# DiNgHy: Null Models for Non-Degenerate Directed Hypergraphs

Maryam Abuissa[1][0009−0004−1454−5164] (✉), Matteo Riondato[2][0000−0003−2523−4420], and Eli Upfal[1][0000−0002−9321−9460]

[1] Brown University, Providence, RI 02912, USA
{maryam_abuissa,eliezer_upfal}@brown.edu
[2] Amherst College, Amherst, MA 01002, USA mriondato@amherst.edu

**Abstract.** Non-degenerate directed hypergraphs, i.e., directed hypergraphs where a node cannot be both in the tail and the head of a hyperedge, model important scenarios, from contact networks for analyzing the spread of information or diseases, to bill cosponsoring graphs for studying the bipartisanship of elected representatives. Existing null models for dihypergraphs allow degeneracy, and most samples drawn from them are degenerate, even when the starting network is not, making these models unrealistic in many cases. An inappropriate null model may lead to wrongly accepting/rejecting a hypothesis when performing statistical hypothesis testing. We introduce the first null models for non-degenerate dihypergraphs, and present DiNgHy, a suite of Markov-Chain-Monte-Carlo algorithms to sample from them. The Markov chain underlying our algorithm is not irreducible in general, so we give mild sufficient conditions for irreducibility. We show that existing methods cannot be used to sample from our null models, and evaluate our algorithms on real and artificial dihypergraphs, comparing the results of hypothesis tests when using our null models versus existing ones that allow degeneracy, and measuring their empirical mixing time.

**Keywords:** Graph Analysis · Hypothesis Testing · Markov Chain Monte Carlo · Network Science

## 1 Introduction

Hypergraphs overcome the limitations of dyadic (i.e., "classic") graphs to model "more-than-binary" relationships between entities [2, 3]. Such relationships are omnipresent in real scenarios, from co-authorship networks [16], to protein interactions [9], to contact networks [4]. There are many data analysis tasks and corresponding algorithms whose input is one or more hypergraphs [14].

The goal of knowledge discovery from data is to use results obtained from data to *identify new facts about the Data Generating Process (DGP)*, of which the available data is only a limited, noisy, random sample [21]. In the framework of statistical hypothesis testing [15], one first formulates a hypothesis, usually expressed as whether known properties of the DGP can sufficiently explain a result obtained from the available data. The hypothesis is then tested by comparing the observed result to the distribution of results over the datasets the DGP may produce. When there is a low probability that the DGP produces datasets with as or more extreme values than the observed value, it is seen as evidence against the hypothesis, i.e., that the known properties of the DGP cannot sufficiently explain the observed result. The DGP is formally captured by a *null model* (see Sect. 3.2 for definitions), a collection of datasets and a distribution over these datasets. The key *algorithmic* challenge is to develop efficient methods to draw samples from the collection according to the distribution. The samples create a distribution over the value of interest, to which the observed value can be compared. The *modeling* challenge is even more important for hypothesis testing: the null model must be *realistic*, i.e., it must capture as many known properties as possible about the DGP. In particular, it should not include any dataset that the DGP cannot produce. Failure to do so may lead to an inaccurate distribution of possible results, changing the outcome of hypothesis testing.

Null models for *undirected* hypergraphs are available (see Sect. 2), but limited attention has been devoted to null models for directed ones (dihypergraphs);

*Contributions* We introduce novel realistic null models for dihypergraphs, and give algorithms to sample from them.[3]

- Our null models are defined over non-degenerate dihypergraphs, i.e., dihypergraphs where a node cannot be present in both the head and the tail of the same hyperedge. This restriction is representative of many real scenarios, from U.S. Congress bill sponsorship, to contagion by contact, and is not captured by existing null models [20]. Thus, our null models are more realistic. They exactly preserve the in-/out-degree and the head-/tail-hyperedge-dimension sequences of an observed hypergraph, extending the popular microcanonical configuration model. We give null models for both edge-ordered and edge-unordered dihypergraphs, defined in Sect. 5.
- We describe and analyze Markov-Chain-Monte-Carlo (MCMC) algorithms, DiNgHy (edge-ordered) and DiNgHy-U (edge-unordered) (for DIrected Non-deGenerate HYpergraphs), to sample from the hypergraph ensembles of our null models according to any user-specified probability distribution, which is a necessary step in statistical hypothesis testing. Our Markov chains use simple transitions; the crux of our analysis is proving an easy-to-check mild condition on the degree and hyperedge-dimension sequences of the observed network that guarantees the irreducibility of the Markov chain.
- The results of our evaluation on real and artificial datasets highlights how the outcome of hypothesis tests may greatly differ between our null model and that of Preti et al. [20], We also give evidence that their algorithm cannot be used with rejection sampling to sample non-degenerate dihypergraphs. Finally, we show faster empirical mixing time of our algorithm compared to a baseline derived from the algorithm by Preti et al. [20].

[3] Theoretical proofs and additional experimental results are in the supplementary materials.

## 2 Related Work

Hypergraph mining has many applications in different settings. Lee et al. [14] survey the area in depth, so here we focus on the works most related to ours.

Many null models are available for dyadic (i.e., non-hyper) graphs, preserving different properties of the observed network, either exactly (microcanonical models) or in expectation (canonical models), together with algorithms, usually MCMC, to sample from these graph ensembles [7, 10, and references therein]. The null model over bipartite graphs with fixed degree sequences (a.k.a. the configuration model) has been deeply studied [10, 11]. We use pairs of bipartite graphs to define an equivalent representation of dihypergraphs, and build on a result by Kannan et al. [11] about the irreducibility of a Markov chain on bipartite graphs to show the irreducibility of our Markov chain on dihypergraphs.

Proving irreducibility is usually the key theoretical challenge in developing MCMC algorithms. For directed graphs, irreducibility is not guaranteed, because edge swaps cannot directly flip a directed triangle. Lamar [13] gives tight conditions on the degree sequence that imply irreducibility for digraphs. Similar issues arise on dihypergraphs, but a novel approach, and conditions on the observed degree and edge-dimension sequences, are required to prove irreducibility.

(Di-)hypergraphs have received relatively little attention, despite their practical importance [2, 3]. Many contributions study the configuration model for undirected hypergraphs [5, 6, 8, 24], or define maximum entropy models [23] or models that preserve higher order constraints [18, 19]. These approaches cannot be adapted to dihypergraphs.

Kim et al. [12] extend the preferential attachment model on hypergraphs [8] to dihypergraphs, preserving, *in expectation*, the node degrees and the hyperedge head- and tail- dimension sequences. As observed by Preti et al. [20], the generated networks do not resemble real ones, despite the adopted mechanism.

Preti et al. [20] introduce two microcanonical null models for (possibly degenerate) diypergraphs, and MCMC methods to sample from them. Both null models allow degeneracy; the first null model maintains the same properties as ours otherwise, and the other preserves additional constraints based on the joint degree distribution. Many DGPs would not create degenerate dihypergraphs (see Sect. 3.1), thus these null models would be unrealistic for such scenarios (see Sect. 6.2), potentially leading to invalid conclusions from statistical hypothesis tests (see Sect. 6.1).

We distinguish between edge-ordered dihypergraphs and edge-unordered dihypergraphs (see Sect. 3.1 for formal definitions). These concepts are not the same as those of vertex- and stub-labeled hypergraphs [5], which relate to the labeling of nodes. Rather, our distinction is related to the concepts of row-order-agnostic and row-order-aware null models introduced by Abuissa et al. [1] for transactional datasets (i.e., for binary matrices), but not immediately extendable to dihypergraphs (see Sect. 5).

# 3 Preliminaries

## 3.1 Directed Hypergraphs

A directed hypergraph (dihypergraph) $G \doteq (N, E)$ has a set $N = \{n_1, \ldots, n_{|N|}\}$ of nodes, and a multiset (a.k.a., a bag) $E = \{\{e_1, \ldots, e_{|E|}\}\}$ of directed hyperedges [20], where each hyperedge $e_i = (\mathsf{t}(e_i), \mathsf{h}(e_i))$ has a *tail* $\mathsf{t}(e_i) \subseteq N$ and a *head* $\mathsf{h}(e_i) \subseteq N$ (i.e., $e_i \in 2^N \times 2^N$). A hyperedge represents a relation from the nodes in the tail to those in the head, matching the graphical representation of directed edges as arrows, with the source being on the tail of the arrow, and the destination being on the arrowhead.[4] The head and tail of a hyperedge are both sets, not multisets, and they are assumed to be non-empty. Although the set constraint may be relaxed, to the best of our knowledge, it is not required or even reasonable to do so in most situations modeled by dihypergraphs. We denote the union of the head and the tail of a hyperedge $e$ as $\mathsf{n}(e)$.

For each $n \in N$, the *out-degree* $\mathsf{odeg}_G(n)$ (resp. the *in-degree* $\mathsf{ideg}_G(n)$) *of* $n$ *in* $G$ is the number of hyperedges in $E$ whose tails (resp. heads) contain $n$, i.e.,

$$\mathsf{odeg}_G(n) \doteq |\{e \in E : n \in \mathsf{t}(e)\}| \enspace .$$

(resp. $\mathsf{ideg}_G(n) \doteq |\{e \in E : n \in \mathsf{h}(e)\}|$). The *degree* $\mathsf{deg}_G(n)$ *of* $n$ *in* $G$ is the sum of its out- and in-degrees, $\mathsf{deg}_G(n) \doteq \mathsf{odeg}_G(n) + \mathsf{ideg}_G(n)$.

For each hyperedge $e \in E$, the *tail-dimension* $\mathsf{tdim}_G(e)$ (resp. *head-dimension* $\mathsf{hdim}_G(e)$) *of* $e$ *in* $G$ is the number of nodes in its tail, i.e., $\mathsf{tdim}_G(e) \doteq |\mathsf{t}(e)|$ (resp. in is head, i.e., $\mathsf{hdim}_G(e) \doteq |\mathsf{h}(e)|$). The *dimension or size* $\mathsf{dim}_G(e)$ *of* $e$ *in* $G$ is the sum of its tail and head dimensions, $\mathsf{dim}_G(e) \doteq \mathsf{tdim}_G(e) + \mathsf{hdim}_G(e)$.

**(Non-)Degenerate Dihypergraphs** We say that a dihypergraph $G = (N, E)$ is *degenerate* if there exist a node $n \in N$ and a hyperedge $e \in E$ s.t. $n$ belongs to both the tail and the head of $e$.[5] Many natural settings impose the requirement that dihypergraphs are not degenerate. For example, dihypergraphs can model U.S. Congress bill sponsorships, where the sponsor is in the tail of a hyperedge, and the cosponsor(s) are in the head [20]. The resulting dihypergraph is non-degenerate, since a representative cannot both sponsor and cosponsor the same bill. Similarly, when modeling contact networks in disease diffusion [20], the same entity cannot be in both the infecting and infected groups of one interaction. On the other hand, in a diypergraph where hyperedges represent citations between groups of authors, degeneracy will arise whenever there are self-citations. In this example, whether self-citation should be included depends on the analysis to be performed. Preti et al. [20] define a null model that includes degenerate dihypergraphs, while ours only includes non-degenerate dihypergraphs.

[4] Preti et al. [20] call "head" what we call "tail" and vice versa. We follow the convention for directed diadic graphs, due to the representation of directed edges as arrows.

[5] In *undirected* hypergraphs, the term "degenerate" denotes that a node appears multiple times in a hyperedge [5]. Our use is related: if we transform a degenerate dihypergraph into an undirected hypergraph by merging the head and tail of each hyperedge, the resulting undirected hypergraph will be degenerate.

**Edge-Ordered and Edge-Unordered Dihypergraphs** It is always assumed that each node of $N$ has a unique identifier, or label, and w.l.o.g., we can assume $N$ has a fixed, arbitrary total order. On the other hand, it may or may not be desirable to distinguish between identical hyperedges (recall that $E$ is a multiset), resulting in dihypergraphs that are effectively different mathematical objects. Assigning a unique identifier to hyperedges is equivalent to imposing that the set of hyperedges $E$ has a *fixed, arbitrary total order*, i.e., it is a sequence $E = \langle e_1, \ldots, e_{|E|} \rangle$. In this case, different orderings of $E$ lead to different dihypergraphs. This fact naturally leads to referring to dihypergraphs whose hyperedges have unique identifiers as *Edge-Ordered Dihypergraphs (EODs)*, and therefore to dihypergraphs whose hyperedges do not have unique identifiers as *Edge-Unordered Dihypergraphs (EUDs)*. The choice of whether to represent a network as an EOD or as an EUD must be deliberate, as they are different objects, which may lead to different outcomes for hypothesis tests performed on the different models, as discussed by Abuissa et al. [1] for binary matrices.

For any EOD $H$, we denote with $\mathsf{ideg}(H)$ (resp. $\mathsf{odeg}(H)$) the sequence of the in-degrees (resp. out-degrees) of its nodes, and with $\mathsf{tdim}(H)$ (resp. $\mathsf{hdim}(H)$) the sequence of the tail dimensions (resp. head dimensions) of its hyperedges.

### 3.2 Null Models and Hypothesis Testing

A *null model* $M = (\mathcal{D}, \pi)$ of dihypergraphs is a representation of the DGP: $\mathcal{D}$ is the collection of dihypergraphs that the DGP may generate, and $\pi$ is a probability distribution over $\mathcal{D}$. The DGP generates $G \in \mathcal{D}$ with probability $\pi(G)$. $\mathcal{D}$ is defined starting from an *observed dihypergraph* $\mathring{G}$ and a set $\mathcal{P}$ of functions over the set of all dihypergraphs. $\mathcal{P}$ represents structural properties of the datasets that the DGP may generate, e.g., the set of nodes, the number of hyperedges, and/or the number of (hypergraph) triangles. $\mathcal{D}$ contains all and only the dihypergraphs with the same values as $\mathring{G}$ for all properties in $\mathcal{P}$, including $\mathring{G}$.[6] $M$ is used to test whether the value $q(\mathring{G})$ of a property $q \notin \mathcal{P}$ can be explained by the properties in $\mathcal{P}$, by measuring the likelihood of observing a value as or more extreme than $q(\mathring{G})$ among the dihypergraphs in $\mathcal{D}$. Formally, we are interested in computing $p$-value of $q(\mathring{G})$, i.e., the probability that $q(G)$ is as or more extreme than $q(\mathring{G})$ when $G$ is sampled from $\mathcal{D}$ according to $\pi$. When the $p$-value is smaller than a critical value $\alpha$ chosen by the user, it allows the user to reject, with a confidence $1 - \alpha$, the hypothesis that $q(\mathring{G})$ is explained only by $\mathcal{P}$.

The $p$-value is hard to compute exactly except in the most simple cases [15], but an empirical $p$-value can be obtained through a Monte Carlo approach by drawing samples from $\mathcal{D}$ according to $\pi$, and using the empirical distribution of $q(\cdot)$ across the samples to approximate its true distribution. From a computational point of view, the key ingredient needed to obtain this approximation is an efficient algorithm that can sample from $\mathcal{D}$ according to $\pi$.

Our goal in this work is to develop such an algorithm for a specific choice of $\mathcal{P}$, where $\mathcal{D}$ is a collection of *non-degenerate* EODs or EUDs.

[6] $\mathcal{D}$, and therefore $M$, depends on $\mathring{G}$ but the notation does not, to keep it light.

## 4 A Null Model for Non-Degenerate EODs

We now introduce a null model for non-degenerate EODs, and present a Markov Chain Monte Carlo (MCMC) algorithm to sample from this model. In Sect. 5 we do the same for EUDs.

Given an observed *non-degenerate* EOD $\mathring{G} = (\mathring{N}, \mathring{E})$, we define the null model $M = (\mathcal{D}, \pi)$ where $\mathcal{D}$ contains all and only the EODs $G = (N, E)$ such that:

- $N = \mathring{N}$; and
- $\mathsf{ideg}(G) = \mathsf{ideg}(\mathring{G})$ and $\mathsf{odeg}(G) = \mathsf{odeg}(\mathring{G})$, i.e., the set of vertices and the in/out degree sequences are preserved; and
- $\mathsf{tdim}(G) = \mathsf{tdim}(\mathring{G})$ and $\mathsf{hdim}(G) = \mathsf{hdim}(\mathring{G})$, i.e., the head/tail dimension sequences are preserved; and
- $G$ is non-degenerate.

Although $\pi$ can be any distribution over $\mathcal{D}$, we focus on the *uniform* distribution in this paper, and present DiNgHy, an MCMC algorithm to sample uniformly from $\mathcal{D}$. This algorithm can be used to sample from $\mathcal{D}$ according to any distribution by using the Metropolis-Hastings approach [17, Ex.10.12].

Preti et al. [20] introduce a null model, along with a sampling algorithm, NuDHy, that preserves the first three properties but not the last one. Thus, even if $\mathring{G}$ is non-degenerate, the null model by Preti et al. [20] may contain both degenerate and non-degenerate dihypergraphs, which is often undesirable. A null model should capture everything known about the DGP as closely as possible. If it is known or assumed that the DGP would never produce a degenerate dihypergraph, then such dihypergraphs should not be included in $\mathcal{D}$, to avoid leading to incorrect outcomes when testing hypotheses. For example, consider the congress cosponsoring case described in Sect. 3.1, whose DGP would never produce degenerate dihypergraphs. Assume that we are interested in studying the likelihood that a U.S. Senator is a cosponsor of a bill whose first sponsor is the other Senator from the same state (each state as exactly two Senators). If we assume the null model by Preti et al. [20], the expectation over $\mathcal{D}$ of this likelihood is roughly doubled, i.e., the distribution of this quantity is completely different depending on the choice of whether to allow degeneracy in the null model. Since the distribution is different, the results of testing the hypothesis may also be different. In Sect. 6 we give experimental evidence of such issues.

One may be tempted to use the algorithm NuDHy by Preti et al. [20] as a subroutine in a rejection sampling scheme to draw samples uniformly from the space of non-degenerate dihypergraphs. This approach would not be successful, since non-degenerate dihypergraphs are extremely sparse within the space including degenerate dihypergraphs, as we show in Sect. 6. In fact, every sample drawn by NuDHy is degenerate, even when $\mathring{G}$ is not. We therefore introduce, in the next section, a new algorithm to sample directly from the space of non-degenerate dihypergraphs.

### 4.1 Sampling uniformly from the null model

Our MCMC algorithm DiNgHy draws uniform samples from $\mathcal{D}$ by running a Markov Chain (MC) whose set of states is $\mathcal{D}$ and whose stationary distribution is uniform. We start by describing the directed graph $\mathcal{G}$ of the states, i.e., for which ordered pairs $(G, G') \in \mathcal{D} \times \mathcal{D}$ the transition probability from $G$ to $G'$ is strictly positive, and we show that this graph is strongly connected (Thm. 1). We then present an algorithm to draw the next state of the MC from the current state, and we prove that the resulting transition probabilities lead to a uniform stationary distribution for the MC. Before these steps, we introduce an equivalent representation of EODs that we use throughout this section.

**EODs as ordered pairs of bipartite graphs** Any EOD $G = (N, E)$ can be represented as an ordered pair $(\mathsf{B}_\mathrm{t}(G), \mathsf{B}_\mathrm{h}(G))$ of bipartite graphs $\mathsf{B}_\mathrm{t}(G) = (N, E, \mathsf{E}_\mathrm{t}(G))$ and $\mathsf{B}_\mathrm{h}(G) = (N, E, \mathsf{E}_\mathrm{h}(G))$, where $N$ and $E$ are the two sets of *vertices* of these bipartite graphs,[7]. There is an edge $(n, e) \in \mathsf{E}_\mathrm{t}(G)$ in the bipartite graph $\mathsf{B}_\mathrm{t}(G)$ iff $n$ is in the *tail* of $e$, for $n \in N$ and $e \in E$, and similarly for $\mathsf{E}_\mathrm{h}(G)$ and the head. Formally, $\mathsf{E}_\mathrm{t}(G) \doteq \{(n, e) \in N \times E : n \in \mathsf{t}(e)\}$ and $\mathsf{E}_\mathrm{h}(G) \doteq \{(n, e) \in N \times E : n \in \mathsf{h}(e)\}$.

**Fact 1** *An EOD G has a unique representation as an ordered pair* $(\mathsf{B}_t(G), \mathsf{B}_h(G))$ *of bipartite graphs.*

On the other hand, not every pair of bipartite graphs $(T, H)$ with the same sets of vertices is a representation of a *non-degenerate* EOD: the edge sets of $T$ and $H$ must be disjoint for the corresponding EOD to be non-degenerate.

We can transform the set of properties $\mathcal{P}$ that define $\mathcal{D}$ into a set of properties $\mathcal{P}'$ over pairs $(T, H)$ of bipartite graphs which have $\mathring{N}$ and $\mathring{E}$ as their sets of left and right vertices:

- for every $n \in \mathring{N}$, the degree of $n$ in $T$ (resp. in $H$) is $\mathsf{odeg}_{\mathring{G}}(n)$ (resp. $\mathsf{ideg}_{\mathring{G}}(n)$); and
- for every $e \in \mathring{E}$, the degree of $e$ in $T$ (resp. in $H$) is $\mathsf{hdim}_{\mathring{G}}(n)$ (resp. $\mathsf{tdim}_{\mathring{G}}(n)$); and
- the EOD corresponding to $(T, H)$ must be non-degenerate, meaning there are no $n \in \mathring{N}, e \in \mathring{E}$ such that $(n, e)$ is an edge of both $T$ and $H$.

We can then define the set $\mathcal{D}_\mathsf{B}$ of pairs of bipartite graphs as

$$\mathcal{D}_\mathsf{B} \doteq \{(\mathsf{B}_\mathrm{t}(G), \mathsf{B}_\mathrm{h}(G)) : G \in \mathcal{D}\} \ .$$

There is a bijection between $\mathcal{D}_\mathsf{B}$ and $\mathcal{D}$. In the rest of this section we define an MC on $\mathcal{D}_\mathsf{B}$ whose stationary distribution is uniform, since a uniform sample from $\mathcal{D}_\mathsf{B}$ corresponds to a uniform sample from $\mathcal{D}$.

[7] To avoid confusion, we use *nodes* for dihypergraphs, and *vertices* for bipartite graphs. Conceptually, the vertices in the bipartite graphs are the nodes in $G$ and the *identifiers* of the hyperedges of $G$.

**The directed graph of the states** To describe the directed graph $\mathcal{G} = (\mathcal{D}_{\mathsf{B}}, \mathcal{E})$ of states, we first define the *swap* (a.k.a. switch [11]), an operation that transforms a bipartite graph $B$ into another bipartite graph $B'$ with the same degree sequences as $B$. We then restrict swaps to define an operation on $\mathcal{D}_{\mathsf{B}}$ that transforms a pair $(T, H) \in \mathcal{D}_{\mathsf{B}}$ into another pair $(T', H') \in \mathcal{D}_{\mathsf{B}}$.

**Definition 1.** *Given a bipartite graph $T = (V, U, W)$, let $v_1, v_2 \in V$, $v_1 \neq v_2$, and $u_1, u_2 \in U$, $u_1 \neq u_2$ such that $e_1 \doteq (v_1, u_1), e_2 \doteq (v_2, u_2) \in W$ and $e'_1 \doteq (v_1, u_2), e'_2 \doteq (v_2, u_1) \notin W$. The* swap $\mathsf{s}_T(e_1, e_2)$ *is the operation that transforms $T$ into the bipartite graph $T' = (V, U, W')$ where $W' = (W \setminus \{e_1, e_2\}) \cup \{e'_1, e'_2\}$. We refer to $\mathsf{s}_T(e_1, e_2)$ as a swap from $T$ to $T'$.*

We now define the *Non-Degenerating Swap (NDS)* operation from $\mathcal{D}_{\mathsf{B}}$ to itself. There are two kinds of NDSs, the the *Tail-Non-Degenerating Swap (TNDS)* and the *Head-Non-Degenerating Swap (HNDS)*.

**Definition 2.** *Let $(T, H) \in \mathcal{D}_{\mathsf{B}}$, with $T = (\mathring{N}, E, Q)$ and $H = (\mathring{N}, E, Z)$. Let $\ell_1 = (n_1, e_1)$, $\ell_2 = (n_2, e_2) \in Q$ be two edges in $T$ such that $\mathsf{s}_T(\ell_1, \ell_2)$ is a swap from $T$ to some $T'$, and such that $(n_1, e_2), (n_2, e_1) \notin Z$. The* Tail-Non-Degenerating Swap (TNDS) $\mathsf{ts}_{T,H}(\ell_1, \ell_2)$ *is the operation that transforms $(T, H)$ into $(T', H) \in \mathcal{D}_{\mathsf{B}}$ by applying $\mathsf{s}_T(\ell_1, \ell_2)$ to $T$.*

**Definition 3.** *Let $(T, H) \in \mathcal{D}_{\mathsf{B}}$, with $T = (\mathring{N}, E, Q)$ and $H = (\mathring{N}, E, Z)$. Let $r_1 = (n_1, e_1)$, $r_2 = (n_2, e_2) \in Z$ be two edges in $H$ such that $\mathsf{s}_H(r_1, r_2)$ is a swap from $H$ to some $H'$, and such that $(n_1, e_2), (n_2, e_1) \notin Q$. The* Head-Non-Degenerating Swap (HNDS) $\mathsf{hs}_{T,H}(r_1, r_2)$ *is the operation that transforms $(T, H)$ into $(T, H') \in \mathcal{D}_{\mathsf{B}}$ by applying $\mathsf{s}_H(r_1, r_2)$ to $H$.*

In the directed graph $\mathcal{G} = (\mathcal{D}_{\mathsf{B}}, \mathcal{E})$, there is an edge from $(T, H) \in \mathcal{D}_{\mathsf{B}}$ to $(T', H') \in \mathcal{D}_{\mathsf{B}}$ (it must hold either $T = T'$ or $H = H'$) if there is a NDS from $(T, H)$ to $(T', H')$. It is evident there can be at most one NDS between any two states. Any NDS is reversible, i.e., if there is a NDS $q$ from $(T, H)$ to $(T', H')$, then there is a NDS $\mathsf{rev}(q)$ (the *reversal* of $q$) of the same type (i.e., head- or tail-) from $(T', H')$ to $(T, H)$. Thus we say that $(T, H)$ and $(T', H')$ are neighbors.

There exist dihypergraphs, like digraphs, where $\mathcal{G}$ is not strongly connected under NDSs (consider flipping a directed triangle). Irreducibility under the edge swap on digraphs requires complex conditions [13]. Dihypergraphs face a similar issue, but the proof for digraphs does not extend to dihypergraphs. We use a novel approach to show that under mild conditions, $\mathcal{G}$ is strongly connected, which is necessary for the MC to have a unique stationary distribution. We first need some technical definitions.

Let $(T, H) \in \mathcal{D}_{\mathsf{B}}$, with $T = (\mathring{N}, E, W)$ and $H = (\mathring{N}, E, Z)$. For an edge $w \in W$ (resp. $z \in Z$), let $\mathsf{tse}_{T,H}(w)$ (resp. $\mathsf{hse}_{T,H}(z)$) be the set of edges $w' \in W$ (resp. $z' \in Z$) such that $\mathsf{ts}_{T,H}(w, w')$ is a TNDS (resp. $\mathsf{hs}_{T,H}(z, z')$ is a HNDS.)

Now let $(T, H') \in \mathcal{D}_{\mathsf{B}}$ be distinct from $(T, H)$ (i.e., $H \neq H'$). For $w \in W$, let $\mathsf{ttse}_{T,H,H'}(w)$ be the set of edges $w'$ in $W$ such that $\mathsf{ts}_{T,H}(w, w')$ is a TNDS on $(T, H)$ *and* a TNDS on $(T, H')$. Let $(T', H) \in \mathcal{D}_{\mathsf{B}}$ be distinct from $(T, H)$ (i.e.,

$T \neq T'$). Similarly, for $z \in Z$, let $\mathsf{hhse}_{H,T,T'}(h)$ be the set of edges $z'$ in $Z$ such that $\mathsf{hs}_{T,H}(z, z')$ is a HNDS on $(T, H)$ *and* a HNDS on $(T', H)$.

**Theorem 1.** *Assume that at least one of the two following pairs of conditions hold for every $(T, H) \in \mathcal{D}_{\mathsf{B}}$:*

**Pair 1** (1) *for every edge $z$ in $H$, $|\mathsf{hse}_{T,H}(z)| \geq 1$; and* (2) *for every edge $z$ in $H$ and for every $(T, H') \in \mathcal{D}_{\mathsf{B}}$, $|\mathsf{ttse}_{T,H,H'}(z)| \geq 1$.*

**Pair 2** (1) *for every edge $w$ in $T$, $|\mathsf{tse}_{T,H}(w)| \geq 1$; and* (2) *for every edge $w$ in $T$ and for every $(T', H) \in \mathcal{D}_{\mathsf{B}}$, $|\mathsf{hhse}_{H,T,T'}(w)| \geq 1$.*

*Then the directed state graph is strongly connected.*

The intuition behind the proof is that for any $(T, H), (T', H') \in \mathcal{D}_{\mathsf{B}}$, we first construct a sequence of NDSs, possibly a mix of TNDSs and HDNSs, from $(T, H)$ to some $(T', H'')$. If $H'' \neq H'$, we then build another sequence of NDSs from $(T', H'')$ to some $(T''', H')$. If $T''' \neq T'$, the final step is a sequence of TNDSs from $(T''', H')$ to $(T', H')$ that "undo" the TNDSs from the second phase, i.e., they are the reversals of TNDSs from the second phase applied in reverse order.

*Proof.* Let us first assume that at least the first pair of conditions hold. We later adapt the proof to the case when only the second pair of conditions hold.

Our proof works in three phases. Given any two $(T, H), (T', H') \in \mathcal{D}_{\mathsf{B}}$, we first construct a sequence of NDSs, possibly a mix of TNDSs and HDNSs, from $(T, H)$ to some $(T', H'')$. We are done iff $H'' = H'$. Otherwise, in the second phase, we build another sequence of NDSs, again possibly a mix of HNDSs and TNDSs, from $(T', H'')$ to some $(T''', H')$. It holds $T''' = T'$ iff only HNDSs appear in this second sequence, in which case we are done. Otherwise, the third phase involves a sequence of TNDSs from $(T''', H')$ to $(T', H')$. The TNDSs in this third sequence are the reversals of TNDSs from the second sequence, applied in reverse order, i.e., the reversals of TNDSs that were applied later in the second sequence are applied earlier in the third sequence.

We use a result by Kannan et al. [11] as a blackbox in the following way. Given any two bipartite graphs $B = (L, R, E)$ and $B' = (L, R, E')$ with the same degree sequences, Kannan et al. [11, Lemma 3.1] show how to obtain a sequence of swaps that transforms $B$ into $B'$ by moving closer to $B'$ at every step, in the sense that the next swap in the sequence transforms the current graph $B'' = (L, R, E'')$ into $B''' = (L, R, E''')$ such that $|E' \cap E''| < |E' \cap E'''|$.

If $T \neq T'$, we start our first phase, and obtain, using the method by Kannan et al. [11], a sequence of swaps that would transform $T$ into $T'$. Let $(T^{\mathrm{c}}, H^{\mathrm{c}})$ be the current pair (at the beginning $(T^{\mathrm{c}}, H^{\mathrm{c}}) = (T, H)$). As long as the next proposed swap in the sequence is a TNDS on $(T^{\mathrm{c}}, H^{\mathrm{c}})$, we apply it. If the proposed swap $\mathsf{s}_T((n_1, e_1), (n_2, e_2))$ is not a TNDS on $(T^{\mathrm{c}}, H^{\mathrm{c}})$, then it must be that at least one of $(n_1, e_2)$ and $(n_2, e_1)$, possibly both, is an edge in $H^{\mathrm{c}}$. By appropriately transforming $H^{\mathrm{c}}$ through one or two HNDSs, the proposed swap will become a TNDS on $(T^{\mathrm{c}}, H^{\mathrm{c}})$. Indeed, if $(n_1, e_2)$ is an edge in $H^{\mathrm{c}}$, we can take any $(n, e) \in \mathsf{hse}_{T^{\mathrm{c}},H^{\mathrm{c}}}((n_1, e_2))$, which exists by the first condition in the hypothesis, and apply the swap $\mathsf{s}_{H^{\mathrm{c}}}((n_1, e_2), (n, e))$ to $H^{\mathrm{c}}$. This swap is a HNDS by definition of $\mathsf{hse}_{T^{\mathrm{c}},H^{\mathrm{c}}}((n_1, e_2))$. We proceed similarly if $(n_2, e_1)$ is an edge of $H^{\mathrm{c}}$. The current pair $(T^{\mathrm{c}}, H^{\mathrm{c}})$ is now such that the swap $\mathsf{s}_T((n_1, e_1), (n_2, e_2))$ is a TNDS on $(T^{\mathrm{c}}, H^{\mathrm{c}})$, so we can apply it. By repeating the process with the next swap proposed in the sequence, we arrive at a pair $(T', H'') \in \mathcal{D}_{\mathsf{B}}$. If all the proposed swaps were TNDSs, then it must be that $H'' = H'$, otherwise it is possible that $H'' \neq H'$.

If $H'' \neq H'$, we enter the second phase. We obtain, using the method by Kannan et al. [11], a sequence of swaps that would transform $H''$ into $H'$. Let $(T^{\mathrm{c}}, H^{\mathrm{c}})$ be the current pair, initialized as $(T^{\mathrm{c}}, H^{\mathrm{c}}) = (T', H'') \in \mathcal{D}_{\mathsf{B}}$. As long as the next proposed swap in the sequence is a HNDS on $(T^{\mathrm{c}}, H^{\mathrm{c}})$, we apply it. Consider now the case when the proposed swap $\mathsf{s}_H((n_1, e_1), (n_2, e_2))$ is not a HNDS on $(T^{\mathrm{c}}, H^{\mathrm{c}})$. It is a swap proposed by the method by Kannan et al. [11], so at least one of $(n_1, e_2)$ and $(n_2, e_1)$ is an edge in $H'$. Assume, w.l.o.g., that $(n_1, e_2)$ is an edge in $H'$. Thus, it must be that $(n_1, e_2) \notin T^{\mathrm{c}}$, because $(T^{\mathrm{c}}, H') \in \mathcal{D}_{\mathsf{B}}$ by construction, and if $(n_1, e_2)$ were in $T^{\mathrm{c}}$, the dihypergraph $G$ such that $\mathsf{B}_{\mathrm{t}}(G) = T^{\mathrm{c}}$ and $\mathsf{B}_{\mathrm{h}}(G) = H'$ would be degenerate, i.e., it would be $(T^{\mathrm{c}}, H') \notin \mathcal{D}_{\mathsf{B}}$, which would be a contradiction. For the swap $\mathsf{s}_H((n_1, e_1), (n_2, e_2))$ not to be a HNDS on $(T^{\mathrm{c}}, H^{\mathrm{c}})$, it must be that $(n_2, e_1)$ is an edge in $T^{\mathrm{c}}$. Let then $(n, e) \in \mathsf{ttse}_{T^{\mathrm{c}},H^{\mathrm{c}},H'}((n_2, e_1))$, which exists by the second condition in the hypothesis, and apply the swap $\mathsf{s}_{T^{\mathrm{c}}}((n_2, e_1), (n, e))$ to $T^{\mathrm{c}}$. This swap is guaranteed to be a TNDS from the definition of $\mathsf{ttse}_{T^{\mathrm{c}},H^{\mathrm{c}},H'}((n_2, e_1))$. The current pair $(T^{\mathrm{c}}, H^{\mathrm{c}})$ is now such that the swap $\mathsf{s}_H((n_1, e_1), (n_2, e_2))$ is a HNDS on $(T^{\mathrm{c}}, H^{\mathrm{c}})$, so we can apply it. By repeating the process with the next swap proposed in the sequence, we arrive at a pair $(T'', H') \in \mathcal{D}_{\mathsf{B}}$. If all the proposed swaps were HNDSs, then it must be that $T'' = T'$, otherwise it is possible that $T'' \neq T'$.

If $T'' \neq T'$, we enter the third phase. In this phase, we apply, in reverse order, the reversals of the TNDSs performed in the second phase, so we end up at $(T', H')$. Consider the ordered sequence $\mathsf{s} = \langle q_1, \ldots, q_\ell \rangle$ of the TNDSs performed in the second phase (if we are in the third phase, this sequence must be nonempty), and now consider the sequence $\mathsf{rs}$ of TNDSs obtained by flipping the order of the sequence, and replacing each TNDSs with its reversal, i.e., $\mathsf{rs} = \langle \mathsf{rev}(q_\ell), \ldots, \mathsf{rev}(q_1) \rangle$. When we applied the TNDS $q_\ell$ during the second phase, we moved from some $(\tilde{T}, \tilde{H})$ to $(T'', \tilde{H})$. The TNDS $q_\ell = \mathsf{ts}_{\tilde{T},\tilde{H}}(z, y)$ belong to $\mathsf{ttse}_{\tilde{T},\tilde{H},H'}(z)$, by construction. This TNDS $q_\ell$ is therefore a TNDS on $(\tilde{T}, \tilde{H})$ and on $(\tilde{T}, H')$, by definition of $\mathsf{ttse}_{\tilde{T},\tilde{H},H'}(z)$. In particular, if we applied it to $(\tilde{T}, H')$, we would move to $(T''', H')$. Thus, by applying $\mathsf{rev}(q_\ell)$ to $(T''', H')$, we move to $(\tilde{T}, H')$, and $\mathsf{rev}(q_\ell)$ is a TNDS on $(T'', H')$ because $q_\ell$ is a TNDS, and every NDS is reversible. We can repeat this reasoning for $q_{\ell-1}$ and $\mathsf{rev}(q_{\ell-1})$: when we apply the TNDS $\mathsf{rev}(q_{\ell-1})$ to $(\tilde{T}, H')$ we obtain $(\hat{T}, H')$, where $\hat{T}$ is such that during the second phase we applied $q_{\ell-1}$ to $(\hat{T}, \hat{H})$ to obtain $(\tilde{T}, \hat{H})$. Continuing this way, when we applied $q_1$ during the second phase we moved from $(T', \check{H})$ to $(\check{T}, \check{H})$, hence when we apply $\mathsf{rev}(q_1)$ to $(\check{T}, H')$ we move to $(T', H')$. Thus, there is a sequence of NDSs from any $(T, H) \in \mathcal{D}_{\mathsf{B}}$

to any other $(T', H') \in \mathcal{D}_{\mathsf{B}}$ if the first pair of conditions in the hypothesis holds, i.e., in this case the graph is strongly connected.

When only the second pair of conditions holds, we can adapt the proof by first "fixing" $H$ to go to $(T'', H')$, then move to $(T', H'')$ in the second phase, and finally to $(T', H')$ in the third phase. □

The following result gives an easy-to-compute condition on the degree and hyperedge dimension sequences of the observed non-degenerate EOD $\mathring{G}$ for the hypothesis of Thm. 1 to hold. It is a corollary of the four technical lemmas that are stated and proved in the supplementary materials. The quantities $\mathsf{odeg}(\cdot)$ and $\mathsf{ideg}(\cdot)$ can be switched to obtain another sufficient condition.

**Corollary 1.** *Let $n^*$ be the node with maximum degree in $\mathring{G}$, and $e^*$ be the hyperedge with maximum dimension in $\mathring{G}$. If the following condition holds:*

$(\mathsf{dim}_{\mathring{G}}(e^*) - 1)(2\mathsf{deg}_{\mathring{G}}(n^*) - 1) + 1 < \left\| \mathsf{ideg}(\mathring{G}) \right\|_1$*; and*

$2(\mathsf{dim}_{\mathring{G}}(e^*) - 1)(2\mathsf{deg}_{\mathring{G}}(n^*) - 1) + 1 < \left\| \mathsf{odeg}(\mathring{G}) \right\|_1$*;*

*then the condition from Thm. 1 holds.*

Our analysis is constrained by the path proposed by Kannan et al. [11]. As a result, we conjecture that the conditions of both Thm. 1 and Corol. 1 could be tightened using a more carefully tailored sequence of switches.

**Drawing the next state of the Markov Chain** We now present an algorithm that, given a pair $(T, H) \in \mathcal{D}_{\mathsf{B}}$ representing the current state of the MC, draws a neighbor $(T', H') \in \mathcal{D}_{\mathsf{B}}$ of $(T, H)$. We then show that the transition probabilities resulting from this algorithm lead to a uniform stationary distribution over $\mathcal{D}_{\mathsf{B}}$.

The pseudocode of the algorithm is presented in Alg. 1. We start by drawing an unordered pair of *distinct* hyperedges $(e_1, e_2) \in E \times E$ uniformly at random (u.a.r.) from the set of such pairs (Alg. 1). We then decide whether to perform a HNDS or a TNDS involving these hyperedges, by flipping a biased coin with a probability of heads $b$, a user-specified parameter (Alg. 1).[8] If the outcome is *heads*, and there is at least one HNDS involving $e_1$ and $e_2$ on $(T, H)$, we select one by first drawing a node $n_1$ u.a.r. from $\mathsf{h}(e_1) \setminus \mathsf{n}(e_2)$ (Alg. 1), and then similarly drawing $n_2$ u.a.r. from $\mathsf{h}(e_2) \setminus \mathsf{n}(e_1)$ (Alg. 1). The sets we draw from ensure that the resulting swap is a HNDS. We then perform the HNDS $\mathsf{hs}_{T,H}((n_1, e_1), (n_2, e_2))$ on $(T, H)$ to obtain $(T', H')$, which is returned (Alg. 1). If there is no HNDS involving $e_1$ and $e_2$, we take a self-loop from the state $(T, H)$ to itself (Alg. 1). If the outcome of the biased coin was tails, we proceed in a similar fashion with a TNDS (lines 10–16). Any neighbor of $(T, H)$ can be returned in output by Alg. 1.

[8] In our experiments we use $b = \left\| \mathsf{odeg}(\mathring{G}) \right\|_1 / (\left\| \mathsf{odeg}(\mathring{G}) \right\|_1 + \left\| \mathsf{ideg}(\mathring{G}) \right\|_1)$, which is a heuristic value to roughly balance the number of HNDSs and TNDSs performed.

**Algorithm 1:** Drawing the next state of the MC

```
Input: the current state $(T, H) \in \mathcal{D}_{\mathsf{B}}$, with $T = (\mathring{N}, E, W)$, and
       $H = (\mathring{N}, E, Z)$, the coin heads probability $b$
Output: the next state $(T', H') \in \mathcal{D}_{\mathsf{B}}$
$(e_1, e_2) \leftarrow$ unordered pair of distinct hyperedges in $E$ chosen u.a.r.
flip $\leftarrow$ outcome of a flip a biased coin with heads probability $b$
if flip is heads then
    if $\mathsf{h}(e_1) \setminus \mathsf{n}(e_2) \neq \emptyset$ and $\mathsf{h}(e_2) \setminus \mathsf{n}(e_1) \neq \emptyset$ then
        $n_1 \leftarrow$ node drawn u.a.r. from $\mathsf{h}(e_1) \setminus \mathsf{n}(e_2)$
        $n_2 \leftarrow$ node drawn u.a.r. from $\mathsf{h}(e_2) \setminus \mathsf{n}(e_1)$
        $(T, H') \leftarrow$ result of applying $\mathsf{hs}_{T,H}((n_1, e_1), (n_2, e_2))$ on $(T, H)$
        return $(T, H')$
    else return $(T, H)$
else
    if $\mathsf{t}(e_1) \setminus (\mathsf{h}(e_2) \cup \mathsf{t}(e_2)) \neq \emptyset$ and $\mathsf{t}(e_2) \setminus (\mathsf{h}(e_1) \cup \mathsf{t}(e_1)) \neq \emptyset$ then
        $n_1 \leftarrow$ node drawn u.a.r. from $\mathsf{t}(e_1) \setminus \mathsf{n}(e_2)$
        $n_2 \leftarrow$ node drawn u.a.r. from $\mathsf{t}(e_2) \setminus \mathsf{n}(e_1)$
        $(T', H) \leftarrow$ result of applying $\mathsf{ts}_{T,H}((n_1, e_1), (n_2, e_2))$ on $(T, H)$
        return $(T', H)$
    else return $(T, H)$
```

**Stationary distribution** This result shows that the transition probabilities arising from Alg. 1 allow us to sample uniformly from $\mathcal{D}_{\mathsf{B}}$. It relies on the transition matrix being doubly-stochastic.

**Theorem 2.** *The unique stationary distribution of the MC is uniform over* $\mathcal{D}_{\mathsf{B}}$.

We can then use the MC as part of our MCMC algorithm DiNgHy to sample uniformly from $\mathcal{D}$, by running the MC starting from $\mathring{G}$ until it mixes, and taking the state of the MC at that point as the sample from $\mathcal{D}$.

## 5 A Null Model for Non-Degenerate Edge-Unordered Hypergraphs

We extend the null model introduced in Sect. 4 to non-degenerate EUDs, to obtain a null model $(\mathcal{D}_{\mathrm{U}}, \pi)$, where $\mathcal{D}_{\mathrm{U}}$ is the set of all the non-degenerate EUDs with the same in-/out-degree and head-/tail-hyperedge dimension sequence as the observed non-degenerate EUD $\mathring{G}$. The algorithm DiNgHy presented in Sect. 4.1 is easily modified as follows, to obtain an algorithm DiNgHy-U for sampling uniformly from $\mathcal{D}_{\mathrm{U}}$.

There is a surjective function $\mathsf{o2u}(\cdot)$ from $\mathcal{D}$ to $\mathcal{D}_{\mathrm{U}}$, mapping any EOD $G \in \mathcal{D}$ to the EUD $\mathsf{o2u}(G) \in \mathcal{D}_{\mathrm{U}}$ obtained by removing the hyperedge identifiers from $G$. For any EUD $G'$, we denote with $\mathsf{o2u}^{-1}(G')$ the inverse image of $G'$ through $\mathsf{o2u}()$, i.e., $\mathsf{o2u}^{-1}(G') = \{\text{EOD } G : \mathsf{o2u}(G) = G'\}$. The following result gives an expression for $\left|\mathsf{o2u}^{-1}(G')\right|$.

**Lemma 1.** *Let $G = (N, E)$ be a EUD. Let $t^* \doteq \max_{e \in E} \mathsf{tdim}_G(e)$ and $h^* \doteq \max_{e \in E} \mathsf{hdim}_G(e)$. For $1 \leq i \leq t^*$ and $1 \leq j \leq h^*$, let $E_{i,j} \doteq \{\!\{e \in E : \mathsf{tdim}_G(e) = i \wedge \mathsf{hdim}_G(e) = j\}\!\}$ be the multiset of hyperedges with tail dimension $i$ and head dimension $j$. Let $\bar{E}_{i,j} \doteq \{e_{i,j,1}, \ldots, e_{i,j,u_{i,j}}\}$ be the* set *of such hyperedges, and $\bar{E}$ be the set version of $E$. For $1 \leq k \leq u_{i,j}$, let $w_{i,j,k} \doteq \mathsf{m}_E(e_{i,j,k})$ be the multiplicity of $e_{i,j,k}$ in $E$. Then, the number $\left|\mathsf{o2u}^{-1}(G)\right|$ of edge-labeled hypergraphs mapped to $G$ by $\mathsf{o2u}(\cdot)$ is*

$$\left|\mathsf{o2u}^{-1}(G)\right| = \prod_{i=1}^{t^*} \prod_{j=1}^{h^*} \binom{|E_{i,j}|}{w_{i,j,1}, \ldots, w_{i,j,u_{i,j}}} = \frac{\prod_{i=1}^{t^*} \prod_{j=1}^{h^*} |E_{i,j}|!}{\prod_{e \in \bar{E}} \mathsf{m}_G(e)!} \quad . \tag{1}$$

We can use Lemma 1 and the Metropolis-Hastings (MH) approach to modify the stationary distribution $\pi$ of the MC over EODS from Sect. 4.1, so that, for every EOD $G \in \mathcal{D}$, it holds

$$\pi(G) = \frac{1}{|\mathcal{D}_{\mathrm{U}}| \left|\mathsf{o2u}^{-1}(\mathsf{o2u}(G))\right|} \quad . \tag{2}$$

The MC is modified as follows to achieve the above. At every step, an EOD $B$ is *proposed* as the next state of the MC by drawing it from the neighbors of the current state $A$ wrt the original transition probabilities of the MC. $B$ is accepted as the next state of the MC with probability

$$\min\left\{1, \frac{\pi(B) p_{B,A}}{\pi(A) p_{A,B}}\right\} = \min\left\{1, \frac{\left|\mathsf{o2u}^{-1}(\mathsf{o2u}(A))\right|}{\left|\mathsf{o2u}^{-1}(\mathsf{o2u}(B))\right|}\right\},$$

as, per Thm. 2, the transition probabilities are symmetric in our original MC (i.e., $p_{B,A} = p_{A,B}$). The correctness of the MH approach guarantees that the stationary distribution of this modified MC is as in eq. (2). Thus we can sample an EOD $G \in \mathcal{D}$ using this modified MC, and consider the EUD $\mathsf{o2u}(G)$ as a uniform sample from $\mathcal{D}_{\mathrm{U}}$, because for every EUD $G' \in \mathcal{D}_{\mathrm{U}}$ the probability that it is sampled is

$$\sum_{G \in \mathsf{o2u}^{-1}(G')} \pi(G) = \frac{1}{|\mathcal{D}_{\mathrm{U}}|} \quad .$$

## 6 Experimental Evaluation

The goal of our experimental evaluation is threefold:

- assess the structural differences between the space of non-degenerate EODS we introduce and the space where degeneracy is allowed, proposed by Preti et al. [20]. In particular, we aim to understand the presence and amount of degeneracy in samples from the latter, and whether the outcome of hypothesis tests differ depending on which null model is used (results in Sect. 6.1);
- study the behavior of NuDHy [20] when the observed EOD is non-degenerate, to evaluate the possibility of using this algorithm as a subroutine in a rejection sampling scheme to sample non-degenerate EOD (results in Sect. 6.1);
- evaluate the empirical mixing time of our algorithm DiNgHy, compared to NuDHy and a baseline DiNgHy-B (results in Sect. 6.2)

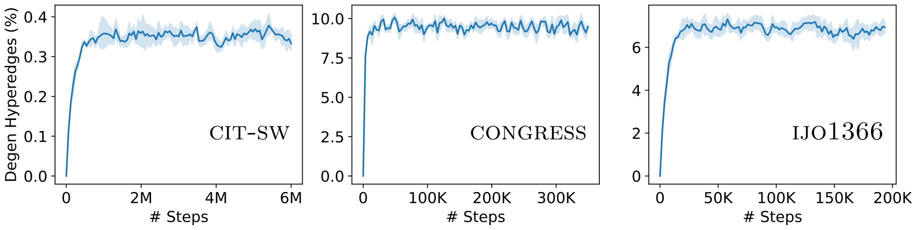


Fig. 1: Mean percentage (over 33 samples, 95% confidence interval shaded) of degenerate hyperdeges as function of the number of steps of the Markov chain.

*Implementation and datasets* We use Preti et al.'s publicly available implementation of NuDHy [20] (specifically, the NuDHy-Degs variant), which is a standard swap chain on the bipartite representation of a dihypergraph. Off of this codebase, we implement our algorithm DiNgHy, and a baseline DiNgHy-B used to evaluate the mixing time.[9] DiNgHy-B uses the same approach as NuDHy with the additional constraint that swaps do not cause degeneracy (i.e., they are NDSs). DiNgHy takes a novel approach to select NDSs (see Sect. 4.1).

We use datasets used by Preti et al. [20] and synthetic datasets.[10]

### 6.1 Difference between the null models

For all datasets, we measured the percentage of samples with at least one degenerate edge on 500 samples from the space that allows degeneracy, drawn using NuDHy. For every dataset, *all* 500 samples included degeneracy, even when the dataset was non-degenerate.

Figure 1 shows the percentage of hyperedges that are degenerate as a function of the number of steps taken by the Markov chain when starting from the observed network. This quantity increases sharply by the first measurement, and stabilizes soon after, without ever disappearing, i.e., at *every* measurement, the current state of the Markov chain was a degenerate EOD.

For each sample, we also count the degenerate hyperedges, and the nodes that participate in any degenerate hyperedge. Results are in Table 1. Samples where $\mathring{G}$ was a real dataset have up to 10% degenerate hyperedges, and up to 99% of nodes participating in degenerate edges. The synthetic datasets demonstrate that high density networks exhibit higher degeneracy.

All these results indicate that the two sample spaces, thus the null models, are very different, which was the first goal of our experimental evaluation (more evidence is given below). In particular, non-degenerate EODs are extremely sparse in the sample space that includes degenerate ones. Thus, one cannot use NuDHy as a rejection sampling subroutine to produce non-degenerate samples, which was

[9] Implementation available from https://github.com/acdmammoths/dinghy-code

[10] Details in the supplementary materials.

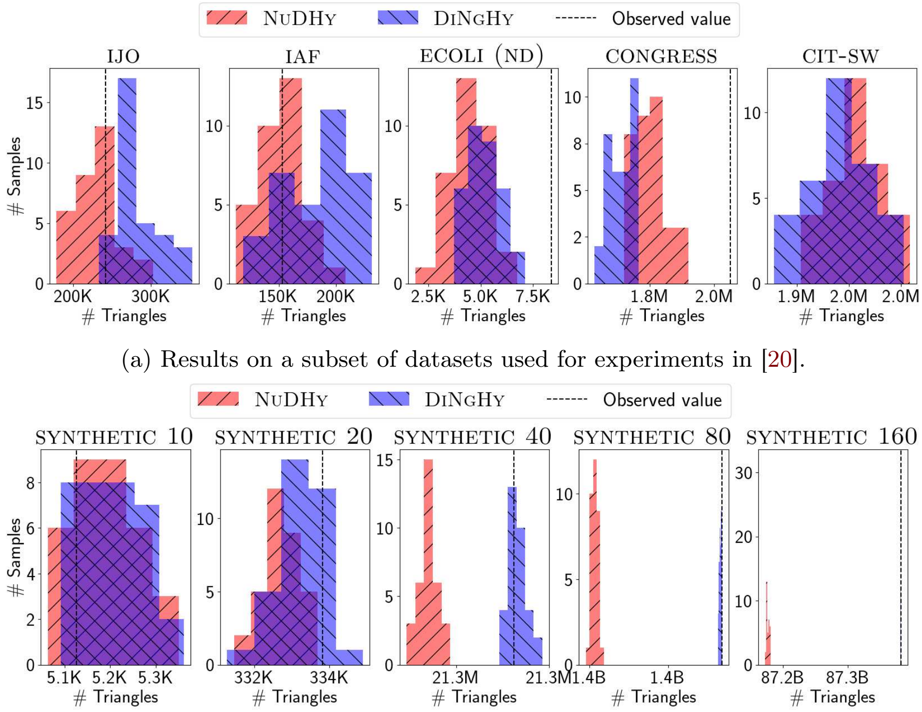
Table 1: Median amount of degeneracy in 33 samples per observed network, obtained by NuDHy.

| | Degenerate edges | | Nodes in degenerate edges | |
|---|---|---|---|---|
| Dataset | Count | Normalized | Count | Normalized |
| congress | 178 | 9.55% | 100 | 99.01% |
| iaf1260b | 129 | 6.19% | 247 | 14.81% |
| ijo1366 | 152 | 6.75% | 294 | 16.29% |
| ecoli (nd) | 29 | 3.17% | 60 | 8.55% |
| cit-sw | 179 | 0.34% | 806 | 4.87% |
| dblp-9 (nd) | 272 | 0.29% | 1023 | 4.87% |
| enron | 379 | 0.25% | 1859 | 3.28% |
| math (nd) | 58 | 0.06% | 228 | 0.66% |
| ord (nd) | 378 | 0.08% | 722 | 0.11% |
| synthetic 160 | 1274 | 99.53% | 1280 | 100% |
| synthetic 80 | 927 | 72.42% | 1280 | 100% |
| synthetic 40 | 351 | 27.42% | 1280 | 100% |
| synthetic 20 | 96 | 7.5% | 1000 | 78.12% |
| synthetic 10 | 27 | 2.11% | 227 | 17.73% |

our second goal, and justifies the need to develop our new algorithm DiNgHy to sample directly from the space of non-degenerate EODs.

To further evaluate the importance of using the right null model when performing statistical hypothesis testing, we compare the distributions of the numbers of directed cycles of sizes 2 and 3 in the digraph projections of samples obtained with NuDHy and with DiNgHy.[11] Figure 2 shows the distributions of the number of directed 3-cycles for some of the datasets. Results for other datasets are in the supplementary materials, and are qualitatively similar. For all datasets, the distributions are clearly different. For iaf1260b and ijo1366, this difference leads to opposite outcomes of hypothesis tests. We compute the *empirical* p-value of a quantity $q(\mathring{G})$ as the ratio of sampled dihypergraphs $G$ where $|q(G) - \mu| \geq |q(\mathring{G}) - \mu|$, where $\mu$ is the mean value of $q$ over the samples. The number of directed 3-cycles in iaf1260b and ijo1366 is marked as non-significant under the degeneracy-allowed null model (NuDHy), with p-values of 0.97 and 0.73 respectively. Under our more appropriate null model, the p-values are 0 and 0.33 respectively, indicating that existing knowledge about the DGP *does not* explain the observed number of directed 3-cycles. The p-values for other real datasets is zero under both null models, since the observed value is outside both distributions, but the distributions are still distinct.

[11] For this experiment, we exclude enron, ord (nd), math (nd), and dblp-9 (nd) due to a prohibitive runtime of more than 22 hours per dataset. Corollary 1 does not hold on some of the dihypergraphs we consider, because it is not tight. We conjecture that the MCs for these cases are still irreducible. If not, the distribution would be uniform over the strongly connected component that includes the observed network.



(a) Results on a subset of datasets used for experiments in [20].

(b) Results on synthetic datasets, identified by their edge size.

Fig. 2: Empirical distributions of the number of 3-cycles in 33 samples. The observed value is omitted when far from both empirical distributions.

On synthetic datasets with higher density, and therefore degeneracy, the distributions are more distinct. On all but the least dense synthetic dataset, the observed number of 3-cycles is found to be significant under the degeneracy-allowed null model, and non-significant under our null model. Results for directed 2-cycles are similar (see the supplementary materials).

In all cases the distributions over the two null models are different, so there exist critical thresholds for which a hypothesis would be rejected under one null model but not under the other. This fact stresses the profound difference between our null models and that of Preti et al. [20], emphasizing the importance of choosing the appropriate null model when testing hypotheses.

### 6.2 Convergence

The perturbation score [22] is a popular measure for the empirical mixing time of MCMC algorithms that sample from null models over dyadic graphs and binary matrices. Given two binary matrices, it is defined as the fraction of entries with value 1 in one matrix that have value 0 in the other. We extend the perturbation score between two dihypergraphs $\mathring{G}$ and $G$ as the average of the perturbation score between the incidence matrices of $\mathsf{B}_t(\mathring{G})$ and $\mathsf{B}_t(G)$, and the perturbation score between the incidence matrices of $\mathsf{B}_h(\mathring{G})$ and $\mathsf{B}_h(G)$.

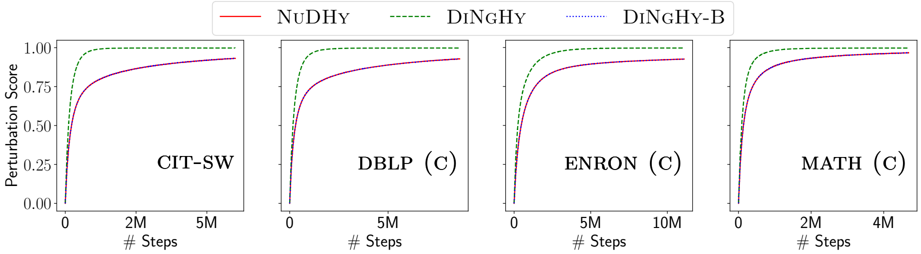


Fig. 3: Perturbation score as function of the steps on the Markov chain. In each case, the curves for NuDHy and DiNgHy-B perfectly overlap.

On the datasets satisfying Corol. 1, we use the same number of steps $s$ as Preti et al. [20] to take a single sample for each dataset (details in the supplementary materials), and we measure the perturbation score between $\mathring{G}$ and the current state of the Markov Chain every $\frac{s}{100}$ steps, for a total of 100 measurements.

Figure 3 shows the perturbation score as function of the number of steps. DiNgHy converges faster than DiNgHy-B and NuDHy, with the latter two showing identical behavior (overlapping curves). Thus, the different approach to choosing NDSs taken by DiNgHy is to be preferred, as it leads to faster mixing.

## 7 Conclusion

We introduce the first null models for edge-ordered and edge-unordered non-degenerate dihypergraphs, capturing important properties of the observed network. By preserving non-degeneracy, our models are more realistic than existing ones in many scenarios, and thus should be preferred for statistical hypothesis testing. Our MCMC algorithms sample from the null models according to any user-specified distribution, and converge quickly. Directions for future work include strengthening the sufficient conditions for irreducibility (Thm. 1 and Corol. 1), and developing more descriptive null models for dihypergraphs, and efficient algorithms to sample from them.

**Acknowledgments.** This work is supported in part by the National Science Foundation awards CAREER-2238693 and AF-2312241, and by the National Science Foundation Graduate Research Fellowship Program under Grant No. 2439559. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the Authors and do not necessarily reflect the views of the National Science Foundation.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

# Bibliography

[1] Abuissa, M., Lee, A., Riondato, M.: ROhAN: Row-order agnostic null models for statistically-sound knowledge discovery. DMKD **37**(4) (2023)

[2] Battiston, F. et al.: Networks beyond pairwise interactions: Structure and dynamics. Phys. Rep. **874**, 1–92 (2020)

[3] Bick, C. et al: What are higher-order networks? SIAM Rev. **65**(3) (2023)

[4] Billings, J.C.W. et al.: Simplex2vec embeddings for community detection in simplicial complexes. arXiv:1906.09068 (2019)

[5] Chodrow, P.S.: Configuration models of random hypergraphs. J. Compl. Netw. **8**(3) (2020)

[6] Choe, M. et al.: Representative and back-in-time sampling from real-world hypergraphs. ACM TKDD **18**(6) (2024)

[7] Cimini, G. et al.: The statistical physics of real-world networks. Nat. Rev. Phys. **1**(1) (2019)

[8] Do, M. et al.: Structural patterns and generative models of real-world hypergraphs. KDD'20

[9] Feng, S. et al.: Hypergraph models of biological networks to identify genes critical to pathogenic viral response. BMC Bioinf. **22**(1) (2021)

[10] Fosdick, B.K. et al.: Configuring random graph models with fixed degree sequences. SIAM Rev. **60**(2) (2018)

[11] Kannan, R. et al.: Simple Markov-chain algorithms for generating bipartite graphs and tournaments. Rand. Struct. Alg. **14**(4) (1999)

[12] Kim, S. et al.: Reciprocity in directed hypergraphs: Measures, findings, and generators. arXiv:2210.05328 (2023)

[13] Lamar, M.D.: On uniform sampling simple directed graph realizations of degree sequences. arXiv:912.3834 (2018)

[14] Lee, G. et al.: A survey on hypergraph mining: Patterns, tools, and generators. ACM Comp. Surv. (2025)

[15] Lehmann, E.L., Romano, J.P.: Testing Statistical Hypotheses. 4 edn. (2022)

[16] Luo, Q. et al.: Sampling hypergraphs via joint unbiased random walk. World Wide Web **27**(2) (2024)

[17] Mitzenmacher, M., Upfal, E.: Probability and Computing (2005)

[18] Miyashita, R. et al: Random hypergraph model preserving two-mode clustering coefficient. DaWaK'23

[19] Nakajima, K., Shudo, K., Masuda, N.: Randomizing hypergraphs preserving degree correlation and local clustering. IEEE TNSE **9**(3) (2021)

[20] Preti, G. et al.: Higher-order null models as a lens for social systems. Phys. Rev. X **14**(3), (2024)

[21] Riondato, M.: Statistically-sound knowledge discovery from data. SDM'23

[22] Strona, G. et al.: A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. Nat. Comm. **5**(1) (2014)

[23] Sun, H., Bianconi, G.: Higher-order percolation processes on multiplex hypergraphs. Phys. Rev. E **104**(3) (2021)

[24] Zeng, Y. et al.: Hyper-null models and their applications. Entropy **25** (2023)

# DiNgHy: Null Models for Non-Degenerate Directed Hypergraphs – Supplementary Materials

Maryam Abuissa[1][0009–0004–1454–5164] (✉), Matteo Riondato[2][0000–0003–2523–4420], and Eli Upfal[1][0000–0002–9321–9460]

[1] Brown University, Providence, RI 02912, USA
{maryam_abuissa,eliezer_upfal}@brown.edu
[2] Amherst College, Amherst, MA 01002, USA mriondato@amherst.edu

## 1 Derivation of sufficient conditions for irreducibility

The following set of lemmas lead to Corollary 1 from the main text. Lemma 1 and Lemma 3 imply Corollary 1 as stated, while Lemma 2 and Lemma 4 imply the unstated version of the Corollary 1 where the roles of $T$ and $H$ are reversed.

**Lemma 1.** *Let $n^*$ be the node with maximum degree in $\mathring{G}$, and $e^*$ be the hyperedge with maximum dimension in $\mathring{G}$. If*

$$(\mathsf{dim}_{\mathring{G}}(e^*) - 1)(2\mathsf{deg}_{\mathring{G}}(n^*) - 1) + 1 < \left\|\mathsf{ideg}(\mathring{G})\right\|_1,$$

*then, for every $(T, H) \in \mathcal{D}_\mathsf{B}$, and every edge $q \in H$, it holds $\mathsf{hse}_{T,H}(q) \geq 1$.*

*Proof.* Let $(T, H) \in \mathcal{D}_\mathsf{B}$, with $T = (\mathring{N}, E, W)$ and $H = (\mathring{N}, E, Z)$. For any edge $z = (n, e) \in Z$, define

$$N_{n,e} \doteq \left\{n' \in \mathring{N} \setminus \{n\} : (n', e) \in W \cup Z\right\}$$

as the set of nodes other than $n$ that connect to $e$ in either $T$ or $H$ (i.e., that belong to either the head or the tail of $e$ in the non-degenerate EOD $G \in \mathcal{D}$ such that $T = \mathsf{B}_\mathrm{t}(G)$, and $H = \mathsf{B}_\mathrm{h}(H)$). It holds $|N_{n,e}| = \mathsf{dim}_G(e) - 1$.

Let also, for any $z = (n, e) \in Z$,

$$E_{n,e} \doteq \{e' \in E \setminus \{e\} : (n, e') \in W \cup Z\}$$

be the set of hyperedges other $e$ that form an edge $n$ in either $T$ or $H$ (i.e., of which $n$ belongs to the head or tail in $G$). It holds $|E_{n,e}| = \mathsf{deg}_G(n) - 1$.

Denote with $\overline{\mathsf{hse}}_{T,H}(z) = Z \setminus \mathsf{hse}_{T,H}(z)$ the set of all edges in $H$ that do *not* form a HNDS with $z$. It holds

$$\begin{aligned}\overline{\mathsf{hse}}_{T,H}(z) =& \{z\} \cup \{(n', e') \in Z : n' \in N_{n,e} \vee e' \in E_{n,e}\} \\ =& \{z\} \cup \left(\bigcup_{e' \in E_{n,e}} \left\{(n'', e') \in Z : n'' \in \mathring{N}\right\}\right) \\ & \cup \left(\bigcup_{n' \in N_{n,e}} \{(n', e'') \in Z : e'' \in E\}\right) \quad .\end{aligned}$$

Thus,

$$\begin{aligned}\left|\overline{\mathsf{hse}}_{T,H}(z)\right| &\le 1 + \sum_{e'\in E_{n,e}} \mathsf{tdim}_G(e') + \sum_{n'\in N_{n,e}} \mathsf{ideg}_G(n') \\ &\le 1 + |E_{n,e}|(\mathsf{dim}_{\mathring{G}}(e^*) - 1) + |N_{n,e}|\mathsf{deg}_{\mathring{G}}(n^*) \\ &= 1 + (\mathsf{deg}_{\mathring{G}}(n) - 1)(\mathsf{dim}_{\mathring{G}}(e^*) - 1) + (\mathsf{dim}_{\mathring{G}}(e) - 1)\mathsf{deg}_{\mathring{G}}(n^*) \\ &\le 1 + (\mathsf{dim}_{\mathring{G}}(e^*) - 1)(2\mathsf{deg}_{\mathring{G}}(n^*) - 1) \\ &< \sum_{n'\in\mathring{N}} \mathsf{ideg}_{\mathring{G}}(n') = |Z|,\end{aligned}$$

where the last inequality follows from the hypothesis. $\overline{\mathsf{hse}}_{T,H}(z)$ and $\mathsf{hse}_{T,H}(z)$ partition $Z$, so it must be $|\mathsf{hse}_{T,H}(z)| \ge 1$. □

The proof for the following lemma proceeds similarly to the previous proof.

**Lemma 2.** *Let $n^*$ and $e^*$ be as in Lemma 1. If*

$$(\mathsf{dim}_{\mathring{G}}(e^*) - 1)(2\mathsf{deg}_{\mathring{G}}(n^*) - 1) + 1 < \left\|\mathsf{odeg}(\mathring{G})\right\|_1,$$

*then, for every $(T,H) \in \mathcal{D}_{\mathsf{B}}$, and every edge $q \in H$, it holds $\mathsf{tse}_{T,H}(q) \ge 1$.*

**Lemma 3.** *Let $v^*$ and $e^*$ be as in Lemma 2. If*

$$2(\mathsf{dim}_{\mathring{G}}(e^*) - 1)(2\mathsf{deg}_{\mathring{G}}(n^*) - 1) + 1 < \left\|\mathsf{odeg}(\mathring{G})\right\|_1,$$

*then, for every $(T,H)$ and $(T,H') \in \mathcal{D}_{\mathsf{B}}$, and every edge $z$ in $T$, it holds $|\mathsf{ttse}_{T,H,H'}(e)| \ge 1$.*

*Proof.* Let $T = (\mathring{N}, E, W)$, $H = (\mathring{N}, E, Z)$, $z = (n,e) \in Z$, $N_{n,e}$, and $E_{n,e}$ be as in the proof for Lemma 1. Let $H' = (\mathring{N}, E, Y)$, and define $N'_{n,e}$ and $E'_{n,e}$ similarly as $N_{n,e}$ and $E_{n,e}$ but on $(T,H')$.

Denote with $\overline{\mathsf{ttse}}_{T,H,H'}(z) = W \setminus \mathsf{ttse}_{T,H,H'}(z)$ the set of all edges in $W$ that do *not* form a TNDS with $z$ in both $(T,H)$ and $(T',H')$. It holds

$$\begin{aligned}\overline{\mathsf{ttse}}_{T,H,H'}(z) =& \{z\} \cup \left\{(n',e') \in W : n' \in N_{n,e} \cup N'_{n,e} \vee e' \in E_{n,e} \cup E'_{n,e}\right\} \\ =& \{z\} \cup \left(\bigcup_{e'\in E_{n,e}\cup E'_{n,e}} \left\{(n'',e') \in W : n'' \in \mathring{N}\right\}\right) \\ & \cup \left(\bigcup_{n'\in N_{n,e}\cup N'_{n,e}} \{(n',e'') \in W : e'' \in E\}\right) .\end{aligned}$$

Thus,

$$\begin{aligned}
|\overline{\mathsf{ttse}}_{T,H,H'}(z)| \leq & 1 + \sum_{e' \in E_{n,e} \cup E'_{n,e}} \left|\left\{(n'', e') \in W : n'' \in \mathring{N}\right\}\right| \\
& + \sum_{n' \in N_{n,e} \cup N'_{n,e}} |\{(n', e'') \in W : e'' \in E\}| \\
\leq & 1 + \sum_{e' \in E_{n,e} \cup E'_{n,e}} \mathsf{tdim}_{\mathring{G}}(e') + \sum_{n' \in N_{n,e} \cup N'_{n,e}} \mathsf{odeg}_{\mathring{G}}(n') \\
\leq & 1 + (|E_{n,e}| + \left|E'_{n,e}\right|)(\mathsf{dim}_{\mathring{G}}(e^*) - 1) + (|N_{n,e}| + \left|N'_{n,e}\right|)(\mathsf{deg}_{\mathring{G}}(n^*)) \\
\leq & 1 + 2(\mathsf{deg}_{\mathring{G}}(n^*) - 1)(\mathsf{dim}_{\mathring{G}}(e^*) - 1) + 2(\mathsf{dim}_{\mathring{G}}(e^*) - 1)\mathsf{deg}_{\mathring{G}}(n^*) \\
< & \sum_{n' \in \mathring{N}} \mathsf{odeg}_{\mathring{G}}(n') = |W|,
\end{aligned}$$

where the last inequality follows from the hypothesis. Because $\overline{\mathsf{ttse}}_{T,H,H'}(z)$ and $\mathsf{ttse}_{T,H,H'}(z)$ partition $W$, the last inequality implies $|\mathsf{ttse}_{T,H,H'}(z)| \geq 1$. □

The proof for the following lemma proceeds similarly to the previous proof.

**Lemma 4.** *Let $v^*$ and $e^*$ be as in Lemma 2. If*

$$2(\mathsf{dim}_{\mathring{G}}(e^*) - 1)(2\mathsf{deg}_{\mathring{G}}(n^*) - 1) + 1 < \left\|\mathsf{ideg}(\mathring{G})\right\|_1,$$

*then, for every $(T', H)$ and $(T, H) \in \mathcal{D}_{\mathsf{B}}$, and every edge $z$ in $H$, it holds $|\mathsf{hhse}_{H,T,T'}(e)| \geq 1$.*

## 2 Proof of stationary uniform distribution

This result is Theorem 2 in the main text.

**Theorem 1.** *The unique stationary distribution of the MC is uniform over $\mathcal{D}_{\mathsf{B}}$.*

*Proof.* We start by showing that the transition probabilities are symmetric, i.e., the probability $p_{S,S'}$ of moving from $S = (T, H)$ to $S' = (T', H')$ in one step is the same as the probability $p_{S',S}$ of moving from $S'$ to $S$ in one step.

Clearly the transition probabilities are zero if $S$ and $S'$ are not neighbors. Assume now that there is a HNDS $q = \mathsf{hs}_{T,H}((\bar{n}_1, \bar{e}_1), (\bar{n}_2, \bar{e}_2))$ on $(T, H)$ leading to $(T', H')$. The proof for the case when there is a TNDS between two neighbors follows the same steps. Since a HNDS only changes the second bipartite graph, it must be $T' = T$.

The probability that the MC moves from $(T, H)$ to $(T, H')$ is the probability that $q$ is performed by our algorithm (Algorithm 1 in the main text). This probability is

$$pS, S' = \binom{|E|}{2}^{-1} b \frac{1}{|\mathsf{h}(\bar{e}_1) \setminus \mathsf{n}(\bar{e}_2)|} \frac{1}{|\mathsf{h}(\bar{e}_2) \setminus \mathsf{n}(\bar{e}_1)|},$$

as it is the product of:

1. the probability of drawing $\bar{e}_1$ and $\bar{e}_2$ as $e_1$ and $e_2$ on line 1 of Algorithm 1;
2. the probability that the coin flip is heads (line 2);
3. the probability of drawing $\bar{n}_1$ as $n_1$ (line 5);
4. the probability of drawing $\bar{n}_2$ as $n_2$ (line 6);

When $q$ is performed on $(T, H)$ to obtain $(T', H')$, the hyperedge $\bar{e}_1$ is modified to become $\bar{e}'_1 = (\mathsf{t}(\bar{e}_1), (\mathsf{h}(\bar{e}_1) \setminus \{\bar{n}_1\}) \cup \{\bar{n}_2\})$,[3] and the hyperedge $\bar{e}_2$ is modified to become $\bar{e}'_2 = (\mathsf{t}(\bar{e}_2), (\mathsf{h}(\bar{e}_2) \setminus \{\bar{n}_2\}) \cup \{\bar{n}_1\})$. The probability that the MC moves from $(T, H')$ to $(T, H)$ is the probability that the HNDS $\mathsf{rev}(q)$, the reversal of $q$, is performed by Algorithm 1. This probability is

$$p_{S',S} = \binom{|E|}{2}^{-1} b \frac{1}{|\mathsf{h}(\bar{e}'_1) \setminus \mathsf{n}(\bar{e}'_2)|} \frac{1}{|\mathsf{h}(\bar{e}'_2) \setminus \mathsf{n}(\bar{e}'_1)|} \quad .$$

We want to show that $p_{S,S'} = p_{S',S}$. From the definition of $\bar{e}'_1$ and $\bar{e}'_2$ above, and the fact that $\mathsf{n}(\bar{e}'_2) = \mathsf{h}(\bar{e}'_2) \cup \mathsf{t}(\bar{e}'_2)$ and these two sets are disjoint or we would have degeneracy, we get

$$\begin{aligned} \mathsf{h}(\bar{e}'_1) \setminus \bar{e}'_2 &= (\mathsf{h}(\bar{e}'_1) \setminus \mathsf{h}(\bar{e}'_2)) \setminus \mathsf{t}(\bar{e}'_2) \\ &= (((\mathsf{h}(\bar{e}_1) \setminus \{\bar{n}_1\}) \cup \{\bar{n}_2\}) \setminus ((\mathsf{h}(\bar{e}_2) \setminus \{\bar{n}_2\}) \cup \{\bar{n}_1\})) \setminus \mathsf{t}(\bar{e}_2) \\ &= (((\mathsf{h}(\bar{e}_1) \setminus \mathsf{h}(\bar{e}_2)) \setminus \{\bar{n}_1\}) \cup \{\bar{n}_2\}) \setminus \mathsf{t}(\bar{e}_2) \quad . \end{aligned}$$

Since $\bar{n_1} \in \mathsf{h}(\bar{e}_1) \setminus \mathsf{h}(\bar{e}_2)$, while $\bar{n_2} \notin \mathsf{h}(\bar{e}_1) \setminus \mathsf{h}(\bar{e}_2)$, then

$$|((\mathsf{h}(\bar{e}_1) \setminus \mathsf{h}(\bar{e}_2)) \setminus \{\bar{n}_1\}) \cup \{\bar{n}_2\}| = |\mathsf{h}(\bar{e}_1) \setminus \mathsf{h}(\bar{e}_2)|,$$

which, with the fact that $\mathsf{t}(\bar{e}_2) = \mathsf{t}(\bar{e}'_2)$, implies $|\mathsf{h}(\bar{e}'_1) \setminus \mathsf{n}(\bar{e}'_2)| = |\mathsf{h}(\bar{e}_1) \setminus \mathsf{n}(\bar{e}_2)|$, i.e., the third factors in the expressions of $p$ and $p'$ are equal. By the same reasoning, we can show that $|\mathsf{h}(\bar{e}'_2) \setminus \mathsf{n}(\bar{e}'_1)| = |\mathsf{h}(\bar{e}_2) \setminus \mathsf{n}(\bar{e}_1)|$, concluding that $p_{S,S'} = p_{S',S}$.

The transition matrix of our Markov chain, which is defined over a finite number of states, is therefore symmetric, hence is doubly-stochastic. Markov chains whose transition matrix is doubly-stochastic have a unique stationary distribution, the uniform [5, Ex. 7.11]. □

## 3 Edge-unordered dihypergraphs

This result is Lemma 1 in the main text.

**Lemma 5.** *Let* $G = (N, E)$ *be a EUD. Let* $t^* \doteq \max_{e \in E} \mathsf{tdim}_G(e)$ *and* $h^* \doteq \max_{e \in E} \mathsf{hdim}_G(e)$. *For* $1 \leq i \leq t^*$ *and* $1 \leq j \leq h^*$, *let*

$$E_{i,j} \doteq \{\!\{ e \in E : \mathsf{tdim}_G(e) = i \wedge \mathsf{hdim}_G(e) = j \}\!\}$$

[3] The hyperedges $\bar{e}_1$ and $\bar{e}'_1$ have the same identifier, but the nodes belonging to this hyperedge changed as a consequence of applying the NDS $q$.

*be the multiset of hyperedges with tail dimension $i$ and head dimension $j$. Let $\bar{E}_{i,j} \doteq \left\{e_{i,j,1}, \ldots, e_{i,j,u_{i,j}}\right\}$ be the* set *of such hyperedges, labeled arbitrarily, and $\bar{E}$ be the set version of $E$. For $1 \leq k \leq u_{i,j}$, let $w_{i,j,k} \doteq \mathsf{m}_E(e_{i,j,k})$ be the multiplicity of $e_{i,j,k}$ in $E$. Then, the number $\left|\mathsf{o2u}^{-1}(G)\right|$ of edge-labeled hypergraphs mapped to $G$ by $\mathsf{o2u}(\cdot)$ is*

$$\left|\mathsf{o2u}^{-1}(G)\right| = \prod_{i=1}^{t^*}\prod_{j=1}^{h^*}\binom{|E_{i,j}|}{w_{i,j,1}, \ldots, w_{i,j,u_{i,j}}} = \frac{\prod_{i=1}^{t^*}\prod_{j=1}^{h^*}|E_{i,j}|!}{\prod_{e\in\bar{E}}\mathsf{m}_G(e)!} \enspace . \tag{1}$$

*Proof.* The argument follows the same structure as the proof for [1, Lemma 3].

Let $G^*$ be any EOD in $\mathsf{o2u}^{-1}(G)$. Any other $G' \in \mathsf{o2u}^{-1}(G)$ can be obtained by appropriately permuting the hyperedge identifiers of $G^*$. Of all the possible permutations, some only differs from each other by the permutation of identifiers of identical hyperedges, resulting in the same $G'$. Thus the total number of permutations (the numerator in eq. (1)) must be divided by the fraction of such identical permutations (the denominator in eq. (1)) to obtain the number of distinct ones, i.e., the number of distinct EODs in $\mathsf{o2u}^{-1}(G)$. □

## 4 Datasets

The real datasets, whose salient statistics are in Table 1, represent a variety of applications. cit-sw [4] and dblp-9 [8] are citation hypergraphs where each hyperedge correspond to a pair of papers where the first cites the second: the tail is the set of authors of the first paper, and the head is the set of authors of the second. iaf1260b and ijo1366 [4] represent chemical reactions among genes as hyperedges, where each gene is a node. enron is a network of emails with the sender in the tail and the recipients in the head. math [4] represents a question-and-answer forum from MathOverflow where each hyperedge is a post, the tail contains the question original poster, and any responders are in the head. ecoli (nd) [7] is constructed using the pathway *eco01100* of Escherichia coli from the Kyoto Encyclopedia of Genes and Genomes (KEGG). congress is a dihypergraph representation of sponsor-cosponsor relationship on bills in the Senate from the 107th U.S. Congress [2]. ord [3] models chemical reactions, with reagents in the hyperedge tail, and products in the head.

The dihypergraphs ecoli, ord, dblp-9, and math were originally degenerate. We remove degeneracy from them as a preprocessing step, by iterating through the hyperedges, fixing the tail and removing any duplicate nodes from the head. If the resulting head is empty, we remove the hyperedge. The resulting non-degenerate datasets have real-world structure, and represent modified settings which are still useful for many applications. For example, dblp-9, once degeneracy is removed, is a citation dataset where self-citations are ignored, and math without degeneracy represents a question-and-answer forum where self-responses are ignored. There are many applications where self-interactions are not meaningful or impossible, thus where it would be appropriate to use

DiNgHy. We add (nd) after a dataset name to denote that degeneracy has been removed.

For the convergence experiment in Sect. 6.3, we further process the datasets dblp-9 (nd), math (nd), enron to remove a few of the largest hyperedges so the resulting datasets fulfill the condition for irreducibility from Corollary 1. We refer to the datasets so obtained as dblp-9 (c), math (c), and enron (c) (see also Table 1). For dblp-9 (c) and math (c), we remove the 1% of hyperedges with highest dimension, while for enron (c) we remove 7% of edges. The resulting datasets maintain sufficient real structural properties for our evaluation of convergence to be meaningful.

Table 1: Datasets statistics. Irreducibility condition from Corollary 1. See text for details.

| Dataset | $\lvert E\rvert$ | $\lvert V\rvert$ | $\overline{\mathsf{tdim}}$ | $\overline{\mathsf{hdim}}$ | $\overline{\mathsf{odeg}}$ | $\overline{\mathsf{ideg}}$ | Irreducibility | $s$ |
|---|---|---|---|---|---|---|---|---|
| enron | 148754 | 56700 | 1.0 | 4.0 | 2.6 | 10.4 | N | 15m |
| enron (c) | 138845 | 48050 | 1.0 | 3.0 | 2.9 | 8.7 | Y | 15m |
| ijo1366 | 2251 | 1805 | 2.3 | 2.0 | 2.8 | 2.5 | N | 194k |
| ecoli (nd) | 914 | 702 | 2.0 | 2.1 | 2.6 | 2.8 | N | 79k |
| cit-sw | 53177 | 16555 | 2.7 | 2.9 | 8.7 | 9.4 | Y | 6.0m |
| ord (nd) | 478084 | 632245 | 4.5 | 1.0 | 3.4 | 0.8 | N | 53m |
| dblp-9 (nd) | 92526 | 20986 | 2.4 | 2.4 | 10.7 | 10.6 | N | 9.3m |
| dblp-9 (c) | 91636 | 20614 | 2.3 | 2.4 | 10.4 | 10.5 | Y | 9.3m |
| congress | 1864 | 101 | 1.0 | 8.4 | 18.5 | 154.7 | N | 178k |
| math (nd) | 90689 | 34578 | 1.0 | 1.8 | 2.6 | 4.6 | N | 5.2m |
| math (c) | 89791 | 33271 | 1.0 | 1.6 | 2.7 | 4.3 | Y | 5.2m |
| iaf1260b | 2083 | 1668 | 2.2 | 2.0 | 2.8 | 2.5 | N | 178k |

In addition to the real and modified datasets, we use five synthetic dihypergraphs to perform the difference in the outcomes of hypothesis tests when using our null models vs. the one by Preti et al. [6]. These dihypergraphs each have a 1280 hyperedges and 1280 nodes, and the hyperedge size is varied to produce different densities. We refer to each of these datasets as synthetic $n$, where $n$ is a hyperedge size in $\{10, 20, 40, 80, 160\}$. In synthetic n, all the hyperedges have tail dimension and head dimension $\frac{n}{2}$, and the nodes have in-degree and out-degree $\frac{n}{2}$, so a larger $n$ means higher density.

## 5 Additional Experiments

We include the results from all experiments on all relevant datasets in this section. See Section 6 in the main text for a discussion of the experimental results.

We show the incidence of degeneracy measured over the course of drawing a single sample for all datasets in Fig. 1.

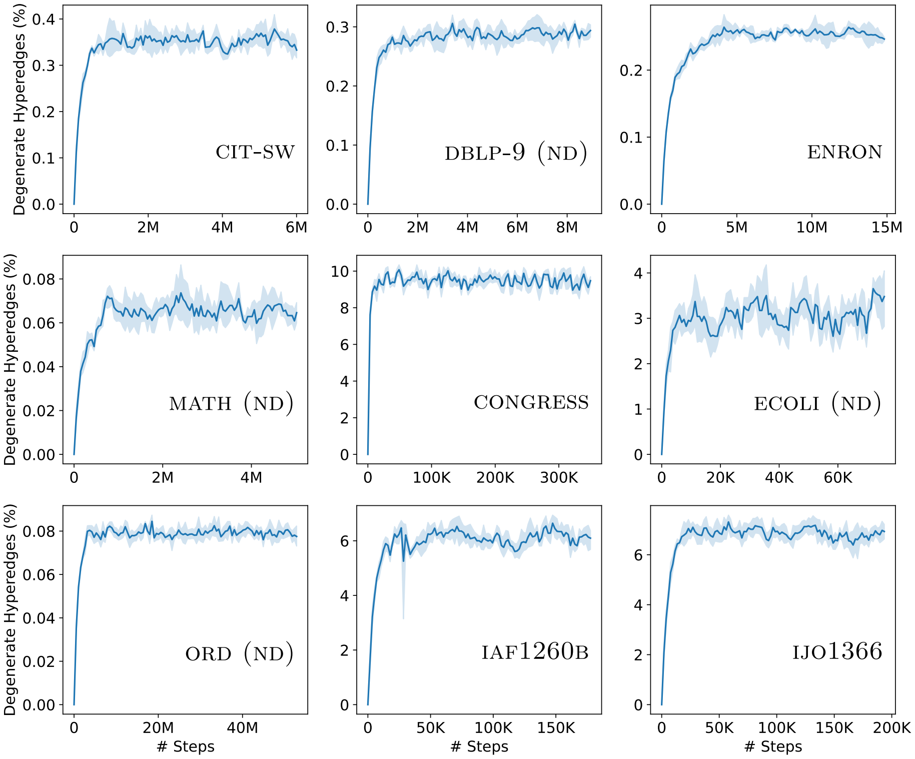
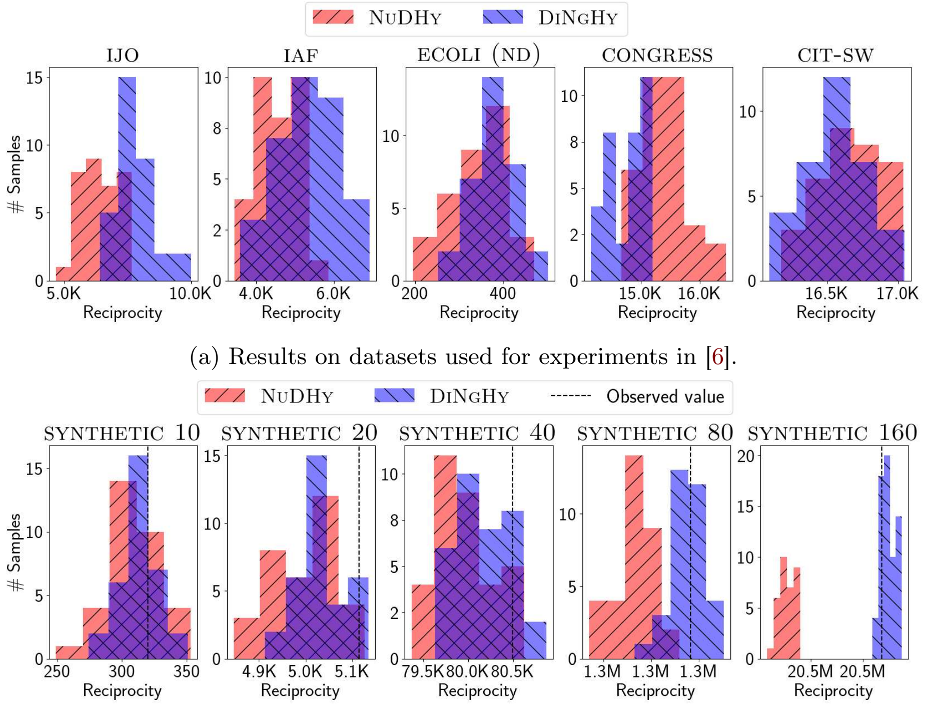


Fig. 1: Shows the percentage of degenerate edges on the y-axis and the number of steps in the Markov Chain on the x-axis. The line is the mean over the samples and the shaded region represents a 95% confidence interval.

We performed the same experiment as the one for directed triangles but to measure reciprocity, or directed 2-cycles. The results are similar, as shown in Fig. 2.

We include exact p-values for the synthetic data in Table 2. All nonzero p-values for real datasets are discussed in the main text.



(a) Results on datasets used for experiments in [6].

(b) Results on synthetic datasets, identified by their edge size, which corresponds to density.

Fig. 2: Histogram that shows the frequency of the number of directed 2-cycles, or reciprocity, in 33 samples from NuDHy and DiNgHy for each dataset. The value of the observed dataset is shown by a black dashed vertical line or omitted when it is extremely far from both sample distributions.

Table 2: P-values for the number of triangles and reciprocity of nondegenerate synthetic datasets compared to 33 samples from NuDHy or DiNgHy.

| | Triangles p-values | | Reciprocity p-values | |
|---|---|---|---|---|
| Dataset | DiNgHy | NuDHy | DiNgHy | NuDHy |
| synthetic 160 | 0.45 | 0 | 0.61 | 0 |
| synthetic 80 | 0.06 | 0 | 0.79 | 0 |
| synthetic 40 | 0.76 | 0 | 0.21 | 0.09 |
| synthetic 20 | 0.30 | 0.03 | 0.15 | 0.15 |
| synthetic 10 | 0.30 | 0.39 | 0.61 | 0.55 |

# Bibliography

[1] Abuissa, M., Lee, A., Riondato, M.: ROhAN: Row-order agnostic null models for statistically-sound knowledge discovery. Data Mining and Knowledge Discovery **37**(4), 1692–1718 (2023)

[2] Fowler, J.H.: Legislative cosponsorship networks in the us house and senate. Social Networks **28**(4), 454–465 (2006), ISSN 0378-8733, https://doi.org/https://doi.org/10.1016/j.socnet.2005.11.003, URL https://www.sciencedirect.com/science/article/pii/S0378873305000730

[3] Kearnes, S.M., Maser, M.R., Wleklinski, M., Kast, A., Doyle, A.G., Dreher, S.D., Hawkins, J.M., Jensen, K.F., Coley, C.W.: The open reaction database. Journal of the American Chemical Society **143**(45), 18820–18826 (2021), https://doi.org/10.1021/jacs.1c09820, URL https://doi.org/10.1021/jacs.1c09820, pMID: 34727496

[4] Kim, S., Choe, M., Yoo, J., Shin, K.: Reciprocity in directed hypergraphs: Measures, findings, and generators (2023), URL https://arxiv.org/abs/2210.05328

[5] Mitzenmacher, M., Upfal, E.: Probability and Computing: Randomized Algorithms and Probabilistic Analysis. Cambridge University Press (2005)

[6] Preti, G., Fazzone, A., Petri, G., De Francisci Morales, G.: Higher-order null models as a lens for social systems. Physical Review X **14**(3), 031032 (2024)

[7] Shen, T., Zhang, Z., Chen, Z., Gu, D., Liang, S., Xu, Y., Li, R., Wei, Y., Liu, Z., Yi, Y., Xie, X.: A genome-scale metabolic network alignment method within a hypergraph-based framework using a rotational tensor-vector product. Scientific Reports **8** (2018), URL https://api.semanticscholar.org/CorpusID:53227644

[8] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 990–998, KDD '08, Association for Computing Machinery, New York, NY, USA (2008), ISBN 9781605581934, https://doi.org/10.1145/1401890.1402008, URL https://doi.org/10.1145/1401890.1402008