Interarrival times of message propagation on directed networks

Tamara Mihaljev[*]
Computational Physics, IfB, ETH Zurich, Schafmattstrasse 6, CH-8093 Zurich, Switzerland

Lucilla de Arcangelis[†]
Department of Information Engineering and CNISM, Second University of Naples, I-81031 Aversa (CE), Italy

Hans J. Herrmann
Computational Physics, IfB, ETH Zurich, Schafmattstrasse 6, CH-8093 Zurich, Switzerland and
Departamento de Física, Universidade Federal do Ceará, 60451-970 Fortaleza, Ceará, Brazil
(Received 29 October 2010; revised manuscript received 6 May 2011; published 15 August 2011)

One of the challenges in fighting cybercrime is to understand the dynamics of message propagation on botnets, networks of infected computers used to send viruses, unsolicited commercial emails (SPAM) or denial of service attacks. We map this problem to the propagation of multiple random walkers on directed networks and we evaluate the interarrival time distribution between successive walkers arriving at a target. We show that the temporal organization of this process, which models information propagation on unstructured peer to peer networks, has the same features as SPAM reaching a single user. We study the behavior of the message interarrival time distribution on three different network topologies using two different rules for sending messages. In all networks the propagation is not a pure Poisson process. It shows universal features on Poissonian networks and a more complex behavior on scale free networks. Results open the possibility to indirectly learn about the process of sending messages on networks with unknown topologies, by studying interarrival times at any node of the network.

DOI: 10.1103/PhysRevE.84.026112 PACS number(s): 89.75.Fb, 05.40.Fb, 64.60.aq

I. INTRODUCTION

A botnet is a network of infected computers which are under command and control of a single person, the "botmaster." Botnets are abstract overlay networks on top of the physical network topology. They are used for sending unsolicited commercial emails (SPAM), viruses, denial of service attacks, for stealing identity data, and for other sorts of cybercrime. Botnets are the primary security threat on Internet today. They grew up to be a global and multimillion dollar business. Fighting botnets is a hard task since their structure is constantly evolving, and their inner working is not known. Understanding the dynamics of communication on such networks is a big challenge and could be crucial for finding effective tools for fighting botnets. This problem can be successfully approached from the network theory perspective.

Botnets can have different structures. The new generations of botnets, which are more robust against attacks and very difficult to track, are based on peer to peer communication [1]. This type of communication can be modeled as a random walk. When the botmaster sends an order to its bots, it sends it only to a fraction of nodes in the botnet, which can then forward it to only those bots whose IP addresses they know. These addresses are randomly assigned to bots. There is no centralized point in such networks since all the nodes are equally important; they are all clients and servers at the same time. This is the reason for botnets' robustness against attacks. Even if a node of the botnet is identified, its communication with the rest of the botnet can be tracked back only to a limited number of bots. Therefore no attack would destroy the whole botnet, or lead to the botmaster. Links in peer to peer networks do not have to be bidirectional, and the unstructured peer to peer networks, often used for constructing botnets since they are the most difficult to track, have random topology. Therefore we will model botnets as random directed networks, and we will study the propagation of random walkers on them in order to attack the problem of understanding the internal mechanisms of message propagation on these networks. Random walks on directed networks are also an interesting fundamental problem, important for understanding the communication in any other peer to peer network [2], wireless sensor networks [3], *ad hoc* networks [4], or different processes on the World Wide Web, such as tagging [5]. The results we present are general, not botnet specific, and are valid for any system in which data packets propagate in a random fashion on the directed network.

Random walks and related stochastic processes have mainly been studied on regular lattices and d-dimensional Euclidean spaces in the past [6], due to their obvious relevance to physical problems. In recent years networks are becoming the preferred model to study complex systems [7,8] and this triggered studies of random walks on them [9–22]. However, most of the previous results concern random walks on undirected networks. Random walks on directed networks have been mainly investigated to find communities in citation networks [23], identify subgraph structures on the World Wide Web [24], or in calculations of the PageRank [25,26]. This is a measure used by the search engine Google (as well as by several other search engines) to determine the prestige of Web pages. When a user submits a query, the hits returned by Google are

[*]tamaram@ethz.ch
[†]dearcangelis@na.infn.it

ranked according to the value of their PageRank. The algorithm determining this value is based on a modified random walk on the web graph, were nodes are Web pages and edges between them are naturally directed hyperlinks connecting Web pages. In each step the modified random walker either follows a randomly chosen outgoing link of the present node, or with a small probability, called the damping factor, it jumps to a randomly chosen node in the network.

We study random walks on directed networks to model peer to peer communication on botnets as the spreading of messages. The spam propagation problem is quite complex since it consists of a first process, where the botmaster gives orders to its bots, and of a second process, where bots send spams to a list of users. Whereas it is reasonable to suppose that the botmaster gives orders in a random fashion to its bots, and therefore this process can be modeled by a random walk propagation on directed networks, very little is known about the second process, which can vary from bot to bot and is continuously updated by botmasters. In this sense, the comparison with spam data is not strictly justified since the simulation models only the first process. However, it may provide an indication on the real bot topology. We are interested in the temporal organization of this dynamic process and therefore we investigate the distribution of interarrival times between two successive messages arriving at a given receiver. The interarrival time distribution has been first introduced to characterize the temporal occurrence of earthquakes [27]. It has been then studied in the context of different stochastic processes, as solar flares [28–31], forest fires [32], or in package transport in computer science [33,34]. The interesting property of this quantity is that it is able to provide information about the temporal organization of processes whose detailed mechanisms are unknown. If the temporal occurrence of events is completely decorrelated (Poisson process), the interarrival time distribution can be derived analytically and is an exponential function. Detection of a nonexponential behavior of interarrival time distribution enlightens the presence of temporal correlations among events. In contrast to exponential distribution in Poisson processes, the presence of a power law regime in the distribution is the indication of an occurrence rate decaying in time as a power law (result derived analytically by Utsu for earthquake occurrence [35]), which is the clear signature of temporal clustering. For earthquakes, this is the well known Omori law for aftershock occurrence: after a large earthquake, the occurrence rate abruptly increases and then decreases in time as a power law. Indeed aftershocks occur close in time just after the mainshock and then their occurrence rate decays as time goes on. The interarrival time distribution for earthquakes presents a power law regime, confirming that events occur in bursts.

A recent paper [36] has investigated the statistical properties of the SPAM delivery interarrival times. Results have suggested that SPAM messages delivered to a given recipient are time correlated: if the interarrival time between two consecutive SPAM messages is small (large), then the next SPAM message will most probably arrive after a small (large) interarrival time. SPAM temporal correlations have been reproduced by a numerical model based on the random superposition of SPAM sequences, each one described by the Omori law [37]. This and other experimental findings [36] suggest that statistical approaches may be used to infer how spammers operate.

Our motivation to study the distribution of message interarrival times on model networks is to detect the eventual presence of temporal correlations and their relation with the network topology. The interarrival time distribution of messages sent only to one or a small fraction of nodes could then provide information about the dynamical process taking place on a real network. In the case of botnets this would imply that we would be able to get information about their organization and structure by studying interarrival times of either SPAM emails or of contaminated packages, both sent by a botnet, by analyzing data even of a single user. These data are easily accessible, cheap, and easy to monitor. Since botnets are difficult to identify, this indirect way of learning about their organization would give a boost in fighting botnets and cybercrime in general.

The paper is organized as follows: In the next section we will describe the model and the implemented networks. In the third section we discuss results obtained for random networks and show their comparison with the real data. In the following two sections we extend our investigation to networks with random topologies without dead ends and to scale free networks. Finally in the last section we discuss the results and give some concluding remarks.

II. MODEL

We start by constructing a randomly connected directed network with a given degree distribution of inputs and outputs. We implement the Poisson distribution, the Poisson without dead ends, and the power law distributions. We choose a random node in the network to be the one where we measure the interarrival times between two successive random walker arrivals, $dt = t_{i+1} - t_i$. We call this node the target node or the receiver. The target node is chosen at random among nodes with a given number of inputs. The number of outputs of the target is not fixed since it does not influence the number of message arrivals. Next, we choose a node from which messages depart, the sender. This node is chosen at random, with a fixed number of outputs. Since only the number of outputs determines the number of different ways a message can leave on its way to the receiver, we do not fix the number of inputs of a sender. In our process the sender is the botmaster sending orders to its bots, and what we measure at the receiver can be compared to the arrivals of messages to a generic user. We have found that increasing the number of senders does not affect the basic properties of the interarrival time distribution of messages reaching the receiver.

After a message departs from the sender, it performs a random walk, namely it follows at each time step one randomly chosen outgoing link of the occupied node. The messages are sent one after the other and the time needed to reach the target is recorded. Messages are also sent with a constant delay. In this case the delay time is added to the measured length of the walk, the two giving together the arrival time. The walker arrival times (with and without delays) are listed and sorted. In this way the sorted list represents correctly the process we simulate. The interarrival times are calculated

from this sorted list in which the first walker is the one with the shortest arrival time (including possible starting delay), no matter when in the simulation this walk has been measured. This procedure simulates the process in which the messages propagate simultaneously. The walk continues until either the target node is reached, or an initially fixed maximal number of steps is exceeded. It is necessary to introduce a limit on the number of steps since, due to the fact that the links are directed, the network may have regions that the walker can enter but cannot escape from. This limit also exists in real internet protocols where it is called the *time to live* (TTL). This is a limit on the number of transmissions that a data package can experience before it should be discarded. In all simulations presented here we fix this value to twice the size of the network.

To simulate the dynamical process typical for a botmaster, who sends a large number of messages through botnets, many random walkers depart from the sender. We send them either one by one or all at once. The walkers are independent. We record the times at which messages arrive to the receiver t_i and calculate the interarrival times between successive messages, $dt = t_{i+1} - t_i$. When the messages are sent all at once, their arrival times depend solely on the length of the paths undertaken on their way to the receiver. This is mainly affected by the network topology. When the messages are sent one after another, each consecutive message has an equal time delay in starting its walk to the target node. This process is more complex than the previous one since a message can arrive to the target before others sent earlier due to a shorter undertaken path. In the beginning of the process the number of messages arriving at the target increases with time, until a stationary state is reached. We are interested in the stationary state of the process, where we measure the distribution of interarrival times $P(dt)$, normalized by the number of received messages and the number of networks in the sample. In our model at each time step a new walker departs from the sender. We have, however, verified that the introduction of a longer time delay between successive departures does not affect the main properties of the distribution.

III. RANDOM NETWORKS

All the networks randomly connect input and output links assigned to the nodes according to some distribution [38,39]. We call, however, the networks random only when the distribution is Poissonian, $p(k) = (\langle k\rangle^k/k!)\cdot e^{\langle k\rangle}$, where k is the number of links and $\langle k\rangle$ is its mean value [40]. When we construct random networks we choose both, the distribution of input and of output links, to be Poissonian with the same mean degree. The number of walkers has to be large enough to provide good statistics for the distribution of interarrival times.

We sample data from 500 different network realizations for a given degree distribution and fixed values of the sender outdegree and the receiver indegree. We fix both these values equal to 4. We have, however, verified that the specific value of these parameters does not affect the behavior of the distribution. The space of possible topologies and possible choices of a sender-receiver couple is extremely large. To get better statistics we also fix the distance between these two nodes, namely the shortest path between them.

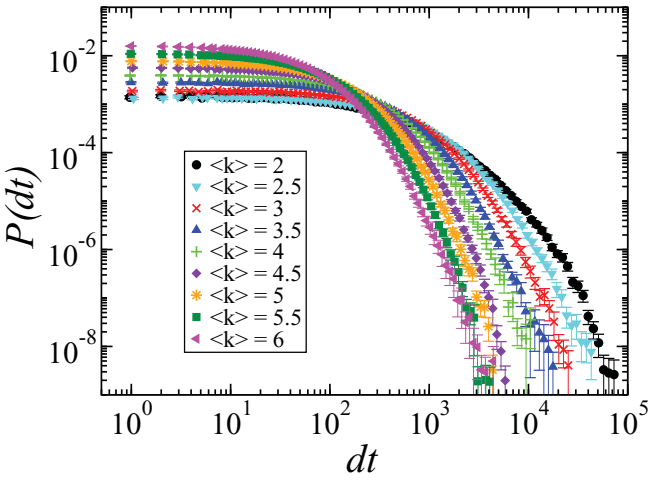

FIG. 1. (Color online) Distribution of interarrival times in networks with Poisson degree distribution and different values of the mean degree. The networks have $N = 10^4$ nodes, the number of sent messages is $M = 10^6$, and the distance between the sender and the receiver is fixed to 8. The messages are sent one by one.

We first study the case when messages are sent one by one. We find that the distributions of interarrival times for networks with Poisson distributed links and different average degree $\langle k\rangle$ exhibit similar behavior (Fig. 1). Only for networks with a small value of the average degree, and therefore a lower level of connectivity, are longer interarrival times measured. The average interarrival time indeed increases exponentially as the average connectivity, $\langle k\rangle$, decreases (inset Fig. 2). In order to check if the distribution is a universal function, solely

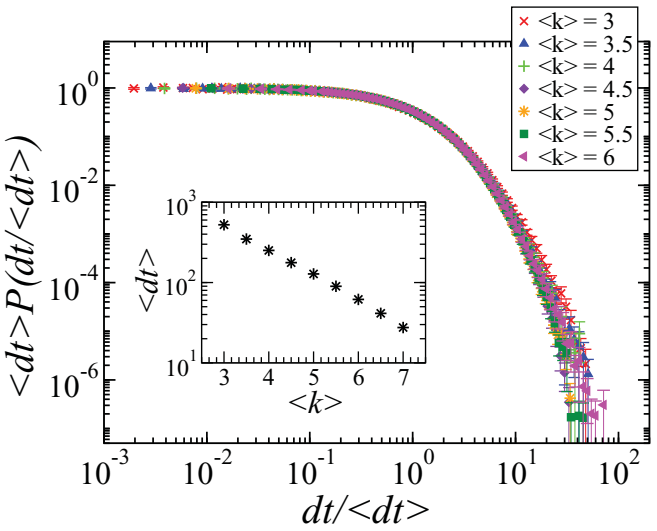

FIG. 2. (Color online) Distributions of interarrival times in networks with Poisson degree distribution and different values of the mean degree, rescaled by the average rate of message arrivals. The networks have $N = 10^4$ nodes, the number of sent messages is $M = 10^6$, and the distance between the sender and the receiver is fixed equal to 8. The messages are sent one by one. The inset shows the dependence of the mean interarrival time on the average degree of the network.

controlled by the average rate R of walkers arriving at the target, we verify the following scaling relation [41]:

$$P(dt) = Rf(Rdt).$$

Therefore we evaluate the average rate for each distribution, as the inverse of the mean interarrival time $R = 1/\langle dt \rangle$, and rescale the interarrival time by the average rate. We find that the different distributions collapse quite well onto a universal curve (Fig. 2), if the total number of links in the network is large enough. Small deviations are observed only for large dt. If the network is too sparse, i.e., if $\langle k \rangle \leqslant 3$, the mean interarrival time increases and the probability of longer dt becomes larger. This effect is caused by the fact that in sparse networks only a small number of messages reaches the target. The space of possible paths leading to it is not fully explored and the trapping regions in sparse networks are more prominent.

The universal scaling function in Fig. 2 exhibits an initial almost constant regime followed by an exponential like decay. If the messages would arrive to the target independently of each other, interarrival times would be distributed exponentially, as it happens in Poissonian processes. In our process the distribution of interarrival times deviates from the exponential, which indicates that the process is more complex, and possibly correlations are present, coming from the networks topology or the message sending process itself. We have also studied the influence of the distance d between the sender and the receiver on the distribution of interarrival times. If the nodes are not too far away (for $d < 8$), the rescaling of the distributions by the average rate of message arrivals provides a good collapse (Fig. 3) with fluctuations at large interarrival times only for $d = 8$. The inset shows that the mean interarrival time grows linearly with the sender-receiver distance.

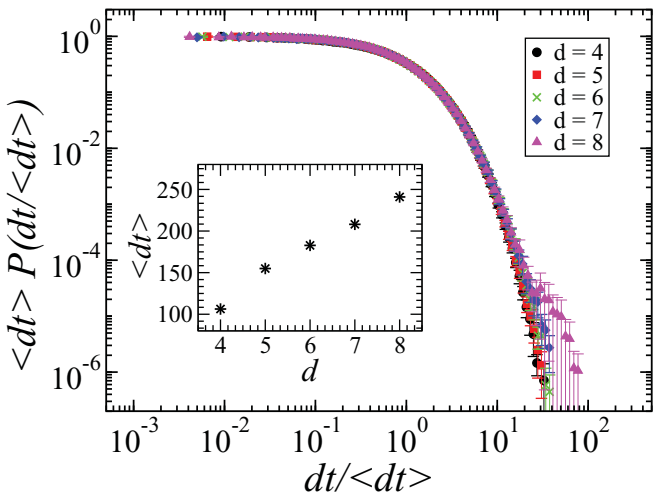

FIG. 3. (Color online) Distributions of interarrival times in networks with Poisson degree distribution with mean degree equal to 4 and different distances between the sender and the receiver. The distributions are rescaled by the average rate of message arrival. The networks have $N = 10^4$ nodes, the number of sent messages is $M = 10^6$. The messages are sent one by one. The inset shows the dependence of the mean interarrival time on the distance between the sender and the receiver.

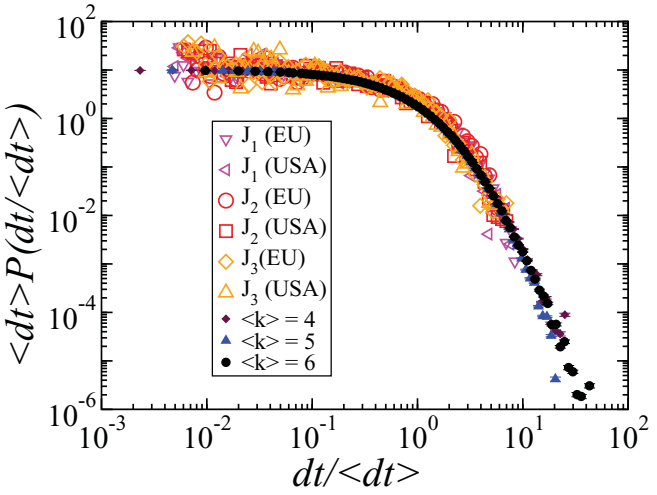

FIG. 4. (Color online) Rescaled distributions of interarrival times of SPAM emails sent from two different domains and sampled in three different junk mailboxes (empty symbols) from Ref. [36], and of our model (full symbols) with messages sequentially sent through random networks with different $\langle k \rangle$ values and the distance between the sender and the receiver fixed to 8. The networks have $N = 10^4$ nodes, the number of sent messages is $M = 10^6$.

The present model simulates propagation of messages in unstructured peer to peer botnets where the botmaster is sequentially sending a large number of messages to the bots. The dynamics of sending orders inside the botnets influences the dynamics of arrival of messages sent from bots to the final destination, which could be a computer of a simple user receiving SPAM. Therefore the fingerprint of the dynamics of message propagation inside the botnets should be visible in the distribution of interarrival times of SPAM emails collected in the mailbox of a single user, although our model does not simulate the whole process of message propagation from the spammer to its victim. We compare our results with the distribution of interarrival times for SPAM data, presented in Ref. [36]. The experimental data in Fig. 4 are sampled from three different junk mailboxes, and the spam emails are selected on the basis of their geographical origin, Europe and the United States. Surprisingly, already our simple model is able to reproduce quite well the basic characteristics of the interarrival time distribution of the real data, as can be seen in Fig. 4.

In order to understand the characteristics of the dynamical process, and thus the behavior of the interarrival time distribution for experimental data, we study in detail different aspects of the dynamics on model networks. If the walkers are sent one after the other the interarrival time does not depend only on the different paths taken by the walkers but also on their starting time. To understand the effect of this time delay on the process we also analyze the case where all the messages are sent at the same time. In this case interarrival times are uniquely determined by the complexity of the undertaken paths. The distribution of interarrival times rescaled by the average rate shows universal behavior, well fitted by a power law with the exponent close to 2 (Fig. 5). We show results for only the networks which are not too sparse, since for sparse networks the distribution shows large statistical fluctuations. The power law behavior suggests that even if the walkers are completely

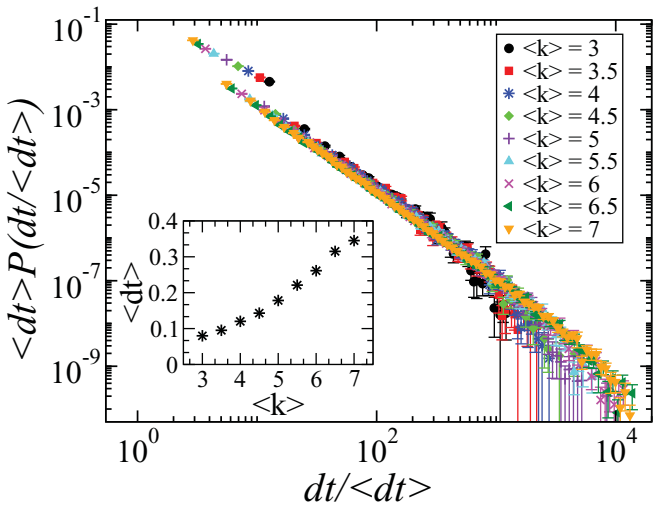

FIG. 5. (Color online) The rescaled distributions of interarrival times in networks with Poisson degree distribution with different mean degrees. All the messages are sent at the same time and the distance between the sender and the receiver is equal to 8. The networks have $N = 10^4$ nodes, the number of sent messages is $M = 10^6$.

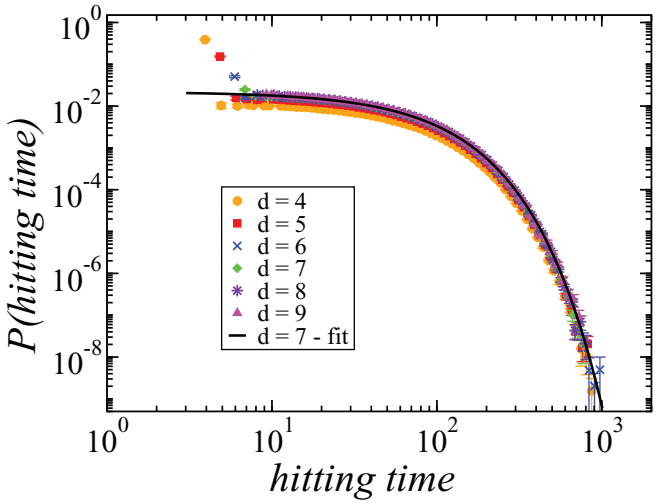

FIG. 6. (Color online) The distributions of hitting times in networks with Poisson degree distribution with mean degree equal to 4 and different distances between the sender and the receiver. The networks have $N = 10^4$ nodes, the number of sent messages is $M = 10^6$. The messages are sent all at once. The fit is a stretched exponential function of the form $y = 0.02 \cdot \exp(-0.02 \cdot x^{0.96})$.

independent during their propagation, they arrive in bursts originating temporal clustering in the process. The probability for the shortest dt depends on the level of connectivity in the network, moreover the higher the average degree the smaller dt is with respect to the average rate. In general, we observe that the average degree controls the extension of the scaling regime: the more interconnected is the network the wider is the variety of possible paths and therefore the range of observed dt.

The maximal value of interarrival times is determined by the TTL in the case where all messages are sent at once. In the case when messages are sent one after the other, the time delay between messages affects the interarrival time, which can become as large as $M + TTL$, since the number of sent messages M is the maximal possible additional delay between two messages. This maximal value corresponds to the particular situation where the first message arrives and the only other message reaching the target is the one sent as last, which takes the longest possible path.

Since the outputs of random networks are distributed according to the Poisson degree distribution, a fraction of nodes in the network has $k = 0$ outputs. Such nodes exist also in real networks. These nodes serve as a trap for the random walker. Similarly to the trapping problem on networks [42], the message reaching such a node cannot proceed any further. In random networks this is a dominant mechanism for preventing messages from finding the target. A large number of messages gets lost and the interarrival times can become extremely large (only up to 1% of messages actually reaches the target). It is highly unlikely that a message will be stopped because the length of its path has reached the limit given by TTL. It rather appears that on random networks messages either reach the target after a relatively short period of time, or they never reach it. We check this point by studying the distribution of hitting times. The hitting time, or the first passage time, is the time that a random walker takes to reach the target for the first time. In our model the distribution of hitting times is equivalent to the distribution of the lengths of the paths taken by the random walkers to reach the target node. This is at the same time the distribution of arrival times in the model where all the messages are sent at the same time. In Fig. 6 we see that only a small number of walkers takes the maximal number of 1000 steps, which is relatively small for the $N = 10^4$ networks. The most probable hitting time has a value equal to the distance between the sender and the receiver, meaning that many messages take the shortest path between the two nodes. Its probability is higher for smaller distances. The other possible paths are distributed according to a stretched exponential distribution (Fig. 6) independent of the sender-receiver distance.

The hitting time, or the first passage time distribution, is an important property of random walks that has been studied in the past [11,43,44]. A number of nice and complete analytical results are known for random walks on undirected networks [9,12]. For simple processes, the interarrival time distribution can be calculated if the first passage time distribution is known. For instance, if $p(dt = t_b - t_a)$ is the probability that the interarrival time is dt, and $p(t_i)$ is the probability that the first passage time of a walker is t_i, i.e., it reaches the target time t_i, then

$$p(dt = t_b - t_a) = p(t_a)\left(1 - \prod_{i=a+1}^{b-1} p(t_i)\right) p(t_b).$$

However, in the present case, where networks are directed and not completely connected, the random walk does not have a stable steady state, which makes analytical calculations much more complex. The properties of random walks are for these networks much different than in the undirected case. Our numerical results confirm that this is due to lost walkers, which are here present mainly due to the existence of nodes without outputs.

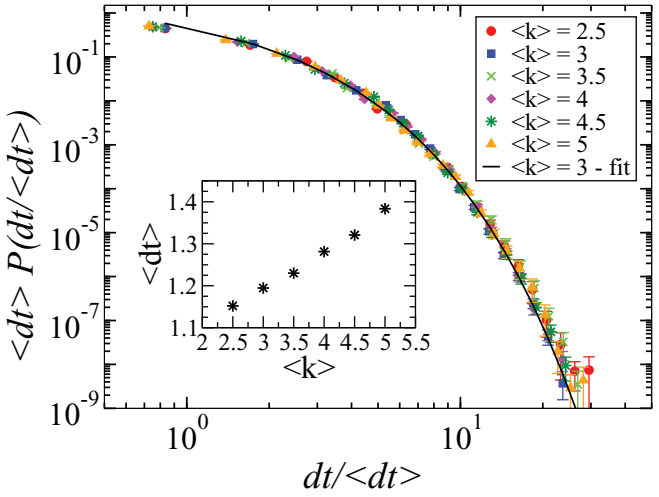

FIG. 7. (Color online) The distribution of interarrival times in networks with Poisson degree distribution without dead ends and with different values of the mean degree, rescaled by the average rate of message arrival. The networks have $N = 10^4$ nodes, the number of sent messages is $M = 10^6$, and the distance between the sender and the receiver is fixed equal to 5. The messages are sent one by one. The fit is a stretched exponential function of the form $y = 2.2 \cdot \exp(-1.5 \cdot x^{0.8})$. The inset shows the dependence of the mean interarrival time on the average degree of the network.

IV. RANDOM NETWORKS WITHOUT DEAD ENDS

In order to better understand the influence of nodes without outputs on the distribution of interarrival times, we study the same dynamical process on networks with slightly different topology. When we assign the number of ingoing or outgoing links to a node, we choose a random number between zero and $N - 1$ from the Poisson distribution, but we assign to the node this number plus 1. In this way there are no nodes with zero ingoing or outgoing links and there are no dead ends in the network. The only trap in the network is now the target node.

For the process where walkers are sent one by one we find that the distributions depend weakly on the average degree and the distance between the sender and the receiver. Indeed, the average rate varies on a much smaller range (insets in Figs. 7 and 8). Very good collapse is therefore observed rescaling the distributions by the average rate (Figs. 7 and 8). The universal function behaves as a stretched exponential and is therefore different than the one in Fig. 2. The main mechanism preventing the message from arriving at the target is now time exceeding the TTL limit. Many messages arrive to the target (up to 90%), which results in small interarrival times. In this case the walkers explore most of the paths existing between the two nodes. Similarly to previous cases, we show the results only for networks which are not too sparse, where the distributions show larger statistical fluctuations.

Conversely, for the process where all messages are sent at once, the interarrival times are in the majority of cases either zero, i.e., two messages arrive to the target at the same time, or equal to 1. This is due to the large number of messages arriving at the target and to the ability of the walkers to explore well the space of all possible paths, with lengths ranging between the shortest path and the TTL. At each time step then at least one walker arrives to the target leading to an interarrival time equal to 1.

The distribution of hitting times when all messages are sent at once is also quite different than in the case where dead ends exist. In this case the walkers have the possibility to sample paths of all lengths and therefore the hitting time can assume values up to the threshold TTL. As we can see in Fig. 9, the distribution has an exponential behavior. Deviations from the exponential function can be seen only for small values

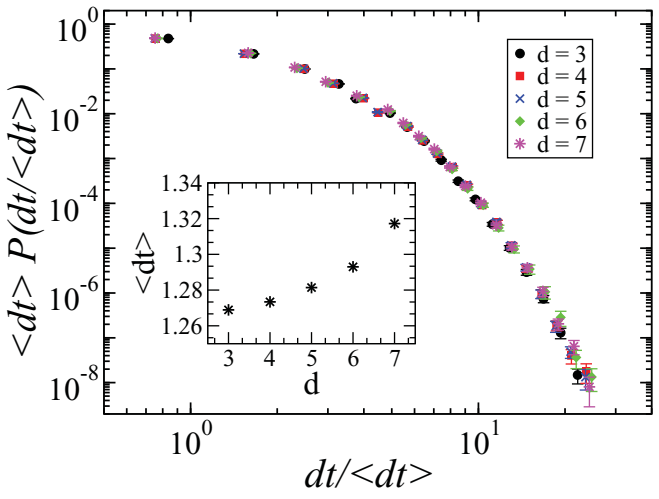

FIG. 8. (Color online) The distribution of interarrival times in networks with Poisson degree distribution without dead ends and with different distances between the sender and the receiver, rescaled by the average rate of message arrivals. The mean of the Poissonian distribution is $\langle k \rangle = 4$. The networks have $N = 10^4$ nodes, the number of sent messages is $M = 10^6$, and the distance between the sender and the receiver is fixed equal to 5. The messages are sent one by one. The inset shows the dependence of the mean interarrival time from the distance between the sender and the receiver.

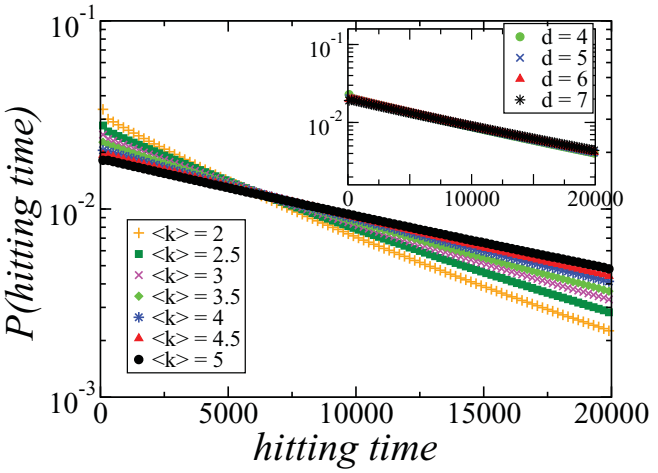

FIG. 9. (Color online) The distributions of hitting times in networks with Poisson degree distribution without dead ends and for different $\langle k \rangle$. The distance between the sender and the receiver is equal to 5. The inset shows the same distribution for the average degree equal to 4 and different distances between the sender and the receiver. The networks have $N = 10^4$ nodes, the number of sent messages is $M = 10^6$.

of the hitting time, with this region getting smaller when the connectivity is larger. The coefficient of the exponential distribution increases with $\langle k \rangle$, namely the probability for longer hitting times is higher for larger average degrees. The walker takes more tortuous paths in a network with a higher level of connectivity. Conversely, the coefficient is independent of the distance between the sender and the receiver (see the inset of Fig. 9).

The results obtained for random networks without dead ends confirm our conclusion that the behavior of the interarrival time distribution for random networks is a consequence of the existence of dead ends, the nodes without outputs which serve as traps for the messages. Since these traps exist in real networks, it is important to understand their influence on propagation of messages through random networks.

V. SCALE-FREE NETWORKS

To explore further the influence of the network topology on the propagation of random walkers, we study this process on scale free networks. We find that in this case the characteristics of the process are much different. In order to measure the distribution we wait for the process to become stationary. Whereas in the case of Poisson distributed links this happens very fast, for scale free networks the stationary regime is reached after a long transient. A large number of messages has to be sent to obtain good statistics.

In Fig. 10 we show the distribution of interarrival times for scale free networks with different exponents of the degree distribution [$P(k) \propto k^{-\gamma}$] and in the case that messages are sent one by one. We observe that sparse networks (high γ) behave differently than well connected networks. In fact, the mean interarrival time rapidly increases for decreasing γ, suggesting that the walker takes very tortuous paths in a well connected network. In contrast to the previous cases, the distributions of interarrival times for scale free networks do not collapse onto a universal curve if the interarrival time is rescaled by the average rate of the process, even when the networks are not sparse. Therefore the average rate is not the only relevant quantity in the process.

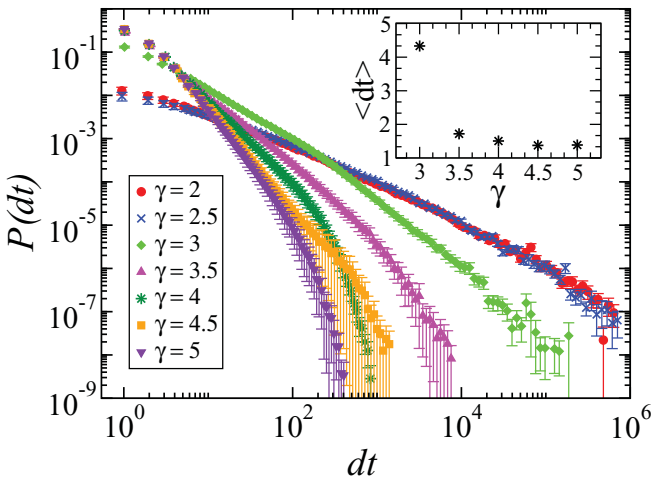

FIG. 10. (Color online) The distribution of interarrival times in networks with power law degree distribution with different values of coefficient γ [$P(k) \propto k^{-\gamma}$]. The networks have $N = 10^4$ nodes, the number of sent messages is $M = 10^6$, and the distance between the sender and the receiver is fixed equal to 30. The messages are sent one by one. The inset shows the dependence of the mean interarrival time on γ.

The number of messages arriving at the target is smaller than in Poisson networks without dead ends, and the interarrival times can be extremely long (the number of messages arriving at the target varies strongly with the exponent of the distribution, and can range from 1 to 90%). In scale free networks there are by definition no dead ends, and the main mechanism for stopping the random walker is here time exceeding the TTL limit. The walkers take many long paths, probably looping through system and being able to visit different parts of the networks through shortcuts whose probability is higher due to hubs. In contrast to Poisson networks without dead ends, although the walkers can explore well the space of possible paths, many of them never reach the target since they are either trapped in loops, or in regions typical for directed networks, where the walker can enter but cannot escape. Moreover, the path to the target could be longer than the TTL limit, which is very probable on scale free networks. Since the number of walkers reaching the target decreases, the probability of longer interarrival times increases.

When messages are sent all at the same time, the distribution of interarrival times shows a power law behavior. However, in contrast to random networks with dead ends, where we find universal power law behavior, a change in the topology by tuning the coefficient γ changes the slope from about 0.8, in well connected, to 2.2, in sparse networks (Fig. 11). As in the case of consecutive departures of messages, also here we do not observe the distribution collapse if interarrival time is rescaled by the average rate.

The hitting times of the dynamical process on scale free networks also show a quite different behavior. In Fig. 12 we see that networks with different power law coefficients have different distributions of hitting times. We also see that, similarly to the case of Poisson networks without dead ends,

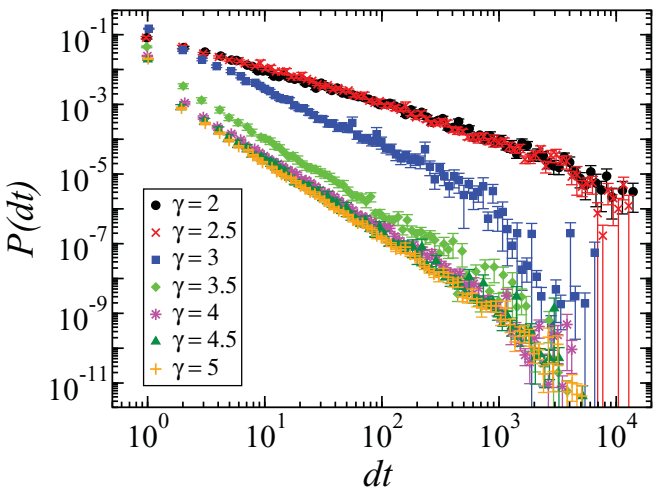

FIG. 11. (Color online) The distributions of interarrival times in networks with scale free distribution with different coefficients of the power law. All the messages are sent at the same time and the distance between the sender and the receiver is 30. The networks have $N = 10^4$ nodes, the number of sent messages is $M = 10^6$.

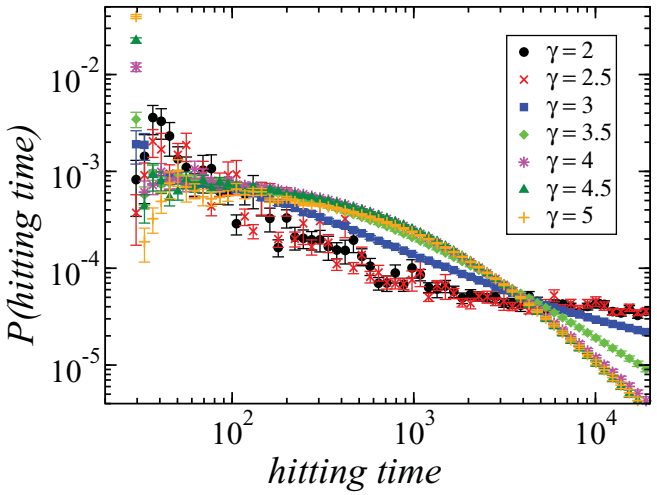

FIG. 12. (Color online) The distributions of hitting times in networks with power law degree distributions with different coefficients γ. The distance between the sender and the receiver is fixed to 30. The networks have $N = 10^4$ nodes, the number of sent messages is $M = 10^6$.

the walkers on scale free networks explore well the space of all possible paths from shortest path to TTL, especially in the case of well connected networks (smaller γ exponents). By increasing γ, the number of walkers which take very long paths to reach the target decreases.

VI. CONCLUSIONS

In this paper we study the distribution of interarrival times of walkers sent through complex networks, with the goal to gain understanding on the process of information spreading inside the new generation of botnets based on peer to peer communication. The dynamics of sending botmaster orders inside the botnet influences the dynamics of arrival of messages sent from bots to the final destination. This can be a computer of a simple user receiving either SPAM or data packages contaminated with viruses. The dynamics of message propagation inside the botnets therefore affects the distribution of interarrival times of SPAM emails collected in the mailbox of a single user. We compare the results obtained by modeling botnets as random directed networks, where messages sent sequentially from the botmaster are random walkers, with the distribution of interarrival times of real SPAM data, presented in Ref. [36]. The comparison shows that this simple model reproduces well the basic features of interarrival time distribution.

To better understand the behavior of the distribution of the interarrival times we study the dynamical process of message propagation on different model networks and by two different procedures. We find that the main ingredients controlling the distribution of interarrival times are the distribution of possible path lengths between the sender and the receiver, and the number of messages not reaching the target. Possible paths between two nodes are determined only by the network topology while the mechanisms preventing messages from reaching the target depend in addition on TTL. In the case of random networks, nodes without outputs represent natural traps for the messages, and most of the sent messages are prevented from reaching the target leading to long interarrival times. Since such nodes exist in real networks, this is the situation that we expect to observe in real peer to peer botnets. The distribution of interarrival times for the sequential sending of messages shows an almost constant regime for small interarrival times, followed by an exponential like cutoff including nonvanishing probability for very long interarrival times. When the messages are sent in parallel the interarrival times are power law distributed up to long interarrival times.

We confirm that the nodes without outputs have a crucial role in processes on random networks by studying networks with the same Poisson distribution of links, but without dead ends. In this case interarrival times are much shorter and are distributed as a stretched exponential for sequentially sent messages. When the messages are sent in parallel only trivial values of interarrival times, 0 or 1, appear. Many messages arrive to the target, the space of possible paths is well explored and only those few messages exceeding the limit of maximal number of steps are prevented from reaching the target.

Networks with a scale free distribution of links also show the important role of the network topology for the behavior of the interarrival time distribution. In these networks long interarrival times appear, but the messages are prevented from reaching the target by different mechanisms. Here the limit on the maximal number of steps together with the distribution of possible paths between the sender and the receiver determine the distributions of interarrival times. The interarrival times of sequentially sent messages can be very long in the case of well connected networks, or much shorter for the less connected networks (higher γ values), but the behavior of the distributions are in all cases different than for the other two network types. For messages sent in parallel this distribution is a power law, as in the case of random networks, but with the slope depending on the network's connectivity. From the three types of networks investigated, only for scale free networks do the distributions not collapse onto a universal curve if dt is rescaled by the average rate of the process.

The change in topology has a different influence on the dynamical process on networks with Poisson and scale free distributed links. For scale free networks we change the topology by varying the γ exponent of the link distribution. This change has a strong influence on the distribution of message interarrival times. The average rate is not the only relevant quantity for the process and the universality class of the distribution of interarrival times depends on the topology. Tuning the exponent γ affects the number of hubs in the network, which in scale free networks play a crucial role in the process of message spreading. Their number changes qualitatively the behavior of the interarrival time distribution. It influences the length of possible paths between nodes in the network not only by the change of local properties, such as the number of links of a node, but also by the creation of long range shortcuts through hubs, which increases the number of possible paths between the nodes.

When the link distribution is Poissonian, changing the mean number of links per node modifies the network topology and therefore the mean rate of walker arrivals. The interarrival time distributions, however, collapse onto a universal scaling

function if interarrival time is rescaled by the average rate. The distance between the sender and the receiver is also not affecting the universality class of the universal feature of the distribution. The behavior of the distribution is also robust with respect to changes of other parameters, such as the time distance between two sequential messages, the number of output links of the sender and input links of the receiver, or even the number of senders. The robustness of the behavior of the interarrival time distribution seems to be typical for networks with Poisson distributed links. Interestingly, the SPAM data, analyzed in terms of the junk mailbox, or by the geographical location of IP addresses of the sender, also collapse onto a universal function when interarrival time is rescaled by the average rate. This new approach in studying such processes using network theory can be employed in many fields. By applying this approach to directed networks we show that we can learn about botnets indirectly. The results are not botnet specific and can be applied to any other system which can be modeled by directed networks and through which the information propagates in random fashion.

ACKNOWLEDGMENTS

We acknowledge financial support from the ETH Competence Center "Coping with Crises in Complex Socio-Economic Systems" (CCSS) through ETH Research Grant No. CH1-01-08-2, and financial support from the Swiss National Science Foundation (Grant No. 200021-126853) and FUNCAP.

[1] J. B. Grizzard, V. Sharma, C. Nunnery, B. B. Kang, and D. Dagon, in *HotBots'07: Proceedings of the First Conference on First Workshop on Hot Topics in Understanding Botnets* (USENIX Association, Berkeley, CA, USA, 2007).
[2] N. Bisnik and A. A. Abouzeid, *Proceedings IEEE INFOCOM 2007* (Anchorage, AK, USA, 2007), pp. 517–25.
[3] Y. Li and Z. Zhang, *Proceedings of the INFOCOM, 2010* (San Diego, California, USA, 2010), p. 1.
[4] Z. Bar-Yossef, R. Friedman, and G. Kliot, *Proceedings of the Seventh ACM International Symposium on Mobile ad hoc Networking and Computing, 2006* (MobiHoc, Florence, Italy, 2006), p. 238.
[5] A. Capocci, A. Baldassarri, V. D. P. Servedio, and V. Loreto, *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia, 2009* (Torino, Italy, 2009), p. 239.
[6] B. D. Hughes, *Random Walks and Random Enviroments*, Vol. 1 of Random Walks (Oxford University Press, New York, 1995).
[7] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).
[8] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical Processes in Complex Networks* (Cambridge University Press, Cambridge, UK, 2008).
[9] J. D. Noh and H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004).
[10] L. K. Gallos, Phys. Rev. E **70**, 046116 (2004).
[11] S. Condamin, v. Bénichou, V. Tejedor, J. Voituriez, and J. Klafter, Nature (London) **450**, 77 (2007).
[12] V. Sood, S. Redner, and D. Ben Avraham, J. Phys. A **38**, 109 (2005).
[13] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, Phys. Rev. E **64**, 046135 (2001).
[14] B. J. Kim, C. N. Yoon, S. K. Han, and H. Jeong, Phys. Rev. E **65**, 027103 (2002).
[15] M. Rosvall, P. Minnhagen, and K. Sneppen, Phys. Rev. E **71**, 066111 (2005).
[16] R. Germano and A. P. S. de Moura, Phys. Rev. E **74**, 036117 (2006).
[17] B. Tadic and J. Rodgers, Adv. Complex Syst. **5**, 445 (2002).
[18] S. Jespersen and A. Blumen, Phys. Rev. E **62**, 6270 (2000).
[19] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[20] H. Zhou, Phys. Rev. E **67**, 041908 (2003).
[21] H. Zhou, Phys. Rev. E **67**, 061901 (2003).
[22] V. Sood and P. Grassberger, Phys. Rev. Lett. **99**, 098701 (2007).
[23] Y. Kim, S. W. Son, and H. Jeong, Phys. Rev. E **81**, 016103 (2010).
[24] B. Tadic, Eur. Phys. J. B **23**, 221 (2001).
[25] S. Fortunato and A. Flammini, Internat. J. Bifur. Chaos. **17**, 2343 (2007).
[26] N. Perra, V. Zlatić, A. Chessa, C. Conti, D. Donato, and G. Caldarelli, Europhys. Lett. **88**, 48002 (2009).
[27] P. Bak, K. Christensen, L. Danon, and T. Scanlon, Phys. Rev. Lett. **88**, 178501 (2002).
[28] M. S. Wheatland and Y. E. Litvinenko, Sol. Phys. **211**, 255 (2002).
[29] L. de Arcangelis, C. Godano, E. Lippiello, and M. Nicodemi, Phys. Rev. Lett. **96**, 051102 (2006).
[30] E. Lippiello, L. de Arcangelis, and C. Godano, Astron. Astrophys. **488**, L29 (2008).
[31] E. Lippiello, L. de Arcangelis, and C. Godano, Astron. Astrophys. **511**, L2 (2010).
[32] A. Corral, L. Telesca, and R. Lasaponara, Phys. Rev. E **77**, 016101 (2008).
[33] P. Varga, *EUNICE 2005: Networks and Applications Towards a Ubiquitously Connected World, IFIP International Federation for Information Processing, 2006*, edited by C. Delgado Kloos, A. Marín, and D. Larrabeiti, Vol. 196 (XIV, 2006), p. 17.
[34] J. Zimmermann, A. Clark, G. Mohay, F. Pouget, and M. Dacier, *First International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE 2005)* (IEEE Computer Society, 2005), pp. 89–104.
[35] T. Utsu, in *International Handbook of Earthquake and Engineering Seismology, Part A*, edited by W. H. K. Lee, H. Kanamori, P. C. Jennings, and C. Kisslinger (Academic, Amsterdam, 2002), p. 719.
[36] M. Pica Ciamarra, A. Cognilio, and L. de Arcangelis, Europhys. Lett. **84**, 28004 (2008).
[37] F. Omori, J. Coll. Sci., Imp. Univ. Tokyo **7**, 111 (1894).
[38] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161 (1995).
[39] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).

[40] B. Bollobás, *Random Graphs* (Academic, Orlando, 1985).
[41] A. Corral, Phys. Rev. Lett. **92**, 108501 (2004).
[42] A. Kittas, S. Carmi, S. Havlin, and P. Argyrakis, Europhys. Lett. **84**, 40008 (2008).
[43] S. Redner, *A Guide to First-Passage Processes* (Cambrige University Press, Cambridge, UK, 2001).
[44] D. Ben Avraham and S. Havlin, *Diffusion and Reactions in Fractals and Disordered Systems* (Cambrige University Press, Cambridge, UK, 2000).