Entropy-based models to randomise real-world hypergraphs

Check for updates

Fabio Saracco [1,2,3] ✉, Giovanni Petri[4], Renaud Lambiotte [5] ✉ & Tiziano Squartini [3,6]

Network theory has often disregarded many-body relationships, solely focusing on pairwise interactions: neglecting them, however, can lead to misleading representations of complex systems. Hypergraphs represent a suitable framework for describing polyadic interactions. Here, we leverage the representation of hypergraphs based on the incidence matrix for extending the entropy-based approach to higher-order structures: in analogy with the Exponential Random Graphs, we introduce the Exponential Random Hypergraphs (ERHs). After exploring the asymptotic behaviour of thresholds generalising the percolation one, we apply ERHs to study real-world data. First, we generalise key network metrics to hypergraphs; then, we compute their expected value and compare it with the empirical one, in order to detect deviations from random behaviours. Our method is analytically tractable, scalable and capable of revealing structural patterns of real-world hypergraphs that differ significantly from those emerging as a consequence of simpler constraints.

Networks provide a powerful language to model interacting systems[1,2]. Within such a framework, the basic unit of interaction, i.e., the *edge*, involves two nodes, and the complexity of the structure as a whole arises from the combination of these units. Despite its many successes, network science disregards certain aspects of interacting systems, notably the possibility that more-than-two constituent units could interact at a time[3]. Yet, it has been increasingly shown that, for a variety of systems, interactions cannot be always decomposed into a pairwise fashion and that neglecting higher-order ones can lead to an incomplete, if not misleading, representation of them[3–5]: examples include chemical reactions involving several compounds, coordination activities within small teams of co-working people and brain activities mediated by groups of neurons. Generally speaking, thus, modelling the joint coordination of multiple entities calls for a generalisation of the traditional edge-centred framework.

While approaches focusing on the so-called *simplicial complexes* have been proposed[6,7], an increasingly popular alternative to support a science of many-body interactions is provided by *hypergraphs*, as these mathematical objects allow nodes to interact in groups without posing restrictions, such as the 'hierarchical" ones characterising the former ones[8], which, in fact, include all the subsets of a given simplex[6].

Several contributions to the definition of analytical tools for their study have already appeared[9–12]: while some pertain to the purely mathematical literature and have considered probabilistic hypergraphs with the aim of studying properties such as the existence of cycles, cliques, etc.[13–15], others

have adopted approaches rooted into statistical physics. Among the latter ones, some have proposed microcanonical approaches[9,11] while others have focused on canonical ones[10,12,16].

The present contribution aims at extending the class of entropy-based benchmarks[17–20] to hypergraphs while providing a coherent framework to formally derive the canonical approaches that have been proposed so far. These models work by preserving a given set of quantities while randomising everything else, hence destroying all possible correlations between structural properties except for those that are genuinely embodied into the constraints themselves[20–22]. The versatility of such an approach allows it to be employed either in presence of full information (to quantify the level of self-organisation of a given configuration by identifying the patterns that are incompatible with simpler, structural constraints[23–27]) or in presence of partial information (to infer the missing portion of a given configuration[28]).

Our strategy for defining null models for hypergraphs is based on the randomisation of their incidence matrix, i.e., the (generally, rectangular) table contains information about the connectivity of nodes (the set of hyperedges they belong to) and the connectivity of hyperedges (the set of nodes they cluster). We will explicitly derive two members of this novel class of models hereby named *Exponential Random Hypergraphs (ERH)*, i.e., the Random Hypergraph Model (RHM) (RHM, generalising the Erdös-Rényi Model) and the Hypergraph Configuration Model (HCM) (HCM, generalising the Configuration Model), and provide an analytical characterisation of their behaviour. To this aim, we will exploit the formal equivalence

[1]'Enrico Fermi' Research Center (CREF), Via Panisperna 89A, Rome, Italy. [2]'Mauro Picone' Institute for Applied Computing (IAC), CNR, Via Madonna del Piano 10, Sesto Fiorentino, Italy. [3]IMT School for Advanced Studies, Piazza San Francesco 19, Lucca, Italy. [4]Network Science Institute, Northeastern University London, Devon House 58 St. Katharine's Way, London, United Kingdom. [5]Mathematical Institute, University of Oxford, Woodstock Road, Oxford, United Kingdom. [6]INdAM-GNAMPA, Istituto Nazionale di Alta Matematica 'Francesco Severi', P.le Aldo Moro 5, Rome, Italy. ✉e-mail: fabio.saracco@cref.it; renaud.lambiotte@maths.ox.ac.uk

between the incidence matrix of a hypergraph and the biadjacency matrix of a bipartite graph[10,12]. Afterwards, we will employ the HCM to assess the statistical significance of a number of patterns characterising several real-world hypergraphs.

Methods
Formalism and basic quantities
A hypergraph can be defined as a pair $H(\mathcal{V}, \mathcal{E}_H)$ where $\mathcal{V}$ is the set of vertices and $\mathcal{E}_H$ is the set of hyperedges. Moving from the observation that the edge set $\mathcal{E}_G$ of a traditional, binary, undirected graph $G(\mathcal{V}, \mathcal{E}_G)$ is a subset of the power set of $\mathcal{V}$, several definitions of the hyperedge set $\mathcal{E}_H$ have been provided: the two most popular ones are those proposed in[29,30], where hyperedges tie one or more vertices, and in[31], where hyperedges are allowed to be empty sets as well. Hereby, we adopt the definition according to which $\mathcal{E}_H$ is a multiset of the power set of $\mathcal{V}$: since the concept of 'multiset' extends the concept of 'set', allowing for multiple instances of (each of) its elements, our choice implies that we are considering non-simple hypergraphs, admitting loops and parallel edges (i.e., hyperedges involving exactly the same nodes) of any size, including 0 (corresponding to empty hyperedges) and $|\mathcal{V}|$ (corresponding to hyperedges clustering all vertices together).

As for traditional graphs, an algebraic representation of hypergraphs can be devised as well. In analogy with the traditional case, we call the cardinality of the set of nodes $|\mathcal{V}| \equiv N$ and the cardinality of the set of hyperedges $|\mathcal{E}_H| \equiv L$: then, we consider the $N \times L$ table known as *incidence matrix*, each row of which corresponds to a node and each column of which corresponds to a hyperedge. If we indicate the incidence matrix with $\mathbf{I}$, its generic entry $I_{i\alpha}$ will be 1 if vertex i belongs to hyperedge α and 0 otherwise. Notice that the number of 1s along each row can vary between 0 and L, the former case indicating an isolated node and the latter one indicating a node that belongs to each hyperedge; similarly, the number of 1s along each column can vary between 0 and N, the former case indicating an empty hyperedge and the latter one indicating a hyperedge that includes all nodes. As explicitly noticed elsewhere[9,10,12], representing a hypergraph via its incidence matrix is equivalent to considering the bipartite graph defined by the sets $\mathcal{V}$ and $\mathcal{E}_H$ - more formally, the function that assigns each hypergraph to the bipartite graph associated with it is a bijection when both nodes and hyperedges are uniquely labelled[9]. For instance, the incidence matrix $\mathbf{I}$ describing the binary, undirected hypergraph shown in Fig. 1 is the following:

$$
\mathbf{I} = \begin{array}{c} \\ \mathbf{n_1} \\ \mathbf{n_2} \\ \mathbf{n_3} \\ \mathbf{n_4} \\ \mathbf{n_5} \\ \mathbf{n_6} \end{array} \begin{array}{c} e_1 \ e_2 \ e_3 \ e_4 \ e_5 \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}. \tag{1}
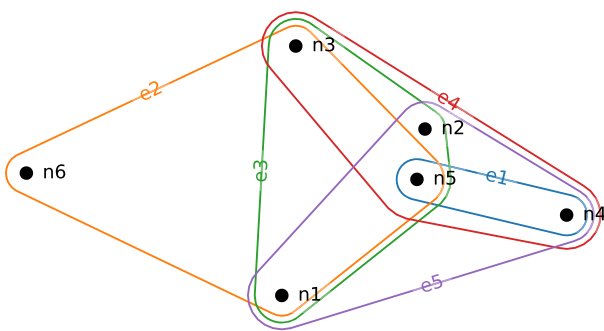$$

Fig. 1 | Cartoon representation of a hypergraph. Black dots represent nodes, while the coloured shapes represent hyperedges. The incidence matrix describing the present hypergraph is defined by Eq. (1).

Once the incidence matrix has been defined, several quantities needed for the description of hypergraphs can be defined quite straightforwardly: for example, the 'degree of node i'' (hereby, *degree*) reads

$$
k_i = \sum_{\alpha=1}^{L} I_{i\alpha} \tag{2}
$$

and counts the number of hyperedges that are incident to it; analogously, the 'degree of hyperedge α'' (hereby, *hyperdegree*) reads

$$
h_\alpha = \sum_{i=1}^{N} I_{i\alpha} \tag{3}
$$

and counts the number of nodes it clusters. Both the sum of degrees and that of hyperdegrees equal the total number of 1s, i.e., $\sum_{i=1}^{N} k_i = \sum_{i=1}^{N} \sum_{\alpha=1}^{L} I_{i\alpha} = \sum_{\alpha=1}^{L} \sum_{i=1}^{N} I_{i\alpha} = \sum_{\alpha=1}^{L} h_\alpha \equiv T$. Importantly, a node degree no longer coincides with the number of its neighbours: instead, it matches the number of hyperedges it belongs to; a hyperdegree, instead, provides information about the hyperedge size. Analogously, T paves the way for the alternative definition of 'density of connections'' reading $\rho = T/NL \equiv h/N$, i.e., the ratio between the (average) number of nodes each hyperedge clusters and the total number of nodes.

Binary, undirected hypergraphs randomisation
An early attempt to define randomisation algorithms for hypergraphs within a framework closely resembling ours can be found in ref. 32. Its authors, however, have just considered hyperedges that are incident to triples of nodes—a framework that has been, later, applied to the study of the World Trade Network[10].

Considering the incidence matrix has two clear advantages over the tensor-based representation employed in[10,32]: *i) generality*, because the incidence matrix allows hyperedges of any size to be handled at once; *ii) compactness*, because the order of the tensor $\mathbf{I}$ never exceeds two, hence allowing any hypergraph to be represented as a traditional, bipartite graph.

In order to extend the rich set of null models induced by graph-specific global and local constraints to hypergraphs, we first need to identify the quantities that can play this role within the novel setting. In what follows, we will consider the total number of 1s, i.e., T, the degree and the hyperdegree sequences, i.e., $\{k_i\}_{i=1}^{N}$ and $\{h_\alpha\}_{\alpha=1}^{L}$—either separately or in a joint fashion; moreover, we will distinguish between microcanonical and canonical randomisation techniques.

Homogeneous benchmarks: the RHM
Microcanonical formulation. The model is defined by just one, global constraint, which, in our case, reads

$$
T = \sum_{i=1}^{N} \sum_{\alpha=1}^{L} I_{i\alpha}; \tag{4}
$$

Its microcanonical version extends the model by Erdös and Rényi[33]—also known as Random Graph Model—to hypergraphs and prescribes to count the number of incidence matrices that are compatible with a given, total number of 1s, say T^*: they are

$$
\Omega_{\text{RHM}} = \binom{V}{T^*} \tag{5}
$$

with $V \equiv NL$ being the total number of entries of the incidence matrix $\mathbf{I}$. Once the total number of configurations composing the microcanonical ensemble has been determined, a procedure to generate them is needed: in the case of the RHM, it simply boils down to reshuffling the entries of the incidence matrix, a procedure ensuring that the total number of 1s is kept fixed while any, other correlation is destroyed.

Canonical formulation. The canonical version of the RHM, instead, extends the model by Gilbert[34] and rests upon the constrained maximisation of Shannon entropy, i.e.

$$\mathscr{L} \equiv S[P] - \sum_{i=0}^{M} \theta_i [P(\mathbf{I})C_i(\mathbf{I}) - \langle C_i \rangle] \tag{6}$$

where $S[P] = -\sum_{\mathbf{I} \in \mathscr{I}} P(\mathbf{I}) \ln P(\mathbf{I})$, $C_0 \equiv \langle C_0 \rangle \equiv 1$ sums up the normalisation condition and the remaining $M-1$ constraints represent proper topological properties. The sum defining Shannon entropy runs over the set $\mathscr{I}$ of incidence matrices described in the introductory paragraph and known as *canonical ensemble*. Such an optimisation procedure defines the ERH framework, described by the expression

$$P(\mathbf{I}) = \frac{e^{-H(\mathbf{I})}}{Z} = \frac{e^{-H(\mathbf{I})}}{\sum_{\mathbf{I} \in \mathscr{I}} e^{-H(\mathbf{I})}} = \frac{e^{-\sum_{i=1}^{M} \theta_i C_i(\mathbf{I})}}{\sum_{\mathbf{I} \in \mathscr{I}} e^{-\sum_{i=1}^{M} \theta_i C_i(\mathbf{I})}}. \tag{7}$$

In the simplest case, the only global constraint is represented by T and leads to the expression

$$P(\mathbf{I}) = \frac{e^{-\theta T(\mathbf{I})}}{\sum_{\mathbf{I} \in \mathscr{I}} e^{-\theta T(\mathbf{I})}} = \frac{e^{-\sum_{i=1}^{N} \sum_{\alpha=1}^{L} \theta I_{i\alpha}}}{\sum_{\mathbf{I} \in \mathscr{I}} e^{-\sum_{i=1}^{N} \sum_{\alpha=1}^{L} \theta I_{i\alpha}}} = \prod_{i=1}^{N} \prod_{\alpha=1}^{L} x^{I_{i\alpha}} \prod_{i=1}^{N} \prod_{\alpha=1}^{L} (1+x)^{-1} \tag{8}$$

that can be rewritten as

$$P(\mathbf{I}) = \prod_{i=1}^{N} \prod_{\alpha=1}^{L} p^{I_{i\alpha}} (1-p)^{1-I_{i\alpha}} = p^{T(\mathbf{I})} (1-p)^{NL-T(\mathbf{I})} \tag{9}$$

with $e^{-\theta} \equiv x$ and $p \equiv x/(1+x)$. The canonical ensemble, now, includes all $N \times L$, rectangular matrices whose number of entries equal to 1 ranges from 0 to NL. According to such a model, the entries of the incidence matrix are i.i.d. Bernoulli random variables, i.e., $I_{i\alpha} \sim \mathrm{Ber}(p)$, $\forall\ i, \alpha$; as a consequence, the total number of 1s, the degrees and the hyperdegrees obey Binomial distributions, being all defined as sums of i.i.d. Bernoulli random variables: specifically, $T \sim \mathrm{Bin}(NL, p)$, $k_i \sim \mathrm{Bin}(L, p)$, $\forall\ i$ and $h_\alpha \sim \mathrm{Bin}(N, p)$, $\forall\ \alpha$, in turn, implying that $\langle T \rangle_{\mathrm{RHM}} = NLp$, $\langle k_i \rangle_{\mathrm{RHM}} = Lp$, $\forall\ i$ and $\langle h_\alpha \rangle_{\mathrm{RHM}} = Np$, $\forall\ \alpha$.

Parameter estimation. In order to ensure that $\langle T \rangle_{\mathrm{RHM}} = T^*$, parameters have to be tuned opportunistically. To this aim, the likelihood maximisation principle can be invoked[19]: it prescribes to maximise the function $\mathcal{L}(\theta) \equiv \ln P(\mathbf{I}^* | \theta)$ with respect to the unknown parameter that defines it. Such a recipe leads us to find

$$p = \rho^* = \frac{T^*}{NL} \tag{10}$$

with $T^* = T(\mathbf{I}^*)$ indicating the empirical value of the constraint defining the RHM.

The RHM (also considered in[12], although without providing any derivation from first principles, and in[16], although without providing any recipe for the estimation of its parameter) is formally equivalent to the Bipartite Random Graph Model[35]. Such an identification is guaranteed by our focus on non-simple hypergraphs.

Estimation of the number of empty hyperedges. Non-simple hypergraphs admit the presence of empty as well as parallel hyperedges. As this type of structures is associated with configurations that may be regarded as problematic (since not observed in empirical data), we evaluate how frequently they appear in the ensembles induced by our benchmarks. Let us denote the number of empty hyperedges, i.e., the number of hyperedges whose hyperdegree equals zero, with $N_\emptyset$: since the hyperdegrees are i.i.d. Binomial

random variables, $N_\emptyset \sim \mathrm{Bin}(L, p_\emptyset)$ where

$$p_\emptyset \equiv (1-p)^N \tag{11}$$

is the probability for the generic hyperedge to be empty or, equivalently, for its hyperdegree to equal zero; the expected number of empty hyperedges reads

$$\langle N_\emptyset \rangle = L p_\emptyset = L(1-p)^N. \tag{12}$$

Let us, now, inspect the behaviour of $p_\emptyset$ the ensemble induced by the RHM as the density of 1s in the incidence matrix, i.e., $p = T/NL$, varies. The two regimes of interest are the *dense* one, defined by $T \to NL$, and the *sparse* one, defined by $T \to 0$. In the dense case, one finds

$$\lim_{T \to NL} p_\emptyset = \lim_{T \to NL} \left(1 - \frac{T}{NL}\right)^N = 0, \tag{13}$$

a relationship inducing $\langle N_\emptyset \rangle \overset{T \to NL}{\to} 0$: in words, the probability of observing empty hyperedges progressively vanishes as the density of 1s increases. Consistently, in the sparse case, one finds

$$\lim_{T \to 0} p_\emptyset = \lim_{T \to 0} \left(1 - \frac{T}{NL}\right)^N = 1, \tag{14}$$

a relationship inducing $\langle N_\emptyset \rangle \overset{T \to 0}{\to} L$: in words, the probability of observing empty hyperedges progressively rises as the density of 1s decreases.

To evaluate the density of 1s in the incidence matrix in correspondence of which the transition from the sparse to the dense regime happens, let us consider the case $N \gg 1$: more formally, this amounts to consider the asymptotic framework defined by letting $N \to +\infty$ while posing $T = O(L)$ - equivalently, defined by posing $p = O(1/N)$. Since $p = T/NL$ and $h = T/L$ remain finite, the probability for the generic hyperedge to be empty obeys the relationship

$$\lim_{N \to +\infty} p_\emptyset = \lim_{N \to +\infty} \left(1 - \frac{h}{N}\right)^N = e^{-h}, \tag{15}$$

i.e., remains finite as well: consistently, the expected number of empty hyperedges becomes

$$\lim_{N \to +\infty} \langle N_\emptyset \rangle = \lim_{N \to +\infty} L p_\emptyset = L e^{-h}; \tag{16}$$

upon imposing $L e^{-h} \leq 1$, i.e., that the expected number of empty hyperedges is *at most* 1, one derives what may be called *filling threshold*, corresponding to $h_f^{\mathrm{RHM}} \equiv \ln L$. In words, a value

$$p > p_f^{\mathrm{RHM}} = \frac{h_f^{\mathrm{RHM}}}{N} = \frac{\ln L}{N} \tag{17}$$

ensures that the expected number of empty hyperedges in our random hypergraph is strictly less than one. As a last observation, let us notice that evaluating $p_\emptyset$ in correspondence with the filling threshold returns the value $1/L$.

A comparison with simple graphs: estimating the number of isolated nodes. A similar line of reasoning can be repeated for traditional graphs, the aim being, now, that of estimating N_0, i.e., the number of isolated nodes. To this aim, let us consider the asymptotic framework defined by letting $N \to +\infty$ while posing $L = O(N)$ - equivalently, defined by posing $q = O(1/N)$. Since $q = 2L/N(N-1)$ and $k \equiv 2L/(N-1)$ remains finite, the probability for the

generic node i to be isolated obeys the relationship

$$\lim_{N \to +\infty} q_0 = \lim_{N \to +\infty} (1-q)^{N-1} = \lim_{N \to +\infty} \left(1 - \frac{k}{N-1}\right)^{N-1} = e^{-k},$$

(18)

i.e., remains finite as well: this, in turn, implies that the expected number of isolated nodes $\langle N_0 \rangle = N q_0$ obeys the relationship

$$\lim_{N \to +\infty} \langle N_0 \rangle = \lim_{N \to +\infty} N q_0 = N e^{-k};$$

(19)

upon imposing $N e^{-k} \le 1$, i.e., that the expected number of isolated nodes is *at most* 1, one derives the *connectivity threshold*, corresponding to $k_c^{\mathrm{RGM}} \equiv \ln N$. In words, a value

$$q > q_c^{\mathrm{RGM}} = \frac{k_c^{\mathrm{RGM}}}{N} = \frac{\ln N}{N}$$

(20)

ensures that the expected number of isolated nodes in our random graph is strictly less than one[1]. As a last observation, let us notice that evaluating q_0 in correspondence with the connectivity threshold returns the value $1/N$.

We also note that N is the only quantity playing a relevant role in the case of graphs, while an interplay between L and N can be observed in the case of hypergraphs: in both cases, however, a condition on connectivity is present, driven by the request that the objects under investigations (nodes on the one hand and hyperedges on the other) have a non-zero number of connections.

Estimation of the number of parallel hyperedges. Let us now move to considering the issue of parallel hyperedges. By definition, two, parallel hyperedges α and β are characterised by identical columns: hence, their Hamming distance, defined as the number of positions at which the corresponding symbols are different, is zero. More formally,

$$d_{\alpha\beta} \equiv \sum_{i=1}^{N} [I_{i\alpha}(1 - I_{i\beta}) + I_{i\beta}(1 - I_{i\alpha})],$$

(21)

a sum whose generic addendum is 1 in just two cases: either $I_{i\alpha} = 1$ and $I_{i\beta} = 0$, or $I_{i\alpha} = 0$ and $I_{i\beta} = 1$. Since $d_{\alpha\beta} \sim \mathrm{Bin}(N, 2p(1-p))$, one finds that

$$p_{/\!/}^{\alpha\beta} \equiv P(d_{\alpha\beta} = 0) = [1 - 2p(1-p)]^N$$

(22)

and that

$$\langle d_{\alpha\beta} \rangle = 2p(1-p)N$$

(23)

$\forall \ \alpha \ne \beta$. Since $p(1-p) = (T/NL)(1 - T/NL)$, one finds that

$$\lim_{T \to NL} p(1-p) = \lim_{T \to 0} p(1-p) = 0,$$

(24)

i.e., $p(1-p)$ vanishes in both regimes, a result further implying that both $P(d_{\alpha\beta} = 0) \overset{T \to NL}{\to} 1$ and $P(d_{\alpha\beta} = 0) \overset{T \to 0}{\to} 1$ and that both $\langle d_{\alpha\beta} \rangle \overset{T \to NL}{\to} 0$ and $\langle d_{\alpha\beta} \rangle \overset{T \to 0}{\to} 0$: in words, the probability of observing parallel hyperedges progressively rises both as a consequence of having many 1s and as a consequence of having few 1s. Analogously, for the expected Hamming distance.

Let us, now, evaluate the expected Hamming distance between any two hyperedges α and β within the asymptotic framework defined by letting $N \to +\infty$ while posing $T = O(L)$ - equivalently, defined by posing $p = O(1/N)$. Since $p = T/NL$ and $h \equiv T/L$ remains finite, one finds that

$$\lim_{N \to +\infty} \langle d_{\alpha\beta} \rangle = \lim_{N \to +\infty} \frac{2h}{N}\left(1 - \frac{h}{N}\right)N = 2h;$$

(25)

upon imposing $2h \ge 1$, i.e., that the expected Hamming distance between any two hyperedges α and β is *at least* 1, one derives what may be called *resolution threshold*, corresponding to $h_r^{\mathrm{RHM}} \equiv 1/2$. In words, a value $p > p_r^{\mathrm{RHM}} = h_r^{\mathrm{RHM}}/N = 1/2N$ ensures that, on average, any two hyperedges α and β differ by at least one element.

To derive a global condition on the total number of parallel hyperedges, note that, although the overlaps between pairs of hyperedges cannot be treated as i.i.d. random variables, the expected number of parallel hyperedges can still be computed explicitly. Upon posing $p_{/\!/} \equiv p_{/\!/}^{\alpha\beta}$, it reads

$$\langle N_{/\!/} \rangle = \sum_{\alpha=1}^{L} \sum_{\substack{\beta = 1 \\ \beta > \alpha}}^{L} p_{/\!/} = \frac{L(L-1)}{2} p_{/\!/}.$$

(26)

Considering that $p_{/\!/} \overset{N \to +\infty}{\to} e^{-2h}$ and imposing $\langle N_{/\!/} \rangle \le 1$, i.e., that the expected number of parallel hyperedges is *at most* 1, one derives what may be called a *multiple resolution threshold*, corresponding to $h_m^{\mathrm{RHM}} \equiv \ln L - \ln\sqrt{2} \lesssim h_f^{\mathrm{RHM}}$. In words, a value

$$p > p_m^{\mathrm{RHM}} = \frac{h_m^{\mathrm{RHM}}}{N} = \frac{\ln L}{N} - \frac{\ln\sqrt{2}}{N} \lesssim \frac{\ln L}{N} = \frac{h_f^{\mathrm{RHM}}}{N} = p_f^{\mathrm{RHM}}$$

(27)

(also) ensures that the expected number of parallel hyperedges in our random hypergraph is strictly less than one.

Estimation of the percolation threshold. The two thresholds derived in the previous subsections emerge in consequence of the attempts to solve the problems related to the appearance of empty as well as parallel hyperedges. Remarkably, a third threshold exists: known as *percolation threshold*, it was first derived in ref. [12], following the definition according to which any two hyperedges are said to be connected if they share at least one node. Here, we re-derive the percolation threshold by considering the 'hypergraph to graph' projection (see also below): more formally, the total number of nodes shared by hyperedge α with any other hyperedge reads

$$\sigma_\alpha = \sum_{i=1}^{N} \sum_{\substack{\beta = 1 \\ \beta \ne \alpha}}^{L} I_{i\alpha} I_{i\beta},$$

(28)

its expected value being

$$\langle \sigma_\alpha \rangle = N(L-1)p^2 \simeq NLp^2;$$

(29)

imposing $\langle \sigma_\alpha \rangle = 1$ leads to find the value

$$p_p^{\mathrm{RHM}} = \frac{h_p^{\mathrm{RHM}}}{N} = \frac{1}{\sqrt{NL}}$$

(30)

which, in turn, induces the value $h_p^{\mathrm{RHM}} \equiv \sqrt{N/L}$. In words, a value $p > p_p^{\mathrm{RHM}}$ ensures that any two hyperedges in our random hypergraph share, on average, at least one node.

Heterogeneous benchmarks: the HCM

Microcanonical formulation. The number of constraints can be enlarged to include the degrees, i.e., the sequence $\{k_i\}_{i=1}^{N}$, and the hyperdegrees, i.e., the sequence $\{h_\alpha\}_{\alpha=1}^{L}$. Although counting the number of configurations on which both sequences match their empirical values is a hard task, numerical recipes that shuffle the entries of a rectangular matrix, while preserving its marginals, exist[9,36–38]. It should be, however, noticed that, if not carefully implemented, algorithms of the kind may lead to a non-uniform exploration of the space of configurations[39,40]; moreover, the issue concerning the time needed to collect a sufficiently large number of configurations should be

addressed even in the presence of an ergodic system. A recent proposal is that of extending the traditional Curveball algorithm to hypergraphs[38].

Canonical formulation. Solving the corresponding problem in the canonical framework is, instead, straightforward. Indeed, Shannon entropy maximisation leads to

$$P(\mathbf{I}) = \frac{e^{-\sum_{i=1}^{N}\alpha_i k_i(\mathbf{I})-\sum_{\alpha=1}^{L}\beta_\alpha h_\alpha(\mathbf{I})}}{\sum_{\mathbf{I}\in\mathscr{I}}e^{-\sum_{i=1}^{N}\alpha_i k_i(\mathbf{I})-\sum_{\alpha=1}^{L}\beta_\alpha h_\alpha(\mathbf{I})}} = \frac{e^{-\sum_{i=1}^{N}\sum_{\alpha=1}^{L}(\alpha_i+\beta_\alpha)I_{i\alpha}}}{\sum_{\mathbf{I}\in\mathscr{I}}e^{-\sum_{i=1}^{N}\sum_{\alpha=1}^{L}(\alpha_i+\beta_\alpha)I_{i\alpha}}}$$
$$= \prod_{i=1}^{N}x_i^{k_i(\mathbf{I})}\prod_{\alpha=1}^{L}y_\alpha^{h_\alpha(\mathbf{I})}\prod_{i=1}^{N}\prod_{\alpha=1}^{L}(1+x_iy_\alpha)^{-1},$$

(31)

an expression that can be re-written as

$$P(\mathbf{I}) = \prod_{i=1}^{N}\prod_{\alpha=1}^{L}p_{i\alpha}^{I_{i\alpha}}(1-p_{i\alpha})^{1-I_{i\alpha}} \tag{32}$$

with $e^{-\alpha_i} \equiv x_i$, $\forall i$, $e^{-\beta_\alpha} \equiv y_\alpha$, $\forall \alpha$ and $p_{i\alpha} \equiv x_iy_\alpha/(1+x_iy_\alpha)$, $\forall i, \alpha$. According to such a model, the entries of the incidence matrix of a hypergraph are independent random variables that obey different Bernoulli distributions, i.e. $I_{i\alpha} \sim \text{Ber}(p_{i\alpha})$, $\forall i, \alpha$. As a consequence, both degrees and hyperdegrees obey Poisson-Binomial distributions, i.e. $k_i \sim \text{PoissBin}(L, \{p_{i\alpha}\}_{\alpha=1}^{L})$, $\forall i$ and $h_\alpha \sim \text{PoissBin}(N, \{p_{i\alpha}\}_{i=1}^{N})$, $\forall \alpha$[24].

Parameter estimation. In this case, solving the likelihood maximisation problem amounts to solving the system of coupled equations

$$k_i^* = \sum_{\alpha=1}^{L}\frac{x_iy_\alpha}{1+x_iy_\alpha} = \sum_{\alpha=1}^{L}p_{i\alpha} = \langle k_i\rangle, \forall i \tag{33}$$

$$h_\alpha^* = \sum_{i=1}^{N}\frac{x_iy_\alpha}{1+x_iy_\alpha} = \sum_{i=1}^{N}p_{i\alpha} = \langle h_\alpha\rangle, \forall \alpha \tag{34}$$

ensuring that $\langle k_i\rangle = k_i^*$, $\forall i$, $\langle h_\alpha\rangle = h_\alpha^*$, $\forall \alpha$ (and, as a consequence, $\langle T\rangle = T^*$). In case hypergraphs are sparse and in the absence of hubs

$$p_{i\alpha} \simeq x_iy_\alpha = \frac{k_i^*h_\alpha^*}{T^*}, \forall i, \alpha. \tag{35}$$

The HCM reduces to a 'partial' Configuration Model[24] when either the degree or the hyperdegree sequence is left unconstrained (see also Supplementary Note 1 of the Supplementary Information). The canonical ensemble of each randomisation model (Supplementary Table 1 in the Supplementary Information sums up the sets of constraints defining them) can be explicitly sampled by considering each entry of $\mathbf{I}$, drawing a real number $u_{i\alpha} \in U[0, 1]$ and posing $I_{i\alpha} = 1$ if $u_{i\alpha} \leq p_{i\alpha}$, $\forall i, \alpha$.

The HCM is formally equivalent to the Bipartite Configuration Model[35]. Such an identification is guaranteed by our focus on non-simple hypergraphs.

Estimation of the number of empty hyperedges. Let us, now, consider the probability for the generic hyperedge α to be empty or, in other terms, that its hyperdegree h_α is zero. Upon remembering that $h_\alpha \sim \text{PoissBin}(N, \{p_{i\alpha}\}_{i=1}^{N})$, one finds

$$p_\emptyset^\alpha \equiv \prod_{i=1}^{N}(1-p_{i\alpha}) \tag{36}$$

while the expected number of empty hyperedges, now, reads

$$\langle N_\emptyset\rangle \equiv \sum_{\alpha=1}^{L}p_\emptyset^\alpha = \sum_{\alpha=1}^{L}\prod_{i=1}^{N}(1-p_{i\alpha}). \tag{37}$$

As previously done, let us inspect the behaviour of the aforementioned quantities on the ensemble induced by the HCM as the density of 1s in the incidence matrix varies. Although it depends on (the heterogeneity of) the sets of coefficients $\{x_i\}_{i=1}^{N}$ and $\{y_\alpha\}_{\alpha=1}^{L}$, general conclusions can be still drawn within a simpler framework. To this aim, let us consider the functional form reading

$$p_{i\alpha} = \frac{zf_ig_\alpha}{1+zf_ig_\alpha}, \forall i, \alpha \tag{38}$$

where the vector of fitnesses $\{f_i\}_{i=1}^{N}$ accounts for the heterogeneity of nodes, the vector of fitnesses $\{g_\alpha\}_{\alpha=1}^{L}$ accounts for the heterogeneity of hyperedges, and z tunes the density of 1s in the incidence matrix - 'partial' Configuration Models are recovered upon posing either $f_i = 1$, $\forall i$ or $g_\alpha = 1$, $\forall \alpha$.) Within such a framework, the fitnesses of the nodes and the fitnesses of the hyperedges can be drawn from any distribution. The dense and sparse regimes are now defined by the positions $z \to +\infty$ and $z \to 0$, respectively. In the dense case, one finds

$$\lim_{z\to+\infty}p_\emptyset^\alpha = \lim_{z\to+\infty}\prod_{i=1}^{N}\left(1-\frac{zf_ig_\alpha}{1+zf_ig_\alpha}\right) = 0, \tag{39}$$

a relationship inducing $\langle N_\emptyset\rangle \xrightarrow{z\to+\infty} 0$: in words, the probability of observing empty hyperedges progressively vanishes as the density of 1s increases. Consistently, in the sparse case one finds

$$\lim_{z\to0}p_\emptyset^\alpha = \lim_{z\to0}\prod_{i=1}^{N}\left(1-\frac{zf_ig_\alpha}{1+zf_ig_\alpha}\right) = 1, \tag{40}$$

a relationship inducing $\langle N_\emptyset\rangle \xrightarrow{z\to0} L$: in words, the probability of observing empty hyperedges progressively rises as the density of 1s decreases.

For what concerns the filling threshold, a derivation that is similar-in-spirit to the one carried out for the case of the RHM can be sketched. Let us pose ourselves in the sparse regime: since $1 - p_{i\alpha} \simeq e^{-p_{i\alpha}}$, the probability for the generic hyperedge to be empty satisfies the chain of relationships

$$p_\emptyset^\alpha = \prod_{i=1}^{N}(1-p_{i\alpha}) \simeq \prod_{i=1}^{N}e^{-p_{i\alpha}} = e^{-\sum_{i=1}^{N}p_{i\alpha}} = e^{-h_\alpha}; \tag{41}$$

consistently, the expected number of empty hyperedges becomes

$$\langle N_\emptyset\rangle = \sum_{\alpha=1}^{L}p_\emptyset^\alpha \simeq \sum_{\alpha=1}^{L}e^{-h_\alpha} \tag{42}$$

and imposing $\langle N_\emptyset\rangle \leq 1$, i.e., that the expected number of empty hyperedges is *at most* 1, one derives a global condition to be satisfied by the hyperdegrees. In general terms, the aforementioned condition leads to require $e^{-h_\alpha} = O(1/L)$, i.e., $h_\alpha = O(\ln L)$, $\forall \alpha$ and $p_{i\alpha} = O(\ln L/N)$, $\forall i, \alpha$.

A comparison with simple graphs: estimating the number of isolated nodes. Coming to traditional graphs, the aim is now to estimate the number of isolated nodes. In this case, $1 - p_{ij} \simeq e^{-p_{ij}}$ and the probability for the generic node i to be isolated reads

$$q_0^i = \prod_{\substack{j=1\\j\neq i}}^{N}(1-p_{ij}) \simeq \prod_{\substack{j=1\\j\neq i}}^{N}e^{-p_{ij}} = e^{-\sum_{j(\neq i)=1}^{N}p_{ij}} = e^{-k_i}; \tag{43}$$

consistently, the expected number of isolated nodes becomes

$$\langle N_0 \rangle = \sum_{i=1}^{N} q_0^i \simeq \sum_{i=1}^{N} e^{-k_i} \qquad (44)$$

and imposing $\langle N_0 \rangle \lesssim 1$, i.e., that the expected number of isolated nodes is *at most* 1, one derives a global condition to be satisfied by the degrees. In general terms, the aforementioned condition leads to require $e^{-k_i} = O(1/N)$, i.e., $k_i = O(\ln N)$, $\forall\ i$ and $p_{ij} = O(\ln N/N)$, $\forall\ i < j$.

Estimation of the number of parallel hyperedges. As in the case of the RHM, we consider the Hamming distance between the columns representing the two hyperedges α and β. Since, now, $d_{\alpha\beta} \sim$ PoissBin (N, $\{q_i^{\alpha\beta}\}_{i=1}^N$), where $q_i^{\alpha\beta} \equiv p_{i\alpha}(1 - p_{i\beta}) + p_{i\beta}(1 - p_{i\alpha})$ with $p_{i\alpha} = zfg_\alpha/(1 + zfg_\alpha)$ and $p_{i\beta} = zfg_\beta/(1 + zfg_\beta)$, one finds that

$$P(d_{\alpha\beta} = 0) = \prod_{i=1}^{N}(1 - q_i^{\alpha\beta}) \qquad (45)$$

and that

$$\langle d_{\alpha\beta} \rangle = \sum_{i=1}^{N} q_i^{\alpha\beta} \qquad (46)$$

$\forall\ \alpha \neq \beta$. Since

$$\lim_{z \to +\infty} q_i^{\alpha\beta} = \lim_{z \to 0} q_i^{\alpha\beta} = 0, \qquad (47)$$

i.e., $q_i^{\alpha\beta}$ vanishes in both regimes, one finds that both $P(d_{\alpha\beta} = 0) \overset{z \to +\infty}{\to} 1$ and $P(d_{\alpha\beta} = 0) \overset{z \to 0}{\to} 1$ and that both $\langle d_{\alpha\beta} \rangle \overset{z \to +\infty}{\to} 0$ and $\langle d_{\alpha\beta} \rangle \overset{z \to 0}{\to} 0$: as in the case of the RHM, the probability of observing parallel hyperedges progressively rises both as a consequence of having many 1s and as a consequence of having few 1s. Analogously, for the expected Hamming distance.

For what concerns the resolution threshold, a derivation that is similar-in-spirit to the one carried out for the case of the RHM can be sketched. Let us pose ourselves in the sparse regime and consider that $1 - q_i^{\alpha\beta} \simeq e^{-q_i^{\alpha\beta}} \simeq e^{-(p_{i\alpha} + p_{i\beta})}$. As a consequence

$$p_{//}^{\alpha\beta} \equiv P(d_{\alpha\beta} = 0) = \prod_{i=1}^{N}(1 - q_i^{\alpha\beta}) \simeq \prod_{i=1}^{N} e^{-q_i^{\alpha\beta}} \simeq e^{-\sum_{i=1}^{N}(p_{i\alpha} + p_{i\beta})} = e^{-(h_\alpha + h_\beta)} \qquad (48)$$

and

$$\langle d_{\alpha\beta} \rangle = \sum_{i=1}^{N} q_i^{\alpha\beta} \simeq \sum_{i=1}^{N}(p_{i\alpha} + p_{i\beta}) = h_\alpha + h_\beta. \qquad (49)$$

Imposing $\langle d_{\alpha\beta} \rangle \gtrsim 1$, i.e., that the expected Hamming distance between any two hyperedges α and β is *at least* 1, amounts to require that $P(d_{\alpha\beta} = 0) \lesssim e^{-1}$ - thus recovering the same condition holding true in the case of the RHM where, in fact, $p_{//} \equiv p_{//}^{\alpha\beta} \overset{N \to +\infty}{\to} e^{-2h}$.

The expected number of parallel hyperedges, now, reads

$$\langle N_{//} \rangle = \sum_{\alpha=1}^{L} \sum_{\substack{\beta=1 \\ \beta > \alpha}}^{L} p_{//}^{\alpha\beta} \simeq \sum_{\alpha=1}^{L} \sum_{\substack{\beta=1 \\ \beta > \alpha}}^{L} e^{-(h_\alpha + h_\beta)}; \qquad (50)$$

upon imposing $\langle N_{//} \rangle \lesssim 1$, i.e., that the expected number of parallel hyperedges is *at most* 1, one derives a global condition to be satisfied by the

hyperdegrees. In general terms, the aforementioned condition leads to require $e^{-(h_\alpha + h_\beta)} = O(1/L^2)$, i.e., $h_\alpha = O(\ln L)$, $\forall\ \alpha$ and $p_{i\alpha} = O(\ln L/N)$, $\forall\ i, \alpha$.

Estimation of the percolation threshold. For what concerns the percolation threshold, the expected value of the total number of nodes shared by hyperedge α with any other hyperedge, now, reads

$$\langle \sigma_\alpha \rangle = \sum_{i=1}^{N} \sum_{\substack{\beta=1 \\ \beta \neq \alpha}}^{L} p_{i\alpha}p_{i\beta} \qquad (51)$$

and imposing $\langle \sigma_\alpha \rangle = 1$ leads to a global condition to be satisfied. In general terms, the aforementioned condition leads to require $p_{i\alpha}p_{i\beta} = O(1/NL)$, i.e., $p_{i\alpha} = O(1/\sqrt{NL})$, $\forall\ i, \alpha$.

Results
Hypergraphs in the dense and sparse regime
Let us start by verifying the correctness of the estimations of the filling, multiple resolution and percolation thresholds provided by our benchmarks: to this aim, we have considered the values $N = 300$ and $L = 1000$.

The RHM. Each quantity has been plotted as a function of $p \in [10^{-6}, 1]$. The dense (sparse) regime is recovered for large (small) values of p. Each dot of Figs. 2–4 represents an average taken over an ensemble of 10^3 configurations explicitly sampled from the RHM and is accompanied by the corresponding 95% confidence interval, calculated via the bootstrap method[41].

The filling threshold. Figure 2a depicts the (analytical) trend of $\langle N_\emptyset \rangle / L = p_\emptyset = (1 - p)^N$ (solid line): its agreement with the numerical estimations (dots) confirms the correctness of our formula.

Although the value of the filling threshold has been determined by inspecting the asymptotic behaviour of $\langle N_\emptyset \rangle$, the quantity showing the neatest transition from the sparse to the dense regime is the probability of observing at least one empty hyperedge

$$P(N_\emptyset > 0) = 1 - P(N_\emptyset = 0) = 1 - (1 - p_\emptyset)^L = 1 - [1 - (1 - p)^N]^L, \qquad (52)$$

where we have exploited the fact that $P(N_\emptyset > 0)$ is nothing but the complementary of the probability that no hyperedge is empty. Since $p_\emptyset \overset{N \to +\infty}{\to} e^{-h}$, evaluating such an expression in correspondence of $h_f^{RHM} = \ln L \simeq 6.907$ returns the value $1/L = 10^{-3}$ (see Fig. 2a). As a consequence, the result $P(N_\emptyset > 0) \overset{N \to +\infty}{\to} 1 - (1 - e^{-h})^L$ is numerically recovered (see Fig. 2b). In words, although the filling threshold ensures that each, single hyperedge is empty with an overall small probability, the likelihood of observing at least one, empty hyperedge is still large (i.e., $\simeq 2/3$): the steepness of the trend of $P(N_\emptyset > 0)$, however, suggests it to quickly vanish as the density of 1s in the incidence matrix crosses the value $p_f^{RHM} = h_f^{RHM}/N = \ln L/N \simeq 0.023$.

The multiple resolution threshold. Figure 3a depicts the trend of $2\langle N_{//} \rangle / L(L - 1) = [1 - 2p(1 - p)]^N = p_{//}$ (solid line): again, its agreement with the numerical estimations (dots) confirms the correctness of our formula.

Evaluating $p_{//} \overset{N \to +\infty}{\to} e^{-2h}$ in correspondence of $h_m^{RHM} \lesssim \ln L \simeq 6.907$ returns the value $1/L^2 = 10^{-6}$ (see Fig. 3a): in words, the aforementioned critical value causes the likelihood of observing any two parallel hyperedges to be almost the square of the probability for each single hyperedge to be empty (i.e., $1/L$, see the comments under Eq. (17)).

As the overlaps between pairs of hyperedges cannot be treated as i.i.d. random variables, evaluating the probability of observing at least one pair of parallel hyperedges forces us to proceed in a purely numerical fashion.

Fig. 2 | Impact of empty hyperedges on the RHM ensemble. More in detail, the trends of $\langle N_\emptyset \rangle / L = p_\emptyset = (1-p)^N \overset{N \to +\infty}{\to} e^{-h}$, i.e., the probability for the generic hyperedge to be empty is represented in (**a**) and the one of $P(N_\emptyset > 0) = 1 - [1 - (1-p)^N]^L \overset{N \to +\infty}{\to} 1 - (1 - e^{-h})^L$, i.e., the probability of observing at least one empty hyperedge is represented in (**b**). Evaluating them in correspondence of $p_f^{\text{RHM}} = h_f^{\text{RHM}} / N = \ln L / N \simeq 0.023$ (vertical line) returns, respectively, the values $1/L = 10^{-3}$ and $1 - (1-1/L)^L \simeq 0.6323$. The dense (sparse) regime is recovered for large (small) values of p. Each dot represents an average taken over an ensemble of 10^3 configurations (explicitly sampled from the RHM) and is accompanied by the corresponding 95% confidence interval.
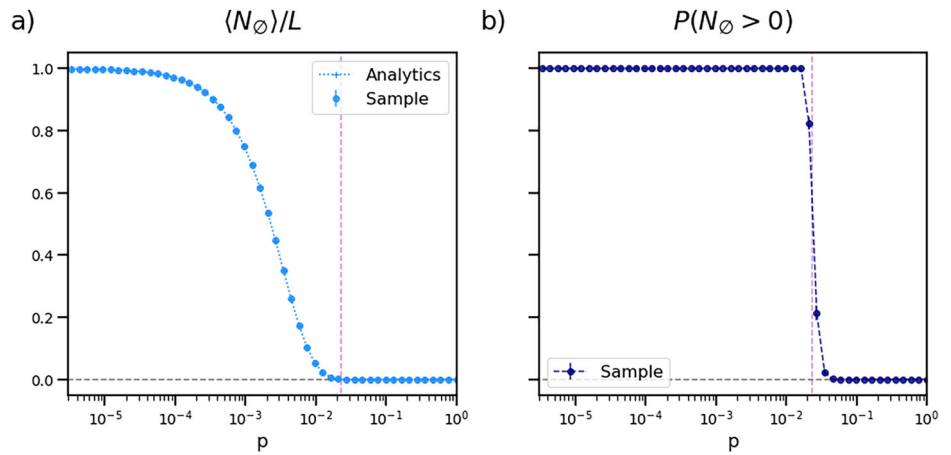
Fig. 3 | Impact of parallel hyperedges on the RHM ensemble. More in detail, the trends of $2\langle N_\parallel \rangle / L(L-1) = [1 - 2p(1-p)]^N = p_\parallel \overset{N \to +\infty}{\to} e^{-2h}$, i.e., the probability for the generic pair of hyperedges to be parallel and $P(N_\parallel > 0)$, i.e., the probability of observing at least one pair of parallel hyperedges, as functions of p, are represented, respectively, in (**a** and **b**). Evaluating them in correspondence of $p_m^{\text{RHM}} = h_m^{\text{RHM}} / N \lesssim \ln L / N \simeq 0.023$ (vertical line) returns, respectively, the values $1/L^2 = 10^{-6}$ and $\simeq 0.6$. The dense (sparse) regime is recovered for large (small) values of p. Each dot represents an average taken over an ensemble of 10^3 configurations (explicitly sampled from the RHM) and is accompanied by the corresponding 95% confidence interval.
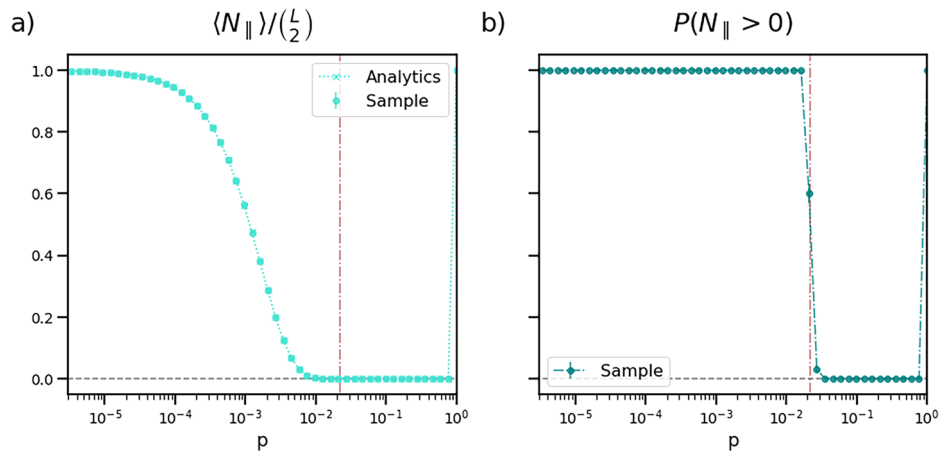

Calculating $P(N_\parallel > 0)$ in correspondence of $p_m^{\text{RHM}} = h_m^{\text{RHM}} / N \lesssim \ln L / N \simeq 0.023$ returns the value 0.6 (see Fig. 3b).

Carrying out such an estimation in the aforementioned regime is, however, instructive as it leads to the expression $P(N_{//} > 0) = 1 - (1 - p_{//})^{L(L-1)/2} = 1 - \{1 - [1 - 2p\,(1-p)]^N\}^{L(L-1)/2} \overset{N \to +\infty}{\to} 1 - [1 - e^{-2h}]^{L(L-1)/2}$ that, evaluated in correspondence of $h_m^{\text{RHM}} \lesssim \ln L \simeq 6.907$, returns the value $P(N_{//} > 0) = 1 - (1 - 1/L^2)^{L(L-1)/2} \simeq 0.393$ - whose difference with 0.6, obtained numerically, lets us fully appreciate the role played by correlations.

The percolation threshold. Let us, now, focus on the projection of our hypergraph onto the layer of hyperedges. The generic hyperedge is isolated either because is not 'connected" with any node or because is a singleton (i.e., it is 'connected" with a node with which no other hyperedge is 'connected"): in symbols,

$$p_0 \equiv \{1 - p[1 - (1-p)^{L-1}]\}^N = [(1-p) + p(1-p)^{L-1}]^N. \quad (53)$$

In order to evaluate p_0 in correspondence of the percolation threshold, let us consider that $L = s^2 N$, with $s^2 = 10/3$; one, then, finds

$$\lim_{N \to +\infty} p_0 = \lim_{N \to +\infty} \left[\left(1 - \frac{1}{Ns}\right) + \frac{1}{Ns}\left(1 - \frac{1}{Ns}\right)^{s^2N-1} \right]^N = e^{-\frac{1-e^{-s}}{s}}, \quad (54)$$

whose numerical value amounts to 0.631 (see Fig. 4a): in words, the value $p_0(p = p_p^{\text{RHM}}) \lesssim 2/3$ implies that the expected number of isolated hyperedges in the projection $\langle N_0 \rangle = L p_0$ tends to $2L/3$ as p tends to p_p^{RHM}.

Pairs of hyperedges cannot be treated as independent. Let us, in fact, consider the bipartite representation of a hypergraph: as adjacent pairs of hyperedges (say α, β and β, γ) may share some neighbours (on the opposite layer), the number of common neighbours of α and β will, in general, covariate with the number of common neighbours of β and γ; therefore, evaluating the probability of observing at least one, isolated hyperedge in the projection forces us to proceed in a purely numerical fashion. Calculating $P(N_0 > 0)$ in correspondence of $p_p^{\text{RHM}} = h_p^{\text{RHM}} / N = 1/\sqrt{NL} \simeq 0.002$ practically returns 1 (see Fig. 4b). From the perspective of a hypergraph connectedness, the percolation threshold is 'less strict" than the filling threshold, allowing for a larger number of disconnected nodes (2/3 of the total versus 1).

As before, estimating the percolation threshold in the regime where the pairs of hyperedges behave as i.i.d. Binomial random variables is instructive. In this case, projecting a bipartite network onto the layer of hyperedges amounts to connect any two of them with probability $1 - (1 - p^2)^N$—i.e., the complementary of the probability $(1 - p^2)^N$ of not sharing any node. The number of isolated nodes, thus, obeys the relationship $N_0 \sim \text{Bin}(L, p_0)$, with

$$p_0 \equiv [(1 - p^2)^N]^L \quad (55)$$

Fig. 4 | Impact of isolated hyperedges on the RHM ensemble. More in detail, trends of $\langle N_0 \rangle/L = p_0$, i.e., the probability for the generic hyperedge to be isolated in the projection, $P(N_0 > 0)$, i.e., the probability of observing at least one, isolated hyperedge in the projection and $|LCC|/N$, i.e., the percentage of nodes belonging to the largest connected component (LCC), are represented as functions of p, respectively in (**a–c**). Evaluating the first two in correspondence of $p_p^{\text{RHM}} = h_p^{\text{RHM}}/N = 1/\sqrt{NL} \simeq 0.002$ (vertical line) returns, respectively, the values $e^{(e^{-s}-1)/s} \simeq 0.631$ and $\simeq 1$. The dense (sparse) regime is recovered for large (small) values of p. Each dot represents an average taken over an ensemble of 10^3 configurations (explicitly sampled from the RHM) and is accompanied by the corresponding 95% confidence interval.
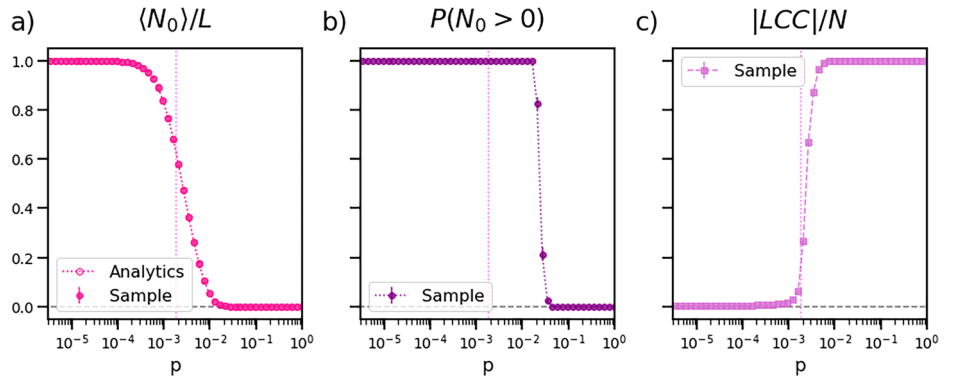

Fig. 5 | Impact of empty hyperedges on the HCM ensemble. More in detail, trends of $\langle N_\emptyset \rangle/L = \overline{p_\emptyset}$, i.e., the probability for the generic hyperedge to be empty and $P(N_\emptyset > 0) \simeq 1 - \prod_{\alpha=1}^{L}(1 - e^{-h_\alpha})$, i.e., the probability of observing at least one empty hyperedge, are represented as functions of the connectance ρ, respectively in (**a** and **b**). Evaluating the latter in correspondence of the filling threshold, reading $p_f^{\text{HCM}} \simeq 0.032$ (vertical line), returns the value $\simeq 0.642$. The dense (sparse) regime is recovered for large (small) values of z. Each dot represents an average taken over an ensemble of 10^3 configurations (explicitly sampled from the HCM) and is accompanied by the corresponding 95% confidence interval.
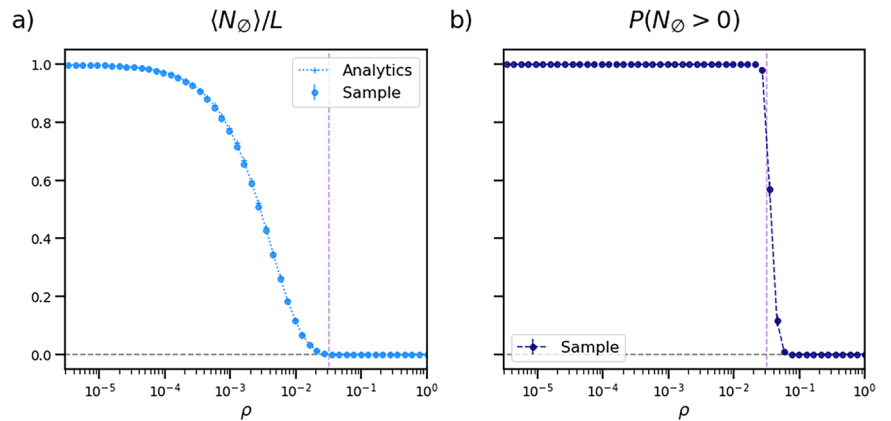
being the probability for the generic hyperedge to be isolated: in words, such an expression returns the probability for the generic hyperedge to not share *any* node - with probability $(1 - p^2)^N$ - with *any* other hyperedge—with a probability amounting to the previous one raised to the power of L. As a consequence, evaluating p_0 in correspondence with the percolation threshold returns a value tending to 1/3, a result further implying that the expected number of isolated hyperedges in the projection $\langle N_0 \rangle = L p_0$ tends to the value $L/3$ - both letting us fully appreciate the role played by correlations.

Within such a context, the probability of observing at least one isolated hyperedge in the projection satisfies the chain of relationships $P(N_0 > 0) = 1 - P(N_0 = 0) = 1 - (1 - p_0)^L = 1 - [1 - (1 - p^2)^{NL}]^L \overset{N \to +\infty}{\to} 1 - (1 - e^{-1})^L$: as the last expression quickly converges to 1 for large values of L, the same qualitative behaviour observed before is thus recovered.

Finally, one may wonder which kind of mesoscale structure is identified by the percolation threshold: the answer is provided by Fig. 4c, showing the appearance of a large connected component - as also pointed out in[12], the presence of a large connected component is inspected by projecting our hypergraph onto the layer of *nodes*, whose connectedness is ensured by requiring the connectedness of hyperedges.

The role of thresholds in a hypergraph evolution. Let us, now, make a couple of observations. The first one concerns the result according to which $h_m^{\text{RHM}} \lesssim h_f^{\text{RHM}}$: such a relationship suggests that, while filling a large portion of the incidence matrix is not required to observe a limited amount of parallel hyperedges (see Fig. 3b), when empty hyperedges are no longer observed, parallel hyperedges are no longer observed as well.

The second one concerns the result according to which either $h_f^{\text{RHM}} \leq h_p^{\text{RHM}}$ or $h_f^{\text{RHM}} \geq h_p^{\text{RHM}}$. By progressively rising the parameter p, two, different thresholds are, thus, met: if $h_f^{\text{RHM}} \leq h_p^{\text{RHM}}$, the filling threshold $p_f^{\text{RHM}} = \ln L/N$ is met before the percolation threshold $p_p^{\text{RHM}} = 1/\sqrt{NL}$, i.e., hyperedges are filled before they start sharing nodes—as a consequence, singletons appear; if $h_f^{\text{RHM}} \geq h_p^{\text{RHM}}$, the percolation threshold $p_p^{\text{RHM}} = 1/\sqrt{NL}$ is met before the filling threshold $p_f^{\text{RHM}} = \ln L/N$, i.e., hyperedges start sharing nodes before they are filled—as a consequence, no singleton appears before the filling threshold is crossed.

Notice that, for simple graphs, $k_p^{\text{RGM}} = 1 \leq k_c^{\text{RGM}} = \ln N$, i.e., by progressively rising the parameter q, the percolation threshold is always met before the connectivity threshold.

The HCM. In order to carry out the numerical simulations in the case of the HCM, we have followed the procedure described in the previous sections and drawn both the fitnesses of nodes and those of hyperedges from a Pareto distribution with $\alpha = 2$ - other fat-tailed distributions were considered: qualitatively, results do not change. Each quantity has been plotted as a function of $\rho(z) = \langle T \rangle/NL \in [10^{-6}, 1]$ where $\langle T \rangle = \sum_{i=1}^{N} \sum_{\alpha=1}^{L} z f_i g_\alpha/(1 + z f_i g_\alpha)$ varies with z. The dense (sparse) regime is recovered for large (small) values of z. Each dot of Figs. 5–7 represents an average taken over an ensemble of 10^3 configurations explicitly sampled from the HCM and is accompanied by the corresponding 95% confidence interval.

The filling threshold. Figure 5a depicts the (analytical) trend of $\langle N_\emptyset \rangle/L = \sum_{\alpha=1}^{L} p_\emptyset^\alpha/L = \overline{p_\emptyset}$ (solid line): as in the case of the RHM, its agreement with the numerical estimations (dots) confirms the correctness of our formula. Deriving an explicit expression for the filling

Fig. 6 | Impact of parallel hyperedges on the HCM ensemble. More in detail, trends of $2\langle N_{\parallel}\rangle/L(L-1) = \overline{p_{\parallel}}$, i.e., the probability for the generic pair of hyperedges to be parallel and $P(N_{\parallel} > 0)$, i.e., the probability of observing at least one pair of parallel hyperedges, are represented as functions of the connectance ρ, respectively, in (**a** and **b**). Evaluating the latter in correspondence with the multiple resolution threshold, reading $p_m^{HCM} \simeq 0.031$ (vertical line), returns 0.493. The dense (sparse) regime is recovered for large (small) values of z. Each dot represents an average taken over an ensemble of 10^3 configurations (explicitly sampled from the HCM) and is accompanied by the corresponding 95% confidence interval.
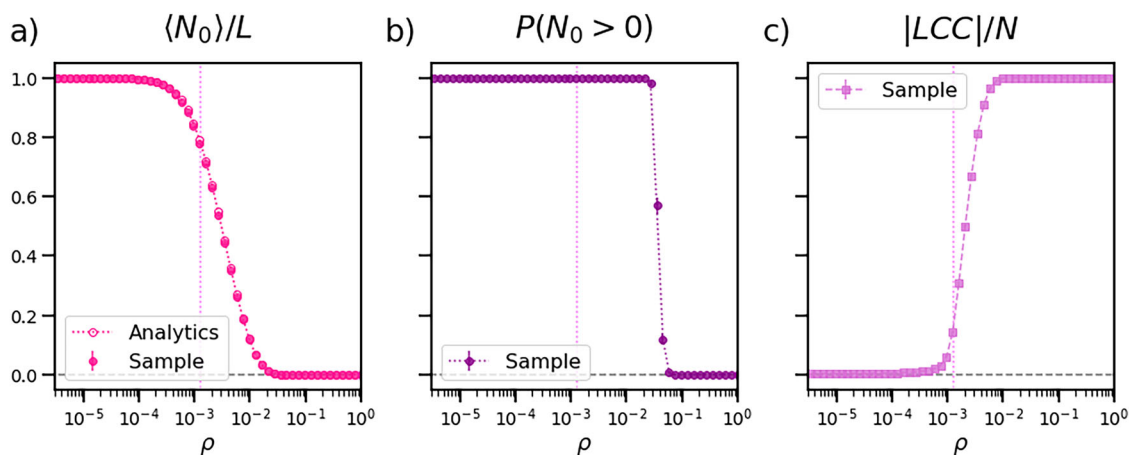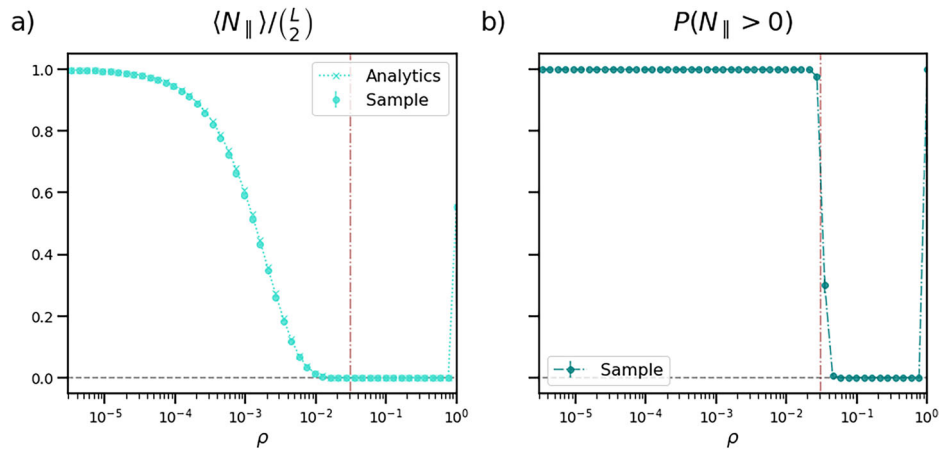
Fig. 7 | Impact of isolated hyperedges on the HCM ensemble. More in detail, trends of $\langle N_0 \rangle/L = \overline{p_0}$, i.e., the probability for the generic hyperedge to be isolated in the projection, $P(N_0 > 0)$, i.e., the probability of observing at least one, isolated hyperedge in the projection, and $|LCC|/N$, i.e., the percentage of nodes belonging to the largest connected component (LCC), are repesented as functions of the connectance ρ, respectively, in (**a**–**c**). Evaluating the first two in correspondence of $p_p^{HCM} \simeq 0.001$ (vertical line), returns the values $\simeq 0.766$ and $\simeq 1$. The dense (sparse) regime is recovered for large (small) values of p. Each dot represents an average taken over an ensemble of 10^3 configurations (explicitly sampled from the HCM) and is accompanied by the corresponding 95% confidence interval.

threshold in the case of the HCM is a rather difficult task; still, we can proceed in a purely numerical fashion and individuate the value of the density of 1s in the incidence matrix guaranteeing that the expected number of empty hyperedges (divided by L) amounts to 1 (divided by L): although its, precise, numerical value depends on the values of the fitnesses, such a threshold still lies in the right tail of the trend induced by the HCM and reads $p_f^{HCM} \simeq 0.032$ (see Fig. 5a).

Even in this case, the quantity showing the neatest transition from the sparse to the dense regime is the probability of observing at least one empty hyperedge

$$P(N_{\emptyset} > 0) = 1 - P(N_{\emptyset} = 0) = 1 - \prod_{\alpha=1}^{L}(1 - p_{\emptyset}^{\alpha})$$
$$= 1 - \prod_{\alpha=1}^{L}\left[1 - \prod_{i=1}^{N}(1 - p_{i\alpha})\right] \quad (56)$$

that, in the sparse regime, can be approximated as $P(N_{\emptyset} > 0) \simeq 1 - \prod_{\alpha=1}^{L}(1 - e^{-h_{\alpha}})$: as Fig. 5b shows, evaluating $P(N_{\emptyset} > 0)$ in correspondence of the filling threshold returns 0.642. Finally, let us

explicitly notice that the value of the filling threshold is shifted on the right with respect to its homogeneous counterpart, an evidence probably due to the presence of small fitnesses that increase the probability of observing at least one, empty hyperedge, hence requiring a larger value of z to let $P(N_{\emptyset} > 0)$ vanish.

The multiple resolution threshold. Figure 6a depicts the (analytical) trend of $2\langle N_{//}\rangle/L(L-1) = 2\sum_{\alpha=1}^{L}\sum_{\substack{\beta=1 \\ \beta > \alpha}}^{L} p_{//}^{\alpha\beta}/L(L-1) = \overline{p_{//}}$ (solid line): as in the case of the RHM, its agreement with the numerical estimations (dots) confirms the correctness of our formula.

Adopting the strategy described in the previous paragraph, i.e., that of requiring that the expected number of parallel hyperedges (divided by $L(L-1)/2$) amounts to 1 (divided by $L(L-1)/2$), we found that $p_m^{HCM} \simeq 0.031$ (see Fig. 6a). Numerically clculating $P(N_{\parallel} > 0)$ in correspondence of the multiple resolution threshold returns the value 0.493 (see Fig. 6b).

Let us notice that the value of the multiple resolution threshold no longer coincides with the value of the filling threshold, although it is still shifted to the right with respect to its homogeneous counterpart.

The percolation threshold. The probability for the generic hyperedge to be isolated in the projection, now, reads

$$
p_0^\alpha \equiv \prod_{i=1}^{N} \left\{ 1 - p_{i\alpha} \left[1 - \prod_{\substack{\beta=1 \\ \beta \neq \alpha}}^{L} (1 - p_{i\beta}) \right] \right\}
$$
$$
= \prod_{i=1}^{N} \left\{ (1 - p_{i\alpha}) + p_{i\alpha} \prod_{\substack{\beta=1 \\ \beta \neq \alpha}}^{L} (1 - p_{i\beta}) \right\} \tag{57}
$$

that, in the sparse regime, can be approximated as $p_0^\alpha \simeq \prod_{i=1}^{N} [(1 - p_{i\alpha}) + p_{i\alpha} e^{-k_i}]$. Adopting the strategy described in the previous paragraphs, i.e., that of requiring that the expected value of the total number of nodes shared by any hyperedge with any other hyperedge amounts to 1 - in symbols, $\overline{\langle \sigma \rangle} = \sum_{\alpha=1}^{L} \langle \sigma_\alpha \rangle / L = 1$ - we found that $p_p^{\text{HCM}} \simeq 0.001$: since $\langle N_0 \rangle = \sum_{\alpha=1}^{L} p_0^\alpha$, evaluating $\langle N_0 \rangle / L = \sum_{\alpha=1}^{L} p_0^\alpha / L = \overline{p_0}$ and $P(N_0 > 0)$ in correspondence of the percolation threshold returns respectively the values 0.766 (see Fig. 7a) and $\simeq 1$ (see Fig. 7b). For what concerns the hypergraph connectedness, the same conclusion drawn in the case of the RHM holds true, as the percolation threshold allows for a larger number of disconnected nodes (3/4 of the total versus 1).

Analogously, the mesoscale structure individuated by the percolation threshold consists of a large connected component constituted (see Fig. 7c).

Solving the HCM on real-world hypergraphs

In order to test our benchmarks on real-world configurations, we have focused on a number of data sets taken from Austin R. Benson's website (https://www.cs.cornell.edu/~arb/data/), i.e., the contact-primary-school, the email-Enron and the NDC-classes ones.

Although the parameters defining the HCM must be numerically determined by solving the system of equations induced by the likelihood maximisation, when the system under analysis is sparse they can be approximated as described in Supplementary Note 1 of the Supplementary Information. Such an approximation leads to the expression

$$
p_{i\alpha} \simeq x_i y_\alpha = \frac{k_i^* h_\alpha^*}{T^*}, \, \forall \, i, \alpha \tag{58}
$$

that, as Supplementary Fig. 1 in the Supplementary Information shows, is quite accurate for each data set considered here—in fact, one can safely assume that $x_i \simeq k_i^*/\sqrt{T^*}, \, \forall \, i$ and $h_\alpha^*/\sqrt{T^*}, \, \forall \, \alpha$.

"Hypergraph to graph" projection

The canonical formalism that we have adopted leads to factorisable distributions, i.e., distributions that can be written as a product of pair-wise probability distributions; this allows the expectation of several quantities of interest to be evaluated analytically.

Let us start by considering the matrix, introduced in[4], reading

$$
\mathbf{W} = \mathbf{I} \cdot \mathbf{I}^T - \mathbf{K} \tag{59}
$$

with $\mathbf{K}$ being the diagonal matrix whose i-th entry reads k_i; according to the definition above, it induces a projection of a hypergraph onto a weighted graph, whose generic entry

$$
w_{ij} = \sum_{\alpha=1}^{L} I_{i\alpha} I_{j\alpha} - \delta_{ij} k_i \tag{60}
$$

returns the number of hyperedges both i and j belong to - more explicitly, $w_{ij} = \sum_{\alpha=1}^{L} I_{i\alpha} I_{j\alpha}, i \neq j$ and $w_{ii} = \sum_{\alpha=1}^{L} I_{i\alpha} I_{i\alpha} - k_i = \sum_{\alpha=1}^{L} I_{i\alpha} - k_i = k_i - k_i = 0$. In other words, $\mathbf{W}$ represents the object most closely resembling a traditional adjacency matrix. The null models discussed so far can be employed to calculate $\langle w_{ij} \rangle, i \neq j$ that, in a perfectly general fashion, reads

$$
\langle w_{ij} \rangle = \sum_{\alpha=1}^{L} \langle I_{i\alpha} I_{j\alpha} \rangle = \sum_{\alpha=1}^{L} \langle I_{i\alpha} \rangle \langle I_{j\alpha} \rangle = \sum_{\alpha=1}^{L} p_{i\alpha} p_{j\alpha}; \tag{61}
$$

as we said, the total number of hyperedges shared by node i with any other node in the hypergraph (in a sense, its 'strength" - see also Fig. 8) can be computed as

$$
\sigma_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{ij} = \sum_{\substack{j=1 \\ j \neq i}}^{N} \sum_{\alpha=1}^{L} I_{i\alpha} I_{j\alpha} \tag{62}
$$

whose expected value reads

$$
\langle \sigma_i \rangle = \sum_{\substack{j=1 \\ j \neq i}}^{N} \langle w_{ij} \rangle = \sum_{\alpha=1}^{L} p_{i\alpha} [\langle h_\alpha \rangle - p_{i\alpha}]. \tag{63}
$$

As further confirmed by Supplementary Fig. 1 in the Supplementary Information, the approximation provided by Eq. (58) allows us to pose $\langle \sigma_i \rangle_{\text{HCM}} \simeq k_i^* \sum_{\alpha=1}^{L} (h_\alpha^*)^2 / T^*, \, \forall \, i$. Interestingly, as Fig. 9a shows, the HCM overestimates the extent to which any two nodes of the email-Enron
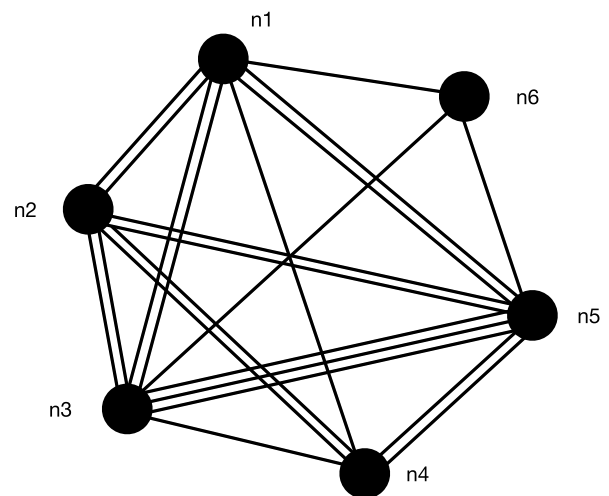

Fig. 8 | Graphical representation of the 'hypergraph to graph' projection. Notice that while the degree of each node (i.e., the number of hyperedges that are incident to it) reads $\{k_i\} = (3, 3, 3, 3, 4, 1)$, the sigma of each node (i.e., the total number of hyperedges shared by it with any other node or, equivalently, the strength on the projection) reads $\{\sigma_i\} = (8, 9, 6, 10, 3)$ and the kappa of each node (i.e., the number of nodes it shares an hyperedge with or, equivalently, the degree on the projection) reads $\{\kappa_i\} = (5, 4, 5, 4, 5, 3)$.

Fig. 9 | Scatter plots between the empirical and the expected values for the `email-Enron` **dataset.** More in detail, $\{\sigma_i\}$ vs. $\{\langle\sigma_i\rangle_{\text{HCM}}\}$ (**a**), $\{\kappa_i\}$ vs. $\{\langle\kappa_i\rangle_{\text{HCM}}\}$ (**b**), $\{Y_i\}$ vs. $\langle\langle Y_i\rangle_{\text{HCM}}\rangle$ (**c**), and $\{\text{CEC}_i\}$ vs. $\{\langle\text{CEC}_i\rangle_{\text{HCM}}\}$ (**d**). The HCM overestimates the extent to which any two nodes overlap, as well as the CEC; the disparity ratio, instead, is underestimated by it. These results can be understood by considering that the HCM just constrains the degree sequences, hence inducing an ensemble where connections are 'distributed" more evenly than observed.
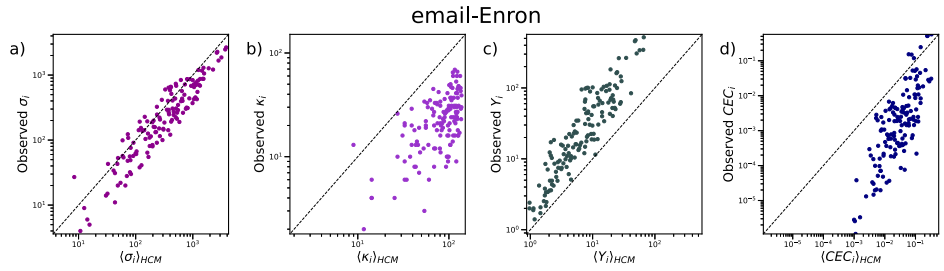
data set overlap: in words, such a real-world hypergraph is more compartmentalised than expected.

Let us, now, extend the concept of *assortativity* to hypergraphs. To this aim, we consider the quantity named *average incident hyperedges degree*, defined as

$$k_i^{nn} = \sum_{\alpha=1}^{L} \frac{I_{i\alpha}h_\alpha}{k_i} = \frac{\sigma_i + k_i}{k_i} = \frac{\sigma_i}{k_i} + 1 \simeq \frac{\sigma_i}{k_i} \quad (64)$$

and representing the arithmetic mean of the degrees of the hyperedges including node i. An analytical approximation of its expected value can be provided as well:

$$\langle k_i^{nn}\rangle \simeq \sum_{\alpha=1}^{L} \frac{p_{i\alpha}[\langle h_\alpha\rangle + 1 - p_{i\alpha}]}{\langle k_i\rangle} = \frac{\langle\sigma_i\rangle + \langle k_i\rangle}{\langle k_i\rangle} = \frac{\langle\sigma_i\rangle}{\langle k_i\rangle} + 1 \simeq \frac{\langle\sigma_i\rangle}{\langle k_i\rangle}. \quad (65)$$

Disparity ratio and degree in the projection

More information about the patterns shaping real-world hypergraphs can be obtained upon defining the ratio $f_{ij} = w_{ij}/\sigma_i$, $i \neq j$ that induces the quantity

$$Y_i = \sum_{\substack{j=1 \\ j\neq i}}^{N} f_{ij}^2 = \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{w_{ij}^2}{\sigma_i^2}, \quad (66)$$

known as *disparity ratio* and quantifying the (un)evenness of the distribution of the weights constituting the strength of node i over the $\kappa_i = \sum_{\substack{j=1 \\ (j\neq i)}}^{N} \Theta[w_{ij}] \equiv \sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij}$ links characterising its connectivity - since $a_{ij} = 1$ if nodes i and j share, at least, one hyperedge, κ_i is the degree of node i in the projection of the hypergraph (see also Figs. 8 and 9b). Since, under the RHM, $w_{ij} \sim \text{Bin}(L, p^2)$, we find that $\langle a_{ij}\rangle = 1 - (1 - p^2)^L$, i.e., the expected value of a_{ij} coincides with the probability of observing a non-zero overlap. Under the HCM, instead, $w_{ij} \sim \text{PoissBin}(L, \{p_{i\alpha}p_{j\alpha}\}_{\alpha=1}^{L})$, hence

$$\langle a_{ij}\rangle = 1 - \prod_{\alpha=1}^{L}(1 - p_{i\alpha}p_{j\alpha}). \quad (67)$$

Let us also notice that

$$Y_i = \frac{1}{\kappa_i} \quad (68)$$

in case weights are equally distributed among the connections established by node i, i.e., $w_{ij} = a_{ij}\sigma_i/\kappa_i$, $i \neq j$. Any larger value signals an excess concentration of weight in one or more links. An analytical approximation of the expected

value of the disparity ratio of node i can be provided as well:

$$\langle Y_i\rangle \simeq \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{\langle w_{ij}^2\rangle}{\langle\sigma_i^2\rangle}. \quad (69)$$

Contrary to what has been previously observed, the expected value of the disparity ratio cannot always be safely decomposed as a ratio of expected values, not even if the 'full" HCM is employed. In fact, while this approximation works relatively well for the `contact-primary-school` data set, it does not for the `email-Enron` and the `NDC-classes` ones (see also Supplementary Fig. 3 in the Supplementary Information). For this reason, the expected value of the disparity ratio has been evaluated by explicitly sampling the ensemble of incidence matrices induced by the 'full' HCM. In any case, as Fig. 9c shows, such a null model underestimates the disparity ratio characterising each node of the `email-Enron` data set: in words, the empirical overlap between any two nodes is (much) less evenly 'distributed' than expected. A similar conclusion can be drawn by considering $\langle\kappa_i\rangle = \sum_{i=1}^{N}\left[1 - \prod_{\alpha=1}^{L}(1 - p_{i\alpha}p_{j\alpha})\right]$: as Fig. 9b shows, the degree of the nodes in the projection tend to be significantly smaller than expected, meaning that hyperedges concentrate on fewer edges than expected. This observation is in line with recent works showing the encapsulation and 'simpliciality' of real-world hypergraphs[42,43].

Eigenvector centrality

Centrality measures for hypergraphs have been defined as well. An example is provided by the *clique motif eigenvector centrality* (CEC), defined in[44] (see also Supplementary Note 3 of the Supplementary Information): CEC_i corresponds to the i-th entry of the Perron-Frobenius eigenvector of $\mathbf{W}$. As Fig. 9d shows, the HCM underestimates the CEC as well: such a result can be understood by considering that the HCM constrains only the degree sequences, hence inducing an ensemble where connections are 'distributed" more evenly than observed, an evidence letting the nodes overlap more, thus causing the entries of $\langle\mathbf{W}\rangle$ to be overall larger and less dissimilar, as well as those of its Perron-Frobenius eigenvector.

Confusion matrix

Let us, now, consider the set of indices constituting the so-called *confusion matrix* (see also Supplementary Note 4 of the Supplementary Information). They are intended to quantify the capability of a given network model in reproducing microscopic properties, such as the position of 1s and 0s by explicitly comparing their empirical location with the one expected under the chosen model. They are named *true positive rate* (TPR), i.e., the percentage of 1s correctly recovered by a given method, whose expected value reads

$$\langle\text{TPR}\rangle = \sum_{i=1}^{N}\sum_{\alpha=1}^{L}\frac{I_{i\alpha}p_{i\alpha}}{T}; \quad (70)$$

Fig. 10 | Validated vs. non-validated 'hypergraph to network" projections of empirical datasets. a–c projections of the `contact-primary-school`, `email-Enron` and `NDC-classes` data sets onto the layer of nodes. **d–f** validated counterparts of the aforementioned projections: any two nodes are linked if they share a significantly large number of hyperedges. Communities have been detected by running the Louvain algorithm.
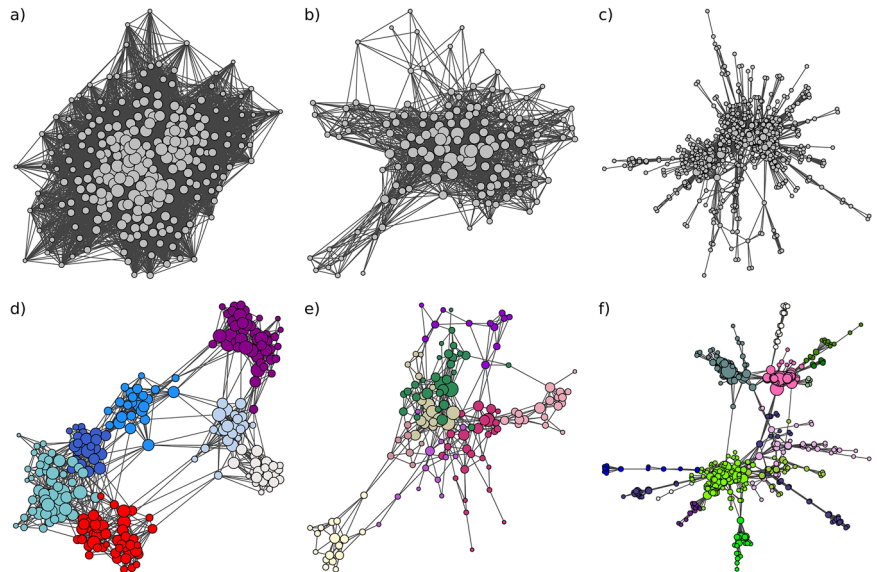
specificity (SPC), i.e., the percentage of 0s correctly recovered by a given method, whose expected value reads

$$\langle \text{SPC} \rangle = \sum_{i=1}^{N} \sum_{\alpha=1}^{L} \frac{(1 - I_{i\alpha})(1 - p_{i\alpha})}{NL - T}; \tag{71}$$

positive predictive value (PPV), i.e., the percentage of 1s correctly recovered by a given method with respect to the total number of 1s predicted by it, whose expected value reads

$$\langle \text{PPV} \rangle = \sum_{i=1}^{N} \sum_{\alpha=1}^{L} \frac{I_{i\alpha} p_{i\alpha}}{\langle T \rangle}; \tag{72}$$

accuracy (ACC), measuring the overall performance of a given method in correctly placing both 1s and 0s, whose expected value reads

$$\langle \text{ACC} \rangle = \frac{\langle \text{TP} \rangle + \langle \text{TN} \rangle}{NL}. \tag{73}$$

Results on the confusion matrix of a number of real-world hypergraphs reveal that the large sparsity of the latter ones makes it difficult to reproduce the TPR and the PPV (see also Supplementary Table 2 in the Supplementary Information); on the other hand, the capability of the HCM (both in its 'full" and approximated version) to reproduce the density of 1s—and, as a consequence, the density of 0s—ensures the SPC to be recovered quite precisely, in turn ensuring the overall ACC of the model to be large (for an overall evaluation of the performance of the HCM in reproducing real-world hypergraphs, see also Supplementary Table 3 in Supplementary Note 5 of the Supplementary Information).

Community detection

Communities are commonly understood as densely connected groups of nodes. Representing an hypergraph via its incidence matrix allows this statement to be made more precise from a statistical perspective: in fact, the null models discussed so far can be employed to test if any two nodes share a significantly large number of hyperedges - hence can be clustered together, should this be the case. In other words, it is possible to devise a 'validation procedure" that filters the projection described by the matrix **W** by removing the entries that do not satisfy the requirement above.

To this aim, we can adapt the recipe proposed in[24] to project bipartite networks and summarised in the following. One, first, computes

$$\text{p}-\text{value}\,(w_{ij}^{*}) = \sum_{x \geq w_{ij}^{*}} f(x) \tag{74}$$

for each pair of nodes; $f(x)$ depends on the chosen null model: in case the RHM is employed, it coincides with the Binomial distribution $\text{Bin}(x|L, p)$; in case the HCM is employed, it coincides with the Poisson-Binomial distribution $\text{PoissBin}\,(x|L, \{p_{i\alpha} p_{j\alpha}\}_{\alpha=1}^{L})$. Second, one implements the FDR procedure, designed to handle multiple tests of hypothesis[45]: in practice, after ranking the p-values in increasing order, i.e., $p\,\text{value}_1 \leq \text{p-value}_2 \leq \cdots \leq \text{p-value}_n$, one individuates the largest integer $\hat{i}$ satisfying the condition

$$\text{p}-\text{value}_{\hat{i}} \leq \frac{\hat{i} t}{n} \tag{75}$$

where $n = N(N-1)/2$ and t is the single-test significance level, set to 0.01 in the present analysis. Third, one links the (pairs of) nodes whose related p value is smaller than the aforementioned threshold.

Figure 10 shows the partitions returned by the Louvain algorithm run on the validated projections: as noticed elsewhere[24,46,47], the detection of mesoscale structures is enhanced if carried out on filtered topologies.

Conclusions

Our paper contributes to current research on hypergraphs by extending the constrained entropy-maximisation framework to incidence matrices, i.e., their simplest, tabular representation. Differently from the currently-available techniques[9], our methodology has the advantage of being analytically tractable, scalable and versatile enough to be straightforwardly extensible to directed and/or weighted hypergraphs.

Beside leading to results whose relevance is mostly theoretical (i.e., the individuation of different regimes for higher-order structures and the estimation of the actual impact of empty and parallel hyperedges on the analysis of empirical systems), our models prove to be particularly useful when employed as benchmarks for real-world systems, i.e., for detecting patterns that are not imputable to purely random effects. Specifically, our results suggest that real-world hypergraphs are characterised by a degree of self-organisation that is absolutely non-trivial (see also Supplementary Note 5 of the Supplementary Information).

This is even more surprising when considering that our results are obtained under a benchmark such as the HCM, i.e., a null model constraining both the degree and the hyperdegree sequences: since it

overestimates the extent to which any two nodes overlap—a result whose relevance becomes evident as soon as one considers the effects that higher-order structures have on spreading and cooperation processes[48–50]—our future efforts will be directed towards the analysis of benchmarks constraining non-linear quantities such as the co-occurrences between nodes and/or hyperedges.

Data availability
All the data used for the analysis are freely available at https://www.cs.cornell.edu/arb/data/.

Code availability
The authors will provide the code used for the analysis upon request.

References

1. Newman, M. *Networks: An Introduction*. (Oxford University Press, 2010).
2. Caldarelli, G. *Scale-Free Networks: Complex Webs in Nature and Technology*. (Oxford University Press, 2010)
3. Lambiotte, R., Rosvall, M. & Scholtes, I. From networks to optimal higher-order models of complex systems. *Nat. Phys.* **15**, 313–320 (2019).
4. Battiston, F. et al. Networks beyond pairwise interactions: structure and dynamics. *Phys. Rep.* **874**, 1–92 (2020).
5. Bick, C., Gross, E., Harrington, H. A. & Schaub, M. T. What are higher-order networks? *SIAM Rev.* **65**, 686–731 (2023).
6. Courtney, O.T., Bianconi, G. Generalized network structures: the configuration model and the canonical ensemble of simplicial complexes. *Phys. Rev. E* **93**, 062311 (2016).
7. Bianconi, G. Satistical physics of exchangeable sparse simple networks, multiplex networks and simplicial complexes. *Phys. Rev. E* **105**, 034310 (2022).
8. Battiston, F. et al. The physics of higher-order interactions in complex systems. *Nat. Phys.* **17**, 1093–1098 (2021).
9. Chodrow, P. S. Configuration models of random hypergraphs. *J. Complex Netw.* **8**, 3 (2020).
10. Nakajima, K., Shudo, K. & Masuda, N. Randomizing Hypergraphs Preserving Degree Correlation and Local Clustering, *IEEE Transactions on Network Science and Engineering*, Vol. 9, 1139–1153 (2022).
11. Musciotto, F., Battiston, F. & Mantegna, R. N. Detecting informative higher-order interactions in statistically validated hypergraphs. *Commun. Phys.* **4**, 218 (2021).
12. Barthelemy, M. Class of models for random hypergraphs. *Phys. Rev. E* **106**, 064310 (2022).
13. Dudek, A., Frieze, A., Ruciński, A. & Šileikis, M. Embedding the Erdös-Rényi hypergraph into the random regular hypergraph and hamiltonicity. *J. Combinatorial Theory, Ser. B* **122**, 719–740 (2017).
14. Boix-Adserà, E., Brennan, M., Bresler, G. The average-case complexity of counting cliques in Erdos-Renyi hypergraphs. *SIAM J. Comput.* 19–39 https://doi.org/10.1137/20M1316044 (2019).
15. Barrett, J., Prałat, P., Smith, A., Théberge, F. *Counting Simplicial Pairs In Hypergraphs* (Springer, 2024).
16. Wegner, A. E., Olhede, S. Atomic subgraphs and the statistical mechanics of networks. *Phys. Rev. E* **103**, 042311 (2021).
17. Jaynes, E. T. Information theory and statistical mechanics. *Phys. Rev.* **106**, 181–218 (1957).
18. Park, J. & Newman, M. E. J. Statistical mechanics of networks. *Phys. Rev. E* **70**, 66117 (2004).
19. Garlaschelli, D. & Loffredo, M. I. Maximum likelihood: extracting unbiased information from complex networks. *Phys. Rev. E* **78**, 1–5 (2008).
20. Squartini, T. & Garlaschelli, D. Analytical maximum-likelihood method to detect patterns in real networks. *N. J. Phys.* **13**, 083001 (2011).
21. Cimini, G. et al. The statistical physics of real-world networks. *Nat. Rev. Phys.* **1**, 58–71 (2018).
22. Squartini, T., Caldarelli, G., Cimini, G., Gabrielli, A. & Garlaschelli, D. Reconstruction methods for networks: the case of economic and financial systems. *Phys. Rep.* **757**, 1–47 (2018).
23. Squartini, T., Lelyveld, I. & Garlaschelli, D. Early-warning signals of topological collapse in interbank networks. *Sci. Rep.* **3**, 3357 (2013).
24. Saracco, F. et al. Inferring monopartite projections of bipartite networks: an entropy-based approach. *N. J. Phys.* **19**, 16 (2017).
25. Becatti, C., Caldarelli, G., Lambiotte, R. & Saracco, F. Extracting significant signal of news consumption from social networks: the case of Twitter in Italian political elections. *Palgrave Commun.* **5**, 1–16 (2019).
26. Caldarelli, G. et al. The role of bot squads in the political propaganda on Twitter. *Commun. Phys.* **3**, 1–15 (2020).
27. Neal, Z. P., Domagalski, R. & Sagan, B. Comparing alternatives to the fixed degree sequence model for extracting the backbone of bipartite projections. *Sci. Rep.* **11**, 23929 (2021).
28. Parisi, F., Squartini, T. & Garlaschelli, D. A faster horse on a safer trail: generalized inference for the efficient reconstruction of weighted networks. *N. J. Phys.* **22**, 053053 (2020).
29. Berge, C. Graphes et Hypergraphes, 1st edn., Dunod, Paris, France. Monographies universitaires de mathématiques №37 p. 502 (1970).
30. Berge, C. Graphes et Hypergraphes, 2nd edn., Dunod, Paris, France. 2 édition, Dunod université, Mathématiques 604 p. 516 (1973).
31. Voloshin, V.I. Introduction to Graph and Hypergraph Theory, p. 287. (Nova Science Publishers, 2009).
32. Ghoshal, G., Zlatic, V., Caldarelli, G. & Newman, M. E. J. Random hypergraphs and their applications. *Phys. Rev. E* **79**, 066118 (2009).
33. Erdös, P. & Rényi, A. On random graphs i. *Publicationes Mathematicae* **6**, 290–297 (1959).
34. Gilbert, E. N. Random graphs. *Ann. Math. Stat.* **30**, 1141–1144 (1959).
35. Saracco, F., Clemente, R. D., Gabrielli, A. & Squartini, T. Randomizing bipartite networks: the case of the world trade web. *Sci. Rep.* **5**, 10595 (2015).
36. Strona, G., Nappo, D., Boccacci, F., Fattorini, S. & San-Miguel-Ayanz, J. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nat. Commun.* **5**, 4114 (2014).
37. Carstens, C. J. Proof of uniform sampling of binary matrices with fixed row sums and column sums for the fast Curveball algorithm. *Phys. Rev. E* **91**, 1–7 (2015).
38. Kraakman, Y.J., Stegehuis, C. Hypercurveball algorithm for sampling hypergraphs with fixed degrees (2024). https://arxiv.org/abs/2412.05100.
39. Coolen, A. C. C., Martino, A. D. & Annibale, A. Constrained Markovian dynamics of random graphs. *J. Stat. Phys.* **136**, 1035–1067 (2009).
40. Roberts, E.S., Coolen, A.C.C. Unbiased degree-preserving randomization of directed binary networks. *Phys. Rev. E* **85**, 046103 (2012).
41. Efron, B. Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7**, 1–26 (1979).
42. LaRock, T. & Lambiotte, R. Encapsulation structure and dynamics in hypergraphs. *J. Phys. Complex.* **4**, 045007 (2023).
43. Landry, N. W., Young, J.-G. & Eikmeier, N. The simpliciality of higher-order networks. *EPJ Data Sci.* **13**, 17 (2024).
44. Benson, A. R. Three hypergraph eigenvector centralities. *SIAM J. Math. Data Sci.* **1**, 293–312 (2019).
45. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **57**, 289–300 (1995).
46. Pratelli, M., Saracco, F. & Petrocchi, M. Entropy-based detection of Twitter echo chambers. *PNAS Nexus* **3**, 177 (2024).

47. Guarino, S., Mounim, A., Caldarelli, G. & Saracco, F. Verified authors shape X/Twitter discursive communities. arXiv preprint arXiv:2405.04896 (2024).
48. Ferraz de Arruda, G. et al. Multistability, intermittency, and hybrid transitions in social contagion models on hypergraphs. *Nat. Commun.* **14**, 1375 (2023).
49. Iacopini, I., Petri, G., Baronchelli, A. & Barrat, A. Group interactions modulate critical mass dynamics in social convention. *Commun. Phys.* **5**, 64 (2022).
50. St-Onge, G. et al. Influential groups for seeding and sustaining nonlinear contagion in heterogeneous hypergraphs. *Commun. Phys.* **5**, 25 (2022).

Acknowledgements

Author contributions

F.S., G.P., R.L. and T.S. conceived the study. F.S. performed the analyses and T.S. wrote the first version of the manuscript. F.S., G.P., R.L. and T.S. discussed the results, revised the draft and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42005-025-02182-2.

Correspondence and requests for materials should be addressed to Fabio Saracco or Renaud Lambiotte.

Peer review information *Communications Physics* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints