Counting simplicial pairs in hypergraphs

Jordan Barrett*, Paweł Prałat†, Aaron Smith‡, François Théberge§

May 22, 2025

Abstract

We present two ways to measure the simplicial nature of a hypergraph: the simplicial ratio and the simplicial matrix. We show that the simplicial ratio captures the frequency, as well as the rarity, of simplicial interactions in a hypergraph while the simplicial matrix provides more fine-grained details. We then compute the simplicial ratio, as well as the simplicial matrix, for 10 real-world hypergraphs and, from the data collected, hypothesize that simplicial interactions are more and more *deliberate* as hyperedge size increases. We then present a new Chung-Lu model that includes a parameter controlling (in expectation) the frequency of simplicial interactions. We use this new model, as well as the real-world hypergraphs, to show that multiple stochastic processes exhibit different behaviour when performed on simplicial hypergraphs vs. non-simplicial hypergraphs.

1 Introduction

Many datasets that are typically represented as graphs would be more accurately represented as hypergraphs. For example, in the graph representation of a collaboration dataset, authors are represented as vertices and an edge exists between two vertices if the corresponding authors wrote a paper together [32]. Using this representation, it is impossible to distinguish between a three-author paper and three separate two-author papers. In contrast, when we represent a collaboration dataset as a hypergraph we can clearly distinguish between a three-author paper (a single hyperedge) and three separate two-author papers (three distinct hyperedges). Hypergraph representations have proven to be useful for studying collaboration datasets [17], protein complexes and metabolic reactions [11], and many other datasets that

*Department of Mathematics, Toronto Metropolitan University, Toronto, ON, Canada; e-mail: jordan.barrett@torontomu.ca

†Department of Mathematics, Toronto Metropolitan University, Toronto, ON, Canada; e-mail: pralat@torontomu.ca

‡Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada; e-mail: asmi28@uOttawa.ca

§Tutte Institute for Mathematics and Computing, Ottawa, ON, Canada; email: theberge@ieee.org

are traditionally represented as graphs [29]. Moreover, after many years of intense research using graph theory in modelling and mining complex networks [10, 15, 21, 31], hypergraph theory has started to gain considerable traction [3, 4, 5, 6, 23, 19, 22]. It is becoming clear to both researchers and practitioners that higher-order representations are needed to study datasets involving higher-order interactions [5, 25, 35, 29].

Similar to hypergraph representations, simplicial complexes provide another way to represent datasets with higher-order interactions and, in some cases, it is not clear what the better model is for a given dataset [24, 36, 38]. The notion of *simpliciality* was first introduced by Landry, Young and Eikmeier in [27] as a way of describing how closely a hypergraph resembles its simplicial closure. In their work, they discover that many hypergraphs built from real-world data, although not actually simplicial complexes, resemble their simplicial closures more closely than random hypergraphs. In a similar but distinct study, LaRock and Lambiotte in [28] find that real-world hypergraphs often contain more instances of hyperedges contained in other hyperedges than in random hypergraphs. In [16], Joslyn et al. propose a measure similar to simpliciality called *inclusiveness* and provide more evidence that real networks are more "simplicial" than random hypergraphs. The results found in these three papers suggest that real-world hypergraphs are organized in a way where many of the small hyperedges live inside larger hyperedges. In our work, we pursue this idea further and define a ratio and a matrix for hypergraphs, which we call the *simplicial ratio* and *simplicial matrix* respectively, based on the number of instances of hyperedges inside other hyperedges compared to that of a null model.

The remainder of the paper is organized as follows. In Sections 1.1 and 1.2 we discuss notation as well as the measures for simpliciality given in [27]. Next, we define the simplicial ratio in Section 2.1, the simplicial matrix in Section 2.2, and temporal variants in Section 2.3. Then, in Section 3.1 we compute the simplicial ratio and simplicial matrix of the same 10 real-world hypergraphs that were studied in [27] and then analyse this data in Section 3.2. In Section 4 we present a new random graph model that allows for more or less instances of hyperedges inside other hyperedges depending on an input parameter $q \in [0, 1]$. In Section 5 we experiment with four stochastic processes, comparing the processes on real-world hypergraphs and on our proposed model for varying q. We conclude and suggest further research in Section 6. Finally, let us mention that this paper is a (long) journal version of the (short) proceeding paper published in [2].

1.1 Notation

A hypergraph G is a pair $(V(G), E(G))$ where $V(G)$ is a set of vertices and $E(G)$ is a collection of hyperedges, i.e., a collection of subsets of vertices. We insist that $\emptyset \notin E(G)$ for any hypergraph G. For $e \in E(G)$, write $|e|$ for the size of e and, for each positive integer k, define

$$E_k(G) := \{e \in E(G), |e| = k\}.$$

In general, for a hypergraph G and hyperedge $e \in E(G)$, it is acceptable that $|e| = 1$. In this paper, however, we forbid such hyperedges and consider only hyperedges of size at least 2.

We write $[n] := \{1, \ldots, n\}$ and typically label the vertices in G as $[n]$. A subhypergraph of a hypergraph G is any hypergraph $H = (V(H), E(H))$ with $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$ (note that, as H is itself a hypergraph, any hyperedge $e \in E(H)$ contains only vertices in $V(H)$). If $E_k(G) = E(G)$ for some $k \geq 2$, then we call G a k-uniform hypergraph. Note that, for any hypergraph G, the hypergraph $G_k := (V(G), E_k(G))$ is a k-uniform subhypergraph of G, and

$$G = \bigcup_{k \geq 2} G_k \,,$$

and thus every hypergraph is the hyperedge-disjoint union of uniform subhypergraphs.

A *multihypergraph* G is a hypergraph that allows hyperedges $e \in E(G)$ with more than one instance of the same vertex (multiset hyperedges) and allows multiple hyperedges $e_1, \ldots, e_k \in E(G)$ that are identical (parallel hyperedges); a hypergraph G is *simple* if it contains no multiset hyperedges or parallel hyperedges. Note that all simple hypergraphs are multihypergraphs. For a multihypergraph G and a vertex v, writing $m_G(v, e)$ for the number of instances of v in e, the *degree of v in G*, denoted $\deg_G(v)$, is defined as

$$\deg_G(v) := \sum_{e \in E(G)} m_G(v, e) \,.$$

If G is simple, we equivalently have

$$\deg_G(v) = \left| \left\{ e \in E(G) \;\middle|\; v \in e \right\} \right| .$$

All hypergraphs in this paper are simple except for the random hypergraphs generated by Algorithm 2 and Algorithm 4.

We use standard notation for probability, i.e., $\mathbb{P}(\cdot)$ for probability, $\mathbb{E}[\cdot]$ for expectation. We write $X \sim \mathcal{U}$ to mean X is sampled from distribution $\mathcal{U}$ and write $X_1, \ldots, X_k \overset{i.i.d.}{\sim} \mathcal{U}$ to mean $X_1, \ldots, X_k$ are sampled independently and identically from distribution $\mathcal{U}$. For a set S, we write $X \in_u S$ to mean that X is chosen uniformly at random from S.

1.2 Measures for simpliciality

In [27], Landry, Young and Eikmeier establish three distinct measures quantifying how close a hypergraph is to a simplicial complex. The first measure they establish is the *simplicial fraction*. Given a hypergraph G, let $S \subseteq E(G)$ be the set of hyperedges such that $e \in S$ if and only if $|e| \geq 3$ and, for all $f \subseteq e$ with $|f| \geq 2$, $f \in E(G)$. Then the *simplicial fraction* of G, written $\sigma_{\mathrm{SF}}(G)$, is defined as

$$\sigma_{\mathrm{SF}}(G) := \frac{|S|}{\left| \bigcup_{k \geq 3} E_k(G) \right|} \,.$$

In words, $\sigma_{\mathrm{SF}}(G)$ is the proportion of hyperedges of size at least 3 in $E(G)$ that satisfy downward closure.

The second and third measures that Landry, Young and Eikmeier establish are the *edit simpliciality* and the *face edit simpliciality*, respectively. For a hypergraph G, define the *k-closure*, written $\overline{G}_k$, as the hypergraph $(V(\overline{G}_k), E(\overline{G}_k))$ where

$$V(\overline{G}_k) = V(G),$$
$$E(\overline{G}_k) = \Big\{ e \subseteq V(G) \,\Big|\, |e| \geq k \text{ and } e \subseteq f \text{ for some } f \in E(G) \Big\}.$$

Then the *edit simpliciality* of G, written $\sigma_{\mathrm{ES}}(G)$, is defined as

$$\sigma_{\mathrm{ES}}(G) := \frac{|E(G)|}{|E(\overline{G}_2)|}\,.$$

Thus, $1 - \sigma_{\mathrm{ES}}(G)$ is the (normalized) number of additional hyperedges needed to turn G into its 2-closure. Similarly, the *face edit simpliciality* of G, written $\sigma_{\mathrm{FES}}(G)$, is the average edit simpliciality across all induced subhypergraphs defined by maximal hyperedges (hyperedges not contained in other hyperedges) in $\bigcup_{k\geq 3} E_k(G)$.

Using the three measures defined above, Landry, Young and Eikmeier show that real-world hypergraphs are significantly more simplicial than hypergraphs sampled from certain random models. However, they also note some unique short-comings of each measure. In the following examples, we show additional ways in which all three measures seem to disagree with common intuitions about which graphs are more simplicial. The first example shows that none of the measures properly capture the *types* of simplicial relationships in a hypergraph.

Example 1.1. Fix n, k with $5 \leq k$ and $3k \leq n$. Let G_1 be a hypergraph on the vertex set $[n]$ and with three hyperedges $\{1, \ldots, k\}, \{k+1, \ldots, 2k\}, \{2k+1, \ldots, 3k\}$ of size k and three hyperedges $\{1, 2, 3\}, \{k+1, k+2, k+3\}, \{2k+1, 2k+2, 2k+3\}$ of size 3. Let G_2 be a hypergraph on the same vertex set and with the same three hyperedges $\{1, \ldots, k\}, \{k+1, \ldots, 2k\}, \{2k+1, \ldots, 3k\}$ of size k, but now with three hyperedges $\{1, \ldots, k-1\}, \{k+1, \ldots, 2k-1\}, \{2k+1, \ldots, 3k-1\}$ of size $k-1$. See Figure 1 for an illustration of G_1 and G_2 with $n = 18$ and $k = 6$.

With G_1 and G_2 as defined above, we have

$$\sigma_{\mathrm{SF}}(G_1) = \sigma_{\mathrm{SF}}(G_2) = 0\,,$$
$$\sigma_{\mathrm{ES}}(G_1) = \sigma_{\mathrm{ES}}(G_2) = \frac{2 \cdot 3}{(2^k - k - 1) \cdot 3} = \frac{2}{2^k - k - 1}\,, \text{ and}$$
$$\sigma_{\mathrm{FES}}(G_1) = \sigma_{\mathrm{FES}}(G_2) = \frac{2}{2^k - k - 1}\,,$$

the value $2^k - k - 1$ coming from the fact that there are 2^k subsets, k of which are subsets of size 1, and 1 of which is the empty set. Thus, by all three measures, G_1 and G_2 are equally simplicial. However, qualitatively the simplicial relationships in G_1 are different than in G_2. Consider, for example, hyperedges e_3, e_5, e_6 in an Erdős-Rényi random hypergraph on n vertices with $|e_3| = 3, |e_5| = 5$ and $|e_6| = 6$. Then, the probability of $e_3 \subset e_6$ (as in G_1) is of order n^{-3}, whereas the probability of $e_5 \subset e_6$ (as in G_2) is of order n^{-5}.

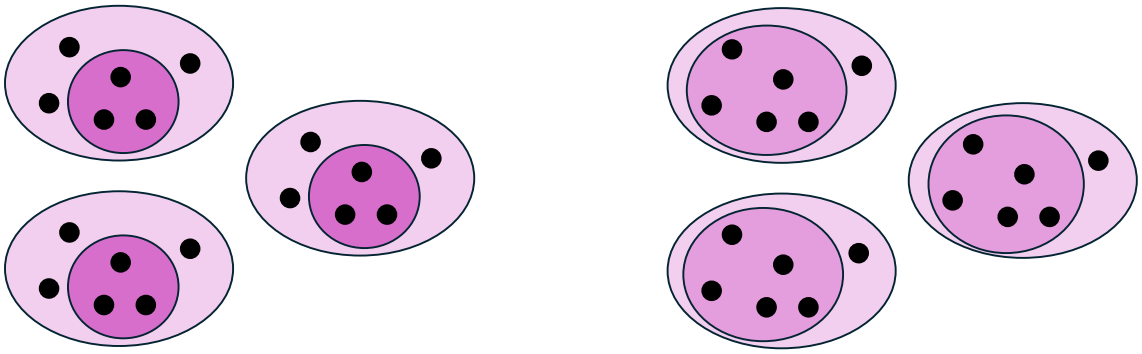


Figure 1: (left) a hypergraph G_1 with 18 vertices, 3 hyperedges of size 6, and 3 hyperedges of size 3, and (right) a hypergraph G_2 with 18 vertices, 3 hyperedges of size 6, and 3 hyperedges of size 5. We have $\sigma_{\mathrm{SF}}(G_1) = \sigma_{\mathrm{SF}}(G_2) = 0$, $\sigma_{\mathrm{ES}}(G_1) = \sigma_{\mathrm{ES}}(G_2) = 2/57$, and $\sigma_{\mathrm{FES}}(G_1) = \sigma_{\mathrm{FES}}(G_2) = 2/57$.

The second example shows that, while the three measures are good indicators of how close a hypergraph is to its 2-closure, they do not always distinguish hypergraphs with many hyperedge-in-hyperedge interactions from hypergraphs with no such interactions. The present work argues that measuring how "surprising" a hypergraph looks (with respect to some baseline given by a null model) often accords more closely with our intuition, at least for the far-from-fully-simplicial hypergraphs seen in most datasets.

Example 1.2. Let G_1 and G_2 be as shown in Figure 2. There is a clear, strong simplicial structure in G_1, and there is clearly no simplicial structure in G_2. However, in both hypergraphs, the simplicial fraction is 0 (none of the hyperedges satisfy downward closure). Moreover, the edit simpliciality of G_1 is $4/57 \approx 0.07$ and of G_2 is $3/41 \approx 0.07$. Likewise, the face edit simpliciality of G_1 is $4/57 \approx 0.07$ and of G_2 is

$$\frac{1}{3}\left(\frac{1}{26} + \frac{1}{11} + \frac{1}{4}\right) \approx 0.13\,.$$

Thus, G_1 and G_2 are equally simplicial according to the simplicial fraction and the edit simpliciality and, more strikingly, G_1 is *less* simplicial than G_2 according to the face edit simpliciality.

As mentioned previously, Examples 1.1 and 1.2 are not issues when we treat the simplicial fraction, edit simpliciality, and face edit simpliciality as measures of how close a hypergraph is to its 2-closure (as was their intended purpose). Instead, these examples suggest that if we want to understand the extent to which hyperedges sit inside other hyperedges in real-world networks then we need a new type of scoring system.

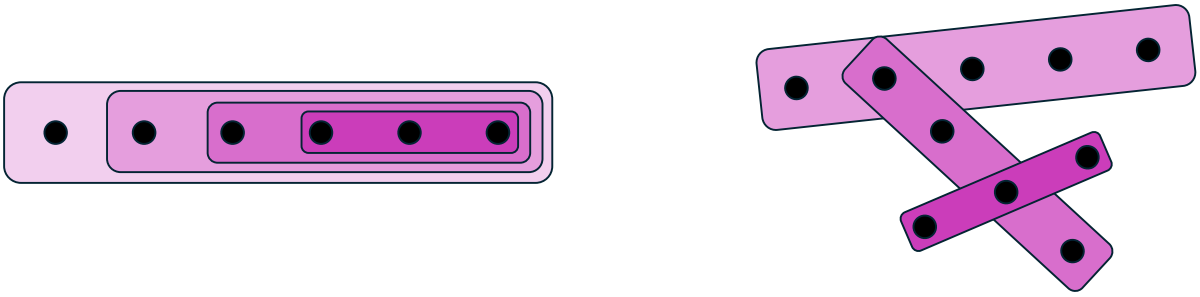


Figure 2: (left) a hypergraph G_1 with 6 vertices and 4 hyperedges, and (right) a hypergraph G_2 with 10 vertices and 3 hyperedges. We have $\sigma_{\text{SF}}(G_1) = 0$, $\sigma_{\text{ES}}(G_1) \approx 0.07$, $\sigma_{\text{FES}}(G_1) \approx 0.07$, and $\sigma_{\text{SF}}(G_2) = 0$, $\sigma_{\text{ES}}(G_2) \approx 0.07$, $\sigma_{\text{FES}}(G_2) \approx 0.13$.

2 A new approach to simpliciality

We aim to quantify a hypergraph based on the frequency and rarity of hyperedges inside other hyperedges when compared to a null model. The metrics we present focus on the regime where data is "slightly" more simplicial than random (and so nearly-complete large simplices are extremely rare), while previous metrics focus on the regime where data is "almost completely" simplicial. The motivation behind these metrics is that the former regime is often more appropriate in real networks.

The hypergraph Chung-Lu model

In the material to come, we frequently reference the hypergraph Chung-Lu model. The original model was defined for graphs [9] and has been extensively studied since then. More recently, the model was generalized to other structures, including geometric graphs [20, 18] (both undirected and directed variants) as well as hypergraphs [19]. We give an algorithm for building the hypergraph model, conditioned on the number of hyperedges, and point the reader to [19] for a full description of the model.

Let $(d_1, \ldots, d_n)$ be a degree sequence on vertex set $[n]$ and let $(m_{k_{\min}}, \ldots, m_{k_{\max}})$ be a sequence of hyperedge sizes where m_k represents the number of hyperedges of size k. Then, writing $p(\cdot)$ for the probability distribution with $p(v) = d_v / \sum_{u \in [n]} d_u$ for all $v \in [n]$, we first give the algorithm that generates a Chung-Lu hyperedge of a given size.

Algorithm 1 Chung-Lu hyperedge.

Require: $(d_1, \ldots, d_n)$, k

1: Sample $e[1], \ldots, e[k] \overset{i.i.d.}{\sim} p(\cdot)$.
2: Return $\{e[1], \ldots, e[k]\}$

We now give the algorithm that generates a Chung-Lu hypergraph.

Algorithm 2 Chung-Lu Model.

Require: $(d_1, \ldots, d_n)$, $(m_{k_{\min}}, \ldots, m_{k_{\max}})$.

Initialize hyperedge list $E = \{\}$.
for $k \in \{k_{\min}, \ldots, k_{\max}\}$ **do**
  for $i \in [m_k]$ **do**
    sample $e \sim$ **Algorithm 1**$\Big((d_1, \ldots, d_n), k\Big)$.
    Set $E = E \cup \{e\}$.
  end for
end for
Return $G = ([n], E)$.

For a hypergraph G with degree sequence $\mathbf{d} = (d_1, \ldots, d_n)$ and hyperedge size sequence $\mathbf{m} = (m_{k_{\min}}, \ldots, m_{k_{\max}})$, we write $\hat{G} \sim \mathrm{CL}(G)$ to mean $\hat{G} \sim \mathrm{CL}(\mathbf{d}, \mathbf{m})$, where $\mathrm{CL}(\mathbf{d}, \mathbf{m})$ is the random hypergraph returned by **Algorithm 2**. A key feature of the Chung-Lu model is that the degree sequence is preserved in expectation.

Lemma 2.1. *Let $\hat{G} \sim \mathrm{CL}(G)$ for some hypergraph G. Then*

$$\mathbb{E}\left[\deg_{\hat{G}}(v)\right] = \deg_G(v)$$

for all $v \in [n]$.

Proof. Let $d_v := \deg_G(v)$ for all $v \in [n]$. First, note that every vertex in every hyperedge of $\hat{G}$ is sampled independently with probability p, where $p(v) = \frac{d_v}{\sum_{u \in [n]} d_u}$. Thus, the expected total occurrence of v in $E(\hat{G})$ is

$$p(v) \sum_{e \in E(G)} |e| = \left(\frac{d_v}{\sum_{u \in [n]} d_u}\right) \sum_{e \in E(G)} |e| = \left(\frac{d_v}{\sum_{u \in [n]} d_u}\right) \sum_{u \in [n]} d_u = d_v \,,$$

the second equality coming from the hypergraph counterpart of the hand-shaking lemma. Given that the total occurrence of v in $E(\hat{G})$ is precisely $\deg_{\hat{G}}(v)$, the lemma follows. □

2.1 The simplicial ratio

For a hypergraph G, a *simplicial pair in G* is a pair of distinct hyperedges $e_1, e_2 \in E(G)$ with $e_1 \subset e_2$. Let $\mathrm{sp}\,(G)$ be the number of simplicial pairs in G.

Let G be a hypergraph and let $\hat{G} \sim \mathrm{CL}(G)$ conditioned on $\hat{G}$ having no multiset hyperedges. Then the *simplicial ratio*, denoted by $\sigma_{\mathrm{SR}}\,(G)$, is defined as

$$\sigma_{\mathrm{SR}}\,(G) := \frac{\mathrm{sp}\,(G)}{\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right]} \,,$$

if $\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right] > 0$, and $\sigma_{\mathrm{SR}}\,(G) := 1$ otherwise. In words, $\sigma_{\mathrm{SR}}\,(G)$ is the ratio of the number of simplicial pairs to the expected number of simplicial pairs.

Remark 2.2. If $\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right] = 0$ then it is necessarily the case that $\mathrm{sp}\left(G\right) = 0$, since it is always true that $\mathbb{P}\left(\hat{G} = G\right) > 0$. Moreover, if $\mathrm{sp}\left(G\right) = 0$ and $\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right] = 0$ then the number of simplicial pairs is as expected and so we define $\sigma_{\mathrm{SR}}\left(G\right) = 1$.

Remark 2.3. We have mentioned already that the sizes of the hyperedges in a simplicial pair are important. For this reason, we condition on $\hat{G} \sim \mathrm{CL}(G)$ having no multiset hyperedges.

Remark 2.4. One could define an analogous notion of "simplicial surprise" by choosing any other reference distribution for the random graph $\hat{G}$ in simplicial ratio: the configuration model, Erdős-Rényi model, Stochastic Block Model, ABCD model, etc. We choose to use the Chung-Lu model as, in our opinion, it achieves the best balance of (a) retaining important features of a hypergraph and (b) allowing for fast approximations of $\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right]$.

Remark 2.5. As mentioned in the previous remark, we *approximate* $\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right]$ rather than compute this expectation exactly. For a hypergraph G, computing $\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right]$ is quite difficult as we discuss in the open problems presented in Section 6.1. We approximate using a Monte Carlo estimator which is detailed in Appendix B.

Examples

Let us revisit Examples 1.1 and 1.2, and introduce a new family of examples for which our measures strongly disagree.

Starting with Example 1.1, the number of simplicial pairs in both hypergraphs is 3. However, in G_1 the expected number of simplicial pairs is ≈ 0.3, and in G_2 this expectation is ≈ 0.008. Thus, $\sigma_{\mathrm{SR}}\left(G_1\right) \approx 10$, whereas $\sigma_{\mathrm{SR}}\left(G_2\right) \approx 380$, suggesting that the number of simplicial relationships in G_2 is far more surprising than in G_1. This result confirms that the simplicial ratio weighs different types of simplicial pairs differently.

Continuing with Example 1.2, we have that $\mathrm{sp}\left(G_1\right) = 6$ and $\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right] \approx 4.3$, meaning $\sigma_{\mathrm{SR}}\left(G_1\right) \approx 1.4$, whereas $\mathrm{sp}\left(G_2\right) = 0$ and $\mathbb{E}\left[\mathrm{sp}\left(\hat{G}_2\right)\right] \approx 0.2 > 0$, meaning $\sigma_{\mathrm{SR}}\left(G_2\right) = 0$. Thus, the simplicial ratio can clearly distinguish G_1 and G_2.

Let us continue with another example highlighting an extreme difference between our measure and the measures of Landry et al.

Example 2.6. Fix a graph size $n \in \mathbb{N}$ and parameters $2 \leq k < n-1$ and $0 < p < 1$. Let $G = (V, E)$ be a hypergraph that contains *all* hyperedges of size $2 \leq t \leq k$, *no* hyperedges of size $t > k$, and $\left\lceil p\binom{n}{k}\right\rceil$ hyperedges of size $t = k$. If we use the Chung-Lu model *conditional on the resulting graph being simple* as the null model for simplicial ratio, these graphs have simplicial ratio of exactly 0.[1] This is in stark contrast to the measures of Landry et al, for which $\sigma_{\mathrm{SF}}(G) = \sigma_{\mathrm{ES}}(G) = \sigma_{\mathrm{FES}}(G) = 1$. We note that the full simplex graph is a special case of this example: Landry et al call this highly simplicial (for obvious reasons), while we say

[1]For our Chung-Lu model, the simplicial ratio is typically small but nonzero.

it is not *surprisingly* simplicial (because there is no other way to build a simple hypergraph with the same number of edges of all orders).

By computing the simplicial ratio of the hypergraphs in Examples 1.1, 1.2 and 2.6, we see a clear distinction between the three measures given in [27] and the simplicial ratio that we present here: the simplicial fraction, edit simpliciality, and face edit simpliciality are all ways of measuring how close a hypergraph is to its induced simplicial complex, whereas the simplicial ratio is a way to measure how *surprisingly simplicial* a hypergraph is. Computing these simplicial ratios also provides examples of the different extremes of $\sigma_{\mathrm{SR}}(G)$: if G contains hyperedges of different sizes but no simplicial pairs then $\sigma_{\mathrm{SR}}(G) = 0$, whereas if G is comprised of simplicial pairs with 2 large hyperedges than $\sigma_{\mathrm{SR}}(G) \gg 1$.

2.2 The simplicial matrix

For a hypergraph G, write $\mathrm{sp}(G, i, j)$ for the number of simplicial pairs (e_1, e_2) in G with $|e_1| = i$ and $|e_2| = j$ with $i < j$. Then, letting $\hat{G} \sim \mathrm{CL}(G)$ conditioned on having no multiset hyperedges, the simplicial matrix of G, denoted by $M_{\mathrm{SR}}(G)$, is the partial matrix with cell (i, j) equaling

$$M_{\mathrm{SR}}(G, i, j) := \frac{\mathrm{sp}(G, i, j)}{\mathbb{E}\left[\mathrm{sp}\left(\hat{G}, i, j\right)\right]}$$

whenever $i < j$ and G contains hyperedges of size i and of size j (and substituting 0 if there are no simplicial pairs of this type), and with cell (i, j) being empty otherwise.

Remark 2.7. We once again approximate $\mathbb{E}\left[\mathrm{sp}\left(\hat{G}, i, j\right)\right]$ via the sampling technique found in Appendix B.

Intuitively, the simplicial matrix "unpacks" the simplicial ratio and shows how powerful the simplicial interactions between hyperedges of all different sizes are. More formally, the simplicial matrix and simplicial ratio of G satisfy the following weighted sum.

$$\sigma_{\mathrm{SR}}(G) = \sum_{i<j} w_{i,j} \cdot M_{\mathrm{SR}}(G, i, j)$$

where

$$w_{i,j} := \frac{\mathbb{E}\left[\mathrm{sp}\left(\hat{G}, i, j\right)\right]}{\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right]}, \qquad \sum_{i<j} w_{i,j} = 1.$$

We will see in Section 3 that the simplicial matrix reveals information about real-world hypergraphs that the simplicial ratio alone does not. In particular, a hypothesis we make in this paper, as suggest by these matrices, is that *the composition of a hyperedge in a real-world network becomes more dependent on simpliciality as the hyperedge size increases.*

Examples

We again revisit Examples 1.1 and 1.2. In Example 1.1, $M_{\mathrm{SR}}(G_1)$ contains one non-empty cell, $(3,6)$, with value ≈ 10, and $M_{\mathrm{SR}}(G_2)$ contains one non-empty cell, $(5,6)$, with value ≈ 380.

Example 1.2 is more interesting as G_1 contains simplicial pairs of various types. For G_1, we have

$$M_{\mathrm{SR}}(G_1) \approx \begin{bmatrix} \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset \\ \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset \\ \emptyset & \emptyset & \emptyset & \mathbf{3.8} & \mathbf{1.7} & \mathbf{1} \\ \emptyset & \emptyset & \emptyset & \emptyset & \mathbf{2.4} & \mathbf{1} \\ \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \mathbf{1} \\ \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset \end{bmatrix},$$

and for G_2 we have

$$M_{\mathrm{SR}}(G_2) \approx \begin{bmatrix} \emptyset & \emptyset & \emptyset & \emptyset & \emptyset \\ \emptyset & \emptyset & \emptyset & \emptyset & \emptyset \\ \emptyset & \emptyset & \emptyset & \mathbf{0} & \mathbf{0} \\ \emptyset & \emptyset & \emptyset & \emptyset & \mathbf{0} \\ \emptyset & \emptyset & \emptyset & \emptyset & \emptyset \end{bmatrix}.$$

The simplicial matrix for G_1 unpacks the information about its simplicial interactions. Indeed, the simplicial ratio simply tells us that the number of simplicial pairs is 1.4 times more than expected. On the other hand, the simplicial matrix tells us that all 3 simplicial pairs involving the hyperedge of size 6 are to be expected, whereas the other three simplicial pairs are at least somewhat surprising. We can also see that the existence of the $(3,4)$ pair in G_1 is more surprising than the existence of the $(3,5)$ pair, which is in turn more surprising than the existence of the $(3,6)$ pair. In general, given a hypergraph G and distinct hyperedge sizes $i < j < k$, if G has the property that $|E_j(G)| \leq |E_k(G)|$ then it follows from the sampling process in **Algorithm 1** that $\mathbb{E}[\mathrm{sp}(G,i,j)] \leq \mathbb{E}[\mathrm{sp}(G,i,k)]$. In the case of Example 1.2, we have that $|E_4(G_1)| = |E_5(G_1)| = |E_6(G_1)| = 1$ and $\mathbb{E}[\mathrm{sp}(G_1,3,4)] \approx 0.26$, $\mathbb{E}[\mathrm{sp}(G_1,3,5)] \approx 0.59$, and $\mathbb{E}[\mathrm{sp}(G_1,3,6)] = 1$.

2.3 Including a temporal element

Many networks are not merely static hypergraphs, but rather evolving processes with hyperedges forming over time. In these evolving processes, we can distinguish between two ways that a simplicial pair can form: either a small hyperedge could form first, followed by a larger (superset) hyperedge, or a large hyperedge could form first, followed by a smaller (subset) hyperedge. In the context of a collaboration hypergraph, a "bottom-up" formation is a group of collaborators who invite more people for a future collaboration, whereas a "top-down" formation is a group who exclude some people for a future collaboration. At least in this context, there is a substantial difference between bottom-up simplicial pairs and top-down simplicial pairs, and it is natural to ask how different networks bias towards or against the two types of simplicial formations. For this reason, we mention several versions of

the simplicial ratio and of the simplicial matrix that accounts for time-stamped hyperedges. In the definitions to come, we assume that no two hyperedges are born at the exact same time, and continue to assume that there are no repeated hyperedges.

To define the temporal simplicial measure that seems closest in spirit to those in the rest of the paper, let G be a hypergraph with an *ordered* set of hyperedges $E(G) = (e_1, \ldots, e_m)$ (typically, this ordering would come from time stamps associated with each hyperedge). Next, let $\mathrm{sp}^{\nearrow}(G)$ be the number of simplicial pairs (e_i, e_j) in G with $i < j$ and $|e_i| < |e_j|$, and let $\mathrm{sp}^{\searrow}(G)$ be the number of simplicial pairs (e_i, e_j) with $i > j$ and $|e_i| < |e_j|$. Finally, sample edges $\hat{G} \sim \mathrm{CL}(G)$ and assign a uniformly random ordering to the hyperedges of $\hat{G}$. Then the bottom-up simplicial ratio and top-down simplicial ratio of G, denoted $\sigma^{\nearrow}_{\mathrm{SR}}(G)$ and $\sigma^{\searrow}_{\mathrm{SR}}(G)$ respectively, are defined as

$$\sigma^{\nearrow}_{\mathrm{SR}}(G) := \frac{\mathrm{sp}^{\nearrow}(G)}{\mathbb{E}\left[\mathrm{sp}^{\nearrow}\left(\hat{G}\right)\right]} \qquad \text{and} \qquad \sigma^{\searrow}_{\mathrm{SR}}(G) := \frac{\mathrm{sp}^{\searrow}(G)}{\mathbb{E}\left[\mathrm{sp}^{\searrow}\left(\hat{G}\right)\right]}.$$

For the temporal version of the simplicial matrix we distinguish between bottom-up and top-down simplicial pairs by their location in the matrix. For a temporal hypergraph G with hyperedge ordering $E(G) = (e_1, \ldots, e_m)$ and for $k < \ell$, write $\mathrm{sp}^{\nearrow}(G, k, \ell)$ for the number of simplicial pairs (e_i, e_j) such that $i < j$, $|e_i| = k$, and $|e_j| = \ell$. Likewise, write $\mathrm{sp}^{\searrow}(G, k, \ell)$ for the number of simplicial pairs (e_i, e_j) such that $i > j$, $|e_i| = k$ and $|e_j| = \ell$. Then the temporal simplicial matrix, denoted $M^{\rightarrow}_{\mathrm{SR}}(G)$, is the partial matrix with cell (k, ℓ) equaling

$$M^{\rightarrow}_{\mathrm{SR}}(G, k, \ell) := \frac{\mathrm{sp}^{\nearrow}(G, k, \ell)}{\mathbb{E}\left[\mathrm{sp}^{\nearrow}\left(\hat{G}, k, \ell\right)\right]},$$

cell (ℓ, k) equaling

$$M^{\rightarrow}_{\mathrm{SR}}(G, \ell, k) := \frac{\mathrm{sp}^{\searrow}(G, k, \ell)}{\mathbb{E}\left[\mathrm{sp}^{\searrow}\left(\hat{G}, k, \ell\right)\right]},$$

for all valid $k < \ell$, and cells (k, ℓ) and (ℓ, k) being empty otherwise.

While this is a natural measure of simpliciality for temporal hypergraphs, we think there are many more natural analogues. We highlight two issues that appear in the temporal graph literature:

1. **Isolation of Temporal Effects:** are we measuring *only* temporal effects, or do they get mixed with non-temporal effects?

2. **Isolation of Time Windows:** are we trying to measure something that is uniform across time, or something that varies?

To deal with the first issue, we can replace the reference measure $\mathrm{CL}(G)$ appearing in our definitions by the the empirical measure $EM(G)$, where a temporal graph $\hat{G} \sim E(G)$ sampled from $EM(G)$ has exactly the same edges as G, but in a uniformly random order.

Informally, the idea is that $\hat{G} \sim EM(G)$ differs from G *only* through the time stamps - the actual hypergraph is the same - and so we can be more confident that we have isolated temporal effects.

Remark 2.8. When sampling $\hat{G} \sim EM(G)$, we have $\sigma_{\mathrm{SR}}^{\nearrow}(G) + \sigma_{\mathrm{SR}}^{\searrow}(G) = 1$. Thus, this choice of null model gives us a well-scaled measurement of the fraction of simpliciality that is due to "up"- and "down"-edges.

Remark 2.9. For both choices of reference measure considered in this section, by symmetry, we have that $\mathbb{E}\left[\mathrm{sp}^{\nearrow}\left(\hat{G}\right)\right] = \mathbb{E}\left[\mathrm{sp}^{\searrow}\left(\hat{G}\right)\right] = \frac{1}{2} \cdot \mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right]$. Thus, in this case, we can equivalently define the bottom-up simplicial ratio and top-down simplicial ratio respectively as

$$\frac{2 \cdot \mathrm{sp}^{\nearrow}(G)}{\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right]} \qquad \text{and} \qquad \frac{2 \cdot \mathrm{sp}^{\searrow}(G)}{\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right]}.$$

To deal with the second issue, many papers in network science assign *weights* to hyperedges according to the time stamps. Perhaps the most common weighting scheme is the temporal "sliding window": fix a time interval, and keep only the hyperedges with timestamps occurring within some fixed interval.

Mathematically, we can analyze these "sliding windows" using exactly the same formulas as in the rest of the paper - we are merely restricting them to a subset of the hyperedges. Practically, the "best" weighting scheme depends very heavily on the hypergraph and intended application, and so is beyond the scope of this paper. See *e.g.* [7, 8, 14] for empirical work on using hyperedges with timestamps effectively.

3 Empirical results

In this section, we show the simplicial ratio and simplicial matrix, both with and without a temporal element where applicable, for the same 10 hypergraphs that were analysed in [27]. We then comment on the data and build some hypotheses about the simplicial nature of real networks.

The 10 hypergraphs are all taken from [26] and full descriptions can be found there. We paraphrase and summarize the descriptions below.

contact-primary-school: a temporal hypergraph where nodes are primary students and hyperedges are instances of contact (physical proximity) between students.

contact-high-school: the same as contact-primary-school except with high-school students.

hospital-lyon: the same as contact-primary-school and contact-high-school except with patients and health-care workers in a hospital.

email-enron: a temporal hypergraph where nodes are email-addresses and hyperedges comprise the sender and receivers of emails.

email-eu: the same as email-enron except built from a different organization.

diseasome: a static (non-temporal) hypergraph where nodes are diseases and hyperedges are collections of diseases with a common gene.

disgenenet: a static hypergraph where nodes are genes and hyperedges are collections of genes found in a disease. In other words, disgenenet is precisely the line-hypergraph of diseasome.

ndc-substances: a static hypergraph where nodes are substances and hyperedges are collections of substances that make up various drugs.

congress-bills: a temporal hypergraph where nodes are US Congresspersons and hyperedges comprise the sponsor and co-sponsors of legislative bills put forth in both the House of Representatives and the Senate.

tags-ask-ubuntu: a temporal hypergraph where nodes are tags and hyperedges are collections of tags applied to questions on the website `askubuntu.com`.

For each hypergraph, we restrict to hyperedges of sizes 2 through 11, as is the case in [27]. We throw away multi-hyperedges, only keeping the first occurrence of each hyperedge in the case of temporal hypergraphs. We approximate $\mathbb{E}\left[\hat{G}\right]$ using the Monte Carlo method presented in Appendix B.

3.1 The data

We first show a table which includes the ratios, as well as useful information about each hypergraph.

G	$\lvert V(G)\rvert$	$\lvert E(G)\rvert$	$[\lvert E_2\rvert, \lvert E_3\rvert, \lvert E_4\rvert, \lvert E_{\geq 5}\rvert]$	$\sigma_{\mathrm{SR}}(G)$	$\sigma^{\nearrow}_{\mathrm{SR}}(G)$	$\sigma^{\searrow}_{\mathrm{SR}}(G)$
disgenenet	1982	760	[157, 139, 93, 371]	28.81	n.a.	n.a.
contact-h.s.	327	7818	[5498, 2091, 222, 7]	6.68	11.19	2.17
diseasome	516	314	[153, 92, 26, 43]	6.49	n.a.	n.a.
email-eu	967	23729	[13k, 5k, 2k, 4k]	5.19	5.77	3.72
email-enron	143	1442	[809, 317, 138, 178]	4.96	6.98	2.94
congress-bills	1715	58788	[14k, 10k, 8k, 27k]	4.46	5.23	3.69
ndc-substances	2740	4754	[1130, 745, 535, 2344]	4.22	n.a.	n.a.
contact-p.s.	242	12704	[7748, 4600, 347, 9]	2.74	4.82	0.66
hospital-lyon	75	1824	[1107, 657, 58, 2]	0.94	1.71	0.17
tags-ask-ubuntu	3021	145053	[28k, 52k, 39k, 25k]	0.69	1.09	0.29

Table 1: The simplicial ratio of 10 real networks and the corresponding bottom-up simplicial ratio and top-down simplicial ratio for the 7 temporal networks. The hypergraphs are ordered according to $\sigma_{\mathrm{SR}}(G)$, from largest to smallest.

In Figure 4 we show the simplicial matrices of these hypergraphs and in Figure 5 we show the temporal matrices of the 7 temporal hypergraphs. For readability we show only the non-empty cells of the partial matrices and omit cells involving hyperedges of size greater than 5. Figure 3 shows the simplified presentation of the simplicial matrix of G_1 from Example 1.2.

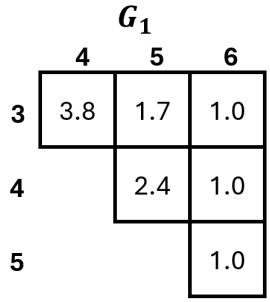


Figure 3: The simplicial matrix of G_1 from Example 1.2, presented in a simplified manner.

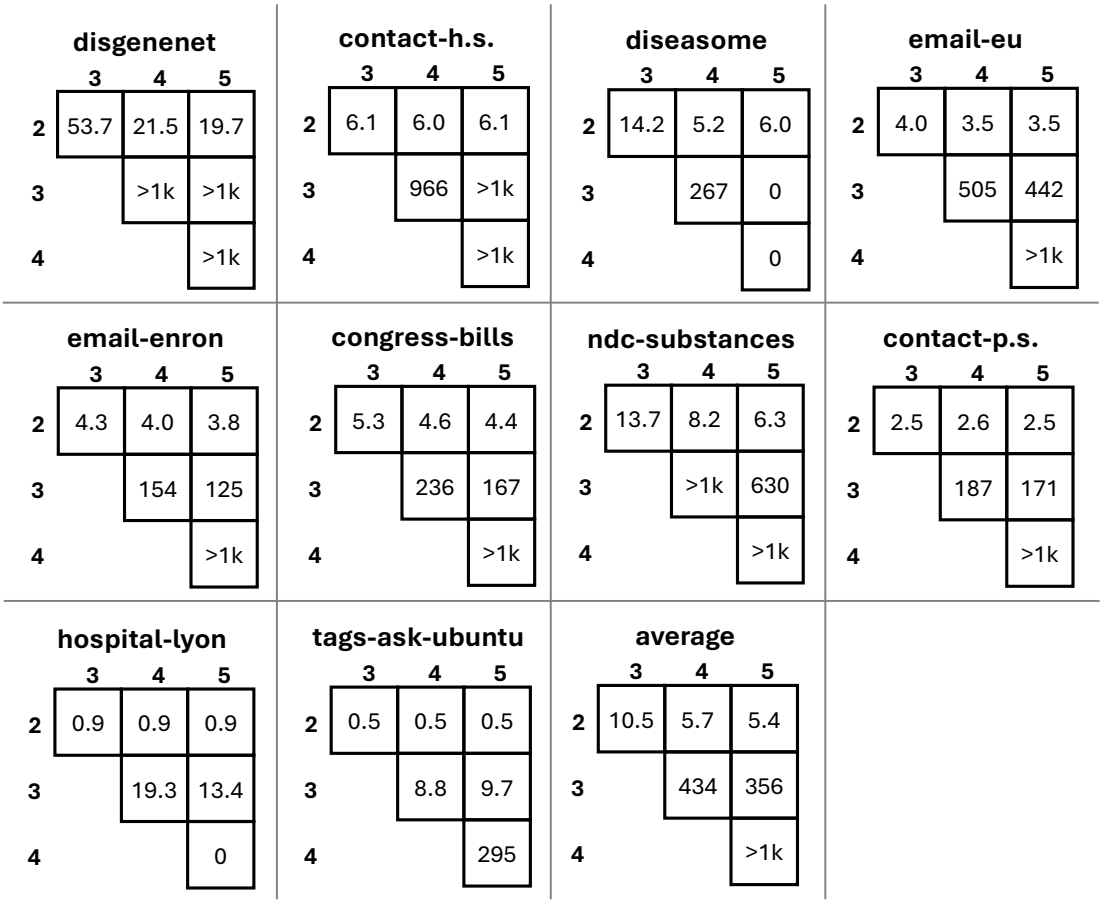


Figure 4: The simplicial matrix of 10 real networks, as well as the cell-wise average matrix. For each hypergraph G, only non-empty cells of $M_{\mathrm{SR}}(G)$ are shown, and cells involving hyperedges of size greater than 5 are omitted. The value of a cell is replaced with "$>$ 1k" whenever the value is above 1000.

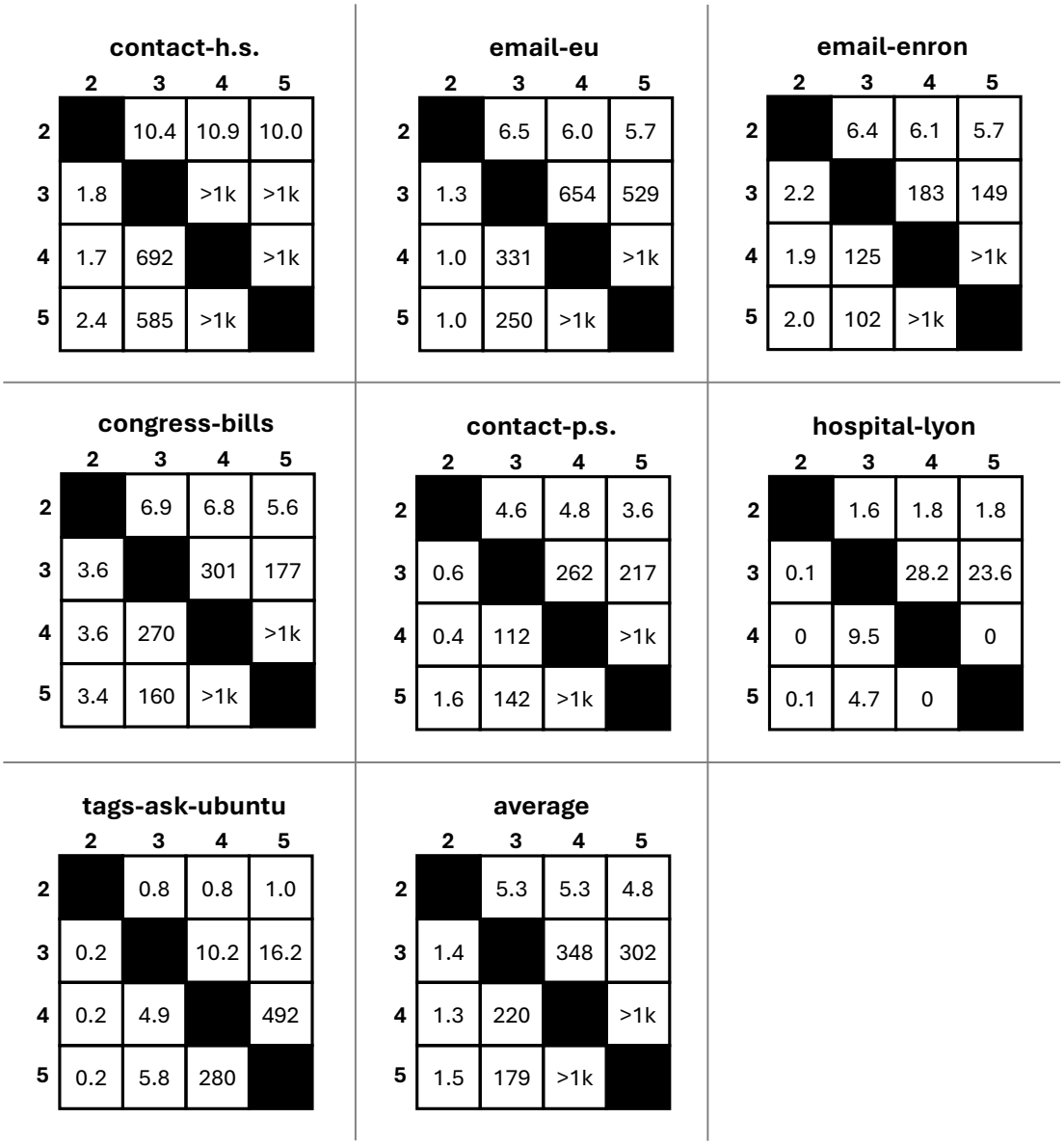


Figure 5: The temporal simplicial matrix of 7 real networks, as well as the cell-wise average matrix. For each hypergraph G, only non-empty cells of $M^{\rightarrow}_{\mathrm{SR}}(G)$ are shown, and cells involving hyperedges of size greater than 5 are omitted. The value of a cell is replaced with "$>$ 1k" whenever the value is above 1000.

3.2 Analysis

Simplicial ratio

Based on our results, we see that that biology networks are, on average, more surprisingly simplicial than contact-based networks and email networks. In contrast, it was shown in [27] that contact-based networks are the closest to their simplicial closures and biological networks

are furthest from theirs. In fact, comparing the ranks of the 3 existing measures (sf, es, fes) and the ranks from our simplicial ratio (sr), we get the following Kendall correlation values.

	sf	es	fes	sr
sf	1.000	0.706	0.989	-0.270
es	0.706	1.000	0.722	-0.256
fes	0.989	0.722	1.000	-0.289
sr	-0.270	-0.256	-0.289	1.000

These values show that our ranking system is negatively correlated with the ranking systems in [27]. A partial explanation for this correlation is that (a) the measures behave differently under different regimes of hyperedge density and (b) the 10 datasets cover a wide range of hyperedge density.

Bottom-up and top-down simplicial ratios

In our testing, we find that every temporal hypergraph contains more bottom-up simplicial pairs than top-down simplicial pairs. This suggests that, in general for many real networks, a small hyperedge leading to a larger (superset) hyperedge is more common than a large hyperedge leading to a smaller (subset) hyperedge. However, this result is *heavily* biased on our choice of keeping only the first instance of a hyperedge. To see this bias, let G be a temporal hypergraph with hyperedges $e_1, e_2 \in E(G)$ such that $e_1 \subset e_2$ and suppose that e_1 appears with multiplicity 5 and that e_2 appears with multiplicity 1. Then there are 6 possible birth orderings for e_2 and the 5 copies of e_1, and only 1 such ordering sees e_2 born before e_1. In most of the temporal networks analysed, the highest frequency of multi-hyperedges are indeed 2-hyperedges, and hence this bottom-up trend is at least partly explained by the above discussion. The topic of temporal simpliciality is one that we intend on exploring further in future works.

Simplicial matrix

Arguably the most immediate take-away from these matrices is that simplicial interactions become more surprising as hyperedge size increases. Although this feature is interesting, there is at least a partial explanation for this phenomenon that we explore in the following example.

Example 3.1. Let $n \in \mathbb{N}$, $\mathbf{d}$ be a uniform degree sequence, and let $\mathbf{m} = (m_2, \dots, m_5)$ be a sequence of hyperedge sizes with $m_2 = m_3 = m_4 = m_5 = n$. Now let $G \sim \mathrm{CL}(n, \mathbf{p}, \mathbf{m})$, and let $e_2, e_3, e_4, e_5 \in E(G)$ be chosen uniformly at random conditioned on $|e_i| = i$ for each $i \in \{2, 3, 4, 5\}$. Then, writing $X_{i,j}$ for the indicator variable which is 1 if $e_i \subset e_j$, we have

$\mathbb{E}[X_{2,3}] \propto n^{-2}$	$\mathbb{E}[X_{2,4}] \propto n^{-2}$	$\mathbb{E}[X_{2,5}] \propto n^{-2}$
	$\mathbb{E}[X_{3,4}] \propto n^{-3}$	$\mathbb{E}[X_{3,5}] \propto n^{-3}$
		$\mathbb{E}[X_{4,5}] \propto n^{-4}$

which implies

$\mathbb{E}[\mathrm{sp}(G,2,3)] \propto 1$	$\mathbb{E}[\mathrm{sp}(G,2,4)] \propto 1$	$\mathbb{E}[\mathrm{sp}(G,2,5)] \propto 1$
	$\mathbb{E}[\mathrm{sp}(G,3,4)] \propto 1/n$	$\mathbb{E}[\mathrm{sp}(G,3,5)] \propto 1/n$
		$\mathbb{E}[\mathrm{sp}(G,4,5)] \propto 1/n^2$

Now, let H be a hypergraph with degree sequence $\mathbf{d}$ and hyperedge-size sequence $\mathbf{m}$, and suppose H has one simplicial pair of each type. Then, based on the above calculations, we get that

$\sigma_{\mathrm{SR}}(H,2,3) \propto 1$	$\sigma_{\mathrm{SR}}(H,2,4) \propto 1$	$\sigma_{\mathrm{SR}}(H,2,5) \propto 1$
	$\sigma_{\mathrm{SR}}(H,3,4) \propto n$	$\sigma_{\mathrm{SR}}(H,3,5) \propto n$
		$\sigma_{\mathrm{SR}}(H,4,5) \propto n^2$

Thus, the above matrix acts as a loose, point-wise lower-bound on the simplicial matrix for sparse hypergraphs with at least one simplicial pair of each type. For many of the hypergraphs analysed, this rough sketch of a simplicial matrix is a good approximation of the actual matrices. In summary, what the simplicial matrix is capturing, above all else, is that (a) real hypergraphs contain simplicial pairs of all types, and (b) synthetic (sparse) models very rarely generate simplicial pairs other than pairs containing 2-hyperedges.

Temporal simplicial matrix

Here, the bias towards bottom-up simplicial pairs is consistent with the cell-wise comparisons. This suggests that the bias is independent, or at least not heavily dependent, on hyperedge size.

Disagreements for Social Datasets

We point out that our measures for the hospital-lyon dataset might be surprising, as they suggest that this dataset is slightly *less* simplicial than a fitted random graph. This conclusion broadly agrees with some prior measurements (see, e.g., [30]), but sharply disagrees with [27]. A similar but less-stark disagreement happens for the contact-p.s. dataset. While it is not possible to give a simple and complete explanation for why measures disagree on real datasets, we believe that example 2.6 gives a stylized explanation for much of the disagreement: the hospital-lyon hypergraph has a large fraction of all possible 2-edges and very few k-edges for $k > 3$, and so (as in that example) this forces the measure in [27] to be large while allowing our measure to be small.

4 A new model that incorporates simpliciality

In this section, we define a random hypergraph model, called the *simplicial Chung-Lu model*, that generalizes the Chung-Lu hypergraph model defined in [19]. We begin with the algorithm that generates a simplicial hyperedge.

Let $(d_1, \ldots, d_n)$ be a degree sequence, k be a hyperedge size, E be a set of existing hyperedges, and $E_k \subseteq E$ be a set of existing hyperedges that are of size k. Recalling that $p(\cdot)$ is the probability distribution governed by $(d_1, \ldots, d_n)$, writing $\binom{S}{k}$ for the collection of k-subsets of S, and recalling that $x \in_u X$ means x is sampled uniformly from X, the algorithm to generate a simplicial hyperedge is as follows.

Algorithm 3 Simplicial hyperedge.

Require: $(d_1, \ldots, d_n)$, k, E.
if $E \setminus E_k = \emptyset$ **then**
  Sample $e \sim$ **Algorithm 1**$\Big((d_1, \ldots, d_n), k\Big)$
else
  Sample $e' \in_u E \setminus E_k$.
  if $|e'| < k$ **then**
    Sample $e'' \sim$ **Algorithm 1**$\Big((d_1, \ldots, d_n), k - |e'|\Big)$.
    Set $e = e' \cup e''$
  else
    Sample $e \in_u \binom{e'}{k}$
  end if
end if
Return e

In words, we first check if there is at least one hyperedge in E *not* of size k to pair e with. If there is no such hyperedge, we return a Chung-Lu hyperedge. Otherwise, we choose an existing hyperedge e' uniformly at random from the set of hyperedges *not* of size k and construct our hyperedge e from e' in one of two ways: if $k < |e'|$ we set e to be a uniform k-subset of e', whereas if $k > |e'|$ we build e by combining e' with a Chung-Lu hyperedge of size $k - |e'|$.

We now give the algorithm to generate a simplicial Chung-Lu hypergraph. Let $(d_1, \ldots, d_n)$ be a degree sequence, $(m_{k_{\min}}, \ldots, m_{k_{\max}})$ be a sequence of hyperedge sizes, and $S = (s_1, \ldots, s_\ell)$ be a random permutation of all available sizes for a hyperedge, i.e., S contains m_k copies of k for each hyperedge size k in some random order. Additionally, let $q \in [0, 1]$ be a parameter governing the number of simplicial hyperedges created during the process.

Algorithm 4 Simplicial Chung Lu model.

Require: $(d_1, \dots, d_n)$, $(m_{k_{\min}}, \dots, m_{k_{\max}})$, q.

Initialize hyperedge list $E = \{\}$ and random hyperedge-size list S.
for $k \in S$ **do**
    Sample $X \sim \text{Bernoulli}(q)$.
    if $X = 1$ **then**
        Sample $e \sim$ **Algorithm 3**$\Big((d_1, \dots, d_n), k, E\Big)$
    else
        Sample $e \sim$ **Algorithm 1**$\Big((d_1, \dots, d_n), k\Big)$
    end if
    Set $E = E \cup \{e\}$.
end for
Return $G = ([n], E)$.

Note that, if $q = 0$, the simplicial Chung-Lu model yields a Chung-Lu model, ensuring that this new model is indeed a generalized Chung-Lu model. Moreover, the following lemma shows that the main feature of the Chung-Lu model is still present in this new model.

Lemma 4.1. *Let G be a random hypergraph generated as a simplicial Chung-Lu model with input parameters $(d_1, \dots, d_n)$, $(m_{k_{\min}}, \dots, m_{k_{\max}})$, and $q \in [0, 1]$. Then, for all $v \in [n]$,*

$$\mathbb{E}\left[\deg_G(v)\right] = d_v\,.$$

Proof. Let us generate a random hyperedge-size list S that will be used to create the simplicial Chung-Lu hypergraph G. We will first prove (by induction on i) the following claim.

Claim: Each vertex v of the i'th hyperedge e_i formed during the construction process of G satisfies

$$\mathbb{P}\left(v = u\right) = p(u) \text{ for all } u \in [n].$$

Note that hyperedges of G are not generated independently; the hypergraph has rich dependence structure. The distribution of e_i is affected by hyperedges generated earlier. It is important to keep in mind that the claim applies to the hyperedge formed at time i but without exposing information about earlier hyperedges.

Firstly, if $i = 1$, then e_1 is necessarily generated via **Algorithm 1** and the claim follows immediately. Now fix $i > 1$ and consider the formation of e_i. On the one hand, if e_i was generated via **Algorithm 1** then the claim is once again immediate. Otherwise, e_i was generated via **Algorithm 3**, i.e., generated constructively from another hyperedge e_j with $j < i$. In this case, if $|e_i| < |e_j|$ then $e_i \in_u \binom{e_j}{|e_i|}$ and, regardless which subset of e_j is selected to form e_i, the claim holds by induction. Otherwise, if $|e_i| > |e_j|$, then e_i is the union of e_j and another hyperedge e'' generated via **Algorithm 1**: the claim holds immediately for vertices in e'', and for vertices in e_j, the claim holds by induction.

Thus, for any $e \in E(G)$, $v \in e$, and $u \in [n]$, we have that $\mathbb{P}(v = u) = p(u)$. Summing over all vertices in all hyperedges, we get that

$$\mathbb{E}[\deg_G(u)] = \left(\sum_{e \in E(G)} \sum_{v \in e} \mathbb{P}(v = u) \right) = \left(p(u) \sum_{e \in E(G)} |e| \right) = \left(p(u) \sum_{v \in [n]} d_v \right) = d_u \,,$$

the first equality following from linearity of expectation, and the third equality following from the generalized handshaking lemma. □

The simplicial Chung Lu model does in fact generate more simplicial pairs as q increases. Figure 6 shows the expected number of simplicial pairs (approximated via 1000 samples) for hypergraphs generated via **Algorithm 4** with q varying from 0 to 1 in 0.1 increments.

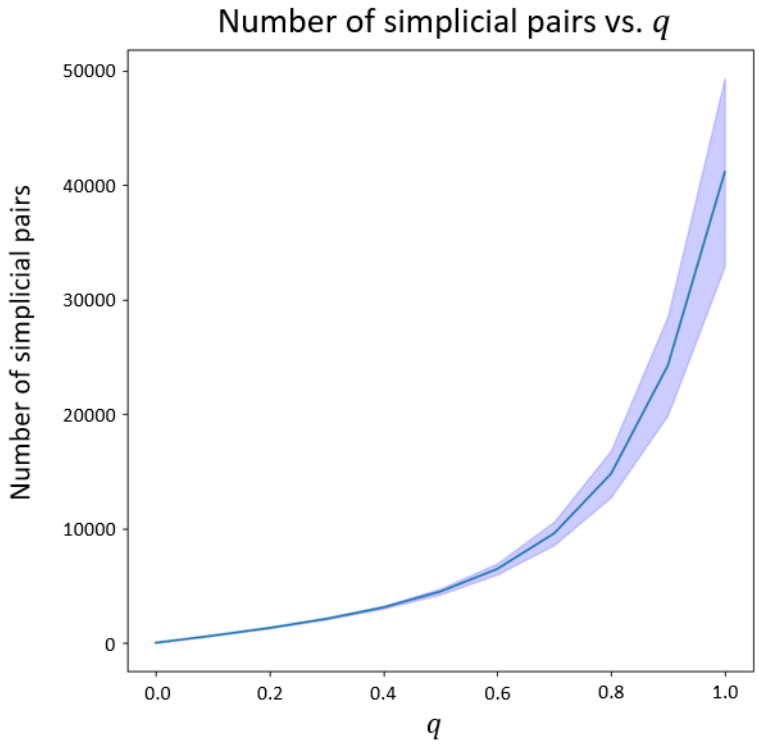


Figure 6: The average number of simplicial pairs (taken over 1000 samples) for simplicial Chung Lu hypergraphs with varying q. For each $q \in [0, 0.1, \ldots, 1]$, G_q is a simplicial Chung Lu hypergraph with $n = 1000$, $\mathbf{d}$ a uniform degree sequence, and $[|E_2|, |E_3|, |E_4|, |E_5|] = [5000, 1000, 100, 10]$. The shaded region represents the standard deviation over the 1000 samples.

5 Experiments

One reason to study simpliciality is that it likely has an impact on the evolution of stochastic processes on the associated hypergraphs. We illustrate this potential impact via two toy processes with varying parameters. The first process is component growth which a standard way to measure the robustness of a network (see, for example, Chapter 8 in [1]). The second type is information diffusion which simulates how quickly a substance (e.g., a disease, a rumour) spreads through a network. Intuitively, both of these processes should be affected by a hypergraph containing a large number of simplicial pairs: in the case of component growth the smaller hyperedge of a simplicial pair does not contribute to component size, and in the case of information diffusion a simplicial pair transfers information less efficiently than two non-overlapping hyperedges.

5.1 Descriptions of the experiments

We perform four experiments (two experiments for each of the two types of stochastic processes) on the real networks and on the corresponding simplicial Chung-Lu hypergraphs for varying $q \in \{0, q^*, 1\}$, where q^* is dataset dependent. For each dataset, we run a bisection method to obtain q^*, the value that best fits the number of simplicial pairs in the real dataset. Note that for two of the datasets (hospital-lyon and ubuntu), we get $q^* = 0$, so the corresponding graphs have one less curve.

Giant component growth with random hyperedge selection: We choose a uniform random order for $E(G)$ and track the size of the largest component as hyperedges are added to G according to a random ordering. We plot the growth up to the point where $\min\{|E(G)|, |V(G)|\}$ hyperedges have been added. We perform this experiment independently 100 times on the real hypergraphs, meaning we shuffle the hyperedge ordering and track the growth 100 times. For the simplicial Chung-Lu models we (a) sample the hypergraph, (b) shuffle the hyperedges, and (c) track the growth, performing steps (a), (b), and (c) independently 100 times.

Giant component growth with adversarial hyperedge selection: We order $E(G)$ in ascending order of betweenness (breaking ties randomly) and track the size of the largest component as hyperedges are added to G according to this adversarial ordering. Note that the betweenness of a hyperedge e in a hypergraph is equivalent to the betweenness of its corresponding vertex v_e in the line-hypergraph (see [12], or any textbook on network science such as [21], for a definition of betweenness for hypergraphs). For the real hypergraphs, we run the experiment only once (the results will be the same every time), and for the Chung-Lu models we sample and track growth 30 times (10 times for the three largest hypergraphs). We sample less here than in the other three experiments due to the time complexity of calculating betweenness.

Information diffusion from a single source: We initialize a function $w_0 : V(G) \to [0, 1]$ with $w_0(v) = 0$ for all vertices, except for one randomly chosen vertex v^* which has

$w_0(v^*) = 1$. Then, in round $i+1$, we choose a random hyperedge e and, for each $v \in e$, set $w_{i+1}(v) = w(e)/|e|$, where $w(e) = \sum_{u \in e} w(u)$ (keeping $w_{i+1}(v) = w_i(v)$ for all $v \notin e$). We track the Wasserstein-1 distance (also known as the "earth mover's distance" [34]) between w_i and the uniform distribution $w_\infty : V(G) \to 1/|V(G)|$. We run the experiment 100 times, re-rolling the Chung-Lu model every time.

Information diffusion from $|V(G)|/10$ sources: This experiment is identical to the previous experiment, except that $w_0(v^*) = 1$ for 10% of the vertices chosen at random, and that $w_\infty : V(G) \to 1/10$.

Insisting on connected hypergraphs

These experiments, and in particular the two diffusion experiments, are highly dependent on connectivity. The real hypergraphs are restricted to their largest component, and so we insist that the random hypergraphs are also connected. To achieve this, we modify the simplicial Chung-Lu model and insist that incoming hyperedges must connect disjoint components, until the point the hypergraph is connected when we continue generating hyperedges as normal. A full description of this algorithm is presented in Appendix B.

5.2 The results

Here, we will show the results for the two hypergraphs: **hospital-lyon** and **disgenenet**. Recall that the **hospital-lyon** hypergraph has a simplicial ratio of approximately 0.94, whereas the **disgenenet** hypergraph has a ratio of approximately 28.81. The full collection of results can be found in Appendix A and the sampling technique can be found in Appendix B. For each curve, the shaded area corresponds to the standard error of the mean in cases where multiple repeated experiments were conducted.

Experiment 1: random growth

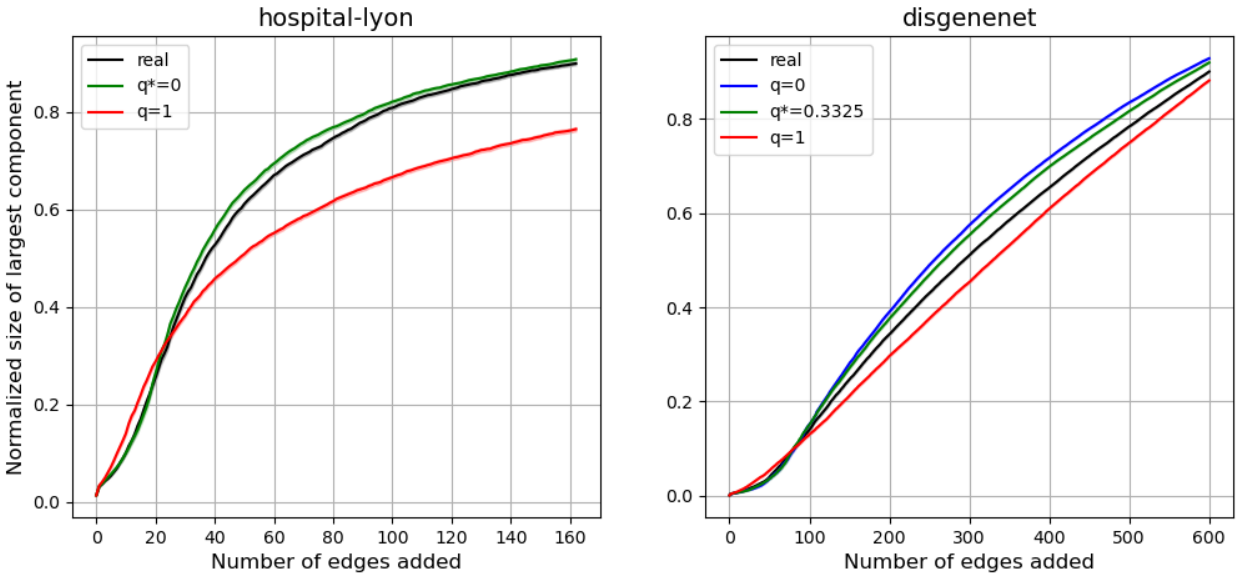


Figure 7: Giant component size (normalized by the number of vertices) vs. number of hyperedges added in the random growth process on the **hospital-lyon** hypergraph (left) and the **disgenenet** hypergraph (right). The curve is the point-wise average across 100 independent experiments: for the real hypergraph the hyperedges are resampled each time, and for the random models the entire hypergraphs are resampled each time.

In the first experiment shown in Figure 5.2, we see the following. For **hospital-lyon** the real hypergraph grows in a near identical way to the random model with $q^* = 0$ whereas the random model with $q = 1$ grows much slower. In contrast, for **disgenenet** the real hypergraph grows at a rate somewhere between the random model with $q^* = 0.3325$ and $q = 1$. Of course, these hypergraphs have very different growth behaviour due to the difference in hyperedge densities. Nevertheless, this result suggests that the high simplicial ratio of **disgenenet** plays a role in slowing down the growth of the hypergraph, whereas the low simplicial ratio of **hospital-lyon** leads it to grow as quickly as a classical Chung-Lu model.

Experiment 2: adversarial growth

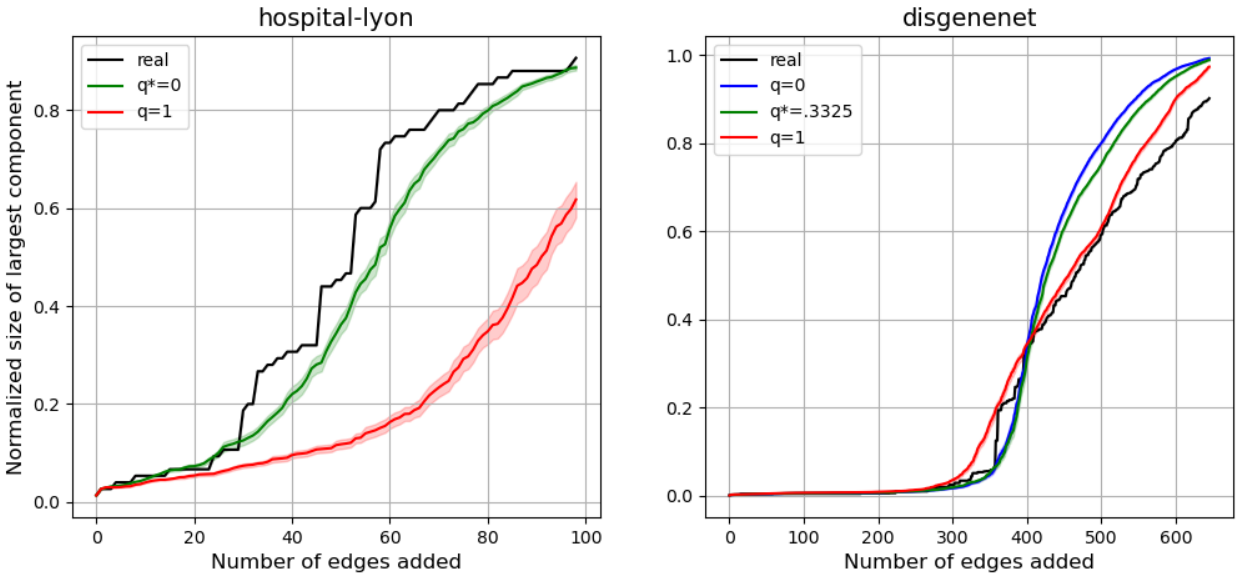


Figure 8: Giant component size vs. number of hyperedges added in the adversarial growth process on the **hospital-lyon** hypergraph (left) and the **disgenenet** hypergraph (right). The curve is the point-wise average across 30 independent experiments: for the real hypergraph the experiment is performed only once as the result will always be the same, and for the random models the hypergraphs are resampled each time.

The results of the second experiment, shown in Figure 8, still show a clear distinction between the real growth vs. the synthetic growth for these two hypergraphs. On the left, we see that the real hypergraph grows faster than all the random models, whereas on the right the real hypergraph grows slower than all the models.

Experiment 3: single-source diffusion

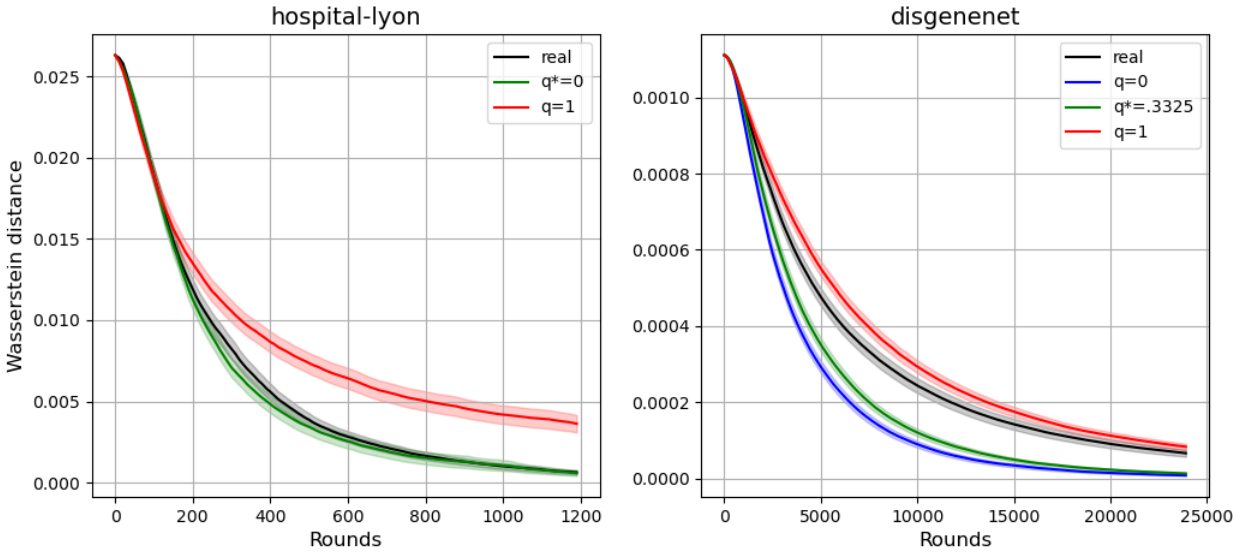


Figure 9: Wasserstein distance to uniform vs. number of rounds in the single-source diffusion process on the **hospital-lyon** hypergraph (left) and the **disgenenet** hypergraph (right). The curve is the point-wise average across 100 independent experiments: for the real hypergraph the chosen hyperedges per round, as well as the location of the initial vertex with weight 1, are resampled each time, and for the random models the entire hypergraphs are resampled each time.

The third experiment is perhaps the most substantial in showing the effect of simpliciality on a random process, namely, that information diffusion is slower on highly simplicial hypergraphs vs. non-simplicial hypergraphs. The results are shown in Figure 9.

Experiment 4: 10% diffusion

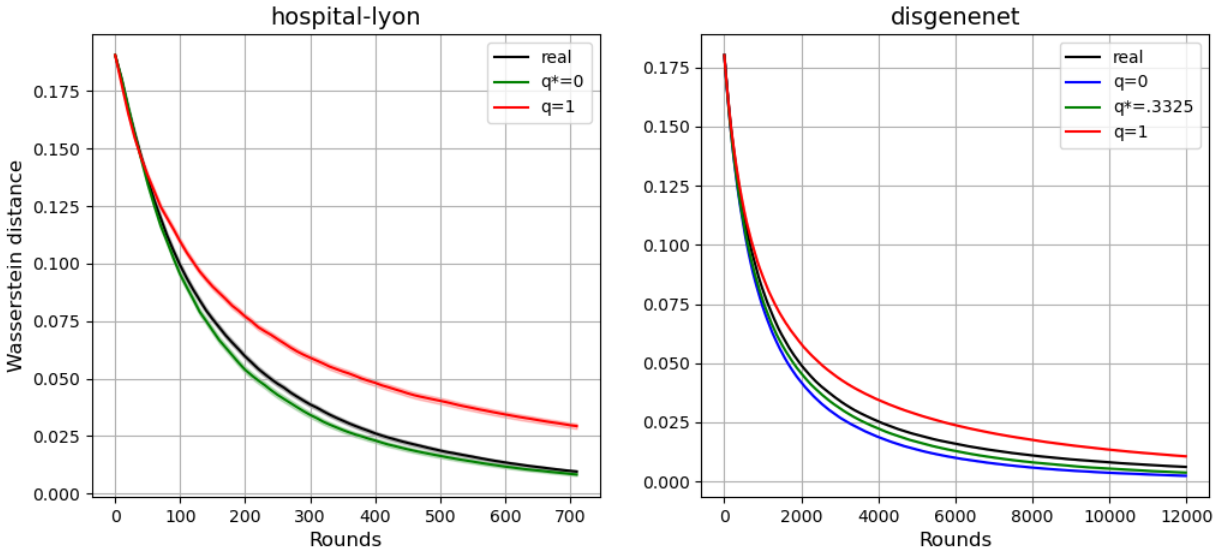


Figure 10: Wasserstein distance to uniform vs. number of rounds in the 10% sprinkled diffusion process on the **hospital-lyon** hypergraph (left) and the **disgenenet** hypergraph (right). The curve is the point-wise average across 100 independent experiments: for the real hypergraph the chosen hyperedges per round, as well as the location of the initial 10% of vertices with weight 1, are resampled each time, and for the random models the entire hypergraphs are resampled each time.

The result of the fourth experiment, shown in Figure 10, mirrors the results of the previous experiment, except of course that the diffusion is much faster.

Our hypergraph model is a simple 1-parameter model, and our process for finding the optimal value q^* is based solely on matching the number of simplicial pairs (and so does not directly try to match the behavior of stochastic processes). As such, we do not think it is plausible to fully capture the behaviour of complex stochastic processes on these hypergraphs. Nonetheless, the experiments show that even a very simple model and an optimization process that does not target stochastic processes can often lead to *substantially* better agreement between stochastic processes, providing some evidence that simplicial surprise alone is very informative.

6 Conclusion

The phenomenon of hyperedges inside of other hyperedges is a feature of hypergraphs not present in graphs, and, based on our results and on the preceding results of Landry, Young and Eikmeier, it is clear that this phenomenon is a key feature of real-world networks with multi-way interactions. The simplicial ratio captures the strength of simplicial interactions

in a hypergraph and, from the collection of 10 real-world networks analysed, we have showed that (a) the simplicial ratio is not at all consistent across the hypergraphs, (b) the simplicial ratio varies significantly even for hypergraphs of a similar type (e.g., **contact high-school**, **contact primary-school**, and **hospital-lyon**), (c) the number of simplicial interactions involving hyperedges of size $k, \ell > 2$ is not at all captured by the Chung Lu model, and (d) the simplicial ratio can affect the outcome of random growth, adversarial growth, and information diffusion. We hope that our work continues to motivate research into the phenomenon of hyperedges inside hyperedges, and we discuss some potential follow-ups to this research.

6.1 Further research

The simplicial ratio involves the parameter $\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right]$ where $\hat{G} \sim \mathrm{CL}(G)$. Instead of approximating $\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right]$ as we do, we could compute $\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right]$ explicitly. For example, given a uniform degree sequence $\mathbf{d}$ and a hyperedge size sequence $(m_{k_{\min}}, \ldots, m_{k_{\max}})$, and conditional on $\hat{G}$ containing no multi-set hyperedges, the probability that e_1, e_2 form a simplicial pair is

$$\binom{|e_2|}{|e_1|} \Big/ \binom{n}{|e_1|}.$$

Thus, by linearity of expectation, conditioning on the event that $\hat{G}$ has no multiset hyperedges, we have

$$\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right] = \sum_{k=k_{\min}}^{k_{\max}-1} \sum_{\ell=k+1}^{k_{\max}} m_k m_\ell \binom{\ell}{k} \Big/ \binom{n}{k}.$$

Thus, for a uniform degree sequence, $\mathbb{E}\left[\mathrm{sp}\left(\hat{G}\right)\right]$ is relatively straightforward to compute. However, trying to compute this expectation if the degree sequence is not uniform is significantly harder. Finding a closed form for this expectation, or even a closed form approximation, would allow a significantly faster algorithm for computing $\sigma_{\mathrm{SR}}(G)$. Such a result would also allow for a better understanding of the nature of the simplicial matrix for both sparse and dense hypergraphs.

Understanding the degree to which hyperedges form simplicial pairs could aid in predicting the composition of future hyperedges, especially large hyperedges, in temporal networks. If a hypergraph has a high simplicial ratio, then a potential new hyperedge should be given more weight based on the number of new simplicial pairs it creates, as well as on the size of the smaller hyperedge in each pairs. For example, when considering the location of a new hyperedge of size 5 in a highly simplicial hypergraph G, a location that creates many $(2, 5)$ pairs should receive more weight, but perhaps a location that creates a single $(4, 5)$ pair should be given *even more* weight. In any case, incorporating simpliciality in the link prediction problem should improve existing algorithms and models (e.g., the models presented in [13, 14, 33]).

Along with the simplicial ratio and simplicial matrix, we introduce temporal variants. In our experiments where only the first instance of a hyperedge is kept in a temporal network,

we find that, typically, more bottom-up pairs are generated than top-down pairs, in part because there are more small multi-hyperedges than large multi-hyperedges. There are of course other ways to measure the difference in frequency between bottom-up pairs and top-down pairs. For example, we could insist that a simplicial pair e_k, e_ℓ is "temporally relevant" if and only if both e_k and e_ℓ were born within the same ϵ-window of time. In this case, we could measure the frequency of e_k pairs followed shortly by e_ℓ pairs, and vice versa. The temporal formation of simplicial pairs could once again be valuable for the task of link prediction.

Acknowledgements

This research is supported in part by the Tutte Institute for Mathematics and Computing (TIMC), a division of the Communications Security Establishment (CSE). TIMC does not specifically endorse the contents of this work, or any other work by the authors. Any opinions or positions represented herein do not represent the official position of CSE or the Government of Canada.

References

[1] Albert-László Barabási and Márton Pósfai. *Network science*. Cambridge University Press, Cambridge, 2016. URL: `http://barabasi.com/networksciencebook/`.

[2] Jordan Barret, Paweł Prałat, Aaron Smith, and François Théberge. Counting simplicial pairs in hypergraphs. In *International Conference on Complex Networks and Their Applications*, page in press. Springer, 2024.

[3] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. Networks beyond pairwise interactions: structure and dynamics. *Physics Reports*, 874:1–92, 2020.

[4] Austin R Benson, Rediet Abebe, Michael T Schaub, Ali Jadbabaie, and Jon Kleinberg. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences*, 115(48):E11221–E11230, 2018.

[5] Austin R Benson, David F Gleich, and Desmond J Higham. Higher-order network analysis takes off, fueled by classical ideas and new data. *arXiv preprint arXiv:2103.05031*, 2021.

[6] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[7] Giulia Cencetti, Federico Battiston, Bruno Lepri, and Márton Karsai. Temporal properties of higher-order interactions in social networks. *Scientific reports*, 11(1):7028, 2021.

[8] Alberto Ceria and Huijuan Wang. Temporal-topological properties of higher-order evolving networks. *Scientific Reports*, 13(1):5885, 2023.

[9] Fan RK Chung and Linyuan Lu. *Complex graphs and networks*. Number 107. American Mathematical Soc., 2006.

[10] David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press, 2010.

[11] Song Feng, Emily Heath, Brett Jefferson, Cliff Joslyn, Henry Kvinge, Hugh D Mitchell, Brenda Praggastis, Amie J Eisfeld, Amy C Sims, Larissa B Thackray, et al. Hypergraph models of biological networks to identify genes critical to pathogenic viral response. *BMC bioinformatics*, 22(1):287, 2021.

[12] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[13] Luca Gallo, Lucas Lacasa, Vito Latora, and Federico Battiston. Higher-order correlations reveal complex memory in temporal hypergraphs. *Nature Communications*, 15, 2024. doi:10.1038/s41467-024-48578-6.

[14] Iacopo Iacopini, Marton Karsai, and Alain Barrat. The temporal dynamics of group interactions in higher-order social networks. *Nature Communications*, 15, 2024. doi:10.1038/s41467-024-50918-5.

[15] Matthew O Jackson. *Social and economic networks*. Princeton university press, 2010.

[16] Cliff A. Joslyn, Sinan G. Aksoy, Tiffany J. Callahan, Lawrence E. Hunter, Brett Jefferson, Brenda Praggastis, Emilie Purvine, and Ignacio J. Tripodi. Hypernetwork science: From multidimensional networks to computational topology. In *Unifying Themes in Complex Systems X*, pages 377–392, Cham, 2021. Springer International Publishing.

[17] Jonas L Juul, Austin R Benson, and Jon Kleinberg. Hypergraph patterns and collaboration structure. *Frontiers in Physics*, 11:1301994, 2024.

[18] Bogumił Kamiński, Łukasz Kraiński, Paweł Prałat, and François Théberge. A multipurposed unsupervised framework for comparing embeddings of undirected and directed graphs. *Network Science*, 10(4):323–346, 2022.

[19] Bogumił Kamiński, Valérie Poulin, Paweł Prałat, Przemysław Szufel, and François Théberge. Clustering via hypergraph modularity. *PloS one*, 14(11):e0224307, 2019.

[20] Bogumił Kamiński, Paweł Prałat, and François Théberge. An unsupervised framework for comparing graph embeddings. *Journal of Complex Networks*, 8(5):cnz043, 2020.

[21] Bogumil Kaminski, Pawel Prałat, and François Théberge. *Mining complex networks*. Chapman and Hall/CRC, 2021.

[22] Bogumił Kamiński, Paweł Prałat, and François Théberge. Hypergraph artificial benchmark for community detection (h–abcd). *Journal of Complex Networks*, 11(4):cnad028, 2023.

[23] Bogumił Kamiński, Paweł Misiorek, Paweł Prałat, and François Théberge. Modularity based community detection in hypergraphs, 2024. URL: https://arxiv.org/abs/2406.17556, arXiv:2406.17556.

[24] Jihye Kim, Deok-Sun Lee, and K.-I. Goh. Contagion dynamics on hypergraphs with nested hyperedges. *Physical Review E*, 108(3), 2023. URL: http://dx.doi.org/10.1103/PhysRevE.108.034313, doi:10.1103/physreve.108.034313.

[25] Renaud Lambiotte, Martin Rosvall, and Ingo Scholtes. Understanding complex systems: From networks to optimal higher-order models. *arXiv preprint arXiv:1806.05977*, 2018.

[26] Nicholas W. Landry, Maxime Lucas, Iacopo Iacopini, Giovanni Petri, Alice Schwarze, Alice Patania, and Leo Torres. XGI: A Python package for higher-order interaction networks. *Journal of Open Source Software*, 8(85):5162, May 2023. URL: https://joss.theoj.org/papers/10.21105/joss.05162, doi:10.21105/joss.05162.

[27] Nicholas W Landry, Jean-Gabriel Young, and Nicole Eikmeier. The simpliciality of higher-order networks. *EPJ Data Science*, 13(1):17, 2024.

[28] Timothy LaRock and Renaud Lambiotte. Encapsulation structure and dynamics in hypergraphs. *Journal of Physics: Complexity*, 4(4):045007, nov 2023. URL: https://dx.doi.org/10.1088/2632-072X/ad0b39, doi:10.1088/2632-072X/ad0b39.

[29] Geon Lee, Fanchen Bu, Tina Eliassi-Rad, and Kijung Shin. A survey on hypergraph mining: Patterns, tools, and generators, 2024. URL: https://arxiv.org/abs/2401.08878, arXiv:2401.08878.

[30] Quintino Francesco Lotito, Federico Musciotto, Alberto Montresor, and Federico Battiston. Hyperlink communities in higher-order networks. *Journal of Complex Networks*, 12(2), March 2024. Publisher Copyright: © The Author 2024. Published by Oxford University Press. All rights reserved. doi:10.1093/comnet/cnae013.

[31] Mark Newman. *Networks*. Oxford university press, 2018.

[32] Tom Odda. On properties of a well-known graph or what is your ramsey number. *Annals of the New York Academy of Sciences*, 328:166 – 172, 12 2006. doi:10.1111/j.1749-6632.1979.tb17777.x.

[33] H. A. Bart Peters, Alberto Ceria, and Huijuan Wang. Higher-order temporal network prediction and interpretation, 2024. URL: https://arxiv.org/abs/2408.05165, arXiv:2408.05165.

[34] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, 1998. doi:10.1109/ICCV.1998.710701.

[35] Hao Tian and Reza Zafarani. Higher-order networks representation and learning: A survey. *ACM SIGKDD Explorations Newsletter*, 26(1):1–18, 2024.

[36] Leo Torres, Ann S. Blevins, Danielle Bassett, and Tina Eliassi-Rad. The why, how, and when of representations for complex systems. *SIAM Review*, 63(3):435–485, 2021. doi:10.1137/20M1355896.

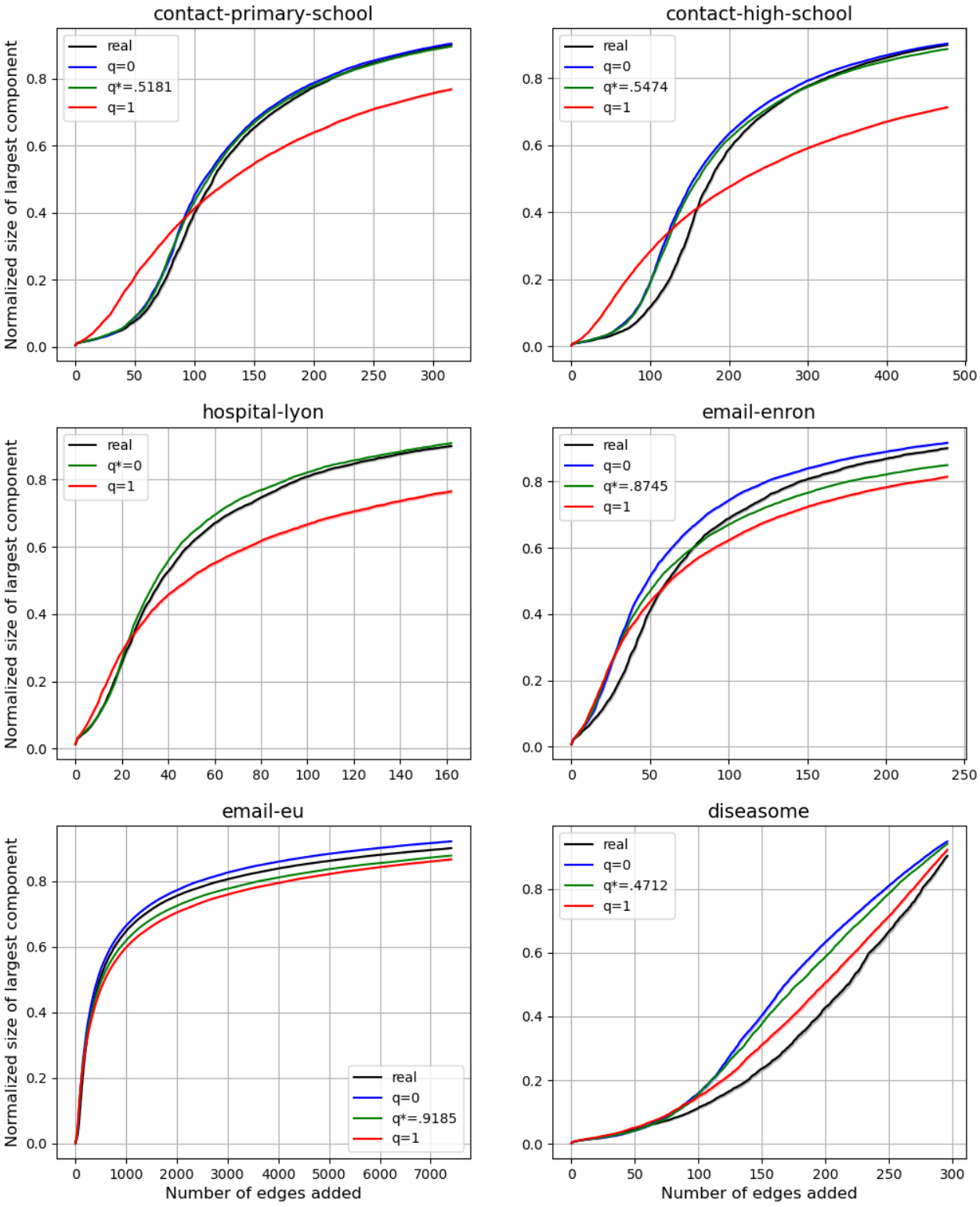
[37] Dominic Yeo. Multiplicative coalescence, 2012. URL: https://api.semanticscholar.org/CorpusID:2555801.

[38] Yuanzhao Zhang, Maxime Lucas, and Federico Battiston. Higher-order interactions shape collective dynamics differently in hypergraphs and simplicial complexes. *Nature Communications*, 14(1), 2023. doi:10.1038/s41467-023-37190-9.

A All experiments

Here we show the results of the random growth, adversarial growth, single-source diffusion, and 10% diffusion experiments. Note that **ubuntu (edge-chopped)** is the subhypergraph of **tags-ask-ubuntu** containing only the first 20,000 hyperedges. The simplicial ratio of this edge-chopped hypergraph is ≈ 0.37 and so this subhypergraph is even less simplicial than the whole hypergraph.

The experiments are presented in the the following order: random growth, adversarial growth, single-source diffusion, and 10% diffusion. Each of the four figures are presented on two separate pages.


contact-primary-school
real
q=0
q*=.5181
q=1
contact-high-school
real
q=0
q*=.5474
q=1
hospital-lyon
real
q*=0
q=1
email-enron
real
q=0
q*=.8745
q=1
email-eu
real
q=0
q*=.9185
q=1
diseasome
real
q=0
q*=.4712
q=1
Normalized size of largest component
Number of edges added

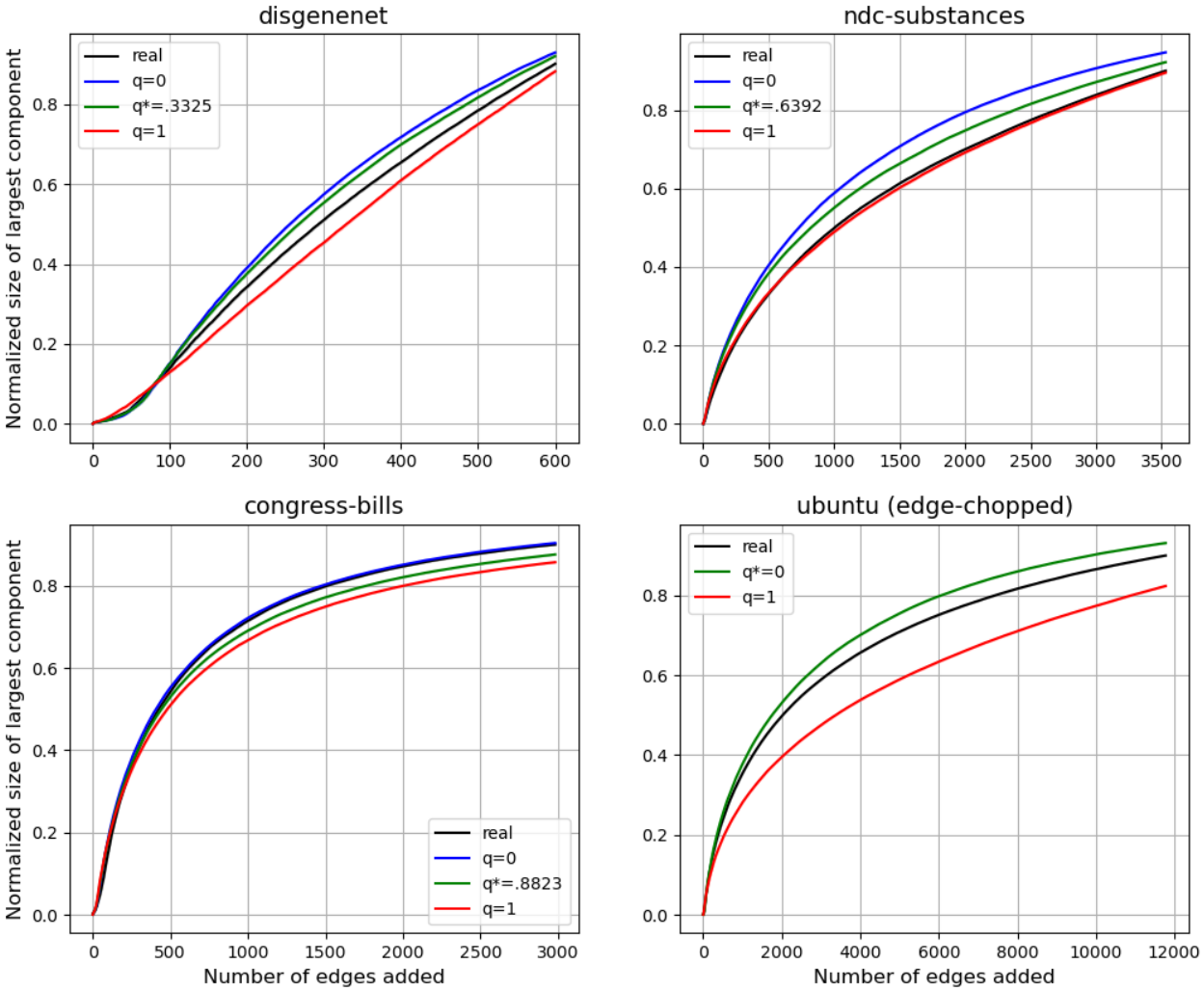
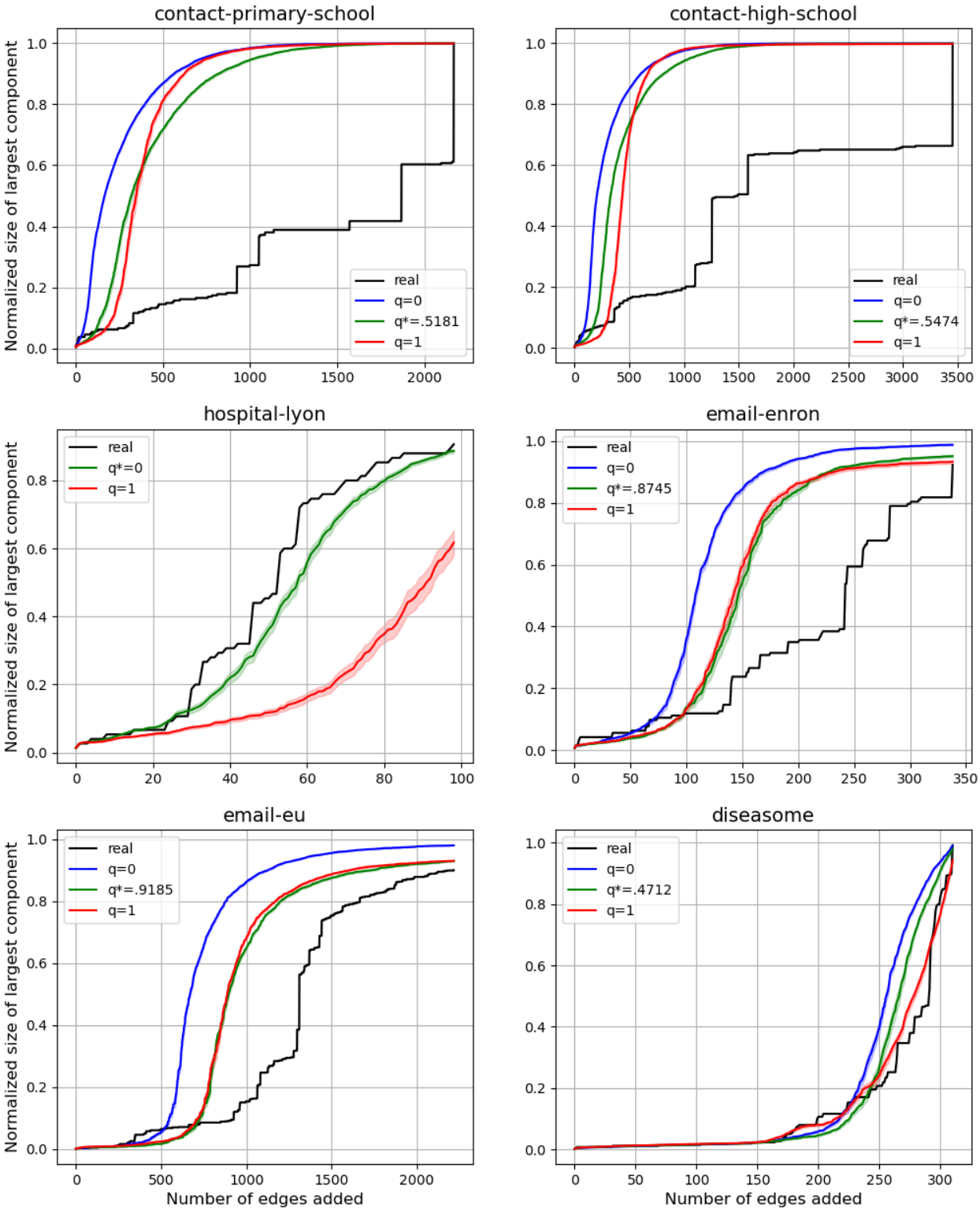


Figure 11: Giant component size (normalized by the number of vertices) vs. number of hyperedges added in the random growth process for all 10 hypergraphs. The curve is the point-wise average across 100 independent experiments: for the real hypergraph the hyperedges are resampled each time, and for the random models the entire hypergraphs are resampled each time.


contact-primary-school
Normalized size of largest component
real
q=0
q*=.5181
q=1
contact-high-school
real
q=0
q*=.5474
q=1
hospital-lyon
real
q*=0
q=1
email-enron
real
q=0
q*=.8745
q=1
email-eu
real
q=0
q*=.9185
q=1
Number of edges added
diseasome
real
q=0
q*=.4712
q=1
Number of edges added

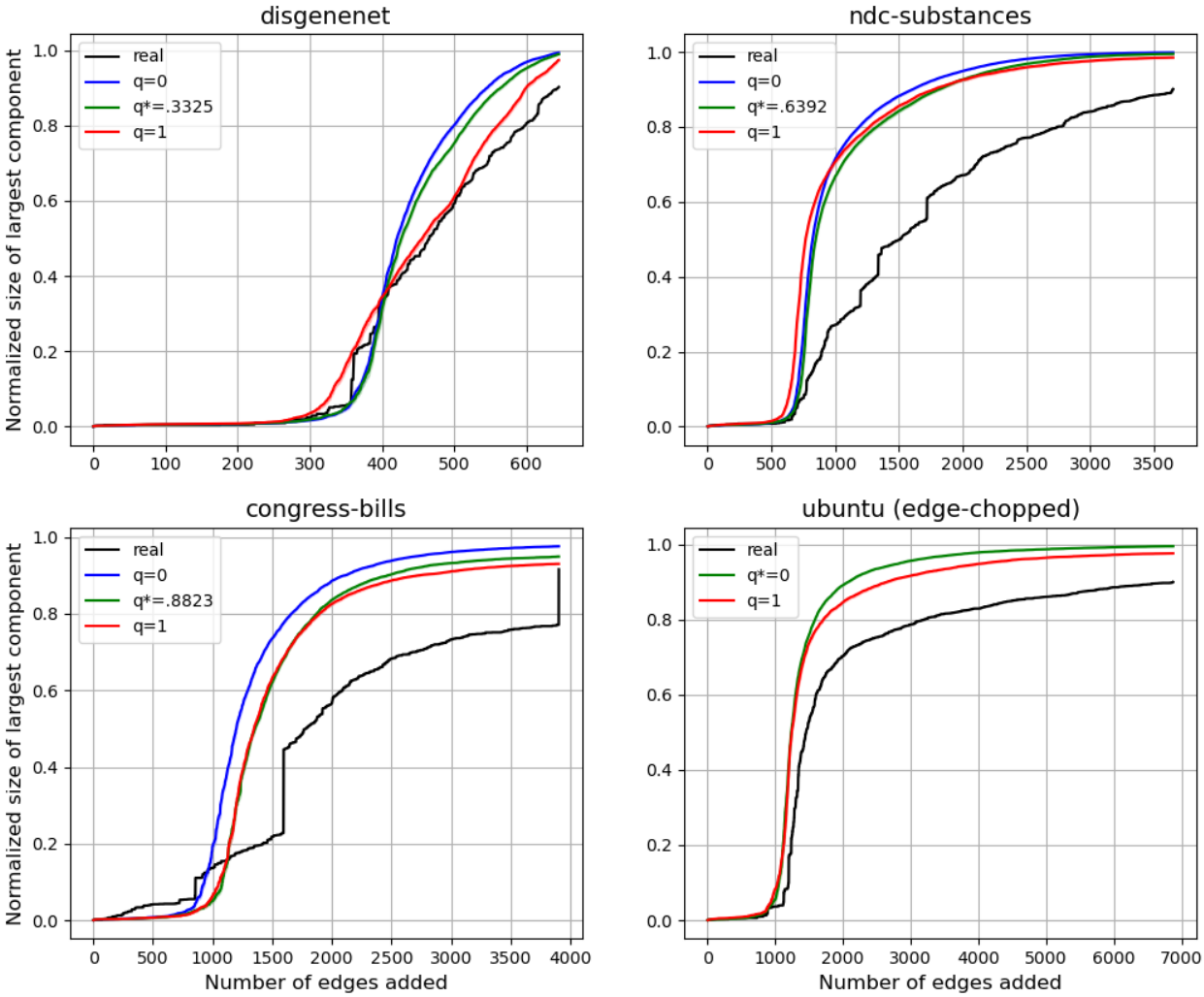
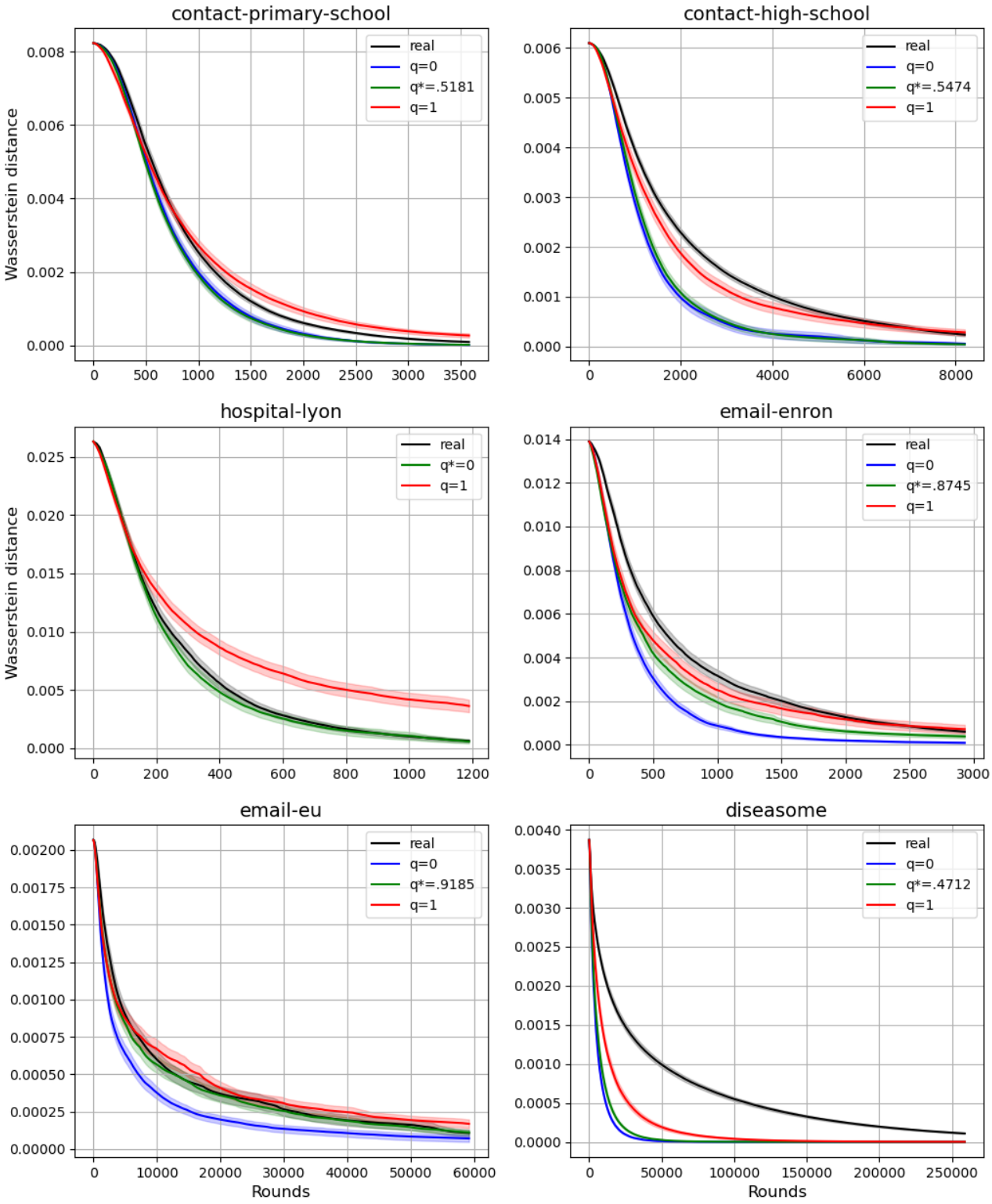


Figure 12: Giant component size (normalized by the number of vertices) vs. number of hyperedges added in the adversarial growth process for all 10 hypergraphs. The curve is the point-wise average across 30 independent experiments (10 for the three largest graphs). For the real hypergraph the experiment is performed only once as the result will always be the same, and for the random models the hypergraphs are resampled each time.


contact-primary-school
real
q=0
q*=.5181
q=1
contact-high-school
real
q=0
q*=.5474
q=1
hospital-lyon
real
q*=0
q=1
email-enron
real
q=0
q*=.8745
q=1
email-eu
real
q=0
q*=.9185
q=1
diseasome
real
q=0
q*=.4712
q=1
Wasserstein distance
Rounds

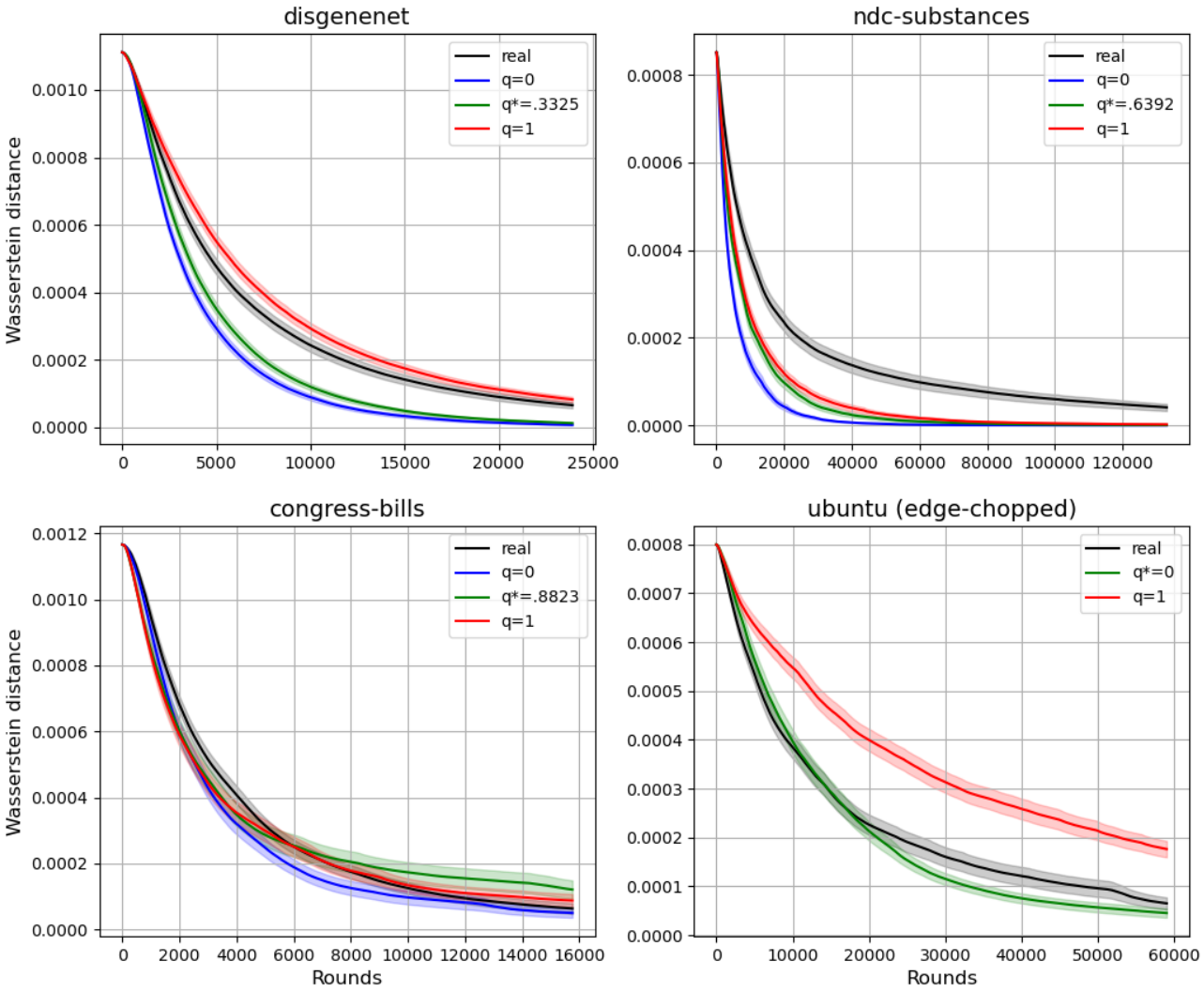
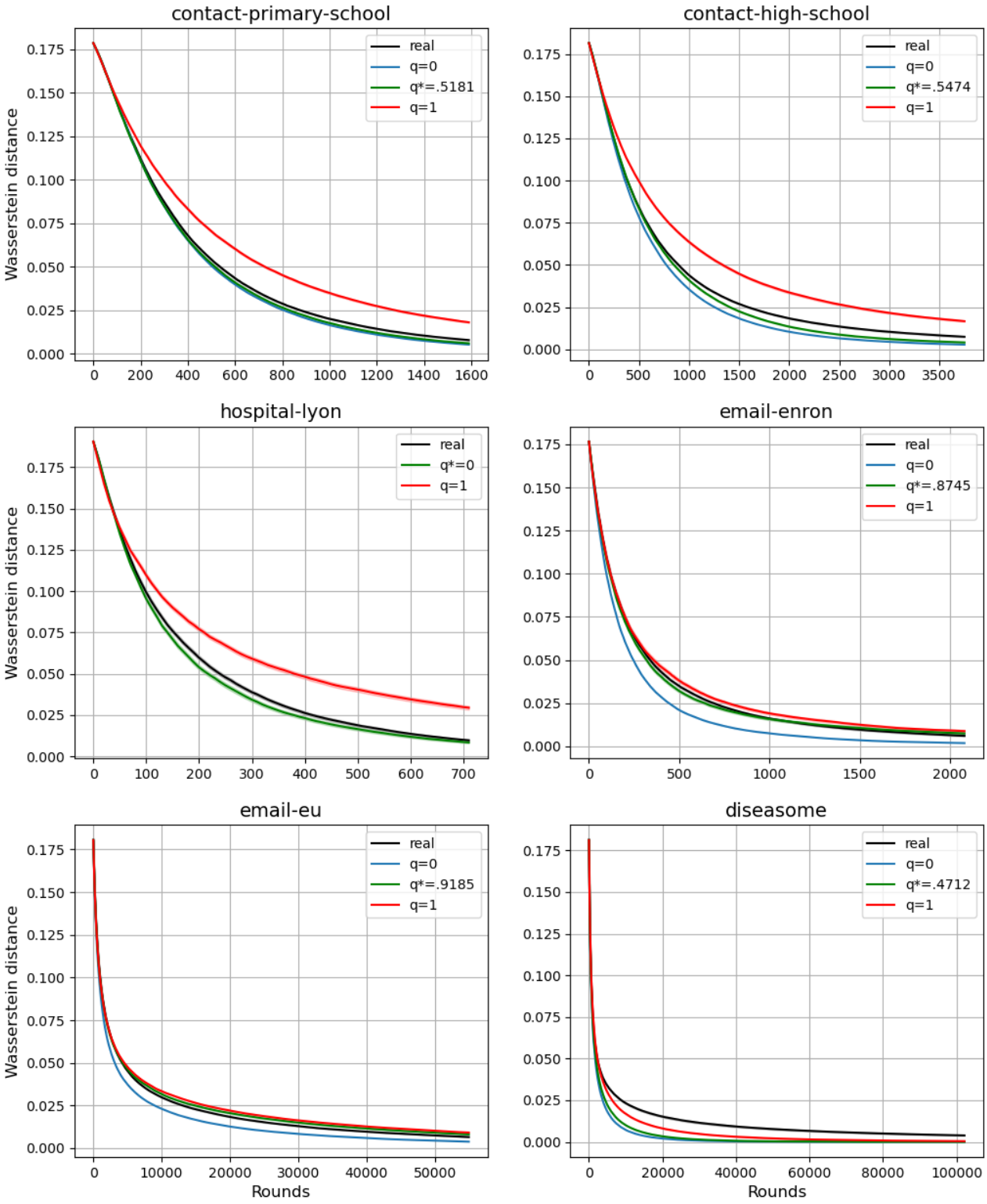


Figure 13: Wasserstein distance to uniform vs. number of rounds in the single-source diffusion process for all 10 hypergraphs. The curve is the point-wise average across 100 independet experiments: for the real hypergraph the chosen hyperedges per round, as well as the location of the initial vertex with weight 1, are resampled each time, and for the random models the entire hypergraphs are resampled each time.


contact-primary-school
real
q=0
q*=.5181
q=1
contact-high-school
real
q=0
q*=.5474
q=1
hospital-lyon
real
q*=0
q=1
email-enron
real
q=0
q*=.8745
q=1
email-eu
real
q=0
q*=.9185
q=1
diseasome
real
q=0
q*=.4712
q=1
Wasserstein distance
Rounds

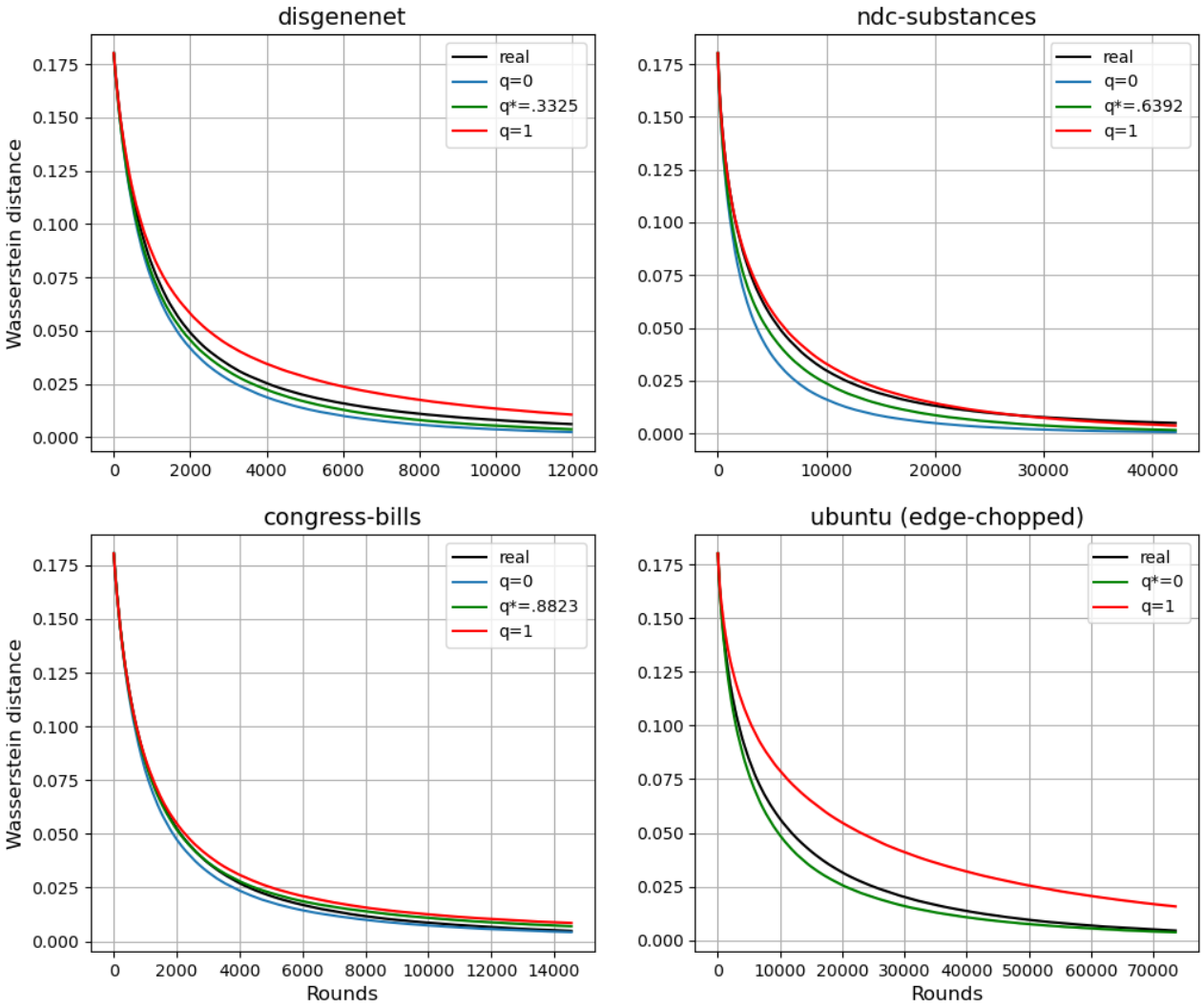


Figure 14: Wasserstein distance to uniform vs. number of rounds in the 10% sprinkled diffusion process for all 10 hypergraphs. The curve is the point-wise average across 100 independent experiments: for the real hypergraph the chosen hyperedges per round, as well as the location of the initial 10% of vertices with weight 1, are resampled each time, and for the random models the entire hypergraphs are resampled each time.

B Algorithms

B.1 Estimating the expected number of simplicial pairs

To compute the simplicial ratio of a hypergraph G, we must first compute the expected number of simplicial pairs in $\hat{G} \sim \mathrm{CL}(G)$. As discussed in Section 6, computing this expectation is quite difficult. In this section, we outline a Monte Carlo approximate technique for this expectation.

For a degree sequence $\mathbf{d} = (d_1, \dots d_n)$ and a hyperedge size k, write $\mathbb{P}(\text{simple} \mid \mathbf{d}, k)$ for the probability that **Algorithm 1** generates a simple hyperedge when given inputs $\mathbf{d}$ and k. For a hypergraph G with degree sequence $\mathbf{d}$, we first approximate $\mathbb{P}(\text{simple} \mid \mathbf{d}, k)$ for all hyperedge sizes k in $E(G)$. To do this, we chose a number of samples s, sample s hyperedges independently as **Algorithm 1**$(\mathbf{d}, k)$, and compute the ratio x/s where x is the number of simple hyperedges generated. In all experiments performed for the paper, we use $s = 1000$.

With $\mathbb{P}(\text{simple} \mid \mathbf{d}, k)$ approximated for all hyperedge sizes k, we can now approximate the number of simplicial pairs. We will show the algorithm for computing the expected number of $(3,5)$-pairs here, as the generalization is straightforward to interpret but difficult to notate. Write $|\mathbf{d}| := \sum_{i \in [n]} d_i$. For a hyperedge $e = \{v_1, \dots, v_5\}$, the probability that a hyperedge e' of size 3 generated by **Algorithm 1** is (a) simple and (b) satisfies $e' \subset e$ is given by

$$\sum_{1 \le a < b < c \le 5} \frac{3!\, d_{v_a} d_{v_b} d_{v_c}}{(|\mathbf{d}|)^3\, \mathbb{P}(\text{simple} \mid \mathbf{d}, 3)}\,. \tag{1}$$

To break this down, consider only the probability that $e' = \{v_1, v_2, v_3\}$. **Algorithm 1** can generate this hyperedge in 3! different orders, and the probability of generating the hyperedge in each case is

$$\frac{d_{v_1} d_{v_2} d_{v_3}}{|\mathbf{d}|^3}\,.$$

It can also happen that **Algorithm 1** generates a multi-hyperedge, requiring us to sample again. Thus, the probability of *eventually* sampling the hyperedge $e' = \{v_1, v_2, v_3\}$ is

$$\begin{aligned}
\sum_{i \ge 0} (1 - \mathbb{P}(\text{simple} \mid \mathbf{d}, 3))^i \frac{3!\, d_{v_1} d_{v_2} d_{v_3}}{|\mathbf{d}|^3} &= \frac{3!\, d_{v_1} d_{v_2} d_{v_3}}{|\mathbf{d}|^3} \sum_{i \ge 0} (1 - \mathbb{P}(\text{simple} \mid \mathbf{d}, 3))^i \\
&= \frac{3!\, d_{v_1} d_{v_2} d_{v_3}}{|\mathbf{d}|^3} \left(\frac{1}{1 - (1 - \mathbb{P}(\text{simple} \mid \mathbf{d}, 3))} \right) \\
&= \frac{3!\, d_{v_1} d_{v_2} d_{v_3}}{(|\mathbf{d}|^3)\, \mathbb{P}(\text{simple} \mid \mathbf{d}, 3)}\,.
\end{aligned}$$

Summing over all $\binom{5}{3}$ possible 3-hyperedges inside e gives us (1).

We now approximate the number of (k, ℓ) simplicial pairs as follows.

1. Choose some sampling number s. Then, sample s independent hyperedges via **Algorithm 1**$(\mathbf{d}, \ell)$.
2. For each hyperedge, compute the probability of generating a (k, ℓ) simplicial pair.

3. Compute the average and multiply this result by $m_k m_l$, where m_k is the number of hyperedges of size k, and similarly for m_ℓ.

B.2 Constructing a connected skeleton of a random hypergraph

We will generate a connected skeleton for our random hypergraph via multiplicative coalescence. In short, multiplicative coalescence is a process in which particles in a space join together at a rate proportional to the product of their masses. We point the reader to [37] for an overview on the multiplicative coalescence process. In the context of generating random hypergraphs, multiplicative coalescence is the process where new hyperedges joining disjoint components are chosen with probability proportional to the product of the weights of the components.

We will describe **Algorithm 5** in words before presenting it as pseudo-code. Let $\mathbf{d} := (d_1, \ldots, d_n)$ be a degree sequence and $\mathbf{m} := (m_{k_{\min}}, \ldots, m_{k_{\max}})$ be a sequence of hyperedge sizes. We construct the skeleton of our hypergraph as follows.

1. Initially, we have an empty hyperedge list $E = \{\}$ and a collection of components, one for each vertex. For component $C = \{v\}$, define the weight of C, written $w(C)$, as $w(C) := d_v$.

2. We generate a random hyperedge-size list S as per **Algorithm 4**, i.e., a uniform permutation containing m_k copies of k for each hyperedge size k.

3. Iteratively until the hypergraph is connected, we do the following.

   (a) Choose a size k from S (iteratively).

   (b) Sample k components independently, each component C being chosen with probability proportional to $w(C)$. If the chosen components $C_1, \ldots, C_k$ are not all unique, discard them all and sample again (repeating until we have a collection of distinct components).

   (c) For each component C chosen in the previous step, randomly sample a designated vertex for C; for $v \in C$, choose v as the designated vertex for C with probability $d_v / \sum_{u \in C} d_u$.

   (d) Construct the hyperedge e consisting of all the designated vertices. Add e to E, remove the chosen components $C_1, \ldots, C_k$, and create a new component $C = \cup_{j \in [k]} C_k$ with $w(C) = \sum_{i \in [k]} C_i$.

If, just before the hypergraph is fully connected, the chosen size k is greater than the number of components c, we generate the last hyperedge of the connected skeleton by connecting the final c components as per step 3 (with k replaced by c) and sampling the remaining $k - c$ vertices as per the usual Chung-Lu sampling technique, i.e., using **Algorithm 1**. We note that, other than potentially the last hyperedge constructed, a hyperedge constructed in step 3 is equivalent to a hyperedge generated by **Algorithm 1** conditioned on this hyperedge

joining k distinct components. We use this observation to simplify **Algorithm 5**. We will simplify **Algorithm 5** by writing "update [collection of components]" after generating a hyperedge.

Algorithm 5 Connected skeleton.

Require: $(d_1, \dots, d_n)$, $(m_{k_{\min}}, \dots, m_{k_{\max}})$

Initialize hyperedge list $E = \{\}$, a random hyperedge-size list S as per **Algorithm 4**, and a collection of components $\mathcal{C} = \big\{C_v := \{v\} \big| v \in [n]\big\}$.
for $k \in S$ **do**
  if $k \leq |\mathcal{C}|$ **then**
    repeat
      Sample $e \sim$ **Algorithm 1**$\Big((d_1, \dots, d_n), k\Big)$.
    until $\big|e \cap C\big| \leq 1$ for all $C \in \mathcal{C}$
    Set $E = E \cup e$ and update $\mathcal{C}$.
  else
    Set $c = |\mathcal{C}|$.
    repeat
      Sample $e' \sim$ **Algorithm 1**$\Big((d_1, \dots, d_n), c\Big)$.
    until $\big|e' \cap C\big| \leq 1$ for all $C \in \mathcal{C}$
    Sample $e'' \sim$ **Algorithm 1**$\Big((d_1, \dots, d_n), k - c\Big)$.
    Set $E = E \cup \{e' \cup e''\}$ and update $\mathcal{C}$.
  end if
  if $|\mathcal{C}| = 1$ **then**
    Return E
  end if
end for
Return E

Once we generate a connected skeleton via **Algorithm 5**, we then update the parameter $(m_{k_{\min}}, \dots, m_{k_{\max}})$ (by subtracting, from m_k, the number of hyperedges of size k that were generated for each k) and generate the rest of the simplicial Chung-Lu hypergraph via **Algorithm 4** with updated parameter $(m_{k_{\min}}, \dots, m_{k_{\max}})$ and initial (non-empty) hyperedge list E.