Hypernetwork Science: From Multidimensional Networks to Computational Topology

Cliff A. Joslyn, Sinan G. Aksoy, Tiffany J. Callahan, Lawrence E. Hunter, Brett Jefferson, Brenda Praggastis, Emilie Purvine, and Ignacio J. Tripodi

Abstract As data structures and mathematical objects used for complex systems modeling, hypergraphs sit nicely poised between on the one hand the world of network models, and on the other that of higher-order mathematical abstractions from algebra, lattice theory, and topology. They are able to represent complex systems interactions more faithfully than graphs and networks, while also being some of the simplest classes of systems representing topological structures as collections of multidimensional objects connected in a particular pattern. In this paper we discuss the role of (undirected) hypergraphs in the science of complex networks, and provide a mathematical overview of the core concepts needed for hypernetwork modeling, including duality and the relationship to bicolored graphs, quantitative adjacency and incidence, the nature of walks in hypergraphs, and available topological relationships and properties. We close with a brief discussion of two example applications: biomedical databases for disease analysis, and domain-name system (DNS) analysis of cyber data.

1 Hypergraphs for Complex Systems Modeling

In the study of complex systems, graph theory has been the mathematical scaffold underlying network science [4]. A graph $\mathcal{G} = \langle V, E \rangle$ comprises a set V of vertices connected in a set $E \subseteq \binom{V}{2}$ of edges (where $\binom{V}{2}$ here means all unordered pairs of $v \in V$), where each edge $e \in E$ is a pair of distinct vertices. Systems studied in biology, sociology, telecommunications, and physical infrastructure often afford a representation as such a set of entities with binary relationships, and hence may be analyzed utilizing graph theoretical methods.

Graph models benefit from simplicity and a degree of universality. But as abstract mathematical objects, graphs are limited to representing *pairwise* relationships

C. A. Joslyn (✉) · S. G. Aksoy · B. Jefferson · B. Praggastis · E. Purvine
Pacific Northwest National Laboratory, Seattle, WA, USA
e-mail: Cliff.Joslyn@pnnl.gov

T. J. Callahan · L. E. Hunter · I. J. Tripodi
University of Colorado Anschutz Medical Campus, Denver, CO, USA

D. Braha et al. (eds.), *Unifying Themes in Complex Systems X*, Springer Proceedings in Complexity, https://doi.org/10.1007/978-3-030-67318-5_25

between entities, while real-world phenomena in these systems can be rich in *multi-way* relationships involving interactions among more than two entities, dependencies between more than two variables, or properties of collections of more than two objects. Representing *group* interactions is not possible in graphs natively, but rather requires either more complex mathematical objects, or coding schemes like "reification" or semantic labeling in bipartite graphs. Lacking multi-dimensional relations, it is hard to address questions of "community interaction" in graphs: e.g., how is a collection of entities A connected to another collection B through chains of other communities?; where does a particular community stand in relation to other communities in its neighborhood?

The mathematical object which *natively* represents multi-way interactions in networks is called a "hypergraph" [6].[1] In contrast to a graph, in a hypergraph $\mathcal{H} = \langle V, E \rangle$ those same vertices are now connected generally in a family E of hyperedges, where now a hyperedge $e \in E$ is an arbitrary subset $e \subseteq V$ of k vertices, thereby representing a k-way relationship for any integer $k > 0$. Hypergraphs are thus the natural representation of a broad range of systems, including those with the kinds of multi-way relationships mentioned above. Indeed, hypergraph-structured data (i.e. hypernetworks) are ubiquitous, occurring whenever information presents naturally as set-valued, tabular, or bipartite data.

Hypergraph models are definitely more complicated than graphs, but the price paid allows for higher fidelity representation of data which may contain multi-way relationships. An example from bibliometrics is shown in Fig. 1.[2] On the upper left is a table showing a selection of five papers co-authored by different collections of four authors. Its hypergraph is shown in the lower left in the form of an "Euler diagram", with hyperedges as colored bounds around groups of vertices. A typical approach to these same data would be to reduce the collaborative structure to its so-called "2-section": a graph of all and only the two-way interactions present, whether explicitly listed (like paper 1) or implied in virtue of larger collaborations (as for paper 2). That co-authorship graph is shown in the lower right, represented by its adjacency matrix in the upper right. It can be seen that the reduced graph form necessarily loses a great deal of information, ignoring information about the single-authored paper (3) and the three-authored paper (2).

Hypergraphs are closely related to important objects in discrete mathematics used in data science such as bipartite graphs, set systems, partial orders, finite topologies, abstract simplicial complexes, and especially graphs proper, which they explicitly generalize: every graph is a 2-uniform hyergraph, so that $|e| = k = 2$ for all hyperedges $e \in E$. Thus they support a wider range of mathematical methods, such as those from

[1] Throughout this paper we will deal only with "basic" hypergraphs in the sense of being undirected, unordered, and unlabeled. All of these forms are available and important [3, 16].

[2] Hypergraph calculations shown in this paper were produced using PNNL's open source hypergraph analytical capabilities HyperNetX (HNX, https://github.com/pnnl/HyperNetX) and the Chapel Hypergraph Library (CHGL, https://github.com/pnnl/chgl); and additionally diagrams were produced in HNX.

Paper #	Authors
1	Andrews, Davis
2	Andrews, Carter, Davis
3	Davis
4	Andrews, Bailey
5	Bailey, Carter

	Andrews	Bailey	Carter	Davis
Andrews		Y	Y	Y
Bailey	Y		Y	
Carter	Y	Y		Y
Davis	Y		Y	

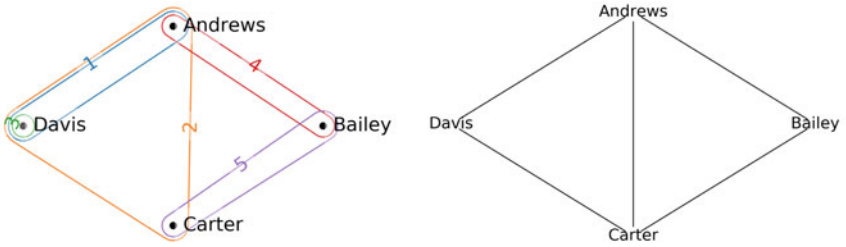


Fig. 1 Bibliometrics example comparing graphs and hypergraphs. (Upper left) Collaborative authorship structure of a set of papers. (Lower left) Euler diagram of its hypergraph. (Upper right) Adjacency matrix of the 2-section. (Lower right) Co-authorship graph resulting from the 2-section

computational topology, to identify features specific to the high-dimensional complexity in hypernetworks, but not available using graphs.

Hypergraph methods are well known in discrete mathematics and computer science, where, for example, hypergraph partitioning methods help enable parallel matrix computations [9], and have applications in VLSI [22]. In the network science literature, researchers have devised several path and motif-based hypergraph data analytics (albeit fewer than their graph counterparts), such as in clustering coefficients [31] and centrality metrics [14]. Although an expanding body of research attests to the increased utility of hypergraph-based analyses [17, 20, 24, 30], and are seeing increasingly wide adoption [19, 28, 29], many network science methods have been historically developed explicitly (and often, exclusively) for graph-based analyses. Moreover, it is common for analysts to reduce data arising from hypernetworks to graphs, thereby losing critical information.

As explicit generalizations of graphs, we must take care with axiomatization, as there are many, sometimes mutually inconsistent, sets of possible definitions of hypergraph concepts which can yield the same results (all consistent with graph theory) when instantiated to the 2-uniform case. And some graph concepts have difficulty extending naturally at all. For example, extending the spectral theory of graph adjacency matrices to hypergraphs is unclear: hyperedges may contain more than two vertices, so the usual (two-dimensional) adjacency matrix cannot code adjacency relations. In other cases, hypergraph extensions of graph theoretical concepts may

be natural, but trivial, and risk ignoring structural properties only in hypergraphs. For example, while edge incidence and vertex adjacency can occur in at most one vertex or edge for graphs, these notions are set-valued and hence *quantitative* for hypergraphs. So while subsequent graph walk concepts like connectivity are applicable to hypergraphs, if applied simply, they ignore high-order structure in not accounting for the "widths" of hypergraph walks.

Researchers have handled the complexity and ambiguity of hypergraphs in different ways. A very common approach is to limit attention to k-uniform hypergraphs, where all edges have precisely k vertices (indeed, one can consider graph theory itself as actually the theory of 2-uniform hypergraphs). This is the case with much of the hypergraph mathematics literature, including hypergraph coloring [11, 25], hypergraph spectral theory [7, 8], hypergraph transversals [2], and extremal problems [32]. k-uniformity is a very strong assumption, which supports the identification of mathematical results. But real-world hypergraph data are effectively *never* uniform; or rather, if presented with uniform hypergraph data, a wise data scientists would be led to consider other mathematical representations for parsimony.

Another prominent approach to handling real-world, and thus non-uniform, hypergraph data is to study simpler graph structures implied by a particular hypergraph. Known by many names, including line graph, 2-section, clique expansion, and one-mode projection, such reductions allow application of standard graph-theoretic tools. Yet, unsurprisingly, such hypergraph-to-graph reductions are inevitably and strikingly lossy [10, 23]. Hence, although affording simplicity, such approaches are of limited utility in uncovering hypergraph structure.

Our research group is dedicated to facing the challenge of the complexity of hypergraphs in order to gain the formal clarity and support for analysis of complex data they provide. We recognize that to enable analyses of hypernetwork data to better reflect their complexity but remain tractable and applicable, striking a balance in this faithfulness-simplicity tradeoff is essential. Placing hypergraph methods in the context of the range of both network science methods on the one hand, and higher-order topological methods on the other, can point the way to such a synthesis.

The purpose of this paper is to communicate the *breadth* of hypergraph methods (which we are exploring in depth elsewhere) to the complex systems community. In the next section we survey the range of mathematical methods and data structures we use. Following that we will illustrate some uses by showing examples in two different contexts: gene set annotations for disease analysis and drug discovery, and cyber analytics of domain-name system relations.

2 Hypergraph Analytics

A **hypergraph** is a structure $\mathcal{H} = \langle V, E \rangle$, with $V = \{v_j\}_{j=1}^{n}$ a set of vertices, and $E = \{e_i\}_{i=1}^{m}$ an indexable family of hyperedges $e_i \subseteq V$. Hyperedges come in different sizes $|e_i|$ possibly ranging from the singleton $\{v\} \subseteq V$ (distinct from the element $v \in V$) to the entire vertex set V.

Table 1 Incidence matrix of example hypergraph Fig. 1

	1	2	3	4	5
a	1	1	0	1	0
b	0	0	0	1	1
c	0	1	0	0	1
d	1	1	1	0	0

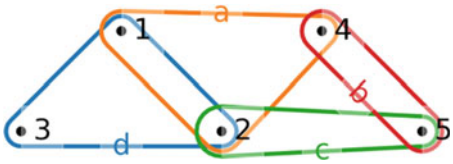


Fig. 2 Dual hypergraph $\mathcal{H}^*$ of our example

A hyperedge $e = \{u, v\}$ with $|e| = 2$ is the same as a graph edge. Indeed, all graphs $\mathcal{G} = \langle V, E\rangle$ are hypergraphs: in particular, graphs are "2-uniform" hypergraphs, so that now $E \subseteq \binom{V}{2}$ and all $e \in E$ are unordered pairs with $|e| = 2$. It follows that concepts and methods in hypergraph theory should explicitly specialize to those in graph theory for the 2-uniform case. But conversely, starting from *graph theory* concepts, there can be many ways of consistently extending them to hypergraphs. The reality of this will be seen in a number of instances below.

Hypergraphs can be represented in many forms. In our example in Fig. 1 above, letting $V = \{a, b, c, d\}$ for the authors and $E = \{1, 2, 3, 4, 5\}$ for the papers, we first represent it as a set system

$$\mathcal{H} = \{\{a, d\}, \{a, c, d\}, \{d\}, \{a, b\}, \{b, c\}\}.$$

We commonly compress set notation for convenience, and when including edge names as well, this yields the compact set system form $\mathcal{H} = \{1{:}ad, 2{:}acd, 3{:}d, 4{:}ab, 5{:}bc\}$. This representation in turn points to the fact that a hypergraph $\mathcal{H}$ is determined uniquely by its Boolean **incidence matrix** $B_{n\times m}$, where $B_{ji} = 1$ iff $v_j \in e_i$, and 0 otherwise. The incidence matrix for the example from Fig. 1 is shown in Table 1. Note that hypergraph incidence matrices are general Boolean matrices, rectangular and non-symmetric, unlike the adjacency matrices typically used to define graphs; while the incidence matrices of graphs are restricted to having precisely two 1's in each column. Also, adjacency structures for hypergraphs are substantially more complicated than for graphs, and in fact are not matrices at all.

The **dual hypergraph** $\mathcal{H}^* = \langle E^*, V^*\rangle$ of $\mathcal{H}$ has vertex set $E^* = \{e_i^*\}_{i=1}^m$ and family of hyperedges $V^* = \{v_j^*\}_{j=1}^n$, where $v_j^* := \{e_i^* : v_j \in e_i\}$. $\mathcal{H}^*$ is just the hypergraph with the transposed incidence matrix B^T, and $(\mathcal{H}^*)^* = \mathcal{H}$. We thus consider that hypergraphs always present as dual *pairs*, which we call the "forward" and the "dual" somewhat arbitrarily, depending on how the data are naturally presented. But this is not true for graphs: the dual $\mathcal{G}^*$ of a graph $\mathcal{G}$ is 2-uniform (and hence still a graph) if and only if $\mathcal{G}$ is 2-regular (all vertices have degree 2), in which case $\mathcal{G}$ is a cycle or disjoint union of cycles. The dual of our example is shown in Fig. 2.

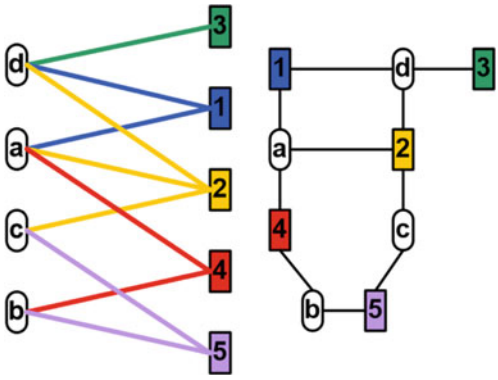


Fig. 3 Two different layouts of the bicolored graph representation of the hypergraph $\mathcal{H}$

There is a bijection between the class of hypergraphs and that of **bicolored graphs**[3] $\mathcal{G} = \langle V, E, \phi \rangle$, where now $E \subseteq \binom{V}{2}$ is a set of unordered pairs of vertices (graph edges), and $\phi : V \to \{0, 1\}$ with $\{v_i, v_j\} \in E$ iff $\phi(v_i) \neq \phi(v_j)$. Conversely, any bicolored graph $\mathcal{G}$ determines a hypergraph $\mathcal{H}$ by associating the vertices and hyperedges of $\mathcal{H}$ with the two colors respectively, and then defining $B_{j,i} = 1$ if and only if $\{v_i, v_j\} \in E$. Figure 3 shows two different layouts of the bicolored graph form of our example hypergraph $\mathcal{H}$.

Moving from the bijection between bicolored graphs and hypergraphs to establishing canonical isomorphisms still requires careful consideration, ideally in a categorical context [12, 15, 33]. While a number of complex network analytics for bipartite graph data can be applied naturally to hypergraphs, and *vice versa*, depending on the semantics of the data being modeled, questions and methods for data with this common structure may be better addressed in one or another form, if only for algorithmic or cognitive reasons. Nor does it mean that graph theoretic methods suffice for studying hypergraphs. Whether interpreted as bicolored graphs or hypergraphs, data with this structure often require entirely different network science methods than (general) graphs. An obvious example is triadic measures like the graph clustering coefficient: these cannot be applied to bicolored graphs since (by definition) bicolored graphs have no triangles. Detailed work developing bipartite analogs of modularity [5], community structure inference techniques [26], and other graph-based network science topics [27] further attests that bipartite graphs (and hypergraphs) require a different network science toolset than for graphs.

In graphs, the structural relationship between two distinct vertices u and v can *only* be whether they are adjacent in a *single* edge ($\{u, v\} \in E$) or not ($\{u, v\} \notin E$); and dually, that between two distinct edges e and f can *only* be whether they are incident at a single vertex ($e \cap f = \{v\} \neq \emptyset$) or not ($e \cap f = \emptyset$). In hypergraphs, both of these concepts are applicable to *sets* of vertices and edges, and additionally become *quantitative*. Define $\mathrm{adj} : 2^V \to \mathbb{Z}_{\geq 0}$ and $\mathrm{inc} : 2^E \to \mathbb{Z}_{\geq 0}$, in both set notation and

[3] Typically the concept of a **bipartite** graph is used here, which is a graph that admits to at least one bicoloring function. The resulting differences are interesting, but not significant for this paper.

(polymorphically) pairwise:

$$\text{adj}(U) = |\{e \supseteq U\}|, \quad \text{adj}(u, v) = |\{e \supseteq \{u, v\}\}|$$

$$\text{inc}(F) = |\cap_{e \in F} e|, \quad \text{inc}(e, f) = |e \cap f|$$

for $U \subseteq V, u, v \in V, F \subseteq E, e, f \in E$. In our example, we have e.g. $\text{adj}(a, d) = 3$, $\text{adj}(\{a, c, d\}) = 1$, $\text{inc}(1, 2) = 2$, $\text{inc}(\{1, 2, 3\}) = 1$. These concepts are dual, in that adj on vertices in $\mathcal{H}$ maps to inc on edges in $\mathcal{H}^*$, and *vice versa*. And for singletons, $\text{adj}(\{v\}) = \deg(v) = |e \ni v|$ is the degree of the vertex v, while $\text{inc}(\{e\}) = |e|$ is the size of the edge e.

This establishes the basis for extending the central concept of graph theory, a **walk** as a sequential visitation of connected nodes, to hypergraphs. Consider a (graph) walk of length ℓ as a sequence $W = v_0, e_0, v_1, e_1, \ldots, e_\ell, v_{\ell+1}$ where v_i, v_{i+1} are adjacent in $e_i, 0 \leq i \leq \ell$, and (dually!) e_i, e_{i+1} are incident on $v_{i+1}, 0 \leq i \leq \ell - 1$. Then W can be equally determined by either the vertex sequence $v_0, \ldots, v_{\ell+1}$, or the edge sequence $e_0, \ldots, e_\ell$. In contrast, with quantitative adjacency and incidence in hypergraphs, sequences of vertices can be adjacent, and sequences of hyperedges incident, in quantitatively different ways, and need not determine each other. Indeed, vertex sequences become hyperedge sequences in the dual, and *vice versa*. For parsimony we work with edgewise walks, and define [1] an ***s*-walk** as a sequence of edges $e_0, e_1, \ldots, e_\ell$ such that $s \leq \text{inc}(e_i, e_{i+1})$ for all $0 \leq i \leq \ell - 1$. Thus walks in hypergraphs are characterized not only by length ℓ, indicating the distance of interaction, but also by "width" s, indicating a *strength* of interaction (see Fig. 4).

For a fixed $s > 0$, we define the ***s*-distance** $d_s(e, f)$ between two edges $e, f \in E$ as the length of the shortest s-walk between them, or infinite if there is none. Note that a graph walk is a 1-walk. We have proved [1] that s-distance is a metric, and can thus define the ***s*-diameter** as the maximum s-distance between any two edges, and an ***s*-component** as a set of edges all connected pairwise by an s-walk. Connected components in graphs are simply 1-components, and our example graph is "connected" in that sense, having a single 1-connected component. But it has *three* 2-components, the hyperedge sets {1, 2}, {4}, and {5}. Other network science methods generalize from graphs to hypergraphs [1], including vertex ***s*-degree**, ***s*clustering coefficients**, and both ***s*-closeness** and ***s*-betweenness centralities**.

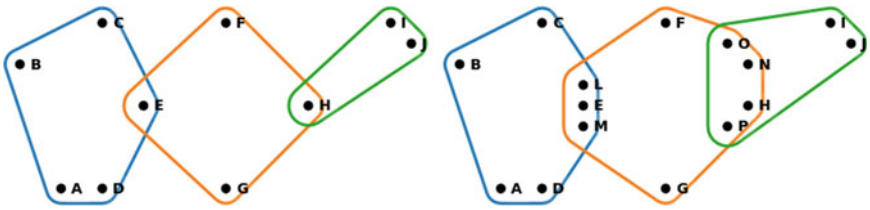


Fig. 4 Two s-walks of length $\ell = 2$. (Left) Lower width $s = 1$. (Right) Higher width $s = 3$

In graphs, two edges e, f are only incident or not, but in hypergraphs, there could additionally be an **inclusion** relationship $e \subseteq f$ or $f \subseteq e$. Indeed, define a **toplex** or **facet** as a maximal edge e such that $\nexists f \supseteq e$, and let $\check{E} \subseteq E$ be the set of all toplexes. Then the **inclusiveness** $I(\mathcal{H}) \in [0, 1)$ of a hypergraph $\mathcal{H}$ is the proportion of included edges, that is, the ratio of non-facets to all edges: $I(\mathcal{H}) := |E \setminus \check{E}|/|E|$. For a hypergraph $\mathcal{H}$, let $\check{\mathcal{H}} = \left\langle V, \check{E} \right\rangle$ be the **simplification** of $\mathcal{H}$, and we call $\mathcal{H}$ **simple** when $\mathcal{H} = \check{\mathcal{H}}$. $I(\mathcal{H}) = 0$ iff $\mathcal{H} = \check{\mathcal{H}}$ is simple, so that all edges are toplexes. In our example, there are three toplexes $\check{E} = \{acd, ab, bc\}$, so that $I(\mathcal{H}) = 2/5$.

Maximal $I(\mathcal{H})$, on the other hand, is more complicated, and the case when all possible sub-edges are present, so that E is closed by subset. This yields $\mathcal{H}$ as an **abstract simplicial complex (ASC)**, so that if $e \in E$ and $f \subseteq e$ then $f \in E$. Let $\widehat{\mathcal{H}} = \left\langle V, \widehat{E} \right\rangle$ be the ASC generated by $\mathcal{H}$, so that $\widehat{E} = \{g \subseteq e\}_{e \in E}$ is the closure of the hyperedges by subset. Each hypergraph $\mathcal{H}$ then maps to a class of hypergraphs we call a **hyperblock** $[\mathcal{H}]$, so that each pair of hypergraphs $\mathcal{H}', \mathcal{H}'' \in [\mathcal{H}]$ have the same ASC: $\widehat{\mathcal{H}'} = \widehat{\mathcal{H}''}$. It follows that they also have the same toplexes: $\widecheck{\mathcal{H}'} = \widecheck{\mathcal{H}''}$.

This results in another representation we call a **simplicial diagram**, shown for our example in Fig. 5. The toplexes $\check{E}$ of $\mathcal{H}$ are shown as a collection of hypertetrahedrons joined where they intersect. This is also sufficient to indicate the ASC $\widehat{\mathcal{H}}$, and, indeed, all the hypergraphs $\mathcal{H}' \in [\mathcal{H}]$ in the hyperblock of $\mathcal{H}$ are included in the diagram. They are distinguished by additionally labeling the hyperedges (and circling singletons) actually included in a particular hypergraph $\mathcal{H}' \in [\mathcal{H}]$, including both their toplexes and their included edges. In our example, these are $3 = \{d\} \subseteq 1 = \{a, d\} \subseteq 2 = \{a, c, d\}$. Contrast with the singleton $\{b\}$ or graph edge $\{a, c\}$, which are only in the ASC $\widehat{\mathcal{H}}$, and not edges in $\mathcal{H}$ itself.

Given a hypergraph $\mathcal{H}$, we can define its ***k*-skeleton** $k\text{-skel}(\mathcal{H}) = \{e \in E : |e| = k\}$ as the set of hyperedges of size k. Each $k\text{-skel}(\mathcal{H})$ is thus a k-uniform subhypergraph of $\mathcal{H}$, and we can conceive of $\mathcal{H}$ as the disjoint union of its k-uniform skeletons: $\mathcal{H} = \bigsqcup_k k\text{-skel}(\mathcal{H})$. Where the k-skeleton is the set of all edges of size k *present* in a hypergraph $\mathcal{H}$, in contrast the ***k*-section** is the set of all edges of size k *implied* by $\mathcal{H}$, that is, all the vertex sets which are sub-edges of some hyperedge. Formally, $\mathcal{H}_k = k\text{-skel}\left(\widehat{\mathcal{H}}\right)$, so that the k-section is the k-skeleton of the ASC of $\mathcal{H}$, and the ASC is the disjoint union of the sections: $\widehat{\mathcal{H}} = \bigsqcup_k \mathcal{H}_k$.

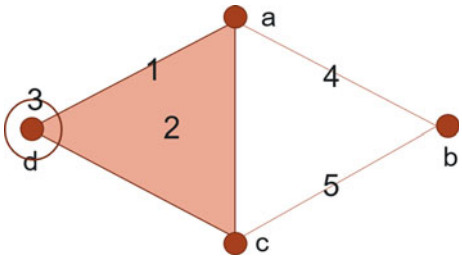
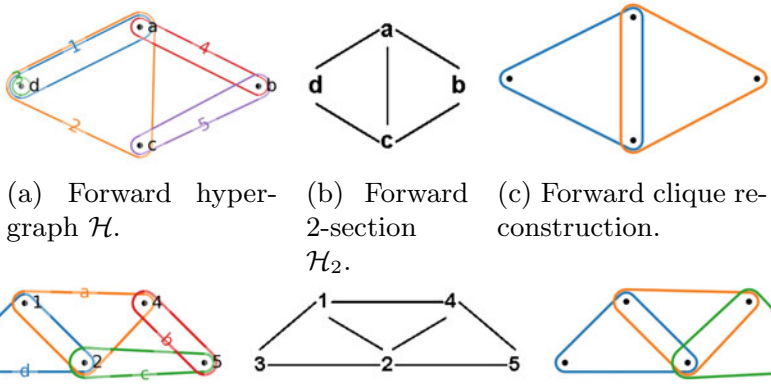


Fig. 5 Example hypergraph $\mathcal{H}$ as a simplicial diagram

Since the k-skeletons are all uniform, and any k-section or union of k-sections is smaller than the entire hypergraph $\mathcal{H}$, there is substantial interest in understanding how much information about a hypergraph is available using only them. The 2-section in particular, which is a graph with adjacency matrix BB^T, can be thought of as a kind of "underlying graph" of a hypergraph $\mathcal{H}$. Also of key interest is the 2-section of the dual hypergraph $\mathcal{H}^*$, called the **line graph** $L(\mathcal{H}) = (\mathcal{H}^*)_2$, which dually is a graph with adjacency matrix $B^T B$. As noted above in the discussion of Fig. 1, these are particularly widely used in studies when confronted with complex data naturally presenting as a hypergraph. The limitations of this are evident in the example in Fig. 6. On the left are our example hypergraph and its dual, and in the center the 2-section $\mathcal{H}_2$ and the line graph $L(\mathcal{H}) = (\mathcal{H}^*)_2$. On the right are the results of taking the maximal cliques of the 2-sections as hyperedges in an attempt to "reconstruct" the original hypergraph $\mathcal{H}$. It is clear how much information is lost.

The ASC $\widehat{\mathcal{H}}$ is additionally a topological complex, that is, a collection of different k-dimensional structures attached together in a particular configuration or pattern. Indeed, the hyperblock $[\mathcal{H}]$ of a (finite) hypergraph $\mathcal{H}$ generates a number of (finite) topological spaces of interest [13]. The most cogent of these is the Alexandrov topology with a sub-base consisting of m open sets $T_j = \{e \supseteq e_j\}_{e \in E}$; that is, for each hyperedge $e_j \in E$, the sub-base element T_j is constructed by collecting all its superedges. The full topology $\mathcal{T}(\mathcal{H})$ is generated by taking all unions of all intersections of these sub-base elements T_j.

The topological space $\mathcal{T}(\mathcal{H})$ will reflect the inherent complexity of the overall "shape" of the hypergraph $\mathcal{H}$. This includes those portions which are connected enough to be contracted, and also the presence of open loops, "holes" or "voids" of different dimension, which can obstruct such contractions. This is



(a) Forward hypergraph $\mathcal{H}$.

(b) Forward 2-section $\mathcal{H}_2$.

(c) Forward clique reconstruction.

(d) Dual hypergraph $\mathcal{H}^*$.

(e) Line graph $L(\mathcal{H}) = (\mathcal{H}^*)_2$.

(f) Dual clique reconstruction.

Fig. 6 2-sections and their clique reconstructions

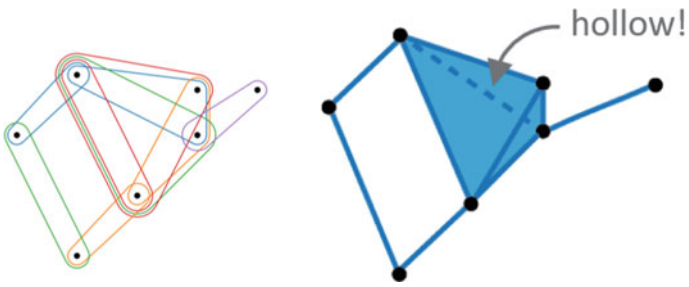


Fig. 7 An example simple hypergraph $\check{\mathcal{H}}$ with $\beta = \langle 1, 1, 1 \rangle$. (Left) Euler diagram of $\check{\mathcal{H}}$. (Right) Simplicial diagram of $\widetilde{\check{\mathcal{H}}}$

called the **homology** of the space $\mathcal{T}(\mathcal{H})$, and is characterized by its **Betti numbers** $\beta_k, 0 \leq k \leq \max |e_j| - 1$, of $\mathcal{H}$, indicating the humber of holes of dimension k present in $\mathcal{H}$. We collect the Betti nubmers to create a **Betti sequence** $\beta = \langle \beta_k \rangle_{k=0}^{\max |e_j|-1}$.

The presence of such gaps may invoke questions or hypotheses: what is stopping the connectivity of these holes, of filling them in? In our example, $\beta_1 = 1$ because of the single open 1-cycle indicated by edges ab, bc, and ca. Contrast this with the similar cycle acd, which is closed in virtue of the hyperedge 2. $\beta_0 = 1$, indicating that $\mathcal{H}$ is 1-connected; while $\beta_2 = 0$, so that $\beta = \langle 1, 1, 0 \rangle$. By comparison, the simplicial diagram of the simple hypergraph $\check{\mathcal{H}}$ shown in Fig. 7 contains a hollow tetrahedron (four triangles surrounding a void), in addition to the open cycle of graph edges on the left. Thus its Betti sequence is $\beta = \langle 1, 1, 1 \rangle$.

3 Example Applications

Here we illustrate some of the mathematical structures and methods introduced above in brief reports of two example case studies.

3.1 Human Gene Set Example

While network science using graph theory methods is a dominant discipline in biomolecular modeling, in fact biological systems are replete with many-way interactions likely better represented as hypergraphs. Genes interact in complex combinations, as recorded in a panoply of biomedical databases. We have begun an exploratory examination of the usefulness of hypergraphs in elucidating the relationships between human genes, *via* their annotations to semantic categories of human

Table 2 Distribution of annotations across biological databases

Database	Annotations
GO biological process	12,305
Reactome pathways	2,291
Chemicals	3,289
Diseases	2,683
Total	**20,568**

biological processes in the Gene Ontology[4] and the Reactome pathway database,[5] chemicals from the Chemical Entities of Biological Interest ontology,[6] and diseases from the Human Disease Ontology.[7] These data were selected to help us better understand potential overlaps between gene sets with causative relations in metabolic rare diseases, like phenylketonuria and Alpha-1 antitrypsin deficiency, and their known biological processes and chemical interactions. We also seek to explore potential overlaps in pathway, biological process, and chemical gene sets as a means to elucidate novel gene targets for drug repurposing, which can then be evaluated in the lab.

Data for this analysis were obtained from PheKnowLator v2.0.0.[8] We compiled a hypergraph $\mathcal{H} = \langle V, E \rangle$ with $|V| = 17{,}806$ human genes as vertices against $|E| = 20{,}568$ annotations as hyperedges, distributed across the source databases as shown in Table 2 (noting that there is substantial overlap among these sources). Of these edges, 8,006 are toplexes, yielding an inclusivity of $I(\mathcal{H}) = 61.1\%$, and the density of the incidence matrix B is 0.000926. Figure 8 shows the distribution of vertex degree $\deg(v) = \text{adj}(\{v\})$ and edge size $|e| = \text{inc}(\{e\})$, with the expected expoential distribution.

Figure 8 shows only the lowest, "first order" distribution of the hypergraph structure, the adjacency and incidence of singletons. Consideration of higher-order interactions would require expensive combinatorial calculations of, for example, k-way intersections and hyperedge inclusions of arbitrary sets of vertices. A modest step towards that goal in our methodology is first to focus on toplexes, which determine the topological structure, and then their pairwise intersections: $\text{inc}(e, f)$ for $e, f \in \check{E}$. This is shown in the left of Fig. 9, which reveals a long tail, indicating a significant number of pairs of annotations with large intersections of genes. Attending to incidences of even higher order would reveal the increasingly rich complex interactions of gene sets.

[4] http://geneontology.org/docs/ontology-documentation.

[5] https://reactome.org.

[6] https://www.ebi.ac.uk/chebi/.

[7] https://disease-ontology.org/.

[8] https://github.com/callahantiff/PheKnowLator, downloaded on 03/05/20, see also https://github.com/callahantiff/PheKnowLator/wiki/v2-Data-Sources.

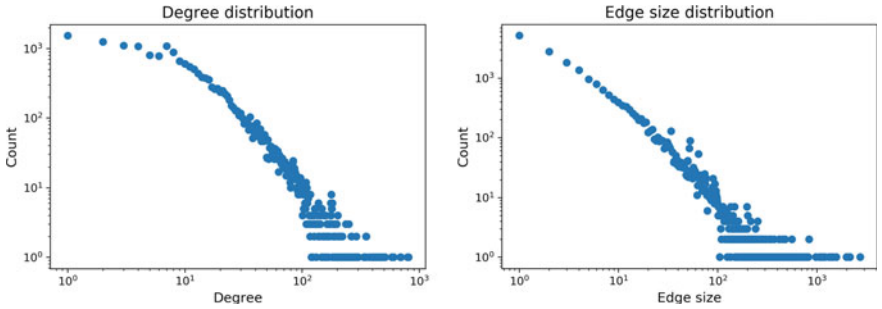


Fig. 8 Distributions of: (Left) Vertex degree $\deg(v) = \text{adj}(\{v\})$; (Right) Edge size $|e| = \text{inc}(\{e\})$

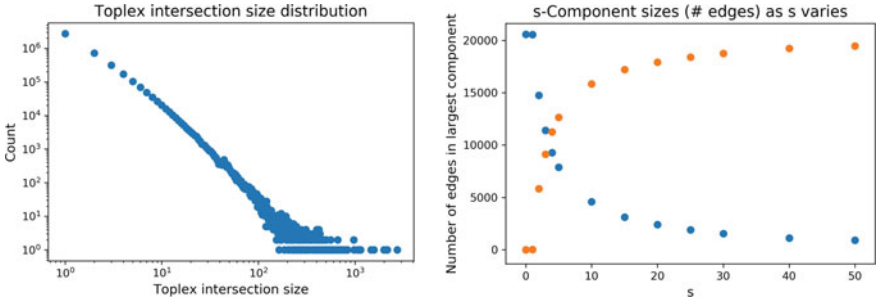


Fig. 9 (Left) Distribution of the size of toplex intersections: $\text{inc}(e, f)$ for $e, f \in \check{E}$. (Right) # components (orange) and size of largest component (blue)

Of even more interest is the right of Fig. 9, which shows distribution information about the connected s-components for different intersection levels s. On the top the number of s-components is shown in orange, and the size of the largest component in blue, all as a function of increasing s. Expectedly these appear monotonic increasing and decreasing respectively, but it's notable that even for large s there persist some very large components, again demonstrating the large multi-way interactions amongst these gene sets.

3.2 DNS Cyber Example

The Domain Name System (DNS) provides a decentralized service to map from domain names (e.g., www.google.com) to IP addresses. Perhaps somewhat counter-intuitively, DNS data present naturally as a hypergraph, in being a many-many relationship between domains and IPs. While typically this relationship is one-to-one, with each domain uniquely identifying a single IP address and *vice versa*, there are a number of circumstances which can violate this, for example domain aliasing,

	2wzi79.com	37b82p.com	5bewa2.com	8sa5zz.com	82jpn3.com	9mtj5d.com	dyewjr.com	ts3fag.com	j4upk1.com	kbeuu8.com	sgbdn1.com	u2594m.com		
103.86.122.130							X			X	X			
103.86.122.148		X	X	X				X				X		
103.86.122.149			X					X				X		
103.86.122.152	X	X		X										
103.86.122.154					X	X	X			X	X			
103.86.122.160							X				X	X		
103.86.122.169			X											
103.86.122.173	X													
103.86.122.181	X			X				X						
103.86.122.192					X									
103.86.122.195						X			X					
103.86.122.220			X											
103.86.122.222		X					X					X		
103.86.122.223					X	X			X	X	X			
103.86.122.225	X	X		X				X						
103.86.122.238									X					
103.86.122.242										X				

Fig. 10 Portion of the incidence matrix for ADNS data

hosting services where one IP serves multiple websites, or duplicated IPs to manage loads against popular domains.

ActiveDNS (ADNS) is a data set maintained by the Astrolavos Lab at the Georgia Institute of Technology.[9] It submits daily DNS lookups for popular zones (e.g., .com, .net, .org) and lists of domain names. Using data from April 26, 2018 as an example, this day consists of 1,200 Avro files with each file containing on average 900K records. Our hypergraph representation coded each domain (hyperedges $e \in E$) as a collection of its IPs (vertices $v \in V$). A small portion of the incidence matrix B is shown in Fig. 10.

To identify some of the simplest hypergraph-specific properties, we looked [21] specifically at the 2-components, and identified the one with maximum 2-diameter (6), which is shown in Fig. 11. The IP addresses in this component all belong to the IP range 103.86.122.0/24 and the domains are registered to GMO INTERNET, INC according to WHOIS records. Moreover, current DNS queries for most of these domains at a later date resolve to IPs in the range 103.86.123.0/24 and have a "time to live" of only 120 s. This pattern of quickly changing of IP address is consistent with the "fast flux" DNS technique which can be used by botnets to hide malicious content delivery sites and make networks of malware more difficult to discover [18].

[9] https://activednsproject.org.

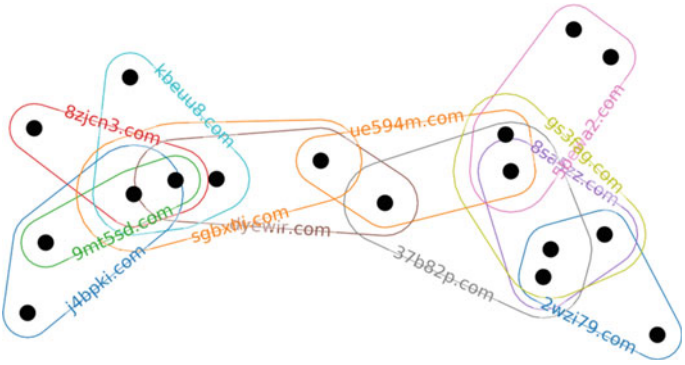


Fig. 11 The 2-component with largest 2-diameter, possibly indicating fast flux behavior

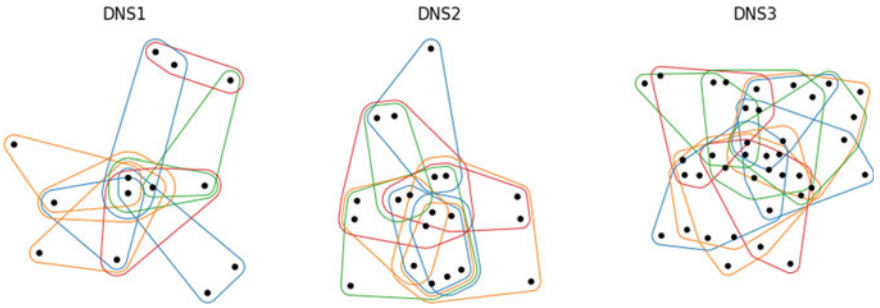


Fig. 12 Three 2-components with non-trivial homologies

Other 2-components reveal non-trivial homologies, three of which are shown in Fig. 12. DNS1 has $\beta = \langle 1, 1, 0, 0, \ldots \rangle$, with a visible hole surrounded by a 1 dimensional loop on the top. DNS2 has $\beta = \langle 1, 1, 2, 0, \ldots \rangle$, and DNS3 has $\beta = \langle 1, 3, 1, 0, \ldots \rangle$, indicating one and three 1-dimensional holes, and two and one 2-dimensional voids, respectively. The open loops in DNS2 and DNS3 are harder to visualize, so Fig. 13 shows a simplicial diagram of one of the two 2-dimensional voids in DNS2. There are two solid tetrahedrons for the domains potterybarn.com and pbteen.com, each with four IPs, three of which (those ending in .160, .9, and .105) they share. Then wshome.com is a triangle in the foreground, and westelm.com a triangle behind (see caption for details). These are effectively "transparent window panes" surrounding a hollow tetrahedral space. Identification of such multi-dimensional open-loop structures affords the opportunity to consider these as hypotheses: on what basis is the Pottery Barn company structuring its multiple domains over their multiple IPs in this complex pattern?

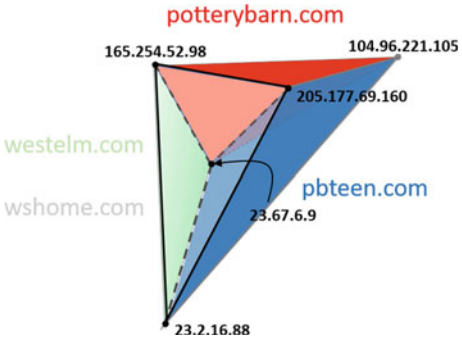


Fig. 13 Simplicial diagram of the complex pattern of IPs shared by some Pottery Barn domains around one of the two 2-dimensional voids in DNS2 of Fig. 12. wshome.com is the transparent foreground triangle with IPs ending in .160, .88 and .98; and westelm.com the background triangle with IPs ending in .98, .88, and .9. Potterybarn.com and pbteen.com are solid tetrahedrons with four IPs each

Acknowledgements PNNL-SA-152208. This work was also partially funded under the High Performance Data Analytics (HPDA) program at the Department of Energy's Pacific Northwest National Laboratory. Pacific Northwest National Laboratory is operated by Battelle Memorial Institute under Contract DE-ACO6-76RL01830.

References

1. Aksoy, S.G., Joslyn, C.A., Marrero, C.O., Praggastis, B., Purvine, E.A.: Hypernetwork science via high-order hypergraph walks. EPJ Data Sci. **9**, 16 (2020)
2. Alon, N.: Transversal numbers of uniform hypergraphs. Graphs Comb. **6**(1), 1–4 (1990)
3. Ausiello, G., Fanciosa, P.G., Frigioni, D.: Directed hypergraphs: Problems, algorithmic results, and a novel decremental approach. In: ICTCS 2001, LNCS, vol. 2202, pp. 312–328 (2001)
4. Barabási, A.L.: Network Science. Cambridge University Press, Cambridge (2016)
5. Barber, M.J.: Modularity and community detection in bipartite networks. Phys. Rev. E **76**(6) (2007). https://doi.org/10.1103/physreve.76.066102
6. Berge, C., Minieka, E.: Graphs and Hypergraphs. North-Holland (1973)
7. Chung, F.: The laplacian of a hypergraph. Expanding graphs (DIMACS series), pp. 21–36 (1993)
8. Cooper, J., Dutle, A.: Spectra of uniform hypergraphs. Linear Algebr. Its Appl. **436**(9), 3268–3292 (2012)
9. Devine, K., Boman, E., Heaphy, R., Bisseling, R., Catalyurek, U.: Parallel hypergraph partitioning for scientific computing. In: Proceedings 20th IEEE International Parallel & Distributed Processing Symposium. IEEE (2006). https://doi.org/10.1109/ipdps.2006.1639359
10. Dewar, M., Healy, J., Pérez-Giménez, X., Prałat, P., Proos, J., Reiniger, B., Ternovsky, K.: Subhypergraphs in non-uniform random hypergraphs. Internet Math. (2018). https://doi.org/10.24166/im.03.2018
11. Dinur, I., Regev, O., Smyth, C.: The hardness of 3-uniform hypergraph coloring. Combinatorica **25**(5), 519–535 (2005)
12. Dörfler, W., Waller, D.A.: A category-theoretical approach to hypergraphs. Archiv der Mathematik **34**(1), 185–192 (1980). https://doi.org/10.1007/bf01224952

13. Edelsbrunner, H., Harer, J.L.: Computational Topology: An Introduction. AMS (2000)
14. Estrada, E., Rodríguez-Velázquez, J.A.: Subgraph centrality and clustering in complex hyper-networks. Phys. A: Stat. Mech. Its Appl. **364**, 581–594 (2006). https://doi.org/10.1016/j.physa.2005.12.002
15. Fong, B., Spivak, D.I.: Hypergraph categories (2019)
16. Gallo, G., Longo, G., Pallottino, S.: Directed hypergraphs and applications. Discret. Appl. Math. **42**, 177–201 (1993)
17. Iacopini, I., Petri, G., Barrat, A., Latora, V.: Simplicial models of social contagion. Nat. Commun. **10**, 2485 (2019)
18. jamie.riden: How Fast-Flux Service Networks Work. http://www.honeynet.org/node/132. Last accessed 26 Nov 2018
19. Javidian, M.A., Lu, L., Valtorta, M., Qang, Z.: On a hypergraph probabilistic graphical model (2018). https://arxiv.org/abs/1811.08372
20. Johnson, J.: Hypernetworks in the Science of Complex Systems. Imperial College Press, London (2013)
21. Joslyn, C.A., Aksoy, S., Arendt, D., Firoz, J., Jenkins, L., Praggastis, B., Purvine, E.A., Zalewski, M.: Hypergraph analytics of domain name system relationships. In: Kaminski, B., et al. (eds.) 17th Workshop on Algorithms and Models for the Web Graph (WAW 2020). Lecture Notes Comput. Sci. **12901**, 1–15. Springer (2020)
22. Karypis, G., Kumar, V.: Multilevel k-way hypergraph partitioning. VLSI Des. **11**(3), 285–300 (2000). https://doi.org/10.1155/2000/19436
23. Kirkland, S.: Two-mode networks exhibiting data loss. J. Complex Netw. **6**(2), 297–316 (2017). https://doi.org/10.1093/comnet/cnx039
24. Klamt, S., Haus, U.U., Theis, F.: Hypergraphs and cellular networks. PLoS Comput. Biol. **5**(5), e1000,385 (2009)
25. Krivelevich, M., Sudakov, B.: Approximate coloring of uniform hypergraphs. J. Algorithms **49**(1), 2–12 (2003)
26. Larremore, D.B., Clauset, A., Jacobs, A.Z.: Efficiently inferring community structure in bipartite networks. Phys. Rev. E **90**(1) (2014). https://doi.org/10.1103/physreve.90.012805
27. Latapy, M., Magnien, C., Vecchio, N.D.: Basic notions for the analysis of large two-mode networks. Soc. Netw. **30**(1), 31–48 (2008). https://doi.org/10.1016/j.socnet.2007.04.006
28. Leal, W., Restrepo, G.: Formal structure of periodic system of elements. Proc. R. Soc. A. **475** (2019). https://doi.org/10.1098/rspa.2018.0581
29. Minas, M.: Hypergraphs as a unifrom diagram representation model. In: Proceedings of the 6th International Workshop on Theory and Applications of Graph Transformations (2000). ftp://ftp.informatik.uni-erlangen.de/local/inf2/Papers/tagt98.pdf
30. Patania, A., Petri, G., Vaccarino, F.: Shape of collaborations. EPJ Data Sci. **6**, 18 (2017)
31. Robins, G., Alexander, M.: Small worlds among interlocking directors: network structure and distance in bipartite graphs. Comput. Math. Organ. Theory **10**(1), 69–94 (2004). https://doi.org/10.1023/b:cmot.0000032580.12184.c0
32. Rödl, V., Skokan, J.: Regularity lemma for k-uniform hypergraphs. Random Struct. Algorithms **25**(1), 1–42 (2004)
33. Schmidt, M.: Functorial approach to graph and hypergraph theory (2019)