Subgraphs in non-uniform random hypergraphs

Megan Dewar[1], John Healy[1], Xavier Pérez-Giménez[2], Paweł Prałat[2], John Proos[1], Benjamin Reiniger[2], and Kirill Ternovsky[2]

[1] The Tutte Institute for Mathematics and Computing, Ottawa, ON, Canada,
[2] Department of Mathematics, Ryerson University, Toronto, ON, Canada

Abstract. Myriad problems can be described in hypergraph terms. However, the theory and tools are not sufficiently developed to allow most problems to be tackled directly within this context. The main purpose of this paper is to increase the awareness of this important gap and to encourage the development of this formal theory, in conjunction with the consideration of concrete applications. As a starting point, we concentrate on the problem of finding (small) subhypergraphs in a (large) hypergraph. Many existing algorithms reduce this problem to the known territory of graph theory by considering the 2-section graph. We argue that this is not the right approach, neither from a theoretical point of view (by considering a generalization of the classic model of binomial random graphs to hypergraphs) nor from a practical one (by performing experiments on two datasets).

Keywords: random graphs, random hypergraphs, subgraphs, subhypergraphs

1 Introduction

Myriad problems can be described in hypergraph terms. However, the theory and tools are not sufficiently developed to allow most problems to be tackled directly within this context. Hypergraphs are of particular interest in the field of knowledge discovery, where most problems currently modelled as graphs would be more accurately modelled as hypergraphs. Researchers in the knowledge discovery field are particularly interested in the generalization of the concepts of modularity and diffusion to hypergraphs. Such generalizations require a firm theoretical basis on which to develop these concepts. Unfortunately, although hypergraphs were formally defined in the 1960s (and various realizations of hypergraphs were studied long before that), the general formal theory is not as mature as required for the applications of interest to many industry partners or governments. The main purpose of this paper is to increase the awareness of this important gap and to encourage the development of this formal theory, in conjunction with the consideration of concrete applications.

In order to illustrate the issue, let us consider the following "toy example." Consider the coauthorship hypergraph in which vertices correspond to researchers and each hyperedge consists of the set of authors listed on a scientific paper. We have two goals for this dataset. As a first goal, we would like

to determine the Erdős number of every researcher (zero for Erdős, one for coauthors of Erdős, two for coauthors of coauthors of Erdős, etc.). Our second goal is to find a minimum set of authors who between them cover all the papers in the subhypergraph consisting only of the seminal papers in a particular field.

Often even though a dataset is naturally represented as a hypergraph we do not work directly on the hypergraph. Instead we reduce the hypergraph to its 2-section graph (the 2-section graph of a hypergraph is obtained by making each hyperedge a clique; see Section 2 for a formal definition) or a weighted version of the 2-section. Taking a 2-section of a hypergraph loses some of the information about hyperedges of size greater than 2. Sometimes losing this information does not affect our ability to answer the questions of interest. For example, the Erdős number of an author is the minimum distance between the author's vertex and Erdős' vertex in the hypergraph and this distance is not changed by taking the 2-section. Other times the information lost when taking the 2-section prevents us answering the question of interest. This is the case for our second goal of finding a minimum set of authors that cover a set of papers. In the hypergraph this goal means finding a minimum set of vertices that are incident with every hyperedge of interest. However, taking the 2-section of the hypergraph loses the information about the set of papers that a particular author covers. In fact, the 2-section does not even retain how many papers exist. Basically, if the composition of the hyperedges of size greater than 2 is important in solving a problem than solving the problem in the 2-section is going to be difficult or impossible.

Besides the information loss, there is another potential downside to working with the 2-section of a hypergraph. Namely, that the 2-section can be much denser than the hypergraph since a single hyperedge of size k implies $\binom{k}{2}$ edges in the 2-section. Depending on the dataset and algorithm being executed the increased density of the 2-section can have a significant detrimental affect on the runtime.

In this paper, we shall be interested in finding subhypergraphs in hypergraphs. While the composition of the hyperedges of size greater than 2 matters when answering this question, it is natural to ask whether 2-section graphs can be used to help answer the question. That is, when determining whether or not a hypergraph H contains H_1 as a subhypergraph, is it useful to look for G_{H_1}, the 2-section of H_1, in G_H, the 2-section of H? Clearly there are many ways that G_{H_1} could appear in G_H without H_1 appearing in H. An obvious technique would be to use the existing graph theoretical tools to find all copies of G_{H_1} in G_H and then simply inspect them, one by one, in the original hypergraph. So perhaps reducing the hypergraph to its 2-section can be used to solve the problem. Maybe in most networks that are considered in practice, any two subhypergraphs inducing the same graph in the 2-section occur with the same probability? This would be desirable, as it would mean that the above technique does not waste a lot of time dealing with subhypergraphs that we are not interested in finding. Of course, even if the 2-section can be used in this way for finding subhypergraphs, the increased density of the 2-section may lead to the graph theoretical tools used being quite inefficient.

In order to deal with the question of the false positive rate of G_{H_1} in G_H, we introduce a natural generalization of Erdős-Rényi (binomial) random graphs to non-uniform random hypergraphs. We study (rigorously, via theorems with proofs) occurrences of a given hypergraph in the random hypergraph. One of the implications of our work is that two hypergraphs H_1, H_2 that induce the same subgraph in the 2-section can have drastically different thresholds for appearance. This suggests that the answer to the latest question is "no," and that we have lost something by considering only the 2-section. Assuming that hyperedges in the network we try to analyze occur randomly, our theorems imply that there might be very few (if any) copies of H_1 (the hypergraph we are looking for in the network) but plenty of copies of H_2 (the hypergraph we do not care about). So the algorithm discovers a lot of potential candidates but none of them is what we are looking for!

We investigate two real-world networks: an email hypergraph and the coauthorship hypergraph that was already mentioned. Not surprisingly, we confirm that hypergraphs that are not distinguishable in the 2-section graph occur with different probabilities (as predicted by the model). Hence we feel that using existing graph algorithms on the 2-section can be and often is lacking and that the research community needs to develop more algorithms that deal with hypergraphs directly.

While non-uniform random hypergraphs might serve as the very first model of the real-world hypergraphs, the assumption that events that occur in the network are independent is likely not reasonable. Perhaps of particular importance is a notion of clustering coefficient; there have been a number of proposals for generalizing clustering coefficient from graphs to hypergraphs, for instance [1, 5, 11, 14, 15]. In the longer journal version of this paper we compute the hypergraph clustering coefficient of [15] for our random hypergraph model and the two real networks we are investigating. With the knowledge and experience we gathered, we feel that we are better prepared to propose a probabilistic model that is more suitable. However, it is left for the forthcoming papers.

Due to space limitation, all proofs and details of a set of experiments we performed in this project are omitted in this proceedings version but will be included in the journal version of this paper.

2 Definitions and Conventions

2.1 Random graphs and random hypergraphs

First, let us recall a classic model of random graphs. The *binomial random graph* $\mathscr{G}(n,p)$ is the random graph G with vertex set $[n] := \{1, 2, \ldots, n\}$ in which every pair $\{i,j\} \in \binom{[n]}{2}$ appears independently as an edge in G with probability p. Note that $p = p(n)$ may (and usually does) tend to zero as n tends to infinity.

In this paper, we are concerned with more general combinatorial objects: hypergraphs. A *hypergraph* H is an ordered pair $H = (V, E)$, where V is a finite set (the *vertex set*) and E is a family of distinct subsets of V (the *hyperedge*

set). A hypergraph $H = (V, E)$ is *r-uniform* if all hyperedges of H are of size r. For a given $r \in \mathbb{N}$, the *random r-uniform hypergraph* $\mathscr{H}_r(n, p)$ has n labelled vertices from a vertex set $V = [n]$, in which every subset $e \subseteq V$ of size $|e| = r$ is chosen to be a hyperedge of H randomly and independently with probability p. For $r = 2$, this model reduces to the model $\mathscr{G}(n, p)$.

The binomial random graph model is well known and thoroughly studied (e.g. [3, 12, 10]). Random hypergraphs are much less understood and, unfortunately, most of the existing papers deal with uniform hypergraphs. For example, Hamilton cycles (both tight ones and loose ones) were recently studied in [7–9]; perfect matchings were investigated in [13] (for a few more examples see the recent book on random graphs [10]).

In this paper, we are concerned with a natural generalization of this model that produces non-uniform hypergraphs. Let $\boldsymbol{p} = (p_r)_{r \geq 1}$ be any sequence of numbers such that $0 \leq p_r = p_r(n) \leq 1$ for each $r \geq 1$. The *random hypergraph* $\mathscr{H}(n, \boldsymbol{p})$ has n labelled vertices from a vertex set $V = [n]$, in which every subset $e \subseteq V$ of size $|e| = r$ is chosen to be a hyperedge of H randomly and independently with probability p_r. In other words, $\mathscr{H}(n, \boldsymbol{p}) = \bigcup_{r \geq 1} \mathscr{H}_r(n, p_r)$ is a union of independent uniform hypergraphs.

Let us mention that there are several natural generalizations that might be worth exploring, depending on a specific application in mind. One possible generalization would be to allow hyperedges to contain repeated vertices (multiset-hyperedge hypergraphs). Another one would be to allow the hyperedges to be chosen with possible repetitions, to get parallel hyperedges.

A vertex of a hypergraph is *isolated* if it is contained in no edge. (In particular, a vertex of degree 1 that belongs only to an edge of size 1 is not isolated.) The *2-section* of a hypergraph H, denoted $[H]_2$, is the graph on the same vertex set as H and an edge uv if (and only if) u and v are contained in some edge of H. In other words, it is obtained by making each hyperedge of H a clique in $[H]_2$.

2.2 Subgraphs

In this paper, we are concerned with occurrences of a given substructure in hypergraphs. However, there are at least two natural generalizations of "subgraph" for hypergraphs.

A hypergraph $H' = (V', E')$ is a *strong subhypergraph* (called *hypersubgraph* by Bahmanian and Sajna [2] and *partial hypergraph* by Duchet [6]) of $H = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$; that is, each hyperedge of H' is also an hyperedge of H. We write $H' \subseteq_s H$ when H' is a strong subhypergraph of H. For $H = (V, E)$ and $V' \subseteq V$, the *strong subhypergraph of H induced by V'*, denoted $H_s[V']$, has vertex set V' and hyperedge set $E' = \{e \in E : e \subseteq V'\}$.

The hypergraph H' is a *weak subhypergraph* of H (called *subhypergraph* by Bahmanian and Sajna) if $V' \subseteq V$ and $E' \subseteq \{e \cap V' : e \in E\}$; that is, each hyperedge of H' can be extended to one of H by adding vertices of $V \setminus V'$ to it. For $V' \subseteq V$, the *weak subhypergraph induced by V'*, denoted $H_w[V']$, has vertex set V' and hyperedge set $E' = \{e \cap V' : e \in E\}$. For this paper however, since we desire our hypergraphs to never contain the empty hyperedge, we tacitly replace

E' by $E' \setminus \{\emptyset\}$. For now, weak subgraphs are assumed not to have multiple hyperedges (E' is a set, not a multiset).

Note that when G is an ordinary (i.e. 2-uniform) graph, strong subhypergraphs are the usual notion of subgraph, and weak subhypergraphs are subgraphs together with possible hyperedges of size 1. Note that each strong subhypergraph is also a weak subhypergraph but not vice versa.

Given hypergraphs H_1 and H_2, a weak (resp. strong) *copy* of H_1 in H_2 is a weak (resp. strong) subhypergraph of H_2 isomorphic to H_1. Most of this paper is concerned with determining the existence of strong or weak copies of a fixed H in $\mathscr{H}(n, \boldsymbol{p})$. With a mild abuse of terminology, we will often say that $\mathscr{H}$ contains H as a weak (strong) subhypergraph when we actually mean that $\mathscr{H}$ contains a weak (strong) copy of H. The precise meaning will always be clear from the context.

Fig. 1. The hypergraph H_1 appears as a weak subhypergraph of H_2 (induced by the dashed vertex subset), but not as a strong subhypergraph.

3 Small subgraphs in $\mathscr{H}(n, p)$.

We are interested in answering questions about the existence of subgraphs within $\mathscr{H}(n, \boldsymbol{p})$. This question was addressed for $\mathscr{G}(n, p)$ by Bollobás in [4]. We are going to generalize his result to hypergraphs but first we we need a few more definitions. Let $H = (V, E)$ be a hypergraph. Denote by $v(H) = |V|$ and by $e(H) = |E|$ the number of vertices and edges of H, respectively. For any $r \geq 1$, we will use $e_r(H) = |\{e \in E : |e| = r\}|$ to denote the number of edges of H of size r.

Define

$$\mu_s(H) = n^{v(H)} \prod_{r \geq 1} p_r^{e_r(H)}. \tag{1}$$

Now we are ready to state our result for the appearance of strong subgraphs of $\mathscr{H}(n, \boldsymbol{p})$. We adopt the convention that $0^0 = 1$ and assume all our hypergraphs have nonempty vertex set.

Theorem 1. *Let H be an arbitrary fixed hypergraph. Let $\boldsymbol{p} = (p_r)_{r \geq 1}$ be any sequence such that $0 \leq p_r = p_r(n) \leq 1$ for each $r \geq 1$. Let $\mathcal{J}$ denote the family of all strong subgraphs of H.*

(a) *If for some* $H' \in \mathcal{J}$ *we have* $\mu_s(H') \to 0$ *(as* $n \to \infty$*), then a.a.s.* $\mathscr{H}(n, \boldsymbol{p})$ *does not contain* H *as a strong subgraph.*
(b) *If for all* $H' \in \mathcal{J}$ *we have* $\mu_s(H') \to \infty$ *(as* $n \to \infty$*), then a.a.s.* $\mathscr{H}(n, \boldsymbol{p})$ *contains* H *as a strong subgraph.*

Let us mention that the result also holds for the multiset setting: that is, when vertices are allowed to be repeated in each hyperedge with some multiplicity. Moreover, if additionally there exists $\varepsilon > 0$ such that $p_r \leq 1 - \varepsilon$, for all r, then the same conditions (that is, conditions (a) and (b) of Theorem 1) determine whether or not H appears as an *induced* strong subgraph.

In view of Theorem 1, we emphasize that the existence of strong copies of H in $\mathscr{H}(n, \boldsymbol{p})$ cannot be determined by translating to graphs via 2-sections. For instance, consider the three hypergraphs H_1, H_2 and H_3 from Figure 2. Each of these has H_1 as its 2-section. However, the expected number of strong copies of H_1, H_2 and H_3 in $\mathscr{H}(n, \boldsymbol{p})$ is, respectively, of order $n^4 p_2^5$, $n^4 p_2^2 p_3$, and $n^4 p_3^2$. So if, say, $p_3 = n^{-5/2}$ and $p_2 = n^{-3/4}$, then we expect many copies of H_1, a constant number of copies of H_2, and $o(1)$ copies of H_3. Moreover, by testing the conditions of Theorem 1 for all the strong subgraphs of H_1, H_2, H_3, we obtain that a.a.s. $\mathscr{H}(n, \boldsymbol{p})$ contains H_1 but not H_3 as a strong subgraph (and the theorem is inconclusive for H_2).

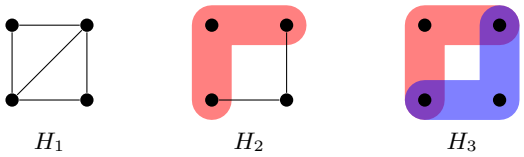

Fig. 2. These three hypergraphs have the same 2-section, which is precisely H_1, but their behaviour as potential strong subgraphs of $\mathscr{H}(n, \boldsymbol{p})$ is different.

Now we move to our result for the appearance of weak subgraphs of $\mathscr{H}(n, \boldsymbol{p})$. For technical reasons, we restrict ourselves to hypergraphs with bounded edge sizes. Formally, for a given $M \in \mathbb{N}$, we say that $H = (V, E)$ is an *M-bounded hypergraph* if $|e| \leq M$ for all $e \in E$. Similarly, $\boldsymbol{p} = (p_r)_{r \geq 1}$ is an *M-bounded sequence* if $p_r = 0$ for $r > M$. We will use $\boldsymbol{p} = (p_r)_{r=1}^{M}$ for an M-bounded sequence instead of an infinite sequence $\boldsymbol{p} = (p_r)_{r \geq 1}$ with a bounded number of non-zero values. Clearly, if $\boldsymbol{p}$ is M-bounded, then so is $\mathscr{H}(n, \boldsymbol{p})$ (with probability 1). For $r \in [M]$, let

$$p'_r = p_r + np_{r+1} + n^2 p_{r+2} + \cdots + n^{M-r} p_M, \tag{2}$$

and, given any fixed hypergraph H, define

$$\mu_w(H) = n^{v(H)} \prod_{r=1}^{M} (p'_r)^{e_r(H)}, \tag{3}$$

which will play an analogous role to $\mu_s(H)$.

Theorem 2. *Let H be an arbitrary fixed hypergraph, and let $\mathcal{J}$ be the collection of all strong subgraphs of H. Let $\boldsymbol{p} = (p_r)_{r=1}^{M}$ be an M-bounded sequence.*

(a) If for some $H' \in \mathcal{J}$ we have $\mu_w(H') \to 0$ (as $n \to \infty$), then a.a.s. $\mathscr{H}(n, \boldsymbol{p})$ does not contain H as a weak subgraph.

(b) If for all $H' \in \mathcal{J}$ we have $\mu_w(H') \to \infty$ (as $n \to \infty$), then a.a.s. $\mathscr{H}(n, \boldsymbol{p})$ contains H as a weak subgraph.

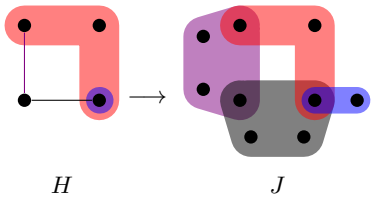

Fig. 3. A hypergraph J and an induced weak hypergraph H with different thresholds for appearance as strong subgraphs.

We shall discuss a few relevant points concerning Theorem 2. First, it is possible that a.a.s. some graph occurs as a weak subgraph but not as a strong one. For example, if

$$p_1 = n^{-0.6}, \quad p_2 = n^{-0.9}, \quad p_3 = n^{-1.7}, \quad \text{and} \quad p_4 = n^{-3.1}, \tag{4}$$

then a.a.s. $\mathscr{H}(n, \boldsymbol{p})$ does not contain graph H (presented on Figure 3) as a strong subgraph but a.a.s. it contains J (also presented on Figure 3) and so a.a.s. it contains H as a weak subgraph.

Next, observe that if we replace $\mathcal{J}$ in the statement of Theorem 2 by the collection $\mathcal{J}_w$ of all *weak* subgraphs of H, the theorem remains valid. This is trivially true for part (b), since $\mathcal{J}_w \supseteq \mathcal{J}$. For part (a), a few easy modifications in the proof are necessary which will be mentioned in the journal version of this paper.

Finally, let us comment on the definition of p'_r, and introduce related parameters p''_r and p'''_r, which will play a role later on. Our particular choice of p'_r in (3) and thus in the statement of Theorem 2 is the simplest function from the equivalence class of all functions of the same order. However, the following one is more natural (as argued below). For $r \in [M]$, let

$$p''_r = p_r + np_{r+1} + \binom{n}{2} p_{r+2} + \cdots + \binom{n}{M-r} p_M. \tag{5}$$

Note that p'_r and p''_r are of the same order. More precisely,

$$(1 + o(1)) \frac{p'_r}{(M-r)!} \leq p''_r \leq p'_r.$$

Hence, p'_r can be replaced in (3) by the more natural (but less simple) p''_r, and Theorem 2 remains valid. It is worth noting that both p'_r and p''_r can be greater than one or even tend to infinity as $n \to \infty$. Indeed, p''_r is not a probability but rather is asymptotic to the expected number of edges to which a given set of size r belongs. In contrast, the probability that such a set belongs to some edge is

$$p'''_r = 1 - (1-p_r)(1-p_{r+1})^{n-r}(1-p_{r+2})^{\binom{n-r}{2}} \cdots (1-p_M)^{\binom{n-r}{M-r}}. \tag{6}$$

Observe that, if $p'_r = o(1)$ (or equivalently $p''_r = o(1)$), then $p''_r, p''_{r+1}, \ldots, p''_M = o(1)$, and therefore

$$\begin{aligned} p'''_r &= 1 - \exp\left(-(1+o(1))\left(p_r + np_{r+1} + \binom{n}{2}p_{r+2} + \cdots + \binom{n}{M-r}p_M\right)\right) \\ &= 1 - \exp\left(-(1+o(1))p''_r\right) \quad \sim \quad p''_r, \end{aligned} \tag{7}$$

so p''_r and p'''_r asymptotically coincide.

4 Induced weak subgraphs

Let us discuss how one can use Theorem 2 to determine whether H appears as an *induced* weak subgraph of $\mathscr{H}(n, \boldsymbol{p})$. This seems to be more complex than in the case of strong subgraphs:the non-edges of H play a crucial role in determining the existence of induced weak copies. Indeed, a weak subgraph H of $\mathscr{H}(n, \boldsymbol{p})$ is induced provided that, for every set e of vertices of H that do not form an edge, e cannot be extended to an edge of $\mathscr{H}(n, \boldsymbol{p})$ by adding vertices not in H.

First, we will give some conditions that forbid a.a.s. the existence of weak induced copies of H in $\mathscr{H}(n, \boldsymbol{p})$ (even if H *does* appear as a weak subgraph).

Proposition 1. *Let H be an arbitrary fixed hypergraph on k vertices with a non-edge of size r ($1 \le r \le k$). Suppose $p''_r \ge (k+\varepsilon)\log n$ for some constant $\varepsilon > 0$. Then, a.a.s. H does not occur as an induced weak subgraph of $\mathscr{H}(n, \boldsymbol{p})$.*

As a result, the condition $p''_r \ge (k+\varepsilon)\log n$ implies that, if H is an induced weak subgraph of $\mathscr{H}(n, \boldsymbol{p})$ of order k, then H must contain all possible edges of size r. Coming back to our example with H from Figure 3 and p_i's from (4), note that $p''_1 \sim \binom{n}{2}p_3 \sim n^{0.3}/2$. Thus, a.a.s. H will not occur as an induced weak subgraph of $\mathscr{H}(n, \boldsymbol{p})$, as not every vertex of H belongs to an edge of size 1.

On the other hand, suppose that $r \ge 1$ is the size of the smallest non-edge of H and assume that

$$\max\{p'''_r, p'''_{r+1}, \ldots, p'''_M\} \le 1 - \varepsilon \tag{8}$$

for some constant $\varepsilon > 0$. Then any given weak copy of H in $\mathscr{H}(n, \boldsymbol{p})$ is also induced with probability bounded away from zero. In that case, the same calculations in the proof of Theorem 2 (that is omitted in this version) are still valid with an extra $\Theta(1)$ factor, and thus the conclusions of that theorem extend to induced weak subgraphs. Since verifying condition (8) may sometimes be slightly unwieldy, we will give a simpler sufficient condition.

Proposition 2. *Let H be an arbitrary fixed hypergraph, and let r be the size of its smallest non-edge. Suppose that $p_r \leq 1 - \varepsilon$ for some constant $\varepsilon > 0$ and that $p'_r = O(1)$ (and, as a result, $p''_r = O(1)$ too). If the conditions in part (b) of Theorem 2 are satisfied, then a.a.s. $\mathscr{H}(n, \boldsymbol{p})$ contains H as an induced weak subgraph.*

Let us come back to our example from Figure 3 and (4) for the last time. Note that $p''_2 \sim np_3 = n^{-0.7} = o(1)$. Hence, if the "missing" edges of size 1 are added to H, then a.a.s. the resulting graph would occur as an induced weak subgraph of $\mathscr{H}(n, \boldsymbol{p})$.

5 The 2-section of $\mathscr{H}(n, \boldsymbol{p})$

We first consider the question of whether a given (2-uniform) graph G appears as a subgraph of the 2-section of $\mathscr{H}(n, \boldsymbol{p})$. We again may assume that G has no isolated vertices.

Let us start with some general observations that apply for any host hypergraph $\mathscr{H}$, not necessarily $\mathscr{H}(n, \boldsymbol{p})$. Observe that $G \subseteq [\mathscr{H}]_2$ if and only if there is a weak subhypergraph H of $\mathscr{H}$ such that G is a spanning subgraph of $[H]_2$. So we may test for $G \subseteq [\mathscr{H}]_2$ by finding every hypergraph H with G a spanning subgraph of $[H]_2$ and applying Theorem 2 to each. We can reduce the number of hypergraphs that need to be tested: if H_1 is a weak subhypergraph of H_2 and H_2 is a weak subhypergraph of $\mathscr{H}$, then H_1 is also a weak subhypergraph of $\mathscr{H}$. Note too that a spanning weak subhypergraph is actually a strong subhypergraph. So it suffices to check only the hypergraphs H that are minimal—with respect to the (strong) subhypergraph relation—that have G as a spanning subgraph of their 2-section.

In $\mathscr{H}(n, \boldsymbol{p})$, one can reduce the number of hypergraphs H to be tested even further. A *subedge system* of a hypergraph H is a hypergraph obtained from H by taking a subset of each edge of H and taking a (strong) subhypergraph of the result. Let H_1 be a subedge system of H_2 and let H_2 be a weak subhypergraph of H. It is not necessarily true that H_1 is a weak subhypergraph of H, but it is true a.a.s. for $H = \mathscr{H}(n, \boldsymbol{p})$.

Proposition 3. *Let H_1 and H_2 be fixed hypergraphs with H_1 a spanning subedge system of H_2, and let $\boldsymbol{p}$ be M-bounded. Let $\mathcal{J}_1$ and $\mathcal{J}_2$ denote the set of all strong subgraphs of H_1 and H_2, respectively. If every $H'_2 \in \mathcal{J}_2$ satisfies $\mu_w(H'_2) \to \infty$, then every $H'_1 \in \mathcal{J}_1$ also satisfies $\mu_w(H'_1) \to \infty$.*

Corollary 1. *Fix a (2-uniform) graph G without isolated vertices. Let $\mathcal{F}$ denote the family of minimal—with respect to the subedge system relation—hypergraphs containing G in their 2-section. Let $\boldsymbol{p}$ be M-bounded.*

(a) If for every $H \in \mathcal{F}$ there is some strong subgraph $H' \subseteq_s H$ with $\mu_w(H') \to 0$, then a.a.s. G is not a subgraph of $[\mathscr{H}(n, \boldsymbol{p})]_2$.
(b) If for some $H \in \mathcal{F}$ every strong subgraph $H' \subseteq_s H$ satisfies $\mu_w(H') \to \infty$, then a.a.s. G is a subgraph of $[\mathscr{H}(n, \boldsymbol{p})]_2$.

We next consider the following problem. Suppose that a copy of G is found in $[\mathscr{H}(n,\boldsymbol{p})]_2$. We would like to estimate the probability that this copy comes from a given weak subhypergraph of $\mathscr{H}(n,\boldsymbol{p})$.

Let G be a fixed 2-uniform graph with no isolated vertices. Let $\mathcal{F}$ denote the family of hypergraphs H on the same vertex set as G such that $G \simeq [H]_2$. Then, G appears as an induced subgraph of $[\mathscr{H}(n,\boldsymbol{p})]_2$ if and only if some $H \in \mathcal{F}$ appears as an induced weak subhypergraph of $\mathscr{H}(n,\boldsymbol{p})$. More precisely, for every set of vertices S inducing a copy of G in $[\mathscr{H}(n,\boldsymbol{p})]_2$, there is exactly one $H \in \mathcal{F}$ such that S induces a weak copy of H in $\mathscr{H}(n,\boldsymbol{p})$. We say in that case that hypergraph H *originates* that particular copy of G. As a result we have the following proposition.

Proposition 4. *Let $\boldsymbol{p} = (p_r)_{r=1}^M$ be an M-bounded sequence. For $r \in [M]$, let p_r''' be defined as in* (6)*. Then, given a copy of G in $[\mathscr{H}(n,\boldsymbol{p})]_2$, the probability that it originates from a given $H \in \mathcal{F}$ is*

$$(1+o(1))\frac{\operatorname{aut}(H)\prod_{r=1}^M (p_r''')^{e_r(H)}(1-p_r''')^{\binom{v(G)}{r}-e_r(H)}}{\sum_{H'\in\mathcal{F}}\operatorname{aut}(H')\prod_{r=1}^M (p_r''')^{e_r(H')}(1-p_r''')^{\binom{v(G)}{r}-e_r(H')}}.$$

Instead of determining which specific $H \in \mathcal{F}$ originates a copy of G in the 2-section of $\mathscr{H}(n,\boldsymbol{p})$, we may take equivalence classes in $\mathcal{F}$ given their r-edge counts. To that end, define the *signature* of $H \in \mathcal{F}$ as the vector $\boldsymbol{e}(H) = (e_1(H), e_2(H), \ldots, e_k(H))$, where $k = v(G)$ (and hence also $k = v(H)$). Let $\boldsymbol{e}(\mathcal{F}) = \{\boldsymbol{e}(H) : H \in \mathcal{F}\}$. For a given signature $\boldsymbol{e} \in \boldsymbol{e}(\mathcal{F})$, let $\mathcal{F}_{\boldsymbol{e}} \subseteq \mathcal{F}$ be the family of hypergraphs in $\mathcal{F}$ with signature $\boldsymbol{e}$. Notice that $\{\mathcal{F}_{\boldsymbol{e}} : \boldsymbol{e} \in \boldsymbol{e}(\mathcal{F})\}$ is a partition of $\mathcal{F}$. Then, the following useful result holds.

Corollary 2. *Let $\boldsymbol{p} = (p_r)_{r=1}^M$ be an M-bounded sequence. For $r \in [M]$, let p_r''' be defined as in* (6)*. Then, given a copy of G in $[\mathscr{H}(n,\boldsymbol{p})]_2$, the probability that it originates from a hypergraph with a given signature $\boldsymbol{e} = (m_1, m_2, \ldots, m_k) \in \boldsymbol{e}(\mathcal{F})$ is*

$$(1+o(1))\frac{\sum_{H\in\mathcal{F}_{\boldsymbol{e}}}\operatorname{aut}(H)\prod_{r=1}^k (p_r''')^{m_r}(1-p_r''')^{\binom{v(G)}{r}-m_r}}{\sum_{H'\in\mathcal{F}}\operatorname{aut}(H')\prod_{r=1}^k (p_r''')^{e_r(H')}(1-p_r''')^{\binom{v(G)}{r}-e_r(H')}}.$$

The following example illustrates how, under natural assumptions on $\boldsymbol{p}$, Corollary 2 implies that a copy of G in $[\mathscr{H}(n,\boldsymbol{p})]_2$ "typically" originates from a hypergraph $H \in \mathcal{F}$ with few but large edges rather than many but small edges. Let $G = K_k$ (i.e. the clique of order k) for a fixed $k \geq 2$, and suppose that $\boldsymbol{p}$ is an M-bounded sequence satisfying $\binom{n}{j}p_j = O(n)$ for all $j \in [M]$. The latter condition is equivalent to assuming that the expected number of edges of each given size is at most linear in the number of vertices, which is a fairly reasonable assumption for many relevant models of hypergraph networks. Suppose additionally that for some r with $k \leq r \leq M$ we also have $\binom{n}{r}p_r = \Omega(n)$. From (7), we obtain that $p_j''' = O(1/n^{j-1})$ for every $j \in [M]$ and $p_k''' = \Theta(1/n^{k-1})$. Consider the signature $\widehat{\boldsymbol{e}} = (0, \ldots, 0, 1)$ corresponding to the hypergraph $\widehat{H}$ on k

vertices with one single edge of size k. A straightforward inductive argument reveals that, for any signature $\boldsymbol{e} = (m_1, m_2, \ldots, m_k) \in \boldsymbol{e}(\mathcal{F})$,

$$\prod_{r=1}^{k} (p_r''')^{m_r} (1 - p_r''')^{\binom{k}{r} - m_r} = \begin{cases} (1 + o(1)) p_k''' = \Theta(1/n^{k-1}) & \text{if } \boldsymbol{e} = \widehat{\boldsymbol{e}} \\ o(1/n^{k-1}) & \text{if } \boldsymbol{e} \neq \widehat{\boldsymbol{e}}. \end{cases}$$

As a result, applying Corollary 2 to all signatures different from $\widehat{\boldsymbol{e}}$, we conclude that, for a given copy of G in $[\mathscr{H}(n, \boldsymbol{p})]_2$, a.a.s. it must originate from $\widehat{H}$.

6 Experiments

We performed a number of experiments on two real-world datasets that are naturally represented as a hypergraph network. Our goal was to compare the results with the corresponding theoretical predictions. Due to space limitation, the details are omitted in this version but will be included in the journal version of this paper.

The experiments we performed confirmed the intuition that the fact that some set of vertices S forms a hyperedge should increase the probability that some proper subset of S belongs to some other hyperedge. Moreover, in many instances, the correlation seems to be so strong that not only having one hyperedge increases substantially the probability that another hyperedge intersects it but it is more likely that there will be another hyperedge intersecting it than not. Of course, such behaviour is not present in our theoretical model in which events are independent. In order to understand the behaviour we experience, some notion of "clustering coefficient" has to be introduced in the hypergraph setting. Again, the details are omitted here but will be included in the journal version of this paper.

7 Conclusions and future work

The goal of the larger project behind this paper is to propose a reasonable model for complex networks using hypergraphs, as they seem more suitable for many existing networks and associated applications. Whereas there are many models using graphs (classic ones such as $\mathscr{G}(n, p)$, random d-regular graphs, and PA model, as well as spatial ones such as random geometric graphs and SPA model), there are very few using hypergraphs. In order to better understand micro-processes that shape macro-properties that are observed in these networks, we introduced the random hypergraphs and investigated some properties of it in order to compare them with two real-world networks. These results are interesting from a pure random graph theory perspective but, of course, we did not expect such models to work well in practice; we did it to learn why they do *not* work. As is common in this field, such an exercise taught us a lot, and we feel that we are now better prepared to design a more suitable model, probably combining both geometry and the "rich get richer" paradigm. However, it is left for the forthcoming papers.

References

1. M. Alexander and G. Robins, Small Worlds among Interlocking Directors: Network Structures and Distance in Bipartite Graphs, *Computational & Mathematical Organization Theory*, 10(1), 69-94, 2004.
2. M. Bahmanian, M. Sajna, Connection and separation in hypergraphs, arXiv:1504.04274v1, 2015.
3. B. Bollobás, *Random Graphs*, Cambridge University Press, Cambridge, 2001.
4. B. Bollobás, Random graphs, in *Combinatorics* (ed. H.N.V. Temperley), London Mathematical Society Lecture Note Series, 52, Cambridge University Press, Cambridge (1981), pp. 80–102.
5. S. Borgatti and M. Everett, Network analysis of 2-mode data, *Social Networks*, 19(3), 243-269, 1997.
6. P. Duchet, Hypergraphs, in *Handbook of combinatorics* (ed. R.L. Graham, M. Grötschel, L. Lovász), Elsevier, Amsterdam, 1995.
7. A. Dudek and A.M. Frieze, Loose Hamilton cycles in random k-uniform hypergraphs, *Electronic Journal of Combinatorics* (2011) P48.
8. A. Dudek and A.M. Frieze, Tight Hamilton cycles in random uniform hypergraphs, *Random Structures and Algorithms* **42** (2012), 374–385.
9. A. Ferber, Closing gaps in problems related to Hamilton cycles in random graphs and hypergraphs, preprint.
10. A.M. Frieze and M. Karoński, *Introduction to Random Graphs*, Cambridge University Press, 2015.
11. JL. Guillaume and S. Le Blond and M. Latapy, Clustering in P2P exchanges and consequences on performance, In Lecture Notes in Computer Science (LNCS) vol. 3640, *Proceedings of the 4-th international workshop on Peer-to-Peer Systems (IPTPS)*, 2005, 193-204.
12. S. Janson, T. Łuczak, A. Ruciński, *Random Graphs*, Wiley, New York, 2000.
13. A. Johansson, J. Kahn, and V. Vu, Factor in random graphs, *Random Structures and Algorithms* **33** (2008), 1–28.
14. M. Latapy and C. Magnien and N. Del Vecchio, Basic notions for the Analysis of Large Two-mode Networks, *Social Networks*, 30(1), 31-48, 2008.
15. W. Zhou and L. Nakhleh, Properties of metabolic graphs: biological organization or representation artifacts?, *BMC Bioinformatics* 2011, 12:132.