The spreading of computer viruses on time-varying networks

Terry Brett,[1] George Loukas,[1] Yamir Moreno,[2,3] and Nicola Perra[1,3,*]

[1]*University of Greenwich, Old Royal Naval College, London, UK*
[2]*Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Zaragoza, Spain*
[3]*ISI Foundation, Turin, Italy*

(Dated: May 10, 2019)

Social networks are the prime channel for the spreading of computer viruses. Yet the study of their propagation neglects the temporal nature of social interactions and the heterogeneity of users' susceptibility. Here, we introduce a theoretical framework that captures both properties. We study two realistic types of viruses propagating on temporal networks featuring Q categories of susceptibility and derive analytically the invasion threshold. We found that the temporal coupling of categories might increase the fragility of the system to cyber threats. Our results show that networks' dynamics and their interplay with users features are crucial for the spreading of computer viruses.

Alongside clear societal and economic benefits, modern technology exposes us to serious challenges. In particular, the spreading of malicious content online, often based on ingenious deception strategies, is one of the most pressing because it poses serious threats to our privacy, finances, and safety [1]. Victims of a typical *social engineering attack* [2] may receive a message containing a malicious link or file, appearing to originate from a friend or other trusted entity. If opened, it may compromise the computer, access personal information, and spread the virus further unbeknownst to the victim. Recent research has shown how the susceptibility of individuals to such attacks is not homogenous and depends on several features such as age, prior training, computer proficiency, familiarity with social network platforms, among others [3–5]. Furthermore, the properties of real networks are known to facilitate the propagation of such processes [6–15]. In particular, the heterogeneity in contact patterns makes socio-technical systems quite fragile to biological and digital threats.

The study of these phenomena has largely neglected the complex temporal nature of online contact patterns in favor of static and time-aggregated approaches [16, 17]. These approximations might be fitting. Indeed, in the past, computer viruses would spread mainly via email networks, targeting the address books of victims, which contain contacts lists [18]. However, not many people create such lists any more and access to them is restricted [7]. In the context of social or biological contagions, neglecting the temporal nature of the networks where the processes unfold has been shown to induce misrepresentations of their spreading potential. In fact, the order and concurrency of connections is key [19–43]. To the best of our knowledge, beside some early work on the spreading of viruses via Bluetooth among mobile phones [44], the study of the propagation of cyber threats considering the temporal nature of social interactions is still missing. Furthermore, with few exceptions [45], the literature devoted to the study of computer viruses unfolding on networks typically neglects that the susceptibility of online users is not homogenous. Conversely, the literature that studies the susceptibility of users to cyber threats traditionally focuses on single users neglecting their connections.

To tackle these limitations, here we introduce a theoretical framework to study the spreading of computer viruses, based on social engineering deception strategies, on time-varying networks. We model users' interactions using a time-varying network model and consider two types of viruses. The first mimics threats that can propagate only via connections activated at each time step. The second, on the contrary, considers viruses able to access also information about past connections. We investigate the impact of different classes of susceptibility considering that they might also influence the link formation process. In all cases, we analytically derive the conditions regulating the spreading of the virus. Interestingly, these are defined by the interplay between the features of the cyber threats, the categories of susceptibility and their time-varying connectivity. Furthermore, in some scenarios, the temporal coupling between categories creates a complex phenomenology that favors the spreading of the virus. These results have the potential to initiate future efforts aimed at describing more realistically the spreading of computer viruses on online social networks.

We consider a population of N online users which exchange messages in a time-varying network. Nodes are assigned to one of Q categories describing their susceptibility to cyber threats measured in terms of their *gullibility* and time needed to recover from successful attacks. Since susceptibility is linked to demographic features, we consider that the membership to a category might influence the link creation process. In fact, homophily is a strong social mechanism known to affect the structure and organization of ties [46]. We model the contact patterns between users with a generalization of the activity-driven framework [21, 47–49]. Here, nodes feature an activity a describing their propensity to initiate communications. Activities are extracted from a distribution $F(a)$ which, as observations in real systems have shown, is typically heterogenous [21, 22, 48, 50]. We select power-law distributions $F(a) \sim a^{-\alpha}$ with $a \in [\epsilon, 1]$ to avoid divergences. At each time step nodes are active with probability $a\Delta t$. Active nodes select m others and create directed (outgoing) links which mimic messages.

In the simplest version of activity-driven networks the selection is random and memoryless [21]. Here, we propose

* n.perra@greenwich.ac.uk

a variation: with probability p each target is selected, at random, among the group of nodes in the same category, and with probability $1-p$ among the nodes in any other category. In other words, p tunes the homophily level in the network with respect to susceptibility to cyber threats. At time $t+\Delta t$ all edges are deleted and the process starts from the beginning. Unless specified otherwise, links have a duration Δt. Without loss of generality we set $\Delta t = 1$. The model is clearly a simplification of real interactions. However, it offers simple, yet non trivial, settings to study the effects of temporal connectivity patterns on contagion processes unfolding at a comparable time-scale with respect to the evolution of connections [20, 21, 47, 51].

We describe the propagation of a computer virus adopting the prototypical SIS model [13, 52]. At each time step t the virus, unbeknownst to the victims, sends a message, with malicious content, to all the nodes genuinely contacted at t (virus type 1) or within $t-\tau$ time-steps (virus type 2). The focus is not defining the optimal set of nodes to maximize/minimize the damage. Thus, we select randomly a small percentage (0.5%) of nodes as initial seeds. In these settings, susceptible nodes of class $x \in [1,\dots,Q]$, that receive a malicious message, become infectious with probability λ_x which defines their gullibility. They recover and become susceptible again with rate μ_x. In the literature of epidemic spreading on static networks we find few studies that consider different classes of infectiousness and/or recovery rates [53–55]. Interestingly, this body of research highlights how heterogeneities in such quantities, especially in case of correlations with topological features such as the degree or in presence of large values of clustering, induce no trivial phenomena that might speed up or slow down the spreading. As shown below, our results confirm this picture. We assume that nodes with the same value of activity and in the same category are statistically equivalent, we group them according to the two features. At each time step, we call S_a^x and I_a^x the number of nodes susceptible and infected in activity class a and category x. Clearly $\int da S_a^x = S^x$, $\int da I_a^x = I^x$, $\sum_x S^x = S$, and $\sum_x I^x = I$. Furthermore, N_a^x describes the number of nodes of activity a in category x, thus $\int da N_a^x = N^x$ and $\sum_x N^x = N$. In these settings, we can represent the variation of the number of infected nodes of activity a in category x as:

$$d_t I_a^x = -\mu I_a^x + \lambda_x m S_a^x \times \left[p \int da' a' \frac{I_{a'}^x}{N^x} + (1-p) \sum_{y \neq x} \int da' a' \frac{I_{a'}^y}{N - N^y} \right]. \quad (1)$$

The first term on the right hand side accounts for the recovery process. The second and third terms capture susceptible nodes that receive messages from active and infected vertices in the same (second) or different (third) category, and get infected as a result. With respect to the typical biological contagion process, here transmission is asymmetric. Only nodes receiving a message from an infected person might be exposed to the virus. Thus, not only the order of connections, but also their direction is a crucial ingredient for the spreading. Since the links are created randomly, each node is selected with a probability pm/N^x by nodes in the same category or $(1-p)m/(N-N^y)$ by nodes in other categories. The total number of nodes is constant thus $S_a^x = N_a^x - I_a^x$ and at the early stages of the spreading we can assume that the number of infected nodes is very small: $S_a^x \sim N_a^x$. By integrating across all activities Eq. 1 we get:

$$d_t I^x = -\mu_x I^x + \lambda_x m \left[p\theta^x + (1-p) N^x \sum_{y \neq x} \theta^y / (N - N^y) \right],$$

where we define $\theta^x = \int da a I_a^x$. By multiplying both sides of Eq. 1 for a and integrating across all the activities we obtain

$$d_t \theta^x = -\mu_x \theta^x + m\lambda_x \langle a \rangle_x \left[p\theta^x + (1-p) N^x \sum_{y \neq x} \theta^y / (N - N^y) \right].$$

The virus is able to spread, if and only if the largest eigenvalue of the Jacobian matrix of the system of differential equations in I^x and θ^x is larger than zero [21]. As shown in details in the Supplementary Material (SM) [56] this implies:

$$R_0 = \frac{p \sum_x \beta_x + \Xi}{\sum_x \mu_x} > 1, \quad (2)$$

where R_0 is the basic reproductive number defined as the average number of infected nodes generated, in a fully susceptible population, by an infected individual [52], $\beta_x = m\lambda_x \langle a \rangle_x$ and Ξ is a function of the interplay between the average activation, infection and recovery rate of each category as well as of the mixing between categories.

To understand the dynamics, let us consider a particular case in which the system is characterized by only two categories. Furthermore, let us consider, as first scenario, that all nodes have the same recovery rate. In these settings we have $\Xi^2 = p^2(\beta_1+\beta_2)^2 + 4\beta_1\beta_2(1-2p)$. The condition for the spreading, even with only two classes, is a non linear function of the average activity of each category, the infection probabilities per contact and the homophily. In the limit $p = 0$, nodes in a category connects only with vertices in the other and the expression reduces to $R_0 = \frac{\sqrt{\beta_1\beta_2}}{\mu}$. In the limit $p = 1$ instead, interactions are only between nodes in the same category. The system is effectively split in two disconnected networks and there are two independent conditions $R_0^x = \beta_x/\mu$. For a general p we found that these two values confine R_0: $\min_x R_0^x \leq R_0(p) \leq \max_x R_0^x$. In fact, any value of $p < 1$ will reduce the spreading power of the category characterized with the largest R_0^x as some connections will be established with nodes where the virus finds it harder to spread (see SM for the proof).

In Fig. 1-A-C, we compare analytical predictions with numerical simulations. We set $\lambda_2 = 0.2$ and use Eq. 2 to estimate the critical value of λ_1 for which $R_0 \equiv 1$. On the y-axis we plot the lifetime of the process defined as the time that the virus needs either to die out or to reach a fraction Y of the population [57]. The lifetime acts as the susceptibility of a second order phase transition and allows a precise numerical estimation of the threshold of SIS processes [57]. In panels A-B we

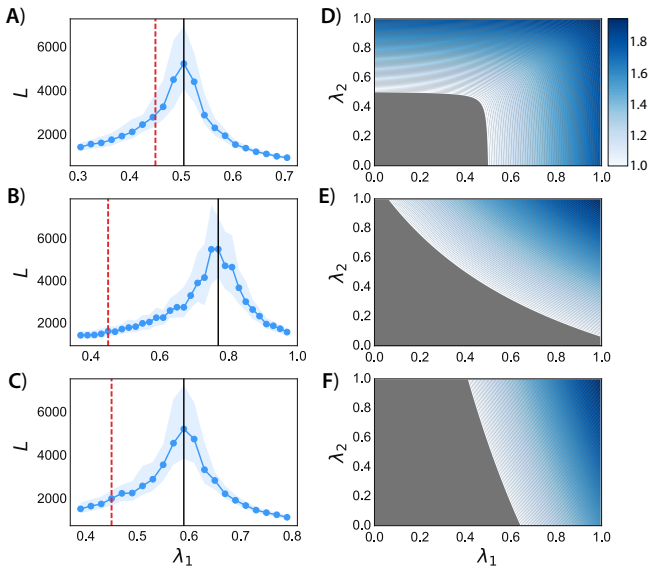

FIG. 1. Lifetime of the SIS process (A-C) and contour plot of $R_0(\lambda_1, \lambda_2)$ (D-F). In A-B-D-E nodes are randomly assigned to two categories, in C-F instead in decreasing order of activity. We set $p = 0.9$ (A-D), $p = 0.4$ (B-C-E-F). In A-C we fix $N = 2 \times 10^5$, $m = 4$, $\alpha = 2.1$, $\mu_1 = \mu_2 = 10^{-2}$, $\lambda_2 = 0.2$, $Y = 0.3$, and 0.5% of random initial seeds. We plot the median and 50% confidence intervals in 10^2 simulations per point. The solid lines come from Eq. 2, and the dashed lines are the analytical threshold in case of a single category.

consider a scenario in which nodes are assigned randomly to one of the two categories. Thus the average activity in the two is the same and set $p = 0.9$ and $p = 0.4$ respectively. The analytical value of the threshold (vertical solid line) perfectly matches the numerical estimation. For $p = 0.9$ the threshold is smaller than for $p = 0.4$ and closer to the threshold of a system with a single category (dashed lines). For smaller values of homophily, instead, the critical conditions are driven by the interplay between the activation rates and gullibility of the two categories. Panels D-E show the analytical value of R_0 as a function of λ_1 and λ_2 for the two values of p. The grey regions are sub-critical, i.e., the virus is not able to spread. Since the average activity in the two categories is the same, the two plots are symmetric. Interestingly, the region where the virus is able to spread is larger for large values of p. This is due to the fact that in these settings the virus will spread if above the threshold in at least one category independently of the other. In the opposite limit, on the contrary, the two categories get intertwined and a small value of the infection probability in one category should be associated to a progressively large value in the other.

In panels C-F we consider that the first category contains a fraction g of nodes selected in decreasing order of activity. Thus, this category contains the gN most active nodes, while the other the $(1-g)N$ least active (see SM). To compare with panel B, we set $g = 0.5$ and $p = 0.4$. First, the analytical threshold nicely matches the numerical simulations. Second, although the other parameters are the same used in panel B, the critical value of the gullibility of the first class is smaller. Thus, correlations between activity and gullibility facilitate the spreading. This is confirmed in panel F where the active phase space features a region in which the spreading is completely dominated by the category of most active nodes. Overall, all the plots show the importance of distinguishing nodes according to their gullibility. Indeed, neglecting the presence of different classes of users might induce a strong misrepresentation of the virus propagation (dashed lines).

Let us next consider a second scenario where categories differentiate also for the time needed to recover from a successful attack. For two categories, we can write $\Xi^2 = (\mu_1 - \mu_2)^2 + p^2(\beta_1 + \beta_2)^2 + 2p(\mu_2 - \mu_1)(\beta_1 - \beta_2) + 4\beta_1\beta_2(1-2p)$. Interestingly, we have the same terms that appeared in the first scenario, plus two that feature the difference between the recovery rates and βs of the two categories. Thus R_0 is a function of the interplay between the activities, gullibilities and recovery rates. In the limit $p = 0$, each category only connects with nodes in the other, the two groups are coupled and the threshold reads $R_0 = \frac{\sqrt{(\mu_1-\mu_2)^2+4\beta_1\beta_2}}{\mu_1+\mu_2}$. In the limit $p = 1$ instead, the two categories are completely decoupled and the threshold becomes, as before, $R_0 = \beta_x/\mu_x$. As shown in Fig. 2-A-B, for a general value of p the reproductive number is not bounded, as before, by the values of R_0^x computed in the two classes separately (see SM). In Fig. 2-A, we assign nodes randomly to each category, fix β_x and μ_x and compute R_0 as a function of p. In the shaded area $\min_x R_0^x \leq R_0(p) \leq \max_x R_0^x$. Interestingly, after a p^* (vertical dashed line), which as shown in the SM can be computed analytically, we enter in a regime where $R_0(p) > \max_x R_0^x$. Thus, only specific values of the coupling between categories might induce the virus to spread faster in the combined system than in each single category in isolation. However, this non linear effect is found only in a small fraction of the phase space see Fig. 2-B. The necessary, but not sufficient condition, is that two categories differentiate both for gullibility and recovery rates in such a way that one is more gullible and recovers faster than the other. In this regime, the right mixing between the two might create a feedback loop that makes the system more fragile.

Fig. 3-A-C shows a good match between the analytical (solid vertical lines) and numerical thresholds in case of nodes are assigned at random (A-B) or in decreasing order of activity (C) to the two categories. We fix two different recovery rates, λ_2, and use λ_1 as order parameter. Panels A-B-C differ in the value of the homophily p. We set $p = 0.9$ in A, while $p = 0.4$ in B-C. The presence of a category of nodes characterized by a smaller value of recovery rate pushes the threshold to smaller values with respect to the first scenarios (Fig. 1). As before, the value of the threshold estimated considering only a single category, characterized by the average recovery rate of the two, (dashed lines) leads to a misrepresentation of the spreading power of the virus, especially for smaller values of homophily (see panel B).

The effect of p on the critical value of λ_1 is similar to the first scenario. In fact, even when categories differentiate by the recovery rates, high values of homophily push the critical point to smaller values. However, here the difference between

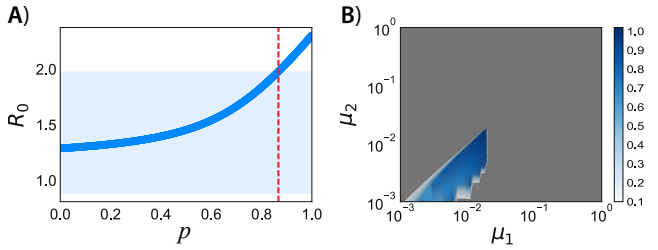

FIG. 2. In A we plot the analytical value of R_0 as function of p. The shaded area describes the region where $\min_x \beta_x/\mu_x \leq R_0 \leq \max_x \beta_x/\mu_x$. The dashed vertical line describes the analytical value of p above which $R_0 > \max_x \beta_x/\mu_x$. We set $\mu_1 = 10^{-2}$ and $\mu_2 = 5 \times 10^{-3}$. In B we plot p^* as function of μ_1 and μ_2. In both plots, we set $m = 4$, $\lambda_1 = 0.9$, $\lambda_2 = 0.2$, and randomly assign nodes to two categories.

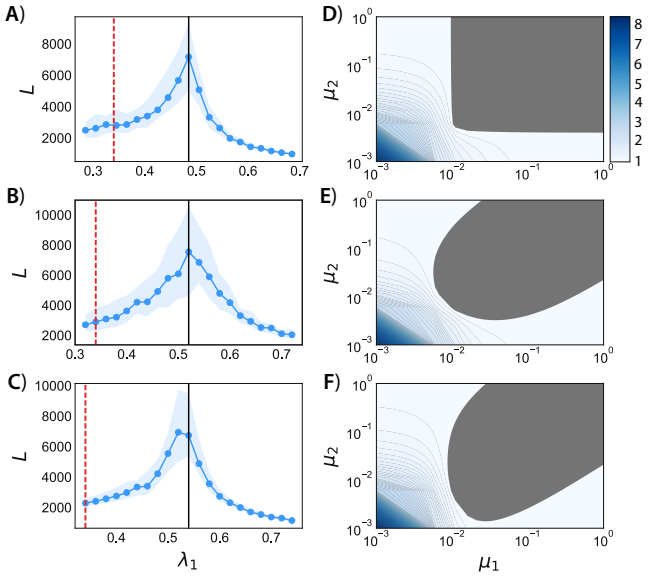

FIG. 3. Lifetime of the process (A-C), $R_0(\mu_1, \mu_2)$ (D-F). In A-B-D-E nodes are randomly assigned to two categories, in C-F instead in decreasing order of activity. We set $p = 0.9$ (A-D), $p = 0.4$ (B-C-E-F). In panels A-C we set $N = 2 \times 10^5$, $m = 4$, $\alpha = 2.1$, $\mu_1 = 10^{-2}$, $\mu_2 = 5 \times 10^{-3}$, $\lambda_2 = 0.2$, $Y = 0.3$, and 0.5% randomly selected seeds. We plot the median and 50% confidence intervals in 10^2 simulations per point. The solid lines come from Eq. 2. The dashed lines are the analytical threshold in case of a single category of recovery rate characterized by the average value of the recovery rates. In the contour plot we set $\lambda_1 = 0.485$ and $\lambda_2 = 0.2$.

the two is less significant than in Fig. 1. In Fig. 3-D-E, we show the analytical value of R_0 as function of μ_1 and μ_2. Interestingly, the sub-critical region, for $p = 0.4$, is smaller than for $p = 0.9$. This is in contrast to what was observed in the corresponding plots for the first scenario and highlights once again the complex phenomenology introduced by the interplay of different recovery rates. In Fig. 3-C-F we investigate a scenario where nodes are assigned to categories of susceptibility in decreasing order of activity. In case the most active nodes are able to recover quickly from the attack, the virus is able to spread only if the gullibility of such users is higher than in the corresponding case in which nodes are assigned to categories randomly (panel B). This is confirmed in panel F, where we see that partitioning nodes according to their activities significantly change the region where the threat is able to spread.

Finally, we turn our attention to a second type of virus able to access also past contacts of infected users within a time window τ. As before, the virus propagates via active infected nodes, but at each time t active users might infect their contacts in a time-window $(t-\tau, t]$. Within a mean-field approximation, we can adopt the same equations described above and change the probability that a node in each activity class receives a message by active and infected nodes. In this case, the out-degree of each active node is not m, but a function of τ: $k^{out}(a) = m\left[a + (\tau - 1)a^2\right]$ (see SM). To grasp the derivation, consider the simplest scenario in which $\tau = 2$. In this case, active nodes might have either m or $2m$ contacts in two time steps. The first class describes nodes that are active at time t but were not active at time $t-1$; whereas the second, nodes that were active in both time steps. Thus the out-degree of these nodes, on average, is $k^{out}(a) = ma(1-a) + 2ma^2$. As shown in the SM, the condition for the spreading has the same structure of Eq. 2 where, however, the value of βs are changed with the following transformation $m \to m\left[\langle a\rangle + (\tau - 1)\langle a^2\rangle\right]$. Thus, the larger the visibility of past connections, from the virus point of view, the larger R_0. Intuitively this is due to the fact that the virus, for large values of τ, is able to access more contacts, which results in a larger spreading potential. This observation nicely shows how neglecting the temporal nature of connectivity patterns in favor of static (or time integrated) approximations might lead to a poor description of the propagation of viruses that do not have access to contacts lists or past connections. In Fig. 4 we show the comparison between analytical (solid lines) and numerical values of the threshold for different values of τ. To isolate the effect of τ we considered two categories, a single recovery rate, and set $p = 0.5$. The analytical value is a good approximation only for small values of τ. The mean-field approximation becomes less accurate as more connections from past time-steps are kept in memory. Thus, the analytical estimation provides only a lower bound, which together with the solution for $\tau = 1$ (dashed lines) −that constitutes an upper bound−, marks the region containing the epidemic threshold (red regions). In other words, for a general value of τ, the threshold will be lower than the analytical value computed for $\tau = 1$, and larger than the corresponding value computed at τ.

Overall our results highlight how the spreading of computer viruses based on social engineering is critically affected by the temporal nature of our interactions and different susceptibilities to cyber threats. Our findings show that networks' dynamics and their interplay with the characteristics of users have to be considered in order to avoid misrepresentation of the spreading power of computer viruses in social networks. We have also quantified the extent to which the previous mismatch is important for three plausible scenarios. We, how-

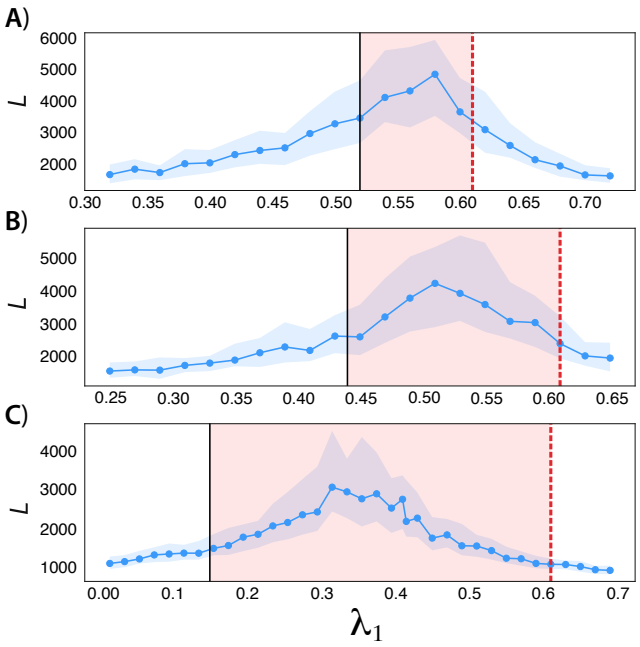

FIG. 4. Lifetime of the SIS process for $\tau = 2, 3, 10$ (A,B,C) for two categories to which nodes are assigned randomly. Simulations are done setting $N = 2 \times 10^5$, $m = 4$, $\alpha = 2.1$, $Y = 0.3$, $\mu = 10^{-2}$, $\lambda_2 = 0.3$, $p = 0.5$, and 0.5% random initial seeds. We plot the median and 50% confidence intervals in 10^2 simulations per point.

ever, note that we have studied a simple network model that neglects a range of properties of real social networks such as the presence of weak and strong ties, high order correlations, and community structures. The study of the impact of these features on the unfolding of computer viruses calls for additional research.

This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF-18-1-0376. Y. M. acknowledges support from the Government of Aragón, Spain through grant E36-17R (FENOL) and by MINECO and FEDER funds (grant FIS2017-87519-P). The authors thanks Andrea Baronchelli and Michele Starnini for useful discussions.

[1] I. Kayes and A. Iamnitchi, Online Social Networks and Media **3**, 1 (2017).
[2] R. Heartfield and G. Loukas, ACM Computing Surveys (CSUR) **48**, 37 (2016).
[3] R. Heartfield and G. Loukas, Computers & Security **76**, 101 (2018).
[4] R. Heartfield, G. Loukas, and D. Gan, in *IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)* (IEEE, 2017) pp. 371–378.
[5] R. Heartfield, G. Loukas, and D. Gan, IEEE Access **4**, 6910 (2016).
[6] A. L. Lloyd and R. M. May, Science **292**, 1316 (2001).
[7] J. Balthrop, S. Forrest, M. E. Newman, and M. M. Williamson, Science **304**, 527 (2004).
[8] R. Pastor-Satorras and A. Vespignani, Phys. Rev, Lett. **86**, 3200 (2001).
[9] Y. Moreno and A. Vázquez, Eur. Phys. J. **31**, 265 (2003).
[10] M. E. J. Newman, Phys. Rev. E **66**, 016128 (2002).
[11] M. Newman, *Networks. An Introduction* (Oxford Univesity Press, 2010).
[12] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Reviews of Modern Physics **87**, 925 (2015).
[13] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks* (Cambridge Univesity Press, 2008).
[14] L.-X. Yang, X. Yang, J. Liu, Q. Zhu, and C. Gan, Applied Mathematics and Computation **219**, 8705 (2013).
[15] L.-X. Yang and X. Yang, Physica A: Statistical Mechanics and Its Applications **396**, 173 (2014).
[16] P. Holme, The European Physical Journal B **88**, 1 (2015).
[17] P. Holme and J. Saramäki, Physics Reports **519**, 97 (2012).
[18] M. E. Newman, S. Forrest, and J. Balthrop, Physical Review E **66**, 035101 (2002).
[19] A. Barrat and C. Cattuto, in *Social Phenomena* (Springer International Publishing, 2015) pp. 37–57.
[20] N. Perra, A. Baronchelli, D. Mocanu, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani, Physical Review Letter **109**, 238701 (2012).
[21] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani, Scientific Reports **2**, 469 (2012).
[22] B. Ribeiro, N. Perra, and A. Baronchelli, Scientific Reports **3**, 3006 (2013).
[23] S. Liu, N. Perra, M. Karsai, and A. Vespignani, Physical Review Letters **112**, 118702 (2014).
[24] S.-Y. Liu, A. Baronchelli, and N. Perra, Physical Review E **87**, 032805 (2013).
[25] G. Ren and X. Wang, Chaos: An Interdisciplinary Journal of Nonlinear Science **24**, 023116 (2014).
[26] M. Starnini, A. Machens, C. Cattuto, A. Barrat, and R. Pastor-Satorras, Journal of Theoretical Biology **337**, 89 (2013).
[27] M. Starnini, A. Baronchelli, A. Barrat, and R. Pastor-Satorras, Physical Review E **85**, 056115 (2012).
[28] E. Valdano, L. Ferreri, C. Poletto, and V. Colizza, Physical Review X **5**, 021005 (2015).
[29] I. Scholtes, N. Wider, R. Pfitzner, A. Garas, C. Tessone, and F. Schweitzer, Nature Communications **5**, 5024 (2014).
[30] M. J. Williams and M. Musolesi, Royal Society Open Science **3**, 160196 (2016).

[31] L. E. Rocha and N. Masuda, New Journal of Physics **16**, 063023 (2014).

[32] T. Takaguchi, N. Sato, K. Yano, and N. Masuda, New Journal of Physics **14**, 093003 (2012).

[33] L. E. Rocha and V. D. Blondel, PLoS computational biology **9**, e1002974 (2013).

[34] G. Ghoshal and P. Holme, Physica A: Statistical Mechanics and its Applications **364**, 603 (2006).

[35] K. Sun, A. Baronchelli, and N. Perra, The European Physical Journal B **88**, 1 (2015).

[36] D. Mistry, Q. Zhang, N. Perra, and A. Baronchelli, Physical Review E **92**, 042805 (2015).

[37] R. Pfitzner, I. Scholtes, A. Garas, C. Tessone, and F. Schweitzer, Physical Review Letter **110**, 19 (2013).

[38] T. Takaguchi, N. Sato, K. Yano, and N. Masuda, New Journal of Physics **14**, 093003 (2012).

[39] T. Takaguchi, N. Masuda, and P. Holme, PloS one **8**, e68629 (2013).

[40] P. Holme and F. Liljeros, Scientific Reports **4**, 4999 (2014).

[41] P. Holme and N. Masuda, PloS one **10**, e0120567 (2015).

[42] Z. Wang, C. T. Bauch, S. Bhattacharyya, A. d'Onofrio, P. Manfredi, M. Perc, N. Perra, M. Salathé, and D. Zhao, Physics Reports **664**, 1 (2016).

[43] B. Gonçalves and N. Perra, *Social phenomena: From data analysis to models* (Springer, 2015).

[44] P. Wang, M. C. González, C. A. Hidalgo, and A.-L. Barabási, Science **324**, 1071 (2009).

[45] S. Peng, G. Wang, Y. Zhou, C. Wan, C. Wang, and S. Yu, IEEE Transactions on Dependable and Secure Computing (2017).

[46] M. McPherson, L. Smith-Lovin, and J. M. Cook, Annual review of sociology **27**, 415 (2001).

[47] M. Karsai, N. Perra, and A. Vespignani, Scientific Reports **4**, 4001 (2014).

[48] E. Ubaldi, N. Perra, M. Karsai, A. Vezzani, R. Burioni, and A. Vespignani, Scientific Reports **6**, 35724 (2016).

[49] M. Tizzani, S. Lenti, E. Ubaldi, A. Vezzani, C. Castellano, and R. Burioni, Physical Review E **98**, 062315 (2018).

[50] M. Tomasello, N. Perra, C. Tessone, M. Karsai, and F. Schweitzer, Scientific Reports **4**, 5679 (2014).

[51] S. Liu, N. Perra, M. Karsai, and A. Vespignani, Physical review letters **112**, 118702 (2014).

[52] M. Keeling and P. Rohani, *Modeling Infectious Disease in Humans and Animals* (Princeton University Press, 2008).

[53] D. Smilkov, C. A. Hidalgo, and L. Kocarev, Scientific reports **4**, 4795 (2014).

[54] J. C. Miller, Journal of the Royal Society Interface **6**, 1121 (2009).

[55] W. Gou and Z. Jin, Infectious Disease Modelling **2**, 353 (2017).

[56] "See supplemental material at [url] for detailed information about the mathematical derivation and sensitivity analysis of the results to the variation of the main parameters.".

[57] M. Boguña, C. Castellano, and R. Pastor-Satorras, Physical Review Letter **111**, 068701 (2013).