

The present analysis of Amazon sales ranks is unauthorized and in no way sponsored by Amazon. We thank Y. Ageon and T. Gilbert for help and useful discussions and M. Rosenthal for important information on the Amazon ranking system. We also thank M. Rosenthal for his permission to reproduce Fig. 3.

[1] D. Ruelle, Phys. Today **57**, 48 (2004).

[2] D. Sornette, "Endogenous versus exogenous origins of crises" in the monograph entitled *Extreme events in Nature and Society*, edited by S. Albeverio, V. Jentsch, and H. Kantz (Springer, Berlin, 2005) (http://arxiv.org/abs/physics/0412026).

[3] D. Sornette, F. Deschâtres, T. Gilbert, and Y. Ageon, Phys. Rev. Lett. **93**, 228701 (2004).

[4] B. M. Roehner, D. Sornette, and J. V. Andersen, Int. J. Mod. Phys. C **15**, 809 (2004).

[5] Of course, we should mention search engines such as Google. To have a look at the most popular queries, see http://www.google.com/press/zeitgeist.html.

[6] http://msdn.microsoft.com/library/default.asp?url=/workshop/browser/webbrowser/webbrowser.asp

[7] In October 2004, Amazon completely redesigned its ranking system. See http://www.fonerbooks.com/surfing.htm for details. Not only the meaning of the ranks have changed, due to the inclusion of e-books and Marketplace sales, but also the infrastructure of the system has been modified: the three tiers are gone, all ranks are updated every cycle, which runs around once an hour. It is thus no longer possible to compare the ranks before to those after October 2004. It is important to realize that Amazon's ranking system may again evolve on a moment's notice.

[8] We indeed found a strong spectral peak at the period equal to one day on the spectrum of the time series of book ranks.
[9] M. Rosenthal, http://www.fonerbooks.com/surfing.htm.
[10] D. Sornette, *Critical Phenomena in Natural Sciences*, Springer Series in Synergetics, 2nd ed. (Springer, Berlin, 2004).
[11] J. Brody, "Push up the weights, and roll back the years," *The New York Times* F7 (June 4, 2002).
[12] D. Sornette and A. Helmstetter, Physica A **318**, 577 (2003).
[13] B. M. Roehner and D. Sornette, Eur. Phys. J. B **16**, 729 (2000).
[14] M. Gladwell, *The Tipping Point: How Little Things Can Make a Big Difference* (Back Bay Books, Boston, 2002).
[15] D. J. Watts, *Six Degree* (W. W. Norton & Company, New York, NY, 2003).
[16] A. L. Barabàsi, *Linked* (Perseus, Cambridge, 2002).