Controlling the spread of deception-based cyber-threats on time-varying networks

Nicolò Gozzi [1] and Nicola Perra [2,3,*]

[1]*ISI Foundation, Turin, Italy*
[2]*School of Mathematical Science, Queen Mary University of London, London, United Kingdom*
[3]*The Alan Turing Institute, London, United Kingdom*

We study the efficacy of strategies aimed at controlling the spread of deception-based cyber-threats unfolding on online social networks. We model directed and temporal interactions between users using a family of activity-driven networks featuring tunable homophily levels among gullibility classes. We simulate the spreading of cyber-threats using classic susceptible-infected-susceptible (SIS) models. We explore and quantify the effectiveness of four control strategies. Akin to vaccination campaigns with a limited budget, each strategy selects a fraction of nodes with the aim to increase their awareness and provide protection from cyber-threats. The first strategy picks nodes randomly. The second assumes global knowledge of the system selecting nodes based on their activity. The third picks nodes via egocentric sampling. The fourth selects nodes based on the outcome of standard security awareness tests, customarily used by institutions to probe, estimate, and raise the awareness of their workforce. We quantify the impact of each strategy by deriving analytically how they affect the spreading threshold. Analytical expressions are validated via large-scale numerical simulations. Interestingly, we find that targeted strategies, focusing on key features of the population such as the activity, are extremely effective. Egocentric sampling strategies, though not as effective, emerge as a clear second best despite not assuming any knowledge about the system. In addition, we find that networks characterized by highly homophilic interactions linked to gullibility might expand the range of transmissibility parameters that allows for macroscopic outbreaks. At the same time, they reduce the reach of these spreading events. Hence, rather isolated patches of the network formed by highly gullible individuals might provide fertile grounds for the propagation and survival of cyber-threats.

Deception-based attacks such as phishing, baiting, and file masquerading have become one of the most diffuse types of cyber-threats [1–8]. These are designed around ingenious strategies that target human nature. For example, the classic phishing scheme tries, via trusted access, to lure victims to open a malicious link and/or to download a malware which might affect personal accounts and devices, reveal sensitive data, and, unbeknownst to the victims, allow the threat to spread further. Worryingly, the recent advancements in generative artificial intelligence are offering new unprecedented opportunities to malicious actors to enhance and scale up their criminal activities [9,10].

The extant literature devoted to modeling these processes and gathering insights to contrast them is vast, but presents two main limitations. First, most previous work neglects the temporal nature of online contact patterns, focusing instead on aggregated networks [11–15]. However, the order and concurrency of interactions are key factors shaping the characteristics of a wide range of spreading processes on networks.

As extensively shown in the case of random walks [16–20], biological [21–34] and social contagion phenomena [35–37] neglecting time-varying patterns in favor of static (i.e., aggregated) representations may lead to a misrepresentation of the spreading potential of these processes. Indeed, as not all contacts might be active at the same time, the possible routes a random walker can take or the possible number of infection chains could be reduced with respect to those available in time-aggregated representations [16–18,21,22,26,37]. Even more, all contacts a node has before it is infectious or is visited by a random walker cannot lead to a further propagation of the process [11,12]. At the same time, it is interesting to note how the complex nature of interactions, in particular possible correlations in activation patterns, could also lead to a speed up of such processes [19,38]. In the context of cyber-threats, a few papers have highlighted the importance of accounting for temporal interactions. Reference [39] modeled the spreading of computer viruses on time-varying networks featuring homophily. The paper shows how the interplay between timescales (one describing the evolution of connectivity patterns and the other the spreading process) might induce highly nontrivial dynamics. Reference [40] studied the spreading of computer viruses via temporal Bluetooth connections. The results highlight how the diffusion of these threats is slowed down by the actual, fewer, opportunities of a contagion with respect to those that would be erroneously considered available in static aggregations of users' contacts.

---

*Contact author: n.perra@qmul.ac.uk

As a second limitation, most of the literature assumes users to be equally susceptible (i.e., gullible). However, recent empirical studies showed that susceptibility to cyber-threats is not homogeneous and may depend on factors such as age, digital proficiency, or familiarity with online social networks, among others [5,41]. Reference [39] is an exception in the literature, as it explicitly accounts for different gullibility classes. The paper shows how interactions across different susceptibility groups might render the system more fragile to the spreading of cyber-threats with respect to homogeneous cases. Hence, considering the presence of such heterogeneities might be critical to gather a better picture of systems' resilience against cyber-threats.

Here, we build on the theoretical framework developed in Ref. [39] and expand it to investigate the effectiveness of different strategies devoted to controlling the spread of deception-based attacks. To this end, we imagine a large corporation or institution facing the challenge of protecting their digital infrastructure against cyber-threats. Following Ref. [39], we model the temporal interactions between users in the corporation adopting a family of activity-driven networks [22,37,42,43] featuring different gullibility classes and tunable homophily levels between them. We model the potential spreading of deception-based cyber-threats within the corporation considering a classic susceptible-infected-susceptible (SIS) model [44].

In these settings, we assume that, in a given time window, the corporation has a limited budget to increase the awareness of their employees via specific training courses, and it faces the following question: Who should be selected for such training? Defining a strategy to select a fraction of employees to better protect a corporation from computer viruses can be formulated as a vaccination problem with limited budget [45]. We investigate the impact of four different strategies. The first acts as a baseline and selects a random fraction of the nodes for training independently from their features. The second focuses on targeting nodes that, possibly due to the nature of their work (e.g., customer service), tend to get in contact with more individuals over time (i.e., they are more active) [23]. As a third option, we consider a strategy based on local egocentric sampling of the connections of a fraction of nodes that act as probes [15,23]. Finally, we study a fourth strategy that targets users based on their knowledge of cyber-threats estimated via prototypical awareness tests (e.g., simulated phishing campaigns) [46].

We derive closed analytical expressions of the epidemic threshold, providing insights about the impact of users' features in all four immunization strategies. We quantify the effectiveness of each strategy via large-scale numerical simulations, which (i) validate the analytical solutions derived and (ii) allow the characterization of the dynamics under consideration. Overall, we find a clear hierarchy among strategies in terms of their effectiveness. The activity-based strategy emerges as the most effective. Indeed, selecting nodes based on their activities might reduce the needed fraction of users to halt the spreading by more than one order of magnitude with respect to other strategies. The egocentric strategy emerges consistently as second best, while the strategy based on security awareness tests is only marginally better than the baseline. Interestingly, we find that highly homophilic interactions

among gullibility classes increase the range of transmissibility parameters that might result in macroscopic outbreaks, but at the same time they reduce the reach of cyber-threats confining them within the most gullible group. This result highlights how cyber-threats might survive in rather isolated parts of networks even if they are not able to spread in most of the others. Identifying these possible breeding grounds might be crucial for elimination campaigns.

The paper is organized as follows. In Sec. I we describe the general structure of the model. In Sec. II we describe the four strategies, and we present the analytical results and the numerical simulations. In Sec. III we present our conclusions.

I. THE MODEL

In this section, we summarize the main features of the model that acts as the building block of our study. As mentioned, we build on the framework proposed in Ref. [39]. We consider a population of N users that exchange directed messages online. They are divided into Q categories describing their susceptibility to cyber-threats (i.e., gullibility classes). In our hypothetical scenario, categories might be linked to the organizational structure of the corporation and describe different departments formed by people with similar computer proficiency. Each node features an activity a describing their propensity to initiate communications per unit time. Activities are extracted from a power-law distribution $F(a) \sim a^{-\alpha}$ with $a \in [\epsilon, 1]$. Following Ref. [39], we assume the same distribution of activity across all gullibility classes, i.e., $F(a) = F_x(a)$ $\forall x$. This choice allows us to simplify the model in terms of the number of parameters. Furthermore, it allows us to quantify the impact of different levels of gullibility on the spreading process without having to account for other complex dynamics that might emerge from the interplay of different activity distributions. The temporal dynamics regulating the interactions between users are the following:

(i) At each time step, with probability $a\Delta t$, nodes are active.

(ii) Active nodes select m others and send them a message. The selection is driven by a parameter p, which regulates the homophily in the system: with probability p each active node selects at random another user within the same gullibility class. With probability $(1 - p)$, instead, the active user sends a message to a user randomly picked among the other classes. In other words, the membership to a category influences the link creation process [39,47]. Indeed, in our scenario, we can imagine that people tend to interact more with others in the same department.

(iii) At the end of each time step, all connections are deleted and the process restarts.

Without loss of generality, we set $\Delta t = 1$.

We describe the propagation of cyber-threats using an SIS model [44]. Hence, users might be compromised (i.e., infected) and recover as susceptible. Imagine a user who falls for the ruse at time t and realizes they have been compromised taking actions to regain full control of their computer at time t'. The threat will attempt to spread further by covertly sending malicious content to all users legitimately contacted between t and t'. In other words, we assume that cyber-threats do not have access to a user's entire communication history. Instead,

they can only piggyback on messages initiated after the user has been compromised and before the threat is detected [14]. Each of the contacted users will be infected, thus falling for the ruse, with a probability λ_x, and recover, becoming susceptible again, at a rate μ_x. As before, x denotes the gullibility class. We stress the asymmetry in the transmission process: an infection can only occur when an infected user contacts a susceptible user, and not the opposite. This asymmetry is the key difference with respect to similar models for biological contagion processes.

As shown in Ref. [39], in these settings it is possible to compute an analytical expression for the basic reproduction number (i.e., R_0) of a cyber-threat unfolding in the system. R_0 is defined as the average number of secondary infections generated by a single compromised account in an otherwise susceptible population [44]. The expression for R_0 reads

$$R_0 = \frac{p \sum_x \beta_x + \Xi}{\sum_x \mu_x}, \tag{1}$$

where $\beta_x = m\lambda_x \langle a \rangle_x$, and Ξ is a function of β_x, μ_x, and $c_{x,y}$ (i.e., the mixing probabilities among different gullibility classes) and has a specific algebraic expression for a fixed number of classes Q. The quantity $\langle a \rangle_x$ denotes the average activity in the gullibility class x. We refer the reader to Ref. [39] and to the Appendix A for more details about the derivation.

II. CONTROL STRATEGIES

We imagine a hypothetical scenario of a large institution that, in a given time-window, has the budget to provide cyber-security training to a fraction of its workforce. We assume, for simplicity, that the training provides complete protection from future deception-based attacks. The key question is how to select users that will be trained. In these settings, the cyber-security training is equivalent to a sterilizing vaccine [45]. Indeed, the training reduces the risk of being compromised to zero, and users who receive it are removed from the propagation process. As a result, we model the impact of the training as an SIS model where a fraction of the nodes is completely removed from the spreading dynamics.

We consider four strategies to select users for training. In the first one, users are selected at random independently from any of their features. The second strategy selects users in decreasing order of activity. The implementation of this strategy requires a complete knowledge about users' activity. This might be challenging to obtain in reality due to computational costs required to monitor all communications, as well as privacy and ethical constraints. The third strategy is based on an egocentric sampling of the network of communication starting from random probes. The fourth targets users based on their knowledge of cyber-threats estimated via prototypical security awareness training (SAT) tests.

In what follows, we imagine that the training of a fraction γ of employees takes place in a given time-window. We then assume that a small fraction of users fall for a deception-based attack, and we study the impact of each of the four control strategies in hampering the spreading potential of the cyber-threat. In other words, we study how each strategy protects the system from the attack.

A. Random strategy

Given the framework discussed in the previous section, in the absence of any training, the equation describing the evolution in time of the number of infected nodes with activity a and in class x can be written as

$$d_t I_a^x = -\mu_x I_a^x + m\lambda_x S_a^x$$
$$\times \left[p \int da' a' \frac{I_{a'}^x}{N^x} + (1-p) \sum_{y \neq x} \int da' a' \frac{I_{a'}^y}{N - N^y} \right]. \tag{2}$$

In particular, the first term on the right-hand side accounts for the recovery process, while the second and third terms in the square brackets account for the possibility of infection due to a compromised message coming, respectively, from inside or outside the gullibility class x. In the first strategy, which acts as a baseline, users are randomly selected for training. Hence, we remove a random fraction γ of individuals at the beginning of the spreading process: $R_a^x = \gamma N_a^x$. During the early stages of the spreading, we assume the number of compromised accounts to be small, i.e., $I_a^x \ll N_a^x$. Hence, we can approximate the number of susceptible individuals in each activity class a and gullibility x as $S_a^x \approx (1-\gamma)N_a^x$. Equation (3) then becomes

$$d_t I_a^x = -\mu_x I_a^x + m\lambda_x (1-\gamma) N_a^x$$
$$\times \left[p \int da' a' \frac{I_{a'}^x}{N^x} + (1-p) \sum_{y \neq x} \int da' a' \frac{I_{a'}^y}{N - N^y} \right]. \tag{3}$$

By defining $\lambda_x^{rnd} = \lambda_x(1-\gamma)$ and integrating Eq. (4) across all activities, we obtain

$$d_t I^x = -\mu_x I^x + m\lambda_x^{rnd} \left[p\theta^x + (1-p) \sum_{y \neq x} c_{x,y} \theta^y \right], \tag{4}$$

where $\theta^x = \int da\, a I_a^x$, $c_{x,y} = N^x/(N - N^y)$, $I^x = \int da I_a^x$, and $N^x = \int da N_a^x$. By multiplying both sides of Eq. (4) by a and integrating across all activities, we obtain

$$d_t \theta^x = -\mu_x \theta^x + m\lambda_x^{rnd} \langle a \rangle_x \left[p\theta^x + (1-p) \sum_{y \neq x} c_{x,y} \theta^y \right]. \tag{5}$$

Equations (4) and (5) define a system of $2Q$ differential equations. The cyber-threat will be able to spread only if the largest eigenvalue of the Jacobian matrix of the system is greater than zero. As shown in the Appendix B, the largest eigenvalue of the system reads

$$\Lambda_{max}^{rnd} = -\sum_x \mu_x + p \sum_x \beta_x^{rnd} + \Xi^{rnd}, \tag{6}$$

where $\beta_x^{rnd} = (1-\gamma)\beta_x$ and Ξ^{rnd} is a function of β_x^{rnd}, μ_x, $c_{x,y}$. Considering the expression of the largest eigenvalue of the system of differential equations, the basic reproduction number in the case of random immunization

strategy becomes

$$R_0^{rnd} = \frac{p \sum_x \beta_x^{rnd} + \Xi^{rnd}}{\sum_x \mu_x}. \tag{7}$$

We notice that its expression is equal to the expression of R_0 without any security training [see Eq. (1)], except for the terms β_x^{rnd} and Ξ^{rnd}, which are affected by the definition of λ_x^{rnd}. To showcase the full expression of the threshold, we consider the cases with a single and two gullibility classes, i.e., $Q = 1$ and $Q = 2$.

Case $Q = 1$. With a single gullibility class, $\Xi^{rnd} = 0$ (and $p = 1$), hence

$$R_0^{rnd} = \frac{\beta^{rnd}}{\mu} = \frac{(1 - \gamma)\beta}{\mu} = (1 - \gamma)R_0. \tag{8}$$

In this case, R_0 is simply rescaled from a factor $(1 - \gamma)$ (i.e., the fraction of nodes not removed). This is the classic result of random immunization/removal of nodes. Indeed, the impact of the strategy on the threshold scales linearly with the fraction of nodes removed [45].

Case $Q = 2$. With two gullibility classes, $(\Xi^{rnd})^2$ takes the following expression:

$$(\Xi^{rnd})^2 = (\mu_1 - \mu_2)^2 + (1 - \gamma)^2$$
$$\times \left[p^2(\beta_1 - \beta_2)^2 + \frac{2p}{1 - \gamma}(\mu_2 - \mu_1)(\beta_1 - \beta_2) \right.$$
$$\left. + 4\beta_1\beta_2(1 - 2p) \right]. \tag{9}$$

If the two gullibility classes feature the same recovery rate (i.e., $\mu_1 = \mu_2$), the expression simplifies to

$$(\Xi^{rnd})^2 = (1 - \gamma)^2 [p^2(\beta_1 - \beta_2)^2 + 4\beta_1\beta_2(1 - 2p)]$$
$$= (1 - \gamma)^2 \Xi^2. \tag{10}$$

Hence, when the two classes are characterized by the same recovery rate, the basic reproduction number R_0^{rnd} is equal to R_0 rescaled by a factor $(1 - \gamma)$. In other words, if the two gullibility groups differ just by the probability of infection, the impact of a random removal strategy scales linearly with γ. Interestingly, in the case $\mu_1 \neq \mu_2$, this simple relation does not hold anymore. Indeed, in this case there is an interplay between timescales regulating the infection period of each class. As shown in Ref. [39], in the absence of any intervention strategy, this interplay can make the system more fragile than it would be if each class were considered separately.

B. Activity-based strategy

This second strategy targets nodes that, possibly due to the nature of their job or personal attitude, are more active. Hence, we remove all nodes of class x that feature an activity higher than a given threshold $a_c(x)$. In practice, this means that all integrals across activities go from ϵ to $a_c(x)$ (and not 1). In this case, the early-stage linearization takes the form $S_a^x \sim (1 - \gamma_x)N_a^x$, where $\gamma_x = \int_{a_c(x)}^1 N_a^x / N^x \, da$ is the fraction of nodes removed in class x. The system of $2Q$ differential

equations defined by Eqs. (4) and (5) can be rewritten as

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

$$d_t \theta^x = -\mu_x \theta^x + m\lambda_x^{act} \langle a \rangle_x^c \left[p\theta^x + (1 - p) \sum_{y \neq x} c_{x,y}\theta^y \right],$$
$$\tag{11}$$

where we define $\lambda_x^{act} = (1 - \gamma_x)\lambda_x$ and $\langle a \rangle_x^c = \int_\epsilon^{a_c(x)} aF_x(a) \, da$. Following the same steps outlined above, we derive the basic reproduction number in the case of activity-targeted immunization strategy as

$$R_0^{act} = \frac{p \sum_x \beta_x^{act} + \Xi^{act}}{\sum_x \mu_x}, \tag{12}$$

where $\beta_x^{act} = m\lambda_x^{act} \langle a \rangle_x^c$, and Ξ^{act} is a function of all β_x^{act}, μ_x, and $c_{x,y}$. We note how the threshold is a function of the groups' activity averages, i.e., $\langle a \rangle_x^c$. Hence, one could estimate the impact of the strategy by knowing just such averages. However, this information would not be enough to actually implement it. Indeed, one would have to know which subset of users to select for training, thus requiring full knowledge about their activity.

Case $Q = 1$. With a single gullibility class, $\Xi^{act} = 0$ (and $p = 1$), hence

$$R_0^{act} = \frac{\beta^{act}}{\mu} = \frac{m\lambda_x^{act} \langle a \rangle_x^c}{\mu} = \frac{m(1 - \gamma)\lambda \langle a \rangle_x^c}{\mu}. \tag{13}$$

In this case, the threshold is not a simple rescaling of the R_0 obtained without any interventions. Indeed, the expression is also modified by the contribution of activity classes which are able to get infected. In doing so, the modulation of R_0 induced by targeting the most active nodes in each activity class is, generally speaking, not linear with the fraction of nodes removed.

Case $Q = 2$. The expression of Ξ^{act} is analogous to Ξ^{rnd}, where, however, the β_x are substituted with β_x^{act}. Hence, also in this case, the impact of the selection strategy is regulated by the transmission rates of each class, i.e., β_x^{act}. As noted above, these are affected by the expression of γ_x and the activity distributions of the nodes possibly affected by the threat, i.e., $\langle a \rangle_x^c$. Hence, the impact of each node removed is again nonlinear.

C. Egocentric sampling strategy

As mentioned, the activity-based strategy requires a complete knowledge of nodes' activities. Due to practical and privacy issues, this information is typically unavailable in real-world scenarios. However, a proxy of nodes' activity can be obtained by sampling the egocentric network of a fraction of nodes [15,23]. Egocentric networks capture the connection that each ego (i.e., a given node in the system) has with their alters (i.e., the first neighbors of each ego). We can sample these egocentric networks by randomly selecting a group of nodes that act as probes. We then observe their connections (i.e., egocentric network) during a time-window of length T, neglecting the direction of links. In other words, we sample

the interactions of each probe taking place within an observation window. Then, for each of the probes we pick one alter at random in their egocentric network and select it for security training. The idea behind this selection strategy is that highly active nodes are more likely to be in the egocentric network of different probes. This local sampling strategy improves the likelihood of selecting high-activity nodes without assuming any global knowledge about the system. In general, the number of probes in different classes can vary depending on their size. Given a total number of probes N_w, assuming a random distribution, the expected number of these in each gullibility class can be written as $N_w^x = N_w \frac{N^x}{N}$. We note how in these settings the fraction of probes in each gullibility class (i.e., $w_x = N_w^x/N_x$) is equal to the total fraction and equal across each class (i.e., $w_x = w \; \forall x$). However, the number of probes in each class could be different.

Let us define P_a^x as the probability that, from the egocentric network of a given probe, we select a node of activity a in the gullibility class x. These are the nodes that will undertake the security training and thus will be immune from cyber-attacks. After one observation time step, we can write

$$
\begin{aligned}
P_a^x = & \; ap \int da' N_{a'}^x w_x \frac{m}{N_x} \\
& + a(1-p) \sum_{y \neq x} \int da' N_{a'}^y w_y \frac{m}{N - N_x} \\
& + \int da' a' p N_{a'}^x w_x \frac{m}{N_x} \frac{1}{m} \\
& + \sum_{y \neq x} \int da' a' (1-p) N_{a'}^y w_y \frac{m}{N - N_y} \frac{1}{m} \\
= & \; apw_x m + a(1-p)m \frac{N_w - N_w^x}{N - N_x} + pw_x \langle a \rangle_x \\
& + (1-p) \sum_{y \neq x} \frac{N^y}{N - N_y} w_y \langle a \rangle_y.
\end{aligned}
\tag{14}
$$

In particular, the first term of Eq. (14) represents the probability that nodes with activity a and in gullibility class x are selected for training (i.e., are removed from the cyber-threat dynamics) because they are active and connect with probes in the same gullibility class; the second term is analogous but considers connections with probes in other gullibility classes; the third and the fourth term, instead, represent the probability that nodes are removed after being reached and selected from probes, respectively, inside (third term) and outside (fourth term) their gullibility class. By assuming this selection dynamics independent across time steps, the probability for a node with activity a and in class x to be removed after T time-steps can be written as $P_a^x(T) = 1 - (1 - P_a^x)^T$. Hence, the number of nodes removed after T periods with activity a and in class x is $R_a^x(T) = N_a^x[1 - (1 - P_a^x)^T]$. We note how this formulation is a clear approximation. Indeed, it does not consider the depletion of nodes in each class due to the immunization process. As such, this expression holds in the regime of small T and when the probability that a probe is selected more than once is small. Furthermore, we note how due to the possible selection of the same targets from different probes, in general, the cardinality of the set of nodes selected

by this strategy for security training, γ, might be smaller than the fraction of probes $\gamma \leqslant w$. As before, at early stages of the spreading, we can write $S_a^x \sim N_a^x - R_a^x(T)$ and repeat similar calculations to those explained above to obtain

$$
d_t I^x = -\mu_x I^x + m\lambda_x \Psi_{0,x}^T \left[p\theta^x + (1-p) \sum_{y \neq x} c_{x,y} \theta^y \right], \tag{15}
$$

$$
d_t \theta^x = -\mu_x \theta^x + m\lambda_x \Psi_{1,x}^T \left[p\theta^x + (1-p) \sum_{y \neq x} c_{x,y} \theta^y \right], \tag{16}
$$

where we define $\Psi_{n,x}^T = \int da\, a^n F_x(a)(1 - P_a^x)^T$. In this case, the basic reproduction number can be written as

$$
R_0^{\text{ego}} = \frac{p \sum_x \beta_x^{\text{ego}} + \Xi^{\text{ego}}}{\sum_x \mu_x} \quad \text{with} \quad \beta_x^{\text{ego}} = m\lambda_x \Psi_{1,x}^T. \tag{17}
$$

Case Q = 1. In the case of a single gullibility class, we have $R_0^{\text{ego}} = \frac{\beta^{\text{ego}}}{\mu} = \frac{m\lambda}{\mu} \int da\, aF(a)(1 - P_a)^T$. We note how the egocentric sampling strategy affects the threshold by decreasing, nonlinearly as function T, the average activity of susceptible nodes.

Case Q = 2. In the case of two gullibility classes, the expression of Ξ^{ego} is analogous to Ξ^{rnd}, where β_x^{rnd} are substituted with β_x^{ego}. Also in this case, the impact of strategy on the dynamics is hidden in the $\Psi_{1,x}^T$ expressions, which lead to nonlinear effects.

D. Security awareness training strategy

In this last strategy, we imagine that the corporation runs a security awareness training (SAT) test in which all the employees (i.e., nodes) receive a fake compromised email and/or message (e.g., a phishing email). These tests are customarily used for cyber-security training and awareness purposes [48]. The strategy consists in estimating the gullibility of employees based on the outcomes of the test. In particular, we implement it as follows. With probability g, a user sees the SAT email and opens it. In general, we set $g < 1$ thus not all employees engage with the SAT. After seeing the email, a node in gullibility class x clicks on the compromised link, thus falling for the ruse, with probability λ_x. We assume that the fraction γ of the users that are selected to receive security training is selected from the pool of employees that did not recognize the potential threat and clicked on the compromised link. In doing so, we aim to select users more in need of security training.

In these settings, the average number of employees with activity a and in gullibility class x that would fall for the ruse can be estimated as $g N_a^x \lambda_x$. The fraction of γ' needed, such that the overall fraction of employees ultimately selected for training is γ, can be obtained solving the following equation: $\gamma = N^{-1} \sum_x \int da\, g N_a^x \lambda_x \gamma'$. This leads to

$$
\gamma' = \frac{\gamma}{g\langle\lambda\rangle}, \tag{18}
$$

where $\langle\lambda\rangle = \sum_x \lambda_x N_x/N$ is the average transmissibility in the system. Hence, the number of employees with activity a and
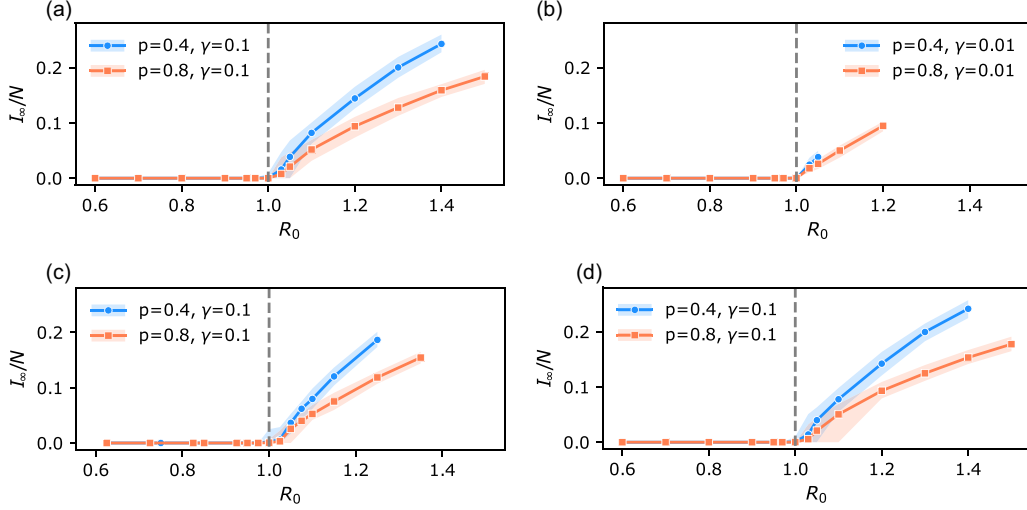
FIG. 1. Numerical validation of the threshold under different strategies. Each panel shows the stationary fraction of infected individuals I_∞/N as a function of R_0 for the four immunization strategies: (a) random, (b) activity-based, (c) egocentric sampling, and (d) SAT. Results are shown for two parameter settings: ($p = 0.4$, $\gamma = 10^{-1}$, $\lambda_2 = 0.5$) in blue and ($p = 0.8$, $\gamma = 10^{-1}$, $\lambda_2 = 0.3$) in orange for panels (a), (c), and (d); and ($p = 0.4$, $\gamma = 10^{-2}$, $\lambda_2 = 0.5$) in blue and ($p = 0.8$, $\gamma = 10^{-2}$, $\lambda_2 = 0.8$) in orange for panel (b). The vertical dashed line indicates the critical value of the threshold computed analytically (i.e., $R_0 = 1$). Solid lines with markers represent mean values, while shaded areas indicate 95% confidence intervals computed in 100 stochastic simulations. Other parameters common to all simulations: $\alpha = 2.1$, $\epsilon = 10^{-3}$, $m = 4$, initial infected percentage 0.5%, $\mu_1 = \mu_2 = 10^{-2}$. In the egocentric sampling strategy case, we set $T = 10$.

in gullibility class x selected for training can be written as

$$R_a^x = g N_a^x \lambda_x \gamma' = N_a^x \frac{\lambda_x}{\langle\lambda\rangle}\gamma. \qquad (19)$$

Interestingly, if a class x features transmissibility equal to the network's average, the fraction of removed nodes in that class is simply that of the random case (i.e., $R_a^x \sim N_a^x \gamma$). If a class x has transmissibility higher (lower) than the average (i.e., is more or less gullible than the average), it will have a higher (lower) fraction of removed nodes with respect to the random case.

At early stages of the spreading, the equation for I_a^x becomes

$$d_t I_a^x = -\mu_x I_a^x + m \lambda_x N_a^x \left(1 - \frac{\lambda_x}{\langle\lambda\rangle}\gamma\right)$$
$$\times \left[p \int da' a' \frac{I_{a'}^x}{N^x} + (1-p) \sum_{y \neq x} \int da' a' \frac{I_{a'}^y}{N - N^y} \right]. \qquad (20)$$

By defining $\lambda_x^{\rm sat} = \lambda_x(1 - \frac{\lambda_x}{\langle\lambda\rangle}\gamma)$, we obtain an equation analogous to the random case. Hence, we can directly write the expression for R_0 as

$$R_0^{\rm sat} = \frac{p \sum_x \beta_x^{\rm sat} + \Xi^{\rm sat}}{\sum_x \mu_x}, \qquad (21)$$

where $\beta_x^{\rm sat} = \beta_x(1 - \frac{\lambda_x}{\langle\lambda\rangle}\gamma)$, and $\Xi^{\rm sat}$ is again a function of $\beta_x^{\rm sat}$, μ_x, $c_{x,y}$.

Case Q = 1. In the case of a single gullibility class, the SAT strategy is equivalent to the random strategy. Indeed, in this case nodes are selected proportionally to their gullibility and not other features. Thus in the case of a single group of nodes,

each one is selected uniformly at random. This aspect of the SAT strategy hints at its difference with respect to the previous two strategies, which, even in the case of one gullibility class, did not lead to the same expression of the baseline strategy.

Case Q = 2. In the case of two gullibility classes, the expression of $R_0^{\rm sat}$ is analogous to the random case. However, as for the other cases, the expression of the β terms is different. The effect of the selection strategy is a function of γ and modulated by the gullibility of each class with respect to the system's average.

Interestingly, as shown in the Appendix E, for any number of gullibility classes and a given fraction of removed nodes γ, the effective average transmissibility in this strategy cannot be larger than in the random case, namely $\langle\lambda^{\rm sat}\rangle \leqslant \langle\lambda^{rnd}\rangle$. This implies that the effective spreading potential in the case in which the subset of nodes is selected via a SAT strategy can only be smaller or equal with respect to a random selection.

E. Numerical simulations

In Fig. 1 we show, for each strategy, the simulated fraction of infected nodes at equilibrium as a function of R_0 for two different values of p and two gullibility classes ($Q = 2$). In all simulations, we assume the process reaches the stationary state when the ratio between the mean and the standard deviation of the prevalence (i.e., the number of currently infected nodes), computed over the last 10^3 simulation steps, falls below a threshold of 0.02. In all scenarios, except those adopting an activity-based strategy, we set $\gamma = 10^{-1}$ (i.e., 10% of employees are enrolled in the security training). The strategy that selects nodes in decreasing order of activity is so effective that, to validate the analytical formulation, we need to consider smaller values of γ (e.g., $\gamma = 10^{-2}$). Indeed, all
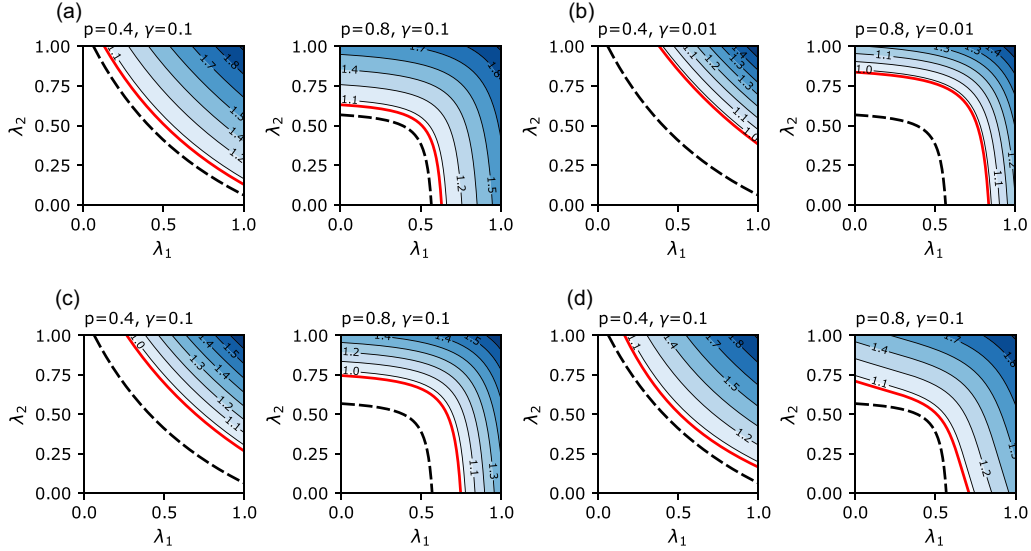
FIG. 2. Phase space of R_0 as a function of λ_1 and λ_2, under different strategies. Each panel corresponds to a specific strategy: (a) random, (b) activity-based, (c) egocentric sampling, and (d) SAT, with two subpanels per strategy representing ($p = 0.4$, $\gamma = 10^{-1}$) and ($p = 0.8$, $\gamma = 10^{-1}$)—except for panel (b), which uses $\gamma = 10^{-2}$. The colored contours show analytically computed R_0 values for each strategy. The red solid contour line marks the critical threshold $R_0 = 1$ under the given intervention strategy, while the black dashed line shows the $R_0 = 1$ threshold in the absence of intervention. Other parameters common to all panels are $\mu_1 = \mu_2 = 10^{-2}$, $m = 4$, $\alpha = 2.1$, $\epsilon = 10^{-3}$.

physical combinations of parameters lead to subcritical states for $\gamma = 10^{-1}$. In all simulations, we fixed the infection probability of the second class (i.e., λ_2) and let λ_1 vary exploring corresponding R_0 values in the range 0.6–1.5. We exclude nonphysical combinations that result in values $\lambda_1 > 1$. Furthermore, we consider a simple case in which the two recovery rates are equal, i.e., $\mu_1 = \mu_2$. As a way to show the validity of the analytical derivation across a wider range of parameters, we set two different values of λ_2 for the two values of p. We use $\lambda_2 = 0.3$ for $p = 0.8$ and $\lambda_2 = 0.5$ for $p = 0.4$. We note how we adopt R_0 as an order parameter rather than λ_1 to fairly compare different parameter combinations. The analytical estimation of the thresholds for all strategies is confirmed by the numerical simulations. Indeed, the analytical thresholds clearly split the phase spaces in two. Below the critical value (i.e., $R_0 = 1$), the cyber-threat is not able to spread into the system. Then, to the right of the critical values, we see a clear transition in the dynamics. Indeed, the fraction of infected nodes reaches an endemic state. The effectiveness of each strategy can be evaluated by looking at the outbreak size for a given R_0. Differences become clear as we move away from the threshold (i.e., $R_0 = 1$). The activity-based strategy emerges are clearly the most effective. Indeed, the values λ_1 above threshold are just extreme values very close to 1. This explains why we have fewer points in that panel [see Fig. 1(b)]. The effectiveness of this strategy is even more striking recalling than in this case we removed only 1% of nodes, rather than 10% as for the other strategies. The baseline and the SAT strategy appear similar and clearly less effective than the activity-based strategy. As mentioned above, the similarity between the two is to be expected by construction. However, the SAT strategy performs marginally better, especially for larger values of p. The egocentric strategy appears to be more effective than both the baseline and the SAT strategy. However, its performance is still far from the most effective. Across the board, we

observe how smaller values of p (i.e., low homophily) result in larger outbreaks, especially for large values of R_0. Hence, well above the threshold, increased mixing across gullibility classes might be detrimental to the whole system in the case of successful attacks.

In the Appendixes, we reproduce Fig. 1 (considering only $p = 0.4$) for other values of parameters. The results show the robustness of the mathematical formulation to the choice of input variables.

In Fig. 2 we show contour plots of the theoretical value of R_0, estimated from the analytical derivations described above, as a function of the gullibility of the two classes, i.e., $R_0(\lambda_1, \lambda_2)$. As for the previous plots, we assume $\mu_1 = \mu_2$. The black dashed lines show the thresholds (i.e., $R_0 = 1$) in the case of $\gamma = 0$ (no security training). The red solid lines, instead, show the thresholds in the case of $\gamma = 10^{-1}$ for all strategies but the activity-based one. As before, we set $\gamma = 10^{-2}$ for this strategy. In each panel, the gap between the two lines quantifies the impact of the training strategy. Indeed, points below each line are subcritical, thus the threat would not be able to spread in those regions of the phase space. The two lines are rather close in the case of the random baseline strategy highlighting the marginal efficacy of this strategy [see Fig. 2(a)]. The effectiveness of the activity-based strategy clearly emerges in Fig. 2(b). Indeed, the gap between the two lines is the largest among all strategies, confirming how selecting nodes based on their activities leads to the best outcomes. We stress one more time how the effectiveness of this strategy is particularly striking when considering that it is the only one for which we removed only 1% of nodes. The egocentric strategy is confirmed to be more effective than both random and SAT strategies with a gap between the two lines closer to the activity-based strategy, though in this case we have $\gamma = 10^{-1}$. The SAT strategy is confirmed to be similar to the random baseline [see Fig. 2(d)], though more effective,
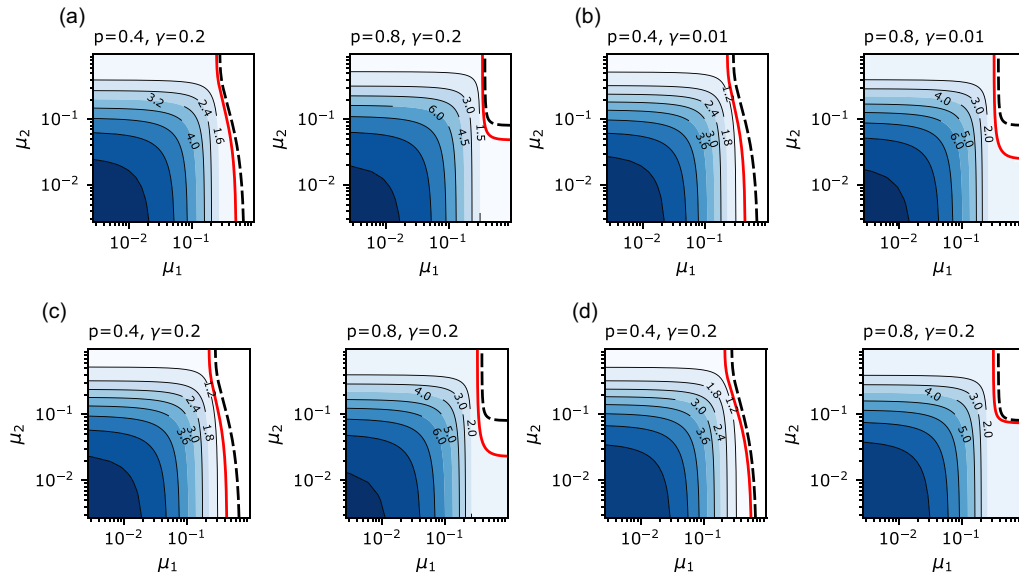
FIG. 3. Phase diagrams of R_0 as a function of μ_1 and μ_2, under different immunization strategies. Each panel corresponds to a specific strategy: (a) random, (b) activity-based, (c) egocentric-sampling, and (d) SAT, with two subpanels per strategy representing ($p = 0.4$, $\gamma = 0.2$) and ($p = 0.8$, $\gamma = 0.2$)—except for panel (b), which uses $\gamma = 10^{-2}$. The colored contours show analytically computed R_0 values for each strategy. The red solid contour line marks the critical threshold $R_0 = 1$ under the given intervention strategy, while the black dashed line shows the $R_0 = 1$ threshold in the absence of intervention. Other parameters common to all panels are $\lambda_1 = 10^{-1}$, $\lambda_2 = 0.8$, $m = 4$, $\alpha = 2.1$, $\epsilon = 10^{-3}$.

especially for larger values of homophily and when the two classes exhibit greater differences in gullibility. Across all strategies, the difference of the phase spaces as a function of p shows how large values of homophily allow macroscopic, yet localized, outbreaks even if one of the two classes is perfectly immune to the threat (e.g., $\lambda_1 = 0$). Indeed, in these scenarios the threat is able to spread, and survive, in one community of the network. On the other hand, smaller values of homophily lead to a larger mix between the two classes, and the dynamics is driven by the interplay between the gullibility of the two classes. Observing the critical value of λ_2 above which the threat would be able to spread even if the other gullibility class is perfectly protected (i.e., $\lambda_1 = 0$) offers another approach to compare strategies. Indeed, higher values of λ_2 highlight better performance in stopping the spreading. This value is the largest in the case of the activity-based strategy ($\lambda_2^c \simeq 0.84$). The egocentric strategy follows with a critical value of $\lambda_2^c \simeq 0.75$. The SAT and random strategy then show values of $\lambda_2^c \simeq 0.71$ and $\lambda_2^c \simeq 0.63$, respectively.

In Fig. 3 we show the phase space of R_0 as a function of μ_1 and μ_2 while fixing the values of λ_1 and λ_2. We set γ in each case setting as $\gamma = 0.2$ for all strategies except for the activity-based strategy, where we use instead $\gamma = 10^{-2}$. Across the board we observe that smaller values of recovery rates result in larger R_0. The trend is to be expected, as the larger the recovery time is (i.e., μ_x^{-1}), the higher is the number of opportunities for each infected node to spread the threat further. Also in this plot, we observe how large values of homophily allow macroscopic outbreaks even if one gullibility class manages to recover immediately after infection (i.e., $\mu_x = 1$), thus limiting the spread of the threat. Furthermore, the plots confirm the hierarchy of efficacy of the four strategies highlighted above.

To further compare the various strategies, we compute the fraction Φ of the (λ_1, λ_2) phase space that results in subcritical dynamics. In other words, Φ is defined as the portion of the (λ_1, λ_2) phase space for which the corresponding R_0 is subcritical (i.e., $R_0 < 1$). For a given range of parameters, the larger Φ is, the smaller is the region of the transmissibility parameters that would allow for a macroscopic outbreak. We show the results for different scenarios in Fig. 4. In all plots, the dashed horizontal lines and gray bars describe the value of Φ in the absence of any control strategy (i.e., $\gamma = 0$). Furthermore, the numerical labels above each bar indicate the value of Φ for each strategy and, in parentheses, the absolute gain with respect to the no-intervention scenario. The panels in the first row are obtained considering the same value of the recovery rates in two gullibility classes, but two different values of the homophily parameter. The second row instead considers scenarios in which the recovery parameters are different. A few observations are in order. First, across the board, the hierarchy of effectiveness of the four strategies confirms previous findings. The activity-based strategy results in the largest increase of Φ. Second, larger values of homophily result in smaller subcritical regions of the phase space. Indeed, as observed above, for large values of p the threat might be able to spread even if one gullibility class is perfectly protected (e.g., $\lambda_1 = 0$). These configurations are not compatible with macroscopic outbreaks in the case of higher mixing levels between gullibility classes (i.e., smaller homophily). We note how these results are not in contrast with the observations we made above, where we noted how, for a given value of R_0, higher levels of homophily corresponded to larger outbreaks. Indeed, Φ quantifies the inactive (i.e., subcritical) region of the phase space. It is agnostic to the prevalence/reach of the cyber-threat in the system. It is
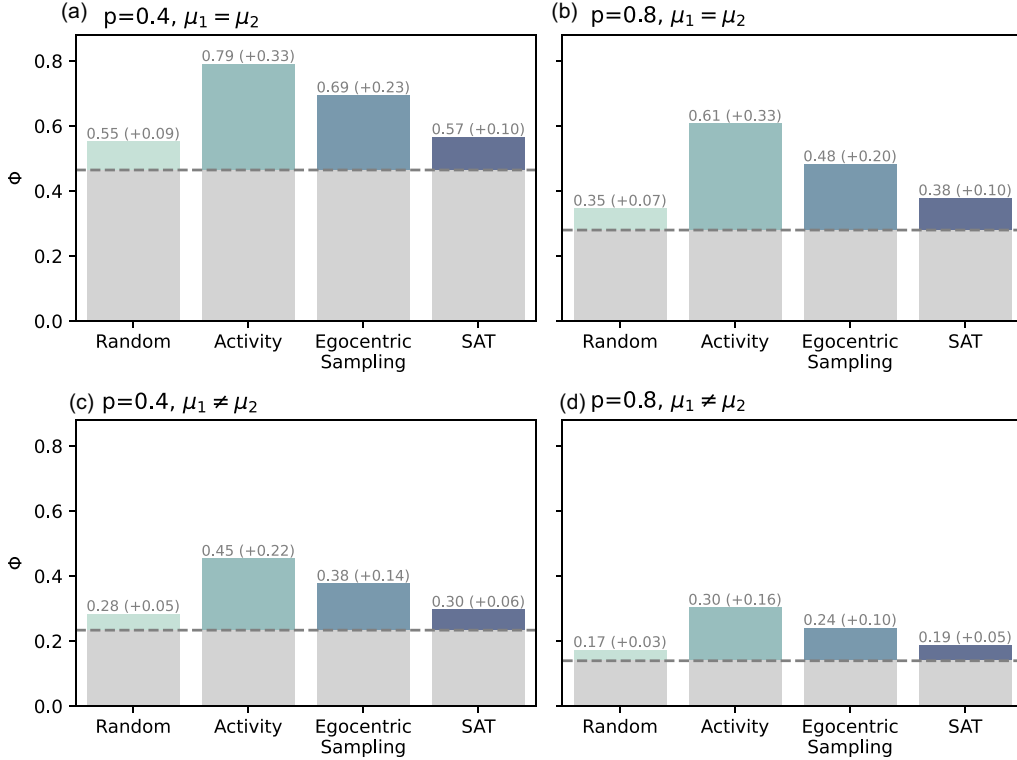
FIG. 4. Subcritical (λ_1, λ_2) phase-space fraction (Φ) under different strategies. Each panel reports results for a specific combination of parameters: (a) $p = 0.4$, $\mu_1 = \mu_2 = 10^{-2}$; (b) $p = 0.8$, $\mu_1 = \mu_2 = 10^{-2}$; (c) $p = 0.4$, $\mu_1 = 10^{-2}$, $\mu_2 = 5 \times 10^{-2}$; (d) $p = 0.8$, $\mu_1 = 10^{-2}$, $\mu_2 = 5 \times 10^{-2}$. The horizontal dashed lines and gray bars represent the value of Φ in the absence of intervention. Numerical labels above the bars indicate the total controlled fraction and, in parentheses, the gain with respect to the no-intervention baseline. Other parameters common to all panels are $m = 4$, $\alpha = 2.1$, $\epsilon = 10^{-3}$.

a measure of the combination of transmissibility parameters that result in subcritical dynamics. It does not provide any information about what happens in supercritical regimes. These two results together suggest how higher values of homophily facilitate the spreading of cyber-threats, but limit their reach in rather isolated groups. Hence, cyber-threats might survive in patches of the network constituted by isolated and highly gullible groups. Third, increasing the recovery rate of even just one class leads to a sensible reduction of the subcritical phase space. Indeed, the values of Φ decrease across the board in Fig. 4(c) and Fig. 4(d). We note, however, how the relative effectiveness of each strategy is preserved also in this case.

III. CONCLUSIONS

We studied the effectiveness of different strategies aimed at containing the spread of deception-based cyber-threats in online social networks. To this end, we modeled the temporal interactions among users using the framework of activity-driven networks. We allowed for the presence of multiple gullibility (i.e., susceptibility) classes describing heterogeneous risk profiles of users. Furthermore, we assumed that the membership to a gullibility class affects the interaction dynamics via a tunable homophily parameter. Finally, we simulated the spreading of cyber-threats using prototypical SIS epidemic models. In these settings, we quantified the efficacy of four strategies aimed at selecting a fraction of nodes to be protected from such threats. The first strategy acts

as a baseline and selects individuals at random. The second assumes complete knowledge of the activity (i.e., propensity of initiating online interactions) of each individual and targets first the most active nodes. The third is based on an egocentric sampling strategy aimed at reaching highly active nodes without assuming any knowledge about their activity. The fourth is based on estimating the gullibility of each node via security awareness tests, which are routinely employed in many institutions to probe the cyber-security awareness of the workforce [48]. We analytically derived the epidemic threshold under each intervention strategy. In doing so, we quantified their effectiveness to control the spreading process. Large-scale numerical simulations validated the analytical expressions across all strategies. The results obtained clearly show the high effectiveness of activity-based strategies, which are able to outperform the others even when protecting a smaller fraction of individuals. The egocentric sampling strategy emerges as second best, confirming the value of local sampling strategies aimed at reaching the most active nodes without global knowledge of the system. The fourth strategy, based on security awareness tests, proved only marginally better than the baseline.

Our findings confirm how identifying the most important nodes in the network allows us to effectively interrupt the spreading of contagion processes [23,45,49]. Furthermore, the results highlight that highly homophilic interactions within gullibility classes expand the transmissibility phase space, thereby fostering conditions for macroscopic outbreaks.

Indeed, in these conditions the cyber-threat may still spread within the most gullible group, even when it cannot propagate through others. At the same time, we find that larger values of homophily ultimately reduce the outbreak size with respect to more mixed scenarios. The modulation effects induced by the mixing levels between gullibility classes highlight the importance of considering heterogeneous susceptibility groups. Indeed, neglecting them in favor of a homogeneous representation of gullibility might lead to a misrepresentation of the spreading potential of cyber-threats and of the efficacy of strategies aimed at hampering them. Furthermore, these results suggest how cyber-threats might survive and propagate in rather isolated groups of gullible individuals, and they highlight the importance of identifying and increasing the awareness of these communities.

The work presented comes with several limitations. First, we neglected more realistic mechanisms driving the interaction between users. Indeed, while we accounted for homophily, we did not consider popularity and social reinforcement mechanisms, among others [12,50]. Second, for simplicity we assumed indefinite and perfect protection granted by cyber-security training. Third, we considered the recovery process as a spontaneous transition function of the gullibility of each node. Hence, we did not account for the possibility that a compromised account might be informed by others, in response to their anomalous behavior. Fourth, we considered only a few groups of gullibility. Future work should explore the impact of more heterogeneity across users. Fifth, we used a simple SIS compartmentalization setup to model cyber-threats. More realistic setups accounting, for example, for possible latent states of the threats are possible alternatives [51]. Sixth, we did not account for possible correlations between activity and gullibility that might lead to the emergence of additional interesting dynamics. Finally, we modeled the probability of falling for a ruse and getting infected as a function only of the gullibility class of each node. Hence, we neglected possible modulations induced by past experiences (i.e., past infection events), recency, and frequency biases [52]. We leave accounting for these limitations to future work.

Overall, our results highlight the striking effectiveness of targeted strategies based on node activity. At the same time, they confirm the effectiveness of local sampling strategies that, although not as efficient as targeted approaches, do not require access to global information about systems' connections. The research contributes to the limited literature devoted to controlling the spread of cyber-threats accounting for both temporal dynamics and heterogeneous susceptibility of users.

APPENDIX A: SPREADING THRESHOLD DERIVATION FOR $\gamma = 0$

The interactions between nodes follow the model proposed in Ref. [39]. A population of N users is divided into Q categories describing their susceptibility to cyber-threats (i.e. gullibility classes). Nodes feature an activity a describing their propensity to initiate communications. Activities are extracted from a power-law distribution $F(a) \sim a^{-\alpha}$ with $a \in [\varepsilon, 1]$. In these settings, at each time step t, a network is generated as follows:

(i) Each node is initially disconnected.

(ii) With probability $a\Delta t$ each node becomes active.

(iii) Active nodes select m others and create directed links (e.g., send them a message). Furthermore, with probability p the new links are created within the same class at random. With probability $1 - p$, links are created with nodes in other gullibility classes, at random.

(iv) At time $t + \Delta t$, all links are deleted and the process restarts.

Without lack of generality, we can set $\Delta t = 1$.

We simulate the spreading of deception-based cyber-threats unfolding on top of these temporal networks using a classic susceptible-infected-susceptible (SIS) model [44]. Nodes in class x get infected with probability λ_x and spontaneously become susceptible again at a rate μ_x. We stress the asymmetry in the transmission process: an infection can only occur when an infected contacts a susceptible, and not the opposite.

Assuming that all nodes in the same gullibility and activity class are statistically equivalent, and considering the continuous limit (i.e., $N \to \infty$), we can write the equation describing the evolution of the number of infected as

$$d_t I_a^x = -\mu_x I_a^x + \lambda_x m S_a^x$$
$$\times \left[p \int da' a' \frac{I_{a'}^x}{N^x} + (1-p) \sum_{y \neq x} \int da' a' \frac{I_{a'}^y}{N - N^y} \right]. \tag{A1}$$

The first term on the right-hand side captures the recovery process. The second term describes susceptible nodes that receive a compromised message coming from their gullibility class and as a result get infected. The third is analogous to the previous term but accounts for compromised messages arriving from other gullibility classes. At early stages, we can assume that the number of compromised nodes is very small, hence we can consider the approximation $S_a^x \sim N_a^x$. The previous equation becomes

$$d_t I_a^x = -\mu_x I_a^x + \lambda_x m N_a^x \left[p \int da' a' \frac{I_{a'}^x}{N^x} + (1-p) \sum_{y \neq x} \int da' a' \frac{I_{a'}^y}{N - N^y} \right]. \tag{A2}$$

We observe that $\int da I_a^x = I^x$, $\int da N_a^x = N^x$. Integrating both members of Eq. (A2) over all activities, we obtain

$$d_t I^x = -\mu_x I_x + \lambda_x m \left[p\theta^x + (1-p)\sum_{y\neq x} c_{x,y}\theta^y \right], \quad \text{(A3)}$$

where we define $\theta^x = \int da' a I_a^x$, $c_{x,y} = N^x/(N - N^y)$. We multiply both members of Eq. (A2) by a and integrate over all activities:

$$d_t \theta^x = -\mu_x \theta^x + \lambda_x m \langle a \rangle_x \left[p\theta^x + (1-p)\sum_{y\neq x} c_{x,y}\theta^y \right], \quad \text{(A4)}$$

where we define $\langle a \rangle_x = \int da\, a N_a^x / N^x$. Finally, we obtain a system of $2Q$ differential equations that describes the evolution of the system:

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

$$d_t \theta^x = -\mu_x \theta^x + \lambda_x m \langle a \rangle_x \left[p\theta^x + (1-p)\sum_{y\neq x} c_{x,y}\theta^y \right] = h^x. \quad \text{(A5)}$$

The threat would be able to spread if the largest eigenvalue of the Jacobian matrix of this system is larger than zero. The Jacobian matrix can be written as

$$J = \begin{bmatrix}
\frac{\partial g^1}{\partial I^1} & \cdots & \frac{\partial g^1}{\partial I^Q} & \frac{\partial g^1}{\partial \theta^1} & \cdots & \frac{\partial g^1}{\partial \theta^Q} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\frac{\partial g^Q}{\partial I^1} & \cdots & \frac{\partial g^Q}{\partial I^Q} & \frac{\partial g^Q}{\partial \theta^1} & \cdots & \frac{\partial g^Q}{\partial \theta^Q} \\
\frac{\partial h^1}{\partial I^1} & \cdots & \frac{\partial h^1}{\partial I^Q} & \frac{\partial h^1}{\partial \theta^1} & \cdots & \frac{\partial h^1}{\partial \theta^Q} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\frac{\partial h^Q}{\partial I^1} & \cdots & \frac{\partial h^Q}{\partial I^Q} & \frac{\partial h^Q}{\partial \theta^1} & \cdots & \frac{\partial h^Q}{\partial \theta^Q}
\end{bmatrix}.$$

Substituting the partial derivatives, we get a block matrix whose structure depends on μ_x, λ_x, $c_{x,y}$, $\langle a \rangle_x$, and p:

$$J = \begin{bmatrix}
-\mu_1 & 0 & \cdots & 0 & p\lambda_1 m & (1-p)\lambda_1 mc_{1,2} & \cdots & (1-p)\lambda_1 mc_{1,Q} \\
0 & -\mu_2 & \cdots & 0 & (1-p)\lambda_2 mc_{2,1} & p\lambda_2 m & \cdots & (1-p)\lambda_2 mc_{2,Q} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & -\mu_Q & (1-p)\lambda_2 mc_{Q,1} & (1-p)\lambda_2 mc_{Q,2} & \cdots & p\lambda_Q m \\
0 & 0 & \cdots & 0 & -\mu_1 + p\beta_1 & (1-p)\beta_1 c_{1,2} & \cdots & (1-p)\beta_1 c_{1,Q} \\
0 & 0 & \cdots & 0 & (1-p)\beta_2 c_{2,1} & -\mu_2 + p\beta_2 & \cdots & (1-p)\beta_2 c_{2,Q} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & (1-p)\beta_Q c_{Q,1} & (1-p)\beta_Q c_{Q,2} & \cdots & -\mu_Q + p\beta_Q
\end{bmatrix},$$

where we define $\beta_x = m\langle a \rangle_x \lambda_x$. Then, the largest eigenvalue $\Lambda_{\max}$ can be written in a general form as [39]

$$\Lambda_{\max} = -\sum_x \mu_x + p\sum_x \beta_x + \Xi, \quad \text{(A6)}$$

where Ξ is an algebraic function of β_x, μ_x, and $c_{x,y}$ and its analytical expression depends on the number of classes Q. Since the cyber-threat is able to spread if $\Lambda_{\max} > 0$, we define the basic reproduction number as

$$R_0 = \frac{p\sum_x \beta_x + \Xi}{\sum_x \mu_x}. \quad \text{(A7)}$$

If $R_0 > 1$, the threat would be able to spread affecting a macroscopic fraction of the population.

APPENDIX B: RANDOM STRATEGY

Here, we provide the details about the threshold derivation for the random strategy. We recall that in this case, the fraction of removed nodes, γ, is picked at random, independently of any of their features. Hence, in the early stages of the spreading we can write

$$S_a^x \sim (1-\gamma)N_a^x. \quad \text{(B1)}$$

Substituting this in the dynamics described by Eq. (A1),

$$d_t I_a^x = -\mu_x I_a^x + \lambda_x (1-\gamma) m N_a^x$$
$$\times \left[p \int da'a' \frac{I_{a'}^x}{N^x} + (1-p)\sum_{y\neq x} \int da'a' \frac{I_{a'}^y}{N - N^y} \right]. \quad \text{(B2)}$$

By defining $\lambda_x^{rnd} = \lambda_x(1-\gamma)$, the equation is analogous to Eq. (A2). Hence, we can directly write the expression of R_0 in this case as

$$R_0^{rnd} = \frac{p\sum_x \beta_x^{rnd} + \Xi^{rnd}}{\sum_x \mu_x}, \quad \text{(B3)}$$

where $\beta_x^{rnd} = (1-\gamma)\beta_x$, and Ξ^{rnd} is a function of β_x^{rnd}, μ_x, and $c_{x,y}$.

APPENDIX C: ACTIVITY-BASED STRATEGY

Here, we provide the detailed derivation of the threshold for the activity-based strategy. In this strategy, we remove all nodes of class x that show an activity higher than a threshold $a_c(x)$. In practice, integrals across activities now spans from ϵ to $a_c(x)$ (and not 1), reflecting the immunization of most active nodes. In the early stages of the spreading, we can write

$$S_a^x \sim (1 - \gamma_x)N_a^x, \tag{C1}$$

where γ_x is the fraction of nodes removed in class x. This quantity can be computed as

$$\gamma_x = \int_{a_c(x)}^1 da N_a^x / N^x. \tag{C2}$$

By repeating the same same calculation presented in Appendix A, we obtain the system of $2Q$ equations:

$$d_t I^x = -\mu_x I^x + m \lambda_x (1 - \gamma_x)$$
$$\times \left[p\theta^x + (1 - p)\sum_{y \neq x} c_{x,y}\theta^y \right], \tag{C3}$$

$$d_t \theta^x = -\mu_x \theta^x + m \lambda_x (1 - \gamma_x)\langle a \rangle_x^c$$
$$\times \left[p\theta^x + (1 - p)\sum_{y \neq x} c_{x,y}\theta^y \right]. \tag{C4}$$

As for the previous strategy, by defining $\lambda_x^{\text{act}} = \lambda_x(1 - \gamma_x)$, we can map this system of equation to the case for $\gamma = 0$. Hence, we can directly write the expression for the basic reproductive number as

$$R_0^{\text{act}} = \frac{p \sum_x \beta_x^{\text{act}} + \Xi^{\text{act}}}{\sum_x \mu_x}, \tag{C5}$$

where $\beta_x^{\text{act}} = m \lambda_x^{\text{act}} \langle a \rangle_x^c$ and Ξ^{act} is a function of all β_x^{act}, μ_x, and $c_{x,y}$.

APPENDIX D: EGOCENTRIC SAMPLING STRATEGY

Here, we provide the detailed derivation of the threshold for the egocentric sampling strategy. In this strategy, a random fraction w of nodes is selected as probes. We observe their interactions (i.e., egocentric network) for T time steps. Then, for each of the probes, we remove, at random, exactly one of their neighbors in the aggregate egocentric network. Let us define N_w as the total number of probes. Assuming a random distribution, the expected number of these in each gullibility class is $N_w^x = N_w \frac{N^x}{N}$. This implies that the fraction of probes in each gullibility class is $w_x = N_w^x / N_x$. In the case of random distribution, the average fraction of probes in each class is the same, and it is equal to the overall fraction $w_x = w \; \forall x$.

Let us define P_a^x as the probability that, from a given probe, we select a node of activity a in the gullibility class x. After one observation time step, we can write

$$P_a^x = ap \int da' N_{a'}^x w_x \frac{m}{N_x} + a(1 - p) \sum_{y \neq x} \int da' N_{a'}^y w_y \frac{m}{N - N_x}$$
$$+ \int da' a' p N_{a'}^x w_x \frac{m}{N_x} \frac{1}{m} + \sum_{y \neq x} \int da' a' (1 - p) N_{a'}^y w_y \frac{m}{N - N_y} \frac{1}{m}$$
$$= apw_x m + a(1 - p)m \frac{N_w - N_w^x}{N - N_x} + pw_x \langle a \rangle_x + (1 - p)\sum_{y \neq x} \frac{N^y}{N - N_y} w_y \langle a \rangle_y. \tag{D1}$$

In particular, the first and the second term represent the probability that the node is removed after reaching one of the probe, respectively, inside and outside its gullibility class; the third and the fourth term, instead, represent the probability that the node is removed after being reached from a probe, respectively, inside and outside its gullibility class. By assuming independent subsequent time steps, the probability for a node with activity a and in class x to be removed after T periods of length 1 is $P_a^x(T) = 1 - (1 - P_a^x)^T$. Hence, the number of nodes removed after T periods with activity a and in class x is $R_a^x(T) = N_a^x[1 - (1 - P_a^x)^T]$. The number of susceptible in each activity and gullibility class can be approximated, at early times, as $S_a^x \sim N_a^x - R_a^x$. By using these two expressions in Eq. (A1) integrating across all activities, we get

$$d_t I^x = -\mu_x I^x + m\lambda_x \Psi_{0,x}^T \left[p\theta^x + (1 - p)\sum_{y \neq x} c_{x,y}\theta^y \right] = h^x, \tag{D2}$$

where we define $\Psi_{n,x}^T = \int da \, a^n F_x(a)(1 - P_a^x)^T$. By multiplying by a and integrating across all activities instead, we obtain

$$d_t \theta^x = -\mu_x \theta^x + m\lambda_x \Psi_{1,x}^T \left[p\theta^x + (1 - p)\sum_{y \neq x} c_{x,y}\theta^y \right] = g^x. \tag{D3}$$

Equations (D2) and (D3) define the system of $2Q$ equations in the case of egocentric network sampling immunization strategies. In this case, the mapping with the simple case with $\gamma = 0$ is not immediate, at least from the system of equations. Hence, in order to obtain the threshold, we can compute the Jacobian matrix. By defining $\lambda_x^{\text{ego}} = \lambda_x \Psi_{0,x}^T$ and $\beta_x^{\text{ego}} = m\lambda_x \Psi_{1,x}^T$,

we obtain

$$
J = \left[
\begin{array}{cccc|ccc}
-\mu_1 & 0 & \cdots & 0 & p\lambda_1^{\text{ego}}m & (1-p)\lambda_1^{\text{ego}}mc_{1,2} & \cdots \\
0 & -\mu_2 & \cdots & 0 & (1-p)\lambda_2^{\text{ego}}mc_{2,1} & p\lambda_2^{\text{ego}}m & \cdots \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\
0 & 0 & \cdots & -\mu_Q & (1-p)\lambda_Q^{\text{ego}}mc_{Q,1} & (1-p)\lambda_Q^{\text{ego}}mc_{Q,2} & \cdots \\
\hline
0 & 0 & \cdots & 0 & -\mu_1 + p\beta_1^{\text{ego}} & (1-p)\beta_1^{\text{ego}}c_{1,2} & \cdots \\
0 & 0 & \cdots & 0 & (1-p)\beta_2^{\text{ego}}c_{2,1} & -\mu_2 + p\beta_2^{\text{ego}} & \cdots \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\
0 & 0 & \cdots & 0 & (1-p)\beta_Q^{\text{ego}}c_{Q,1} & (1-p)\beta_Q^{\text{ego}}c_{Q,2} & -\mu_Q + p\beta_Q^{\text{ego}}
\end{array}
\right].
$$

The structure of the Jacobian is very similar to the cases discussed above. Indeed, we obtain

$$
R_0^{\text{ego}} = \frac{p\sum_x \beta_x^{\text{ego}} + \Xi^{\text{ego}}}{\sum_x \mu_x}. \tag{D4}
$$

APPENDIX E: SECURITY AWARENESS TEST STRATEGY

Here, we provide the detailed derivation of the threshold in the case of the security awareness test strategy. In this strategy, we imagine that all nodes receive a fake compromised email and/or message meant to test their awareness of and susceptibility to cyber-threats. With probability g a node opens the message, and with probability λ_x it gets infected, falling for the ruse. The fraction of nodes to be selected from training is picked from this subset of nodes that (1) opened the message and (2) did not recognize it as a cyber-threat. In these settings, during the early stages of the spreading, we can write $R_a^x \sim N_a^x g\lambda_x\gamma'$, where γ' is the fraction of nodes that fall for the attack that gets selected for training. For comparability with other strategies, we now derive the expression of γ' as a function of γ. By definition of γ we have that $\sum_x \int da \frac{R_a^x}{N} = \gamma$. By solving with respect to γ', we obtain

$$
\gamma' = \frac{\gamma}{g\langle\lambda_x\rangle}, \tag{E1}
$$

where $\langle\lambda\rangle = \sum_x \lambda_x N_x/N$ is the weighted average of the transmissibility parameter across different gullibility classes. Using this expression, we get

$$
R_a^x \sim N_a^x \frac{\lambda_x}{\langle\lambda_x\rangle}\gamma. \tag{E2}
$$

Interestingly, if a class x has a transmissibility parameter equal to the average, the fraction of removed nodes in that class is simply that of the random case (i.e., $R_a^x \sim N_a^x\gamma$). If a class x has transmissibility higher (lower) than the average (i.e., is more or less gullible than the average), it will have a higher (lower) fraction of removed nodes with respect to the random case. During the early stages of the spreading, the equation for I_a^x can be written as

$$
d_t I_a^x = -\mu_x I_a^x + m\lambda_x N_a^x\left(1 - \frac{\lambda_x}{\langle\lambda\rangle}\gamma\right)\left[p\int da'a'\frac{I_{a'}^x}{N^x}\right.
$$

$$
\left. + (1-p)\sum_{y\neq x}\int da'a'\frac{I_{a'}^y}{N-N^y}\right]. \tag{E3}
$$

By defining $\lambda_x^{\text{sat}} = \lambda_x(1 - \frac{\lambda_x}{\langle\lambda\rangle}\gamma)$, we can map the equations to the simple case $\gamma = 0$ for which we have already derived the solution. Hence, we can directly write

$$
R_0^{\text{sat}} = \frac{p\sum_x \beta_x^{\text{sat}} + \Xi^{\text{sat}}}{\sum_x \mu_x}, \tag{E4}
$$

where $\beta_x^{\text{sat}} = \beta_x(1 - \frac{\lambda_x}{\langle\lambda\rangle}\gamma)$, and Ξ^{sat} is again a function of β_x^{sat}, μ_x, $c_{x,y}$.

It can be shown that, for a given fraction of removed nodes γ, the average transmissibility across gullibility classes (i.e., a proxy for the effectiveness of the immunization strategy) in the case of the social awareness test cannot be larger than in the random case, namely $\langle\lambda^{\text{sat}}\rangle \leqslant \langle\lambda^{rnd}\rangle$.

As mentioned, the average transmissibility across gullibility classes is

$$
\langle\lambda^{\text{sat}}\rangle = \sum_x \lambda_x^{\text{sat}}\frac{N_x}{N} \tag{E5}
$$

$$
= \sum_x \lambda_x\left(1 - \frac{\lambda_x}{\langle\lambda\rangle}\gamma\right)\frac{N_x}{N} \tag{E6}
$$

$$
= \sum_x \lambda_x\frac{N_x}{N} - \sum_x \frac{\lambda_x^2}{\langle\lambda\rangle}\gamma\frac{N_x}{N} \tag{E7}
$$

$$
= \langle\lambda\rangle - \gamma\frac{\langle\lambda^2\rangle}{\langle\lambda\rangle}. \tag{E8}
$$

In the random case, instead,

$$
\langle\lambda^{rnd}\rangle = \sum_x \lambda_x^{rnd}\frac{N_x}{N} \tag{E9}
$$

$$
= \sum_x \lambda_x(1-\gamma)\frac{N_x}{N} \tag{E10}
$$

$$
= \sum_x \lambda_x\frac{N_x}{N} - \gamma\sum_x \lambda_x\frac{N_x}{N} \tag{E11}
$$

$$
= \langle\lambda\rangle - \gamma\langle\lambda\rangle. \tag{E12}
$$

By comparing the two, we obtain that $\langle\lambda^{\text{sat}}\rangle \leqslant \langle\lambda^{rnd}\rangle$ if the following conditions hold:

$$
\langle\lambda^2\rangle \geqslant \langle\lambda\rangle^2. \tag{E13}
$$

This is always verified. Even more, the equality holds only if λ_x is constant across x. In other words, every time there is
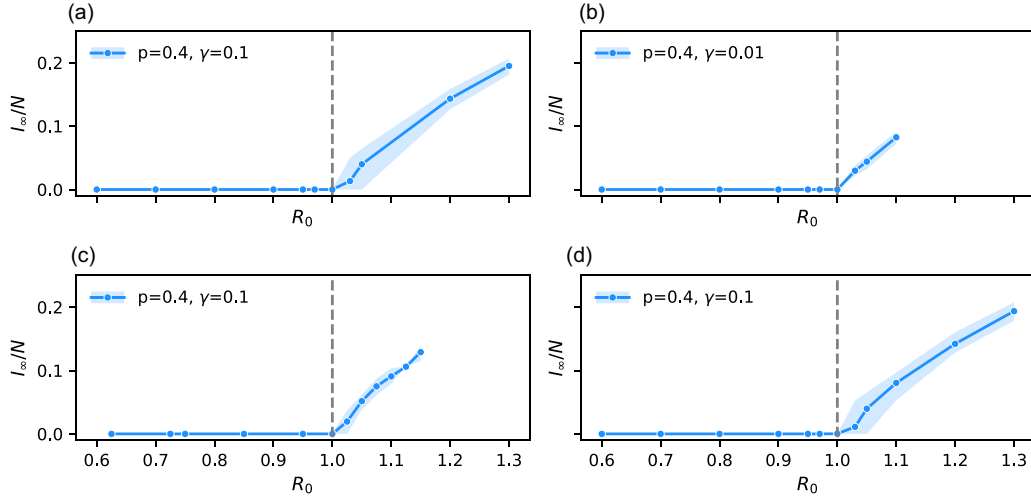
FIG. 5. Numerical validation of the threshold under different strategies (sensitivity analysis). Each panel shows the stationary fraction of infected individuals I_∞/N as a function of R_0, for the four immunization strategies: (a) random, (b) activity-based, (c) egocentric sampling, and (d) SAT. Results are shown for ($p = 0.4$, $\gamma = 10^{-1}$, $\lambda_2 = 0.5$) for panels (a), (c), and (d) and ($p = 0.4$, $\gamma = 10^{-2}$, $\lambda_2 = 0.5$) for panel (b). The vertical dashed line indicates the critical value of the threshold computed analytically (i.e., $R_0 = 1$). Solid lines with markers represent mean values, while shaded areas indicate 95% confidence intervals computed in 100 stochastic simulations. Other parameters common to all simulations are $\alpha = 2.3$, $\epsilon = 10^{-3}$, $m = 8$, initial infected percentage 0.5%, $\mu_1 = \mu_2 = 0.015$. In the egocentric sampling strategy case, we set $T = 10$.

heterogeneity across gullibility classes, the social awareness test immunization is more efficient than the random case in reducing the spread.

APPENDIX F: SENSITIVITY ANALYSIS

In Fig. 5 we show results of a sensitivity analysis regarding the choice of the parameters used in Fig. 1 in the main text. In particular, we reproduce the results of the numerical

simulations validating the analytical epidemic thresholds derived for the four strategies changing the following parameters: $\alpha = 2.3$ (instead of 2.1 used in Fig. 1), $m = 8$ (instead of 4 used in Fig. 1), and $\mu_1 = \mu_2 = 0.015$ (instead of 0.01 used in Fig. 1). We can see that the fraction of infected nodes at the equilibrium grows as a function of R_0 and shows a clear phase transition at $R_0 = 1$, confirming the validity of analytical derivations also in the case of a different parameter choice.

[1] Global threat report, elastic security labs, 2024, https://www.elastic.co/explore/security-without-limits/global-threat-report.

[2] I. Kayes and A. Iamnitchi, Online Social Netw. Media **3-4**, 1 (2017).

[3] B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, Neural Comput. Appl. **28**, 3629 (2017).

[4] R. Heartfield and G. Loukas, Int. J. Cyber Situational Awareness (IJCSA) **1** (2016).

[5] R. Heartfield, G. Loukas, and D. Gan, IEEE Access **4**, 6910 (2016).

[6] V. Zimmermann and K. Renaud, Int. J. Human-Comput. Stud. **131**, 169 (2019).

[7] R. Heartfield and G. Loukas, ACM Comput. Surv. **48**, 1 (2016).

[8] R. Heartfield and G. Loukas, Versatile Cybersecurity **72**, 99 (2018).

[9] P. Falade, Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol. **9**, 5 (2023).

[10] D. V. Grbic and I. Dujlovic, in *2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH)* (IEEE, 2023), pp. 1–5.

[11] P. Holme and J. Saramäki, Phys. Rep. **519**, 97 (2012).

[12] P. Holme, Eur. Phys. J. B **88**, 234 (2015).

[13] M. E. J. Newman, S. Forrest, and J. Balthrop, Phys. Rev. E **66**, 035101(R) (2002).

[14] J. Balthrop, S. Forrest, M. E. Newman, and M. M. Williamson, Science **304**, 527 (2004).

[15] R. Cohen, S. Havlin, and D. ben Avraham, Phys Rev Lett. **91**, 247901 (2003).

[16] B. Ribeiro, N. Perra, and A. Baronchelli, Sci. Rep. **3**, 3006 (2013).

[17] N. Perra, A. Baronchelli, D. Mocanu, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. Lett. **109**, 238701 (2012).

[18] M. Starnini, A. Baronchelli, A. Barrat, and R. Pastor-Satorras, Phys. Rev. E **85**, 056115 (2012).

[19] I. Scholtes, N. Wider, R. Pfitzner, A. Garas, C. Tessone, and F. Schweitzer, Nat. Commun. **5**, 5024 (2014).

[20] L. E. Rocha and N. Masuda, New J. Phys. **16**, 063023 (2014).

[21] A. Barrat and C. Cattuto, in *Social Phenomena* (Springer International, 2015), pp. 37–57.

[22] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani, Sci. Rep. **2**, 469 (2012).

[23] S. Liu, N. Perra, M. Karsai, and A. Vespignani, Phys. Rev. Lett. **112**, 118702 (2014).

[24] S.-Y. Liu, A. Baronchelli, and N. Perra, Phys. Rev. E **87**, 032805 (2013).

[25] G. Ren and X. Wang, Chaos **24**, 023116 (2014).

[26] M. Starnini, A. Machens, C. Cattuto, A. Barrat, and R. Pastor-Satorras, J. Theor. Biol. **337**, 89 (2013).

[27] L. E. Rocha and V. D. Blondel, PLoS Comput. Biol. **9**, e1002974 (2013).

[28] K. Sun, A. Baronchelli, and N. Perra, Eur. Phys. J. B **88**, 1 (2015).

[29] T. Takaguchi, N. Masuda, and P. Holme, PLoS One **8**, e68629 (2013).

[30] P. Holme and F. Liljeros, Sci. Rep. **4**, 4999 (2014).

[31] P. Holme and N. Masuda, PLoS One **10**, e0120567 (2015).

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

[33] B. Prakash, H. Tong, M. Valler, and C. Faloutsos, Mach. Learn. Knowl. Discovery Databases Lecture Notes Comput. Sci. **6323**, 99 (2010).

[34] E. Valdano, L. Ferreri, C. Poletto, and V. Colizza, Phys. Rev. X **5**, 021005 (2015).

[35] D. Mistry, Q. Zhang, N. Perra, and A. Baronchelli, Phys. Rev. E **92**, 042805 (2015).

[36] M.-X. Liu, W. Wang, Y. Liu, M. Tang, S.-M. Cai, and H.-F. Zhang, Phys. Rev. E **95**, 052306 (2017).

[37] M. Karsai, N. Perra, and A. Vespignani, Sci. Rep. **4**, 4001 (2014).

[38] L. E. Rocha, F. Liljeros, and P. Holme, PLoS Comput. Biol. **7**, e1001109 (2011).

[39] T. Brett, G. Loukas, Y. Moreno, and N. Perra, Phys. Rev. E **99**, 050303(R) (2019).

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

[41] 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.

[42] E. Ubaldi, N. Perra, M. Karsai, A. Vezzani, R. Burioni, and A. Vespignani, Sci. Rep. **6**, 35724 (2016).

[43] M. Tizzani, S. Lenti, E. Ubaldi, A. Vezzani, C. Castellano, and R. Burioni, Phys. Rev. E **98**, 062315 (2018).

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

[45] M. Newman, *Networks. An Introduction* (Oxford Univesity Press, 2010).

[46] K. Jansson and R. von Solms, Behav. Inf. Tech. **32**, 584 (2013).

[47] M. McPherson, L. Smith-Lovin, and J. M. Cook, Annu. Rev. Sociol. **27**, 415 (2001).

[48] M. M. Al-Daeef, N. Basir, and M. M. Saudi, in *Proceedings of the World Congress on Engineering* (WCE, 2017), Vol. 1, pp. 5–7.

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

[50] K. Sun, E. Ubaldi, J. Zhang, M. Karsai, and N. Perra, in *Temporal Network Theory* (Springer, 2023), pp. 313–333.

[51] A. Chernikova, N. Gozzi, N. Perra, S. Boboila, T. Eliassi-Rad, and A. Oprea, Appl. Network Sci. **8**, 52 (2023).

[52] E. A. Cranford, C. Gonzalez, P. Aggarwal, M. Tambe, S. Cooney, and C. Lebiere, Cognit. Sci. **45**, e13013 (2021).