Abstract

We propose high-order hypergraph walks as a framework to generalize graph-based network science techniques to hypergraphs. Edge incidence in hypergraphs is quantitative, yielding hypergraph walks with both length and width. Graph methods which then generalize to hypergraphs include connected component analyses, graph distance-based metrics such as closeness centrality, and motif-based measures such as clustering coefficients. We apply high-order analogs of these methods to real world hypernetworks, and show they reveal nuanced and interpretable structure that cannot be detected by graph-based methods. Lastly, we apply three generative models to the data and find that basic hypergraph properties, such as density and degree distributions, do not necessarily control these new structural measurements. Our work demonstrates how analyses of hypergraph-structured data are richer when utilizing tools tailored to capture hypergraph-native phenomena, and suggests one possible avenue towards that end.

1 Introduction

In the study of complex systems, graph theory is often perceived as the mathematical scaffold underlying network science [6]. Systems studied in biology, sociology, telecommunications, and physical infrastructure often afford a representation as a set of entities (“vertices”) with binary relationships (“edges”), and hence may be analyzed utilizing graph theoretic methods. Graph models benefit from simplicity and a degree of universality. But as abstract mathematical objects, graphs are limited to representing pairwise relationships between entities. However, 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.

Hypergraphs are generalizations of graphs in which edges may connect any number of vertices, thereby representing k-way relationships. As such, hypergraphs are 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. Additionally, as finite set systems, hypergraphs have identities related to a number of other mathematical structures important in data science, including finite topologies, simplicial complexes, and Sperner systems. This enables use of a wider range of mathematical methods, such as those from computational topology, to identify features specific to the high-dimensional complexity in hypernetworks, but not available using graphs. Although an expanding body of research attests to the increased utility of hypergraph-based analyses, many network science methods have been historically developed explicitly (and often, exclusively) for graph-based analyses. Moreover, it is common that data arising from hypernetworks are reduced to graphs.

In spite of this incongruity, effectively extending graph theoretical tools to hypergraphs has sometimes lagged or proven elusive. A critical aspect of this is one of axiomatization: as a generalization there are many, sometimes mutually inconsistent, sets of possible definitions of hypergraph concepts which can yield the same results consistent with graph theory when instantiated to the graph case. In some cases, developing any coherent hypergraph analog poses significant theoretical obstacles. For example, extending the spectral theory of graph adjacency matrices to hypergraphs poses an immediate challenge in that hyperedges may contain more than two vertices, thereby rendering the usual (two-dimensional) adjacency matrix insufficient for encoding adjacency relations. In other cases, graph theoretical concepts may be
trivially extended to hypergraphs, but in doing so ignore structural nuance native to hypergraphs which are unobservable in graphs. For instance, 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. Consequently, while subsequent graph walk based notions, such as connectedness, are immediately applicable to hypergraphs, they ignore high-order structure in failing to account for the varying “widths” associated with hypergraph walks.

Due to these challenges, scientists seeking tools to study hypergraph-structured data are frequently left to contend with disparate approaches towards hypergraph research. One approach for grappling with hypergraph complexity is to limit attention to hypergraphs with only uniformly sized edges containing the same number of vertices. Much of the hypergraph research in the mathematics literature, such as in hypergraph coloring, [21] 22, the aforementioned spectral theory of hypergraphs [12] 13, hypergraph transversals [3], and extremal problems [20], focus on this $k$-uniform case only. While imposing this assumption facilitates more mathematically sophisticated and structurally faithful analysis of the hypergraphs in question, real-world hypergraph data is unfortunately very rarely $k$-uniform. Consequently, such tools are problematic in lacking applicability to real hypernetwork data. Another approach towards hypergraph research is to limit attention to transformations of (potentially non-uniform) hypergraphs to graphs. Sometimes called the hypergraph line graph, 2-section, clique expansion, or one-mode projection, such transformations clearly enable the application of graph-theoretic tools to the data. Yet, unsurprisingly, such hypergraph-to-graph reductions are inevitably lossy [21] 38. Hence, although affording simplicity, such approaches are of limited utility in uncovering hypergraph structure.

To enable analyses of hypernetwork data that better reflect their complexity but remain tractable and applicable, we believe striking a balance between this faithfulness-simplicity tradeoff is essential. With this goal at heart, we study how a number of graph analytic tools popular in network science extend to hypergraphs under the framework of high-order hypergraph walks. We characterize a hypergraph walk as an “$s$-walk”, where the order $s$ controls the minimum walk “width” in terms of edge overlap size. High-order $s$-walks ($s > 1$) are possible on hypergraphs whereas for graphs, all walks are 1-walks. The hypergraph walk-based methods we consider include connected component analyses, graph-distance based metrics such as closeness centrality, and motif-based measures such as clustering coefficients. As each of these methods is based fundamentally on the graph-theoretic notion of a walk, we extend them to hypergraphs by using hypergraph walks. Ultimately, our goal is not only to formulate these generalizations in a cogent manner, but to probe whether these tools reveal prevalent and meaningful structure in real hypernetwork data. To the latter end, we compute these measures based on hypergraph walks on three real datasets from different domains and discuss the results.

Our work is organized as follows: in Section 2 we provide background definitions and review preliminary topics relevant to hypernetwork theory. In Section 3 we define the $s$-walk notion underpinning our subsequent work and discuss related prior research. In Section 4 we introduce $s$-walk based analytical measures, apply them to the aforementioned datasets, and briefly analyze the results. In Section 5 we consider three generative hypergraph models, and experimentally test the extent to which the structural properties observed in Section 4 can be replicated by synthetic models. Finally, in Section 6 we conclude and outline several directions for future research.

2 Preliminaries

Hypergraphs are generalizations of graphs in which edges may link any number of vertices together. Just as “network” is often used to refer to processes or systems which yield data streams which are graph-structured, we will use the term “hypernetwork” to refer to those yielding hypergraph-structured data. More formally, we define a hypergraph as follows:

**Definition 1.** A hypergraph $H = (V, E)$ is a set $V = \{v_1, \ldots, v_n\}$ of elements called vertices, and an indexed family of sets $E = \{e_1, \ldots, e_m\}$ called edges in which $e_i \subseteq V$ for $i = 1, \ldots, m$.

The degree of a vertex is the number of hyperedges to which it belongs, $d(v) = |\{e : v \in e\}|$, and the size of an edge is its cardinality, $|e|$. A hypergraph in which all hyperedges have size $k$ is called $k$-uniform, and a 2-uniform hypergraph is simply a graph. We note that definitions of hypergraphs given in the literature may differ slightly from author to author. For instance, Bretto’s definition [10] of a hypergraph is identical to ours, apart from prohibiting empty edges ($e_i$ such that $e_i = \emptyset$). Berge [8] similarly prohibits empty edges, as well as isolated vertices ($v_i$ such that $v_i \notin \cup_{i=1}^m e_i$). In contrast, Katona [46] allows empty edges and isolated vertices, but defines $E = \{e_1, \ldots, e_m\}$ as a set and explicitly prohibits pairs of duplicated edges $e_i = e_j$ for $i \neq j$. In defining $E$ as an (indexed) family of sets, we allow
for duplicated edges but require edges be distinguishable by index. The overall generality of Definition\(^1\) in permitting isolated vertices, as well as empty, duplicated, and singleton edges is intended to facilitate the application of hypergraphs to real data, which commonly possess such features.

**Definition 2.** The incidence matrix \(S\) of a hypergraph \(H = (V, E)\), is a \(|V| \times |E|\) matrix defined by

\[
S(i, j) = \begin{cases} 
1 & \text{if } v_i \in e_j \\
0 & \text{otherwise.}
\end{cases}
\]

Under Definition\(^1\), any rectangular binary matrix uniquely defines a labeled hypergraph\(^2\) conversely any labeled hypergraph uniquely defines an incidence matrix. Consequently, it can be easily seen there is a bijection between hypergraphs and bicolored graphs. Recall a bicolored graph is a triple \((V, E, f)\) where \(V\) is a set of vertices, \(E\) is a set of unordered pairs of vertices, and \(f : V \to \{0, 1\}\) satisfies \(f(v_i) \neq f(v_j)\) for all \(v_i, v_j \in V\) such that \(\{v_i, v_j\} \in E\). Indexing the rows and columns by vertices such that \(f(v_i) = 0\) and \(f(v_j) = 1\), respectively, bicolored graphs may also be uniquely associated with incidence matrices (and hence hypergraphs) by defining \(S(i, j) = 1\) if and only if \(\{v_i, v_j\} \in E\). To avoid confusion, it is worth emphasizing bicolored graphs differ from *bipartite* graphs in explicitly specifying a fixed bicoloring \(f\). In contrast, a bipartite graph is defined as a graph admitting *some* bicingoloring \(f\). Accordingly, since a bipartite graph with \(k\) connected components has \(2^k\) possible bicolorings, a single bipartite graph may correspond to up to \(2^k\) distinct hypergraphs, depending on the bicingoloring. In applications, however, it is often the case that a natural bicingoloring is already given explicitly by the data (e.g. in an author-paper network) and hence the terms “bipartite” and “bicolored” graphs have been used synonymously.

As an upshot of the hypergraph-bicolored graph correspondence, a number of complex network analytics for bipartite graph data extend naturally to hypergraphs, and *vice versa*. However, interpreting this correspondence in light of the fact that bicolored graphs are graphs does not mean that graph theoretic methods suffice for studying hypergraphs. To the contrary, bicolored graph data 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 \([7]\), community structure inference techniques \([22]\), and other graph-based network science topics \([23]\) further attests that bipartite graphs (and hypergraphs) require a different network science toolset than for graphs.

While graph theoretic methods should not be conflated with those of bicolored graphs, it is natural to ask why we gear our exposition towards hypergraphs. We utilize the language of hypergraphs because of the fundamentally set-theoretic nature of our approach. Our focus in this work is on hyperedge incidences and hyperwalks that arise from sequences of incident hyperedges. Hyperedges themselves are defined explicitly for hypergraphs, but only implicitly for bicolored graphs (as the neighborhood of a vertex in the color class designated for hyperedges). For this reason, framing our exposition using the language of arbitrary set systems is natural, whereas adopting the constrained language of bicolored graphs would be cumbersome and confusing.

Another important topic highlighted by the bicolored graphs-hypergraph correspondence is the inherent duality of hypergraphs. That is, just as it may be arbitrary to label one partition in a bicolored graph “left” and the other partition “right”, which class of objects one designates as “vertices” versus “hyperedges” in a hypernetwork may also be arbitrarily chosen. However, hypergraph properties and methods may be vertex-based or edge-based, and hence differ depending on which choice is made. To avoid limiting one’s analysis toward either a vertex-centric or edge-centric approach, it may be prudent to consider both the hypergraph and its *dual hypergraph*. Loosely speaking, the dual of a hypergraph is the hypergraph constructed by swapping the roles of vertices and edges. More precisely:

**Definition 3.** Let \(H = (V, E)\) be a hypergraph with vertex set \(V = \{v_1, \ldots, v_n\}\) and family of edges \(E = (e_1, \ldots, e_m)\). The dual hypergraph of \(H\), denoted \(H^* = (E^*, V^*)\), has vertex set \(E^* = \{e^*_1, \ldots, e^*_m\}\) and family of edges \(V^* = \{v^*_1, \ldots, v^*_n\}\), where \(e^*_k := \{v_i \in e_k\}\).

Put equivalently, the dual of a hypergraph with incidence matrix \(S\) is the hypergraph associated with the transposed incidence matrix, \(S^T\). Clearly, \((H^*)^* = H\). Furthermore, observe that two vertices belonging to the same set of edges in \(H\) correspond to multi-edges in the \(H^*\) and isolated vertices in \(H\) correspond to empty edges in \(H^*\). Thus, the generality of our Definition\(^1\) in permitting multi-edges, empty edges, and isolated vertices ensures the dual of a hypergraph is also a hypergraph. Indeed, as a formal matter, one could go so far as to always consider that hypergraphs present in dual pairs. Note,

\(^1\)By “labeled hypergraph” we mean a hypergraph in which each vertex and edge are distinguishable via an assignment of distinct labels – this is not meant to be confused with so-called attributed hypergraphs in which the vertices and edges have associated metadata.
however, this is not necessarily true when restricted to the graph case: for a graph $G$, its dual $G^*$ is 2-uniform (and hence still a graph) if and only if $G$ is 2-regular, in which case $G$ is a cycle or disjoint union of cycles.

Continuing this line of observation, in the complex networks literature, one of the most oft-used tools for studying hypergraph data is its line graph. In the line graph of a hypergraph, each vertex represents a hyperedge, and each edge represents an intersection between a pair of hyperedges. More formally:

**Definition 4.** Let $H = (V, E)$ be a hypergraph with vertex set $V = \{v_1, \ldots, v_n\}$ and edges $E = \{e_1, \ldots, e_m\}$. The **line graph** of $H$, denoted $L(H)$, is the graph on vertex set $\{e_1^*, \ldots, e_m^*\}$ and edge set $\{(e_i^*, e_j^*) : e_i \cap e_j \neq \emptyset \text{ for } i \neq j\}$.

In order to additionally capture information about the size of hyperedge edge intersections, line graphs of hypergraphs may be defined with additional edge weights where $\{e_i^*, e_j^*\}$ has weight $|e_i \cap e_j|$. By definition of matrix multiplication, it is easy to see that for a hypergraph with incidence matrix $S$, its line graph has edge-weighted adjacency matrix $S^T S$ with diagonal entries all converted to zero. Figure 1 gives an example of a hypergraph, its dual, and their respective line graphs. All the hypergraph visualizations in this paper were created using HyperNetX (HNX) [56], a recently released Python library echoing NetworkX [28] for exploratory data analytics and visualization of hypergraphs.

Hypergraph line graphs are also referred to by a plethora of other names. Berge [8] refers to $L(H)$ as both the “line graph” and “representative graph” of $H$; Naik [49, 50] refers to $L(H)$ the “intersection graph” of $H$. In the complex networks literature on bipartite graphs or “2-mode” graphs, the aforementioned “one-mode projections” are equivalent to hypergraph line graphs. For instance, $L(H)$ and $L(H^*)$ are referred to as the “top and bottom” projections in [44], similarly, [24] dubs these the “column and row” projections. Moreover, $L(H^*)$ is commonly referred to as the “2-section”, “clique graph”, or “clique expansion” of the hypergraph $H$, since the edge set of $L(H^*)$ is generated by taking all 2-element subsets of each edge in $H$, hence vertices within each hyperedge $H$ form a clique in $L(H^*)$. Consequently, if $G$ is a graph, $L(G^*)$ is identical to $G$.

Line graphs play an important role in hypernetwork science. Due to the relative dearth of hypergraph analytic tools, line graphs are often analyzed in place of the hypergraphs they were derived from so that classical network science techniques can be applied. However, as has been noted, hypergraph line graphs are fundamentally limited in several ways. First, line graphs are lossy representations of hypergraphs in the sense that distinct hypergraphs can have identical line graphs. We note that such structural loss does not occur for graphs, as Whitney’s theorem [67] states, apart from the triangle and 4-vertex star graph, any pair of connected graphs with isomorphic line graphs must be isomorphic. In the case of hypergraph line graphs, Kirkland [38] recently illustrated the structural loss in a severe sense by giving an example of two distinct 19 × 19 incidence matrices $S$ and $R$ respectively, such that both

\[
S^T S = R^T R, \quad SS^T = RR^T.
\]

Put equivalently in the language of hypergraphs: although non-isomorphic, the weighted line graphs of the hypergraphs represented by $S$ and $R$, as well as those of their dual hypergraphs, are both identical. Using tools from combinatorial matrix theory, Kirkland also constructed infinite families of such pairs of hypergraphs and showed they constitute a vanishingly small proportion of hypergraphs. Accordingly, while one isn’t likely to encounter such pairs of hypergraphs in empirical data, Kirkland’s work illustrates structural properties of hypergraphs may be lost even when simultaneously accounting for hypergraph duality and using weighted line graphs. Consequently, depending on the properties under consideration,
the extent to which line graphs faithfully represent hypergraphs may be unclear. Nonetheless, researchers have offered preliminary evidence that some meaningful, albeit incomplete, hypergraph structure can be extracted from their line graphs [24].

Lastly, as has been previously noted [11, 62], another important limitation of line graphs of hypergraphs is computational: sparse hypergraphs can still yield relatively dense line graphs that may be difficult to analyze or store in computer memory. Indeed, this can be easily seen by observing that \( k \)-way edge intersections (guaranteed by a vertex of degree \( k \)) in the hypergraph yield \( \binom{k}{2} \) edges in its line graph. Particularly if the hypergraph is large and its vertex degree and edge cardinality distributions heavily skewed (common features in real world network data), its line graphs may be too dense to analyze computationally or even construct at all.

## 3 From Graph Walks to Hypergraph Walks

One of the most fundamental concepts in graph theory, underpinning a myriad of areas including Hamiltonian and Eulerian graphs, distance and centrality measures, stochastic processes on graphs and PageRank, is that of a walk. For a graph \( G = (V, E) \), a walk of length \( k \) is a sequence of vertices \( v_0, v_1, \ldots, v_k \), such that each pair of successive vertices are adjacent. By definition of a (simple) graph, two adjacent vertices belong to exactly one edge, and conversely, two incident edges intersect in exactly one vertex. Consequently, any valid graph walk can be equivalently expressed as either a sequence of adjacent vertices or as a sequence of incident edges, i.e.

\[
\begin{align*}
v_0, & \quad v_1, \quad \ldots, \quad v_{k-1}, \quad v_k \\
e_0 \cap v_1, & \quad e_0 \cap v_2, \quad \ldots, \quad e_{k-1} \cap v_k, \quad e_k
\end{align*}
\]

In the setting of hypergraphs, this simple observation no longer holds. Hypergraph edge incidence and vertex adjacency is *set-valued* and *quantitative* in the sense that two hyperedges can intersect at any number of vertices, and two vertices can belong to any number of shared hyperedges. This motivates two walk concepts for hypergraphs that are dual but nonetheless distinct: walks on the vertex level (consisting of successively adjacent vertices), and walks on the edge level (consisting of successively intersecting edges). For ease of presentation, and to be consistent with the presentation of similar notions appearing in the literature, we limit our exposition to edge-level hypergraph walks. Nonetheless, we emphasize that both notions are captured when duality is considered, as a vertex-based walk on a hypergraph \( H \) is simply an edge-walk on the dual hypergraph \( H^* \). We define a hypergraph walk as an “\( s \)-walk” on a hypergraph, where \( s \) controls for the size of edge intersection, as follows:

**Definition 5.** For a positive integer \( s \), an \( s \)-walk of length \( k \) between hyperedges \( f \) and \( g \) is a sequence of hyperedges,

\[
f = e_{i_1}, e_{i_1}, \ldots, e_{i_k} = g,
\]

where for \( j = 1, \ldots, k \), we have \( s \leq |e_{i_{j-1}} \cap e_{i_j}| \) and \( i_{j-1} \neq i_j \).

In this way, a graph walk is simply a 1-walk, whereas \( s \)-walks for \( s > 1 \) are only possible for hypergraphs. When interpreted on the dual hypergraph \( H^* \), an \( s \)-walk corresponds to a sequence of adjacent vertices in which each successive pair of vertices belong to at least \( s \) shared hyperedges. As will become apparent in subsequent sections, a number of basic, yet important, properties of walks in graphs immediately extend to \( s \)-walks on hypergraphs. For instance, just as any graph walk ending at vertex \( v_k \) can be concatenated with any walk starting at vertex \( v_k \) to form another walk, any \( s \)-walk ending at a particular edge can be concatenated to any other \( s \)-walk starting at the edge. Consequently, the existence of an \( s \)-walk between hyperedges defines an equivalence relation under which hyperedges can be partitioned into \( s \)-connected components, which we explore in Section 4.2. Furthermore, this also ensures the length of the shortest \( s \)-walk between edges, called \( s \)-distance (Section 4.3), satisfies the triangle inequality and defines a bona-fide distance metric on the hypergraph. Finally, in Section 4.4 we explore how one may distinguish between different kinds of \( s \)-walks in a hierarchical way, and how the subsequent notions of \( s \)-traces, \( s \)-meanders, \( s \)-paths, and \( s \)-cycles lend themselves to discerning substructures native to hypergraphs, such as \( s \)-triangles.

Many researchers have considered different notions of “high-order walks” on hypergraphs, abstract simplicial complexes, and related set systems. Indeed, a number of concepts closely related to \( s \)-walks have for long appeared in the mathematics literature. Bermond, Heydemann, and Sotteau [9] introduced and analyzed \( k \)-line graphs of uniform hypergraphs, which are derived from hypergraphs by representing each hyperedge as a vertex, and linking two such vertices if their corresponding hyperedges intersect in at least \( k \) vertices. In this way, a (graph) walk on their line graphs corresponds to an \( s \)-walk on a hypergraph.
In [47], Lu and Peng define higher order walks on hypergraphs for \( k \)-uniform hypergraphs as sequences of intersecting edges in which the vertices within each edge are ordered. Their work is related to a rich literature on Hamiltonian cycles in \( k \)-uniform hypergraphs (e.g. [30, 37]) and takes a spectral approach: they introduce these generalized walks as a means for studying a generalized \( s \)-Laplacian matrix. Wang and Lee [64] define hypergraph paths as edge sequences in which no successive intersection is a subset of any other. Their motivation is to prove enumeration formulas for certain cycle structures in hypergraphs.

In a series of three recent papers [15, 16, 17], Kang, Cooley, Koch, and others consider a notion of \( s \)-walk between \( s \)-tuples of vertices. They conduct a rigorous mathematical analysis of the asymptotic \( s \)-walk properties of binomial random \( k \)-uniform hypergraphs, considering hitting times, the evolution of high-order \( s \)-components, and high-order “hypertree” structures. Lastly, in [33, 67], authors of the present work briefly considered the \( s \)-walk based notion of \( s \)-distance as applied to Domain Name System (DNS) cyber data, and the Enron email dataset, respectively.

Like all of the work above, the present work considers high-order hypergraph walks, but for different ends. Most importantly, our approach towards higher-order hypergraph walks differs in being geared first and foremost towards the application and analysis of real hypernetwork data. Accordingly (in contrast to all the work mentioned above, apart from that of Wang and Lee [64]) we do not assume \( k \)-uniformity, as real hypernetworks are frequently non-uniform. Furthermore, while our notion of \( s \)-walk is more general in the sense of applying to non-uniform hypergraphs, it is also simpler than some of those mentioned above in order to ensure our methods are more computationally tractable in light of the combinatorial explosion inherent in hypergraphs. For instance, the aforementioned work of Lu and Peng [47] defines \( s \)-walks between arbitrary ordered \( s \)-tuples of vertices; consequently, the \( s \)-Laplacian matrix they study is \( n^s \times n^s \), where \( n \) denotes the number of vertices and \( n^s = \binom{n}{s} k! \) denotes the falling factorial. Even for a modestly sized hypergraph on \( n = 10^4 \) vertices with \( s = 20 \), this matrix has size \( n^s \approx 10^{70} \), approximately the number of atoms in the known universe. Thus, while this matrix is amenable to theoretical analysis, computational methods are likely infeasible in practice. By defining our notion of \( s \)-walk between pairs of unordered hyperedges (or when working with the dual, pairs of single vertices), we develop \( s \)-walk based methods that are more tractable in application to real data.

In short, our goals for the present work are threefold: (1) to investigate how walk-based network science measures for graphs generalize to hypergraphs under the \( s \)-walk framework; (2) to apply these methods to real hypernetworks and briefly discuss insights they reveal; and (3) to experimentally test the degree to which existing generative hypergraph null models are able to replicate the properties seen in real data captured by these methods.

## 4 Hypergraph Walk Framework

In this section, we explore how analytic tools from network science extend to hypergraphs in the hypergraph walk framework. Within each subsection, we focus on one such particular topic (e.g. \( s \)-distance, a hypergraph geodesic), and introduce relevant methods in the “Methods” section. In the “Application to Data” section, we apply these methods to real hypernetworks, and briefly analyze the results. Here, we emphasize our goal is not to explain why the observed structure exists using domain-specific analyses. Rather, we identify abstract structural properties revealed by these measures, and highlight how these properties differ within each dataset (as we vary the walk order \( s \)), as well as across datasets. In doing so, we illuminate particular properties these measures capture, and point out additional insights revealed by considering \( s \)-walk based metrics. While we take a broader, methods-based viewpoint here, we believe such an approach may be more useful in guiding future application-specific studies of these methods across multiple domains. To that end, we consider three datasets from three domains: corporate governance, biology, and text analysis.

### 4.1 Test Data Sets

For each of our data sets, we briefly define the associated hypergraph, review prior graph or hypergraph analyses of these (or closely related) datasets, and discuss basic properties of the data, summarized in Figure 2. In this figure and throughout, we use the same notation as in Definition 4 to refer to the dual hypergraph associated with a data set (e.g. LesMis* refers to the dual hypergraph of LesMis, in which the roles of the vertices and edges are swapped relative to how they are defined below). Figure 2 presents plots of the edge cardinality distribution and pairwise edge intersection cardinality distribution. For instance, the point \((x, y) = (3, 100)\) on the “edges” distribution means there are 100 edges which consist of exactly 3 vertices in that hypergraph; the same point on the “pairwise edge intersection” distribution...
Figure 2: Plots: The edge and pairwise edge intersection distributions for LesMis, Diseasome, and CompBoard (top row) and their dual hypergraphs (bottom row). Table: Basic statistics on edge and multi-edge counts, the maximum edge and pairwise edge intersection cardinality, and density for each dataset and their dual hypergraphs.

Table:

|                    | LesMis | LesMis* | Diseasome | Diseasome* | CompBoard | CompBoard* |
|--------------------|--------|---------|-----------|------------|-----------|------------|
| # Edges            | 402    | 80      | 516       | 903        | 4,573     | 32,189     |
| # Multi-edges      | 211    | 5       | 108       | 422        | 0         | 22,537     |
| Max edge size      | 9      | 137     | 41        | 11         | 35        | 11         |
| Max edge inter.    | 8      | 31      | 5         | 4          | 12        | 5          |
| Density            | 2.68e-2 | 2.68e-2 | 3.33e-3  | 3.33e-3    | 2.67e-4  | 2.67e-4   |

means there are 100 distinct, unordered pairs of hyperedges whose intersection contains exactly 3 vertices. We remind the reader that the edge cardinality distribution of the dual hypergraph $H^*$ is the same as the vertex degree distribution of $H$.

The table in Figure 2 highlights some basic statistics for each hypergraph. The number of “multi-edges” is the number of edges that duplicate (in the sense of set equality) another edge, i.e. $\{|i : e_i = e_j \text{ for } j < i\}$. The maximum edge size and edge intersection sizes, reported below the multi-edge counts, are particularly pertinent because they determine the range of interest for our measures: the former determines the largest value of $s$ for which $s$-walk based measures are defined while the latter determines the maximum value of $s$ for which $s$-walk based measures are non-trivial. Finally, “density” measures the number of vertex-hyperedge memberships relative to the number of possible vertex-hyperedge memberships. Put equivalently, this is the number of nonzero entries in the incidence matrix $S$ divided by the product $|V| \cdot |E|$. Note that, by definition, density is always the same for the hypergraph and its dual, whereas the other reported values are edge-based and hence may differ.

CompBoard

Data set. A company-board network. Vertices represent board directors (i.e. members of the board), and hyperedges represent company boards. A vertex belongs to a hyperedge if that person sits on the company board. The data consists of 4,573 companies and 32,189 directors that sit on their boards. The companies are identified by their ticker symbols, excluding any location or exchange code suffixes (e.g. Vodafone group is represented solely by VOD, not VOD.L or VOD.O) and were taken from the NYSE, AMEX, and NASDAQ stock exchange listings on 10/1/2018. The data was collected from publicly available board director information listed on Reuters. Board director names were cross referenced against age data to better ensure that different people with the same full name were distinguished.

3 List of companies on these exchanges obtained from https://www.nasdaq.com/screening/company-list.aspx
4 https://www.reuters.com/finance/
Prior work. Company-board network studies are historically rooted in corporate elite theory, focusing on companies which share a common board member, called interlocking directorates. Many such studies focus on the line graph representations of the network, linking companies whose boards interlock. For instance, Conyon and Muldoon [14] studied the small-world properties of company-board networks from the US, UK, and Germany, focusing on the clustering coefficient and average path length of the line graphs. In [53], Newman compares the clustering coefficient of a company-board network line graph to that of a random model. Measuring hypergraph clustering on line graphs rather than the hypergraph has been rightly noted to be potentially misleading (see, for instance, [48, 54]) since the cliques generated by the line graph heavily skew the number of triangles. Levine and Roy [46] appear to be among the first to analyze bipartite representations of company-board networks directly, rather than solely line graphs. They considered topics such as the average path length, connected component sizes, and proposed a “rubber-band model” to cluster the bipartite network. Later, Robins and Alexander devised a bipartite global clustering coefficient, based on the ratio of bipartite 4-cycles to 3-paths, in order to measure “the extent to which directors re-meet one another on two or more boards” [58]. In Section 4.4, we propose a new notion of hypergraph clustering coefficients and explain how it compares to that of Robins and Alexander, as well graph clustering coefficients measuring on the line graph. More generally, since an “interlocking directorate” is represented by a hyperedge intersection, our methods can be interpreted in this context as not only based on the existence of interlocks (i.e. a pure line graph analysis) but also the size and relative set relationships of that interlock.

Basic properties. The edge size distribution shows that the sizes of company boards are tightly concentrated around 7-10 members and drop off sharply at either end: only about 3% of companies have fewer than 4 members, and 3% have more than 14. In contrast, the edge size distribution of the dual hypergraph is monotonically decreasing, showing more than 99% of board members belong to between 1-3 company boards. The pairwise edge intersection distribution for the hypergraph and its dual similarly exhibit a sharp decrease, and the range of these distributions imply that different companies share up to a maximum of 12 board members, while different members serve on up to maximum of 5 common company boards. Among the three datasets, CompBoard is the sparsest: it contains about 0.03% of possible vertex-hyperedge memberships, as opposed to 0.33% and 2.68% for Diseasome and LesMis, respectively.

Diseasome

Data set. A human gene-disease network from [27]. Vertices represent genes and hyperedges represent genetic disorders. A vertex belongs to a hyperedge if mutations in that gene are implicated by that disease. The data consists of 903 genes and 516 diseases.

Prior work. Goh et al. [27] collected the list of genes, disorders, and their associations from the Online Mendelian Inheritance in Man (OMIM) [29] compendium in 2005. Their study considered the line graphs of hypergraph and its dual, which they dubbed the Human Disease Network and Disease Gene Network. One of their structural observations was that the size of the largest connected component in these networks differed with those generated by random models. In Section 4.2 we study a generalized notion of high-order connected components and compare these against those of random hypergraph models in Section 5.1. For a broader discussion of the potential applications of hypergraphs and hypergraph statistics in biology and genomics, see [39].

Basic properties. The edge size distributions of Diseasome and its dual show that the most genes implicated by a disease is 41 while the most diseases implicated by a gene is 11. The pairwise edge intersection size distribution show that, among pairs of diseases that implicate a common gene, the 94% of such pairs share exactly 1 gene; conversely, examining the same distribution for the dual reveals that among genes associated with a common disease, 98% of such pairs are associated with exactly 1 common disease. Among the three datasets, Diseasome and its dual features the narrowest range of pairwise edge intersection sizes, with maximum edge sizes of 5 and 4, respectively.

LesMis

Data set. A character-scene network from [40]. Vertices represent characters and hyperedges represent scenes from Victor Hugo’s novel, Les Misérables. There are 80 characters and 402 scenes.

5 Described as a physical device consisting of two horizontal bars that support “hooks” representing companies and board member nodes, with rubber bands that “join the appropriate hooks and physically represent the inclusion between persons and boards”
Table 1: The s-line graphs for two hypergraphs.

| Hypergraph $H$ | $L_1(H)$ | $L_2(H)$ | $L_3(H)$ | $L_4(H)$ | $L_5(H)$ |
|----------------|-----------|-----------|-----------|-----------|-----------|
| ![Hypergraph](image) | ![Line graph](image) | ![Line graph](image) | ![Line graph](image) | ![Line graph](image) | ![Line graph](image) |

Prior work. This dataset was collected by Donald Knuth [10] and can be structured according to different granularities in which hyperedges represent the scenes, chapters, books, or volumes of the novel. The line graph of the LesMis hypergraph, often dubbed the Les Mis co-appearance network, has appeared frequently in network science literature for the purpose of demonstrating clustering or modularity methods [25, 52], or centrality and ranking methods [4]. With regard to the latter, we apply our proposed hypergraph centrality measure to rank LesMis characters in Section 4.3.

Basic properties. In LesMis and its dual, the largest hyperedge features 9 characters and 137 scenes, respectively, with the latter hyperedge (unsurprisingly) corresponding to the protagonist, Jean Valjean. Compared against other datasets, the edge intersection size distributions are particularly distinct. LesMis* features the largest range of edge intersection sizes across all datasets. Both LesMis and its dual are also notable for featuring the most edge intersections relative to the number of possible edge intersections: respectively, 22% and 8% of all pairs of edges in LesMis and its dual intersect, whereas this ratio is an order of magnitude smaller for Diseasome and its dual and two orders of magnitude smaller for CompBoard and its dual.

4.2 Connected Components

Methods. Under Definition 1, the graph notions of connectedness and connected components extend naturally to the s-walk framework.

Definition 6. For a hypergraph $H = (V, E)$, a subset of hyperedges $C \subseteq E$ is called s-connected if there exists an s-walk between all $f, g \in C$ and is further called an s-connected component if there is no s-connected $J \subseteq E$ such that $C \subseteq J$.

Since for any $e \in E$, there can be no s-walk from $e$ to any other hyperedges if $|e| < s$, the order of an s-connected component is bounded above by $|E_s|$, where $E_s = \{e \in E : |e| \geq s\}$. More precisely, for any positive integer $s$ and hypergraph $H$, the edges in $E_s$ can always be partitioned into s-connected components. We call a hypergraph s-connected if $E_s$ is s-connected.

Observe that, while an s-connected component of a hypergraph $H$ is an equivalence class of edges, a vertex-based notion of s-connected components is obtained by simply applying the above definition to the dual hypergraph $H^*$. In comparing these edge and vertex-based notions, note that the total number of 1-connected components for $H$ and $H^*$ are always the same: it is straightforward to see that, in either case, the number of such components is the same as the number of nontrivial connected components (i.e. excluding isolated vertices) in the bipartite graph representation of the hypergraph. In this sense, edge and vertex-based connectedness are equivalent for $s = 1$ and whenever $H$ is a graph. However, for $s \geq 2$, the number of s-connected components for $H$ and $H^*$ may differ. Hence, s-connectedness in hypergraphs is richer and more varied for high-orders, yielding dual but distinct vertex and edge-based notions.

An effective way of visualizing and studying basic properties of s-connected components is via its s-line graph. As previously mentioned, s-line graphs were studied for k-uniform hypergraphs by Bermond, Heydemann, and Sotteau [9] as early as 1977. A definition for the general case may be stated as follows:

Definition 7. Let $H = (V, E)$ be a hypergraph with vertex set $V = \{v_1, \ldots, v_n\}$ and edge set $E \supseteq E_s$ where $E_s = \{e \in E : |e| \geq s\} = \{e_1, \ldots, e_k\}$ for an integer $s \geq 1$. The s-line graph of $H$, denoted $L_s(H)$, is the graph on vertex set $\{e_1^*, \ldots, e_k^*\}$ and edge set $\{|e_i^* \cap e_j^*| \geq s \text{ for } i \neq j\}$.

In other words, each vertex in the s-line graph represents a hyperedge with at least $s$ vertices in the hypergraph, and two vertices are linked in the s-line graph if their corresponding hyperedges intersect in at least $s$ vertices in the hypergraph. In this way, the connected components of the s-line graph
Table 2: The $s$-connected components of the datasets. For a full-sized image, see Appendix A.

|       | 1-components | 2-components | 3-components | 4-components | 5-components |
|-------|--------------|--------------|--------------|--------------|--------------|
| LesMis* | ![Image] | ![Image] | ![Image] | ![Image] | ![Image] |
| Diseasome | ![Image] | ![Image] | ![Image] | ![Image] | ![Image] |
| CompBoard | ![Image] | ![Image] | ![Image] | ![Image] | ![Image] |

represent the $s$-connected components of the hypergraph, and the 1-line graph is simply the line graph from Definition 4. In Table 1 we give an example of two hypergraphs and their associated $s$-line graphs. Observe that both hypergraphs have identical 1-line graphs. Nonetheless, comparing their $s$-line graphs for $s = 2, 3, 4$ suggests differences otherwise lost when solely considering the (usual) line graph.

Although more general, $s$-line graphs are still subject to the same limitations underlying (the usual) hypergraph line graphs, and do not uniquely identify a hypergraph, up to isomorphism. Nevertheless, $s$-line graphs can be utilized to determine a number of $s$-walk properties, including $s$-distance, which we explore in the next section. It is worth stressing, however, that the study of high-order $s$-walks in hypergraphs is not limited to studying $s$-line graphs. As we will see in Section 4.4, $s$-line graphs cannot be used to distinguish between finer classes of $s$-walks, such as $s$-meanders and $s$-paths, and consequently cannot be used to compute $s$-clustering coefficients, for example.

**Application to Data** In Table 2 we visualize the $s$-components of the LesMis, Diseasome, and Company Board datasets for $s = 1, \ldots, 5$ by plotting their $s$-line graphs. A page-sized version of this table is included in Appendix A.

A number of qualitative differences are readily apparent from the visualization. In the case of LesMis, we observe the majority of hyperedges are contained within a giant component for $s = 1, \ldots, 5$. This means that one can link the majority of characters with each other via a pathway of characters co-occurring in at least one to five scenes together. In the case of Diseasome, for $s = 1$ we similarly observe a giant component; however, for $s \geq 2$, this giant component fragments into small, roughly equally sized components. In this case, as $s$ increases from one to five, many of the shared-gene pathways linking diseases for $s = 1$ break down. By $s = 5$, the $s$-components consist almost entirely of isolated hyperedges (apart from a single pair of closely related diseases, “Diabetes Mellitus” and “Mature onset diabetes of the young (MODY)” meaning that diseases associated with 5 or more genes do not share 5 or more of those genes with other disorders. The most dramatic fragmentation occurs for the CompBoard dataset. For $s = 1$, we note that 74% of the companies are contained within the giant component (pictured in the lower-left hand corner), while for $s = 2$, this drops to 0.5%. This affirms that shared board-member pathways linking companies almost always rely on single shared board members.

In order to quantify these observations about the changes in $s$-connected component sizes more rigorously and completely, we compute several entropy-based measures on the underlying $s$-connected component size probability distributions. The $s$-connected component size probability distribution, $p_s = (p_{s1}, \ldots, p_{sk})$, is defined by taking $p_j$ to be the fraction of hyperedges in $E_s$ that are in the $j$'th $s$-component.

For a discrete probability distribution $p$, its Shannon entropy is given by $H(p) = -\sum_{i=1}^{k} p_i \log_2(p_i)$. However, direct comparisons of Shannon entropy on our data may be problematic, as we note the number of hyperedges in $E_s$ and number of $s$-components varies not only between datasets, but also as $s$ varies within each dataset, thereby complicating cross-comparisons of the (unitless) Shannon entropy. In order
Figure 3: The $p$-smoothness (normalized entropy) and dispersion values of the $s$-connected components for LesMis, Diseasome, and CompBoard (top row) and their dual hypergraphs (bottom row).

To facilitate a more meaningful entropy comparison, we consider the normalized entropy

$$H(p) = \frac{H(p)}{\log_2(k)} \in [0, 1],$$

called $p$-smoothness by the authors [34]. Note that $H$ achieves its maximum value of $\log_2(k)$ for the uniform distribution, $\langle 1/k, \ldots, 1/k \rangle$, and a minimum value of 0 for a fully skewed distribution, e.g. $\langle 0, \ldots, 0, 1 \rangle$. For the special case in which $k = 1$, one takes the limiting definition as $k \to 1$, and defines $H(p) := 1$.

We consider the $p$-smoothness of the $s$-component size distribution $p_s$, which we denote $\tilde{H}_s = \tilde{H}(p_s)$. In our context, if all $s$-components are equally sized, then $\tilde{H}_s$ is 1, whereas if the disparity between component sizes is maximal (e.g. $|E_s| - 1$ hyperedges in one $s$-component, and 1 hyperedge in the other), then $\tilde{H}_s$ approaches 0. In this sense, $p$-smoothness reflects how smooth or uniform the $s$-component sizes are, but may not reflect how dispersedly the hyperedges are distributed among $s$-connected components.

For that purpose, we consider an additional measure, again from [34], aptly called dispersion. In our context, the dispersion of the $s$-component size distribution compares the number of $s$-components to the number of possible $s$-components on a logarithmic scale, i.e.

$$D_s = \frac{\log_2(|C_s|)}{\log_2(|E_s|)} \in [0, 1],$$

where $C_s$ denotes the set of $s$-connected components and $E_s$ denotes the set of $s$-hyperedges.

In Figure 3 we plot the $p$-smoothness and dispersion for $s = 1, \ldots, s_{\text{max}}$, where $s_{\text{max}} = \max_{f,g \in E} |f \cap g|$. For $s > s_{\text{max}}$, the $s$-components are either all isolated hyperedges, or non-existent. In all datasets, we observe that both dispersion and $p$-smoothness tend to increase in $s$, although, as evident from LesMis, this increase is not always monotonic. In the case of LesMis*, we observe lower values of $p$-smoothness for each of $s = 1, \ldots, 5$ relative to those for corresponding values of $s$ in the other datasets, consistent with the highly skewed distribution associated with the large component we observed in the visualization. For CompBoard, we observe a large separation between between $p$-smoothness and dispersion for $s = 1$. In this case, while the component size distribution is still skewed – and hence has low $p$-smoothness – the remaining $s$-components consist of many isolated hyperedges, reflected in the high dispersion value. Lastly, for the Diseasome dataset, $p$-smoothness is maximal while dispersion is minimal for $s = 1$, while for $s \geq 2$ both $p$-smoothness and dispersion closely coincide at values near 1. This reflects the fragmentation of a single giant component into many $s$-components (hence the high dispersion) that are equally sized (hence the high $p$-smoothness).

4.3 Distance and Centrality

Methods Under Definition 1 it is straightforward to show the length of the shortest $s$-walk serves as a distance metric function over a set of hyperedges. More precisely
Figure 4: The maximum $s$-diameter over all $s$-components and average $s$-efficiency for LesMis, Diseasome, and CompBoard (top row) and their dual hypergraphs (bottom row). Note the $s$-diameter values are in linear scale whereas the $s$-efficiency values are in logarithmic scale.

**Proposition 1.** Let $H = (V, E)$ be a hypergraph and $E_s = \{ e \in E : |e| \geq s \}$. Define the $s$-distance function $d_s : E_s \times E_s \to \mathbb{Z}_{\geq 0}$ by

$$d_s(f, g) = \begin{cases} 
\text{length of the shortest $s$-walk} & \text{if an $s$-walk between $f, g$ exists} \\
\infty & \text{otherwise}
\end{cases}.$$

Then $(E_s, d_s)$ is a metric space.

We omit the proof, as the triangle inequality can proved constructively, and the other metric space axioms follow immediately from Definition 8. If the hypergraph $H$ is 2-uniform (i.e. is a graph), then the graph distance between vertices $x$ and $y$ in $H$ is equivalent to the 1-distance between hyperedges $x^*$ and $y^*$ in $H^*$. With $s$-distance serving as hypergraph geodesic distance, hypergraph $s$-analogs of local and global distance-based graph invariants easily extend.

**Definition 8.** Let $H = (V, E)$ be a hypergraph.

(i) The $s$-eccentricity of a hyperedge $f$ is $\max_{g \in E_s} d_s(f, g)$.

(ii) The average $s$-distance of $H$ is $(|E_s|^{-1}) \sum_{f, g \in E_s} d_s(f, g)$.

(iii) The $s$-closeness centrality of a hyperedge $f$ is $\frac{|E_s|^{-1}}{\sum_{g \in E_s} d_s(f, g)}$.

Important caveats arise when applying Definition 8 to real data. As we observed in the previous section, it may often be the case that $H$ contains more than one $s$-component for some values of $s$, in which case the $s$-distance between some pairs of edges is infinite. Consequently, the $s$-eccentricity of every edge (and hence $s$-diameter and $s$-radius) and mean $s$-distance are all infinite; similarly, the $s$-closeness centrality of every edge is trivially 0. Similar to how these issues are sometimes addressed for graphs, one alternative may be to compute these measures on only the largest $s$-component. Depending on the analyst’s aims, such an approach might be satisfactory, particularly if the majority of hyperedges in $E_s$ are contained within the largest $s$-component, as was seemingly the case in LesMis$^*$. However, restricting to the largest component may be unsatisfactory in cases where the largest $s$-component does not constitute the overwhelming majority of edges in $E_s$, as we observed in the Company Board network for $s \geq 2$. In such cases, one may wish to compute $s$-eccentricity on a per-component basis, taking the extrema over all $s$-components as the $s$-diameter and $s$-radius. One may similarly compute mean $s$-distance or $s$-closeness per-component, however, it is unclear how to properly synthesize these values in order to obtain (in the former case) a single global numerical measure or (in the latter case) a ranking over all hyperedges in the entire network. Instead of a per-component approach, an elegant
alternative for averaging graph distances in disconnected graphs, advocated by Newman [51], is to use the harmonic mean instead of the arithmetic mean. This approach was adopted by Latora and Marchiori [45] as the reciprocal of the harmonic mean path length, proposed as a quantitative measure of small-worldness. Latora and Marchiori termed this measure “efficiency”, suggesting an alternative for averaging graph distances in disconnected graphs, advocated by Newman [51], to define network efficiency as the reciprocal of the harmonic mean path length, proposed as a quantitative measure of small-worldness. Latora and Marchiori termed this measure “efficiency” in reference to how efficiently information might be exchanged over the network. Later, a similar approach was adopted by Rochat [50] to define the harmonic closeness centrality index of vertices in a disconnected graph. Extending these notions to the hypergraph context, a more practical definition of the aforementioned s-distance based notions is given by:

Definition 9. Let $H = (V, E)$ be a hypergraph and let $C_s$ denote the set of its $s$-connected components.

(i) The $s$-eccentricity of a hyperedge $f \in C$ where $C \subseteq C_s$ is $\max_{g \in C} d_s(f, g)$.

- The $s$-diameter is the maximum $s$-eccentricity over all edges in $E_s$, while the $s$-radius is the minimum.

(ii) The average $s$-efficiency of $H$ is $\frac{1}{|E_s|} \sum_{f,g \in E_s} \frac{1}{s(f,g)}$.

(iii) The harmonic $s$-closeness centrality index of a hyperedge $f$ is $\frac{1}{|E_s|} \sum_{g \in E_s} \frac{1}{s(f,g)}$.

In this way, we take the limiting value of 0 for the summand in (ii) and (iii) when $f$ and $g$ are in different $s$-components. Both (ii) and (iii) are numerical quantities between 0 and 1, with larger values indicative of closer $s$-distances between hyperedges, either globally (in the former case) or locally (in the latter case). Next, we apply the notions in the above definition to real data.

Application to Data 

Turning our attention to the data, we compute three of the aforementioned $s$-distance-based measures: the average $s$-efficiency index, the $s$-diameter (i.e., the maximum $s$-eccentricity over all $s$-components) and the harmonic $s$-closeness centrality index.

In Figure 5, we plot the maximum $s$-diameter (over all $s$-components) and the average $s$-efficiency for the hypergraph and its dual of each of our datasets. We recall that, by definition, larger values of average $s$-efficiency imply smaller $s$-distances among the hyperedges in question. For a number of the datasets (e.g., LesMis, Diseasome, Diseasome*) we observe that average $s$-efficiency and $s$-diameter tends...
to decrease as $s$ increases. In these networks, the shortest $s$-walks linking hyperedges tend to become longer (or infinite) as $s$ is increased. However, in the case of CompBoard*, we observe that average $s$-efficiency increases in $s$ for each $s \geq 2$. This suggests that, among company board members who sit on multiple boards, those who sit on more boards tend to (on average) be closer to one another in $s$-distance. We observe a similar phenomena regarding average $s$-efficiency for the LesMis* dataset, where for characters appearing in at least $s \geq 22$ scenes, the more scenes they appear in, the closer they are to each other in $s$-distance.

Turning our attention to $s$-diameter, we remark it is possible for $s$-diameter (taken as the maximum over all $s$-components) to increase or decrease in $s$. In the former scenario, as $s$ is increased, shorter $s$-walks linking hyperedges may disappear, and those edges may only be linked via longer $s$-walks, thereby increasing $s$-diameter. We observe this most prominently for LesMis, where $s$-diameter increases from 5 to 9 as $s$ increases from 1 to 2. On the other hand, if increasing $s$ eliminates all $s$-walks between pairs of hyperedges, then these hyperedges are separated into different $s$-components in which hyperedges may be closer to each other. In such cases, one may observe the $s$-diameter decrease, as in Diseasome. Consistent with our intuition from the Diseasome visualization in Figure 2, this $s$-diameter drop reflects the fragmentation of the network into small components; accordingly average $s$-efficiency also drops because of the infinite $s$-distances between edges in different $s$-components.

Lastly, in Figure 5 (top row), we list the top 15 hyperedges in the CompBoard, Diseasome, and LesMis* datasets for $s = 1, 2, 3$, as ranked according to their harmonic $s$-closeness centrality. Boxes enclosing hyperedges indicate a tie in $s$-closeness centrality. Comparing the ordinal rankings across the datasets, we make several observations. In some cases the top 15 ranked hyperedges for $s = 1$ remain within the top ranked for $s = 2$ (e.g. for LesMis*, 10 remain in the top 15) whereas in other cases, the top ranked hyperedges may change completely (e.g. in CompBoard, none of the top 15 companies with highest 1-closeness centrality remain in the top 15 for 2-closeness centrality).

A drop in a hyperedge's rank from $s = 1$ to $2$ may indicate that short pathways linking that hyperedge to others rely on sparse hyperedge intersections. To illustrate this with an example, in Diseasome we observe that while “Colon cancer” and “Breast cancer” remain in the top 3 ranked hyperedges for $s = 1, 2, 3$, “Diabetes mellitus” drops from having the second largest 1-centrality, to having the 34th largest 2-centrality. “Diabetes mellitus” shares genes with 24 other diseases and hence this hyperedge intersects with 24 other hyperedges. However, of these 24 diseases, “Diabetes mellitus” shares at least 2 genes with only two diseases: “Obesity” and “Mature Onset Diabetes of the Young (MODY)”. Thus, any 2-walk between “Diabetes mellitus” and another disease can only go through one of these diseases, which (in this case) results in larger average 2-distance between diabetes and other diseases, relative to the average 2-distance between other pairs of diseases. In contrast, “Breast cancer” shares at least 2 genes with 9 other diseases, and (on average) can be linked to other diseases via a shorter 2-walk than for “Diabetes mellitus”.

In order to more rigorously explore these the changes in ordinal rankings by $s$-closeness, we compute Kendall’s $\tau_B$ rank correlation coefficient between the top $k$ ranked hyperedges for one value of $s$ and the rankings of those same hyperedges under another value of $s$. We compute this coefficient for each of $k = 10, \ldots, |E|$ and for each ordered pair of $s$-values from $\{1, 2, 3\}$. Hyperedges with equal $s$-closeness centrality are considered tied in rank, and we assign the minimum $s$-closeness centrality score of 0 to any hyperedge with fewer than $s$ vertices. We remind the reader that values of Kendall’s $\tau_B$ range from -1 (if the ordinal rankings are perfectly inverted) to 1 (if the ordinal rankings are identical), and that that Kendall’s $\tau_B$ is explicitly formulated to handle ties in rank [1]. We plot the results in Figure 5 for the CompBoard, Diseasome, and LesMis* datasets. For CompBoard, we see an absence of correlation for the 1-closeness rankings when compared against 2 or 3, and a stronger correlation for the 3-closeness rankings compared against the 2-closeness rankings. When all hyperedges in the network are considered (i.e. for $k = |E|$, given by the rightmost points in each plot), the 1-closeness rankings of LesMis* exhibits the strongest correlations between the 2 and 3-closeness rankings.

### 4.4 Paths, Cycles, and Clustering Coefficients

**Methods** So far, we’ve centered our methods solely around the base definition of an $s$-walk. However, just as graph walks may be distinguished into finer classes such as trails, paths, circuits and cycles, $s$-walks may also be distinguished from each other and organized in a hierarchical way. As we will soon see, doing so allows us to distinguish between high-order substructures native to hypergraphs, such as $s$-triangles, that cannot be determined from their $s$-line graphs.

**Definition 10.** For a hypergraph $H = (V, E)$, let the sequence of hyperedges $\omega = (e_i_0, e_i_1, \ldots, e_i_k)$ be an $s$-walk of length $k$. For ease of notation let $I_j = e_{i_{j-1}} \cap e_{i_j}$ be the $j$’th intersection. The $s$-walk $\omega$ may
be further defined as:

(i) An s-trace if \( i_x \neq i_y \) for all \( x \neq y \) (all hyperedges are pairwise distinct by label).

(ii) An s-meander if \( \omega \) is an s-trace in which \( I_x \neq I_y \) for all \( x \neq y \) (all intersections are pairwise distinct).

(iii) An s-path if \( \omega \) is an s-meander in which \( I_x \setminus I_y \neq \emptyset \) for all \( x \neq y \) (no intersection is included in another).

If the hypergraph is 2-uniform (i.e. is a graph), the s-meander condition requiring successive edge intersections to be pairwise distinct is equivalent to the s-path condition requiring no intersection is a subset of any other. Hence, s-meanders and s-paths are equivalent for graphs, and both reduce to the usual graph path. We note that Wang and Lee [14] also define hypergraph paths using the same subset condition stated in [10] above.

Observe that the notions of s-walk, s-trace, s-meander, and s-path form a nested hierarchy: every s-trace, s-meander, or s-path is an s-walk; every s-meander and s-path is an s-trace; and every s-path is an s-meander. However, in each case, the reverse may not be true (e.g. an s-meander may not be an s-path). With regard to s-distance (Section [4.3]), it is straightforward to show constructively that if there exists an s-walk (resp. s-trace, s-meander) of length \( k \) between two hyperedges, there exists an s-trace (resp. s-meander, s-path) of length at most \( k \). This implies the length of the shortest s-walk between two hyperedges is equivalent to the length of the shortest s-path; consequently, s-distance as given by the length of the shortest s-walk is equivalent to that given by the length of the shortest s-path.

While not having ramifications for the notion of s-distance, the finer classes of s-walks above provide a means, within the s-walk framework, to define high-order substructures or motifs that cannot be determined from the s-line graph. To define an example of these substructures, we require the notion of a closed walk. Analogous to its usage in graph theory, we call an s-walk is closed if \( i_0 = i_n \), and call a closed s-path an s-cycle. As a point of clarification, closed s-traces, meanders, or paths are still considered valid s-traces, meanders or paths (that is, only the terminal edges are exempt from the s-trace requirement that all edges be distinct by label). Using s-cycles, we define hypergraph s-analogs of triadic measures commonly applied to graph data. We note that whereas graph triadic notions like the local clustering coefficient [56] are defined for vertices, the s-analogs below are defined for hyperedges, keeping consistent with the rest of our presentation. We remind the reader that vertex-based notions are obtained by simply applying the below definition to the dual hypergraph, \( H^* \).

**Definition 11.** For a hypergraph \( H \), an s-triangle is a closed s-path of length 3 and an s-wedge is an s-path of length 2. For an s-wedge \( e_0, f, e_2 \), we say \( f \) is the center of the s-wedge.

(i) The s-local clustering coefficient of a hyperedge \( f \in E_2 \) is given by

\[
\text{s-LCC}(f) = \begin{cases} 
\frac{\text{number of s-triangles containing } f}{\text{number of s-wedges centered at } f} & \text{if } f \text{ is the center of an s-wedge} \\
0 & \text{otherwise}.
\end{cases}
\]

(ii) The s-global clustering coefficient of a hypergraph \( H \) is given by

\[
\text{s-GCC}(H) = \frac{3 \cdot \text{total number of s-triangles}}{\text{total number of s-wedges}}.
\]

In the same way as for the LCC of graphs, one may obtain a global measure for the s-LCC of a hypergraph by taking the mean s-local clustering coefficient over all edges in \( E_2 \). If \( H \) is a graph (i.e. is 2-uniform), then the above triadic notions reduce to their usual graph counterparts on \( H^* \) (e.g. the 1-LCC of hyperedge \( x^* \) in \( H^* \) is equivalent to the usual LCC of vertex \( x \)).

In Figure [6] we give examples of three different hypergraphs induced by a closed s-walk \( e_1, e_2, e_3, e_4 \) of length 3; namely, in [6a] a closed s-trace that is not an s-meander, in [6b] a closed s-meander that is not an s-path, and in [6c] a closed s-path of length 3 (i.e. an s-triangle). Observe that the 1-line graphs of all three of these hypergraphs consists of a single triangle, while only the hypergraph in [6c] is a 1-triangle (as well as a 2-triangle). This illustrates that s-triangles cannot be determined from line graphs (a fact that is unsurprising, given that the s-line graphs do not encode the subset relationships stipulated for s-paths).

Given that many other definitions of hypergraph clustering coefficients have appeared in the complex networks literature, it is worth clarifying how these notions compare to ours. Estrada [23] proposes
a global hypergraph clustering coefficient as a ratio of (non-\textit{s}-walk based) hypergraph triangles to hypergraph wedges. More precisely, Estrada defines a hypertriangle as an alternating vertex-hyperedge sequence with three distinct vertices and three distinct hyperedges such that for each subsequence \( v_i, e_k, v_j \), we have that \( v_i, v_j \in e_k \) (put equivalently, these are 6-cycles in the bipartite representation of the hypergraph). Thus, returning to the example in Figure 6c, the alternating sequence given by interlacing the pair of vertex and hyperedge triples \((v_1, v_3, e_6)\) and \((e_1, e_2, e_3)\) constitutes a triangle, as does the same pair with \( v_1 \) replaced with \( v_2 \). Indeed, it is easy to see that the existence of an \( s \)-triangle implies the existence of at least one such hypertriangle as defined by Estrada; however, the converse is not necessarily true (e.g. while the hypergraph in Figure 6b contains many such hypertriangles, neither this hypergraph nor its dual contain any \( s \)-triangles). In this sense, Estrada’s notion of clustering and ours are fundamentally different.

Other proposed notions of hypergraph clustering differ to ours in being based on averaging various \textit{pairwise} set theoretic measures between pairs of hyperedges or vertices. For instance, Latapy, Magnien and Vecchio [14] propose a pairwise clustering coefficient between hyperedges \( e_i, e_j \) as \( \frac{|e_i \cap e_j|}{|e_i| |e_j|} \). In other words, this is the Jaccard similarity coefficient between the sets of vertices constituting the two hyperedges (or, when applied to the dual, the Jaccard similarity between the sets of hyperedges to which two vertices belong). They then define a local and global notion of hypergraph clustering by averaging this quantity. Zhou and Nakhleh [68] propose local and global hypergraph clustering coefficients based on the pairwise \textit{excess overlap} between hyperedges. As described and studied further by the authors in [20], excess overlap measures the proportion of the vertices in exactly one of the edges that are neighbors of vertices in only the other edge. Lastly, notions of bipartite graph clustering proposed in the literature, (which via the bicolored graph-hypergraph correspondence mentioned in Section 2 apply naturally in the hypergraph setting) are frequently based on bipartite 4-cycles [2, 58]. In the language of hypergraphs, a bipartite 4-cycle is a subhypergraph on two hyperedges and two vertices. Hence (in addition to again not being based in high-order \textit{s}-walks) these bipartite 4-cycle based notions of clustering differ from our \textit{s}-triangle based notions in involving only pairs (rather than triples) of hyperedges.

**Application to Data** In Figure 7 we plot the mean \( s \)-LCC and \( s \)-GCC (left block) as well as the proportion of triangles and wedges in \( s \)-line graph that correspond to \( s \)-triangles and \( s \)-wedges in the hypergraph (right block) for each of our datasets. We recall that every triangle and wedge in the \( s \)-line graph represents a closed \( s \)-walk of length 3 and \( s \)-trace of length 2; which, in turn, may or may not be an \( s \)-triangle or \( s \)-wedge, respectively. For all three datasets, we observe that a higher proportion of wedges in the \( s \)-line graph correspond to \( s \)-wedges compared with the proportion of triangles in \( L_s(H) \) that correspond \( s \)-triangles.

On the other hand, the datasets exhibit different behavior regarding the absolute size of these proportions, as well as how these proportions vary as \( s \) varies. For LesMis\textsuperscript{*}, a relatively large proportion of triangles in \( L_s \) correspond to \( s \)-triangles than for CompBoard\textsuperscript{*}. Furthermore, the proportions of \( s \)-triangles to \( s \)-wedges are much greater, both on average locally (given by the mean \( s \)-LCC) as well as globally (given by the \( s \)-GCC) than for CompBoard\textsuperscript{*}. In contrast, CompBoard\textsuperscript{*} exhibits an extremely small proportion of triangles in \( L_s(H) \) corresponding to \( s \)-triangles. This means that whenever there is a triad of board members such that each pair of members belong to common company boards, it is almost always the case that for at least one pair of board members, the set of companies in common are either identical (i.e. forming \( s \)-trace that is not an \( s \)-meander) or subsets of each other (i.e. an \( s \)-meander that is not an \( s \)-path). Indeed, \( s \)-triangles in both CompBoard and its dual are scarce, which is reflected in the extremely low \( s \)-LCC and \( s \)-GCC coefficients. Here, it is worth mentioning that (in the context
Figure 7: The mean local and global $s$-clustering coefficients of each hypergraph (left block), and proportion of triangles and wedges in the $s$-line graph that correspond to $s$-triangles and $s$-wedges for LesMis, Diseasome, and CompBoard, and their dual hypergraphs.

of a company-board network) an $s$-wedge is a generalization of a different representatives’ interlock, a topic which features prominently in the corporate governance literature [5, 46, 58]. Furthermore, pairs of hyperedges having an $s$-distance of 1 and 2 represent so-called “direct interlocks” and “third company interlocks,” respectively, between competing companies. This parallel illustrates how $s$-path and $s$-distance based notions may provide a generalized framework for describing and measuring phenomenon already important to particular domains. In the case of CompBoard, given that the aforementioned interlocks between competing companies are regulated by Section 8 of the Clayton Act [5], it is unsurprising that $s$-wedges (and hence $s$-triangles) are relatively rare.

5 Comparison with Generative Hypergraph Null Models

Graph generation serves far-ranging purposes across scientific disciplines. In particular, generative graph models are used for benchmarking, algorithm testing, and creating surrogate graphs to protect the anonymity of restricted data. Here, we apply hypergraph generative models as null models to experimentally test the significance of the high-order properties explored in Section 4. By “null model”, we mean a generative model that controls for certain basic features of the data. Such models may be utilized to test whether observed measurements in the data are necessarily consistent with the controlled features. For example, in the Erdős-Rényi graph model, the user specifies the desired number of vertices $n$ and edge-probability $p$; hence, by controlling $n$ and $p$ one can generate ensembles of random graphs with the same expected edge density. A subsequent comparison of measurements on given graph data against those on random graphs with the same edge density tests whether the measured features can be explained as sole consequences of edge density. To the extent to which the properties of the real and synthetic graphs diverge, the measured features are deemed significant.

---

6 Defined in [5] as “the linking of two companies by a third company having different representatives on the board of the two companies.”

7 Defined in [5] as “the linking of two companies by a third company having different representatives on the board of the two companies.”

8 Defined in [5] as “the linking of two companies by one company having a director on the board of a second, ... which has directors in common with a third company, ... which in turn has directors on the board of a competitor of the first company.”
graphs tend to diverge, this provides evidence that the properties under question cannot be explained as sole consequences of the structural properties preserved by the model.

In comparison to their graph counterparts, generative hypergraph models are relatively few. Nonetheless, researchers have recently begun developing a wider variety of hypergraph models, both for the case of uniform hypergraphs \cite{15, 16, 17, 55} and non-uniform hypergraphs \cite{11, 14, 20, 26, 35}. In the present work, we consider three generative hypergraph models from \cite{2}, which can be thought of as hypergraph interpretations of the graph models Erdős-Rényi (ER) \cite{22}, Chung-Lu (CL) \cite{13}, and Block Two-Level Erdős-Rényi (BTER) \cite{11, 63}. These models were originally presented as “bipartite models” in \cite{2}, with similar acknowledgment of the bicolored graph-hypergraph correspondence we discussed in Section 2. While these models were inspired from their graph counterparts and named as such, there may be multiple ways of conceiving these models in the hypergraph/bicolored graph setting, as is often the case with graph-to-hypergraph extensions. In fact, others have proposed non-uniform hypergraph analogs of Erdős-Rényi and Chung-Lu (see \cite{20} and \cite{35}, respectively) differing to those considered here with regard to the inputs required, the model itself, and the underlying definition of hypergraph assumed.

We’ve chosen these particular models for several reasons. First, these models can generate non-uniform hypergraphs in accordance with the full generality of Definition 1. Notably, all three of these models can generate hypergraphs with duplicated edges, which occur frequently in hypergraph-structured data and are highly prevalent on our particular data (see Figure 2). This is in contrast to the non-uniform hypergraph ER and CL models proposed by \cite{20} and \cite{35}, which do not permit duplicated hyperedges, but treat hyperedges themselves as multisets. Secondly, taken as a suite, these models have the advantage of providing tiered control over three fundamental properties: (1) vertex-hyperedge density, (2) vertex degree and edge cardinality distributions, and lastly, (3) metamorphosis coefficients, a measure of community structure from \cite{2} which we comment on further shortly. More specifically, ER controls for vertex-hyperedge density, CL controls for density as well as degree distributions, and BTER controls for all three of the aforementioned properties. In fact, taken in sequence, ER, CL and BTER can each be conceived formally as a generalization of the previous model. Lastly, all three models afford scalable implementations and \cite{2} \cite{81} report results on hypergraphs generated using these models with hundreds of millions vertex-hyperedge memberships. We note that open source implementations of all three of the above generative hypergraph models are available as part of The Chapel HyperGraph Library (CHGL, \cite{31}), a prototype HPC library \cite{82} for large-scale hypergraph generation and analysis written in the emerging programming language of Chapel.

5.1 Three Generative Hypergraph Models

We define each of the generative models we consider below. Then, we briefly discuss and compare the model properties.

1. Erdős-Rényi, ER(n, m, p). The user specifies three scalar parameters: the desired number of vertices n, desired number of hyperedges m, and vertex-hyperedge membership probability, p \in [0, 1]. For each of the nm vertex-hyperedge pairs, the probability of membership is the same,

\[
Pr(v \in e) = p.
\]

2. Chung-Lu, CL(\(\vec{d}_v, \vec{d}_e\)). The user specifies a desired vertex degree sequence \(\vec{d}_v = (d_{v_1}, \ldots, d_{v_n})\) and desired hyperedge size sequence \(\vec{d}_e = (d_{e_1}, \ldots, d_{e_m})\), which (in order to be realizable by a hypergraph) must satisfy \(c = \sum_{i=1}^{n} d_{v_i} = \sum_{i=1}^{m} d_{e_i}\). The probability that a vertex belongs to a hyperedge is proportional to the product of the desired vertex degree and edge size, i.e.

\[
Pr(v_i \in e_j) = \frac{d_{v_i} \cdot d_{e_j}}{c}.
\]

In order to ensure this probability is always less than 1, one may further require the input sequences satisfy \(\max_{i,j} d_{v_i} d_{e_j} < c\).

3. Block Two-Level Erdős-Rényi, BTER(\(\vec{d}_v, \vec{m}_v, \vec{m}_e\)). In addition to the desired vertex degree and edge size sequences mentioned in Chung-Lu, the user also specifies desired vertex and edge metamorphosis coefficients, \(\vec{m}_v\) and \(\vec{m}_e\), which, as clarified further in the discussion below, are measures of community structure based on the prevalence of small, dense substructures in the hypergraph. The BTER model is designed to output a hypergraph that matches the input degree

\[https://github.com/pnnl/chgl\]
distribution and metamorphosis coefficients. The BTER model proceeds in two phases: in the first, metamorphosis coefficients are approximately matched by grouping vertices and hyperedges into small, disjoint sets called affinity blocks and applying the Erdős-Rényi model on each block. In the second, the degree distributions are matched by running the Chung-Lu model on the excess desired degrees, thereby linking the blocks. As formal details of the BTER model are complicated, the reader is referred to [2] for a complete specification.

Before proceeding, we briefly discuss the three models. For the ER model, the expected number of vertex-hyperedge memberships is $p_{nm}$, and hence this simple model can be used to generate random hypergraphs with a specified vertex-hyperedge membership density. We reported this density for our datasets in Figure 2. For the CL model, each vertex $v$ achieves its user-specified desired degree $d_v$ in expectation since

$$E(\text{deg}(v)) = \sum_e \Pr(v \in e) = \frac{d_v}{c} \sum_e d_e = d_v.$$  

An identical argument also shows each hyperedge $e$ achieves its desired size $d_e$ in expectation. In this way, CL not only matches the desired vertex-hyperedge membership density in expectation like ER, but additionally matches the vertex degree and edge size distributions in expectation. We reported these degree distributions for our datasets in Figure 2.

The CL model is a generalization of the ER model in the sense that the ER can be obtained from CL by taking the degree and edge size sequences to be constant, i.e.

$$\text{CL}(\underbrace{mp, mp, \ldots, mp}_{n \text{ times}}, \underbrace{np, np, \ldots, np}_{m \text{ times}}) = \text{ER}(n, m, p).$$

Lastly, the BTER model (which, as explained in [2], utilizes the CL model as a subroutine) can be understood as a generalization of CL. The BTER model is designed to match not only vertex and edge size distributions, but also per-degree metamorphosis coefficients. A complete definition of metamorphosis coefficients is involved; interested readers are referred to [2] for full details. Nonetheless, to elucidate how metamorphosis coefficients are interpreted in the hypergraph setting, we provide a high-level description here.

Metamorphosis coefficients are measures of network community structure based on counts of bipartite 4-cycles, also called butterflies, and bipartite 3-paths, also called caterpillars. In the language of hypergraphs, a butterfly is a subhypergraph consisting of two vertices and two edges intersecting in those two vertices; a caterpillar is an edge with two vertices intersecting with another edge in one of those vertices. The authors in [2] define metamorphosis coefficients for vertices within each of the two partitions of a bipartite graph, based on the ratios of butterfly to caterpillar counts those vertices participate in. Stated equivalently, this defines metamorphosis for the vertices and hyperedges of a hypergraph. If a hyperedge $e$ has a large metamorphosis coefficient, this means a large proportion of the edges that $e$ intersects with intersect in (at least) 2 vertices; dually, if vertex $v$ has large metamorphosis, then a large proportion of vertices $v$ shares an edge with share (at least) 2 edges. The BTER model is designed to match degree distributions, as well as the average metamorphosis coefficients for vertices and hyperedges of a given degree and cardinality, respectively.

Taken as a suite, these three models serve particularly well as null-models since each provides successively more control over hypergraph structure than the previous, providing the flexibility to choose different tiers of structural nuance for the generated hypergraphs. In the next section, we run each model multiple times on each dataset, and compare their $s$-walk properties. By (for example) “running CL on LesMis”, we mean extracting the model inputs (in this case, the vertex degrees and hyperedge sizes) from the data, and using the Chung-Lu model to generate a hypergraph under these inputs.

### 5.2 Comparison

In Figure 8, we compare $s$-walk based properties of LesMis*, Diseasome, and CompBoard against those of synthetic hypergraphs generated by ER, CL, and BTER. For each dataset, we generate 100 instances of each synthetic model and compute the properties in question for each instance. The plot reports the average values observed over the 100 trials, for each $s$.

In the leftmost column in Figure 8 we use Kolmogorov-Smirnov (KS) distance to compare the $s$-component size probability distributions of the original and synthetic hypergraphs. We remind the reader
that KS distance is normalized between 0 and 1, with smaller KS-values indicating greater similarity. Comparing the models, we observe that as \( s \) increases, the ER model tends to exhibit higher KS distance than for CL and BTER, indicating \( s \)-component size distributions that are more dissimilar to the original. All three models seem to exhibit larger KS distances for larger \( s \) values, although in some cases (e.g. for CL and BTER on LesMis*) this increase is not monotonic in \( s \). One notable exception is CompBoard, in which all models exhibit much larger KS distance for \( s = 1 \) than \( s = 2 \). This can likely be attributed to the large number of isolated hyperedges observed in 1-components of CompBoard: in contrast, all three models tend to output hypergraphs in which the majority of hyperedges are contained within a single giant 1-component.

Turning our attention to our \( s \)-distance based measures, we compare the original \( s \)-diameter (center right) and average \( s \)-efficiency (far right) of the three models against those of the synthetic model’s hypergraphs. As the average \( s \)-efficiency plots are in log-scale, average values of 0 (which occur whenever no two hyperedges intersect in \( s \) vertices) are not plotted. Comparing the three models, ER tends to have lower \( s \)-diameter and average \( s \)-efficiency as \( s \) increases, when compared to CL and BTER. For LesMis* and Diseasome, CL and BTER seem to perform comparably; for CompBoard, however, BTER does noticeably better than both in matching average \( s \)-efficiency for \( s \geq 2 \), although still diverging from the original values considerably for \( s \geq 5 \). Lastly, we consider the model’s performance with regard to the mean \( s \)-local clustering coefficients (center left). For all three datasets, ER produces smaller clustering coefficients than observed in the original data, for all \( s \). For CL and BTER, we observe that for smaller values of \( s \), the mean \( s \)-local clustering coefficients sometimes exceed those of the original data (e.g. for \( s \leq 2 \) on Diseasome), while for some larger values of \( s \) (e.g. for \( 13 \leq s \leq 19 \) on LesMis*), BTER and CL produce smaller local clustering coefficients than those of the original data.

---

10For example, the green point at \((13, 0.2)\) in the LesMis* plot means that, over 100 trials, the average KS-distance between the 13-component size distribution of original LesMis* dataset, and that of a CL hypergraph, is 0.2. In cases where the synthetic graph had no hyperedges containing at least \( s \) vertices (and hence an empty \( s \)-component size distribution) we define KS distance between the original \( s \)-comp distribution as 1 (the maximum).
Taking a broader view of these results, note that none of these three models are able to provide a consistent, close match across values of \( s \). This suggests the \( s \)-walk-based measures in question cannot be explained as sole consequences of the model inputs (e.g. degree distributions for the Chung-Lu model), that are preserved in expectation in the output hypergraphs. Nonetheless, this experiment should not be extrapolated to provide generalized guidance on which model best preserves certain \( s \)-walk properties. Indeed, depending on the properties of the data in question, it may be the case that ER (the least accurate model on our data) provides a closer match than CL or BTER. In order to provide more a comprehensive approach to such questions, it would be of interest to determine conditions on model inputs under which certain \( s \)-walk properties of the output hypergraphs can be tightly bounded or controlled. While such work is outside the scope of the present paper, the aforementioned research by Kang, Cooley, and Koch \[15, 16, 17\] illustrates that proving analytical guarantees on even basic high-order walk based properties in random hypergraphs (such as the size of the largest \( s \)-component) requires sophisticated probabilistic analysis.

6 Conclusion

The prevalence and complexity of hypernetwork data necessitates analytic methods that are both applicable and able to capture hypergraph-native phenomena. We have proposed that hypergraph \( s \)-walks provide a framework under which a number of graph analytic tools popular in network science extend more meaningfully to hypergraphs. In applying these measures to real data, we’ve explored how they may reveal varied, interpretable, and significant structural properties of the data otherwise lost when analyzing hypergraphs under the lens of the usual graph walk. The methods that we’ve focused on – connected component analyses, distance-based measures, high-order motifs and clustering coefficients – are meant to illustrate the breadth of tools to which this approach is relevant. However, ours is clearly far from a comprehensive exploration. We conclude by briefly outlining several promising lines of future work that highlight the limits of our approach.

One immediate open question concerns how the methods we’ve developed here may be generalized further. For instance, it would be of both theoretical and practical interest to develop tractable \( s \)-walk based measures for weighted hypergraphs (with real-valued vertex and/or edge weights), directed hypergraphs (in which each edge’s vertices are either in its “head” or “tail”), ordered hypergraphs (in which each edge’s vertices are totally ordered), or temporal hypergraphs (consisting of sequences of hypergraphs). Another open direction lies in devising efficient computational methods for the \( s \)-walk measures put forth here. Indeed, we did not explore the algorithmic aspects underlying these methods. In some cases, the methods we utilized – while sufficient on our data – were not scalable to massive hypergraph data (e.g. computing \( s \)-centrality via the \( s \)-line graph quickly becomes infeasible for large hypergraphs with skewed degree distributions, as the density of \( s \)-line graphs increase quadratically in the maximum vertex degree). Developing algorithms that leverage the sparsity of the hypergraph (rather than resorting to computation on dense \( s \)-line graphs) would help facilitate the application of these methods to larger-scale data. Furthermore, just as researchers have begun developing efficient schemes for computing atomic bipartite graph motifs such as cycles of length 4 \[61, 65\], work in a similar vein would prove useful for enabling large-scale \( s \)-triangle counting in hypergraphs.

Acknowledgements: We would like to thank numerous colleagues for helpful discussions, including Marcin Zalewski, Francesca Grogan, Katy Nowak, Dustin Arendt, Stephen Young, Brett Jefferson, and Louis Jenkins.

Funding: This work was partially funded under the High Performance Data Analytics (HPDA) program at the Department of Energy’s Pacific Northwest National Laboratory. PNNL Information Release: PNNL-SA-144766. Pacific Northwest National Laboratory is operated by Battelle Memorial Institute under Contract DE-AC06-76RL01830.

References

[1] A. Agresti, *Analysis of Ordinal Categorical Data (Wiley Series in Probability and Statistics Book 656)*, Wiley, 2012.

[2] S. G. Aksoy, T. G. Kolda, and A. Pinar, *Measuring and modeling bipartite graphs with community structure*, Journal of Complex Networks, 5 (2017), pp. 581–603.
[3] N. Alon, Transversal numbers of uniform hypergraphs, Graphs and Combinatorics, 6 (1990), pp. 1–4.

[4] A. J. Alvarez-Socorro, G. C. Herrera-Almarza, and L. A. González-Díaz, Eigencentrality based on dissimilarity measures reveals central nodes in complex networks, Scientific Reports, 5 (2015).

[5] S. M. Axinn, P. A. Proger, and N. Yoerg, Interlocking Directorates Under Section 8 of the Clayton Act (Monograph / American Bar Association, Section of Antitrust Law, 10) (5030057), Amer Bar Assn, 1984.

[6] A.-L. Barabási, Network Science, Cambridge University Press, 2016.

[7] M. J. Barber, Modularity and community detection in bipartite networks, Physical Review E, 76 (2007).

[8] C. Berge, Hypergraphs: Combinatorics of Finite Sets (North-Holland Mathematical Library), North Holland, 1984.

[9] J.-C. Bermond, M.-C. Heydemann, and D. Sotteau, Line graphs of hypergraphs i, Discrete Mathematics, 18 (1977), pp. 235–241.

[10] A. Bretto, Hypergraph Theory, Springer International Publishing, 2013.

[11] P. S. Chodrow, Configuration models of random hypergraphs and their applications, arXiv preprint arXiv:1902.09302, (2019).

[12] F. Chung, The laplacian of a hypergraph, Expanding graphs (DIMACS series), (1993), pp. 21–36.

[13] ———, Complex graphs and networks, no. 107, American Mathematical Soc., 2006.

[14] M. J. Conyon and M. R. Muldoon, The small world network structure of boards of directors, SSRN Electronic Journal, (2004).

[15] O. Cooley, W. Fang, N. Del Giudice, and M. Kang, Subcritical random hypergraphs, high-order components, and hypertrees, arXiv preprint arXiv:1810.08107, (2018).

[16] O. Cooley, M. Kang, and C. Koch, Evolution of high-order connected components in random hypergraphs, Electronic Notes in Discrete Mathematics, 49 (2015), pp. 569–575.

[17] ———, Threshold and hitting time for high-order connectedness in random hypergraphs, Electr. J. Comb., 23 (2016), p. P2.48.

[18] J. Cooper and A. Dutle, Spectra of uniform hypergraphs, Linear Algebra and its Applications, 436 (2012), pp. 3268–3292.

[19] R. W. R. Darling and J. R. Norris, Structure of large random hypergraphs, The Annals of Applied Probability, 15 (2005), pp. 125–152.

[20] M. Dewar, J. Healy, X. Pérez-Giménez, P. Prałat, J. Proos, B. Reingner, and K. Ter-novsky, Subhypergraphs in non-uniform random hypergraphs, Internet Mathematics, 49 (2015), pp. 569–575.

[21] I. Dinur, O. Regev, and C. Smyth, The hardness of 3-uniform hypergraph coloring, Combinatorica, 25 (2005), pp. 519–535.

[22] P. Erdős and A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci, 5 (1960), pp. 17–60.

[23] E. Estrada and J. A. Rodríguez-Velázquez, Subgraph centrality and clustering in complex hyper-networks, Physica A: Statistical Mechanics and its Applications, 364 (2006), pp. 581–594.

[24] M. Everett and S. Borgatti, The dual-projection approach for two-mode networks, Social Networks, 35 (2013), pp. 204–210.

[25] G. C. Garriga, E. Junttila, and H. Mannila, Banded structure in binary matrices, Knowledge and Information Systems, 28 (2010), pp. 197–226.

[26] G. Ghoshal, V. Zlatić, G. Caldarelli, and M. E. J. Newman, Random hypergraphs and their applications, Physical Review E, 79 (2009).

[27] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, The human disease network, Proceedings of the National Academy of Sciences, 104 (2007), pp. 8685–8690.

[28] A. Hagberg, P. Swart, and D. S Chult, Exploring network structure, dynamics, and function using networkx, tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
[29] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, *Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders*, Nucleic acids research, 33 (2005), pp. D514–D517.

[30] H. Han and M. Schacht, *Dirac-type results for loose hamilton cycles in uniform hypergraphs*, Journal of Combinatorial Theory, Series B, 100 (2010), pp. 332–346.

[31] L. Jenkins, T. Bhuiyan, S. Harun, C. Lightsey, D. Mentgen, S. Aksoy, T. Stavcnger, M. Zalewski, H. Medal, and C. Joslyn, *Chapel hypergraph library (chgl)*, in 2018 IEEE High Performance Computing Conference (HPEC), IEEE, 2018, pp. 1–6.

[32] L. Jenkins, T. Stavenger, M. Zalewski, C. Joslyn, S. Aksoy, and H. Medal, https://github.com/pnnl/chgl.

[33] C. Joslyn, S. Aksoy, D. Arendt, L. Jenkins, B. Praggastis, E. Purvine, and M. Zalewski, *High performance hypergraph analytics of domain name system relationships*, in HICSS 2019 Symposium on Cybersecurity Big Data Analytics, 2019.

[34] C. Joslyn and E. Purvine, *Information measures of frequency distributions with an application to labeled graphs*, in Association for Women in Mathematics Series, Springer International Publishing, 2016, pp. 379–400.

[35] B. Kaminski, V. Poulin, P. Pralat, P. Szufel, and F. Theberge, *Clustering via hypergraph modularity*, arXiv preprint arXiv:1810.04816, (2018).

[36] G. O. H. Katona, *Extremal problems for hypergraphs*, in Combinatorics, Springer Netherlands, 1975, pp. 215–244.

[37] G. Y. Katona and H. A. Kierstead, *Hamiltonian chains in hypergraphs*, Journal of Graph Theory, 30 (1999), pp. 205–212.

[38] S. Kirkland, *Two-mode networks exhibiting data loss*, Journal of Complex Networks, 6 (2017), pp. 297–316.

[39] S. Klamt, U.-U. Haus, and F. Theis, *Hypergraphs and cellular networks*, PLoS Computational Biology, 5 (2009), p. e1000385.

[40] D. E. Knuth, *The Stanford GraphBase: a platform for combinatorial computing*, AcM Press New York, 1993.

[41] T. G. Kolda, A. Pinar, T. Plantenga, and C. Seshadhri, *A scalable generative graph model with community structure*, SIAM J. Sci. Comput., 36 (2014), pp. C424–C452.

[42] M. Krivelevich and B. Sudakov, *Approximate coloring of uniform hypergraphs*, Journal of Algorithms, 49 (2003), pp. 2–12.

[43] D. B. Larremore, A. Clauset, and A. Z. Jacobs, *Efficiently inferring community structure in bipartite networks*, Physical Review E, 90 (2014).

[44] M. Latapy, C. Magnien, and N. D. Vecchio, *Basic notions for the analysis of large two-mode networks*, Social Networks, 30 (2008), pp. 31–48.

[45] V. Latora and M. Marchiori, *Efficient behavior of small-world networks*, Physical Review Letters, 87 (2001).

[46] J. H. Levine and W. S. Roy, *A study of interlocking directorates: Vital concepts of organization*, in Perspectives on Social Network Research, Elsevier, 1979, pp. 349–378.

[47] L. Lu and X. Peng, *High-ordered random walks and generalized laplacians on hypergraphs*, in WAW, Springer, 2011, pp. 14–25.

[48] J. Nacher and T. Akutsu, *On the degree distribution of projected networks mapped from bipartite networks*, Physica A: Statistical Mechanics and its Applications, 390 (2011), pp. 4636–4651.

[49] R. N. Naik, *Recent advances on intersection graphs of hypergraphs: A survey*, arXiv preprint arXiv:1809.08472, (2018).

[50] R. N. Naik, S. Rao, S. Shrikhande, and N. Singhi, *Intersection graphs of k-uniform linear hypergraphs*, European Journal of Combinatorics, 3 (1982), pp. 159–172.

[51] M. E. J. Newman, *The structure and function of complex networks*, SIAM Review, 45 (2003), pp. 167–256.

[52] M. E. J. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Physical Review E, 69 (2004).
[53] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Random graphs with arbitrary degree distributions and their applications, Physical Review E, 64 (2001).

[54] T. Opsahl, Triadic closure in two-mode networks: Redefining the global and local clustering coefficients, Social Networks, 35 (2013), pp. 159–167.

[55] O. Parczyk and Y. Person, On spanning structures in random hypergraphs, Electronic Notes in Discrete Mathematics, 49 (2015), pp. 611–619.

[56] B. Praggastis, D. Arendt, C. Joslyn, E. Purvine, S. Aksoy, and K. Monson, Hypernetx, https://github.com/pnnl/HyperNetX, 2019.

[57] E. Purvine, S. Aksoy, C. Joslyn, K. Nowak, B. Praggastis, and M. Robinson, A topological approach to representational data models, in International Conference on Human Interface and the Management of Information, Springer, 2018, pp. 90–109.

[58] G. Robins and M. Alexander, Small worlds among interlocking directors: Network structure and distance in bipartite graphs, Computational & Mathematical Organization Theory, 10 (2004), pp. 69–94.

[59] Y. Rochat, Closeness centrality extended to unconnected graphs: The harmonic centrality index, tech. rep., 2009.

[60] V. Rödl and J. Skokan, Regularity lemma for k-uniform hypergraphs, Random Structures & Algorithms, 25 (2004), pp. 1–42.

[61] S.-V. Sanei-Mehr, A. E. Sariyuce, and S. Thirthapura, Butterfly counting in bipartite networks, in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD ’18, ACM Press, 2018.

[62] A. E. Sariyuce and A. Pinar, Peeling bipartite networks for dense subgraph discovery, in Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM ’18, ACM Press, 2018.

[63] C. Seshadhri, T. G. Kolda, and A. Pinar, Community structure and scale-free collections of erdős-rényi graphs, Physical Review E, 85 (2012).

[64] J. Wang and T. T. Lee, Paths and cycles of hypergraphs, Science in China Series A: Mathematics, 42 (1999), pp. 1–12.

[65] K. Wang, X. Lin, L. Qin, W. Zhang, and Y. Zhang, Vertex priority based butterfly counting for large-scale bipartite networks, arXiv preprint arXiv:1812.00283, (2018).

[66] D. J. Watts and S. H. Strogatz, Collective dynamics of ‘small-world’ networks, Nature, 393 (1998), pp. 440–442.

[67] H. Whitney, Congruent graphs and the connectivity of graphs, American Journal of Mathematics, 54 (1932), p. 150.

[68] W. Zhou and L. Nakhleh, Properties of metabolic graphs: biological organization or representation artifacts?, BMC Bioinformatics, 12 (2011).
### Appendix

|   | LesMis<sup>∗</sup> | Disease | ComplBoard |
|---|-------------------|---------|------------|
| 1-components | ![Diagram](image1) | ![Diagram](image2) | ![Diagram](image3) |
| 2-components | ![Diagram](image4) | ![Diagram](image5) | ![Diagram](image6) |
| 3-components | ![Diagram](image7) | ![Diagram](image8) | ![Diagram](image9) |
| 4-components | ![Diagram](image10) | ![Diagram](image11) | ![Diagram](image12) |
| 5-components | ![Diagram](image13) | ![Diagram](image14) | ![Diagram](image15) |