HAL open science

Influential groups for seeding and sustaining nonlinear contagion in heterogeneous hypergraphs

Guillaume St-Onge, Iacopo Iacopini, Vito Latora, Alain Barrat, Giovanni Petri, Antoine Allard, Laurent Hébert-Dufresne

▶ **To cite this version:**

Guillaume St-Onge, Iacopo Iacopini, Vito Latora, Alain Barrat, Giovanni Petri, et al.. Influential groups for seeding and sustaining nonlinear contagion in heterogeneous hypergraphs. Communications Physics, 2022, 5 (1), pp.25. 10.1038/s42005-021-00788-w . hal-03533806

HAL Id: hal-03533806

https://hal.science/hal-03533806v1

Submitted on 19 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Influential groups for seeding and sustaining nonlinear contagion in heterogeneous hypergraphs

Guillaume St-Onge,[1,2,*] Iacopo Iacopini,[3,4,5,6] Vito Latora,[6,7,8] Alain Barrat,[4,9] Giovanni Petri,[10,11] Antoine Allard,[1,2,12] and Laurent Hébert-Dufresne[1,12,13,†]

[1] *Département de physique, de génie physique et d'optique, Université Laval, Québec (Québec), Canada, G1V 0A6*

[2] *Centre interdisciplinaire en modélisation mathématique, Université Laval, Québec (Québec), Canada, G1V 0A6*

[3] *Department of Network and Data Science, Central European University, 1100 Vienna, Austria*

[4] *Aix Marseille Univ, Université de Toulon, CNRS, CPT, Marseille, France*

[5] *Centre for Advanced Spatial Analysis, University College London, London, W1T 4TJ, United Kingdom*

[6] *School of Mathematical Sciences, Queen Mary University of London. Mile End Road, London E1 4NS, United Kingdom*

[7] *Dipartimento di Fisica ed Astronomia, Università di Catania and INFN, I-95123 Catania, Italy*

[8] *Complexity Science Hub, Josefstädter Strasse 39, A 1080 Vienna, Austria*

[9] *Tokyo Tech World Research Hub Initiative (WRHI), Tokyo Institute of Technology, Tokyo, Japan*

[10] *Mathematics and Complex Systems Research Area, ISI Foundation, Turin, Italy*

[11] *ISI Global Science Foundation, New York, USA*

[12] *Vermont Complex Systems Center, University of Vermont, Burlington, VT 05401, USA*

[13] *Department of Computer Science, University of Vermont, Burlington, VT 05401, USA*

(Dated: November 11, 2021)

Abstract

Contagion phenomena are often the results of multibody interactions—such as superspreading events or social reinforcement—describable as hypergraphs. We develop an approximate master equation framework to study contagions on hypergraphs with a heterogeneous structure in terms of group size (hyperedge cardinality) and of node membership (hyperdegree). By mapping multibody interactions to nonlinear infection rates, we demonstrate the influence of large groups in two ways. First, we characterize the phase transition, which can be continuous or discontinuous with a bistable regime. Our analytical expressions for the critical and tricritical points highlight the influence of the first three moments of the membership distribution. We also show that heterogeneous group sizes and nonlinear contagion promote a mesoscopic localization regime where contagion is sustained by the largest groups, thereby inhibiting bistability. Second, we formulate an optimal seeding problem for hypergraph contagion and compare two strategies: allocating seeds according to node or group properties. We find that, when the contagion is sufficiently nonlinear, groups are more effective seeds than individual hubs.

I. INTRODUCTION

Mathematical models of contagion processes enhance our understanding of spreading dynamics and our predictive capabilities [1]. To account for the interconnected nature of real-world systems, the last two decades of network science and computational epidemiology research have focused on modeling frameworks of increasing complexity [1–3]. From the spreading of infectious diseases to rumors and innovations [4–6], a crucial aspect remains the interplay between the contact patterns enabling transmission and the dynamics that unfolds on top. As a representation for these contacts, graphs have been widely exploited to better represent real-world patterns with increasing levels of accuracy [1, 2, 7].

While graphs remain a reference for the representation of complex systems, they come with a fundamental limitation: they can only encode pairwise interactions (represented by edges). Groups are instead the foundational units of many biological, ecological and social systems, whose processes can involve multibody interactions between any number of elements. To overcome this limitation, higher-order network representations [8] are becoming more and more popular [9–12] to describe the structure of interacting systems [13, 14], their growth [15–18] and dynamics in groups [19–21]. Simplicial complexes and hypergraphs can encode relationships and interactions between any number of elements. They have been used to investigate the implications of higher-order interactions for landmark dynamical processes like synchronization [22–24], diffusion [25–27], and other social

* guillaume.st-onge.4@ulaval.ca
† Laurent.Hebert-Dufresne@uvm.edu

dynamics [28–30]. Battiston *et al.* [11] provides a comprehensive review of early efforts in this direction.

Recently, a model of simplicial contagion has been proposed [31]. This is a standard Susceptible-Infected-Susceptible (SIS) compartmental model in which a susceptible individual can become infected through different transmission channels, beyond infectious edges. In models of simple contagion, the transition from susceptible to infected happens independently with each exposure to an infectious edge [1]. In models of complex contagion instead [32], the transition requires multiple infectious edges or is reinforced by multiple exposures, thus accounting for the empirically observed mechanisms of social influence [33–36]. In simplicial contagions, or more generally hypergraph contagions, a susceptible individual can become infected because of a multibody process, i.e., through an exposure to an infectious group [31]. In this way, the node is simultaneously exposed to the state of the entire group, whose effect can be interpreted as a mechanism of social reinforcement [33]. In addition, the study of higher-order contagion models has applications as well in biological sciences: it has indeed recently been shown that the combination of multibody interactions, heterogeneous temporal activity, and the concept of a minimal infective dose lead to nonlinear infection kernels in a model of biological contagions [37].

The analytical approaches derived so far have confirmed the rich phenomenology emerging from the contagion dynamics, characterized by discontinuous phase transitions, bistability and critical mass effects [31, 38–42]. Most approaches follow a heterogeneous mean-field (HMF) framework in which nodes are divided into hyperdegree classes. These descriptions are analytically tractable, but do not consider the details of the structure and ignore the dynamical correlations within groups, which are especially important for hypergraph contagions since multibody interactions naturally reinforce these correlations.

Other approaches like quenched mean-field theory (QMF) [43] and microscopic Markov-chains (MMCA) [44] can explicitly take the entire contact patterns into account. Along this line, the microscopic epidemic clique equations (MECLE) capture dynamical correlations as well, thereby highlighting the important impact these correlations have on critical points [45]. The analytical tractability of these approaches is, however, sacrificed in favor of a more precise description of the structure. To fully understand the consequences of multibody interactions in contagions on higher-order networks, we thus need a framework that is both analytically tractable and captures dynamical correlations.

In particular, we are interested in better understanding the notion of influence in hypergraph contagions. In classic contagion models on random networks, individual hubs are influential in the sense that they are the best seeds of contagions, but they are also the most apt at sustaining seeded contagions [46]. However, in hypergraph contagions we must consider the influence of both individuals and groups, because dynamical correlations can allow groups to be more influential than sets of uncorrelated hubs. Regimes of bistability and hysteresis also imply a potential decoupling between the ability of nodes to seed a contagion and their ability to sustain it. We thus wish to determine (i) which set of groups can best maintain the stationary state of a hypergraph contagion, and when this becomes a dominant effect (ii) which set of groups and their configuration offer optimal initial conditions for a contagion, as compared to the classic notion of influential spreaders as individual hubs, and (iii) whether or not these two notions of group influence are aligned.

In this work, we use approximate master equations (AMEs) [20, 21, 47–50] to study hypergraph contagions, capturing exactly the inner dynamics of groups. We consider a model where the infection rate is a nonlinear function of the number of infected nodes in groups. First, we provide a detailed characterization of the phase transition and derive analytical expressions for the critical and tricritical points. We find that large values for the third moment of the membership (hyperdegree) distribution suppress the emergence of a discontinuous phase transition. Furthermore, we show that heterogeneous group sizes and superlinear contagion facilitates the onset of a *mesoscopic localization* regime [49, 50], where contagion is sustained by the largest groups, and, incidentally, inhibit bistability. Second, we define and solve an influence maximization problem based on two strategies: allocating seeds according to either node individual properties or according to group properties. When the contagion is sufficiently nonlinear, the latter is more effective. Altogether, our results highlight the role of influential groups to drive both the stationary state of contagion on hypergraphs and its behavior in the transient state.

II. RESULTS

A. Hypergraph contagion model

In the original version of the simplicial contagion model [31], the spreading process takes place on a simplicial complex [see Fig. 1] and obeys the following rule. If all nodes in a d-simplex are infected except a susceptible one, this remaining node gets infected at a rate β_d, and also receives contributions from the lower-dimensional simplices included in the d-simplex with rates respectively equal to $\beta_{d-1}, \ldots, \beta_1$. For instance, in Fig. 1, two of the three nodes composing a 2-simplex are infected, hence the susceptible node gets infected at rate $2\beta_1 + \beta_2$, considering the contributions both from the two edges and from the "triangle".

A simplicial complex is a specific type of hypergraph, and thus we can always map the former on the latter—note that the reverse direction is not always possible.

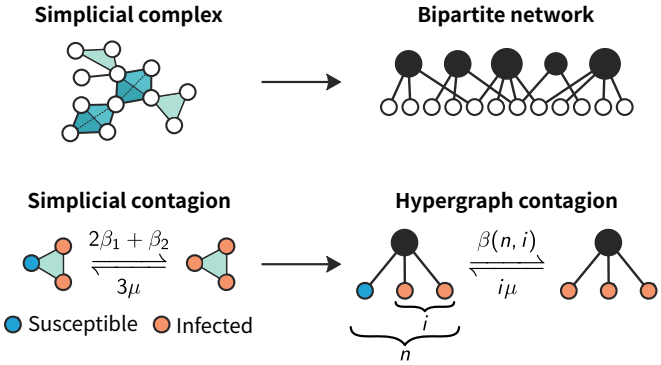

FIG. 1. **Mapping of the simplicial contagion model to a hypergraph contagion.** We use a bipartite representation, where nodes (white circles) belong to groups (black circles). A facet of dimension $n-1$ is mapped onto a group of size n. In the simplicial contagion model, contributions from higher-order interactions are taken into account by additional transmission rates (e.g., β_2 for the 2-simplex) when all but one node of the simplex are infected [31]. With our hypergraph representation, infections within a group are simply modeled by a general infection function $\beta(n, i)$ that depends on both the size n of the group and the number i of infected nodes in the group (with $i \leq n$).

To do so, each facet—a simplex that is not a face of any larger simplex—is represented by a single hyperedge (group). In this paper, we relax the requirement of the simplicial complex in favor of a more general hypergraph structure. We make use of the bipartite representation of hypergraphs [8, 11], in which the two sets of nodes of the bipartite graph correspond respectively to the sets of nodes and groups of the original hypergraph, as illustrated in Fig. 1.

The *size* n of a group corresponds to the number of nodes belonging to this group, and it is therefore equivalent to the *hyperedge cardinality*. Note that a d-simplex consists of $d+1$ nodes and is therefore mapped onto a group of size $n = d+1$. Similarly, the *membership* m of a node, the *node hyperdegree*, corresponds to the number of groups to which it belongs, regardless of their size.

The hypergraph contagion model is defined as follows: for a group of size n, where $i \leq n$ members are infected, each of the $n-i$ susceptible nodes gets infected at rate $\beta(n, i)$. For susceptible nodes that belong to multiple groups, their total transition rate to the infected state is simply the sum of the infection rates associated with each group to which they belong—in other words, the infection processes are independent. All infected nodes transition back to the susceptible state at the same constant recovery rate μ.

Notice that the hypergraph contagion model allows to represent any type of simplicial contagion. For instance, in the simplicial contagion model case of Fig. 1, we would use $\beta(3, 2) = 2\beta_1 + \beta_2$. In fact, the description offered by the infection rate function $\beta(n, i)$ yields a variety of models more general than the original simplicial contagion—in which a function $\beta(i)$ would be sufficient to encode contributions from facets of any dimension.

In all our case studies, we will use an infection rate function of the form

$$\beta(n, i) = \lambda i^{\nu} \ . \tag{1}$$

However, many results we derive hold true for a general infection rate function $\beta(n, i)$. The parameter ν controls the nonlinearity of the contagion. A linear contagion is recovered by setting $\nu = 1$, which is equivalent to a standard SIS model on networks, where each group is a clique [20]. We intentionally chose the infection rate function independent of n to focus on the impact of a nonlinear dependence on i; it would be straightforward to generalize the results by considering $\beta(n, i) \mapsto \Lambda(n) i^{\nu}$.

The infection rate function in Eq. (1) is the simplest nonlinear generalization of standard epidemiological models, where $\beta(n, i) \propto i$. Moreover, we can motivate the choice of exponents $\nu \neq 1$ in the context of social contagions, by comparing our approach to the original formulation of the simplicial contagion model. A value of $\beta_2 > 0$ in Fig. 1 represents social reinforcement [31], and to correctly map the infection rate for a triangle, we need to use an exponent $\nu > 1$ in our model. Similarly, a value $\beta_2 < 0$ represents social inhibition, and this case can be obtained with an exponent $\nu < 1$.

Another motivation for the infection rate function at Eq. (1) is a recent study that shows this general form emerges in the occurrence of heterogeneous temporal patterns [37]. More specifically, if you consider that the participation time of nodes—representing individuals—to higher-order interactions is distributed according to a power law, and that individuals become infected according to a threshold mechanism based on the dose received in the interaction, then the probability for a node to get infected in a group is $\propto i^{\nu}$, where ν is related to the temporal heterogeneity. In the continuous time limit, one recovers the infection rate function defined at Eq. (1).

The infection mechanism is motivated in the context of biological contagions [37], where the infective dose received could represent viral particles for instance, and the threshold would correspond to the minimal infective dose to develop a disease. While such type of complex contagions are rarely used in the context of biological contagions, they could help explain certain observed phenomena, such as super-exponential spread for certain diseases [51]. Moreover, threshold models are very common in social contagions [33, 52–54], thus Eq. (1) could be interpreted as an effective mechanism of social spread accounting for heterogeneous temporal patterns.

1. *Group-based Approximate Master Equations*

To describe hypergraph contagions, we make use of group-based Approximate Master Equations [20, 21, 49,

50]. This means that we do not rely on specific hypergraph realizations. Instead, we assume that the structure is drawn from a random hypergraph ensemble described by the distributions p_n, for the size n of a group, and g_m, for the membership m of a node. Each of the m membership stubs of a node is assigned uniformly at random to a group available spot. Therefore, the membership m of a node and the sizes of the groups to which it belongs are uncorrelated.

To track the evolution of a contagion process on this ensemble of hypergraphs, we define two sets of quantities: $s_m(t)$, representing the fraction of nodes with membership m that are susceptible at time t and $f_{n,i}(t)$, the fraction of groups of size n having i infected members at time t. The last quantity can also be interpreted as a conditional probability (of observing i infected nodes in a group of size n) satisfying the normalization condition $\sum_i f_{n,i} = 1$.

We further define two mean-field quantities. First, let us take a random susceptible node. The mean-field infection rate resulting from a random group to which it belongs is defined as

$$r(t) = \frac{\sum_{n,i} \beta(n,i)(n-i) f_{n,i} p_n}{\sum_{n,i} (n-i) f_{n,i} p_n} \ . \tag{2}$$

Indeed, the joint distribution for the size n and the number of infected nodes i in this group is proportional to $(n-i) f_{n,i} p_n$, and we just average $\beta(n,i)$ over this distribution.

Second, let us randomly choose a susceptible node inside a group. The mean-field infection rate caused by all the external groups to which the susceptible node belongs (excluding the one from which we picked the node) can be written as

$$\rho(t) = r(t) \frac{\sum_m m(m-1) s_m g_m}{\sum_m m s_m g_m} \ . \tag{3}$$

To obtain $\rho(t)$, we assume that infections coming from different groups are independent processes. We multiply $r(t)$ with the mean excess membership of a susceptible node, i.e., if we pick a susceptible node in a group, it is the expected number of other groups to which it belongs. Since the membership distribution of a susceptible node picked in a group is proportional to $m s_m(t) g_m$, we simply average $m-1$, its excess membership, over this distribution.

Using the definitions in Eqs. (2) and (3) we can write the following system of AMEs

$$\frac{\mathrm{d}s_m}{\mathrm{d}t} = \mu(1-s_m) - m r s_m \ , \tag{4a}$$

$$\begin{aligned} \frac{\mathrm{d}f_{n,i}}{\mathrm{d}t} =& \mu(i+1) f_{n,i+1} - \mu i f_{n,i} \\ &- (n-i)\left[\beta(n,i) + \rho\right] f_{n,i} \\ &+ (n-i+1)\left[\beta(n,i-1) + \rho\right] f_{n,i-1} \ . \end{aligned} \tag{4b}$$

This system is composed of $\mathcal{O}(n_{\max}^2 + m_{\max})$ equations. It is approximate in the sense that the evolution of the fractions of infected nodes of membership m, s_m, are treated in a mean-field fashion (still considering dynamic correlations between pairs node-group), while the evolution equations of the $f_{n,i}$ are treated as master equations. In the right-hand side of Eq. (4b), the first two terms are due to the recovery process, and the last two to the infection. The infection rate due to infected nodes inside the group is exact, while the infection rates associated to external groups (the terms making use of ρ) are treated in an approximate way. Without loss of generality (up to a change of time scale) we set $\mu \equiv 1$ for the remainder of the document.

From Eq. (4), we can calculate the global prevalence

$$I(t) = \sum_m [1 - s_m(t)] g_m \ , \tag{5}$$

and the group prevalence

$$I_n(t) = \sum_i \frac{i}{n} f_{n,i}(t) \ , \tag{6}$$

which correspond to the average fraction of infected nodes in the whole system and within groups of size n respectively.

In the stationary state, we obtain the following self-consistent relations:

$$s_m^* = \frac{1}{1 + m r^*} \ , \tag{7a}$$

$$\begin{aligned} (i+1) f_{n,i+1}^* =& \left\{ i + (n-i)\left[\beta(n,i) + \rho^*\right] \right\} f_{n,i}^* \\ &- (n-i+1)\left[\beta(n,i-1) + \rho^*\right] f_{n,i-1}^* \ . \end{aligned} \tag{7b}$$

The latter equation can be simplified by noting that $f_{n,i}^*$ must respect *detailed balance* in the stationary state, i.e.,

$$(i+1) f_{n,i+1}^* = (n-i)\left[\beta(n,i) + \rho^*\right] f_{n,i}^* \ ,$$

expressing that the probability to decrease the number of infectious nodes in a group of size n from $i+1$ to i by a recovery process is equal to the probability of the opposite change (from i to $i+1$ infectious nodes) obtained through a contagion event. We thus finally obtain

$$f_{n,i}^* = f_{n,0}^* \frac{n!}{(n-i)!i!} \prod_{j=0}^{i-1} \left[\beta(n,j) + \rho^*\right] \quad \forall\, i \in \{1, \ldots, n\} \ , \tag{8}$$

where $f_{n,0} = 1 - \sum_{i>0} f_{n,i}$.

2. *Comparison with simulations*

We provide a comparison of our approach with Monte Carlo simulations. To do this, we consider empirical

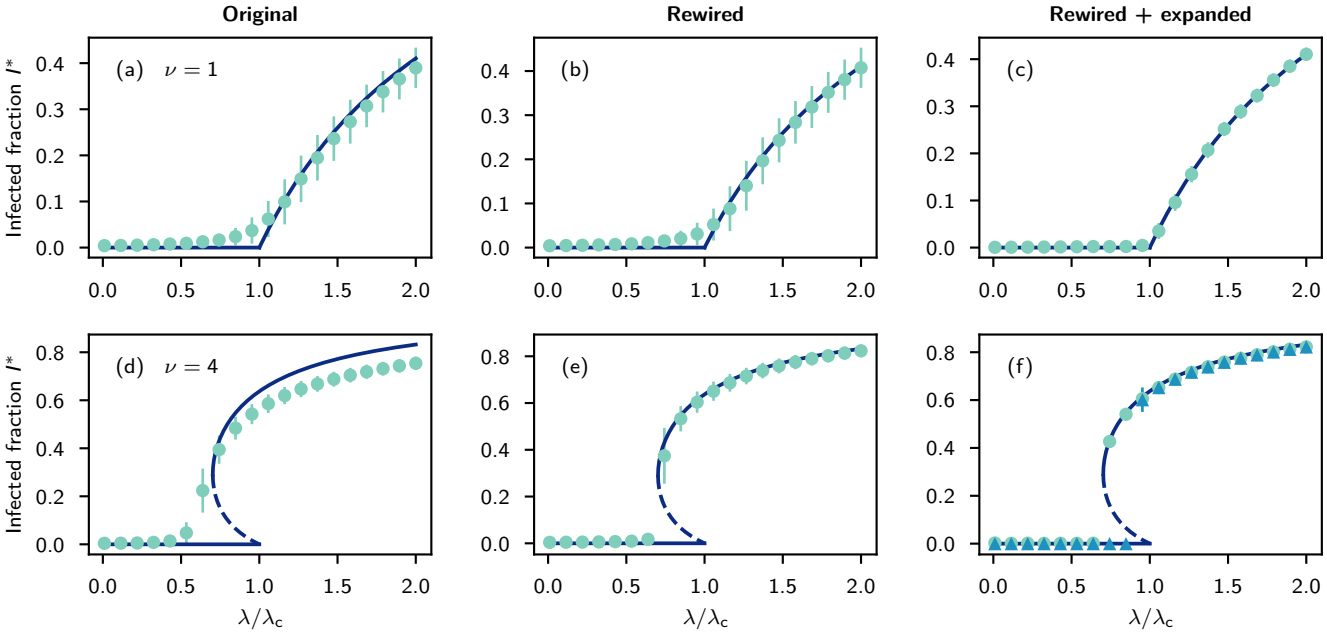

FIG. 2. **Phase transition on a hypergraph constructed from high-resolution face-to-face contact data.** See "Contagion on real-world hypergraphs" in the Methods. The hypergraph contains 242 nodes and 1188 groups. Both the membership and the group size distributions are homogeneous, with $\langle m \rangle \approx 11.79$, $\langle n \rangle \approx 2.40$, $m_{\max} = 32$ and $n_{\max} = 5$. We compare the results of Monte Carlo simulations (symbols) with the predictions of our approach (solid and dashed lines for stable and unstable solutions respectively) for the fraction I^* of infectious nodes in the stationary state. Note that in the "Phase transition" subsection in the main text, we omit the asterisk for stationary quantities to simplify the notation. The infection rate function is $\beta(n, i) = \lambda i^\nu$. We rescale λ with the invasion threshold λ_c, which is computed using Eq. (15). The symbols represent the average infected fraction measured over long runs and averaged over randomized hypergraphs when this applies. The error bars (sometimes too small to be seen) correspond to one standard deviation. The green circles are obtained with the quasistationary-state method, starting with a large fraction of infected nodes ($I = 0.8$) to sample the upper branch of the hysteresis loop when the phase transition is discontinuous. We either use the original hypergraph [panel (a) and (d)], 10 uniformly randomized versions of the original hypergraph [panel (b) and (e)] or 10 uniformly randomized versions of the original hypergraph with its size expanded by a factor 10 [panel (c) and (f)]. In panel (f), the blue triangles are obtained by ordinary simulations, starting with a small fraction of infected nodes ($I = 0.02$) to sample the lower branch of the hysteresis loop. We only do it for the expanded hypergraphs because finite-size effects makes unreliable the estimation of the lower branch with data of small size.

higher-order structures constructed from real-world data, and their randomized counterparts. Details on the simulations and the data collection and aggregation are provided in the subsection "Contagion on real-world hypergraphs" in the Methods.

The motivation for this *a priori* validation is threefold: first, it allows us to illustrate the validity and accuracy of our analytical framework when our assumptions are met—namely when the structure is drawn from an ensemble of uncorrelated hypergraphs with fixed g_m and p_n. Second, because of the excellent agreement with simulations for random hypergraphs, we omit further comparison with Monte Carlo simulations for the many results we present in the following sections. Finally, it showcases the possible sources of discrepancies—and how our results could vary—when considering real datasets.

In Fig. 2, we show the phase diagram, i.e., the fraction I^* of infectious nodes in the stationary state, of contagion dynamics on hypergraphs that encode higher-order social (face-to-face) interactions between individuals from a primary school in Lyon. Both the membership and group size distributions are homogeneous for this dataset. We considered linear contagion ($\nu = 1$), equivalent to the standard SIS model, and superlinear contagion ($\nu = 4$). In both cases, our analytical formalism (continuous lines) agrees quite well with the Monte Carlo simulations (symbols) on the original (empirical) hypergraph [Figs. 2(a) and 2(d)]. The main source of errors for the observed discrepancy can be attributed to structural correlations, which do not appear to affect the threshold values, but reduce the stationary prevalence. Indeed, in Figs. 2(b) and 2(e), the agreement improves by randomizing the hypergraph while preserving the memberships and the group sizes. The remaining discrepancies are due to the fact that simulations are affected by finite-size effects, while our formalism assumes an infinite size system: the agreement becomes almost perfect in Figs. 2(c) and 2(f) by additionally increasing the size of the hypergraph by

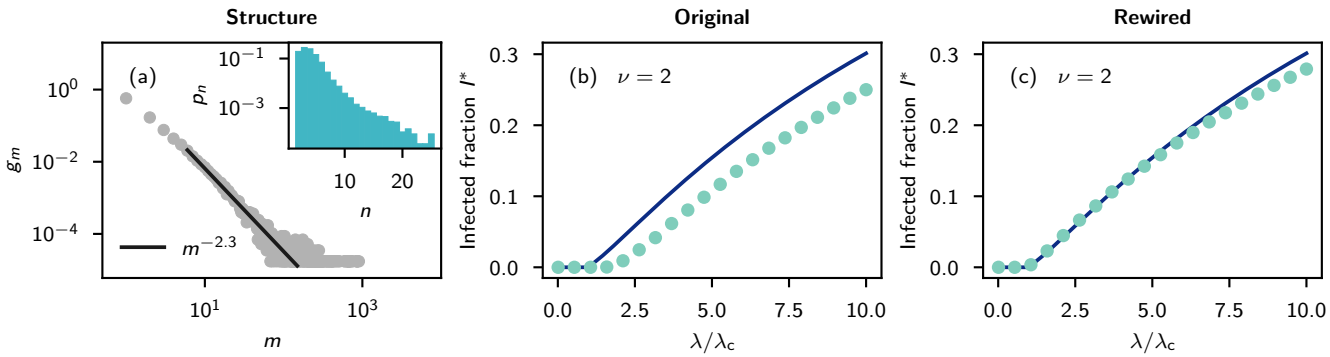

FIG. 3. **Phase transition on a hypergraph constructed from coauthorship data.** See "Contagion on real-world hypergraphs" in the Methods. The hypergraph has been obtained by using a breadth-first search to extract a subhypergraph of 57 501 nodes and 55 204 groups; the original data contained more than 10^6 nodes and groups. (a) The membership distribution is heterogeneous, approximately of the form $g_m \sim m^{-\gamma_m}$ with $\langle m \rangle \approx 3.75$ and $m_{\max} = 903$, while the group size distribution (inset) is more homogeneous, with $\langle n \rangle \approx 3.90$ and $n_{\max} = 25$. (b),(c) We compare the results of Monte Carlo simulations (circle markers) with the predictions of our approach (solid lines). The infection rate function is $\beta(n, i) = \lambda i^\nu$. We rescale λ with the invasion threshold λ_c, which is computed using Eq. (15). The infected fraction I^* has been obtained as averages over time with long runs (and averaging over randomized hypergraphs when this applies). The error bars (too small to be seen) correspond to one standard deviation. We used ordinary simulations, starting with $I = 0.8$. (b) We use the original hypergraph. (c) We use 10 uniformly randomized versions of the original hypergraph.

a factor 10.

Let us remark that the error is more important for $\nu = 4$ on the original hypergraph [Fig. 2(d)], which suggests that structural correlations have a greater effect on nonlinear contagions. In the Supplementary Note 1, we show how to generalize our framework to account for structural correlations.

We also considered a completely different dataset, which represents coauthorship relations in computer science publications, obtained from major journals and proceedings in the field. The resulting hypergraph is considerably larger than the previous one, and it also presents a very different structure. The results are shown in Fig. 3, where we plot the same phase diagram curves as in Fig. 2, using a superlinear contagion ($\nu = 2$). In this case, however, the membership distribution is heterogeneous [Fig. 3(a)], approximately of the form $g_m \sim m^{-\gamma_m}$ with $\gamma_m \approx 2.3$, while the group size distribution is more homogeneous [inset of Fig. 3(a)], but still extends to rather large values with $n_{\max} = 25$. By comparing the results for the original hypergraph [Fig. 3(b)] against those for a randomized ensemble [Fig. 3(c)], we see that structural correlations account for the major part of the discrepancies between simulations and theory, affecting both the invasion threshold and the stationary prevalence. However, Fig. 3(c) shows that structural correlations are not the only source of errors at high prevalence.

Let us recall that our formalism correctly captures the dynamical correlations within a group through a master equation description [Eq. (4b)] of the possible states, but it does not capture the dynamical correlations around nodes, since we use a heterogeneous mean-field description [Eq. (4a)]. These correlations become especially important in the presence of hubs with a large membership, which is the case for the hypergraph considered in Fig. 3. In fact, when we use the same group size distribution as in Fig. 3, but with a more homogeneous membership distribution, the discrepancy at high prevalence disappears (see the Supplementary Note 2).

B. Phase transition

In this section, we unveil the important role of influential groups on the phase transition of hypergraph contagions. We first derive general expression for critical points, marking the limit of the domain of validity of a stationary solution. Secondly, we obtain an expression for tricritical points, indicating when the phase transition switches from continuous to discontinuous, and a bistable regime appears. These results are valid for any infection rate function such that $\beta(n, i) > 0$ for all $i > 0$ (see Supplementary Note 3 for a consideration of threshold models).

We then concentrate our study on the infection rate function $\beta(n, i) = \lambda i^\nu$. This allows us to define important threshold values, the *invasion threshold* λ_c above which the disease-free solution $I^* = 0$ is unstable, and the *bistability threshold* ν_c, the smallest nonlinear exponent allowing for a discontinuous phase transition. In particular, we will show how heterogeneous membership and group size distributions alter these thresholds, especially in the presence of a mesoscopic localization driven by influential groups.

In the following, we assume that the system has

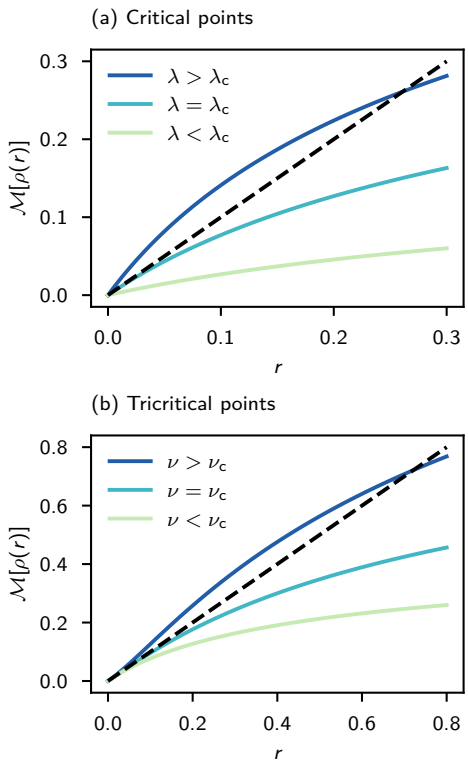

FIG. 4. **Critical behavior of the function $\mathcal{M}[\rho(r)]$.** Each intersection with the dashed line represents a stationary solution of Eq. (7). Results refer to the infection rate function $\beta(n,i) = \lambda i^{\nu}$ and a hypergraph with $g_m = (\delta_{m,2}+\delta_{m,3})/2$ and $p_n = (\delta_{n,4}+\delta_{n,5})/2$. (a) We fix $\nu = 1$ and vary λ. For $\lambda > \lambda_c$, a nontrivial solution emerges. (b) We fix $\lambda = \lambda_c$ and vary ν. Note that the slope of all solid lines is 1 at the origin. A nonlinear exponent $\nu > \nu_c$ is associated with a discontinuous phase transition, since $\partial_r^2 \mathcal{M} > 0$ in this case.

reached the stationary state and we drop the asterisk to simplify the notation throughout this section.

1. *Critical points and the invasion threshold*

Equations (7a) and (8) imply that each s_m and $f_{n,i}$ can be written in terms of r and ρ. In turn, r and ρ can be written in terms of s_m and $f_{n,i}$ through Eqs. (2) and (3), which means the stationary solutions are determined by a set of self-consistent equations.

Satisfying all self-consistent equations means we can reexpress all quantities $(s_m, f_{n,i}, \rho, r)$ as functions of a single mean-field quantity, which we choose to be r. $s_m(r)$ is given by Eq. (7a), and $\rho(r)$ is given by Eq. (3), which is rewritten as

$$\rho(r) = r\frac{\sum_m m(m-1)s_m(r)g_m}{\sum_m m s_m(r) g_m} \,. \tag{9}$$

$f_{n,i}$ is more simply written as a composite function, $f_{n,i}[\rho(r)]$, given by Eq. (8). Finally, r itself must satisfy Eq. (2), which we can write as $r = \mathcal{M}[\rho(r)]$ where

$$\mathcal{M}(\rho) = \frac{\sum_{n,i} \beta(n,i)(n-i) f_{n,i}(\rho) p_n}{\sum_{n,i}(n-i) f_{n,i}(\rho) p_n} \,. \tag{10}$$

This relation is used to solve numerically for the fixed points and evaluate their stability (see the Supplementary Note 4).

In Fig. 4, the stationary solutions correspond to the intersections of $\mathcal{M}[\rho(r)]$ (solid lines) and r (dashed line). We see that $r = 0$ is always a solution, while the solution $r > 0$ only exists for certain values of the parameter λ. This indicates the presence of a critical point.

Critical points mark the limit of the domain of validity of a solution to the equation $r = \mathcal{M}[\rho(r)]$. They arise when r is tangent to $\mathcal{M}[\rho(r)]$, which implies

$$\frac{\mathrm{d}\mathcal{M}}{\mathrm{d}r} = \frac{\mathrm{d}\mathcal{M}}{\mathrm{d}\rho}\frac{\mathrm{d}\rho}{\mathrm{d}r} = 1 \,, \tag{11}$$

as can be seen in Fig. 4(a), where $\mathcal{M}[\rho(r)]$ is tangent to r at the point $r = 0$ for some value $\lambda = \lambda_c$.

When the tangent point $r > 0$, Eq. (11) needs to be solved numerically. However when the solution $r \to 0$, we are able to obtain an analytical expression. In this limit, we expect to have $s_m \to 1$, $\rho \to 0$ and $f_{n,i} \to \delta_{i,0}\,\forall n$, i.e., all nodes are susceptible. Therefore, from Eq. (9) we have

$$\left.\frac{\mathrm{d}\rho}{\mathrm{d}r}\right|_{r\to 0} = \frac{\langle m(m-1)\rangle}{\langle m\rangle} \,, \tag{12}$$

where $\langle\cdots\rangle$ stands for the expectation value with respect either to g_m or p_n. From Eq. (10) (and using the fact that $\beta(n,0) \equiv 0$) we also obtain

$$\left.\frac{\mathrm{d}\mathcal{M}}{\mathrm{d}\rho}\right|_{\rho\to 0} = \frac{1}{\langle n\rangle}\sum_{n,i}\beta(n,i)(n-i)p_n \underbrace{\left.\frac{\mathrm{d}f_{n,i}}{\mathrm{d}\rho}\right|_{\rho\to 0}}_{h_{n,i}} \,. \tag{13}$$

To evaluate $h_{n,i}$, we apply the derivative $\mathrm{d}/\mathrm{d}\rho$ to Eq. (8). We obtain $h_{n,1} = n$ and

$$h_{n,i} = \frac{n!\prod_{j=1}^{i-1}\beta(n,j)}{(n-i)!\,i!} \quad \forall\, i \in \{2,\ldots,n\} \,, \tag{14}$$

Also, $\sum_i h_{n,i} \equiv 0$ as $\sum_i f_{n,i} = 1$, hence

$$h_{n,0} = -\sum_{i>0} h_{n,i} \,.$$

Combining Eqs (12), (13) and (14) with Eq. (11), we obtain the following implicit expression for critical points in the limit $r \to 0$

$$\frac{\langle m(m-1)\rangle}{\langle m\rangle\langle n\rangle}\left\langle \sum_{i>0}\frac{n!}{(n-i-1)!\,i!}\prod_{j=1}^{i}\beta(n,j)\right\rangle = 1 \,. \tag{15}$$

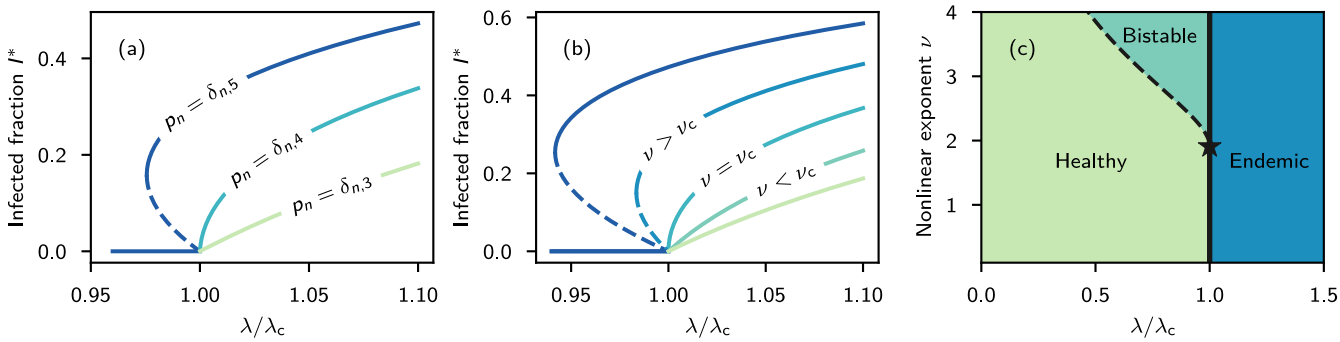

FIG. 5. **Phase transition in regular hypergraphs.** The membership and group size distributions are of the form $(g_m, p_n) = (\delta_{m,m_0}, \delta_{n,n_0})$. The infection rate function is $\beta(n,i) = \lambda i^\nu$. Solid and dashed lines in (a) and (b) represent stable and unstable solutions respectively for the stationary fraction of infected nodes. (a) $\nu = 1.8$, but different structures were used, with a node always having the same number of different neighbors. From bottom to top, we have $(g_m, p_n) = (\delta_{m,6}, \delta_{n,3})$, $(\delta_{m,4}, \delta_{n,4})$ and $(\delta_{m,3}, \delta_{n,5})$. Note that for the middle curve $(\delta_{m,4}, \delta_{n,4})$, $\nu_c \approx 1.81$. (b) The hypergraph structure is fixed with $g_m = \delta_{m,3}$ and $p_n = \delta_{n,4}$. Values of $\nu \in \{1.5, 1.7, \nu_c, 2.1, 2.3\}$ (bottom to top curves) were used, with $\nu_c \approx 1.91$. (c) Phase diagram for the same hypergraph as in (b). The dashed critical line ($\lambda = \lambda_p$) and the solid critical line ($\lambda = \lambda_c$) coalesce at the tricritical point (λ_c, ν_c) indicated by the star marker.

For the infection rate function $\beta(n,i) = \lambda i^\nu$, Eq. (15) defines the invasion threshold λ_c, i.e., the critical value of λ marking the limit of the validity for a solution $r \to 0$. This solution is not always stable, but we always have that the trivial solution $r = 0$, corresponding to $I = 0$, is unstable for all $\lambda > \lambda_c$. This is illustrated in Fig. 5(a) and 5(b) for regular hypergraph structures with fixed membership and group size. The invasion threshold depends on both the structure (g_m, p_n) and the nonlinear exponent ν. The resulting phase diagram in the (λ, ν) plane is shown in Fig. 5(c). The invasion threshold spans the critical line (solid line) $\lambda = \lambda_c$ in the phase diagram.

The dashed critical line $\lambda = \lambda_p$ is associated with the limit of validity of a solution where $\mathcal{M}[\rho(r)]$ is tangent to r for some $r > 0$—it is thus solved numerically. We call λ_p the *persistence threshold* as it is the smallest value of λ such that a nontrivial solution is locally stable. For continuous phase transitions, $\lambda_c = \lambda_p$, but for discontinuous phase transitions, $\lambda_p < \lambda_c$.

2. *Tricritical points and the bistability threshold*

Depending on the structure and the form of $\beta(n,i)$, we can have a continuous or a discontinuous phase transition, as can be seen in Fig. 5. When the phase transition is continuous, we have possibly two solutions for the stationary fraction of infected nodes, $I_1 = 0$ and $I_2 > 0$. When I_2 exists (for instance when $\lambda > \lambda_c$), I_1 is unstable.

When the phase transition is discontinuous, we have typically three solutions, $I_1 = 0$ and $0 < I_2 < I_3$. In the bistable regime [for instance when $\lambda \in (\lambda_p, \lambda_c)$], all three solutions coexist, I_2 is unstable, and I_1 and I_3 are locally stable. In the endemic regime [for instance when $\lambda \geq \lambda_c$], only I_1 and I_3 exist, and only I_3 is locally stable.

We are interested in knowing when the phase transition changes from continuous to discontinuous. In Fig. 5(c), the bistable regime starts at a tricritical point (star marker), where two critical lines meet. For the infection rate function $\beta(n,i) = \lambda i^\nu$ and a fixed hypergraph structure, the tricritical point happens at $(\lambda, \nu) = (\lambda_c, \nu_c)$, where ν_c is what we call the *bistability threshold*, since a bistable regime only exists for $\nu > \nu_c$.

To get some insights on the properties of tricritical points, we show in Fig. 4(b) the function $\mathcal{M}[\rho(r)]$ at $\lambda = \lambda_c$ and for values of ν below, at and above the bistability threshold. For $\nu < \nu_c$, we have $\partial_r^2 \mathcal{M}|_{r\to 0} < 0$ and I_3 does not exist. For $\nu > \nu_c$, we have $\partial_r^2 \mathcal{M}|_{r\to 0} > 0$ and there exists a solution $I_3 > 0$. At the tricritical point, $I_3 = I_2 \to 0$, hence the nontrivial solution is degenerate, which is possible only if

$$\left.\frac{\mathrm{d}^2\mathcal{M}}{\mathrm{d}r^2}\right|_{r\to 0} = \left.\left[\frac{\mathrm{d}\mathcal{M}}{\mathrm{d}\rho}\frac{\mathrm{d}^2\rho}{\mathrm{d}r^2} + \left(\frac{\mathrm{d}\rho}{\mathrm{d}r}\right)^2 \frac{\mathrm{d}^2\mathcal{M}}{\mathrm{d}\rho^2}\right]\right|_{r\to 0} = 0\,.$$

Since a tricritical point is also a critical point, from Eq. (11) $\mathrm{d}\mathcal{M}/\mathrm{d}\rho = (\mathrm{d}\rho/\mathrm{d}r)^{-1}$, so the condition can be rewritten as

$$\left.\left[\frac{\mathrm{d}^2\rho}{\mathrm{d}r^2} + \left(\frac{\mathrm{d}\rho}{\mathrm{d}r}\right)^3 \frac{\mathrm{d}^2\mathcal{M}}{\mathrm{d}\rho^2}\right]\right|_{r\to 0} = 0\,.$$

The derivatives on ρ with respect to r at a critical point where $r \to 0$ can be easily evaluated, and the condition now becomes

$$2\left(\frac{\langle m^2\rangle^2}{\langle m\rangle^2} - \frac{\langle m^3\rangle}{\langle m\rangle}\right) + \left(\frac{\langle m(m-1)\rangle}{\langle m\rangle}\right)^3 \left.\frac{\mathrm{d}^2\mathcal{M}}{\mathrm{d}\rho^2}\right|_{\rho\to 0} = 0\,. \tag{16}$$

To evaluate the last term of Eq. (16), let us rewrite

$$\mathcal{M}(\rho) = \frac{\sum_{n,i} \beta(n,i)(n-i) f_{n,i}(\rho) p_n}{\sum_{n,i} (n-i) f_{n,i}(\rho) p_n} \equiv \frac{u(\rho)}{v(\rho)} \, .$$

In the limit $\rho \to 0$, $f_{n,i} \to \delta_{i,0}$, which implies $u(\rho) \to 0$ and $v(\rho) \to \langle n \rangle$, therefore

$$\left.\frac{\mathrm{d}^2\mathcal{M}}{\mathrm{d}\rho^2}\right|_{\rho\to 0} = \frac{1}{\langle n \rangle} \left.\frac{\mathrm{d}^2 u}{\mathrm{d}\rho^2}\right|_{\rho\to 0} - \frac{2}{\langle n \rangle^2} \left.\frac{\mathrm{d}u}{\mathrm{d}\rho}\frac{\mathrm{d}v}{\mathrm{d}\rho}\right|_{\rho\to 0} \, . \tag{17}$$

First order derivatives can be evaluated using

$$\left.\frac{\mathrm{d}u}{\mathrm{d}\rho}\right|_{\rho\to 0} = \sum_{n,i} \beta(n,i)(n-i) p_n h_{n,i} \, , \tag{18a}$$

$$\left.\frac{\mathrm{d}v}{\mathrm{d}\rho}\right|_{\rho\to 0} = \sum_{n,i} (n-i) p_n h_{n,i} \, . \tag{18b}$$

For the second-order derivative, let us define $l_{n,i} \equiv \mathrm{d}^2 f_{n,i}/\mathrm{d}\rho^2|_{\rho\to 0}$, so that we can write

$$\left.\frac{\mathrm{d}^2 u}{\mathrm{d}\rho^2}\right|_{r\to 0} = \sum_{n,i} \beta(n,i)(n-i) p_n l_{n,i} \, . \tag{19}$$

Finally, we can apply the second order derivative to Eq. (7b) to obtain the recurrence relation

$$\begin{aligned}(i+1) l_{n,i+1} =& 2\left[(n-i)h_{n,i} - (n-i+1)h_{n,i-1}\right] \, , \\ &+ \left[i + (n-i)\beta(n,i)\right] l_{n,i} \, , \\ &- (n-i+1)\beta(n,i-1) l_{n,i-1} \, .\end{aligned} \tag{20}$$

Again, $l_{n,0} = -\sum_{i>0} l_{n,i}$ by definition.

Even though it is possible to express the $l_{n,i}$ in an explicit form, the expression does not give us more intuition, and it is simpler to calculate the $l_{n,i}$ using the recurrence equation just given. After rewriting $\mathrm{d}^2\mathcal{M}/\mathrm{d}\rho^2|_{\rho\to 0} \equiv F[p_n, \beta(n,i)]$, tricritical points are obtained by solving the equation

$$\frac{\langle m^2 \rangle^2}{\langle m \rangle^2} - \frac{\langle m^3 \rangle}{\langle m \rangle} + \frac{\langle m(m-1) \rangle^3}{2\langle m \rangle^3} F[p_n, \beta(n,i)] = 0 \, . \tag{21}$$

Tricritical points result from an intricate relation between the structure (g_m, p_n) and the infection rate $\beta(n,i)$. Fig. 5 shows that either changing the structure [Fig. 5(a)] or the shape of the infection rate function [Fig. 5(b)-(c)] can lead to a change of behavior, from a continuous phase transition to a discontinuous one with a bistable regime.

A first hypothesis we can make from these simple examples is that more nonlinear infection rates (larger ν) and larger groups promote bistability. However, we will see that this intuition does not hold in general for heterogeneous structures due to the onset of mesoscopic localization.

3. *Heterogeneous memberships*

In this section we investigate the effects of a heterogeneous membership distribution g_m while keeping $p_n = \delta_{n,n_0}$ homogeneous to disentangle the impact of the different types of heterogeneity. A first remark we can make about the invasion threshold [Eq. (15)] is that it is coherent with heterogeneous pair-approximation frameworks [55] on random networks when only dyadic interactions are considered, i.e., when $p_n = \delta_{n,2}$. In this case, we can set $\nu = 1$ without loss of generality, thus recovering the standard SIS model. The associated threshold is

$$\lambda_{\mathrm{c}}^{\mathrm{SIS}} = \frac{\langle m \rangle}{\langle m(m-1) \rangle} \, ,$$

where g_m can now be interpreted as the standard degree distribution of graphs. This threshold, although quite accurate for most structures, does not capture the hub reinfection mechanism [56], and thus could be inaccurate for graphs with hubs of a very large degree.

More generally, for group interactions ($p_n \neq \delta_{n,2}$) we can see that a larger average excess membership $\langle m(m-1) \rangle / \langle m \rangle$ always leads to a smaller invasion threshold λ_{c}, akin to the standard SIS model, but the relationship is now nonlinear. To see this, let us rewrite Eq. (15) as

$$\frac{1}{\langle n \rangle} \left\langle \sum_{i>0} \frac{n!}{(n-i-1)!\, i!} \prod_{j=1}^{i} \beta(n,j) \right\rangle = \frac{\langle m \rangle}{\langle m(m-1) \rangle} \, . \tag{22}$$

Since $\beta(n,i)$ is a monotonically increasing function of λ for all n, i, then the left-hand side of Eq. (22) is a monotonically increasing function (of λ) as well. Consequently, if the right-hand side decreases, λ_{c} must decrease as well.

Assessing the impact of membership heterogeneity on the bistability threshold ν_{c} is more complicated. In fact, Eq. (21) explicitly depends on the first three moments of g_m, but it also depends on the first two moments implicitly through λ_{c}, at which F must be evaluated.

In order to build our intuition, let us assume that we are able to keep fixed the first two moments $\langle m \rangle$ and $\langle m^2 \rangle$ while increasing $\langle m^3 \rangle$. This means that λ_{c} would not change, hence the only dependence on g_m would be explicit in Eq. (21). Since the term depending on $\langle m^3 \rangle$ is negative, increasing the third moment implies that F must increase if we want to balance Eq. (21). But since $\mathrm{d}^2\mathcal{M}/\mathrm{d}r^2|_{r\to 0}$ increases with ν [see Fig. 4], and thereby F as well, we can conclude that increasing $\langle m^3 \rangle$ while keeping the first two moments fixed leads to an increase of the bistability threshold ν_{c}. This is validated in the Supplementary Note 5, where we consider two distributions g_m sharing the same first two moments, but a different third moment. The larger third moment suppresses the emergence of a bistable regime.

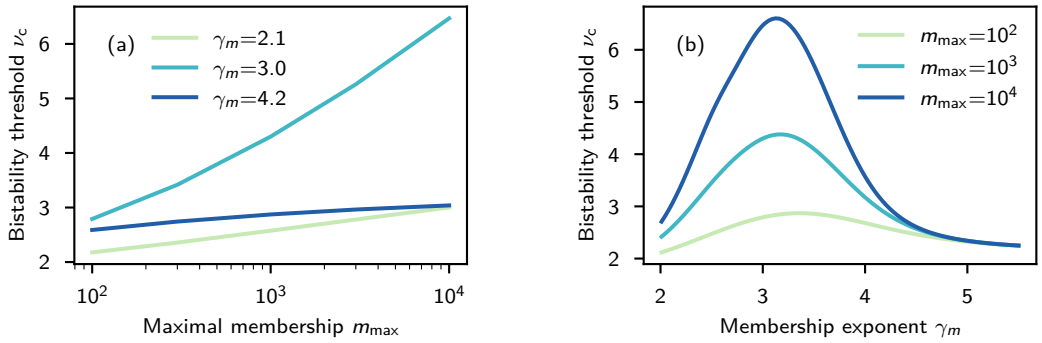

FIG. 6. **Impact of heterogeneous memberships on the bistability threshold.** We considered hypergraphs with power-law membership distributions $g_m \sim m^{-\gamma_m}$ with various exponents γ_m, different maximal values $m_{\max}$, a minimal value $m_{\min} = 2$, and a regular group size distribution $p_n = \delta_{n,4}$. We obtain the bistability threshold by solving Eq. (21). Lower values of γ_m imply a more heterogeneous membership distribution. (a) The bistability threshold grows logarithmically or faster for $\gamma_m \leq 4$, but converges for $\gamma_m > 4$. (b) The bistability threshold shows a maximum near $\gamma_m = 3$, suggesting that this is the maximum in the limit $m_{\max} \to \infty$.

A corollary of this argument is that for certain structures, it is impossible to have bistability. To see this, let us consider a power-law membership distribution of the form $g_m \sim m^{-\gamma_m}$. In this case, since the bistability threshold depends on the third moment of g_m, while the invasion threshold only depends on the first two, by setting the exponent $3 < \gamma_m < 4$, the invasion threshold converges to a value $\lambda_c > 0$, but ν_c does not exist. In other words, it is impossible to have a discontinuous phase transition.

This second statement is validated in 6(a), where we show that that the bistability threshold ν_c appears to grow without bound as $m_{\max} \to \infty$ for $\gamma_m \leq 4$. Instead, for $\gamma_m \geq 4$, the bistability threshold appears to converge, as expected, since the first three moments of g_m converge as well. What is more surprising is the nonmonotonic behavior of ν_c with respect to γ_m, which we present in Fig. 6(b). The bistability threshold has a well defined maximum at a value of γ_m that appears to converge to $\gamma_m = 3$ for $m_{\max} \to \infty$. In other words, $\gamma_m = 3$ is the optimal value of membership exponent in suppressing the emergence of a discontinuous phase transition and the related bistability.

This can be understood from Eq. (21): for $\gamma_m > 3$, the invasion threshold does not vary much since the first two moments of g_m are finite. Hence maximizing the third moment maximizes ν_c, which corresponds to $\gamma_m \to 3$. One could still be surprised that the bistability threshold grows more slowly with $m_{\max}$ in the range $2 < \gamma_m < 3$, since the invasion threshold λ_c tends toward zero. In this case, the bistable regime exists, but its width (λ_p, λ_c) simply vanishes as $\lambda_c \to 0$.

4. *Heterogeneous group sizes*

Let us now consider hypergraphs with heterogeneous group size distribution p_n, and homogeneous membership distribution, namely $g_m = \delta_{m,m_0}$. In this case, the invasion threshold, as defined by Eq. (15), depends on the whole distribution p_n, which makes drawing general conclusions on the impact of a heterogeneous distribution p_n more difficult.

To get some intuitions, let us consider the standard SIS model, i.e., the case $\nu = 1$ in Eq. (1). With our AMEs, it was shown that [50]

$$\lambda_c^{-1} \simeq \left(\frac{\langle m(m-1)\rangle}{\langle m\rangle}\right)\left(\frac{\langle n(n-1)\rangle}{\langle n\rangle}\right) + n_{\max} \ , \qquad (23)$$

for power-law distributions $p_n \sim n^{-\gamma_n}$ with large cut-offs $n_{\max}$. The first term on the right-hand side of Eq. (23) suggests that more heterogeneous groups-size distributions p_n (smaller values of γ_n) lead to smaller invasion thresholds. Intuition tells us that we should expect this behavior for $\nu \neq 1$ as well. We have therefore investigated numerically in Fig. 7(a) the invasion threshold as a function of the group size exponent for different values of ν, confirming that more heterogeneous group sizes (smaller γ_n) do lead to a smaller invasion threshold, even for nonlinear infection functions ($\nu \neq 1$). However, this effect is mitigated when larger values of ν are considered. For large ν and large $n_{\max}$, the value of the invasion threshold is dominated by the cut-off, and scales as $\lambda_c \sim n_{\max}^{-\nu}$, as illustrated in Fig. 7(b).

This behavior can be attributed to the onset of mesoscopic localization [49, 50]. It was shown analytically for $\nu = 1$ that, for certain combinations of (γ_m, γ_n), the epidemic near the invasion threshold [$\lambda = \lambda_c(1+\epsilon)$ with $\epsilon \ll 1$] is dominated by the largest most influential groups. In these scenarios, the second term on the right-hand side in Eq. (23) dominates the first one and, near

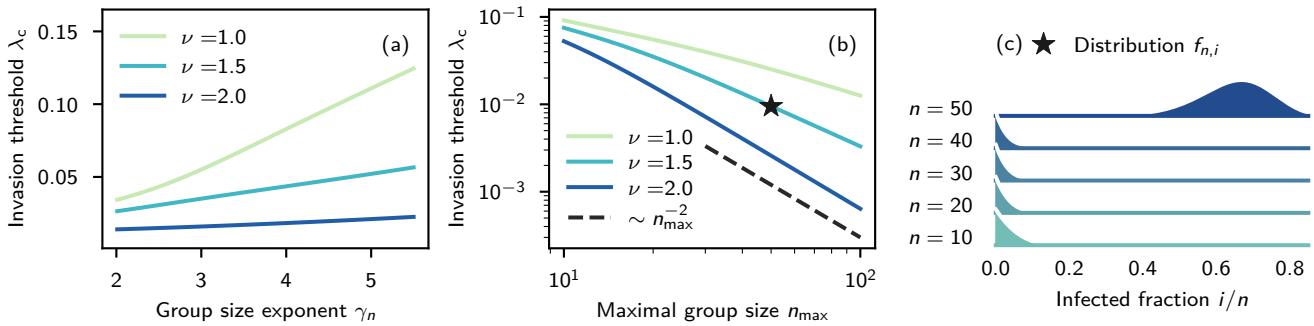

FIG. 7. **Impact of heterogeneous group sizes on the invasion threshold.** We considered hypergraphs with power-law group size distributions $p_n \sim n^{-\gamma_n}$ with various exponents γ_n, a regular membership distribution $g_m = \delta_{m,4}$, and various exponents ν for the infection function in Eq. (1). We obtain the invasion threshold using Eq. (15). Lower values of γ_n implies a more heterogeneous group size distribution. (a) For a fixed $n_{\max} = 20$, the invasion threshold increases with larger γ_n, but the effect is more limited for larger ν. (b) For a fixed $\gamma_n = 3$, the invasion threshold decreases like $n_{\max}^{-\nu}$ for large $n_{\max}$, indicating the onset of mesoscopic localization [49, 50]. (c) Stationary distribution $f_{n,i}$ for the starred case in (b), i.e., $\gamma_n = 3$ and $\nu = 1.5$, with $\lambda = 1.1\lambda_c$, illustrating localization in the largest groups.

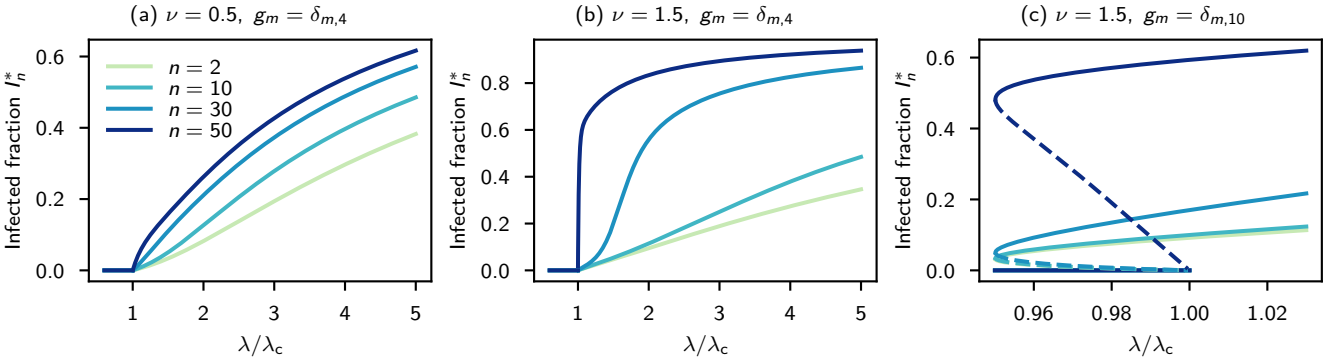

FIG. 8. **Mesoscopic localization in large groups.** We illustrate I_n^*, the average stationary fraction of infected nodes in groups of various size n. We considered hypergraphs with power-law group size distributions $p_n \sim n^{-\gamma_n}$ with $\gamma_n = 3$, regular membership distributions of the form $g_m = \delta_{m,m_0}$, and various exponents ν for the infection function in Eq. (1). Near the invasion threshold λ_c, I_n^* is larger for larger groups in (a) with $\nu = 0.5$, but localization is much more pronouced for (b) with $\nu = 1.5$. (c) For discontinuous phase transitions, mesoscopic localization is still possible, but near λ_c we must look at the unstable solution for I_n^* in the bistable regime.

λ_c, the group prevalence I_n grows exponentially with n, i.e., $I_{n_{\max}}/I_2 = \Omega\left(e^{an_{\max}}\right)$ for some constant a. While an analytical characterization of mesoscopic localization in the general case of $\nu \neq 1$ is out of the scope of this paper, we provide clear numerical evidence of localization phenomena in Fig. 7(c). The stationary distributions of the fraction of infected nodes in groups of increasing size n is concentrated in the largest group ($n = 50$) near the invasion threshold λ_c.

Since mesoscopic localization was characterized using a linear contagion ($\nu = 1$) and a continuous phase transition [49, 50], two natural questions arise: How does $\nu \neq 1$ affects localization? And what happens in the context of discontinuous phase transitions? In Fig. 8, we present the phase diagram of the group prevalence I_n for different scenarios. Comparing Fig. 8(a) and Fig. 8(b), we see that increasing ν from 0.5 to 1.5 (while keeping $g_m = \delta_{m,4}$) strengthens localization effects, which is expected since reinforcement effects are more important when the group prevalence is high. In Fig. 8(c), we show a similar diagram, but for a discontinuous phase transition. We see that the concentration of infected nodes in the largest groups is still possible, but the phenomenon is now associated with the unstable solution near the invasion threshold [$\lambda = \lambda_c(1 - \epsilon)$ with $\epsilon \ll 1$]. Therefore, mesoscopic localization affects both continuous and discontinuous phase transition with a bistable regime, but the exponential growth of I_n with n near λ_c concerns the stable solution in the former and the unstable solution in the latter.

If we now reinterpret the results of Fig. 7 in light of these considerations, larger values of ν facilitate the onset

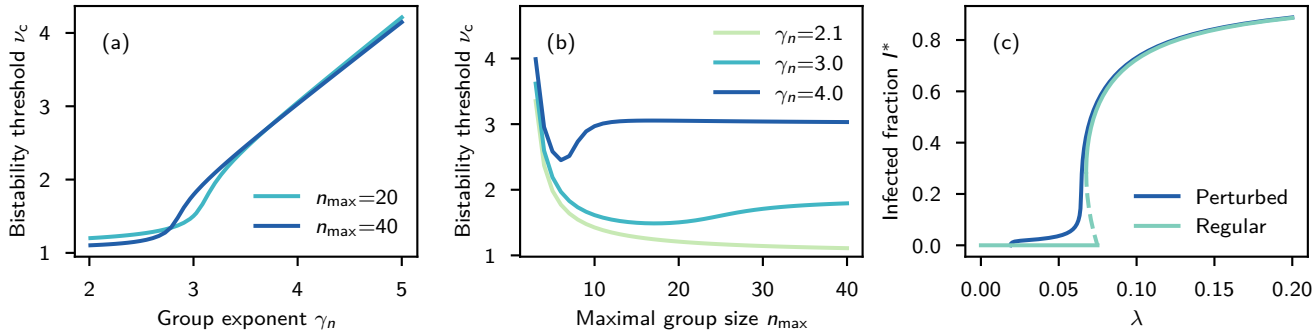

FIG. 9. **Impact of heterogeneous group sizes on the bistability threshold.** We considered hypergraphs with power-law group size distributions $p_n \sim n^{-\gamma_n}$ with various exponents γ_n and a regular membership distribution $g_m = \delta_{m,4}$. We solve the bistability threshold using Eq. (21). Lower values of γ_n implies a more heterogeneous group size distribution. (a) For a fixed $n_{\max}$, the bistability threshold increases with larger γ_n. (b) For a fixed γ_n, the bistability threshold has a nonmonotonic relationship with $n_{\max}$. (c) Phase transition using $g_m = \delta_{m,4}$ and $\nu = 2.3$. We use $p_n = \delta_{n,4}$ for the regular case, and $p_n = (1-\epsilon)\delta_{n,4} + \epsilon\delta_{n,15}$ with $\epsilon = 10^{-3}$ for the perturbed case. Mesoscopic localization [49, 50] inhibits bistability in the perturbed case.

of mesoscopic localization, where the largest groups drive the onset of the endemic phase, and make the invasion threshold scale as $\lambda_c \sim n_{\max}^{-\nu}$. This explains why λ_c varies only slightly with γ_n for $\nu = 2$ in Fig. 7(a).

In Fig. 9, we finally investigate the role of heterogeneous group sizes on the bistability threshold by varying the group exponent γ_n and the maximal group size $n_{\max}$. From Fig. 9(a), we see that a more heterogeneous group distribution, thereby increasing the fraction of larger groups, decreases the value of the bistability threshold ν_c. This is consistent with our observation on regular structures [Fig. 5], for which larger groups appear to promote bistability. However, Fig. 9(b) brings some nuance to this statement: for a fixed exponent γ_n, there is a nonmonotonic relationship between ν_c and the largest group $n_{\max}$. As such, the presence of larger groups does not always promote bistability.

We can again attribute this behavior to localization effects. In fact, we are able to illustrate this via a very simple example in Fig. 9(c). We look at the phase transition for a regular hypergraph with a fixed group size, $p_n = \delta_{n,4}$, and a perturbed version of it, where we introduce a small proportion of larger groups, $p_n = (1-\epsilon)\delta_{n,4} + \epsilon\delta_{n,15}$ with $\epsilon = 10^{-3}$. For the regular distribution, the phase transition is discontinuous, while for the perturbed distribution it is continuous, with the contagion localized in the largest groups near the invasion threshold. The bistability threshold ν_c is larger for the perturbed distribution since mesoscopic localization reduces considerably the invasion threshold λ_c. The largest most influential groups drive and self-sustain an endemic state for smaller values of λ, hence preventing a bistable regime.

C. Influence maximization

Influence maximization broadly refers to the problem of selecting a subset of nodes to initially spark a diffusion process in order to maximize the effect. The process could represent the spread of information, the diffusion of innovations or a viral marketing campaign [6, 57].

There is a large body of literature on influence maximization in complex networks, where various models have been used: threshold models [58–61], independent cascade [58], and simple contagion models (SI, SIS, SIR) [62–65], to name a few. Recently, these ideas have been also exported to higher-order networks [66, 67].

The effectiveness of an influence maximization procedure is often measured by the fraction of affected nodes (in the limit $t \to \infty$) for processes that terminate. However, because the final epidemic size in the SIS dynamics does not depend on the seeds (other than for stochastic extinction), we will consider the simpler task of maximizing $\dot{I}(0)$, the initial spreading speed. This is often a straightforward task to solve for graphs. Considering the SIR model for instance, one just needs to maximize the number of outgoing edges from infected to susceptible nodes, which implies that nodes of maximal degree would be optimal influencers. However, we will show that additional considerations need to be accounted for in higher-order networks. More specifically, our goal is to use our formalism to answer the following question: Should we focus on finding influential nodes, or seed the spread from influential groups?

In this section, to simplify the notation, all dynamic quantities are evaluated at $t = 0$, e.g. $I(0) \equiv I$.

Let us assume that we are given a fixed hypergraph and an initial fraction of nodes that can be infected at the initial time $I = \epsilon \ll 1$ (the seeds of the contagion).

Our task is to invade the system as fast as possible by maximizing $\dot{I}$ for a hypergraph contagion, which is equivalent to maximizing the objective function

$$\Phi[\mathcal{S},\mathcal{F}] = r\langle m s_m\rangle \ , \tag{24}$$

where we define the initial node states $\mathcal{S} \equiv \{s_m\}_{m=1}^{m_{\max}}$ and the initial group states $\mathcal{F} \equiv \{f_{n,i}|0 \le i \le n\}_{n=2}^{n_{\max}}$. The optimization problem is also constrained by

$$0 \le s_m \le 1 \ \forall m \ , \tag{25a}$$

$$0 \le f_{n,i} \le 1 \ \forall n, i \ , \tag{25b}$$

$$\sum_m s_m g_m = 1 - \epsilon \ , \tag{25c}$$

$$\sum_i f_{n,i} = 1 \ \forall n \ , \tag{25d}$$

$$\frac{\langle m s_m\rangle}{\langle m\rangle} = \frac{1}{\langle n\rangle}\sum_{n,i}(n-i) f_{n,i} p_n \ . \tag{25e}$$

While the first four constraints come from the definitions of the variables, the last one is less straightforward. Equation (25e) ensures the consistency between $\mathcal{S}$ and $\mathcal{F}$, more specifically that the fraction of all memberships stubs belonging to susceptible nodes [left-hand side of Eq. (25e)] matches the fraction of susceptible nodes in groups [right-hand side of Eq. (25e)].

By combining the constraint of Eq. (25e) with the definition of r as given by Eq. (2), the objective function can be simplified as

$$\Phi[\mathcal{S},\mathcal{F}] \propto \sum_{n,i} \beta(n,i)(n-i) f_{n,i} p_n \ . \tag{26}$$

Although it appears to be independent of $\mathcal{S}$, it depends on it implicitly through Eq. (25e).

It is worth stressing that our formalism assumes that the membership stubs of nodes are assigned to groups uniformly at random, and thus we cannot engineer both $\mathcal{S}$ and $\mathcal{F}$, i.e., choose at the same time the seeds according to their membership and the repartition of the seeds among the various group sizes. Indeed, if we decide for instance to infect only nodes of a certain membership m' and we try to engineer $\mathcal{F}$, there are no guarantees we can achieve such configuration in practice—e.g. we cannot infect a node in a group if none of its nodes have membership m'.

We therefore compare two strategies to optimize the early spread:

A) The influential spreaders strategy: we engineer $\mathcal{S}$, i.e., we choose the fraction of seeds to assign to each membership class, and we assume a random configuration for the groups, i.e., all $\{f_{n,i}\}_{i=0}^{n}$ are binomial distributions with probability q (to be determined).

B) The influential groups strategy: we engineer $\mathcal{F}$, i.e., we assign a certain number of seeds in the groups depending on their sizes, and assume that nodes are infected at random through the group to which they belong.

1. *Influential spreaders*

In this strategy, we are free to engineer $\mathcal{S}$ in order to maximize Φ, with respect to the constraints of Eq. (25). Let us assume that $f_{n,i}$ is a binomial distribution,

$$f_{n,i} = \binom{n}{i} q^i (1-q)^{n-i} \ .$$

Using Eq. (25e), we can identify

$$q = 1 - \frac{\langle m s_m\rangle}{\langle m\rangle} \ .$$

An optimal solution $\mathcal{S}^\star$ can be found by first finding the value $q^\star$ that maximizes the objective function Eq. (26), and then identifying any set $\mathcal{S}$ that satisfies the relation for $q = q^\star$ above.

There are in general many optimal solutions possible, but they collapse into a single one when q is sufficiently small, which is reasonable for $\epsilon \ll 1$. In this case, we simply have that $\Phi \approx q$, and the optimal solution is intuitive: one needs to infect nodes of maximal membership first in order to maximize q. This is true irrespective of $\beta(n,i)$, p_n and g_m.

The infection function and the structure affect the maximal value of ϵ such that this solution is unique and optimal. For example, in the simplest case of linear contagion, where $\beta(n,i) \propto i$, it is possible to show that this strategy is optimal up to $q = 1/2$ for all g_m and p_n, and we expect even higher values for $\nu > 1$. For all practical purposes, targeting nodes of highest membership is optimal, and this is the case in all experiments we considered.

2. *Influential groups*

In this second strategy, we want to engineer $\mathcal{F}$ in order to maximize Φ with respect to the constraints Eq. (25). Let us assume that we can do so by choosing a certain number of groups and infecting a certain portion of their nodes. Following this procedure, one can realize that not all sets $\mathcal{F}$ satisfying Eq. (25) are allowed. For instance, if we decide to infect i nodes in all groups of size n, the outcome is different from just having $f_{n,i} = 1$. Indeed, nodes belong to more than one group, hence we need to account for spillover effects—groups of size $n' \neq n$ would have some infected nodes as well, and more than i nodes could be infected in some groups of size n.

To do so, let us first define $\tilde{f}_{n,i}$ as the fraction of all the groups of size n for which we infect i nodes at random. Note that if a node belongs to multiple groups, it

can be chosen more than once for infection, but the duplicates have no effect. Spillovers are taken into account by considering that each of the $n-i$ nodes that have not been chosen for infection in a group of size n could have been infected in another group, with probability u (to be determined). Therefore, we can write

$$f_{n,i} = \sum_{j=0}^{i} \tilde{f}_{n,i-j} B_{n-i+j,j} \ , \tag{27}$$

where

$$B_{k,j} = \binom{k}{j} u^j (1-u)^{k-j} \ .$$

Second, let us define η as the fraction of all spots in groups that have been chosen for infection,

$$\eta \equiv \frac{1}{\langle n \rangle} \sum_{n,i} i \tilde{f}_{n,i} p_n \ . \tag{28}$$

Since nodes within groups are chosen at random, a node of membership m is susceptible if it has not been chosen for infection in any of the groups to which it belongs, i.e.,

$$s_m = (1-\eta)^m \ .$$

As a consequence, η is constrained by Eq. (25c),

$$\langle (1-\eta)^m \rangle = 1 - \epsilon \ .$$

The probability u still needs to be obtained. It corresponds to the fraction of all memberships that are not matched with a spot chosen for infection in a group, but that are still associated with an infected node:

$$u = \frac{\langle m(1-s_m) \rangle - \eta \langle m \rangle}{(1-\eta)\langle m \rangle} \ .$$

With this formulation, we engineer $\mathcal{F}$ indirectly through $\tilde{\mathcal{F}} = \{\tilde{f}_{n,i} | 0 \leq i \leq n\}_{n=2}^{n_{\max}}$. The objective function can be rewritten

$$\begin{aligned} \Phi &\propto \sum_{n,i} \sum_{j=0}^{i} \beta(n,i)(n-i) \tilde{f}_{n,i-j} B_{n-i+j,j} p_n \ , \\ &= \sum_{n,i} \sum_{j=0}^{n-i} \beta(n,i+j)(n-i-j) \tilde{f}_{n,i} B_{n-i,j} p_n \ . \end{aligned}$$

Since the objective function is a linear function of each $\tilde{f}_{n,i}$, the optimization problem can be solved using linear programming.

However, there is an intuitive and more efficient way to solve this problem exactly. We just need to identify the most cost-effective $\tilde{f}_{n,i}$ by looking at the effect on Φ of increasing $\tilde{f}_{n,i}$,

$$E \equiv \frac{\partial \Phi}{\partial \tilde{f}_{n,i}} \propto \sum_{j=0}^{n-i} \beta(n,i+j)(n-i-j) B_{n-i,j} p_n \ ,$$

versus the *cost* of increasing $\tilde{f}_{n,i}$, i.e., the variation of η

$$C \equiv \frac{\partial \eta}{\partial \tilde{f}_{n,i}} \propto i p_n \ .$$

The most cost-effective $\tilde{f}_{n,i}$ maximizes the ratio

$$R(n,i) = \frac{E}{C} = \frac{1}{i} \sum_{j=0}^{n-i} \beta(n,i+j)(n-i-j) B_{n-i,j} \ . \tag{29}$$

Obviously, $i=0$ is always the most cost-effective for all n (since it has zero cost), but to satisfy Eq. (28), we must also fill some $\tilde{f}_{n,i}$ with $i>0$.

Optimal solutions tend to fill the $\tilde{f}_{n,i}$ with $i>0$ that maximizes $R(n,i)$, especially for sufficiently small ϵ. A general solution can be obtained using algorithm 1 presented in the "Influential groups solutions" subsection of the Methods, building on this idea of cost effectiveness. In the worst case, the computational complexity to obtain an optimal solution $\mathcal{F}^\star$ under the influential groups strategy is $\mathcal{O}(m_{\max} + n_{\max}^3)$, which is much more efficient than using a general-purpose linear-programming method.

Equation (29) also gives us an intuition of what defines influential groups when trying to maximize the early spread. If $\epsilon \ll 1$, then $u \ll 1$, hence we have

$$\begin{aligned} R(n,i) &\approx \frac{\beta(n,i)(n-i)}{i} \ , \\ &\propto i^{\nu-1}(n-i) \ , \end{aligned}$$

when considering $\beta(n,i) = \lambda i^\nu$. For simple contagions ($\nu = 1$), picking the largest group with a single seed ($i=1$) is always optimal. For hypergraph contagions with $\nu > 1$, the largest groups are the most influential as well, but the optimal number of seeds is generally $i > 1$. Hence, beyond its size, the initial configuration of a group determines whether or not it is influential.

3. *Experiments*

To compare the influential spreaders and the influential groups strategies, we measure the ratio

$$\zeta \equiv \frac{\Phi_{\mathcal{F}}^\star}{\Phi_{\mathcal{S}}^\star} \ , \tag{30}$$

where $\Phi_{\mathcal{F}}^\star$ and $\Phi_{\mathcal{S}}^\star$ are the values of the objective function for the optimal solution of the influential groups and influential spreader strategies, respectively. Therefore, $\zeta > 1$ indicates that the influential groups strategy is better to maximize $\dot{I}$, and vice versa if $\zeta < 1$.

In the Supplementary Note 6, we show that

$$\lim_{\epsilon \to 0} \zeta = \frac{\beta(n',i')(n'-i')\langle n \rangle}{i' \langle \beta(n,1) n(n-1) \rangle m_{\max}} + \frac{\langle m(m-1) \rangle}{\langle m \rangle m_{\max}} \ , \tag{31}$$

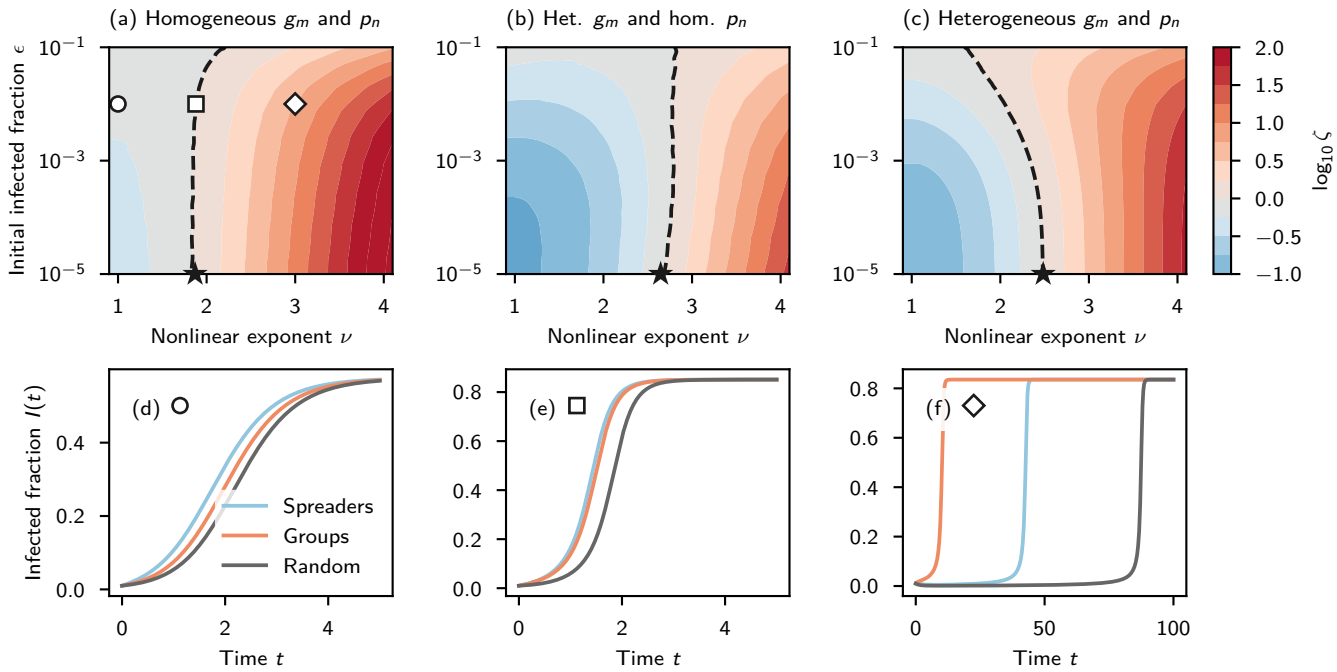

FIG. 10. **Comparison of the influential spreaders and influential groups strategies.** (a)-(c) If $\log_{10}\zeta > 0$, this indicates that the influential groups strategy is better to maximize $\dot{I}$, and vice versa if $\log_{10}\zeta < 0$. We use different combinations of homogeneous and heterogeneous distributions. More specifically, for homogeneous distributions we use $g_m \propto a^m \mathrm{e}^{-a}/m!$ and $p_n \propto a^n \mathrm{e}^{-a}/n!$, with $a = 5$, $m_{\min} = 1$, $n_{\min} = 2$, and $m_{\max} = n_{\max} = 20$. For heterogeneous distributions we use $g_m \propto m^{-\gamma_m}$ and $p_n \propto n^{\gamma_n}$, with $\gamma_m = \gamma_n = 3$, $m_{\min} = n_{\min} = 2$, and $m_{\max} = n_{\max} = 100$. The star markers indicate when $\zeta = 1$ in the limit $\epsilon \to 0$, using Eq. (32). Irregularities of the level curves are due to the discrete nature of g_m and p_n, not to numerical errors. (d)-(f) Time evolution of the fraction of infected nodes for different strategies, with $\epsilon = 10^{-2}$ and $\nu \in \{1, 1.88, 3\}$, corresponding to the three empty markers in (a). We use $\lambda = 3\lambda_c$ in each case. The random strategy corresponds to $s_m = 1 - \epsilon$ for all m, and $f_{n,i}$ is a binomial distribution with probability ϵ.

where (n', i') is the pair that maximizes the ratio $R(n, i)$, restricted to $i > 0$, in the limit $\epsilon \to 0$. For general ϵ, we need to solve numerically the optimization problem as discussed in the previous sections.

With $\beta(n, i) = \lambda i^\nu$, ζ is independent of λ, since $\Phi \propto \lambda$. As a consequence, ζ is agnostic to the underlying phase of the system (healthy, bistable, or endemic). Equation (31) simplifies to

$$\lim_{\epsilon \to 0} \zeta = \frac{i'^{\nu-1}(n' - i')\langle n \rangle}{\langle n(n-1) \rangle m_{\max}} + \frac{\langle m(m-1) \rangle}{\langle m \rangle m_{\max}} \,. \tag{32}$$

In Fig. 10, we illustrate how ζ varies as we change ν, ϵ, and the underlying structure. For homogeneous memberships and group sizes [Fig. 10(a)], we see that the influential groups strategy performs better as soon as the contagion process is sufficiently nonlinear ($\nu \approx 2$); for highly nonlinear contagions ($\nu \approx 4$), the influential group strategy is much more effective, with ζ up to 100. When considering heterogeneous memberships, but still homogeneous group sizes [Fig. 10(b)], the influential spreaders strategy performs better for moderately nonlinear contagions ($\nu \lesssim 2.8$), otherwise the influential groups strategy is still a better choice. Finally, considering a heterogeneous p_n as well [Fig. 10(c)] helps the performance of the influential groups strategy, especially for larger ϵ.

When picking a pair (ϵ, ν) far from $\zeta = 1$ [dashed lines in Figs. 10(a), 10(b) and 10(c)], the strategy that maximizes $\dot{I}$ invades the system faster, as can be seen in Figs 10(d) and 10(f). However, sufficiently close to $\zeta = 1$, maximizing $\dot{I}$ does not necessarily imply that $I(t)$ will be larger for all $t > 0$. For instance, in Fig. 10(e), $\zeta \approx 1$, but the influential spreader strategy is slightly better. Therefore, one must be careful when interpreting the results of Fig. 10. One way to improve on our approach would be to consider higher-order temporal derivatives of I to assess which strategy performs best, or refine the optimization procedure by trying to maximize these higher-order derivatives as well.

Figures 10(d)–(f) suggest that the initial speed, $\dot{I}(0)$, roughly correlates with the time taken by the disease to infect a given fraction of the population, a metric that has been used to measure influence for SI and SIR dynamics [68–70].

Figure 10(f) also illustrates a particular feature of highly nonlinear contagions: the time to reach the stationary state can be excessively long for suboptimal

strategies, despite $\lambda = 3\lambda_c$ (see the Supplementary Note 7). In this regime, the initial conditions have a much more important impact on the capacity of the contagion to invade the system, especially considering the possibility of stochastic extinction in real systems due to finite size.

These results again highlight the importance of considering an accurate description of the inner dynamics of groups when studying hypergraph contagions. In the context of influence maximization, optimizing group configurations is a crucial component; one should not focus exclusively on identifying the most central nodes. Ultimately, an optimal strategy would capitalize on the synergy of these two important aspects.

III. DISCUSSION

We have introduced group-based AMEs to describe hypergraph contagions. Our framework is analytically tractable, allowing us to obtain closed form implicit expressions for the critical and tricritical points. In addition, we have shown that it describes the dynamical process with remarkable accuracy when compared with Monte Carlo simulations. Our formulation in terms of an infection rate function $\beta(n, i)$ makes it extremely flexible, allowing us to consider arbitrary group distribution with large group interactions, contrarily to existing HMF theories [31, 39, 40] which instead require specifying the rule for each different type of interaction separately.

Motivated by simplicity and recent results [37], we analyzed in depth the consequences of a nonlinear infection rate function $\beta(n, i) = \lambda i^{\nu}$, highlighting the important role of influential groups in hypergraph contagions.

With our analytical results about the invasion and bistability thresholds, we were able to perform an exhaustive analysis of the phase transition and better understand the influence of a heterogeneous structure, both in terms of membership m and group size n. We found that the third moment of the membership distribution g_m plays a crucial role, with large $\langle m^3 \rangle$ suppressing the onset of a discontinuous phase transition with a bistable regime, in line with other approaches [39, 40]. This is best exemplified for power-law membership distributions $g_m \sim m^{-\gamma_m}$, where $\gamma_m = 3$ most suppresses bistability, and in the limit $m_{\max} \to \infty$, a discontinuous phase transition is only possible for $\gamma_m > 4$.

The phenomenon of mesoscopic localization [49, 50], driven by the most influential groups, also has important consequences on the phase diagram, with the effects being enhanced by superlinear infection ($\nu > 1$). In this case, the invasion threshold scales as $\lambda_c \sim n_{\max}^{-\nu}$, and for λ close to λ_c, infected nodes are found almost exclusively in the largest groups. This localization of the contagion thereby inhibits bistability by enforcing an endemic state with a very small global fraction of infected nodes.

Our approach, furthermore, provided insights concerning the problem of influence maximization for hypergraph contagions. We focused on the problem of maximizing the early spread, and proposed two strategies: allocating seeds to the influential spreaders (engineering s_m), or to the influential groups (engineering $f_{n,i}$). For various types of structures, the latter strategy performs better for contagions that are sufficiently nonlinear, highlighting the key role of influential groups on the transient state of the system.

For the process we considered, the notion of influential groups to seed and sustain hypergraph contagions are mostly aligned—in both cases, the largest groups typically have a dominant role. In the case of influence maximization, however, we showed that a careful seed allocation is also essential to determine whether or not a group is influential. Moreover, a more realistic infection function $\beta(n, i)$ that actually depends on n could affect which groups are most influential in both scenarios.

Our work constitutes a first step towards a better understanding of the role of higher-order interactions on the outreach of information spreading [6], and resonates with other recent theoretical findings on higher-order naming games, where big groups facilitate takeover of committed minorities in social convention [30]. AMEs thus provide an analytical avenue to study recent empirical results showing how social contagions and movements defy classic influence maximization. As one example, networked counterpublics [71] are public spaces used by underrepresented groups to gather legitimacy and form tight-knit communities. Therein, nondominant forms of knowledge can still spread and reach widespread attention through dense communities (influential groups) despite the limited connectivity of their members (noninfluential spreaders). These results provide one more addition to the mounting evidence that groups of elementary elements are the foundational unit of many complex systems.

Many avenues are now left open to explore and broaden the applicability of our group-based AMEs. While we restrained ourselves to a particular nonlinear infection rate function $\beta(n, i)$ and a constant recovery rate, other dynamical processes could be considered, each having their own phenomenology and a rich dynamical behavior. In the Supplementary Note 3 for instance, we briefly discuss how our framework can be applied to threshold models of the form $\beta(n, i) = \delta_{n-1,i}$, but one could consider other traditional dynamical processes, such as voter models [72].

In the Supplementary Note 1, we provide a roadmap to include structural two-point correlations, but a thorough characterization of the impact of correlation patterns on bistability, mesoscopic localization and influence maximization is still lacking. The inclusion of dynamical correlation around nodes is a more tedious task that would require a fusion between degree-based [47, 48, 73] and group-based [20, 21, 49, 50] AMEs. This would allow to

describe almost exactly *short-range* dynamical correlations, namely correlations between the states of nodes and their direct neighbors. Incorporating *long-range* correlations—beyond first neighbors—in AME frameworks, without a prohibitive computational time due to combinatorial explosion, is still an open problem.

Finally, many directions could be taken with regards to the influence maximization problem on hypergraphs. One avenue would be to analyze the notion of influential groups and influential spreaders from the perspective of centrality measures for hypergraphs [74]. Another would be to investigate the closely related problem of targeted immunization [1, 75].

IV. METHODS

A. Contagion on real-world hypergraphs

1. *Simulation of contagions*

We used a standard Gillespie algorithm for the simulation of contagions on hypergraphs. We decompose the whole process into events $j \in J$, that each happens at rate ω_j. The next event to happen is chosen with probability

$$P(j) = \frac{\omega_j}{\sum_{j\in J} \omega_j} \ ,$$

and the time step between two events Δt is distributed exponentially with mean $\langle \Delta t \rangle = 1/\sum_{j\in J} \omega_j$.

There are two types of events: infection and recovery. On the one hand, all susceptible nodes in a group can be considered equivalent with regards to infection. Consequently, each group is chosen for an infection event with rate

$$\omega_{\text{inf}}(n, i) = (n - i)\beta(n, i) \ .$$

Once a group is chosen for an infection event, one of the $(n - i)$ susceptible nodes is chosen uniformly at random to become infected. On the other hand, all infected nodes perform a recovery event with rate $\omega_{\text{rec}} = 1$.

We store all possible events in an efficient data structure called a `SamplableSet` [76], where insertion, deletion and sampling of elements (events) all have a computational complexity $\mathcal{O}\left[\log\log\left(\omega_{\max}/\omega_{\min}\right)\right]$ [77], where $\omega_{\max}$ and $\omega_{\min}$ are respectively the maximal and minimal rate among $\{w_j\}_{j\in J}$. This makes the sampling and the updating of the data structure extremely fast, which is especially useful when $\{w_j\}_{j\in J}$ spans multiple scales.

Once an event is performed—for instance a node recovers—we need to update the rate ω_{inf} of all groups to which this node belongs. This is the most costly part of the algorithm, which unfortunately cannot be overcome. This essentially means the simulation procedure is slower for hypergraphs with large average excess membership $\langle m(m-1)\rangle/\langle m\rangle$.

In Fig. 2 and 3, we compare the stationary state solutions from our formalism with estimates from Monte Carlo simulations. To compute estimates, we let the system relax during a burnin period $\tau_{\text{b}} \in [10^2, 10^4]$ then we sample $\mathcal{N} \in [10, 10^4]$ states, both depending on the size of the hypergraph and if multiple randomized hypergraphs are being used. Sampled states are separated by a decorrelation period $\tau_{\text{d}} = 1$.

To simulate contagions in the stationary state, we used two approaches, ordinary simulations and the quasistationary-state method.

a. Ordinary simulation method. With this approach, we simply let the simulation run and do not intervene. This is usually not the method of choice, especially for small hypergraphs near the invasion threshold, because finite size systems all eventually reach the absorbing state where all nodes are susceptible. This is, however, more practical to obtain the lower branch in Fig. 2(f), or faster for large hypergraphs, as in Fig. 3.

b. Quasistationary-state method. This approach aims at sampling the quasistationary distribution of the contagion process [78], which is defined as the probability distribution for all states in the limit $t \to \infty$, except for the absorbing state. We used a state-of-the-art method[78], where we keep a history of past states (in our case up to 50 states). We update the history by removing one uniformly at random and storing the current state after each decorrelation period $\tau_{\text{d}} \in [0.1, 1]$. Each time the absorbing state is reached during the simulation, we pick a state from the history uniformly at random to replace the current one. This method is well suited for finite-size analysis, and especially useful for simulations on small hypergraphs, such as in Fig. 2.

2. *Datasets*

The simulations shown in Fig. 2 and Fig. 3 run on two different empirical social structures that encode different types of social higher-order interactions. Here, we briefly describe the nature of these two datasets and the techniques used to construct the associated higher-order structures.

a. Face-to-face interactions in a French primary school. Originally collected as part of the SocioPatterns collaboration, this dataset [79] contains information of face-to-face interactions between children of a primary school in Lyon recorded over 2 days. Participants are given wearable sensors (placed on their chests), and a contact is detected whenever two sensors are in close proximity (1.5m). The initial temporal resolution of this dataset is 20s, but contacts have been further preprocessed in order to construct a static hypergraph from the temporal sequence of interactions [31]. In particular,

considering each child as a node, we aggregated different snapshots using a temporal window of 15 minutes and computed all the maximal cliques appearing in each window. Cliques were then aggregated across the entire time range, retaining only those that appeared at least 3 times, and finally "promoted" to groups. Some properties of the obtained structure are reported in the caption of Fig. 2.

b. Co-authorship in computer science. DBLP is an online bibliography containing information on major computer science journals and proceedings. This dataset, already preprocessed [80] (from the release 3, 2017), consists in a list of publications and respective authors that naturally calls for higher-order representations [81]. In particular, each author corresponds to a node and any collaboration of n (co-)authors in a single publication corresponds to a group of size n. We constructed a hypergraph by aggregating all the resulting groups together, but without considering single author publications (these have been removed in order to avoid disconnected nodes). In addition, the original dataset contained 1 831 127 nodes and 2 954 518 groups, which is too large to perform simulations on a personal computer in a reasonable time. Therefore, we obtained a subhypergraph by performing a breadth-first search, starting from a random group, then visiting all groups at a maximum distance of 2 when considering the one-mode projection of the original hypergraph on the groups. This ensures that the resulting subhypergraph is connected. Some properties of the obtained structure are reported in the caption of Fig. 3.

The authors state that in the country where the work was performed additional ethical approval or a license to reuse the datasets is not required given that the datasets are in the public domain.

3. Randomization and data augmentation

In Fig. 2 and 3, we make use of randomized versions of the original hypergraph. In Fig. 2, we also use *expanded* versions, where the size of the network is increased by a factor x. In all cases, we use the same procedure ($x = 1$ if the hypergraph is not expanded).

Let us first note $\mathbf{m} = [m_1, m_2, \dots]$ and $\mathbf{n} = [n_1, n_1, \dots]$ the membership sequence and the group size sequence of the original hypergraph, i.e., the list for the membership of each node and the list for the size of each group. From these sequences, we create two expanded sequences $\mathbf{m}'$ and $\mathbf{n}'$, which are formed of x copies of $\mathbf{m}$ and $\mathbf{n}$ respectively. This can be seen as the membership and group size sequences for a hypergraph that is x times larger.

For each expanded sequence, we create a stub list. For instance, for the node j of the expanded hypergraph, we include m'_j copies of the label j in the stub list for the nodes. Similarly, we include n'_ℓ copies of the label ℓ in the stub list for the groups. By definition, these two stub lists are of the same length, M', which corresponds to the number of edges in the bipartite representation of the hypergraph. We can thus shuffle them and match the entries of both lists, thereby assigning nodes to groups—or equivalently creating edges between nodes and groups in the bipartite representation of the hypergraph.

We then remove multi-edges (nodes assigned multiple times to a same group) by performing edge swaps [82]. We then perform M' additional edge-swap attempts at random—picking two random edges, swapping the groups, and accepting the swap if it does not create multi-edges. This ensures the uniformity of the generation process (see the Supplementary Note 8). The resulting hypergraph is a randomized version of the original hypergraph, expanded by a factor x.

B. Influential groups solutions

An intuitive approach to solve the problem would be to sort all pairs (n, i) in decreasing order of their $R(n, i)$ values (for $i > 0$), then fill $\tilde{f}_{n,i}$ up to 1 following this order, until $I = \epsilon$, or more directly until η reaches the value prescribed by ϵ. However, this approach does not account for the fact that one may encounter multiple times a same n value before the condition $I = \epsilon$ is reached. For instance, let us assume (n, i) is the next the pair with the highest value $R(n, i)$, but there exists a pair (n, i') with $R(n, i') \geq R(n, i)$ and we have already assigned $\tilde{f}_{n,i'} = 1$. What is the best option?

1. Discard the (n, i) pair.
2. Fill the associated $\tilde{f}_{n,i}$ up to 1 and decrease the value of $\tilde{f}_{n,i'}$ accordingly.

It turns out that an optimal solution is constructed by choosing one or the other depending on certain conditions. Option 1 is chosen whenever $i < i'$, because it can only reduce the total contribution to Φ. If $i > i'$, we assign a new cost-effective ratio to the pair (n, i), accounting for the fact that we need to decrease $\tilde{f}_{n,i'}$:

$$\hat{R}(n, i) = \frac{iR(n, i) - i'R(n, i')}{i - i'} \ .$$

This can be interpreted as the cost-effective ratio for the additional infected nodes $(i - i')$ that we add to the configuration. Note that $\hat{R}(n, i)$ can be negative, which is not a problem: this only means that infecting these nodes decreases the objective function Φ. If $\hat{R}(n, i)$ is still the highest ratio when compared with the ratios from other available pairs, option 2 is chosen.

The procedure to find an optimal set $\tilde{\mathcal{F}}^\star$ is presented in algorithm 1, where we use a priority queue Q to keep sorted the pairs (n, i) in decreasing order of their cost-effective ratio. Using a binary heap for Q, inserting

any pair (n, i) or removing the pair with maximum cost-effective ratio takes $\mathcal{O}(\log |Q|)$ elementary operations in the worst case, where $|Q|$ is the number of elements in the queue. Since $|Q| \leq n_{\max}^2$, insertions and removals have a complexity $\mathcal{O}(\log n_{\max})$. In the worst case, we will perform $\mathcal{O}(n_{\max}^2)$ insertions and removals before the condition $I = \epsilon$ is reached, hence the worst-case complexity for algorithm 1 is $\mathcal{O}(n_{\max}^2 \log n_{\max})$.

We also need to precompute $R(n, i)$ for all $n, i > 0$ and get $\mathcal{F}^\star$ from $\tilde{\mathcal{F}}^\star$, both $\mathcal{O}(n_{\max}^3)$. Moreover, solving for η is $\mathcal{O}(m_{\max})$, so the overall computational complexity of the influential groups optimization procedure is $\mathcal{O}(m_{\max} + n_{\max}^3)$.

Algorithm 1 Optimization for influential groups

Input: p_n, $R(n, i)$, η
Output: $\tilde{\mathcal{F}}^\star$

$\tilde{f}_{n,i} \leftarrow 1$ if $i = 0$, else 0, $\forall n, i$
$\psi \leftarrow \eta\langle n \rangle$ ▷ Initial resources
$i_{\text{opt}} \leftarrow \emptyset$ ▷ Associative array
$Q \leftarrow \emptyset$ ▷ Priority queue
for all (n, i) s.t $n \leq n_{\max}$, $p_n > 0$ and $1 \leq i \leq n$ **do**
  $R \leftarrow R(n, i)$
  $Q \leftarrow (n, i)$ with priority R
end for
while Q is not empty and $\psi > 0$ **do**
  $(n, i) \leftarrow Q$ ▷ Highest priority element
  if $n \notin i_{\text{opt}}$ **then**
    $\tilde{f}_{n,i} \leftarrow \min[1, \psi/(ip_n)]$
    $\tilde{f}_{n,0} \leftarrow \tilde{f}_{n,0} - \tilde{f}_{n,i}$
    $\psi \leftarrow \psi - \tilde{f}_{n,i} i p_n$
    $i_{\text{opt}}[n] \leftarrow i$
  else if $i > i_{\text{opt}}[n]$ **then**
    $i' = i_{\text{opt}}[n]$
    $\hat{R} \leftarrow [iR(n, i) - i'R(n, i')]/(i - i')$
    if $\hat{R} > R$ for all $R \in Q$ **then**
      $\tilde{f}_{n,i} \leftarrow \min\{1, \psi/[p_n(i - i')]\}$
      $\tilde{f}_{n,i'} \leftarrow \tilde{f}_{n,i'} - \tilde{f}_{n,i}$
      $\psi \leftarrow \psi - \tilde{f}_{n,i} p_n (i - i')$
      $i_{\text{opt}}[n] \leftarrow i$
    else
      $Q \leftarrow (n, i)$ with priority $\hat{R}$
    end if
  end if
end while

ACKNOWLEDGMENTS

The authors acknowledge Calcul Québec for computing facilities. This work was supported by the Fonds de recherche du Québec – Nature et technologies (G.S.), the Natural Sciences and Engineering Research Council of Canada (G.S., A.A.), the Sentinelle Nord program of Université Laval, funded by the Canada First Research Excellence Fund (G.S., A.A.), the United Kingdom Regions Digital Research Facility (RDRF)-Urban Dynamics Lab under EPSRC Grant No. EP/M023583/1 (I.I.), the James S. McDonnell Foundation 21st Century Science Initiative Understanding Dynamic and Multi-scale Systems (I.I.), the Agence Nationale de la Recherche (ANR) project DATAREDUX [ANR-19-CE46-0008] (A.B. and I.I.), the National Science Foundation under grant No. OIA-2019470 (L.H.-D.), Google Open Source under the Open-Source Complex Ecosystems and Networks (OCEAN) project (L.H.-D.), and the National Institutes of Health 1P20 GM125498-01 Centers of Biomedical Research Excellence Award (L.H.-D.).

DATA AVAILABILITY

The hypergraphs analysed during the current study are available in the "Influential groups data" repository on Zenodo [83]: https://doi.org/10.5281/zenodo.5662206

The hypergraphs can also be obtained from the original sources.

- Face-to-Face data [79]: http://www.sociopatterns.org/datasets/primary-school-temporal-network-data/
- Coauthorship data [80]: https://github.com/arbenson/ScHoLP-Data/tree/master/coauth-DBLP

See the subsection "Contagion on real-world hypergraphs" of the Methods for the preprocessing of the data.

CODE AVAILABILITY

The code used to produce all results is available on Zenodo [84]: https://doi.org/10.5281/zenodo.5662446

COMPETING INTERESTS

The authors declare that they have no competing interests. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

G.S., I.I., V.L., A.B., G.P., A.A., and L.H.-D. conceptualized the study. G.S. and L.H.-D. developed the model. G.S. derived the theoretical results, implemented the algorithms, and performed the numerical experiments. G.S., I.I., V.L., A.B., G.P., A.A., and L.H.-D. analyzed the results and contributed to writing the manuscript.

REFERENCES

[1] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," Rev. Mod. Phys. **87**, 925–979 (2015).

[2] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks* (Cambridge University Press, 2008).

[3] I. Z. Kiss, J. C. Miller, and P. L. Simon, *Mathematics of Epidemics on Networks: From Exact to Approximate Models* (Springer, 2017).

[4] D. J. Daley and D. G. Kendall, "Epidemics and Rumours," Nature **204**, 1118 (1964).

[5] Y. Moreno, M. Nekovee, and A. F. Pacheco, "Dynamics of rumor spreading in complex networks," Phys. Rev. E **69**, 066130 (2004).

[6] E. M. Rogers, *Diffusion of Innovations*, 4th ed. (Simon and Schuster, 2010).

[7] M. E. J. Newman, *Networks* (Oxford University Press, 2018).

[8] L. Torres, A. S. Blevins, D. S. Bassett, and T. Eliassi-Rad, "The why, how, and when of representations for complex systems," (2020), arXiv:2006.02870.

[9] V. Salnikov, D. Cassese, and R. Lambiotte, "Simplicial complexes and complex systems," Eur. J. Phys. **40**, 014001 (2018).

[10] R. Lambiotte, M. Rosvall, and I. Scholtes, "From networks to optimal higher-order models of complex systems," Nat. Phys. (2019), 10.1038/s41567-019-0459-y.

[11] 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**, 1–92 (2020).

[12] F. Battiston, E. Amico, A. Barrat, G. Bianconi, G. Ferraz de Arruda, B. Franceschiello, I. Iacopini, S. Kéfi, V. Latora, Y. Moreno, M. Murray, T. Peixoto, F. Vaccarino, and G. Petri, "The physics of higher-order interactions in complex systems," Nat. Phys. **17**, 1093–1098 (2021).

[13] G. Petri and A. Barrat, "Simplicial Activity Driven Model," Phys. Rev. Lett. **121**, 228301 (2018).

[14] G. Cencetti, F. Battiston, B. Lepri, and M. Karsai, "Temporal properties of higher-order interactions in social networks," Sci. Rep. **11**, 7028 (2021).

[15] L. Hébert-Dufresne, A. Allard, V. Marceau, P.-A. Noël, and L. J. Dubé, "Structural Preferential Attachment: Network Organization beyond the Link," Phys. Rev. Lett. **107**, 158702 (2011).

[16] J.-G. Young, L. Hébert-Dufresne, A. Allard, and L. J. Dubé, "Growing networks of overlapping communities with internal structure," Phys. Rev. E **94**, 022317 (2016).

[17] G. Bianconi and C. Rahmede, "Emergent Hyperbolic Network Geometry," Sci. Rep. **7**, 41974 (2017).

[18] O. T. Courtney and G. Bianconi, "Weighted growing simplicial complexes," Phys. Rev. E **95**, 062301 (2017).

[19] T. House and M. J. Keeling, "Deterministic epidemic models with explicit household structure," Math. Biosci. **213**, 29–39 (2008).

[20] L. Hébert-Dufresne, P.-A. Noël, V. Marceau, A. Allard, and L. J. Dubé, "Propagation dynamics on networks featuring complex topologies," Phys. Rev. E **82**, 036115 (2010).

[21] D. J. P. O'Sullivan, G. J. O'Keeffe, P. G. Fennell, and J. P. Gleeson, "Mathematical modeling of complex contagion on clustered networks," Front. Phys. **3**, 71 (2015).

[22] C. Bick, P. Ashwin, and A. Rodrigues, "Chaos in generically coupled phase oscillator networks with nonpairwise interactions," Chaos **26**, 094814 (2016).

[23] P. S. Skardal and A. Arenas, "Abrupt Desynchronization and Extensive Multistability in Globally Coupled Oscillator Simplexes," Phys. Rev. Lett. **122**, 248301 (2019).

[24] A. P. Millán, J. J. Torres, and G. Bianconi, "Explosive Higher-Order Kuramoto Dynamics on Simplicial Complexes," Phys. Rev. Lett. **124**, 218301 (2020).

[25] M. T. Schaub, A. R. Benson, P. Horn, G. Lippner, and A. Jadbabaie, "Random Walks on Simplicial Complexes and the Normalized Hodge 1-Laplacian," SIAM Rev. **62**, 353–391 (2020).

[26] T. Carletti, F. Battiston, G. Cencetti, and D. Fanelli, "Random walks on hypergraphs," Phys. Rev. E **101**, 022308 (2020).

[27] J. J. Torres and G. Bianconi, "Simplicial complexes: Higher-order spectral dimension and dynamics," J. Phys. Complex. **1**, 015002 (2020).

[28] L. Neuhäuser, A. Mellor, and R. Lambiotte, "Multibody interactions and nonlinear consensus dynamics on networked systems," Phys. Rev. E **101**, 032310 (2020).

[29] U. Alvarez-Rodriguez, F. Battiston, G. Ferraz de Arruda, Y. Moreno, M. Perc, and V. Latora, "Evolutionary dynamics of higher-order interactions in social networks," Nat. Hum. Behav. , 1–10 (2021).

[30] I. Iacopini, G. Petri, A. Baronchelli, and A. Barrat, "Group interactions modulate critical mass dynamics in social convention," (2021), arXiv:2103.10411.

[31] I. Iacopini, G. Petri, A. Barrat, and V. Latora, "Simplicial models of social contagion," Nat. Commun. **10**, 2485 (2019).

[32] S. Lehmann and Y.-Y. Ahn, eds., *Complex Spreading Phenomena in Social Systems*, Computational Social Sciences (Springer, 2018).

[33] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," Science **329**, 1194–1197 (2010).

[34] M. Karsai, G. Iñiguez, K. Kaski, and J. Kertész, "Complex contagion process in spreading of online innovation," J. R. Soc. Interface **11**, 20140694 (2014).

[35] N. O. Hodas and K. Lerman, "The Simple Rules of Social Contagion," Sci. Rep. **4**, 4343 (2014).

[36] 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).

[37] G. St-Onge, H. Sun, A. Allard, L. Hébert-Dufresne, and G. Bianconi, "Universal nonlinear infection kernel from heterogeneous exposure on higher-order networks," Phys. Rev. Lett. **127**, 158301 (2021).

[38] A. Barrat, G. Ferraz de Arruda, I. Iacopini, and Y. Moreno, "Social contagion on higher-order structures," (2021), arXiv:2103.03709.

[39] B. Jhun, M. Jo, and B. Kahng, "Simplicial SIS model in scale-free uniform hypergraph," J. Stat. Mech. **2019**, 123207 (2019).

[40] N. W. Landry and J. G. Restrepo, "The effect of heterogeneity on hypergraph contagion models," Chaos **30**, 103117 (2020).

[41] G. Ferraz de Arruda, M. Tizzani, and Y. Moreno, "Phase transitions and stability of dynamical processes on hyper-

graphs," Commun. Phys. **4**, 24 (2021).
[42] P. Cisneros-Velarde and F. Bullo, "Multi-group SIS Epidemics with Simplicial and Higher-Order Interactions," (2021), arXiv:2005.11404.
[43] G. Ferraz de Arruda, G. Petri, and Y. Moreno, "Social contagion models on hypergraphs," Phys. Rev. Research **2**, 023032 (2020).
[44] J. T. Matamalas, S. Gómez, and A. Arenas, "Abrupt phase transition of epidemic spreading in simplicial complexes," Phys. Rev. Research **2**, 012049 (2020).
[45] Giulio Burgio, Alex Arenas, Sergio Gómez, and Joan T Matamalas, "Network clique cover approximation to analyze complex contagions through group interactions," Commun. Phys. **4**, 1–10 (2021).
[46] R. Pastor-Satorras and C. Castellano, "Eigenvector Localization in Real Networks and Its Implications for Epidemic Spreading," J. Stat. Phys. **173**, 1110–1123 (2018).
[47] V. Marceau, P.-A. Noël, L. Hébert-Dufresne, A. Allard, and L. J. Dubé, "Adaptive networks: Coevolution of disease and topology," Phys. Rev. E **82**, 036116 (2010).
[48] J. P. Gleeson, "High-Accuracy Approximation of Binary-State Dynamics on Networks," Phys. Rev. Lett. **107**, 068701 (2011).
[49] G. St-Onge, V. Thibeault, A. Allard, L. J. Dubé, and L. Hébert-Dufresne, "Social Confinement and Mesoscopic Localization of Epidemics on Networks," Phys. Rev. Lett. **126**, 098301 (2021).
[50] G. St-Onge, V. Thibeault, A. Allard, L. J. Dubé, and L. Hébert-Dufresne, "Master equation analysis of mesoscopic localization in contagion dynamics on higher-order networks," Phys. Rev. E **103**, 032301 (2021).
[51] S. V. Scarpino, A. Allard, and L. Hébert-Dufresne, "The effect of a prudent adaptive behaviour on disease transmission," Nat. Phys. **12**, 1042 (2016).
[52] M. Granovetter, "Threshold Models of Collective Behavior," Am. J. Sociol. **83**, 1420–1443 (1978).
[53] D. J. Watts, "A simple model of global cascades on random networks," Proc. Natl. Acad. Sci. U.S.A. **99**, 5766–5771 (2002).
[54] P. S. Dodds and D. J. Watts, "Universal Behavior in a Generalized Model of Contagion," Phys. Rev. Lett. **92**, 218701 (2004).
[55] A. S. Mata, R. S. Ferreira, and S. C. Ferreira, "Heterogeneous pair-approximation for the contact process on complex networks," New J. Phys. **16**, 053006 (2014).
[56] G. St-Onge, J.-G. Young, E. Laurence, C. Murphy, and L. J. Dubé, "Phase transition of the susceptible-infected-susceptible dynamics on time-varying configuration model networks," Phys. Rev. E **97**, 022305 (2018).
[57] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, 2001) p. 57.
[58] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03 (Association for Computing Machinery, New York, NY, USA, 2003) pp. 137–146.
[59] P. Dodds and D. J. Watts, "Threshold models of social influence," in *The Oxford Handbook of Analytical Sociology* (Oxford University Press, 2009).
[60] F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," Nature **524**, 65–68 (2015).
[61] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *2010 IEEE international conference on data mining* (IEEE, 2010) pp. 88–97.
[62] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," Nat. Phys. **6**, 888–893 (2010).
[63] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," Physica A **391**, 1777–1787 (2012).
[64] Ş. Erkol, C. Castellano, and F. Radicchi, "Systematic comparison between methods for the detection of influential spreaders in complex networks," Sci. Rep. **9**, 15095 (2019).
[65] G. Poux-Médard, R. Pastor-Satorras, and C. Castellano, "Influential spreaders for recurrent epidemics on networks," Phys. Rev. Research **2**, 023332 (2020).
[66] F. Amato, V. Moscato, A. Picariello, and G. Sperlí, "Influence maximization in social media networks using hypergraphs," in *Green, Pervasive, and Cloud Computing* (Springer International Publishing, Cham, 2017) p. 207.
[67] J. Zhu, J. Zhu, S. Ghosh, W. Wu, and J. Yuan, "Social influence maximization in hypergraph in social networks," IEEE Transactions on Network Science and Engineering **6**, 801 (2019).
[68] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki, "Small but slow world: How network topology and burstiness slow down spreading," Phys. Rev. E **83**, 025102 (2011).
[69] M. Starnini, A. Machens, C. Cattuto, A. Barrat, and R. Pastor-Satorras, "Immunization strategies for epidemic processes in time-varying contact networks," J. Theor. Biol. **337**, 89–100 (2013).
[70] G. Lawyer, "Understanding the influence of all nodes in a network," Sci. Rep. **5**, 8665 (2015).
[71] S. J. Jackson and B. Foucault Welles, "Hijacking #myNYPD: Social Media Dissent and Networked Counterpublics," J. Commun. **65**, 932–952 (2015).
[72] V. Sood and S. Redner, "Voter Model on Heterogeneous Graphs," Phys. Rev. Lett. **94**, 178701 (2005).
[73] J. P. Gleeson, "Binary-State Dynamics on Complex Networks: Pair Approximation and Beyond," Phys. Rev. X **3**, 021004 (2013).
[74] F. Tudisco and D. J. Higham, "Node and edge eigenvector centrality for hypergraphs," (2021), arXiv:2101.06215.
[75] R. Pastor-Satorras and A. Vespignani, "Immunization of complex networks," Phys. Rev. E **65**, 036104 (2002).
[76] G. St-Onge, "Samplableset," `https://github.com/gstonge/SamplableSet` (2018).
[77] G. St-Onge, J.-G. Young, L. Hébert-Dufresne, and L. J. Dubé, "Efficient sampling of spreading processes on complex networks using a composition and rejection algorithm," Comput. Phys. Commun. **240**, 30–37 (2019).
[78] M. M. de Oliveira and R. Dickman, "How to simulate the quasistationary state," Phys. Rev. E **71**, 016129 (2005).
[79] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems, "High-Resolution Measure-

ments of Face-to-Face Contact Patterns in a Primary School," PLoS One **6**, e23176 (2011).

[80] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, "Simplicial closure and higher-order link prediction." Proc. Natl. Acad. Sci. U.S.A. **115**, E11221–E11230 (2018).

[81] A. Patania, G. Petri, and F. Vaccarino, "The shape of collaborations," EPJ Data Sci. **6**, 18 (2017).

[82] B. K. Fosdick, Dl B. Larremore, J. Nishimura, and J. Ugander, "Configuring random graph models with fixed degree sequences," SIAM Rev. **60**, 315 (2018).

[83] Guillaume St-Onge, Iacopo Iacopini, Vito Latora, Alain Barrat, Giovanni Petri, Antoine Allard, and Laurent Hébert-Dufresne, "Influential groups data," (2021).

[84] Guillaume St-Onge, Iacopo Iacopini, Vito Latora, Alain Barrat, Giovanni Petri, Antoine Allard, and Laurent Hébert-Dufresne, "gstonge/influential-groups," (2021).