

Random walks on hypergraphs

Timoteo Carletti

naXys, Namur Institute for Complex Systems, University of Namur, Belgium

Federico Battiston

Department of Network and Data Science, Central European University, Budapest 1051, Hungary

Giulia Cencetti

Mobs Lab, Fondazione Bruno Kessler, Via Sommarive 18, 38123, Povo, TN, Italy

Duccio Fanelli

*Dipartimento di Fisica e Astronomia, Università di Firenze,
INFN and CSDC, Via Sansone 1, 50019 Sesto Fiorentino, Firenze, Italy*

In the last twenty years network science has proven its strength in modelling many real-world interacting systems as generic agents, the nodes, connected by pairwise edges. Yet, in many relevant cases, interactions are not pairwise but involve larger sets of nodes, at a time. These systems are thus better described in the framework of hypergraphs, whose hyperedges effectively account for multi-body interactions. We hereby propose a new class of random walks defined on such higher-order structures, and grounded on a microscopic physical model where multi-body proximity is associated to highly probable exchanges among agents belonging to the same hyperedge. We provide an analytical characterisation of the process, deriving a general solution for the stationary distribution of the walkers. The dynamics is ultimately driven by a generalised random walk Laplace operator that reduces to the standard random walk Laplacian when all the hyperedges have size 2 and are thus meant to describe pairwise couplings. We illustrate our results on synthetic models for which we have a full control of the high-order structures, and real-world networks where higher-order interactions are at play. As a first application of the method, we compare the behaviour of random walkers on hypergraphs to that of traditional random walkers on the corresponding projected networks, drawing interesting conclusions on node rankings in collaboration networks. As a second application, we show how information derived from the random walk on hypergraphs can be successfully used for classification tasks involving objects with several features, each one represented by a hyperedge. Taken together, our work contributes to unveiling the effect of higher-order interactions on diffusive processes in higher-order networks, shading light on mechanisms at the hearth of biased information spreading in complex networked systems.

INTRODUCTION

From social systems and the World Wide Web to economics and biology, networks define a powerful tool to describe many real-world systems [1–3]. Over the last twenty years of network science [4, 5], many interacting systems with different functions were shown to exhibit surprisingly similar structural properties, at different scales. Interestingly, the complex architecture of real-world networks was found to significantly interfere with the dynamical processes hosted on them, from social dynamics [6] to synchronisation [7]. As a consequence, properly tailored dynamical processes are now routinely employed to extract information on the a priori unknown structure of the underlying graphs architectures.

Networks materialise as pairwise interactions, represented by edges, among generic agents, the nodes: by their very first definition they are thus bound to encode binary relationships among units. However, an increasing amount of data indicates that, from biological to social systems, real-world interactions often occur among more than two nodes at a time. This phenomenon is not properly described by the traditional paradigm constrained on

pairwise interactions, and highlight the need for extended notions in the realm of network theory. In recent years, an emerging stream of research has been focusing on developing higher-order network models that account for the diverse kinds of higher-order dependencies, as found in complex systems.

Let us here observe that the current “high-order framework” bears some ambiguity, as it has been occasionally assumed to embrace features which are more specifically stemming from the interactions [8], as e.g. temporal and/or memory effects [9, 10], or reflect the multiplex nature of the examined system [11–13]. Here, the term higher-order is exclusively meant to refer to agents interacting in groups of arbitrary numerosity [14–17], a process often modelled via simplicial complexes [18–20] or hypergraphs [21–23], non trivial mathematical generalisation of the ordinary networks.

Our focus is on hypergraphs, where relationships among agents are described as collections of nodes assembled in sets, called hyperedges, made by any number of nodes. Hypergraphs provide a natural representation for many higher-order real-world networks [24, 25]. In social systems they can for instance be suited to describe

collaboration networks, where nodes denote authors and hyperedges stand for groups of authors, who have written papers together. Alternatively, hypergraphs can be invoked to describe face-to-face social networks where individuals can interact in groups of arbitrary sizes [26]. In biology, hypergraphs allow to properly model biochemical reactions simultaneously involving more than two species, or conveniently describe higher-order interactions among different families of proteins [15]. Crucially, in all these examples, interactions among agents occur in groups of arbitrary size, and cannot be split into disjoint pairwise interactions. Differently from simplicial complexes, a higher-order interaction described by an hypergraph (e.g. a single three body interaction) does not require the existence of all lower order interactions (e.g. the three pairwise interactions associated to the same triangle) [27]. Heterogeneous hypergraphs have been sometimes studied by mapping the nodes belonging to a hyperedge into a clique of suitable size. However, the drawback of this procedure is that it eventually yields a *projected* network, e.g. shown in Fig. 1, where only pairwise interactions are ultimately accounted for (see Appendix A).

Linear dynamics [28–30], and specifically random walks [31], constitute a simple, although powerful tool to extract information on the relational structure of interacting systems. In particular, random walks on complex networks [32] have been proven useful to compute centrality scores [33], finding communities [34] and providing a taxonomy of real-world networks [35]. In the simplest case, at each time step, a walker jumps from the node where it belongs to one of its adjacent neighbours, traveling across one of the available edges, chosen at random with uniform probability. Many variations of this fundamental process have since then been considered. These include more sophisticated dynamical implementations, which allow to targeting the walks towards nodes with given structural features [36], let them interact at the nodes of the network [37], investigate non-linear transition probabilities [38] and crowded conditions [39], consider the temporal [40–42] or multilayer [43, 44] dimensions of the edges under different network topologies.

Random walks have been defined on simplicial complexes [45], but because of the cumbersome involved combinatorics, applications have been limited to higher-order interactions of the lowest dimensions, i.e. triangles. Moreover, walkers are in general allowed to hop between edges or even high-order structures. This is at variance with the setting that we here aim at exploring, where hops can solely occur among nodes which join in a given high-order structure. In parallel, also random walks on hypergraph have been considered by assuming that all the hyperedges are made by an identical – and constant – number of nodes [46, 47]. The first random walk Laplacian defined on hypergraphs can be probably traced back to the seminal paper by Zhou and collaborators [48]. Each hyperedge is endowed with an arbitrary weight, acting as a veritable bias to the walkers dynam-

ics. As observed by the authors of [48], assigning the weights is an outstanding open problem, which deserves to be properly addressed. In this work, we will prove that a physically motivated choice for the aforementioned weights naturally emerge, when framing the problem on solid microscopic grounds.

Interestingly, more complicated nonlinear dynamics have been also recently studied on simplicial complexes [16, 20, 49, 50] or in a pure multi-body frame [51]. Once again, however, the focus is placed on low-dimensional simplicial complexes (triangles). Recently, several dynamics, including epidemic spreading [16, 49, 52] and synchronisation [38], have been shown to produce new collective behaviours when higher-order interactions are assumed to shape the networked arrangement.

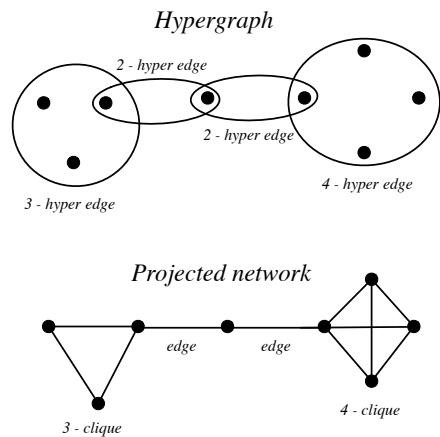


FIG. 1. **Hypergraph and projected network.** Hypergraph (top) and corresponding projected network (bottom). In the projected network each hyperedge E_α becomes a complete clique of size $|E_\alpha|$, with thus $|E_\alpha|(|E_\alpha| - 1)/2$ pairwise interactions.

Starting from this setting, and by further elaborating on the above, we propose in this work a new class of random walks, evolving on generic heterogeneous hypergraphs as dictated by a plausible physical model, and without any limitation on the sizes of the hyperedges. In this framework, multi-body proximity is associated to highly probable exchanges among agents belonging to the same hyperedge, and walkers mitigate their inclination to explore the system with a tendency to naturally spend more time in highly clustered cliques and communities. This feature is reminiscent of bias in information spreading, which is known to be affected by the phenomenon of echo chambers [53]. Similarly to the standard random walk, at each time step a walker sitting on a node, selects a hyperedge among the ones containing the origin node, with a probability proportional to the size of the hyperedge; then the walker jumps with uniform probability onto any node contained in the selected hyperedge. In this way, higher-order interactions between a group of nodes drive the process and the weights postulated in

[48] take non trivial values, as stemming from the microscopic dynamics.

We shall in particular provide an analytical description of the process, by deriving a general formula for the stationary distribution of the walk, and show that the dynamics is driven by a generalised Laplace operator, that reduces to the standard random walk Laplacian when all hyperedges have size 2, and the hypergraph results in a traditional network.

As already stated, random walks can be used to rank nodes, based on the stationary occupancy probability of walkers across the network. Because of these implications, it is therefore interesting to compare the stationary distribution, as obtained within the newly introduced framework, with that displayed by standard random walkers on the corresponding projected network. Because of the tight interactions among agents belonging to the same hyperedge, the probability to find a walker on a given node is in principle different, when confronting the outcome of the two aforementioned processes. As a consequence, we expect a different order in the ranking to be obtained for the same node, depending on the dynamical process employed in the analysis. This observation opens up the way to a new definition of centrality for systems where the high-order structure is known to be relevant. In particular, we will provide a direct evidence for our claims working with co-authorship networks, as extracted from the arXiv on-line preprint server. Our second application, to which we alluded above, concerns a classification task which is borrowed by [48]. Indeed it is well known that one can model a dataset by resorting to networks and then make use of the associated Laplacian Eigenmaps [54] to embed the data on a lower dimensional space, while hopefully preserving relevant information, in the spirit of a generalised principal component analysis. Working in the lower dimensional space allows one to cluster together objects. However, when objects to be classified share annotated features, the use of binary relationships, i.e. usual network, results in a dramatic loss of information. One can thus obtain a better embedding via hypergraphs and invoke the spectral characteristics of the associated Laplacian to achieve more effective clustering scores [48, 54]. By replicating the analysis in [48], we will here consider the problem of separating the animals listed in the UCI Machine Learning Depository in distinct class, e.g. mammals, birds, ..., by using at the scope a set of annotated features, e.g. tail, hair, legs and so on. Here nodes are animals and hyperedges features. We will show that the presence of high-order interactions among features as encoded via the proposed Laplacian operator, yields a very effective embedding with just a few of the most significative directions, a result which is in line with that reported in [48].

Summing up, we here introduce and discuss a novel generalisation of the random walk picture to higher-order networked systems, where hyperedge weights are naturally assigned and thus removing any ambiguity in their values. Finally we hint at important exploitations of this

novel dynamical framework working along two paradigmatic directions, ranking and classification of data.

MODEL

Incidence and hyper adjacency matrices. Let us consider an hypergraph $\mathcal{H}(V, E)$, where $V = \{1, \dots, n\}$ is the set of n nodes, $E = \{E_1, \dots, E_m\}$ the set of m hyperedges, with E_α an unordered collection of nodes, i.e. $E_\alpha \subset V$, $\forall \alpha = 1, \dots, m$. We observe that whenever $E_\alpha = \{i, j\}$, i.e. $|E_\alpha| = 2$, then the hyperedge is actually a “standard” edge, denoting a binary interaction among nodes i and j . An hypergraph where $|E_\alpha| = 2 \forall \alpha$ reduces to a network.

We can define the associated *hyper incidence matrix* $e_{i\alpha}$, carrying the information about how nodes are shared among hyperedges, as

$$e_{i\alpha} = \begin{cases} 1 & i \in E_\alpha \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

We note that the same matrix exists for networks. However, while in regular networks each column can have only two non zero entries, as each edge can contain two nodes only [55], in hypergraphs each column can display several non zeros entries (i.e. an hyperedge can contain several nodes).

Starting from the above matrix, one can construct the $n \times n$ *hyper adjacency matrix*, $A = ee^T$, whose entry A_{ij} represents the number of hyperedges containing both nodes i and j . We note that often the adjacency matrix is defined by setting to 0 the main diagonal. Let us also define the $m \times m$ *hyperedges matrix*, $C = e^T e$, whose entry $C_{\alpha\beta}$ counts the number of nodes in $E_\alpha \cap E_\beta$. Observe that in the literature the number of nodes in a given hyperedge, $C_{\alpha\alpha}$, is often called the degree of the hyperedge, while the node degree stands for the number of hyperedges containing the node, $\sum_\alpha e_{i\alpha} e_{i\alpha}$.

Transition probability. To describe a random walk process, we need to define the transition probability to pass from a state, hereby represented by the node on which the walker belongs to, to any other state, compatible with the former, in one time step. In the case of simple unbiased random walks on networks, one assumes the walker to take with equal probability any link emerging from the node that is initially occupied. Hence, the transition probability can be readily computed as A_{ij}/k_i , where $k_i = \sum_j A_{ij}$ is the degree of the origin node. When dealing with hypergraphs, by choosing with uniform probability any of the neighbouring nodes, namely all the nodes belonging to hyperedges connected with the origin node, is not a sensible choice. In this way, in fact, the real structure of the systems is not incorporated into the dynamical picture. On the contrary, nodes belonging to the same hyperedge exhibit a higher-order interaction and we consequently assume that spreading among them is more probable than with nodes associated to other hyperedges; because of this the information can thus spend

long periods inside the same hyperedge. For instance, a gossip can spread faster because of group interaction among individuals, than as follows successive binary encounters; similarly, ideas can circulate more effectively among collaborators, the coauthors of a joined publication, as compared to the setting where exchanges in pairs are solely allowed for. To compute the transition probability to jump from i to j , we count the number of nodes, excluding i itself, belonging to the same hyperedge of i and j . Recalling the definition of the matrix C , this can be written as

$$k_{ij}^H = \sum_{\alpha} (C_{\alpha\alpha} - 1) e_{i\alpha} e_{j\alpha} = (e\hat{C}e^T)_{ij} - A_{ij} \quad \forall i \neq j, \quad (2)$$

where $\hat{C}$ is a matrix whose diagonal coincides with that of C and it is zero otherwise (see Appendix B). By normalising so as to impose a uniform choice among the connected hyperedges, we get the following expression for the transition probabilities:

$$T_{ij} = \frac{(e\hat{C}e^T)_{ij} - A_{ij}}{\sum_{\ell} k_{i\ell}^H} = \frac{(e\hat{C}e^T)_{ij} - A_{ij}}{\sum_{\ell} (e\hat{C}e^T)_{i\ell} - k_i^H}, \quad (3)$$

where $k_i^H = \sum_{\ell} A_{i\ell}$ is the hyperdegree of the node i , a synthetic measure reminiscent of the node degree, which takes into account both the number and the size of hyperedges i in which i is involved.

When the hypergraph is a network, all hyperedges have 2 nodes. Hence

$$(e\hat{C}e^T)_{ij} = \sum_{\alpha} C_{\alpha\alpha} e_{i\alpha} e_{j\alpha} = 2 \sum_{\alpha} e_{i\alpha} e_{j\alpha} = 2A_{ij}, \quad (4)$$

and Eq. (3) reduces to the standard transition probability for random walk on networks

$$T_{ij} = \frac{2A_{ij} - A_{ij}}{2k_i^H - k_i^H} = \frac{A_{ij}}{k_i}, \quad (5)$$

where we used the fact that, under this assumption, $k_i^H = k_i$.

Stationary solution. Having computed the transition probabilities, we can proceed further by formulating the dynamical equation which rules the temporal evolution of the probability $\mathbf{p}(t) = (p_1(t), \dots, p_n(t))$ of finding the walker on a given node after $t > 0$ steps. The process is governed by the following equation:

$$p_i(t+1) = \sum_j p_j(t) T_{ji}, \quad (6)$$

where the right hand side term combines the probability to be in any node j at time t and the probability to perform a jump towards the target node i , during the next time of iteration. As $\sum_j T_{ij} = 1$ for all i , the stationary probability distribution, $\mathbf{p}^{(\infty)}$, is thus the left eigenvector associated with the eigenvalue $\lambda_1 = 1$ of $\mathbf{T}$.

Given $\mathbf{T}$, it is possible to obtain an exact analytical solution for the stationary state $\mathbf{p}^{(\infty)}$ which encapsulates the higher-order structure of the system. Indeed, a straightforward computation (see Appendix C), yields:

$$p_j^{(\infty)} = \frac{\sum_{\ell} (e\hat{C}e^T)_{j\ell} - k_j^H}{\sum_{m\ell} [(e\hat{C}e^T)_{m\ell} - k_m^H]}, \quad (7)$$

for all $j = 1, \dots, n$. In the case the hypergraph is indeed a network, we recover the well known expression $q_j^{(\infty)} = k_j / \sum_l k_l$ for the stationary solution of the walk. Let us observe that

$$L_{ij} = \delta_{ij} - T_{ij} = \delta_{ij} - \frac{k_{ij}^H}{\sum_{\ell} k_{i\ell}^H}, \quad (8)$$

is a new *random walk* Laplacian that generalises the random walk one for networks. Moreover the former reduces to the latter in the case $|E_{\alpha}| = 2$ for all α .

We observe that the formalism readily extends to the case of continuous-time random walks, where the evolution of the probability is given by

$$\dot{p}_i(t) = \sum_j p_j(t) T_{ji} - \sum_j p_i T_{ij}.$$

Similarly to the case of networks, as $\sum_j T_{ij} = 1$, it is possible to rewrite the latter as

$$\dot{p}_i = \sum_j p_j (T_{ji} - \delta_{ij}) = - \sum_j p_j L_{ji},$$

where $\mathbf{L}$ is the above defined Laplace matrix. In the following, for a sake of for the sake of definiteness we limit our analysis to explore the properties of the discrete-time random walks on synthetic and real-world hypergraphs, leaving the continuous time case to a further work.

Denote by $\mathbf{D}$, the diagonal matrix with entries $d_i^H = \sum_j k_{ij}^H$ and by $\mathbf{K}^H$, the matrix characterised by elements k_{ij}^H . We can introduce the *symmetric* Laplacian $\mathbf{L}^{sym}$ as:

$$\mathbf{L}^{sym} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{K}^H \mathbf{D}^{-1/2},$$

which is well defined since $k_{ij}^H \geq 0$. $\mathbf{L}^{sym}$ is similar to the operator introduced via relation (8), indeed $\mathbf{L} = \mathbf{D}^{-1/2} \mathbf{L}^{sym} \mathbf{D}^{1/2}$. The newly introduced operator $\mathbf{L}$ is hence a properly defined Laplacian: it is in fact non-negative definite, it displays real eigenvalues and the smallest eigenvalue is identically equal to zero, as it readily follows by virtue of the proven similarity with $\mathbf{L}^{sym}$.

Before turning to discussing the applications, we will briefly draw a comparison with the setting proposed by Zhou [48] and show how this materialises in a natural solution for the problem of weights determination. The Laplacian operator $\mathbf{L}^z$ introduced in [48] can be cast in the form:

$$L_{ij}^z = \delta_{ij} - \sum_{\alpha} \frac{w_{\alpha}}{W_i C_{\alpha\alpha}} e_{i\alpha} e_{j\alpha}, \quad (9)$$

where w_α identifies the undetermined weight of the hyperedge E_α , $W_i = \sum_\alpha w_\alpha e_{i\alpha}$ is the total weight of the hyperedges containing the node i , i.e. weighted node degree, and $C_{\alpha\alpha}$ stands for the number of nodes in the

hyperedge E_α . A simple calculation, as detailed in the following, shows that operator $\mathbf{L}$ can be eventually recovered from $\mathbf{L}^z$ by imposing the non trivial weights $w_\alpha = C_{\alpha\alpha}(C_{\alpha\alpha} - 1)$. In fact:

$$\begin{aligned} L_{ij}^z &= \delta_{ij} - \sum_\alpha \frac{C_{\alpha\alpha}(C_{\alpha\alpha} - 1)}{C_{\alpha\alpha} \sum_\beta C_{\beta\beta}(C_{\beta\beta} - 1) e_{i\beta}} e_{i\alpha} e_{j\alpha} = \delta_{ij} - \sum_\alpha \frac{(C_{\alpha\alpha} - 1)}{\sum_\beta C_{\beta\beta}(C_{\beta\beta} - 1) e_{i\beta}} e_{i\alpha} e_{j\alpha} \\ &= \delta_{ij} - \frac{k_{ij}^H}{\sum_\beta \sum_\ell e_{\ell\beta}(C_{\beta\beta} - 1) e_{i\beta}} = \delta_{ij} - \frac{k_{ij}^H}{\sum_\ell k_{i\ell}^H} = L_{ij}, \end{aligned}$$

where used has been made of definition (2) for k_{ij}^H and the fact that $C_{\beta\beta} = \sum_\ell e_{\ell\beta}$. As anticipated, a natural choice for the weights as postulated in [48] can be envisaged, which follows a sensible microscopic modelling of the random walk dynamics.

By invoking Theorem 4 in [56], we can finally conclude that our process is equivalent to a random walk on a weighted projected network, where the weights of the link ij is given by k_{ij}^H , that is the weights scale extensively with the region of influence of the nodes, namely the size of the hyperedge they belong to. It is indeed quite remarkable that a properly weighted binary network encapsulates the higher order information, as stemming for the corresponding hypergraph representation. Observe that authors in [56] also consider an extension of the Zhou et al. model, where nodes bear a given weight, tuned so as to reflect the hyperedge characteristics. Again, the introduced weights are abstract quantities, and do not reflect a physically motivated choice.

RESULTS

Since the Page-Rank [57, 58], random walks on networks are routinely applied to compute centrality scores [33]. Indeed they can be used to rank nodes according to the probability to be visited by the walker, the larger the latter the more “important”/“central” the node. In this section, we show that high-order interactions can strongly modify the ranking, as resulting from a random walk process on hypergraphs, with respect to the homologous estimate as computed for the corresponding projected network. This fact can thus bear relevant implications for ranking real data, stemming from a dynamical process which is better explained in terms of hypergraphs. In this case, in fact, the applications of ranking tools tailored to pairwise interactions might produce misleading results (see Appendices C and E).

To illustrate the effect of a non-trivial higher-order structure, we consider a simple hypergraph made by m hyperedges of size 2 all intersecting in a common node, h ; a different node, say c , belongs to one of such 2-hyperedges and to a hyperedge of size k (see Fig. 2 panel a) for the case $m = 7$ and $k = 6$).

The random walk on the projected network will rank nodes according to their degree, i.e. $q_i^{(\infty)} \sim k_i$. Hence for $m > k$, the node h with $k_h = m$, is ranked first, followed by the c node, $k_c = k$, and all other ones (see green curves in panel c Fig. 2). In contrast, the random walk on the hypergraph ranks nodes taking into account higher-order relations. Since from Eq. (7) we get $p_h^{(\infty)} \sim m$ and $p_c^{(\infty)} \sim 1 + (k - 1)^2$, thus h is the top node as long as $m > 1 + (k - 1)^2$ (see orange curves in panel c Fig. 2). In conclusion, for a fixed size of the hyperedge k , if the “hub” node is too small (see panel d), $m < \hat{k} = k + 1$, or the hub is very large (see panel f), $m > \tilde{k} = 1 + (k - 1)^2$, then both processes will rank nodes in the same way. However, there exists a range of intermediate values, $\hat{k} < m < \tilde{k}$, for which the top ranked node on the hypergraph is the c node while the random walk on the projected network returns the h node as top rank (see panel e). This phenomenon of ranking inversion will be further discussed in Appendix C. In the aim of maximising the probability of occupancy of a given node, it is preferable for this latter to be connected to nodes organised into few large hyperedges, than to many parcelled units.

To further characterise the impact of the high-order interactions on diffusion on larger systems, we consider a second synthetic model where all nodes have the same number of neighbours, which are arranged in a tuneable number of triangles, i.e. hyperedges of size $|E_\alpha| = 3$. The model interpolates between the case where the number of triangles is zero, $f = 0$, meaning that all interactions involve simple pairs, and the case where there are no pairwise interactions but only 3-body ones, $f = 1$. More precisely, we start with a 1D regular lattice where nodes are connected to 4 neighbours (2 on the left and 2 on the right). Each nodes has hence degree 4 and takes part to 2 distinct triangles, i.e. hyperedges with size 3, and $f = 1$. Then with probability p we iteratively swap the ending points of the links with a “criss-cross” rewire, i.e. preserving the nodes degree, progressively eliminating 3-hyperedges, hence triangles. In the limit of high rewire triangles have a negligible probability to be formed, and one eventually obtains a regular random graph with degree $k = 4$. In the process, we control that no hyperedge of size greater than 3 is created, so that competition is

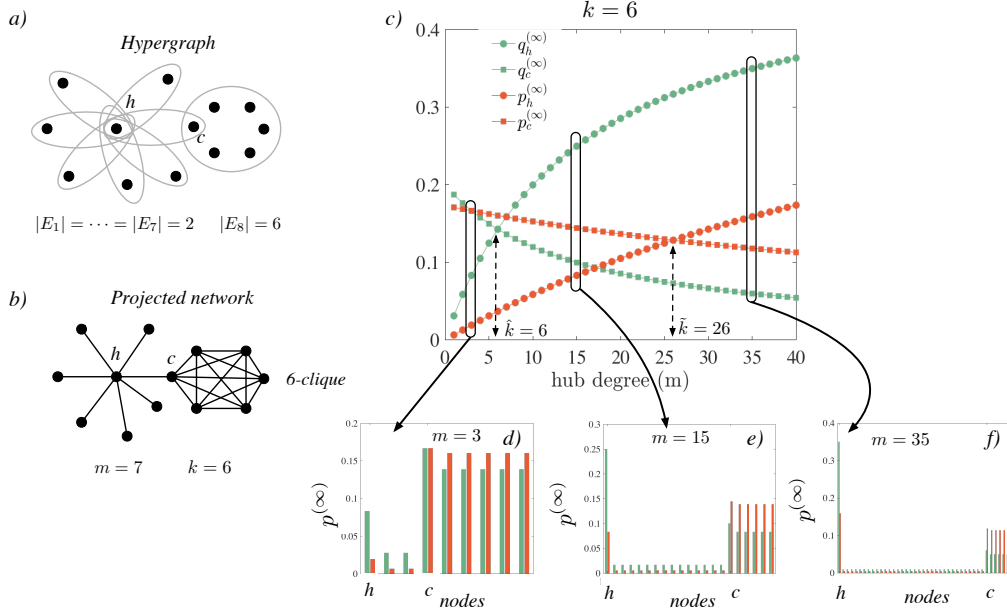


FIG. 2. **The (m, k) -star-clique network.** Panel a: hypergraph made by $m + k = 13$ nodes, divided into $m = 7$ hyperedges of size 2 and one large hyperedge of size $k = 6$. The node h belongs to all the 2-hyperedges, while the node c belongs to one 2-hyperedge and to the 6-hyperedge. Panel b: the projected network where hyperedges are mapped into complete cliques, the 6-hyperedge becomes thus a 6-clique. Panel c: we show the dependence on m of the asymptotic probability of finding the walker on the node h (circle) or on the node c (square), in the projected network (green symbols) and in the hypergraph (orange symbols). Panels d, e and f: we report the asymptotic probabilities $q_i^{(\infty)}$ and $p_i^{(\infty)}$ for three values of m : $m = 3 < \hat{k}$, $\hat{k} < m = 15 < \tilde{k}$ and $\tilde{k} < m = 35$, where $\hat{k} = 6$ and $\tilde{k} = 26$.

only between 2-body and 3-body interactions.

As the degree sequence is unchanged throughout this process and every node shares the same number of links, the asymptotic distribution of walkers on the projected network is uniform and given by $q_i = 1/N$ for all i , where N is the number of nodes, set to 500 in the example below, no matter the value of f . This is also the case for the random walk on hypergraph, in the two limiting cases $f = 0$ and $f = 1$; indeed in the former case the hypergraph and the projected network do coincide because all the hyperedges have size 2. In the latter setting, all nodes are involved in the same number of higher-order interactions and thus they are all equivalent. However, for the walk on hypergraphs the stationary state changes at the intermediate stages of f . In order to quantify the heterogeneity of the stationary state we rely on the Gini coefficient, which is defined as the average absolute difference between all pairs of elements in the vector p , divided by the average:

$$G(p) = \frac{\sum_{i=1}^N \sum_{j=1}^N |p_i - p_j|}{2N \sum_{i=1}^N p_i}. \quad (10)$$

The Gini coefficient for the stationary state of random walk on the above described hypergraph is reported in top panel of Fig. 3. For the limiting values $f = 0$ and $f = 1$ the stationary state on the hypergraph coincides with

the one on the projected network and the Gini index is 0 being the asymptotic solution homogeneous. However, high-order structures arising for intermediate values of the fraction of triangles induce a heterogeneity in the occupation of the different nodes at equilibrium, which is thus different from the one obtained for the associated projected network.

A standard metric to compare lists is the Jaccard index, a measure of the fraction of elements that are common between two lists with respect to the total number of involved elements, $J(A, B) = |A \cap B| / |A \cup B|$. As the Jaccard index does not take into account the order of the elements as appearing in the two confronted lists, we compare the rankings of the two stationary distributions by means of a modified Jaccard index, $\hat{J}$, recently introduced in [59]. Here differences at the top of the ranking induce a stronger change, than differences associated to the lower ranked elements. Let us observe also that the Jaccard index is unable to detect a permutation in the order of the elements in a list, while the modified one does. In the bottom panel of Fig. 3, we show the average modified Jaccard index, $\hat{J}$, for the M -top ranking, $M = 100, 300, 500$, as a function of the fraction f of 3-hyperedges existing in the system. The results are in agreement with the ones obtained via the Gini coefficient; for $f = 0$ and $f = 1$ the rankings do coincide and thus $\hat{J}$ achieves its maximum value, i.e. 1, while for intermediate

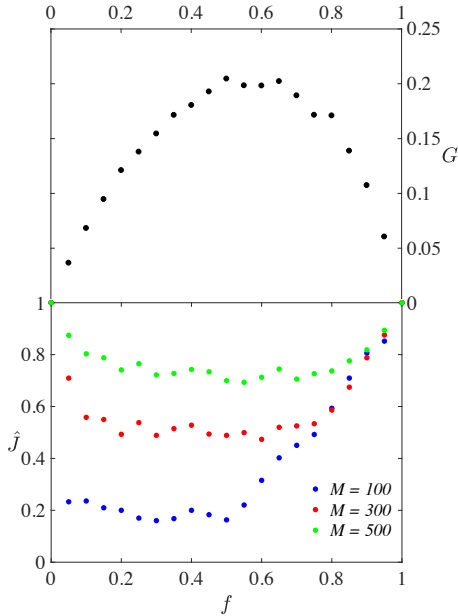


FIG. 3. **Impact of the 3-body interaction on the asymptotic solution of the random walk on the hypergraph.** The top panel reports the Gini coefficient for the stationary state of the random walk on hypergraphs as a function of the fraction of hyperedges of size three, f . Recall that the model does not allow for hyperedges of size larger than 3. The bottom panel shows for the same networks the modified Jaccard index in order to compare the rankings of nodes for the hypergraph and the projected network. Different colours (blue, red and green), correspond to different numbers of nodes chosen for the comparison, i.e. the top 100, the top 300 and all the 500 nodes, respectively.

values of f the index $\hat{J}$ drops down reflecting differences among the rankings. Moreover, we can appreciate the presence of a large turnover in the top lists: indeed $\hat{J}$ associated to small M , i.e. comparing relatively few nodes in the top list, are much smaller than that for large M , i.e. longer lists.

To take one step forward, we consider a synthetic model where high-order structures are not limited to 3-body but larger hyperedges are allowed for. We thus build a third model which interpolates from a 1D ring to a fully connected network. More precisely, we start from a 1D ring where all the nodes have degree 2, and then progressively increase its density as measured by the total number of links, l , until the process terminates with a complete network, corresponding to a hypergraph with a single hyperedge containing all the nodes. Links are added at random avoiding self-loops and multiple links. We note that differently from the previous case, at intermediate values of l , this model presents a much wider variety in the size of the hyperedges (or cliques in the projected network), which are not anymore limited to 2-body and 3-body interactions. For this reason, the structure of the ranking difference is definitely more complex

and rich than what one could eventually guess by just looking at the number of 3-hyperedges, 4-hyperedges or 5-hyperedges (see Fig. 4).

In the initial configuration of a 1D ring, the stationary solutions of the hypergraph and the projected network coincide, because of the absence of higher-order interactions. Similarly, they are also equivalent in the opposite limit, i.e. when the fully connected network is generated. For intermediate number of added links, the two processes result instead in different rankings. In Fig. 4, we report $\hat{J}$ as a function of the total number of links l , to compare the M -top rankings, as obtained by using the random walk on the hypergraph and on the projected network, respectively. We reports in particular results for three values, $M = 5, 10, 20$. The behaviour of the three curves is qualitatively similar. Indeed they all reach the value 1, i.e. perfect matching of the respective rankings for $l = 20$ (initial 1D ring). Then, even the addition of just few links makes the rankings to change abruptly and $\hat{J}$ consequently drops to low values. This is associated with the creation of small hyperedges with size equal to 3 (see bottom panel of Fig. 4). Adding more links reduces the differences, namely $\hat{J}$ increases, up to $l = 190$ (complete network) where again the rankings do coincide and the index equals 1. This is associated with the birth of larger hyperedges. Let us remark that $\hat{J}$ for $M = 5$ is much smaller than the same quantity computed with $M = 10$ (rank half of the nodes) and $M = 20$ (rank all the nodes) meaning that there is a strong turnover in the top positions. The heterogeneity in the stationary solutions of this model, as well as the star-clique example, is further investigated in the Appendix D where the corresponding Gini coefficients are shown.

APPLICATIONS

Node ranking

In the previous section we have shown that hypergraph and the projected network can exhibit different stationary solutions because of ranking inversion (see Appendix C). We thus decided to analyse the impact of this observation in real networks of scientific collaborations, in our opinion one of the most representative examples of high-order structures in human interactions. The analysed data have been gathered from the arXiv database (see Appendix E for more details). Human collaborations are often schematised as resorting to pairwise interaction, a working ansatz which amounts to ignoring the organisation in teams. At variance, we have instead built a hypergraph where researchers (i.e nodes) co-authoring an article are part of the same hyperedge.

We have then determined the largest connected component of the hypergraph and that of the projected network, considered maximal and unique hyperedges (to have a fair comparison with the cliques) and computed: (i) the stationary distribution $\mathbf{p}^{(\infty)}$ for the random walk

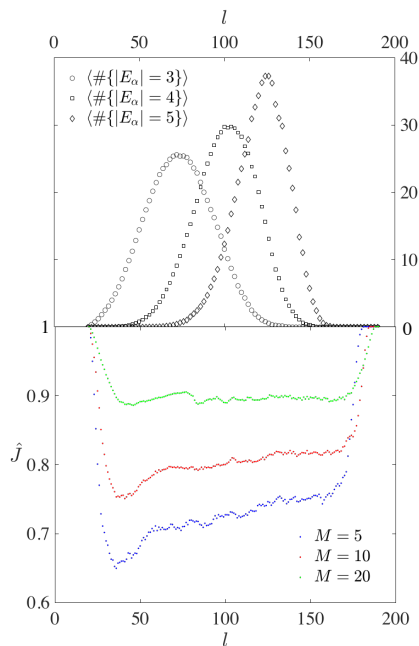


FIG. 4. **Impact of high-order structures on the asymptotic distribution of walkers for the random walk on the hypergraph and on the projected network.** Using the algorithm presented in the text, by iteratively adding links we create hypergraphs that interpolate from a regular 1D ring (where $N = 20$ nodes are connected each one with its two neighbours) to a complete graph. We then perform the random walk process on respectively the hypergraphs and the associated projected network and compare the resulting ranking (the top 5 blue, the top 10 red and the top 20 green, i.e. the whole set of nodes) using $\hat{J}$ (bottom panel). For a small number of available links, l , the hypergraphs does not present many hyperedges and thus the ranking are very close, $\hat{J} \sim 1$. As l starts to increase, few hyperedges of size 3 are created (see circles in the top panel) and the rankings estimated with the two alternative methods deviate, the values of $\hat{J}$ dropping in turn. However, as l increases even more, larger high-order structures, e.g. 4 and 5 hyperedges, emerge (see square and diamond symbols in the top panel) and $\hat{J}$ steadily increases. For a large ensemble of added links, $l \gtrsim 170$, the rankings become similar and $\hat{J} \sim 1$.

on the associated hypergraph, (ii) the stationary distribution $\mathbf{q}^{(\infty)}$ for the random walk on the corresponding projected network. We then normalise the computed stationary probabilities by their relative maximum so as to favour a comparative visualisation. In Fig. 5 we plot $p_i^{(\infty)} / \max_j p_j^{(\infty)}$ vs $q_i^{(\infty)} / \max_j q_j^{(\infty)}$ for the case arXiv-astro and arXiv-physics. In Fig. 15 the same comparison is drawn for the complete arXiv dataset.

Author are ranked differently, according to the two criteria, the one based on hypergraphs being more sensitive to the organisation in groups. If the computed rankings were (almost) the same, the data would (almost) lie on the main diagonal; deviation from this, re-

sults in novel information conveyed by the random walk on the hypergraph. The unitary square in the plane $(q_i^{(\infty)}, p_i^{(\infty)})$ can be divided into four smaller squares (see Fig. 5). The majority of the authors lies in the bottom left square, $[0, 1/2] \times [0, 1/2]$: these authors have therefore written a few papers with a small number of co-authors. Three other regions can be however identified which roughly correspond to the bounded squares: $[1/2, 1] \times [0, 1/2]$ (bottom right), $[0, 1/2] \times [1/2, 1]$ (top left) and $[1/2, 1] \times [1/2, 1]$ (top right). Authors in the top right square are top ranked in both processes: they have hence written a large number of papers with different collaborators (large degree), but they have also contributed to a relevant number of papers with many co-authors, i.e. large hyperedge size. Scholars in the bottom right square are better ranked by the random walk on the network; this means that they have written several papers but with a small number of co-authors (see e.g. right panel corresponding to physics in Fig. 5). Finally, researchers in the top left square manifest a complementary attitude: they have participated to a small number of papers, but written by many authors (see e.g. left panel corresponding to astro in Fig. 5).

As a further consideration, we can bring to the fore different “habits” of publication and writing papers that authors exhibit in each domain, despite the distribution of node degrees, i.e. number of different collaborators per author, and of hyperedges size, i.e. number of co-authors in papers, shows a quite similar shape across domains, as e.g. broad tails (see annexed supplementary information). This is particularly relevant for the High Energy Particle (hep) archive, one among the oldest ones and divided into four subcategories, experimental (ex), lattice (lat), phenomenology (ph) and theory (th) (see Fig. 6). Indeed hep-ex and hep-ph populates mainly the top right square, while hep-lat and hep-th are more present in the top right and bottom right squares. Researchers belonging to the former community tend therefore to write several papers with many co-authors, while those associated to the latter have papers with many different collaborators, each one co-authored by a small number of scholars. This is also confirmed by the largest degree found in the four subcategories (see Table I) which is as large as ~ 1200 for hep-ex and hep-ph, while it is almost 4 times smaller for hep-lat and hep-th.

The comparison drawn may allow to introduce apt corrections to usual bibliographic indicators, by properly weighting the participation to large collaborations, as opposed to research activities carried out in small groups. Recall that the hypergraph Laplacian is equivalent to the Laplacian obtained from a properly weighted projected network, which inherits of the high-order structures of the hypergraph [56]. Assessing the higher order ranking amounts therefore to applying the usual tools to this latter weighted binary graph, a conclusion which points to an immediate operative translation of the newly introduced methods.

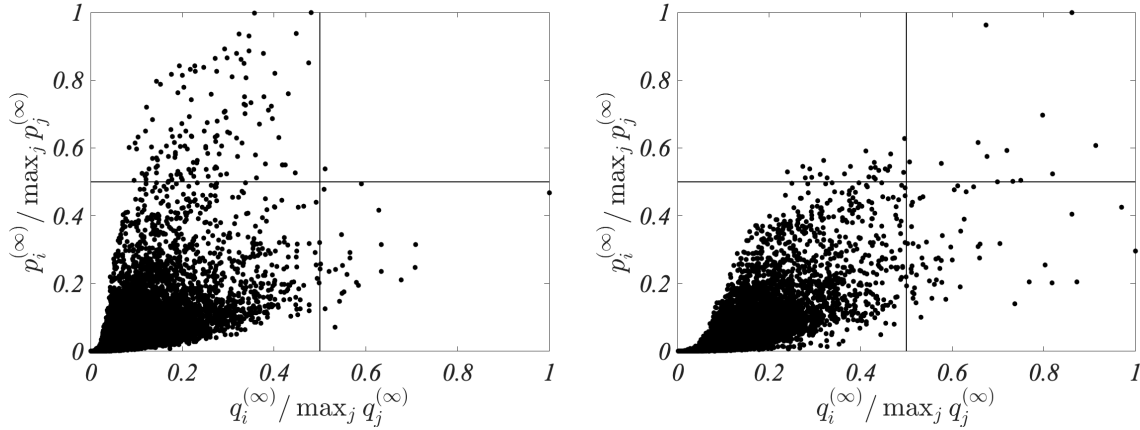


FIG. 5. **Comparing the rankings in the arXiv community: the case of Astro and Physics.** We report the scatter plot of the normalised rankings obtained with the RW on network, $q_i^{(\infty)}$, and the one computed using the random walk on hypergraphs, $p_i^{(\infty)}$ for the arXiv-astro (left panel) and the arXiv-physics (right panel).

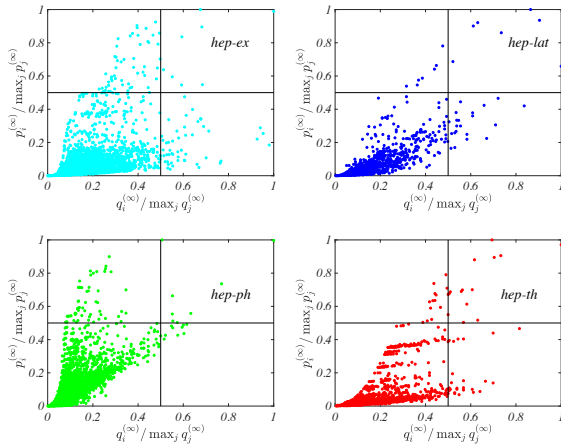


FIG. 6. **Publication habits arXiv-hep.** We report the scatter plot of the normalised rankings obtained with the random walk on the projected network, $q_i^{(\infty)}$, and the one computed using the RW on hypergraphs, $p_i^{(\infty)}$ for the four subdomains of the arXiv-hep domain.

Classification task

To further test the interest of a generalised random walk process biased to account for hyperedged communities within a plausible microscopic framework, we consider the classification task studied by Zhou et al. [48]. We anticipate that the obtained classification outper-

forms that obtained under the usual random walk framework, which ignores the annotated hyper structures.

A standard pipeline to analyse a dataset starts with the determination of pairwise similarities between the objects to be eventually classified. This implies defining a network that can be studied by means of standard spectral methods. However, similarities involve often groups of objects. In this respect, hypergraphs define the ideal mathematical platform to account for the inherent complexity of the classification problem. More precisely, one can make use of spectral methods based on the hypergraph Laplace matrix to eventually obtain a classification which effectively accounts for high-order interaction as displayed in the data [60].

Following [48] we consider an ensemble of animals from a zoologically heterogeneous set. Specifically, we used the zoo database taken from the UCI Machine Learning Depository [61], containing 101 animals, each one endowed with 16 features, such as tail, hair, legs and so on. To each animal we associate its corresponding class, e.g. mammals, birds, etc. (see Appendix F). Here nodes are animals and hyperedges features; we will show that the presence of high-order interactions among features allow to obtain a very satisfying embedding using only 2 or 3 dimensions, a result which is in line with that reported in Zhou et al. [48] for an *ad hoc* choice of the free weights parameters. To this end we build a hypergraph using the above recipe, we compute its random walk Laplacian and eventually its ensuing spectrum. We list the eigenvalues in ascending order and rename ac-

cordingly the eigenvectors. We use the first left eigenvectors [62], associated to the smallest eigenvalues, as coordinates of a Euclidean space where to embed the data (see Fig. 7). Let us observe that since we use a random walk Laplacian, the first eigenvector, i.e. the one associated to the 0 eigenvalue, is not homogeneous and it already contains non trivial information on the structure of the examined sample. Classes are identified by different colours: mammal (yellow circle), bird (magenta up triangle), reptile (cyan left triangle), fish (red right triangle), amphibian (green diamond), bug (blue square) and invertebrate (black down triangle). One can visually appreciate homologous symbols do cluster in space, hence suggesting that the embedding yields an accurate classification. Indeed, the ground-truth partition of animals into these seven classes, and the one obtained by performing a K-means clustering in this 3-dimensional space have an Adjusted Rand Index (ARI) [63] equal to 0.54. In Appendix F we show the results that we obtain when using the eigenvectors of the Laplacian of the projected network. In this case the method is less performant. The classification task is hard to be reached with such a small number of dimensions if the hyperedges are not at play, and the found clusters are not correlated to the ground-truth partition, returning a value $ARI = -0.03$.

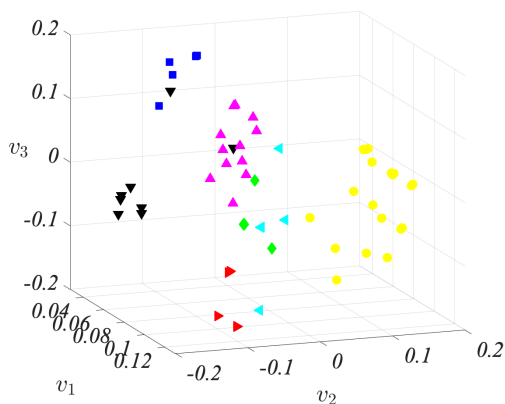


FIG. 7. **Classification of the animals according to their features.** We report a 3D embedding of the zoo dataset, namely using the first three eigenvectors. Each combination colour/symbol refer to a know class and one can appreciate by eyeball analysis the resulting clusters.

CONCLUSIONS

Summing up, we have here introduced a new class of random walks on hypergraphs which take into account the presence of higher-order interactions. We provided an analytical expression for the ensuing stationary distribution, based on the structural features of the networked system, and compared it to the distribution associated to a traditional random walk performed on the corresponding projected network. More precisely, we proposed a self-consistent recipe grounded on a microscopic physical random process biased by the hyperedges sizes to assign weights to hyperedges. We further characterised the dynamics by comparing the two processes on several synthetic and real-world networks, both by means of numerical simulations and analytical arguments. We show that our process produces stationary distributions different from those obtained for the corresponding projected network, and that prove sensitive to higher-order structure in a networked architecture. Our framework is applied to collaboration networks, yielding new insights on node ranking and centrality measure, which allow for a richer characterisation of individual performances, as compared to traditional methods. Moreover, we show that information embedded in the higher-order walk can be used to achieve accurate classification. In particular, we applied our method to successfully cluster into different families, animals with different features, each one representing an hyperedge. The same procedure fails if a simple random walk on the corresponding projected network is considered. Importantly, the proposed Laplacian is equivalent to that stemming from a properly tuned weighted network [56]. Higher order rankings and refined classifications could be hence immediately obtained by supplying to conventional tools and analysis schemes the weighted adjacency matrix that characterises the graph with pairwise edges associated to the hypergraph construction. Taken all together, our work sheds new light on dynamical processes on networks which are not limited to pairwise interactions, and on the complex interplay between the structure and dynamics of higher-order interaction networks. Future applications to machine learning based approaches to classification are also envisaged.

[1] Mark EJ Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).

[2] Albert-László Barabási *et al.*, *Network science* (Cambridge university press, 2016).

- [3] Vito Latora, Vincenzo Nicosia, and Giovanni Russo, *Complex networks: principles, methods and applications* (Cambridge University Press, 2017).
- [4] Réka Albert and Albert-László Barabási, “Statistical mechanics of complex networks,” *Reviews of modern physics* **74**, 47 (2002).
- [5] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang, “Complex networks: Structure and dynamics,” *Physics reports* **424**, 175–308 (2006).
- [6] Claudio Castellano, Santo Fortunato, and Vittorio Loreto, “Statistical physics of social dynamics,” *Reviews of modern physics* **81**, 591 (2009).
- [7] Alex Arenas, Albert Díaz-Guilera, Jurgen Kurths, Yamir Moreno, and Changsong Zhou, “Synchronization in complex networks,” *Physics reports* **469**, 93–153 (2008).
- [8] Renaud Lambiotte, Martin Rosvall, and Ingo Scholtes, “From networks to optimal higher-order models of complex systems,” *Nature physics*, 1 (2019).
- [9] Ingo Scholtes, Nicolas Wider, René Pfitzner, Antonios Garas, Claudio J Tessone, and Frank Schweitzer, “Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks,” *Nature communications* **5**, 5024 (2014).
- [10] Martin Rosvall, Alcides V Esquivel, Andrea Lancichinetti, Jevin D West, and Renaud Lambiotte, “Memory in network flows and its effects on spreading dynamics and community detection,” *Nature communications* **5**, 4630 (2014).
- [11] Manlio De Domenico, Clara Granell, Mason A Porter, and Alex Arenas, “The physics of spreading processes in multilayer networks,” *Nature Physics* **12**, 901 (2016).
- [12] Mikko Kivelä, Alex Arenas, Marc Barthélemy, James P Gleeson, Yamir Moreno, and Mason A Porter, “Multilayer networks,” *Journal of complex networks* **2**, 203–271 (2014).
- [13] Federico Battiston, Vincenzo Nicosia, and Vito Latora, “The new challenges of multiplex networks: Measures and models,” *The European Physical Journal Special Topics* **226**, 401–416 (2017).
- [14] Austin R Benson, David F Gleich, and Jure Leskovec, “Higher-order organization of complex networks,” *Science* **353**, 163–166 (2016).
- [15] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, “Simplicial closure and higher-order link prediction,” *Proceedings of the National Academy of Sciences* **115**, E11221 (2018).
- [16] Iacopo Iacopini, Giovanni Petri, Alain Barrat, and Vito Latora, “Simplicial models of social contagion,” *Nature communications* **10**, 2485 (2019).
- [17] Jacopo Grilli, György Barabás, Matthew J Michalska-Smith, and Stefano Allesina, “Higher-order interactions stabilize dynamics in competitive network models,” *Nature* **548**, 210 (2017).
- [18] Karel Devriendt and Piet Van Mieghem, “The simplex geometry of graphs,” *Journal of Complex Networks* **7**, 469–490 (2019).
- [19] Owen T Courtney and Ginestra Bianconi, “Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes,” *Physical Review E* **93**, 062311 (2016).
- [20] Giovanni Petri and Alain Barrat, “Simplicial activity driven model,” *Physical Review Letters* **121**, 228301 (2018).
- [21] Claude Berge, *Graphs and hypergraphs*, North-Holland Pub. Co. (American Elsevier Pub. Co, 1973).
- [22] Ernesto Estrada and Juan A Rodríguez-Velázquez, “Complex networks as hypergraphs,” *arXiv preprint physics/0505137* (2005).
- [23] Gourab Ghoshal, Vinko Zlatić, Guido Caldarelli, and Mark EJ Newman, “Random hypergraphs and their applications,” *Physical Review E* **79**, 066118 (2009).
- [24] Giovanni Petri, Paul Expert, Federico Turkheimer, Robin Carhart-Harris, David Nutt, Peter J Hellyer, and Francesco Vaccarino, “Homological scaffolds of brain functional networks,” *Journal of The Royal Society Interface* **11**, 20140873 (2014).
- [25] Giovanni Petri and Alain Barrat, “Simplicial activity driven model,” *Physical review letters* **121**, 228301 (2018).
- [26] Alice Patania, Giovanni Petri, and Francesco Vaccarino, “The shape of collaborations,” *EPJ Data Science* **6**, 18 (2017).
- [27] Owen T. Courtney and Ginestra Bianconi, “Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes,” *Phys. Rev. E* **93**, 062311 (2016).
- [28] Robert M May, “Will a large complex system be stable?” *Nature* **238**, 413 (1972).
- [29] Stefano Allesina and Si Tang, “Stability criteria for complex ecosystems,” *Nature* **483**, 205 (2012).
- [30] Louis M Pecora and Thomas L Carroll, “Master stability functions for synchronized coupled systems,” *Physical review letters* **80**, 2109 (1998).
- [31] Sidney Redner, *A guide to first-passage processes* (Cambridge University Press, 2001).
- [32] Jae Dong Noh and Heiko Rieger, “Random walks on complex networks,” *Physical review letters* **92**, 118701 (2004).
- [33] Mark EJ Newman, “A measure of betweenness centrality based on random walks,” *Social networks* **27**, 39–54 (2005).
- [34] Martin Rosvall and Carl T Bergstrom, “Maps of random walks on complex networks reveal community structure,” *Proceedings of the National Academy of Sciences* **105**, 1118–1123 (2008).
- [35] Vincenzo Nicosia, Manlio De Domenico, and Vito Latora, “Characteristic exponents of complex networks,” *EPL (Europhysics Letters)* **106**, 58005 (2014).
- [36] Jesús Gómez-Gardenes and Vito Latora, “Entropy rate of diffusion processes on complex networks,” *Physical Review E* **78**, 065102 (2008).
- [37] Giulia Cencetti, Federico Battiston, Duccio Fanelli, and Vito Latora, “Reactive random walkers on complex networks,” *Physical Review E* **98**, 052302 (2018).
- [38] Per Sebastian Skardal and Sabina Adhikari, “Dynamics of nonlinear random walks on complex networks,” *Journal of Nonlinear Science*, 1–26 (2018).
- [39] Malbor Asllani, Timoteo Carletti, Francesca Di Patti, Duccio Fanelli, and Francesco Piazza, “Hopping in the crowd to unveil network topology,” *Physical review letters* **120**, 158301 (2018).
- [40] Michele Starnini, Andrea Baronchelli, Alain Barrat, and Romualdo Pastor-Satorras, “Random walks on temporal networks,” *Physical Review E* **85**, 056115 (2012).
- [41] Julien Petit, Martin Gueuning, Timoteo Carletti, Ben Lauwens, and Renaud Lambiotte, “Random walk on temporal networks with lasting edges,” *Physical Review*

- E **98**, 052307 (2018).
- [42] Julien Petit, Renaud Lambiotte, and Timoteo Carletti, “Classes of random walks on temporal networks,” arXiv preprint arXiv:1903.07453 (2019).
 - [43] Manlio De Domenico, Albert Solé-Ribalta, Sergio Gómez, and Alex Arenas, “Navigability of interconnected networks under random failures,” *Proceedings of the National Academy of Sciences* **111**, 8351–8356 (2014).
 - [44] Federico Battiston, Vincenzo Nicosia, and Vito Latora, “Efficient exploration of multiplex networks,” *New Journal of Physics* **18**, 043035 (2016).
 - [45] Michael T Schaub, Austin R Benson, Paul Horn, Gabor Lippner, and Ali Jadbabaie, “Random walks on simplicial complexes and the normalized hodge laplacian,” arXiv preprint arXiv:1807.05044 (2018).
 - [46] Linyuan Lu and Xing Peng, “High-ordered random walks and generalized laplacians on hypergraphs,” in *International Workshop on Algorithms and Models for the Web-Graph* (Springer, 2011) pp. 14–25.
 - [47] Amine Helali and Matthias Löwe, “Hitting times, commute times, and cover times for random walks on random hypergraphs,” *Statistics and Probability Letters* **154**, 1–6 (2019).
 - [48] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf, “Learning with hypergraphs: Clustering, classification, and embedding,” in *Advances in neural information processing systems* (2007) pp. 1601–1608.
 - [49] Joan T. Matamalas, Sergio Gómez, and Alex Arenas, “Abrupt phase transition of epidemic spreading in simplicial complexes,” arXiv preprint:1910.03069v1 [physics.soc-ph] (2019).
 - [50] Bukyoung Jhun, Minjae Jo, and B. Kahng, “Simplicial sis model in scale-free uniform hypergraph,” arXiv preprint:1910.00375v1 [physics.soc-ph] (2019).
 - [51] Leonie Neuhäuser, Andrew Mellor, and Renaud Lambiotte, “Multi-body interactions and non-linear consensus dynamics on networked systems,” arXiv preprint arXiv:1910.09226 (2019).
 - [52] Guilherme Ferraz de Arruda, Giovanni Petri, and Yamir Moreno, “Social contagion models on hypergraphs,” arXiv preprint arXiv:1909.11154 (2019).
 - [53] Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi, “Echo chambers: Emotional contagion and group polarization on facebook,” *Scientific Reports* **6** (2016).
 - [54] M Belkin and P Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Computation* **15**, 1373–1396 (2002).
 - [55] We do not consider here hyperedges with size 1, because they correspond to isolated nodes, i.e. nodes that cannot take part to the examined process.
 - [56] Uthsav Chitra and Benjamin J. Raphael, “Random walks on hypergraphs with edge-dependent vertex weights,” arXiv preprint arXiv:1905.08287 (2019).
 - [57] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” in *Seventh International World-Wide Web Conference (WWW 1998)* (1998).
 - [58] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, *The PageRank Citation Ranking: Bringing Order to the Web.*, Technical Report 1999-66 (Stanford InfoLab, 1999) previous number = SIDL-WP-1999-0120.
 - [59] Floriana Gargiulo, Auguste Caen, Renaud Lambiotte, and Timoteo Carletti, “The classical origin of modern mathematics,” *EPJ Data Science* **5**, 26 (2016).
 - [60] Loc Hoang Tran, Linh Hoang Tran, Hoang Trang, and Le Trung Hieu, “Combinatorial and random walk hypergraph laplacian eigenmaps,” *International Journal of Machine Learning and Computing* **5**, 462 (462).
 - [61] Dheeru Dua and Casey Graff, “UCI machine learning repository,” (2017).
 - [62] In principle also the right eigenvectors can be used for classification purposes.
 - [63] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of Classification* **2**, 193–218 (1985).

ACKNOWLEDGEMENTS

T.C. would like to thank Floriana Gargiulo for her help in the data collection process. This research used resources of the “Plateforme Technologique de Calcul Intensif (PTCI)” (<http://www.ptci.unamur.be>) located at the University of Namur, Belgium, which is supported by the FNRS-FRFC, the Walloon Region, and the University of Namur (Conventions No. 2.5020.11, GEQ U.G006.15, 1610468 et RW/GEQ2016). The PTCI is member of the “Consortium des Équipements de Calcul Intensif (CÉCI)” (<http://www.ceci-hpc.be>).

Appendix A: About the projected network

A hypergraph is simple if each hyperedge does not contain any other hyperedge. We report in Fig. 8 two examples, the hypergraph H_1 with nodes $V = \{1, 2, 3\}$ and hyperedges $E_1 = \{1, 2\}$ and $E_2 = \{1, 3\}$ is simple because either $E_1 \not\subset E_2$ nor $E_2 \not\subset E_1$. On the other hand the hypergraph H_2 with nodes $W = \{a, b, c\}$ and hyperedges $E_3 = \{a, b, c\}$ and $E_4 = \{a, b\}$, is not simple because $E_4 \subset E_3$.

Once we build the projected network, $\pi(H_2)$, starting from the latter hypergraph we get a complete 3-clique, loosing thus information on the existence of hyperedge E_4 (see left panel Fig. 9). Hence, we cannot get back to H_2 , by inverting the construction, $\pi^{-1}\pi(H_2) \neq H_2$. A possible way to overcome this difficulty is to consider a weighted projection (see right panel Fig. 9) where edges inherit a weight counting the number of different hyperedges they belong to. Observe however that for large hyperedge sizes the inversion can be computationally costly because of the combinatorial structure of the problem.

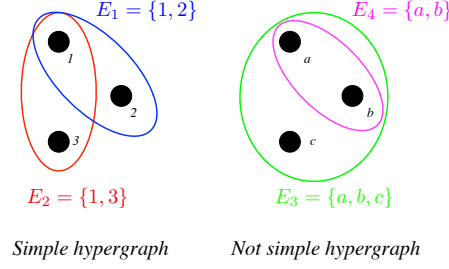


FIG. 8. **Simple and not simple hypergraphs.** The hypergraph shown on the left with nodes $V = \{1, 2, 3\}$ and hyperedges $E_1 = \{1, 2\}$ and $E_2 = \{1, 3\}$ is simple, while the one on the right, with nodes $W = \{a, b, c\}$ and hyperedges $E_3 = \{a, b, c\}$ and $E_4 = \{a, b\}$, is not simple.

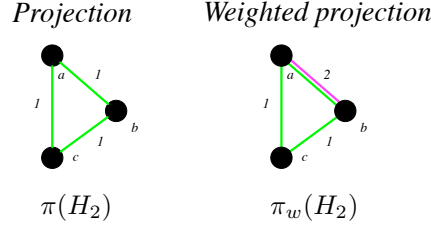


FIG. 9. **Weighted projection of hypergraphs.** We propose a standard projection (left panel) and a weighted projection (right panel) of the hypergraph H_2 shown in Fig. 8. In the latter case, the edge (a, b) has weight 2 because it belongs to two different hyperedges in H_2 .

Appendix B: Transition probability

The aim of this section is to provide some details about the calculation of the generalised transition probabilities which take into account the high-order structure of the hyperedges. To compute the transition probability to jump from i to j , we first count the number of nodes, excluding node i itself, belonging to the same hyperedge of i and j :

$$k_{ij}^H = \sum_{\alpha} (C_{\alpha\alpha} - 1) e_{i\alpha} e_{j\alpha} \quad i \neq j \text{ and } k_{ii}^H = 0 \quad \forall i; \quad (\text{B1})$$

namely for each hyper edge E_{α} we consider the number of its nodes minus one, i.e. $C_{\alpha\alpha} - 1$. Then, this quantity is added to k_{ij}^H if and only if $e_{i\alpha} = e_{j\alpha} = 1$, that is if and only if both i and j belong to E_{α} .

Secondly, we normalise this quantity by considering a uniform choice among the connected hyperedges. Hence, we obtain a first formula for the transition probability T_{ij} to jump from node i to node j :

$$T_{ij} = \frac{k_{ij}^H}{\sum_l k_{il}^H} = \frac{\sum_{\alpha} (C_{\alpha\alpha} - 1) e_{i\alpha} e_{j\alpha}}{\sum_l \sum_{\alpha} (C_{\alpha\alpha} - 1) e_{il} e_{l\alpha}} \quad (\text{B2})$$

so that $\sum_j T_{ij} = 1 \quad \forall i$.

The latter can be rewritten in an equivalent form, which allows one to draw a comparison with the transition probability for unbiased random walks on networks. Indeed. By recalling the definition of $C_{\alpha\beta} = (e^T e)_{\alpha\beta} = \sum_l e_{\alpha l}^T e_{l\beta} = \sum_l e_{l\alpha} e_{l\beta}$, we get $C_{\alpha\alpha} = \sum_l e_{l\alpha} e_{l\alpha}$ and then

$$\sum_{\alpha} C_{\alpha\alpha} e_{i\alpha} e_{j\alpha} = \sum_{\alpha} e_{i\alpha} C_{\alpha\alpha} e_{\alpha j}^T = (e \hat{C} e^T)_{ij}, \quad (\text{B3})$$

where $\hat{C}$ is a diagonal matrix: the diagonal of $\hat{C}$ coincides with that of C and its off-diagonal are identically equal to zero. This allows us to rewrite Eq. (B1) in a more compact way, Eq. (2) in the main text:

$$\begin{aligned} k_{ij}^H &= \sum_{\alpha} (C_{\alpha\alpha} - 1) e_{i\alpha} e_{j\alpha} = (e \hat{C} e^T)_{ij} - (e e^T)_{ij} \\ &= (e \hat{C} e^T)_{ij} - A_{ij} \quad \forall i \neq j, \end{aligned}$$

where in the last step we used the definition of the hyper adjacency matrix. We thus eventually get Eq. (3) as reported in the main text

$$T_{ij} = \frac{(e\hat{C}e^T)_{ij} - A_{ij}}{\sum_l k_{il}^H} = \frac{(e\hat{C}e^T)_{ij} - A_{ij}}{\sum_l (e\hat{C}e^T)_{il} - k_i^H},$$

where $k_i^H = \sum_l A_{il}$ is the hyperdegree of the node i .

Let us observe that this equation remains valid even for not simple hypergraphs. For instance using again the hypergraph H_2 shown in Fig. 8, where the hyperedge E_4 is properly included into E_3 , we get

$$k_{ab} = (E_3 - 1) + (E_4 - 1) = 2 + 1 \text{ and } k_{ac} = E_3 - 1 = 2,$$

and thus the following transition probabilities

$$T_{ab} = \frac{3}{5} \text{ and } T_{ac} = \frac{2}{5},$$

so the transition from a to b is 1.5 more probable than to c because a and b share two hyperedges. Among not simple hypergraphs, one has to account for the fact that hyperedges are repeated several times. The theory here proposed holds true also for weighted hyperedges.

1. Nonlinear transition rates

In deriving the transition rates Eq. (3), we assumed that the size of the hyperedge linearly correlates with the probability for the walker to perform a jump, one can of course relax this assumption and introduce nonlinear transition rates. In other words, one can add a bias in (B1) in the selection rule for a target node j , as operated by a walker sitting on node i . For example, one can posit:

$$k_{ij}^{(H,\gamma)} = \sum_{\alpha} (C_{\alpha\alpha} - 1)^{\gamma} e_{i\alpha} e_{j\alpha} \quad i \neq j \text{ and } k_{ii}^{(H,\gamma)} = 0 \quad \forall i. \quad (\text{B4})$$

In this way large hyperedges are even more favoured, if $\gamma > 0$, while the opposite happens if $\gamma < 0$, and we eventually get for the transition probabilities

$$T_{ij}^{(\gamma)} = \frac{k_{ij}^{(H,\gamma)}}{\sum_l k_{il}^{(H,\gamma)}}.$$

Clearly, other choices are possible but exploring further generalisations is left for future investigations.

Appendix C: Stationary solution and ranking

Given the transition probability stored in the matrix $\mathbf{T}$, we can obtain the analytical solution for the stationary state $\mathbf{p}^{(\infty)}$ defined by $\mathbf{p}^{(\infty)} = \mathbf{p}^{(\infty)}\mathbf{T}$. By recalling Eq. (7) as reported in the main text

$$p_j^{(\infty)} = \frac{\sum_l (e\hat{C}e^T)_{jl} - k_j^H}{\sum_{ml} [(e\hat{C}e^T)_{ml} - k_m^H]},$$

we can straightforwardly verify that it solves the fixed point equation for the governing dynamics. To this end one needs to plug the above equation into Eq. (6) and recall the definition (3) for the T_{ij}

$$\sum_j \left(\sum_l (e\hat{C}e^T)_{jl} - k_j^H \right) \left(\frac{(e\hat{C}e^T)_{ji} - A_{ji}}{\sum_l (e\hat{C}e^T)_{jl} - k_j^H} - \delta_{ji} \right) = \sum_j \left[\left((e\hat{C}e^T)_{ji} - A_{ji} \right) - \left(\sum_l (e\hat{C}e^T)_{il} - k_i^H \right) \right] = 0, \quad (\text{C1})$$

where the last step has been obtained by observing that $A_{ij} = A_{ji}$ and $(e\hat{C}e^T)_{ij} = (e\hat{C}e^T)_{ji}$.

As stated in the main text, random walks can be used to rank nodes. This is achieved by evaluating the asymptotic probability to get the walkers on the selected node: the larger the probability, the more central the node. The analytical

expression for $\mathbf{p}^{(\infty)}$ indicates that the ranking provided by the random walkers on the hypergraph is proportional to $d_i^H = \sum_j k_{ij}^H$, while it is well known that the ranking that follows the usual random walks on the projected network scales proportionally to the node degree, k_i . Two nodes, say i and j , are thus ranked differently by the two processes, if $k_i > k_j$ but $d_i^H < d_j^H$. As we will now show, the presence of high-order structures can induce a rankings inversion.

A simple example where this occurs is shown in the left panel of Fig. 10. The node i belongs to the intersection of three 2-hyperedges. Thus its degree (in the projected network) is given by $k_i = 3$. Moreover, $d_i^H = 3$, because, locally, the hypergraph reduces to a standard network; on the other hand the node j belongs to a 3-hyperedge, hence $k_j = 2$, because it is part of a 3-clique, but $d_j^H = 4$. Hence, $k_i > k_j$ but $d_i^H < d_j^H$. Nodes j will be consequently ranked above i using the generalised random walks on the hypergraph, while the contrary happens if one relies on random walks on the projected network.

The above construction can be readily generalised, as shown by the example presented on the right panel of Fig. 10. Here, i belongs to a 3-hyperedge and to two 2-hyperedges, hence $k_i = 4$ and $d_i^H = 6$; node j instead belongs to a 4-hyperedge, thus $k_j = 3$ and $d_j^H = 9$. So again $k_i > k_j$ while $d_i^H < d_j^H$.

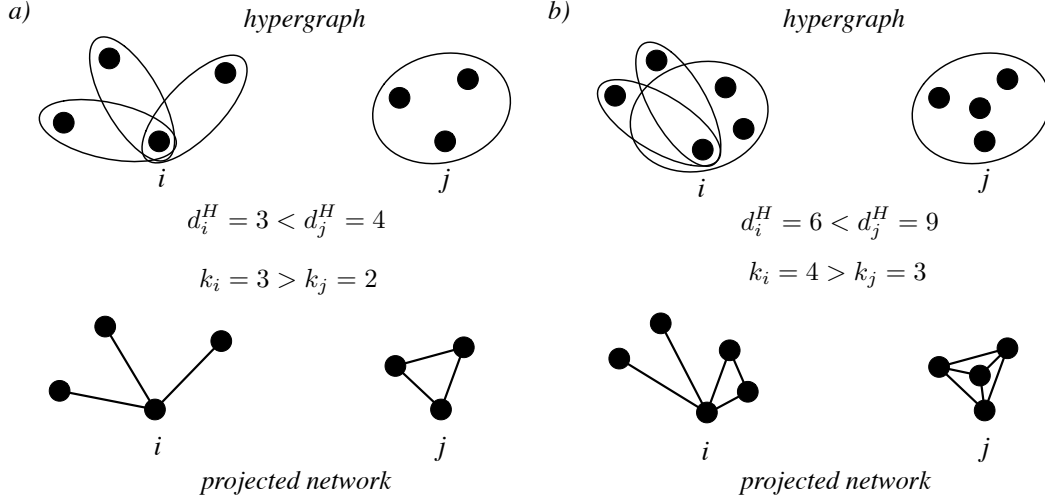


FIG. 10. **Examples with ranking inversion.** We propose two typical examples of high-order structures that locally produce two different rankings. On the left panel an example involving three 2-hyperedges and one 3-hyperedge, while on the right panel the case with one 3-hyperedge and two 2-hyperedges compared with a 4-hyperedge. In both cases the first configuration will be ranked above the second one, when using the random walks on hypergraphs, while the opposite holds when the random walkers run on the projected networks.

Appendix D: Heterogeneity of stationary solution

The stationary solution that we obtain from random walk on a hypergraph is very different from the one we can get from the corresponding projected network, the first one being more sensitive to the organisation in groups. The heterogeneity of the state, i.e. the difference among the occupation probability of the different nodes at equilibrium can be quantified by making use of the Gini coefficient.

Fig. 11 reports on the ratio between the coefficient G computed for the hypergraph and for the projected network of Fig. 2 of the main text, at varying m , the size of the star, and k , the size of the clique.

Fig. 12 instead shows the heterogeneity for the model which goes from a 1D lattice to a fully connected network, by subsequently adding the links (see Fig. 4 of the main text). The red points show the Gini coefficient for the hypergraph, while the green ones are plotted for the projected network, at varying l , the number of links in the graph.

From the results presented in these Figures one can appreciate that the Gini coefficient associated to the stationary solution for the random walk on the hypergraph is always larger than the same quantity computed for the random walk on the projected network. This implies thus that the distribution of walkers on the hypergraph is more heterogeneous than for the projected network.

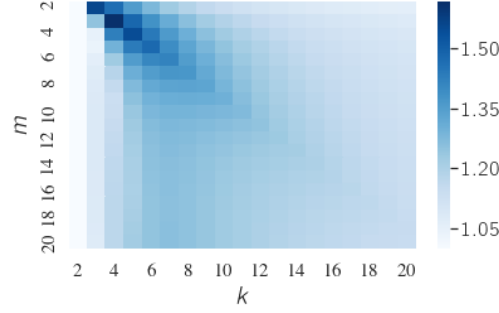


FIG. 11. **Star-clique model.** Ratio between the Gini coefficient of the stationary state on the hypergraph and the projected network, at varying of the size of the clique (k) and of the star (m).

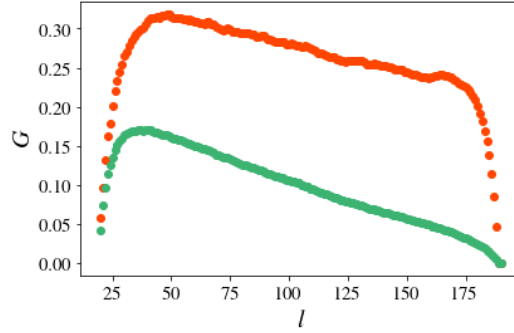


FIG. 12. **Lattice to fully-connected model** Gini coefficient for the stationary state of the random walk on the hypergraph (red) and on the corresponding projected network (green).

Appendix E: Co-authorship networks from arXiv

The collaboration network is one of the most representative examples of hypergraph; nodes are authors and hyperedges are groups of authors that collaborated to accomplish a task, e.g. write a scientific paper. For this reason we decided to applied the method that we developed to the co-authorship networks extracted from the online preprints platform *arXiv* and hence analyse the nodes ranking obtained using the two processes.

In this section we report some results for the co-authorship hypergraph for the subdomains of *arXiv*, since their existence up to 2018 included (second column Table I). In each subdomain, we gathered all the papers and then extracted the authors names, so creating a hyperedge whose nodes are the authors. We thus obtain a set of nodes $V^{(1)}$ and hyperedges $E^{(1)}$, and also the edges of the associated projected network, $E_q^{(1)}$. Such quantities are reported in parentheses in the third, fourth and fifth column of the Table I. Once the hypergraph has been built we identify the *largest connected component* that will contain the nodes $V^{(cc)}$; then we identify all the maximal, i.e. not properly contained in any other larger hyperedge, and unique hyperedges $E^{(cc)}$ and the edges of the associated projected network $E_q^{(cc)}$. Columns 3-4 and 5 of the Table show such values. Finally, we compute the largest hyperedge and the largest node degree in the maximal connected component (columns 6 and 7). For instance in the arXiv-cs there is a node that belongs to a hyperedge of size 65 and that is linked to other 406 nodes: this means that this researcher has signed a paper with 64 other researchers and in total he/she had 406 different collaborators with whom he/she has written a paper. Let us also observe that because of the maximality and uniqueness assumptions, we do not know if he/she has co-authored other papers with a subset of the 64 scholars. Moreover, because we used unweighted networks, we also cannot estimate how many papers he/she wrote with her 406 collaborators. Let us recall that the need for the maximality and uniqueness is only to compare the results with the projected network, while our method works also without these assumptions.

Authors and articles in each subdomain follow different “rules” and “habits” of publication and writing papers. However, the distribution of node degrees, i.e. number of different collaborators per author, and of hyperedges size, i.e. number of co-authors in papers, exhibit quite similar shapes across the domains, as e.g. broad tails (see Fig. 13

arXiv	period	nodes	hyperedges	links	$\max E_\alpha $	$\max k_i$
astro-ph	1992-2018	185579 (195729)	136918 (201270)	4602315 (4617912)	81	2732
cond-mat	1992-2018	221415 (243749)	141611 (207939)	1520895 (1551863)	63	1426
cs	1993-2018	136146 (187689)	84184 (139334)	534462 (607560)	65	406
econ	2017-2018	113 (1147)	63 (612)	214 (1295)	5	36
gr-qc	1992-2018	32088 (40316)	25321 (45378)	216355 (228811)	80	511
hep-ex	1992-2018	48460 (55634)	12310 (23249)	1418268 (1435372)	83	1228
hep-lat	1992-2018	10275 (12483)	7439 (14143)	85194 (87254)	72	346
hep-ph	1992-2018	62885 (70324)	50403 (86150)	814746 (823705)	74	1244
hep-th	1992-2018	41814 (51045)	42410 (74136)	144710 (154737)	57	206
math	1992-2018	112203 (159595)	106583 (194312)	279891 (313402)	60	336
nlin	1993-2018	19491 (30445)	12428 (23503)	52089 (64890)	46	312
physics	1996-2018	188142 (240866)	68805 (116611)	1859156 (1950143)	80	891
q-bio	2003-2018	23630 (45103)	9926 (21191)	93127 (142136)	54	176
q-fin	2008-2018	3136 (8721)	2155 (6042)	6851 (13078)	11	66
stat	2008-2018	39422 (57955)	23377 (39366)	130665 (158435)	65	228

TABLE I. **Some figures for the arXiv subdomains.** The first column shows the subdomain of the arXiv server, while the second one stands for the period of time for which we have extracted the information. The columns 3, 4 and 5 display respectively the number of nodes, the number of maximal unique hyperedges and the number of links in the largest connected component, while in parenthesis we show the same values for the whole hypergraph/network. In column 6, we report the size of the largest hyperedge and in the 7th the maximum degree.

for the degree distribution and - see Fig. 14 - for the hyperedges sizes distribution).

As already stated the random walk on the hypergraph gives more relevance to the size of the hyperedge, i.e. on the number of co-authors, while the same process on a network emphasises the number of different collaborators. Let us remember that we hereby considered unweighted hypergraphs and networks. We can thus use these approaches to distinguish the different “publication habits” in the considered subdomains. To this aim we first normalise the stationary probabilities $p_i^{(\infty)}$ for the hypergraph and $q_i^{(\infty)}$ for the projected network, with respect to their maximum value, to be able to compare sets containing different amount of data, and then we report in the plane with coordinates $(q_i^{(\infty)} / \max_j q_j^{(\infty)}, p_i^{(\infty)} / \max_j p_j^{(\infty)})$, the scatter plot of the data (each point is an author in the maximal connected component of the hypergraph), separated into different subdomains (see Fig. 15).

If the computed rankings were (almost) the same, the data would (almost) lie on the main diagonal; deviation from this, results in novel information conveyed by the random walk on the hypergraph. Beside the region delimited by $q_i^{(\infty)} / \max_j q_j^{(\infty)} \leq 1/2$ and $p_i^{(\infty)} / \max_j p_j^{(\infty)} \leq 1/2$, associated to authors having written few articles (low degree) and in small group, we identify three interesting zones associated (roughly speaking) to the squares: $[1/2, 1] \times [0, 1/2]$ (bottom right), $[0, 1/2] \times [1/2, 1]$ (top left) and $[1/2, 1] \times [1/2, 1]$ (top right). Authors in the top right square are top ranked in both processes: they have hence written a large number of papers with different collaborators, i.e. large degree, but also they have participated to a relevant number of papers with many co-authors, i.e. large hyperedge size. Scholars in the bottom right square are better ranked by the random walk on the network. This means that they have written several papers but with a small number of co-authors (see e.g. panel physics in Fig. 15). Finally scholars in the top left square behave in the opposite way: they have participated to a small number of papers but written by many authors (see e.g. panels gr-qc, q-bio or stat in Fig. 15).

Appendix F: The zoo UCI database

The zoo dataset from the UCI Machine Learning Depository [61], contains 101 animals, each one endowed with 15 boolean features, whose value is thus yes/not, e.g. hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic and catsize. There is also a further class that reports on the number of legs, i.e. 0, 2, 4, 6, 8. To homogenise the dataset we decided to introduce five new boolean features to replace the latter one; the new ones being: “has 0 legs”, “has 2 legs”, “has 4 legs”, “has 6 legs”, “has 8 legs”. The dataset is manually annotated, hence for each animal we have the right class it belongs to, e.g. Mammal, Bird, Reptile, Fish, Amphibian, Bug and Invertebrate.

This dataset has been created to provide a benchmark for machine learning tools, to test their capacity to correctly

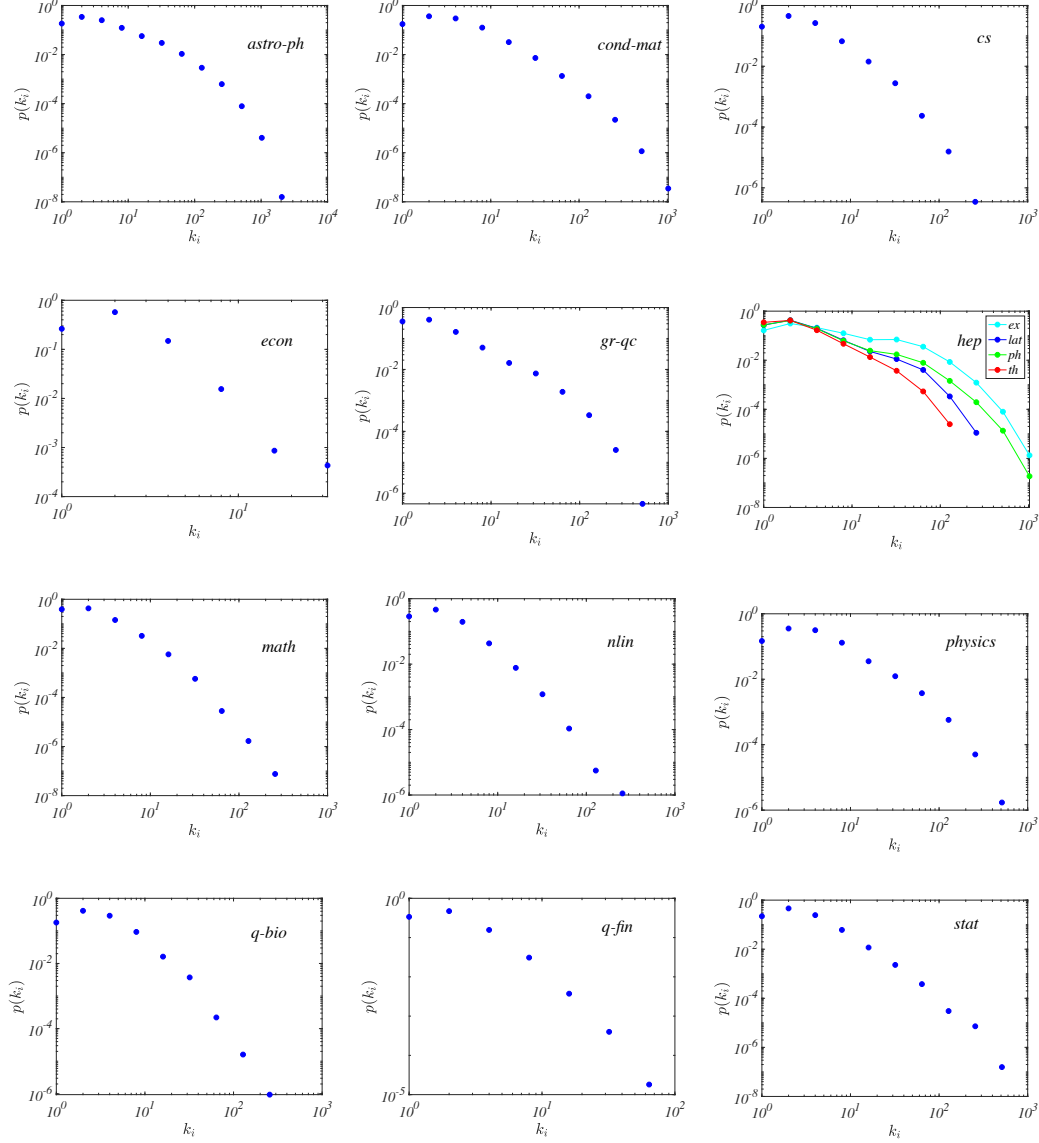


FIG. 13. **Degree distribution.** We report for the arXiv subdomains the probability distribution of node degrees, $p(k_i)$, associated to the maximal connected component. In all the cases, we observe a broad distribution; Notice that the arXiv-econ has a relatively small number of papers and authors because of its young age (2017-2018) and thus also the maximal degree, i.e. number of papers written by an author, is quite small.

assign each animal to the right class based on the associated features.

Animals are the nodes of the hypergraph and the features are the hyperedges, hence all animals sharing the same feature are put in the same hyperedge. The projected network is obtained by making a complete clique from each hyperedge, that is to create a link between all the nodes sharing the same property. Let us observe that this can also be seen as the projection of the bipartite network where there are two kinds of nodes, animals and features, each one linked only to nodes of the other kind. We thus computed the spectrum of the hypergraph Laplacian and the one for the projected network and we ranked eigenvalues in ascending order, being 0 the smallest one. We then accordingly rename the associated left eigenvector and we use the first few to embed the data set in small dimensional Euclidean space. Results reported in Fig. 16 visually show that classification performances are significantly worse as

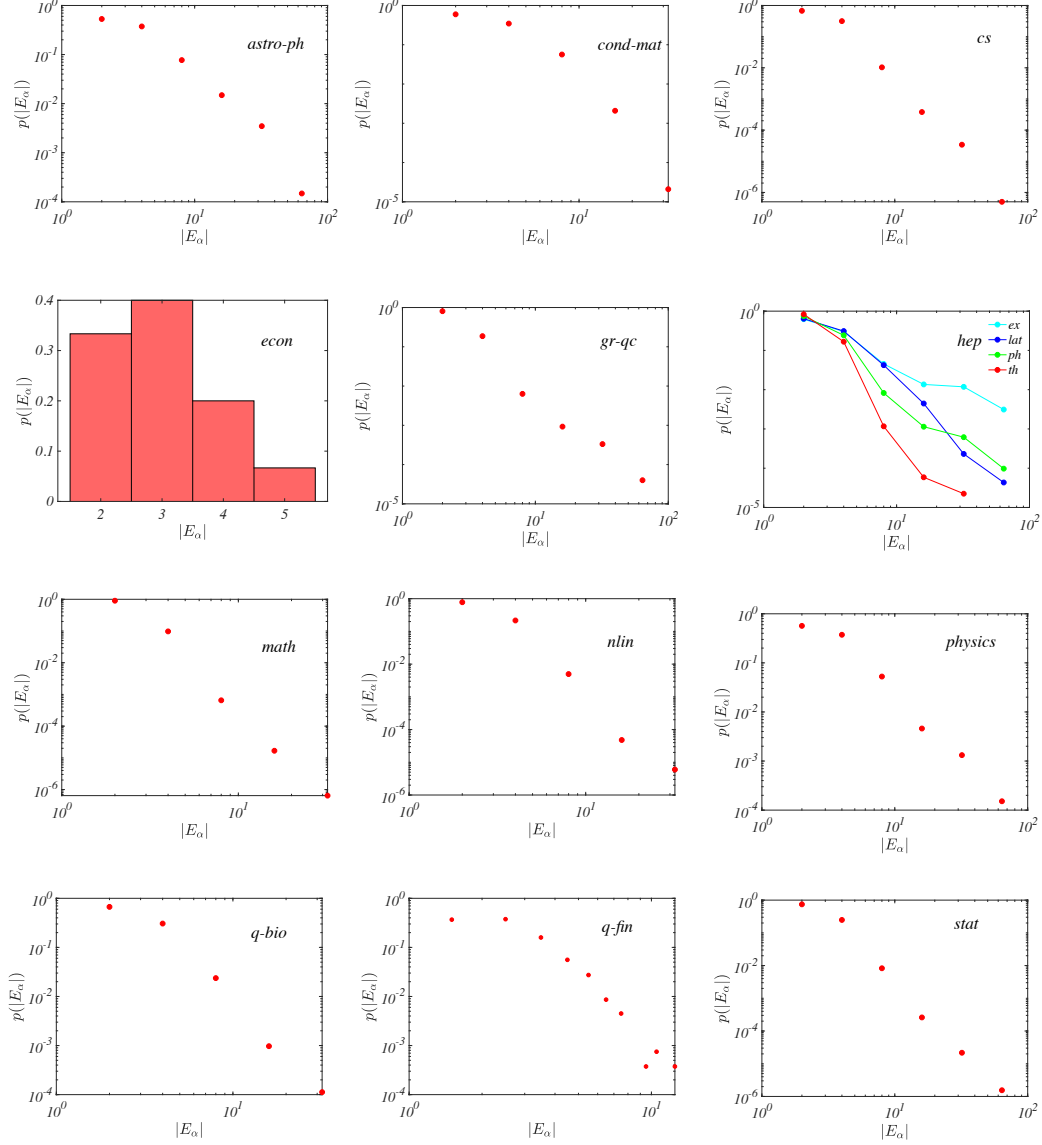


FIG. 14. **Hyperedges size distribution.** We report for the arXiv subdomains the probability distribution of hyperedges size, $p(|E_\alpha|)$, associated to the maximal connected component. In all the cases we observe a broad distribution, except for the arXiv-econ for which the number of papers and authors is relatively small because of its young age (2017-2018) and thus also the maximal hyperedge size, i.e. number of co-authors of a paper, is quite small. For this reason we report data in the form of a histogram.

those obtained when preserving the high order information (see Fig. 7 in the main text). In Fig. 16 we report 2D projections, but similar conclusion (as testified by the quantitative ARI scores) are obtained for a 3D embedding of the data. In particular, this method is less sensitive to the differences among nodes and as a consequence, multiple nodes do overlap. Moreover, it is evident that, while some nodes are correctly clustered (like the yellow ones), the others appear confusingly mixed together (see the magenta triangle, which is at the top of a pile of differently coloured symbols).

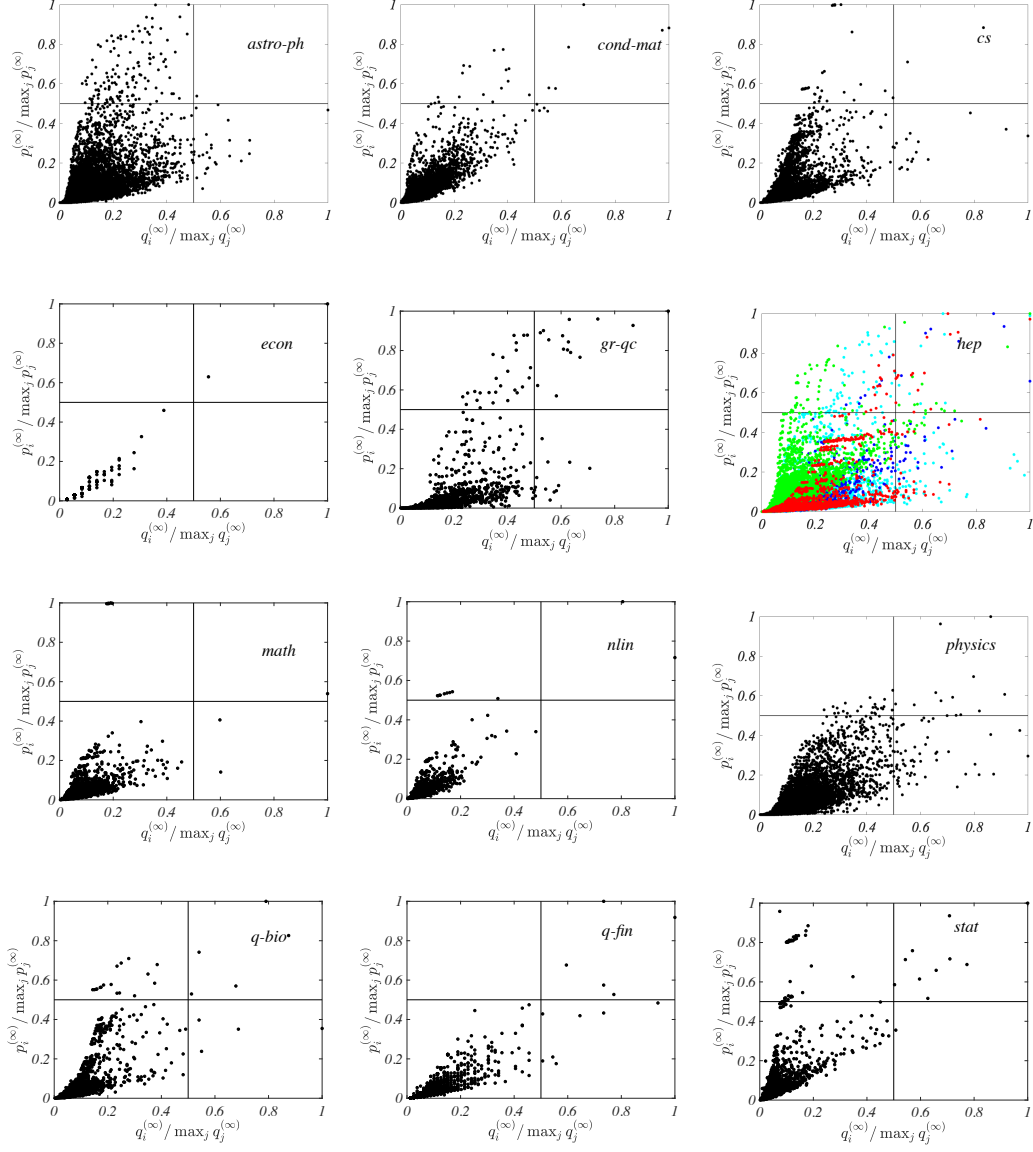


FIG. 15. **Comparing the rankings in the arXiv community.** We report the scatter plot of the normalised rankings obtained with the RW on network, $q_i^{(\infty)}$, and the one computed using the random walk on hypergraphs, $p_i^{(\infty)}$.

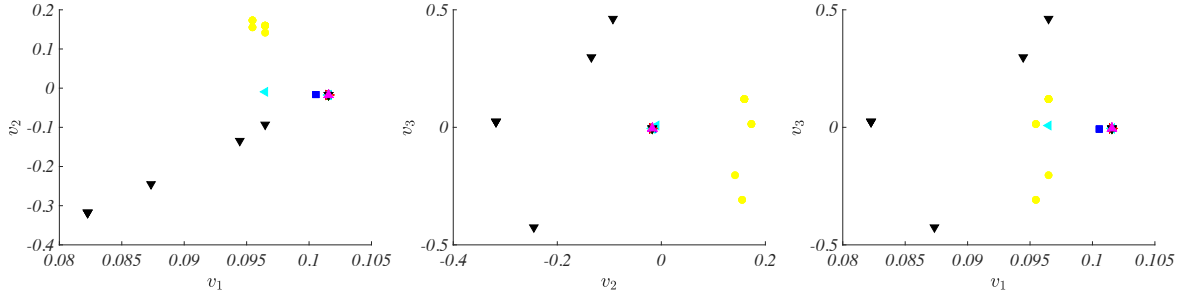


FIG. 16. **Classification of the animals according to their features using the projected network.** We report three $2D$ embedding of the zoo dataset, namely using the first three eigenvector of the random walk Laplacian computed for the projected network. Each combination colour/symbol refer to a know class and one can visually appreciate the resulting clusters.