# Distinguishing Simple and Complex Contagion Processes on Networks

Giulia Cencetti,[1] Diego Andrés Contreras,[2] Marco Mancastroppa,[2] and Alain Barrat[2]
[1]Fondazione Bruno Kessler, Trento, Italy
[2]Aix-Marseille Univ, Université de Toulon, CNRS, Centre de Physique Théorique, Turing Center for Living Systems, Marseille, France

(Received 27 January 2023; revised 25 April 2023; accepted 17 May 2023; published 15 June 2023)

Contagion processes on networks, including disease spreading, information diffusion, or social behaviors propagation, can be modeled as simple contagion, i.e., as a contagion process involving one connection at a time, or as complex contagion, in which multiple interactions are needed for a contagion event. Empirical data on spreading processes, however, even when available, do not easily allow us to uncover which of these underlying contagion mechanisms is at work. We propose a strategy to discriminate between these mechanisms upon the observation of a single instance of a spreading process. The strategy is based on the observation of the order in which network nodes are infected, and on its correlations with their local topology: these correlations differ between processes of simple contagion, processes involving threshold mechanisms, and processes driven by group interactions (i.e., by "higher-order" mechanisms). Our results improve our understanding of contagion processes and provide a method using only limited information to distinguish between several possible contagion mechanisms.

DOI: 10.1103/PhysRevLett.130.247401

Many phenomena can be described as contagions, such as disease spreading, information diffusion, or propagation of social behaviors [1–6]. Modeling contagion processes in a population typically includes two main steps. First, one describes how the state of the hosts (individuals who receive and propagate the disease, or information, or behavior) can evolve. For instance, one often assumes that they can only be in one of few possible states, such as susceptible (healthy), infectious (having the disease or information and able to transmit it), or recovered (cured and immunized). Second, the propagation is described along the structure of interactions between hosts, often encoded through a network in which nodes represent hosts and links represent their interactions. The resulting network epidemiology framework has been applied to the spread of human and computer viruses [4,6,7], rumors [8,9], innovations [10–14], or behavior [15].

Depending on the phenomenon, the fundamental propagation mechanisms are different. To describe the spread of infectious diseases, models of simple contagion, in which it is enough to have a single interaction between a susceptible and an infectious to lead to a transmission event, are adequate [1,6]. In social contagion of behaviors, peer influence and reinforcement mechanisms can play an important role, and empirical evidence indicates that single interactions are not sufficient to cause transmission [15–20]. These cases are hence better described by so-called complex contagion models, in which each transmission event requires interactions with multiple infectious hosts [12,18,21–23].

For both simple and complex contagions, most studies start from a propagation mechanism and design models to represent it and study how the structure of the interaction network impacts the spread [6]. In general, these investigations focus on averages over realizations of the process, and compare the phenomenology of processes and how they depend on the network structure. However, when empirical data related to a spreading process are observed, it concerns a single instance and one cannot average over multiple instances to obtain overall statistics. Therefore, here we address the issue of determining, from the observation of a single instance of a contagion process on a network, whether it is governed by a simple or complex contagion, and whether threshold or group effects are involved. Previous works have tried to identify the footprint of different contagion models on real or simulated processes. Evidence of complex contagion has been found in real data, observing the temporal evolution of the number of infectious [18,20,24,25], or by investigating how the contagion probability of a node depends on infectious neighbors [17–19,26]. Other rather data demanding techniques involve using deep learning [27] or comparing spreading processes on different network structures [15,28]. However, we still lack a clear identification of the main features distinguishing simple and complex processes.

Here, we put forward a new method based on the correlations between the order in which successive nodes in the network are reached by the spread and their basic local properties. We show that paradigmatic models of simple and complex contagion lead to different correlation patterns, and how to exploit them to build a classification tool able to determine whether a given instance of a spread is due to a simple contagion, a threshold model, or a higher-order contagion model. We investigate the

 

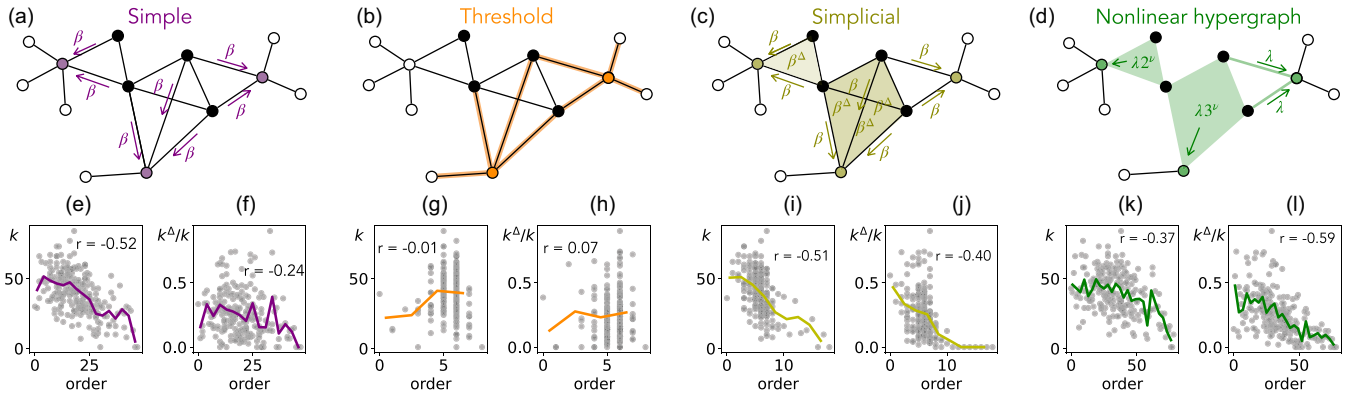


FIG. 1. The first row reports a toy network at an intermediate stage of the process (in which contagion events have taken place, resulting in four infected nodes, shown in black) and how the different models of propagation would imply contagion of further nodes (colored). Contagion events occur, respectively, (a) in simple contagion, along the network edges, with probability $\beta$ per unit time for each edge; (b) when a susceptible node sees a fraction of infected neighbors that is above a threshold $\theta$ (here $\theta = 0.5$); (c) both along network edges (rate $\beta$) and if a susceptible is part of a simplex of three nodes in which the two others are infected (rate $\beta^{\Delta}$); (d) along hyperedges, a susceptible node sharing a hyperedge with $n$ infected becoming infected at rate $\lambda n^{\nu}$. The second row gives scatterplots of $k_i$ vs $o_i$ and $k_i^{\Delta}/k_i$ vs $o_i$ for single numerical realizations of each model on the workplace dataset, where $o_i$ is the order in which the node $i$, with degree $k_i$ and belonging to $k_i^{\Delta}$ hyperedges of size 3, has been reached by the propagation in that realization. The values of the corresponding correlation coefficients are given in the plots. Parameters: $\beta = 0.005$ in panels (e) and (f), $\theta = 0.007$ in (g) and (h), $\beta = 0.005$, $\beta^{\Delta} = 0.8$ in (i) and (j), $\nu = 4$, $\lambda = 0.001$ in (k) and (l). The colored curves report the mean of all $k_i$ or $k_i^{\Delta}/k_i$ for each $o_i$.

robustness of our classifier with respect to incomplete data, and the possibility to apply it to a process taking place on an unknown network, i.e., different from the one(s) on which the classifier has been calibrated, as knowing the detailed structure of the network on which a contagion occurs is often challenging [29–31].

*Models of contagion.*—We consider four contagion processes on networks. For simplicity, we use SI models, i.e., each node can only be in two states, susceptible (S) or infected (I), and infected nodes do not recover. We consider processes in discrete time, differing in the mechanism determining how a node can switch from the S to the I state.

We first consider a simple contagion process [Fig. 1(a)]: every susceptible node can be infected independently by each of its infected neighbors with a probability per unit time $\beta$. The disease spreads thus along the pairwise links among nodes.

The second process [Fig. 1(b)] is the deterministic threshold model [12]: a susceptible node becomes infected when the fraction of its neighbors that are infected reaches a threshold $\theta$, to mimic the fact that an individual may adopt an innovation only if enough friends are already adopters.

We also consider two models of complex contagion that involve higher-order contagion mechanisms, i.e., interactions among groups of nodes [32]. First, the simplicial model takes place on simplicial complexes and the higher-order contagion is regulated by a parameter $\beta^{\Delta}$ [Fig. 1(c)]. Second, the nonlinear hypergraph (NLH) model takes place on hypergraphs and is regulated by parameters $\lambda$ and $\nu$ [Fig. 1(d)]. See Appendix and Fig. 1 for details.

Given an observed single realization of one of these models on a network, our goal is to devise a method to determine which model it corresponds to. To this aim, we consider several empirical datasets as the substrates on which the processes unfold. We use data representing physical or online interactions between individuals in several contexts: a workplace [33], educational contexts [34–37], a scientific conference [33], a hospital [38], and an email dataset [39,40]. These data are temporally resolved but we consider the aggregated network $G_{\mathcal{D}}$, the aggregated hypergraph $H_{\mathcal{D}}$, and its projection on hyperedges of size at most 3 $H_{\mathcal{D}}^{3}$ (defined in the Appendix). The degree $k_i$ of an individual is the number of links involving $i$ in $G_{\mathcal{D}}$, while we denote by $k_i^{\Delta}$ the number of hyperedges of size 3 to which $i$ belongs in $H_{\mathcal{D}}^{3}$. We consider here for simplicity unweighted networks and hypergraphs, but each link or hyperedge can be weighted by the number of times that the corresponding interaction has been observed during the data collection. We discuss the case of weighted networks and hypergraphs in the Supplemental Material (SM). In the main text, we give mostly results obtained with the workplace dataset and refer to the SM for the other datasets.

*Results.*—Simple contagion processes on networks are characterized by hierarchical dynamics: large degree nodes are reached early, and a cascade follows towards small degree nodes [46]. In general, the order in which nodes are infected can be influenced by their degrees, as illustrated in Fig. 1 for single instances of each process. This is explored further in Fig. 2(a), where we show the distribution of the Spearman correlation $C_1 = \mathrm{corr}(o, k)$ between the order in which nodes are infected and their degree $k$ [47], computed for each numerical realization of each model and for a wide range of parameter values. The distributions are similar for the simple and higher-order contagion processes, with

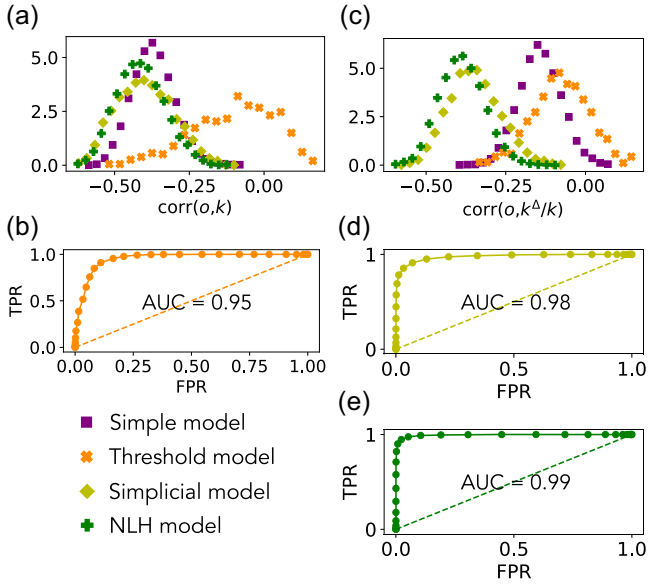


FIG. 2. Results for the workplace dataset. (a) Distributions of $C_1$ for the four contagion models. (b) ROC curve when using $C_1$ to classify threshold model processes against the other three. (c) Distributions of $C_2$ for the four contagion models. (d) ROC curve when using $C_2$ to classify simplicial against simple and threshold models. (e) ROC curve when using $C_2$ to classify NLH against simple and threshold models. For the stochastic models (simple, simplicial, and NLH) 1000 realizations are implemented for each parameter setting, always starting with one random infected node. For the deterministic threshold model we use only one realization for each different initial condition, i.e., one for each network node, for each parameter value. Parameters: $\beta \in \{0.005, 0.008, 0.014, 0.023, 0.039\}$ for both simple and simplicial models, $\beta^{\Delta} = 0.8$, $\lambda \in \{0.0001, 0.001, 0.006, 0.011, 0.015\}$, $\nu = 4$, $\theta \in \{0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10\}$.

similar ranges of (only negative) correlation values, while the distribution of $C_1$ obtained for the threshold process has a broader support including positive values. We thus consider the possibility to use the value of $C_1$ to identify whether a given realization results from the threshold model or from another model. We use the parametric Receiver Operating Characteristic (ROC) curve (see Appendix) to summarize the quality of the classifier as the area under the ROC curve (AUC), which is 0.5 for a random classification and 1 for a perfect one. The ROC curve for the workplace dataset is shown in Fig. 2(b), with a very high AUC of 0.95.

To identify processes involving higher-order mechanisms, we need to take into account the participation of nodes to higher-order structures. We thus consider the correlation between the order in which nodes are infected and the ratio $k^{\Delta}/k$ between the number of hyperedges of size 3 to which they belong and their degree: $C_2 = \mathrm{corr}(o, k^{\Delta}/k)$ [48]. Figures 1(f), 1(h), 1(j), and 1(l) illustrate the correlation $C_2$ on specific instances of each process, and Fig. 2(c) shows its distribution over multiple instances. The distributions are similar for the simplicial and NLH contagion cases on the one hand, and for the simple and threshold models on the other hand. Very good classification performances are attained, as quantified by the ROC curves obtained when using $C_2$ to classify instances of the simplicial model [Fig. 2(d)] or of the NLH model [Fig. 2(e)] against instances of simple and threshold processes. We study in the SM [40] how this performance depends on the model parameters.

We now combine the previous results using $C_1$ and $C_2$ to build a global classifier. We consider in addition the correlations between the order $o$ of infection of nodes with $k^{\Delta}$ and with $k^{|}$, their number of purely pairwise links (excluding connections part of higher-order interactions): respectively, $C_3 = \mathrm{corr}(o, k^{\Delta})$ and $C_4 = \mathrm{corr}(o, k^{|})$. We use a Random Forest (RF) classifier [49], a standard machine learning method, to perform the overall classification of instances of the four models. The performance of this classification task can be assessed by the confusion matrix depicted in Fig. 3(a): it gives in row $x$ and column $y$ the number of instances of a model $x$ that are classified as

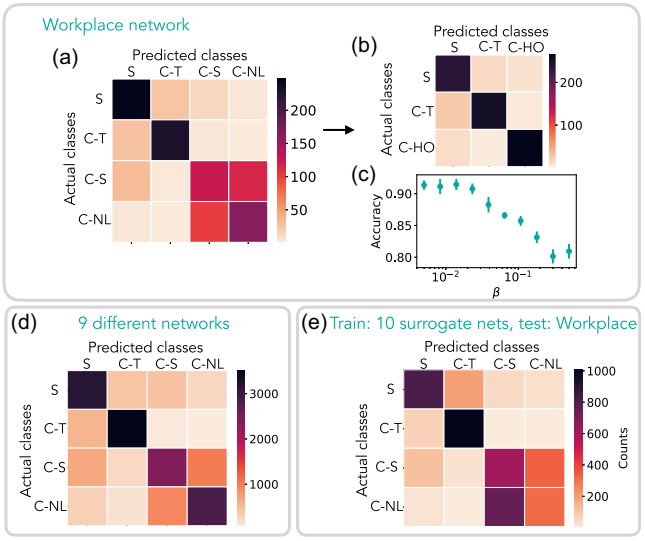


FIG. 3. (a) Confusion matrix with four classes: S (simple), C-S (complex simplicial), C-NL (complex NLH), and C-T (complex threshold). 3288 instances used for training and 1096 for testing (with approximately the same number of instances for each model). (b) Confusion matrix merging C-S and C-NL into C-HO (complex higher-order). 2466 instances used for training and 822 for testing. (c) Accuracy of RF classification with classes S, C-T, C-HO at varying $\beta$ in simple and simplicial. 2247 instances used for training and 750 for testing for each value of $\beta$. (d) Confusion matrix obtained by combining results on nine different networks. The relative accuracy assembling in classes S, C-T, C-HO is 0.84. 51 609 instances used for training and 17 203 for testing. (e) Confusion matrix obtained by training the classifier on ten surrogate networks (obtained with SDC method, see Ref. [40]) with statistical properties similar to the workplace dataset and testing using processes run on the real data. The accuracy when considering three classes S, C-T, C-HO is 0.85. 33 220 instances used for training and 4384 for testing. The parameters are set as in Fig. 2.

resulting from model $y$. Simple and threshold model instances are identified almost perfectly, while simplicial and NLH model instances are confused more easily. Merging the higher-order model realizations as one unique class of Complex higher-order (C-HO) processes results in a very high accuracy [Figs. 3(b) and 3(c)]. The classifier yields similar results for simulations implemented on each of the nine interaction networks we consider [40].

Up to now, we have considered that the classifier is trained using data coming from one network, and applied on processes run on the same network, and that the order of contagion and the local properties of all nodes are known. We now examine less idealized conditions. In particular, we start by relaxing the hypothesis of a full knowledge of the network structure, as measuring the full detailed structure of networks on which spreading processes can occur, such as contact networks, is a much more challenging task [29,31] than getting information on purely local properties [50]. First, we examine the case in which instances of the contagion processes occurring on different datasets are mixed: one can thus consider that the process to be classified has taken place on an unknown network, but that network has been used among others to train the classifier. The resulting accuracy is still very high [Fig. 3(d)] for the distinction between simple, threshold-based, and higher-order based contagion processes. To understand further the generalizability of the classifier, we also consider the case of a process observed on a completely "new" network, while the classifier has been trained using processes run on other network(s). A first case consists in using one or several of the available datasets for training, and another for testing. The resulting accuracy depends on the datasets chosen for training and testing [40], and remains high in many cases, which indicates a certain generalizability. These results also have limitations: the unfolding of a spread depends on the network structure, so that a classifier trained using one network cannot be blindly applied to a completely different one. However, recent works have also shown that statistical properties describing a spreading process can be obtained even from limited information on the network it unfolds on [51,52]. We thus assume that the detailed structure of the dataset $H_{\mathcal{D}}$ on which the spread to be classified has occurred is unknown. However, using the information on the degrees in $H^3_{\mathcal{D}}$ (assumed known anyway, as they represent the minimum information needed to compute the correlations fed into the classifier), we can generate surrogate data, i.e., synthetic networks having similar degree distributions as $H^3_{\mathcal{D}}$. In addition, we envision the case in which the group structure of the data is known (e.g., classes in a school), as it is known to be relevant to spreading processes [51,52] and also build surrogate data reproducing this structure. We consider in the SM [40] three possible ways to built such surrogate datasets. In each case we train the classifier using processes run on the surrogate data and classify processes run on the original data. We show that large values of the accuracy are recovered, as long as the algorithm for creating the surrogate data performs sufficiently well [see Fig. 3(e) for the workplace dataset, and [40] for a more extensive analysis concerning all datasets and surrogate data types].

We finally report in the SM [40] results obtained when relaxing the assumption of a complete observation of the spreading instance. First, we assume that only a fraction of the nodes (chosen arbitrary at random) can be observed. The values of the correlations are thus computed using only the order of infection and degrees of the observed nodes (both for training and testing). In this case, the accuracy of the classifier remains high, with values above 0.7 even when only 20% of the nodes are observed. This can be understood by the robustness of the correlations when randomly removing a fraction of the data points. If instead only the order of infection and the degrees of the first $h$ infected are known, the performances are more impacted. To observe, e.g., the occurrence of a cascade from large to small degrees, the first data points might indeed not be sufficient, and having information beyond the initial phase brings more accurate results.

*Discussion.*—We have developed a framework for classifying contagion processes on networks through the observation of correlations between the order in which nodes are reached and their local structural properties. The classification task (i) uses only local information, without the need to access the whole network structure, (ii) does not use any information on the infection status of the nodes' neighbors, (iii) is applied on single instances of a process, and (iv) can distinguish between a simple contagion process, a process driven by a threshold mechanism, and a process involving contagion on higher-order interactions. The proposed classifier yields a very good accuracy on several real-world networks, remaining robust against partial observation of the process. Moreover, although it cannot be trained with and applied to processes run on networks with very different properties, it can be applied on a process occurring on an unknown network as long as it is possible to generate surrogate data with similar statistical properties to produce the training set.

Our Letter has several limitations worth discussing. First, we have assumed to have access to the precise values of the degrees for all observed nodes, as well as the precise ordering of infection. To be more realistic, one could simulate the use of estimated values by inserting noise in the degree values. As the classifier is based on the measure of correlations, we expect that its accuracy should not be strongly impacted. It could however be affected by, e.g., extreme errors such as hubs classified mistakenly as low degree nodes or vice versa. Second, we have considered SI models, where nodes do not recover, and all nodes are finally reached if the network is connected. More realistic models consider that nodes do not remain

infectious at all times. A preliminary study using the SIR framework yields similar results [40], but we leave further investigations of the role of the recovery parameters and of more complex models to future works (e.g., including a latency period and/or asymptomatic state).

Finally, we have considered a limited series of datasets as substrate, and the classifier performance depends on the network characteristics and on the networks used to produce the training set. Other datasets might have different properties, such as, e.g., geometric embeddings, which could impact the spreading properties; additional features might then be added to the classifier. Overall, it is expected that the classifier cannot be fully general (trained using a randomly chosen network and tested on another one), since the properties of a spread depend on the network's structure. As obtaining detailed knowledge of the structure of networks on which spreading processes occur is challenging [29,31], we have started to explore the generalizability of the classifier to a new dataset, showing the possibility to train it using surrogate datasets which respect the known statistical properties of the new data. These considerations open the door to future studies investigating how different network structures affect the classifier's performance, which properties are the most important for building surrogate data, and to find new algorithms to this aim.

A. B. and M. M. acknowledge support from the Agence Nationale de la Recherche (ANR) project DATAREDUX (ANR-19-CE46-0008).

*Appendix A: Higher-order models of contagion.*—Higher-order models of contagion can be defined on hypergraphs or simplicial complexes, in which a hyperedge of size $m$ represents an interaction among a group of $m$ nodes (simplicial complexes are hypergraphs $\mathcal{H}$ such that, for each hyperedge—simplex—$e = \{i_1, \ldots, i_m\}$, all subsets of $e$ are also hyperedges of $\mathcal{H}$). We use the simplicial contagion [21] model, considering simplices up to the second order, i.e., interactions between three nodes, and neglecting structures of higher orders (which thus appear only as decomposed into second order simplices). Each susceptible node can be infected by an infectious neighbor with rate $\beta$ (as in the simple contagion), but also if it belongs to a simplex of three nodes in which the two others are infectious. This second process happens with rate $\beta^{\Delta}$ [Fig. 1(c)] and is thus specific to the existence of hyperedges (no such process takes place on "empty" triangles which are cliques in the projected networks but not hyperedges). In addition, we consider the nonlinear hypergraph (NLH) model [23], which includes contagion processes in interactions of arbitrary sizes: if in a hyperedge of size $m$ there are $n$ infected individuals, each of the remaining $m-n$ susceptible nodes is independently infected with probability $\lambda n^{\nu}$ at each time step [Fig. 1(d)], with $\lambda$ and $\nu$ free parameters. The case $\nu = 1$ reduces to a simple contagion, while the nonlinearity for $\nu \neq 1$ leads to reinforcement (for $\nu > 1$) or inhibition (for $\nu < 1$) effects and thus to a complex contagion phenomenology, as explored in [23].

*Appendix B: Aggregated networks, hypergraphs, and degrees.*—All the data that we use are temporally resolved, giving the specific time of each interaction and, for each dataset $\mathcal{D}$, we consider the aggregated network $G_{\mathcal{D}}$: each link in $G_{\mathcal{D}}$ represents the fact that the two corresponding individuals have been in contact at least once during the data collection. The degree $k_i$ of an individual is then the number of links involving $i$ in $G_{\mathcal{D}}$. Similarly, we define the aggregated hypergraph $H_{\mathcal{D}}$: a hyperedge of size $m$ represents a simultaneous group of $m$ nodes observed at least once: the availability of temporally resolved data allows us thus to distinguish in the aggregated network between hyperedges and cliques. We also consider the hypergraph $H^3_{\mathcal{D}}$ restricted to hyperedges of size at most 3: each hyperedge of larger size in $H_{\mathcal{D}}$ is simply decomposed into all its groups of nodes of size 3, and we denote by $k_i^{\Delta}$ the number of hyperedges of size 3 to which $i$ belongs in $H^3_{\mathcal{D}}$.

*Appendix C: ROC curve.*—The receiver operating characteristic curve is a parametric method to assess the possibility to classify data. Let us consider the case of classification between threshold and nonthreshold models observing instances of correlation $C_1$. The curve is built as follows: given the parameter $c \in [-1, 1]$, we classify each instance having $C_1 \geq c$ as resulting from a threshold model. If the instance was really produced by the threshold model, it is a true positive (TP), and else a false positive (FP). If the instance instead has a correlation $C_1 < c$, it is classified as resulting from one of the simple, simplicial, or NLH processes: if this is correct, it is a true negative (TN), while if it was a threshold process it is a false negative (FN). The ROC curve presents, as $c$ varies, the true positive ratio $\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, i.e., the fraction of instances of threshold model that are correctly classified, versus the false positive ratio $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$, i.e., the fraction of instances of the other models that are wrongly classified. This example of classification leads to the resulting curves reported in Figs. 2(b), 2(d), and 2(e).

[1] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, Oxford, 1992).
[2] M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, Princeton, NJ, 2011).
[3] D. Centola and M. W. Macy, Complex contagions and the weakness of long ties, Am. J. Sociol. **113**, 702 (2007).
[4] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical Processes on Complex Networks* (Cambridge University Press, Leiden, 2008).

[5] C. Castellano, S. Fortunato, and V. Loreto, Statistical physics of social dynamics, Rev. Mod. Phys. **81**, 591 (2009).
[6] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Epidemic processes in complex networks, Rev. Mod. Phys. **87**, 925 (2015).
[7] R. Pastor-Satorras and A. Vespignani, Epidemic Spreading in Scale-Free Networks, Phys. Rev. Lett. **86**, 3200 (2001).
[8] D. J. Daley and D. G. Kendall, Epidemics and rumours, Nature (London) **204**, 1118 (1964).
[9] Y. Moreno, M. Nekovee, and A. F. Pacheco, Dynamics of rumor spreading in complex networks, Phys. Rev. E **69**, 066130 (2004).
[10] T. W. Valente, *Network Models of the Diffusion of Innovations* (Hampton Press, Cresskill, 1995).
[11] E. Rogers, *Diffusion of Innovations, 5th Edition* (Free Press, New York, 2003).
[12] D. J. Watts, A simple model of global cascades on random networks, Proc. Natl. Acad. Sci. U.S.A. **99**, 5766 (2002).
[13] D. J. Watts and P. S. Dodds, Influentials, networks, and public opinion formation, J. Consum. Res. **34**, 441 (2007).
[14] I. Iacopini, S. Milojević, and V. Latora, Network Dynamics of Innovation Processes, Phys. Rev. Lett. **120**, 048301 (2018).
[15] D. Centola, The spread of behavior in an online social network experiment, Science **329**, 1194 (2010).
[16] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, Competition among memes in a world with limited attention, Sci. Rep. **2**, 335 (2012).
[17] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg, Structural diversity in social contagion, Proc. Natl. Acad. Sci. U.S.A. **109**, 5962 (2012).
[18] M. Karsai, G. Iniguez, K. Kaski, and J. Kertész, Complex contagion process in spreading of online innovation, J. R. Soc. Interface **11**, 20140694 (2014).
[19] B. Mønsted, P. Sapieżyński, E. Ferrara, and S. Lehmann, Evidence of complex contagion of information in social media: An experiment using twitter bots, PLoS One **12**, e0184148 (2017).
[20] D. Notarmuzi, C. Castellano, A. Flammini, D. Mazzilli, and F. Radicchi, Universality, criticality and complexity of information propagation in social media, Nat. Commun. **13**, 1308 (2022).
[21] I. Iacopini, G. Petri, A. Barrat, and V. Latora, Simplicial models of social contagion, Nat. Commun. **10**, 2485 (2019).
[22] G. Ferraz de Arruda, M. Tizzani, and Y. Moreno, Phase transitions and stability of dynamical processes on hypergraphs, Commun. Phys. **4**, 24 (2021).
[23] G. St-Onge, I. Iacopini, V. Latora, A. Barrat, G. Petri, A. Allard, and L. Hébert-Dufresne, Influential groups for seeding and sustaining nonlinear contagion in heterogeneous hypergraphs, Commun. Phys. **5**, 25 (2022).
[24] C. Fink, A. C. Schmidt, V. Barash, J. Kelly, C. Cameron, and M. Macy, Investigating the observability of complex contagion in empirical social networks, in *Proceedings of the Tenth International AAAI Conference on Web and Social Media* (AAAI, 2016).
[25] D. A. Sprague and T. House, Evidence for complex contagion models of social contagion from observational data, PLoS One **12**, e0180802 (2017).
[26] L. M. Aiello, A. Barrat, C. Cattuto, R. Schifanella, and G. Ruffo, Link creation and information spreading over social and communication ties in an interest-based online social network, EPJ Data Sci. **1**, 12 (2012).
[27] C. Murphy, E. Laurence, and A. Allard, Deep learning of contagion dynamics on complex networks, Nat. Commun. **12**, 4720 (2021).
[28] N. Horsevad, D. Mateo, R. E. Kooij, A. Barrat, and R. Bouffanais, Transition from simple to complex contagion in collective decision-making, Nat. Commun. **13**, 1442 (2022).
[29] K. Eames, S. Bansal, S. Frost, and S. Riley, Six challenges in measuring contact networks for use in modelling, Epidemics **10**, 72 (2015).
[30] M. Newman, Network structure from rich but noisy data, Nat. Phys. **14**, 542 (2018).
[31] N. Oliver *et al.*, Mobile phone data for informing public health actions across the covid-19 pandemic life cycle, Sci. Adv. **6**, eabc0764 (2020).
[32] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri, Networks beyond pairwise interactions: Structure and dynamics, Phys. Rep. **874**, 1 (2020).
[33] M. Génois and A. Barrat, Can co-location be used as a proxy for face-to-face contacts?, EPJ Data Sci. **7**, 11 (2018).
[34] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems, High-resolution measurements of face-to-face contact patterns in a primary school, PLoS One **6**, e23176 (2011).
[35] P. Sapiezynski, A. Stopczynski, D. D. Lassen, and S. Lehmann, Interaction data from the copenhagen networks study, Sci. Data **6**, 1 (2019).
[36] R. Mastrandrea, J. Fournet, and A. Barrat, Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys, PLoS One **10**, e0136497 (2015).
[37] D. J. Toth, M. Leecaster, W. B. Pettey, A. V. Gundlapalli, H. Gao, J. J. Rainey, A. Uzicanin, and M. H. Samore, The role of heterogeneity in contact timing and duration in network models of influenza spread in schools, J. R. Soc. Interface **12**, 20150279 (2015).
[38] P. Vanhems, A. Barrat, C. Cattuto, J.-F. Pinton, N. Khanafer, C. Régis, B.-a. Kim, B. Comte, and N. Voirin, Estimating potential infection transmission routes in hospital wards using wearable proximity sensors, PLoS One **8**, e73970 (2013).
[39] B. Klimt and Y. Yang, The enron corpus: A new dataset for email classification research, in *European Conference on Machine Learning* (Springer, New York, 2004), pp. 217–226.
[40] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.130.247401, which includes Refs. [41–45] for details about the data sets and contagion models used, as well as results on all the data sets, sensitivity analysis and details on the creation of surrogate data sets.

[41] L. Breiman, Random forests, Mach. Learn. **45**, 5 (2001).
[42] M. Mancastroppa, I. Iacopini, G. Petri, and A. Barrat, Hyper-cores promote localization and efficient seeding in higher-order processes, arXiv:2301.04235.
[43] N. W. Landry and J. G. Restrepo, Hypergraph assortativity: A dynamical systems perspective, Chaos **32**, 053113 (2022).
[44] M. Génois, C. Vestergaard, C. Cattuto, and A. Barrat, Compensating for population sampling in simulations of epidemic spread on temporal contact networks, Nat. Commun. **8**, 8860 (2015).
[45] N. Ruggeri, F. Battiston, and C. De Bacco, A principled, flexible and efficient framework for hypergraph benchmarking, arXiv:2212.08593.
[46] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, Velocity and Hierarchical Spread of Epidemic Outbreaks in Scale-Free Networks, Phys. Rev. Lett. **92**, 178701 (2004).
[47] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. **100**, 441 (1987).
[48] It is in principle possible to consider the number of hyperedges of larger sizes to which each node participates; for the sake of simplicity, we limit here to size 3, and, if a node belongs to a hyperedge of size $m > 3$, we decompose the hyperedge into triangles to compute $k^{\Delta}$.
[49] T. K. Ho, Random decision forests, in *Proceedings of 3rd International Conference on Document Analysis and Recognition* (IEEE, New York, 1995), Vol. 1, pp. 278–282.
[50] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, and W. J. Edmunds, Social contacts and mixing patterns relevant to the spread of infectious diseases, PLoS Med. **5**, e74 (2008).
[51] A. Machens, F. Gesualdo, C. Rizzo, A. E. Tozzi, A. Barrat, and C. Cattuto, An infectious disease model on empirical networks of human contact: Bridging the gap between dynamic network data and contact matrices, BMC Infect. Dis. **13**, 185 (2013).
[52] D. A. Contreras, E. Colosi, G. Bassignana, V. Colizza, and A. Barrat, Impact of contact data resolution on the evaluation of interventions in mathematical models of infectious diseases, J. R. Soc. Interface **19**, 20220164 (2022).