From unbiased to maximal-entropy random walks on hypergraphs

Pietro Traversa [1,2,3] Guilherme Ferraz de Arruda [3] and Yamir Moreno [1,2,3]
[1]*Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, 50018 Zaragoza, Spain*
[2]*Department of Theoretical Physics, University of Zaragoza, 50018 Zaragoza, Spain*
[3]*CENTAI Institute, 10138 Turin, Italy*

(Received 16 June 2023; revised 21 February 2024; accepted 11 April 2024; published 13 May 2024)

Random walks have been intensively studied on regular and complex networks, which are used to represent pairwise interactions. Nonetheless, recent works have demonstrated that many real-world processes are better captured by higher-order relationships, which are naturally represented by hypergraphs. Here we study random walks on hypergraphs. Due to the higher-order nature of these mathematical objects, one can define more than one type of walks. In particular, we study the unbiased and the maximal entropy random walk on hypergraphs with two types of steps, emphasizing their similarities and differences. We characterize these dynamic processes by examining their stationary distributions and associated hitting times. To illustrate our findings, we present a toy example and conduct extensive analyses of artificial and real hypergraphs, providing insights into both their structural and dynamical properties. We hope that our findings motivate further research extending the analysis to different classes of random walks as well as to practical applications.

DOI: 10.1103/PhysRevE.109.054309

I. INTRODUCTION

One of the main frameworks used to study and describe complex systems is network theory, which has been greatly developed during the past two decades. Despite this development and its success in representing and understanding a plethora of real systems, most network methods are constrained to systems with pairwise interactions. Recently, attention was raised to higher-order interactions, arguing that rich data are revealing more complex relationships among nodes that may not be captured by models based on pairwise interactions [1,2]. This claim has been supported through a series of works. In Refs. [3,4], through linear stability analysis, hypergraphs' stability was evaluated, emphasizing some of the key differences and similarities between graphs and hypergraphs. From a modeling point of view, the need to consider higher-order interactions has also been recently reinforced by theoretical approaches involving phenomena, such as social contagion [5–7], evolutionary game dynamics [8], synchronization [9–11], and random walks [12–15], the latter being the main focus of this contribution.

The study of hypergraphs is also important in other fields of research beyond physics or mathematics. For instance, in the area of machine learning research, hypergraphs have been used in classification, clustering, and embedding techniques [12,15]. Moreover, a hypergraph convolutional neural network (HGCN) has been proposed [16,17]. Of relevance for the present work, the authors of Ref. [13] hypothesized that machine learning algorithms could benefit from further studies of hypergraphs and, more specifically, on random walks. The aforementioned list of works is not exhaustive but it shows the increasing interest that data-rich and higher-order approaches are attracting. Nonetheless, despite this interest, the study of higher-order systems is arguably in its infancy. Thus, it is of utmost importance to build most of the theoretical tools that will allow us to study and develop more complex and realistic processes in the near future. In this context, a random walk process is simple enough to provide new insights and results while capturing this type of system's higher-order nature.

Random walks are paradigmatic, being interesting both from theoretical and practical points of view. They are probably the most fundamental stochastic processs [18], serving as a model for a variety of phenomena, including diffusion, social interactions, and opinions [18], and providing handy insights that can be used in many different contexts. In network theory, this process and its variants are reasonably well studied [18–23]. However, in hypergraphs, this process just got recent attention due to its applications in machine learning and physics. As Carletti *et al.* mentioned in Ref. [14], probably the first random walk defined on hypergraphs was proposed by Zhou *et al.* [12]. In this paper, the authors were concerned about using such a process in machine learning techniques such as clustering, classification, and embedding. In Ref. [15], also focusing on machine learning applications, Koby Hayashi *et al*. proposed a clustering framework using hypergraph-structured data-based and random walks.

Here we are interested in the physical aspects and insights that random walks can bring to the analysis of hypergraphs' structure and dynamics. We focus on two classes of random walks, the unbiased random walk (URW)[1] and the maximal entropy random walk (MERW). Formally, the walker in the URW makes a succession of uniformly random decisions using only local information (the node degree). On the other hand, in the MERW, the walker uniformly chooses a path that maximizes the entropy among all the possible paths of fixed

[1]In the literature, this class of random walks is also called a general random walk.

length [19,24]. The construction of such a process requires complete knowledge about the structure, here expressed by the leading eigenvector of the adjacency matrix. Comparatively, the first is a local process while the second is nonlocal. In network analysis, MERW was applied to the analysis of networks with limited information [25], community detection [26], link predictions [27], and on the definition of centrality measures [28]. However, due to the higher-order nature of hypergraphs, more than one type of step can be defined for each class of random walk, depending on the adjacency matrix definition. One can define a random walk using the adjacency matrix defined by Banerjee [29] or by Battiston *et al.* [30]. Carletti *et al.* [31] proposed a whole spectrum of possible adjacency matrices, and thus random walks, using a free parameter. The absence of a unique definition of the adjacency matrix is the main difference between hypergraphs and pairwise graphs. Under such a varied possibility of choices, in this paper, we highlight the difference between the types of random walks.

Specifically, one of our main contributions is the generalization of maximal entropy random walks to hypergraphs. From a theoretical point of view, we define the probability transition matrix and the hitting times for such type of random walk defined on top of the most common hypergraph projections present in the literature. We establish our derivations from the observation that random walks on complex networks are equivalent to the same processes on hypergraphs, up to some small details and constraints [13–15]. Another major contribution of our work is the numerical experiments. We provide a series of examples, ranging from a small toy example emphasizing the key peculiarities of random walks in hypergraphs, to artificial and real cases. For the artificial experiments, we remark that we were able to evaluate heterogeneity in both the cardinality distribution and the degree distribution.

Our paper is divided as follows. In the next section, we define hypergraphs and discuss their representations. Next, in Sec. III, we define the random walks, focusing on the unbiased random walks, in Sec. III A, and the maximal entropy, in Sec. III B. Complementary, in Sec. III C, we discuss the particularities of uniform and regular hypergraphs, while in Sec. IV, we numerically evaluate synthetic and real hypergraphs. Specifically, in Sec. IV A 1, we present a simple toy example that allows us to discuss the main differences among the classes of random walks studied here, and in Sec. IV A 2 we describe the model we used to generate the artificial hypergraphs. In Sec. IV A 3, we numerically compare the hitting times for the four classes of random walks on different artificial hypergraphs and, in Sec. IV B, on a real hypergraph. To summarize, in Sec. V, we present a short discussion about our findings and their relation with the literature, followed by the conclusions.

II. HYPERGRAPHS: DEFINITIONS AND REPRESENTATION

A hypergraph, $\mathcal{H} = \{\mathcal{V}, \mathcal{E}\}$, is a mathematical structure that extends the concept of a graph. It is composed of a set of nodes, $\mathcal{V} = \{v_i\}$, and a multiset of hyperedges $\mathcal{E} = \{e_j\}$, where e_j is a nonempty subset of $\mathcal{V}$ with arbitrary cardinality $|e_j|$. The maximum cardinality of the hyperedges is given by $e_{\max} = \max(|e_j|)$. The number of nodes in the hypergraph is denoted as $N = |\mathcal{V}|$ and the number of hyperedges as $M = |\mathcal{E}|$. We also denote $\mathcal{E}_i$ as the multiset of hyperedges that contain the node i. A hypergraph is considered to be simple if there are no repeated hyperedges, i.e., if $\mathcal{E}$ is a set rather than a multiset. If $e_{\max} = 2$, then the hypergraph reduces to a standard graph, whereas one recovers a simplicial complex if, for each hyperedge with $|e_j| > 2$, its subsets are also contained in $\mathcal{E}$. The degree of node i, k_i, is defined as the number of hyperedges that contain this node. Conversely, the number of neighbors of node i, n_i is defined as the number of unique nodes that share a hyperedge with i. We remark that these two concepts coincide in graphs, but they might be different in hypergraphs.

A hypergraph $\mathcal{H}$ can be represented by the *incidence matrix*, $\mathcal{I} \in \mathbb{R}^{N\times M}$, which is defined as

$$\mathcal{I}_{ij} = \begin{cases} 1 & \text{if } v_i \in e_j \\ 0 & \text{if } v_i \notin e_j \end{cases}. \tag{1}$$

From this representation, we can define the *counting adjacency matrix* [14,30], $\mathbf{A}^{\text{count}} \in \mathbb{R}^{N\times N}$, given as

$$\mathbf{A}^{\text{count}} = \mathcal{I}\mathcal{I}^T - \mathbf{D}, \tag{2}$$

where $\mathbf{D} = \text{diag}(k_i)$. Each element $\mathbf{A}^{\text{count}}_{ij}$ is the number of hyperedges shared by nodes i and j. The *normalized adjacency matrix* [29], $\mathbf{A}^{\text{norm}} \in \mathbb{R}^{N\times N}$, is defined as

$$\mathbf{A}^{\text{norm}}_{ik} = \sum_{\substack{e_j\in\mathcal{E} \\ v_i, v_k \in e_j \\ v_i \neq v_k}} \frac{1}{|e_j| - 1}. \tag{3}$$

In this formulation, the degree of node i is computed as $k_i = \sum_{k=1}^{N} \mathbf{A}^{\text{norm}}_{ik}$. We remark that in the original definition in Ref. [29], this matrix was not referred to as the normalized adjacency matrix. However, to emphasize its differences with respect to the counting adjacency matrix, we have referred to it as such here.

Both matrices can be considered as projections of the hypergraph onto a graph, where hyperedges are represented as cliques. The information about the hyperedges is retained in the edge weights of the projected graph, although the counting and normalized projections assign different meanings to these weights.

III. RANDOM WALKS

A *walk* [29] $v_{i_0} - v_{i_l}$ of length l between two vertices $v_{i_0}, v_{i_l} \in \mathcal{V}$ in a hypergraph $\mathcal{H}$ is an alternating sequence $v_{i_0} e_{j_1} v_{i_1} e_{j_2} \ldots v_{i_{l-1}} e_{j_l} v_{i_l}$ of distinct pairs of vertices and hyperedges, such that $v_{i_{k-1}}, v_{i_k} \in e_{j_k}$ for $k = i, \ldots, l$. A *step* is a walk of length one. A *random walk* (RW) is a stochastic process, which describes a walk consisting of a succession of random steps. Here we focus on Markovian random walks, where the next step is dependent only on the current state of the process. At each time step t, the random walk process proceeds as follows:

(i) pick an edge $e \in \mathcal{E}_{i_t}$ with some probability $p_{v_{i_t}}(e)$,
(ii) pick a vertex $v \in e$ with some probability $p_e(v)$,
(iii) move to $v_{i_{t+1}} = v$ at time $t + 1$.

TABLE I. Definition of the fundamental matrices used for the two types of random walks considered: The PRW and the HORW. The probability transition matrices for the unbiased and maximal entropy cases are analyzed in detail in Secs. III A and III B, respectively.

Type	PRW	HORW
Adjacency	$\mathcal{A} = \mathbf{A}^{\text{count}}$	$\mathcal{A} = \mathbf{A}^{\text{norm}}$
Diagonal	$\mathcal{D} = \text{diag}(\sum_{k=1}^{N} \mathbf{A}_{ik}^{\text{count}})$	$\mathcal{D} = \text{diag}(\sum_{k=1}^{N} \mathbf{A}_{ik}^{\text{norm}})$
Laplacian	$\mathcal{L} = \mathcal{D} - \mathbf{A}^{\text{count}}$	$\mathcal{L} = \mathcal{D} - \mathbf{A}^{\text{norm}}$
Probability transition (URW)	$\mathcal{P}^{\text{URW}} = \mathcal{D}^{-1}\mathcal{A}$	
Probability transition (MERW)	$\mathcal{P}_{ij}^{\text{MERW}} = \frac{\mathcal{A}_{ij}}{\lambda}\frac{\psi_j}{\psi_i}$	

The probabilities associated with choosing an edge and a vertex may vary depending on the type of random walk considered. For instance, the walker at node v_{i_t} first chooses uniformly a hyperedge e, then, inside this hyperedge, it chooses uniformly a different node $v_{i_{t+1}} \in e \setminus \{v_{i_t}\}$. This process is related to the graph-projection given by Eq. (3). We denote this as the higher-order step and the generated process as a higher-order random walk (HORW). This type of process was initially studied in Refs. [13,15,29] and also explored as an unbiased random walk. Another possible choice is to consider the next step probability for a walker at node v_{i_t} to be proportional to the number of hyperedges between v_{i_t} and $v_{i_{t+1}}$. We call this the projected step and the generated process as a projected random walk (PRW). In this case, the walk takes place in the graph-projection defined by Eq. (2). This type of higher-order step was initially explored in Ref. [14] in the form of unbiased random walks. We stress that this is not a two-event process, as we have no information on which hyperedge the walk is moving through, but only on how many hyperedges two nodes share. As a consequence, this type of higher-order step may not be sensitive to some higher-order structures. Note that, in pairwise relations, both formulations fall into the standard definitions of random walks in graphs.

As we defined a random walk where the nodes are the states, all our quantities of interest can be derived in terms of $N \times N$ matrices. This argument was formally provided in Ref. [13], Theorem 16, where, by using the time reversibility property of Markov chains, the authors proved that the nonlazy[2] random walk on a hypergraph is equivalent to the nonlazy random walk on a graph, provided that there are trivial vertex weights. Hence, using the mapping between random walks in hypergraphs and graphs (under the constraints mentioned above), the theory of random walks in weighted networks can be reinterpreted in our context.

In the following sections, we define two types of random walks: unbiased and maximal entropy random walks. We can also use the two types of steps previously discussed, the projected and higher-order steps for each class of random walks. Thus, to make our notation lighter, we define the adjacency, $\mathcal{A}$, the Laplacian, $\mathcal{L}$, and the probability transition matrix, $\mathcal{P}$, accordingly. Their definitions are given in Table I. We also use superscripts to denote the type of step and random walks when necessary.

[2]The lazy random walk allows the walker to stay at the current node, while the nonlazy random walk does not allow it. In other words, $\mathcal{P}_{ii} \geqslant 0$ for the lazy and $\mathcal{P}_{ii} = 0$ for the nonlazy.

A. Unbiased random walks

We analyze the URW on hypergraphs with the projected and higher-order step. In this process, given a step definition, there is no bias towards a given direction. Since random walks on hypergraphs can be mapped onto random walks on graphs, we can use known literature results to write analytically the expressions for the stationary distribution and mean hitting times. From the Markovian formulation, the stationary distribution is expressed as

$$\pi^T = \pi^T \mathcal{P}^{\text{URW}}, \tag{4}$$

where $\mathcal{P}^{\text{URW}} = \mathcal{D}^{-1}\mathcal{A}$ is the probability transition matrix, $\mathcal{D} = \text{diag}(\sum_{j=1}^{N} \mathcal{A}_{ij})$ is a diagonal matrix, and π is the normalized eigenvector associated to the leading eigenvalue, $\sum_{i=1} \pi_i = 1$. Explicitly, this distribution is given as

$$\pi_i = \frac{\sum_{j=1}^{N} \mathcal{A}_{ij}}{\sum_{\substack{i=1 \\ j=1}}^{N} \mathcal{A}_{ij}}. \tag{5}$$

Aside from the stationary distribution, other quantities of interest are the hitting times. We denote by $0 = \sigma_1 < \sigma_2 \leqslant \cdots \leqslant \sigma_N$ the N eigenvalues of the Laplacian matrix, $\mathcal{L}$, and by $\mu_1, \mu_2, \ldots, \mu_N$ their corresponding normalized eigenvectors, whose components are $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iN})^\top$ for $i = 1, 2, \ldots, N$. For a random walker starting from node v_i, the expected time to hit node v_j is expressed as [20,23]

$$T_{ij}^{\text{URW}} = \sum_{z=1}^{N} k_z \sum_{k=2}^{N} \frac{1}{\sigma_k}\left(\mu_{ki}\mu_{kz} - \mu_{ki}\mu_{kj} - \mu_{kj}\mu_{kz} + \mu_{kj}^2\right). \tag{6}$$

Complementary, assuming that the node v_j is the target node, the partial mean hitting time is given as [20,23]

$$T_j^{\text{URW}} = \frac{N}{N-1} \sum_{k=2}^{N} \frac{1}{\sigma_k}\left(2E \times \mu_{kj}^2 - \mu_{kj} \sum_{z=1}^{N} k_z \mu_{kz}\right), \tag{7}$$

where $E = \sum_{\substack{i=1 \\ j=1}}^{N} \mathcal{A}_{ij}$. Finally, the global mean first passage time can be obtained as [20,23]

$$\langle T^{\text{URW}} \rangle = \frac{2E}{N-1} \sum_{k=2}^{N} \frac{1}{\sigma_k}. \tag{8}$$

Interestingly, the hitting times are fully characterized using only the spectral properties of the Laplacian matrix, while the stationary distribution depends only on the probability transition matrix.

B. Maximal entropy random walk

Another interesting case is the maximal entropy random walk, where the walker is biased towards the direction that maximizes the entropy of possible trajectories.

This type of random walk on graph was studied in Ref. [19] where the probability transition matrix is defined as

$$\mathcal{P}_{ij}^{\mathrm{MERW}} = \frac{\mathcal{A}_{ij}}{\lambda_1}\frac{\psi_{1j}}{\psi_{1i}}. \tag{9}$$

We denote by $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_N$ the eigenvalues of $\mathcal{A}$ and by $\psi_1 \geqslant \psi_2 \geqslant \cdots \geqslant \psi_N$ the associated eigenvectors, whose normalized components are $\psi_i = (\psi_{i1}, \psi_{i2}, \ldots, \psi_{iN})^\top$ for $i = 1, 2, \ldots, N$. The stationary distribution is obtained as

$$\phi_i = \psi_{1i}^2, \tag{10}$$

where the normalization $\sum_{i=1} \psi_{1i}^2 = 1$ must hold.

Mathematically, the expected time to hit v_j, starting from v_i is obtain in Ref. [21] as

$$T_{ij}^{\mathrm{MERW}} = \frac{1}{\psi_{1j}^2}\sum_{k=2}^{N}\frac{\lambda_1}{\lambda_1-\lambda_k}\left(\psi_{kj}^2 - \psi_{ki}\psi_{kj}\frac{\psi_{1j}}{\psi_{1i}}\right). \tag{11}$$

The partial mean hitting time to reach j is

$$T_j^{\mathrm{MERW}} = \frac{1}{\psi_{1j}^2(N-1)}\sum_{k=2}^{N}\frac{\lambda_1}{\lambda_1-\lambda_k} \times \left(N\psi_{kj}^2 - \psi_{kj}\psi_{1j}\sum_{i=1}^{N}\frac{\psi_{ki}}{\psi_{1i}}\right), \tag{12}$$

and the global mean hitting time is

$$\langle T^{\mathrm{MERW}}\rangle = \frac{1}{N(N-1)}\sum_{j=1}^{N}\frac{1}{\psi_{1j}^2}\sum_{k=2}^{N}\frac{\lambda_1}{\lambda_1-\lambda_k} \times \left(N\psi_{kj}^2 - \psi_{kj}\psi_{1j}\sum_{i=1}^{N}\frac{\psi_{ki}}{\psi_{1i}}\right). \tag{13}$$

In contrast with the unbiased case, in the maximal entropy random walk, both the hitting times and the stationary distribution depend only on the eigenvalues and eigenvectors of adjacency matrix $\mathcal{A}$ [21,23].

C. Uniform and regular hypergraphs

In *uniform hypergraphs*, all the hyperedges have the same cardinality, i.e., $|e_j| = c$ for all $j \in \{1, 2, \cdots M\}$. From the spectral viewpoint, the spectra of both the counting and the normalized adjacency matrices are the same, with a difference of only a scaling factor $(c-1)$. Similar arguments can also be applied to the Laplacian and the probability transition matrices. Consequently, for a given class of random walks, both the projected and the higher-order steps have the same stationary distributions and hitting times. Although uniform hypergraphs are relatively simpler structures, it does not imply that they are trivial. Even without heterogeneity in the distribution of cardinalities, the degree distribution can still be heterogeneous. We numerically explored this case in Sec. IV A 3.

Furthermore, the MERW and the URW are equivalent in uniform and regular hypergraphs, where all the nodes in $\mathcal{H}$ have the same degree. In this case, the spectra of the Laplacian matrix can be described by the spectra of the adjacency matrix up to a scale and translation, thus implying that both classes of random walks present the same behavior. Finally, it is worth mentioning that, in Ref. [32], the authors formally derived the cover times for a class of regular and uniform hypergraphs, providing exact expression as well as asymptotic results.

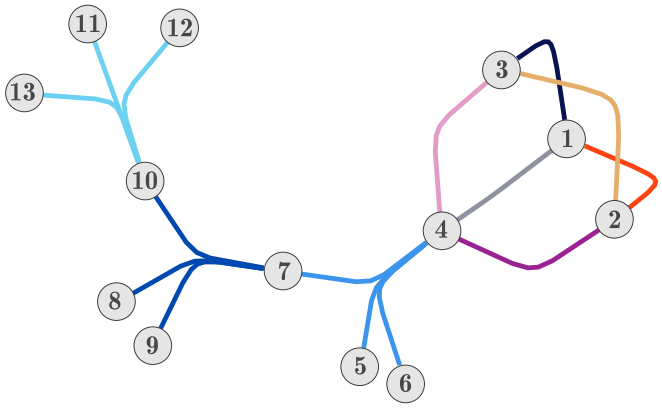

FIG. 1. A toy hypergraph, $\mathcal{H} = \{\mathcal{V}, \mathcal{E}\}$, with $N = 13$ nodes, $M = 9$ hyperedges, where $\mathcal{E} = \{e_1, e_2, ..., e_9\}$ and $e_1 = \{1, 2\}$, $e_2 = \{1, 3\}$, $e_3 = \{1, 4\}$, $e_4 = \{2, 3\}$, $e_5 = \{2, 4\}$, $e_6 = \{3, 4\}$, $e_7 = \{4, 5, 6, 7\}$, $e_8 = \{7, 8, 9, 10\}$, and $e_9 = \{10, 11, 12, 13\}$. The hyperedges are color coded.

IV. NUMERICAL EXPERIMENTS

In this section, we complement our analysis with numerical experiments. First, in Sec. IV A, we focus on artificial hypergraphs, evaluating both the distribution of cardinalities and the degree distribution. Next, in Sec. IV B we show an example of a real hypergraph, extending our analysis to cases where nontrivial correlations are present.

A. Artificial hypergraphs

First, we present a toy example that allows us to comment on both the differences between the step definitions and the random walks. Next, in Secs. IV A 2 and IV A 3, we describe and evaluate a series of synthetic hypergraphs with different levels of heterogeneity.

1. A toy example

We analyze a small toy example of a hypergraph to highlight the differences among the random walks studied here. We consider a hypergraph with $N = 13$ nodes and $M = 9$ hyperedges, as depicted in Fig. 1. This hypergraph is "nearly symmetric" with respect to node 7, as the structure $\mathcal{E}^{\boxtimes} = \{e_1, e_2, e_3, e_4, e_5, e_6\}$ is a projected clique of a hyperedge with cardinality 4 and composed by nodes 1, 2, 3, and 4. The motivation behind the toy model is to emphasize if different types of random walks are able to distinguish a higher-order structure (a hyperedge) from its pairwise counterpart (a clique here denoted by $\mathcal{E}^{\boxtimes}$). Notably, nodes 4, 7, and 10 act as bridges, connecting hyperedges to each other. While both node 7 and node 10 are connecting two hyperedges, node 4 functions as a bridge connecting a hyperedge with a projected

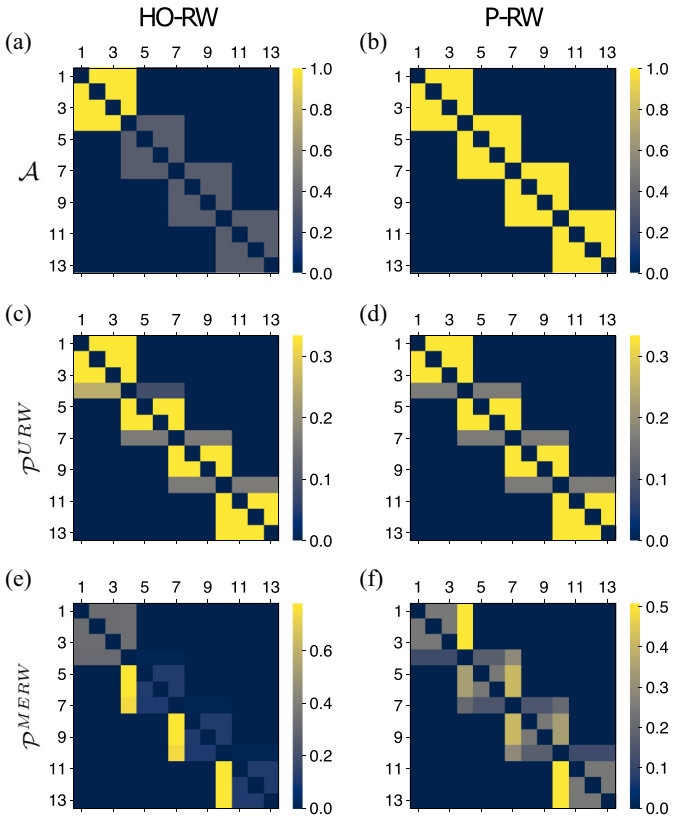

FIG. 2. Graphical representation of the adjacency and probability transition matrices for the toy's example. In (a), (c), and (e) we show the matrices related to the higher-order step, while in (b), (d), and (f) we show those related to the projected step. In (a) and (b) the adjacency matrices, $\mathbf{A}^{\text{norm}}$ and $\mathbf{A}^{\text{count}}$, respectively. In (c) and (d) the unbiased random walk probability transition matrix, $\mathcal{P}^{\text{URW}}$, while in (e) and (f) the maximum entropy random walk transition matrix, $\mathcal{P}^{\text{MERW}}$.

clique. As a result, we expect to observe differences across the random walks due to this asymmetry.

Figure 2 shows the adjacency and probability transition matrices. We observe that the counting adjacency matrix in Fig. 2(b) does not distinguish between the hyperedge $e_9 = \{10, 11, 12, 13\}$ and the projected clique $\mathcal{E}^{\boxtimes}$. On the other hand, the representation given by Eq. (3), in Fig. 2(a), weights these two types of structures differently. Regarding the transition matrices $\mathcal{P}^{\text{URW}}$ in Figs. 2(c) and 2(d), the differences are relatively small, with the bridge node 4 playing a slightly different role. In the URW with the projected step, node 4 is equivalent to other bridge nodes. On the contrary, the URW with the higher-order step has a higher transition probability to move from node 4 towards the clique than moving towards the hyperedge. However, for the $\mathcal{P}^{\text{MERW}}$ in Figs. 2(e) and 2(f), the differences are more pronounced. Specifically, in the higher-order case, the walker is more likely to remain within the set of hyperedges $\mathcal{E}^{\boxtimes}$, in contrast to the projected step. This example also highlights the nonlocality of the MERW: The walker is biased to move towards the clique even if it is far away. This effect was not present in the URW, where the effect of the clique was perceived only locally at node 4.

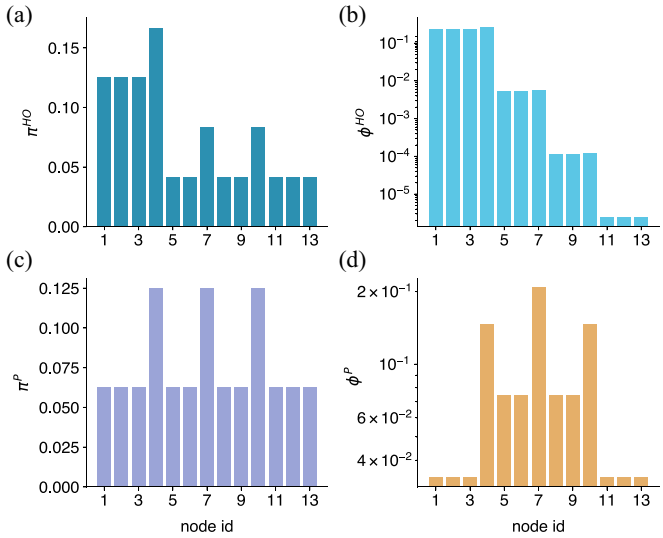

FIG. 3. Toy's example stationary distribution for the unbiased random walk with the higher-order step in (a), the maximal entropy with the same step in (b), the unbiased random walk with projected step in (c), and the maximal entropy random walk with the same step in (d).

In Fig. 3, we examine the stationary distributions of the four types of random walks considered. Figures 3(a) and 3(b) clearly show the effects of asymmetry in the higher-order step cases. In contrast, the projected step creates a symmetry that is reflected in the stationary distributions, which can be seen in Figs. 3(c) and 3(d). This effect is evident in the probabilities of nodes 4, 7, and 10 in both the unbiased and maximum entropy random walks.

Complementary, the expected time to hit v_j, starting from v_i is represented in Fig. 4. In the projected case, Fig. 4(c) and 4(d), both the URW and MERW exhibit a similiar behavior. Additionally, it is evident that the higher-order step creates a

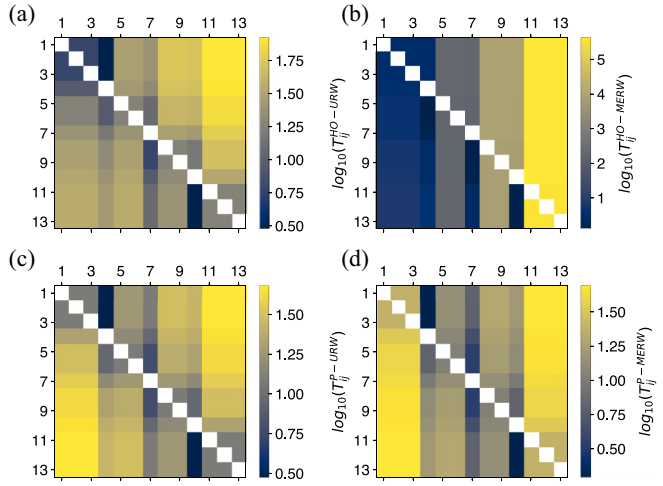

FIG. 4. Toy's example hitting times, T_{ij}, for the different types of random walk studied. In (a) and (c) for the unbiased random walk and in (b) and (d) for the maximum entropy random walk. Complementary, in (a) and (b) the higher-order step, while in (c) and (d) the projected step.

bias towards the clique $\mathcal{E}^{\boxtimes}$. This effect is even more apparent in the maximum entropy case, where the time to reach the nodes 1, 2, 3, and 4 in the clique is orders of magnitude shorter than that for nodes 10, 11, 12, and 13. Moreover, we remark on the particular role of nodes 4, 7, and 10, which serve as bridges. Thus, the probabilities of getting to these nodes typically present a strong dependency on the origin node.

2. Hypergraph models

Here we describe how to generate a random hypergraph with a fixed number of nodes N and a single connected component.

We begin by considering the random uncorrelated model, where we generate hypergraphs with an arbitrary distribution of cardinalities without controlling the degree distribution. We fix the number of nodes, N, and hyperedges, M. The cardinalities are then sampled from a given distribution, $P(|e_j|)$. To generate a hyperedge e_j, we uniformly sample $|e_j|$ nodes from the vertex set $\mathcal{V}$.

This procedure, which we refer to as a single trial, results in a random hypergraph that may have multiple connected components. To ensure a single connected component, we use a trial-and-error algorithm that performs $K = 200$ trials to find a connected component[3] with N nodes. If all K trials are unsuccessful, then the algorithm increases or decreases the number of nodes by $\Delta N = 10$ and repeats the procedure another K times.

The algorithm terminates once the largest connected component forms a hypergraph with N nodes. Even if hypergraphs are generated with a different number of nodes than N, we select the subhypergraph corresponding to a giant connected component with N nodes. In this way, the algorithm effectively maintains the number of nodes, while the number of hyperedges may slightly fluctuate.

Using the above-described model, we constructed both Poisson, $P(|e_j|) \sim \text{Poisson}(\beta)$, and power-law (PL) distributions, $P(|e_j|) \sim |e_j|^{-\gamma}$. In terms of cardinalities, the Poisson distribution generates more homogeneous hypergraphs, where we can control the average cardinality. On the other hand, the PL distribution represents a class of heterogeneous hypergraphs, whose heterogeneity can be controlled by the parameter γ. With this method, we can also set the minimum and maximum cardinality respectively to $e_{\min} = 2$ and $e_{\max} = \sqrt{N}$. While more sophisticated methods, as proposed in Ref. [33], may be available, they can incur higher computational costs, making them impractical for large sample sizes. Here, for each experiment, we considered $n_{\text{runs}} = 10^3$ independently generated hypergraphs. We remark that we decided to keep the number of nodes N and hyperedges M as fixed parameters. In this way, we are sure to avoid that the size of the hypergraph has an effect on the measures. As a continuation of this work, other valid alternatives could be explored, such as keeping the average cardinality or the average degree fixed and letting the number of hyperedges vary.

[3]A connected component in a hypergraph can be obtained by using the standard graph algorithms in the count or normalized adjacency matrices.

In addition to the heterogeneity of cardinalities, we also consider uniform hypergraphs with controlled degree distribution. To produce a homogeneous degree distribution, we use the previous algorithm with the distribution $P(|e_j|) \sim \mathbb{1}_{\{|e_j|=c\}}$, where $\mathbb{1}_{\{|e_j|=c\}}$ is the indicator function which equals one if $|e_j| = c$ and zero otherwise. To produce a heterogeneous degree distribution, we use the algorithm proposed in Ref. [34], which generates uniform hypergraphs with PL distributions with $P(k) \sim k^{-\gamma}$ and $\gamma = 1 + \frac{1}{\nu}$. The algorithm associates each node i with the probability $p_i = \frac{i^{-\nu}}{\zeta_N(\nu)}$, where $\zeta_N(\nu) = \sum_{j=1}^{N} j^{-\nu}$ and $0 < \nu < 1$. Next, for each hyperedge, we select c nodes following the probabilities p_i. This algorithm is a generalization of the static model defined in Refs. [35,36] for graphs and fixes the average degree to be $c\frac{M}{N}$. We again use the same brute force algorithm to ensure a single connected component, which may result in fluctuations similar to those observed in the random uncorrelated model.

Last, we generate hypergraphs with both degree and cardinality distribution following a power law. This model is inherently more complex than the previous ones due to the challenging task of matching the two distributions. We used an algorithm based on three simple steps: (i) an unrestricted matching, (ii) a brute-force fixing algorithm that swaps repeated nodes on the hyperedges, and (iii) a random swap step (using the swap proposed in Ref. [33]) that ensures that the final hypergraph is uniformly sampled from the space of possible hypergraphs. We note that we cannot formally guarantee that our hypergraph is uniformly sampled because, to the best of our knowledge, there is no lower bound on the number of necessary swaps in the general case. However, we perform 10^4 swaps in hypergraphs with $N = 10^3$, which we hope will be sufficient. This algorithm has been proposed and systematically tested in Ref. [37].

3. Numerical results

Figure 5 illustrates the mean hitting time, $\langle T \rangle$, for the four classes of random walks on different Poisson distributions of cardinality distributions and various values of β. Regardless of the type of random walk, $\langle T \rangle$ decreases as we increase β. We remark that $\langle T \rangle$ is lower bounded by $(N - 1)$ as we need at least $(N - 1)$ steps to visit all the nodes. Comparing Figs. 5(a) and 5(c) with Figs. 5(b) and 5(d), we observe that the URW has a smaller $\langle T \rangle$ than the MERW for both steps. Moreover, in general, $\langle T \rangle$ in the projected step is larger than the higher-order one. As the structure is reasonably homogeneous, the URW is very similar for both steps. Although all curves show a similar trend, the MERW has a larger variance, particularly for small values of β. The extreme case is the P-MERW, where some hypergraphs have an average hitting time with a different order of magnitude, as shown in Fig. 5(d).

Complementary, in Fig. 6, we evaluate the impact of heterogeneity using a PL distribution of cardinalities. Except for the P-URW in Fig. 6(c), all the other cases present a mean hitting time that increases with γ. We also observe that the variance is relatively higher if compared to the Poisson case in Fig. 5. Similar comments as before also apply here, such as the projected step imposing a higher mean hitting time. Furthermore, we remark that the variance is more considerable for higher values of γ, which is particularly evident in

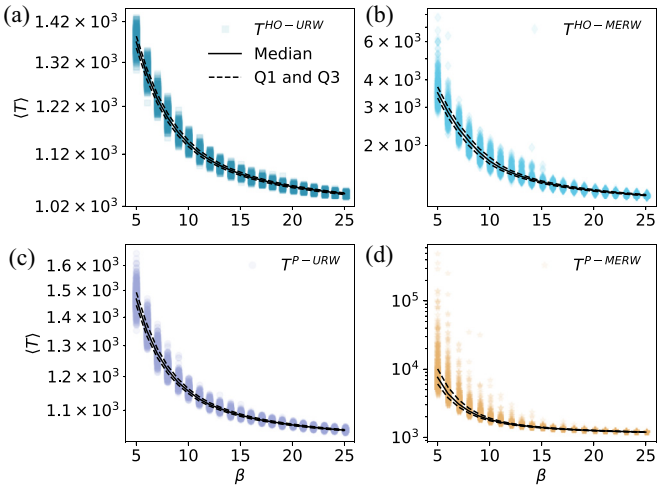

FIG. 5. Mean hitting time on random hypergraphs following a Poisson distribution of cardinalities, $P(|e_j|) \sim \text{Poisson}(\beta)$, $N = 10^3$, and $M = 10^3$. For each parameter, we have 10^3 independently generated hypergraphs, each one being a point in each panel. The median is represented by a black continuous line and the first and third quartiles are the dashed lines.

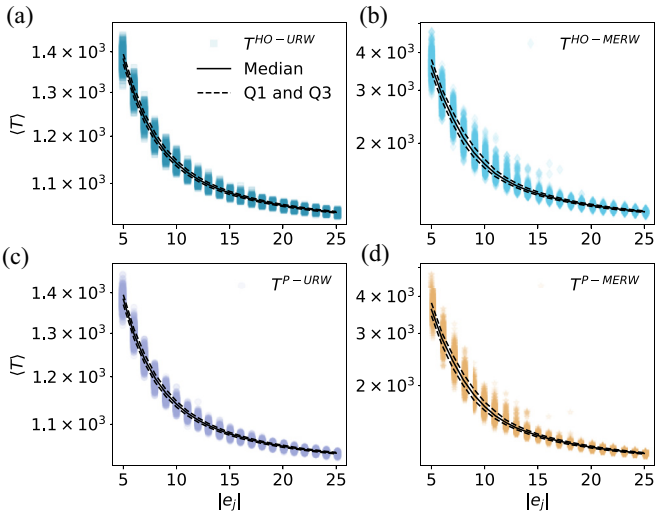

FIG. 7. Mean hitting time on uniform homogeneous random hypergraphs, $N = 10^3$ and $M = 10^3$. For each parameter, we have 10^3 independently generated hypergraphs, each one being a point in each panel. The median is represented by a black continuous line and the first and third quartiles are the dashed lines.

the MERW cases. This phenomenon arises because increasing γ with a fixed number of hyperedges M results in a sparser projected hypergraph. Similar effects can be observed by reducing the parameter β in Fig. 5, where the average size of each hyperedge decreases while the number of hyperedges is fixed, leading to a sparser projected hypergraph.

Next, we investigate the impact of degree distribution heterogeneity on uniform hypergraphs with fixed cardinalities. Figures 7 and 8 show results for homogeneous and power-law degree distributions, respectively. We note that both types of steps in the URW produce the same output, due to the uniformity of the cardinalities. In this case, the counting and normalized adjacency matrices have the same spectral distributions, up to a scale. For the uniform homogeneous case in Fig. 7, we observe that increasing the magnitude of the edges $|e_j|$ leads to a decrease in the mean hitting time, as also observed in the Poisson case, Fig. 5. In the PL case with $|e_j| = 20$ in Fig. 8, we find that, as we increase γ, the mean hitting time, $\langle T \rangle$, decreases. However, it is worth noting that the mean hitting times in the uniform homogeneous cases are considerably smaller, suggesting that degree distribution heterogeneity may also contribute to longer mean hitting times.

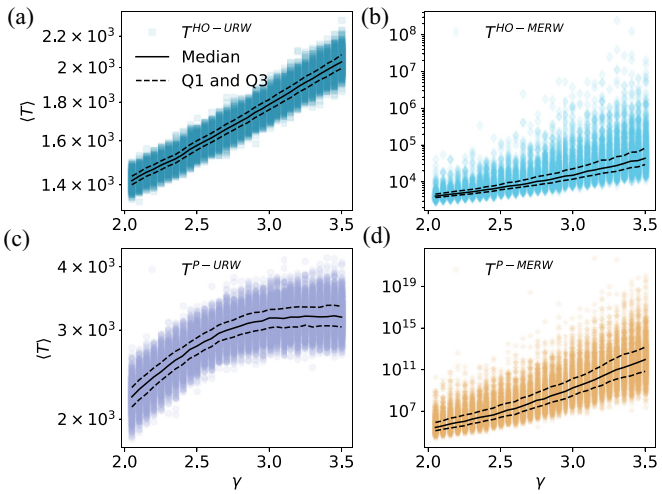

FIG. 6. Mean hitting time on random hypergraphs following a power-law distribution of cardinalities, $P(|e_j|) \sim |e_j|^{-\gamma}$, $N = 10^3$, and $M = 10^3$. For each parameter, we have 10^3 independently generated hypergraphs, each one being a point in each panel. The median is represented by a black continuous line and the first and third quartiles are the dashed lines.

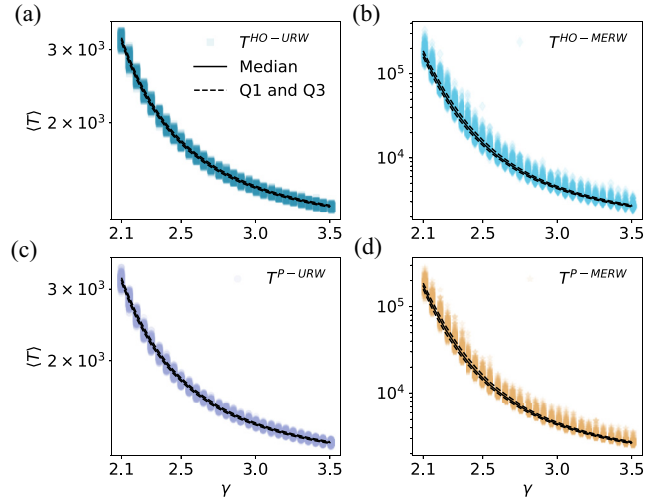

FIG. 8. Mean hitting time on uniform random hypergraphs following a power-law degree distribution, $P(k) \sim k^{-\gamma}$, $N = 10^3$, $|e_j| = 20$, and $M = 10^3$. The average degree is 20. For each parameter, we have 10^3 independently generated hypergraphs, each one being a point in each panel. The median is represented by a black continuous line and the first and third quartiles are the dashed lines.

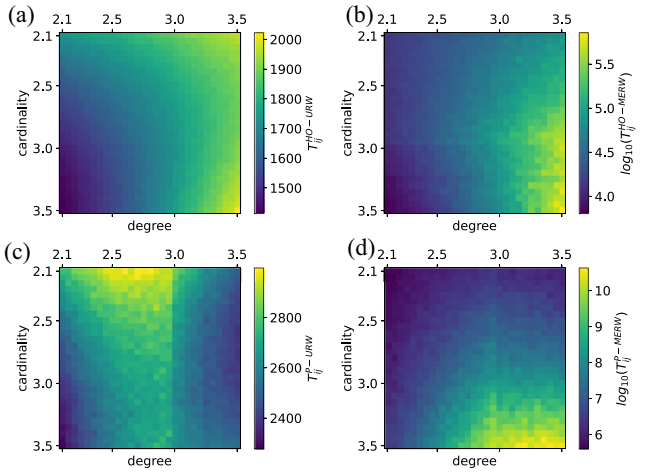

FIG. 9. Heatmaps showing the average mean hitting time for the four types of random walks. Each value of the heatmap is the median of the average mean hitting time of 100 independently generated hypergraphs.

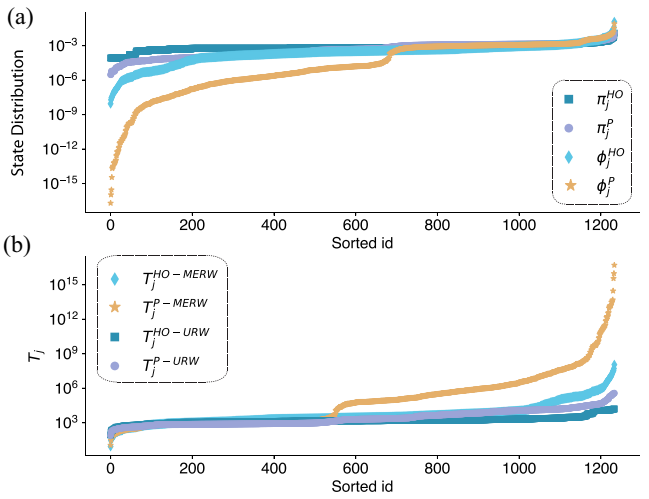

FIG. 10. Analysis of the *cat-edge-vegas-bars-reviews* real hypergraph. In (a) the stationary distribution for the four types of random walks, while in (b) partial mean hitting time, T_j, for the same four types of random walks.

To conclude our analysis of the impact of heterogeneity we investigate synthetic hypergraphs generated with a power-law degree distribution and a power-law cardinality distribution. We refer to the power-law exponent of the degree distribution as γ_d and to the one of the cardinality distribution as γ_c. Figure 9 shows the mean hitting times of the four types of random walks as the power-law distribution exponent is varied from 2.1 to 3.5 for both cardinality and degree distribution. All types of random walks present a particular behavior as the power-law exponent varies.

The unbiased random walk with the higher-order step has a mean hitting time that increases as the degree exponent increases. In particular, for a degree exponent larger than 3, the hitting times become longer, signaling the absence of big hubs. Regarding the cardinality exponent, it seems to play a bigger role when associated with a small degree exponent, reducing consistently the hitting times. Indeed, the minimal mean hitting time is obtained for $\gamma_c = 3.5$ and $\gamma_d = 2.1$.

The unbiased random walk with the projected step displays a completely different scaling. The average mean hitting time behaves as a *saddle* with a maximum for values of $\gamma_d \in (2.5, 3)$. Also, we observe a sharp change in the average mean hitting time as soon as the degree exponent reaches the value of 3.0.

Concerning the maximal entropy random walk, the mean hitting times scale considerably with the exponents [Figs. 9(b) and 9(d) are in log scale]. We can notice that the region with $\gamma_c > 3.0$ and $\gamma_d > 3.0$ is where the mean hitting time is larger. This behavior is probably due to the absence of hubs, which usually make the system easier to navigate. Finally, there is a similar sharp change when the $\gamma_c = 3.0$ for the HO-MERW and one when $\gamma_d = 3.0$ for the P-MERW.

B. Real hypergraphs

In this section, we present the analysis of the real hypergraph *cat-edge-vegas-bars-reviews* [38]. Nodes are Yelp users, and hyperedges are users who reviewed an establishment of a particular category, which are different types of restaurants in Las Vegas, NV, United States. This data were collected for a month. We focus our analysis on this hypergraph, as it presented a very rich behavior, serving as an example of the different types of behavior introduced by correlations. Additional real hypergraphs are analyzed in Appendix.

Figure 10 shows the stationary distribution and the partial mean hitting times in Figs. 10(a) and 10(b), respectively. In this figure, the x axis is sorted independently for each type of random walks studied, for visualization purposes. We observe that the four types of random walks have very different behaviors. This outcome is especially evident for the P-MERW, in which some nodes are much harder to reach and the stationary distribution is more localized in fewer nodes, not just in the hubs, but from node 700 onwards. We found a similar trend exploring the stationary distribution of artificial hypergraphs with a power-law cardinality distribution. However, the real hypergraph *cat-edge-vegas-bars-reviews* exhibits a peculiar fluctuation in the stationary distribution of the P-MERW around sorted id 700 [Fig. 10(a)] that we could not find in any other synthetic hypergraphs. Additionally, in Fig. 11, we can verify that the stationary distributions and their respective partial mean hitting times are strongly correlated. However, this correlation is not linear as the Pearson correlation, ρ^P, is very small. However, the Spearman correlation is high, thus suggesting that there is a monotonous relationship. From the analytical viewpoint, this is particularly evident for the MERW case, as in Eq. (12), T_j is inversely proportional to the stationary distribution $\phi_j = \psi_j^2$. For the URW case, this argument is slightly more complex since it is not obvious that Eq. (7) depends on any individual structural feature. It is known that the recurrence time (T_{ii}) of the URW is inversely proportional to its stationary probability [18],

$$T_{ii} = \frac{1}{\pi_i}, \tag{14}$$

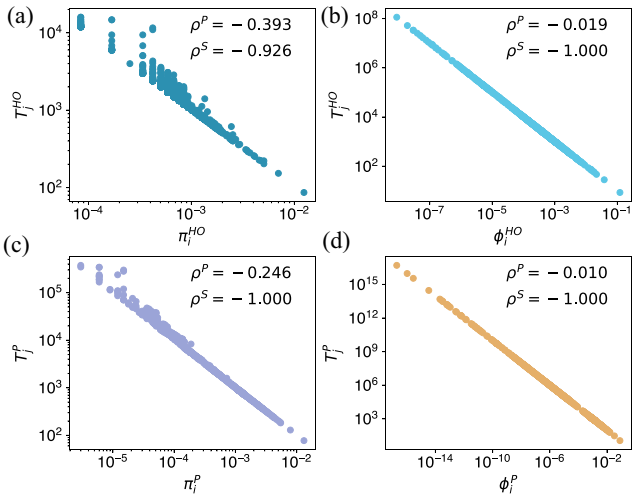

FIG. 11. Comparative analysis of the different stationary distributions with their respective partial mean hitting time for the *cat-edge-vegas-bars-reviews* real hypergraph. We report the Pearson, ρ^P, and Spearman, ρ^S, correlations in the upper right corner of each panel.

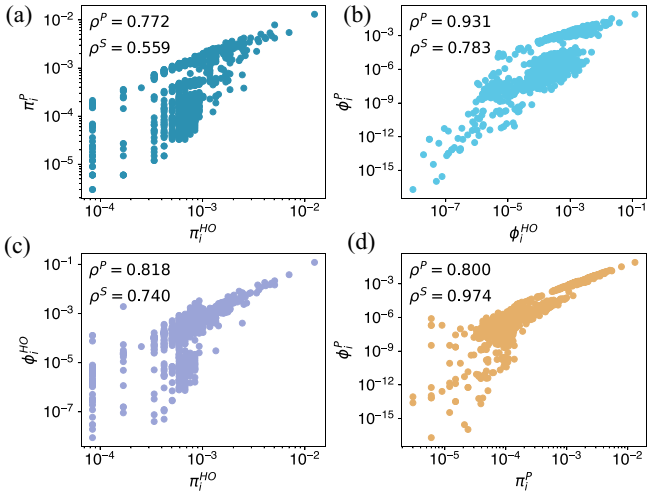

FIG. 12. Comparative analysis of the different stationary distributions for the *cat-edge-vegas-bars-reviews* real hypergraph. We report the Pearson, ρ^P, and Spearman, ρ^S, correlations in the upper left corner of each panel.

and similar behavior can be found for the hitting times T_{ij} with $i \neq j$ (see Sec. 3.2.5 in Ref. [18] for the details of the calculation). Another explanation is given in Ref. [20], where the authors elegantly derived a lower bound for the partial mean hitting time T_j using the Cauchy inequality and the property that the Laplacian components $\mathcal{L}_{ij}$ can be rewritten as

$$\mathcal{L}_{ij} = \sum_k \sigma_k \mu_{ki} \mu_{kj}, \tag{15}$$

and $\mathcal{L}_{ii} = k_i$, the node degree. Finally, the authors arrived at the bound

$$T_j \geqslant \frac{N}{N-1} \frac{(2E - k_j)^2}{2E \times k_j - \sum_z k_z \mathcal{L}_{jz}}. \tag{16}$$

From this bound and remembering that for a URW the stationary distribution $\pi_j \propto k_j$, we can get the intuition behind the nonlinear correlation found in Fig. 11. This is a heuristic argument since the equality holds only for special graphs, like the complete graph or the star graph.

Finally, in Fig. 12, we compare the different classes of random walk in terms of their stationary distributions. In all the evaluated cases, the correlations are positive. Interestingly, comparing the two different steps with the same type of random walk, Figs. 12(a) and 12(b), we observe that the Spearman correlation is considerably lower than the Pearson correlation. This result suggests that the relationship between different steps is nontrivial and that their rankings are not the same. Next, comparing the unbiased with the maximal entropy for each step, Figs. 12(c) and 12(d), we observe a strong correlation. However, their relationship is not trivial.

We also compared Fig. 12 and Fig. 11 with artificial hypergraphs, finding again many similarities with the power-law cardinality distribution case. The most notable difference is in Fig. 12(b), where we found a much weaker correlation (around 0.5) between the projected (ϕ_i^P) and higher-order (ϕ_i^{HO}) stationary distributions for artificial hypergraphs with power-law cardinality distributions. The comparison between random walks on real and artificial hypergraphs is outside the scope of this article and is left as a future work.

V. ANALYSIS AND DISCUSSION

At first glance, random walks on hypergraphs might seem like an abstract problem. However, similar to random walks in other contexts, this abstraction might provide insights into hypergraphs' structural organization. For instance, in Ref. [31], random walks were used to detect community structures. As an application, the authors considered a hypergraph where nodes represent animals and hyperedges represent features, and they used random walks to group "similar animals" into communities [31]. Another possible abstraction is the calculation of the probability and time necessary for a message to travel from node i to node j. These examples support our argument that random walks are more general than their simplistic interpretation of a walker following a physical path in the hypergraph.

Here we have shown that random walks can have different interpretations depending on the type of step adopted. We have considered what we call the projected and the higher-order steps. In the projected step, the hypergraph is effectively projected on a graph, where the walk takes place. This is exemplified by the toy hypergraph in Fig. 1, where we note that projected steps do not distinguish between a hyperedge with $|e_j| > 2$ and a clique with the $|e_j|$ nodes. In the higher-order step, the random walk can be interpreted as a sequence of two dependent processes, first a uniform choice of the next hyperedge, then a uniform choice of the next node. This construction allows for a distinction between the structures mentioned above. We remark that the projected and the higher-order steps are two possibilities, and many other processes can be defined following similar arguments. Indeed, although not explored here, in Ref. [31] the authors have

formulated the random walk following the adjacency matrix

$$K_{ij}^{\tau} = \begin{cases} \sum_{\alpha: e_\alpha \in \mathcal{E}} (\mathcal{B}_{\alpha\alpha} - 1)^{\tau} \mathcal{I}_{i\alpha} \mathcal{I}_{j\alpha} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}, \tag{17}$$

where $\mathcal{B} = \mathcal{I}^T \mathcal{I}$ and τ is a real parameter. Note that the element $\mathcal{B}_{ij}$ is the number of nodes in the intersection between e_i and e_j, i.e., $|e_i \cap e_j|$. If $\tau = 0$, then we recover the counting adjacency matrix and the projected step. On the other hand, if $\tau = -1$, then we recover our normalized adjacency matrix and thus the higher-order step. We can apply the same expressions introduced in Sec. III for the times and stationary distributions.

All these projections have in common that hyperedges of size $|e|$ are projected onto $|e|$ cliques. Another widely used and flexible alternative is to represent a hypergraph of N vertices and M hyperedges as a bipartite graph of $N + M$ nodes. The two classes of nodes in the bipartite graph are the vertices of the hypergraph and the hyperedges.

The corresponding adjacency matrix is

$$\mathbf{A}^{\text{bipartite}} = \begin{pmatrix} \mathbf{0} & \mathcal{I} \\ \mathcal{I}^T & \mathbf{0} \end{pmatrix}, \tag{18}$$

where $\mathcal{I}$ is the hypergraph incidence matrix. It is worth noting that for this representation, a walker to go from a vertex v_i to a vertex v_j necessarily has to move through a hyperedge that both vertices have in common, meaning that it will require two steps. The resulting walk is an alternating sequence of distinct pairs of vertices and hyperedges. Therefore, a step with the projected or higher-order step should be compared to two steps on the bipartite representation, as they both represent a movement from one vertex to another in the hypergraph.

It is important to comment on the unbiased and maximal entropy random walk defined on this representation. We show that a nontrivial mapping exists between random walks defined on the bipartite representation of the hypergraph and random walks defined on top of the projections considered in Sec. II. Specifically, the URW on the bipartite is equivalent to the URW on the normalized adjacency projection, while the MERW on the bipartite is equivalent to the MERW on the counting adjacency projection. The former equivalence was also pointed out in Ref. [39]. One can indeed notice that the weights appearing in Eq. (3) represent a uniform choice between $|e| - 1$ possible arrival nodes. Here we show the equivalence between the MERW on the bipartite representation and the MERW with the projected step. We first follow the reasoning in Ref. [19] to derive the transition probabilities for a nonlazy MERW on the bipartite representation. Finally, we demonstrate that they are equivalent to the MERW with the projected step. We denote as $\gamma^{2t}_{v_{i_0}, v_{i_t}}$ the trajectory of length $2t$ corresponding to the sequence $v_{i_0} e_{j_1}, v_{i_1} \ldots, e_{j_t} v_{i_t}$. The MERW is the random walk that maximizes the entropy of the set of sequences of length $2t$,

$$S_t = -\sum_{v_{i_0}, e_{j_1}, \ldots, v_{i_t}} \mathbb{P}\left(v_{i_0}, e_{j_1}, \ldots, v_{i_t}\right) \ln \mathbb{P}\left(v_{i_0}, e_{j_1}, \ldots, v_{i_t}\right). \tag{19}$$

The probability of the sequence is $\mathbb{P}(v_{i_0}, e_{j_1}, \ldots, v_{i_t}) = \pi_0 \mathbb{P}(\gamma^{2t}_{v_{i_0}, v_{i_t}})$ and π_0 is the probability of being in node v_0. This quantity is maximized when the sequence is chosen with uniform probability. We denote by N_{2t} the number of all the possible sequences of length $2t$, then the entropy simplifies to

$$S_t = \ln N_{2t}. \tag{20}$$

The number of paths of length $2t$ between a pair of nodes v_{i_0}, v_{i_t} in the bipartite graph is simply given by $[(\mathbf{A}^{\text{bipartite}})^{2t}]_{i_0 i_t}$. Since we are working with nonlazy random walks, the above expression can be shown to be modified to $[(\mathbf{A}^{\text{count}})^t]_{i_0 i_t}$. As a consequence, the Shannon entropy in Eq. (19) is maximized for

$$S_t = \ln \sum_{v_{i_0}, v_{i_t}} [(\mathbf{A}^{\text{count}})^t]_{i_0 i_t} \underset{t \to \infty}{\sim} t \ln \lambda, \tag{21}$$

where λ is the leading eigenvalue of $\mathbf{A}^{\text{count}}$. It can be shown that the correct choices for the transition probabilities of a nonlazy maximal entropy random walk on the bipartite projections are

$$P(v_i \to e) = \frac{\mathcal{I}_{ie}}{\lambda} \frac{Z_i(e)}{\psi_i}, \tag{22}$$

$$P(e \to v_k | v_i) = \frac{\psi_k \mathcal{I}^T_{ek} (1 - \delta_{ik})}{Z_i(e)}, \tag{23}$$

$$Z_i(e) = \sum_{k \neq i} \mathcal{I}^T_{ek} \psi_k. \tag{24}$$

$P(e \to v_k | v_i)$ is the probability to transition from hyperedge e to node v_k given that at the previous step the walker was in v_i and ψ_i is the i component of the eigenvector corresponding to the leading eigenvalue λ. This type of random walk maximizes the Shannon entropy in Eq. (19) in the limit $t \to \infty$ and is equivalent to the maximal entropy random walk defined on top of the counting projection. Indeed, by considering the two-step probability

$$\begin{aligned} P(v_i \to v_j) &= \sum_e P(v_i \to e) P(e \to v_k | v_i) \\ &= \sum_e \frac{\mathcal{I}_{ie}}{\lambda} \frac{\mathcal{I}^T_{ej} (1 - \delta_{ij}) \psi_j}{\psi_i} = \frac{\mathbf{A}^{\text{count}}_{ij}}{\lambda} \frac{\psi_j}{\psi_i}, \end{aligned} \tag{25}$$

which is precisely the expression in Eq. (9) with the projected step.

VI. CONCLUSIONS

In this paper, we introduced maximal entropy random walks in hypergraphs, which to the best of our knowledge, has not been previously explored. Besides, we complement the results in Refs. [13–15,19] by allowing for different types of random walk steps. We explore the projected and higher-order steps to construct the unbiased and the maximal entropy random walks and characterize their stationary distribution, hitting times, partial, and mean hitting times.

Our numerical experiments consider homogeneous and heterogeneous hypergraphs in terms of cardinality and degree distributions. We observe that, regardless of the type of random walk, increasing the average cardinality tends to decrease the average hitting time as seen in the numerical experiments reported in Figs. 5 and 7 for homogeneous cases. Furthermore,

TABLE II. Structural characterization of the real hypergraphs. The hypergraphs are characterized by the number of nodes, N, number of hyperedges, M, average cardinality, $\langle|e_j|\rangle$, standard deviation of the cardinalities std$(|e_j|)$, and maximal cardinality, max(e_j). The normalized adjacency matrix metrics are the average, standard deviation minimum and maximum degree, $\langle k^{\mathrm{HO}}\rangle$, std$(k^{\mathrm{HO}})$, $k^{\mathrm{HO}}_{\mathrm{min}}$, and $k^{\mathrm{HO}}_{\mathrm{max}}$, while the respective metrics from the counting adjacency matrix are $\langle k^{P}\rangle$, std(k^{P}), k^{P}_{min}, and k^{P}_{max}. In the hypergraphs marked with a "$*$" we have repeated hyperedges. For more, see Appendix A 1.

Name	N	M	$\langle\|e_j\|\rangle$	std$(\|e_j\|)$	max(e_j)	$\langle k^{\mathrm{HO}}\rangle$	std(k^{HO})	$k^{\mathrm{HO}}_{\mathrm{min}}$	$k^{\mathrm{HO}}_{\mathrm{max}}$	$\langle k^{P}\rangle$	std(k^{P})	k^{P}_{min}	k^{P}_{max}
cat-edge-algebra-questions	420	1267	6.519	6.579	107	19.664	34.091	1	375	239.076	352.769	1	3362
cat-edge-geometry-questions	580	1193	10.465	15.647	230	21.526	36.264	1	260	707.334	1066.547	1	6711
cat-edge-vegas-bars-reviews	1234	1194	9.937	13.817	73	9.615	7.371	1	147	270.665	295.724	1	4388
cat-edge-madison-restaurant-rev.	565	601	7.656	7.281	43	8.143	7.217	1	59	110.588	104.189	2	716
cat-edge-music-blues-reviews	1104	693	15.147	14.716	83	9.508	10.723	1	127	270.447	279.523	2	3393
phs-email-Enron	4423	15 653	4.119	4.458	25	14.576	101.395	1	4869	115.795	494.400	1	15471
phs-email-W3C	13 351	19 351	2.219	0.953	25	3.217	24.751	1	958	5.237	31.534	1	1293
contact-high-school*	327	172 035	2.050	0.234	5	1078.648	816.639	7	4495	1161.639	883.960	7	4655
contact-primary-school*	242	106 879	2.096	0.310	5	925.612	446.772	125	2234	1056.744	530.606	131	2640

our experiments suggest that heterogeneity increases the mean hitting time for uniform hypergraphs with power-law-degree distribution. Notably, when both the degree and cardinality distribution are power law, the four classes of random walks exhibit distinct behaviors as the heterogeneity changes. The average hitting time for the MERW increases as the distribution exponent rises, while the URW with the projected step behaves differently, presenting a sort of *saddle*.

In general, we observe that hitting times for the projected step are typically larger than those for the higher-order step, and hitting times for the URW are smaller than those for the MERW.

We also evaluate a real hypergraph with different types of correlations, emphasizing the complementary nature of the four classes of processes studied here, (P/HO)-(U/ME)RW, and providing different insights about the underlying structure. We discuss other possible types of steps found in the literature, particularly the walk on the bipartite representation of the hypergraph. We comment that a nontrivial mapping exists between random walks defined on the bipartite representation of the hypergraph and random walks defined on top of the projections considered in this paper. Specifically, the URW on the bipartite is equivalent to the URW on the normalized adjacency projection, while the MERW on the bipartite is equivalent to the MERW on the counting adjacency projection. Overall, our work contributes to a better understanding of random walks on hypergraphs and provides a versatile tool for analyzing complex systems in various domains.

Our results highlight the importance of the localization properties of the adjacency matrix and suggest that this feature might play a crucial role in other processes such as social contagion and information diffusion on hypergraphs. We hope that our findings will motivate further research with the potential to provide valuable insights for various applications, including the analysis of real higher-order systems and the development of novel methods in related fields such as artificial intelligence and behavioral sciences.

Custom code that supports the findings of this study is available [40].

ACKNOWLEDGMENTS

Y.M. was partially supported by the Government of Aragón, Spain and "ERDF A way of making Europe" through Grant No. E36-23R (FENOL), and by Ministerio de Ciencia e Innovación, Agencia Española de Investigación (MCIN/AEI/10.13039/501100011033) Grant No. PID2020-115800GB-I00. We acknowledge the use of the computational resources of COSNET Lab at Institute BIFI, funded by Banco Santander (Grant Santander-UZ 2020/0274) and by the Government of Aragón (Grant No. UZ-164255). The funders had no role in study design, data collection and analysis, the decision to publish, or the preparation of the paper.

APPENDIX: ADDITIONAL EXPERIMENTS: REAL HYPERGRAPHS

In this section, we present additional experiments on real hypergraphs to complement the results of the main text. Here we focus on the stationary distributions and the partial mean hitting times. First, in the next section, we briefly describe the databases, while in Sec. A 2, we briefly compare their results.

1. Databases

The database used here is available [41]. Additionally, in Table II, we provide a brief structural characterization of these hypergraphs. We kept the dataset names on the repository to facilitate its identification, reproduction, and further studies. For more information about a specific hypergraph, please see the provided references. We also provide a short description of each hypergraph as follows:

(i) *cat-edge-algebra-questions* [38]: A hypergraph where nodes are users on MathOverflow and hyperedges are sets of users who answered a certain question category. This dataset

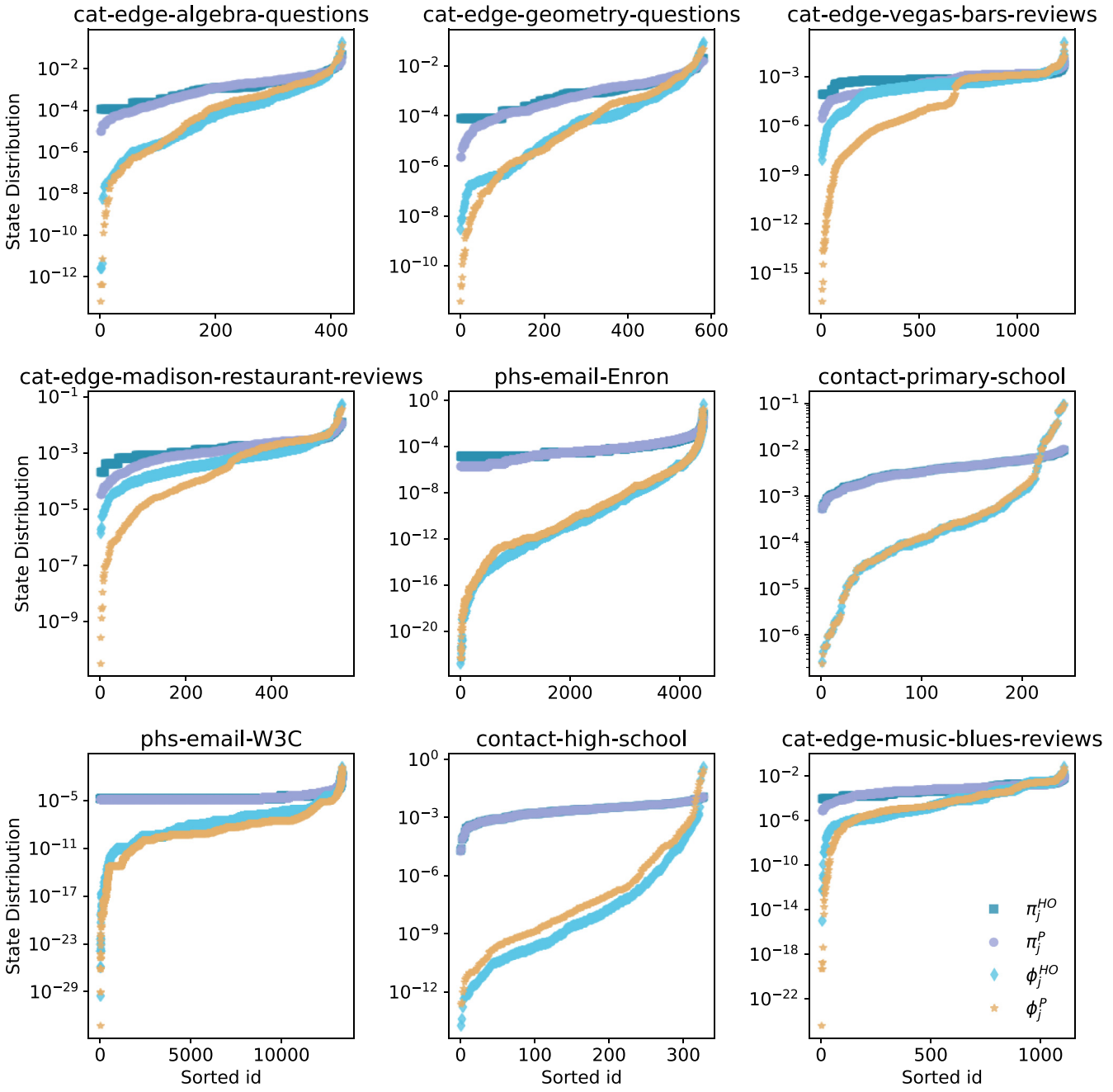

FIG. 13. Stationary distribution for the four types of random walks studied for the database described in Section A 1.

was collected from different tags involving algebra, and it was derived from the Stack Exchange data dump;

(ii) *cat-edge-geometry-questions* [38]: A hypergraph where nodes are users on MathOverflow and hyperedges are sets of users who answered a particular question category. This dataset was collected from different tags involving geometry, and it was derived from the Stack Exchange data dump;

(iii) *cat-edge-vegas-bars-reviews* [38]: A hypergraph where the nodes are Yelp users and hyperedges are users who reviewed a bar of a particular category. This dataset is restricted to bars in Las Vegas, NV, and within a month's timeframe. The data were obtained from the Yelp Kaggle competition data;

(iv) *cat-edge-madison-restaurant-reviews* [38]: A hypergraph where the nodes are Yelp users and hyperedges are users who reviewed a restaurant of a particular category. This dataset is restricted to restaurants in Madison, WI, and within a month's timeframe. The data were obtained from the Yelp Kaggle competition data;

(v) *cat-edge-music-blues-reviews* [42]: A hypergraph where nodes are Amazon reviewers and hyperedges are reviewers who reviewed a specific type of blues music within a month timeframe. The dataset was compiled from the product reviews collected by Jianmo Ni, Jiacheng Li, and Julian McAuley;

(vi) *phs-email-W3C* [43,44]: A hypergraph where nodes correspond to email addresses with a w3c.org domain and

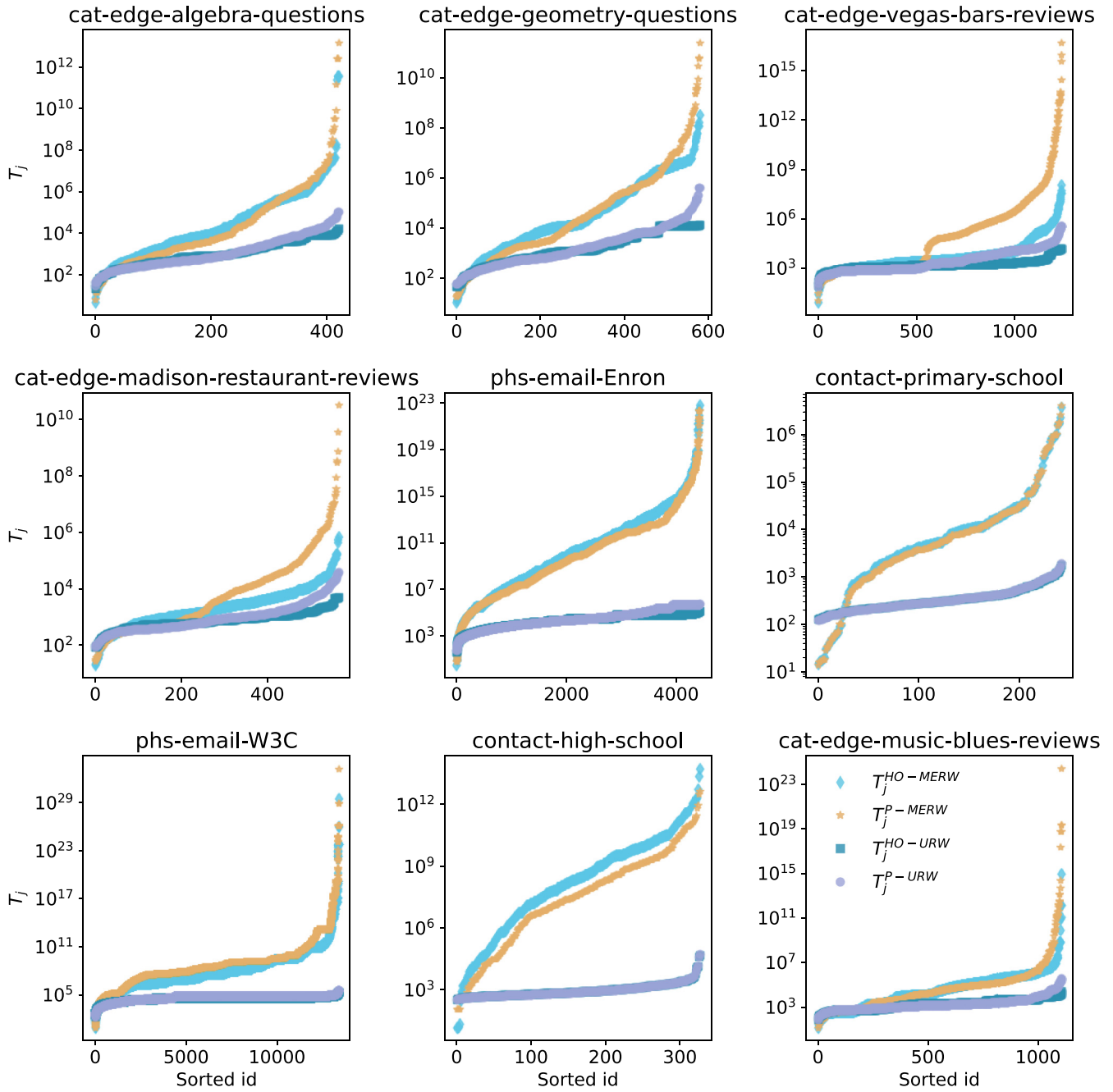

FIG. 14. Partial mean hitting time, T_j, for the four types of random walks studied for the database described in Section A 1.

hyperedge consists of a set of email addresses, which have all appeared on the same email. This dataset was originally used on the analysis of core-fringe structures in Ref. [44];

(vii) *phs-email-Enron* [44]: A hypergraph where nodes correspond to email addresses and hyperedges consist of sets of email addresses, which have all appeared on the same email. This dataset was originally used on the analysis of core-fringe structures in Ref. [44], where core nodes correspond to email addresses of the individuals whose email inboxes were released as part of the investigation by the Federal Energy Regulatory Commission;

(viii) *contact-high-school* [45,46]: This dataset is a temporal sequence of timestamped hyperedges, which are composed of people. It is constructed from interactions recorded by sensors worn by people at a high school. The resolution of these sensors is 20 s;

(ix) *contact-primary-school* [46,47]: This dataset is a temporal sequence of timestamped hyperedges, which are composed of people. It is constructed from interactions recorded by sensors worn by people at a primary school. The resolution of these sensors is 20 s;

We remark that *contact-high-school* and *contact-primary-school* have repeated hyperedges, thus the generated hyper-

graph is not simple. In the following section, we will use this version with repeated hyperedges to emphasize that our results also apply to this type of hypergraph. Thus, the interpretation of the random walk in these cases also slightly changes.

2. Additional experiments

Figure 13 shows the stationary distribution for the database discussed in Appendix A 1 and the four types of random walks studied. Note that, for visualization purposes, the x axis is the sorted id, which is done independently for each curve, implying that the rankings might be different. Perhaps the most evident difference among the distributions is observed in the class of the process, URW or MERW, and later on the step's definition. Moreover, the MERW presents a higher variance of states, spanning orders of magnitude. Finally, this random walk, only in some cases, also presented "lumps," which are most visible for *cat-edge-madison-restaurant-reviews* and *cat-edge-vegas-bars-reviews* hypergraphs. These observations might suggest that these structures present some form of localization, imprisoning the walkers into "entropic wells."

Figure 14 shows the partial mean hitting time for the database discussed in Appendix A 1 and the four types of random walks studied. Again, we highlight that the x axis is independently sorted for each curve. For the partial mean hitting times, similar comments, as for Fig. 13, also apply. As a particular observation for this measurement, we observe that T_j is typically higher for the MERW. However, some nodes present a lower partial mean hitting time if compared to the URW. This result might suggest that these nodes play a notably different role, maybe serving as bridges.

[1] F. Battiston, E. Amico, A. Barrat, G. Bianconi, G. Ferraz De Arruda, B. Franceschiello, I. Iacopini, S. Kéfi, V. Latora, Y. Moreno *et al.*, Nat. Phys. **17**, 1093 (2021).

[2] R. Lambiotte, M. Rosvall, and I. Scholtes, Nat. Phys. **15**, 313 (2019).

[3] R. Mulas, C. Kuehn, and J. Jost, Phys. Rev. E **101**, 062313 (2020).

[4] G. Ferraz De Arruda, M. Tizzani, and Y. Moreno, Commun. Phys. **4**, 24 (2021).

[5] Á. Bodó, G. Y. Katona, and P. L. Simon, Bull. Math. Biol. **78**, 713 (2016).

[6] I. Iacopini, G. Petri, A. Barrat, and V. Latora, Nat. Commun. **10**, 2485 (2019).

[7] G. F. de Arruda, G. Petri, and Y. Moreno, Phys. Rev. Res. **2**, 023032 (2020).

[8] U. Alvarez-Rodriguez, F. Battiston, G. F. de Arruda, Y. Moreno, M. Perc, and V. Latora, Nat. Hum. Behav **5**, 586 (2021).

[9] A. P. Millán, J. J. Torres, and G. Bianconi, Phys. Rev. Lett. **124**, 218301 (2020).

[10] M. Lucas, G. Cencetti, and F. Battiston, Phys. Rev. Res. **2**, 033410 (2020).

[11] L. V. Gambuzza, F. Di Patti, L. Gallo, S. Lepri, M. Romance, R. Criado, M. Frasca, V. Latora, and S. Boccaletti, Nat. Commun. **12**, 1255 (2021).

[12] D. Zhou, J. Huang, and B. Schölkopf, in *Proceedings of the 19th International Conference on Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2006), pp. 1601–1608.

[13] U. Chitra and B. J. Raphael, in *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, Vol. 97 of *Proceedings of Machine Learning Research*, edited by K. Chaudhuri and R. Salakhutdinov (PMLR, Long Beach, CA, 2019), pp. 1172–1181.

[14] T. Carletti, F. Battiston, G. Cencetti, and D. Fanelli, Phys. Rev. E **101**, 022308 (2020).

[15] K. Hayashi, S. G. Aksoy, C. H. Park, and H. Park, in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Association for Computing Machinery, New York, NY, 2020), pp. 495–504.

[16] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019* (AAAI Press, Washington, DC, 2019), pp. 3558–3565.

[17] N. Yadati, M. Nimishakavi, P. Yadav, V. Nitin, A. Louis, and P. Talukdar, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019), Vol. 32.

[18] N. Masuda, M. A. Porter, and R. Lambiotte, Phys. Rep. **716-717**, 1 (2017).

[19] Z. Burda, J. Duda, J. M. Luck, and B. Waclaw, Phys. Rev. Lett. **102**, 160602 (2009).

[20] Y. Lin, A. Julaiti, and Z. Zhang, J. Chem. Phys. **137**, 124104 (2012).

[21] Y. Lin and Z. Zhang, Sci. Rep. **4**, 5365 (2014).

[22] Q. Guo, E. Cozzo, Z. Zheng, and Y. Moreno, Sci. Rep. **6**, 37641 (2016).

[23] Y. Lin and Z. Zhang, Comput. J. **62**, 63 (2018).

[24] J. Duda, J. Phys.: Conf. Ser. **361**, 012039 (2012).

[25] R. Sinatra, J. Gómez-Gardeñes, R. Lambiotte, V. Nicosia, and V. Latora, Phys. Rev. E **83**, 030103(R) (2011).

[26] J. K. Ochab and Z. Burda, Eur. Phys. J.: Spec. Top. **216**, 73 (2013).

[27] R.-H. Li, J. X. Yu, and J. Liu, in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (Association for Computing Machinery, New York, 2011), pp. 1147–1156.

[28] J.-C. Delvenne and A.-S. Libert, Phys. Rev. E **83**, 046117 (2011).

[29] A. Banerjee, Linear Algebra and its Applications **614**, 82 (2021).

[30] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri, Networks beyond pairwise interactions: Structure and dynamics, Phys. Rep. **874**, 1 (2020).

[31] T. Carletti, D. Fanelli, and R. Lambiotte, J. Phys.: Complex. **2**, 015011 (2021).

[32] C. Cooper, A. Frieze, and T. Radzik, in *Structural Information and Communication Complexity*, edited by A.

Kosowski and M. Yamashita (Springer, Berlin, 2011), pp. 210–221.
[33] P. S. Chodrow, J. Complex Netw. **8**, cnaa018 (2020).
[34] B. Jhun, M. Jo, and B. Kahng, J. Stat. Mech.: Theory Exp. (2019) 123207.
[35] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **87**, 278701 (2001).
[36] D.-S. Lee, K.-I. Goh, B. Kahng, and D. Kim, Nucl. Phys. B **696**, 351 (2004).
[37] G. Ferraz de Arruda, H. Ferraz de Arruda, and Y. Moreno, (unpublished).
[38] I. Amburg, N. Veldt, and A. R. Benson, *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)* (SIAM, Philadelphia, 2022), pp. 145–153.
[39] R. Mulas, C. Kuehn, T. Böhle, and J. Jost, Discr. Appl. Math. **317**, 26 (2022).
[40] https://github.com/pietrotraversa/random_walks_on_hypergraphs.
[41] https://www.cs.cornell.edu/~arb/data/.
[42] J. Ni, J. Li, and J. McAuley, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International JointConference on Natural Language Processing (EMNLP-IJCNLP)*, edited by K. Inui, J. Jiang, V. Ng, and X. Wan (Association for Computational Linguistics, Hong Kong, China, 2019), pp. 188–197.
[43] N. Craswell, A. P. de Vries, and I. Soboroff, in *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, 15-18 November 2005*, edited by E. M. Voorhees and L. P. Buckland (National Institute of Standards and Technology (NIST), 2005), Vol. 500-266 of NIST Special Publication.
[44] I. Amburg, J. Kleinberg, and A. R. Benson, J. Phys.: Complex. **2**, 035004 (2021).
[45] R. Mastrandrea, J. Fournet, and A. Barrat, PLoS ONE **10**, e0136497 (2015).
[46] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, Proc. Natl. Acad. Sci. USA **115**, E11221 (2018).
[47] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. V. den Broeck, C. Régis, B. Lina *et al.*, PLoS ONE **6**, e23176 (2011).