Hypergraphs for predicting essential genes using multiprotein complex data

Florian Klimm*[1,2,3], Charlotte M. Deane[1], and Gesine Reinert[1]

[1]Department of Statistics, University of Oxford, Oxford OX1 3LB, United Kingdom

[2]Department of Mathematics, Imperial College London, London, SW7 2AZ, United Kingdom

[3]MRC Mitochondrial Biology Unit, University of Cambridge, Cambridge Biomedical Campus Hills Road, Cambridge, CB2 0XY, United Kingdom

April 2, 2020

Abstract

Protein–protein interactions are crucial in many biological pathways and facilitate cellular function. Investigating these interactions as a graph of pairwise interactions can help to gain a systemic understanding of cellular processes. It is known, however, that proteins interact with each other not exclusively in pairs but also in polyadic interactions and they can form *multiprotein complexes*, which are stable interactions between multiple proteins. In this manuscript, we use *hypergraphs* to investigate multiprotein complex data. We investigate two random null models to test which hypergraph properties occur as a consequence of constraints, such as the size and the number of multiprotein complexes. We find that assortativity, the number of connected components, and clustering differ from the data to these null models. Our main finding is that projecting a hypergraph of polyadic interactions onto a graph of pairwise interactions leads to the identification of different proteins as hubs than the hypergraph. We find in our data set that the hypergraph degree is a more accurate predictor for gene-essentiality than the degree in the pairwise graph. We find that analysing a hypergraph as pairwise graph drastically changes the distribution of the local clustering coefficient. Furthermore, using a pairwise interaction representing multiprotein complex data may lead to a spurious hierarchical structure, which is not observed in the hypergraph. Hence, we illustrate that hypergraphs can be more suitable than pairwise graphs for the analysis of multiprotein complex data.

Keywords: Gene essentiality; Protein interaction networks; Hypergraphs; Null models; Hierarchical exponent; Centrality; Clustering coefficient

*Corresponding author. Email: f.klimm@gmail.com

1 Introduction

Protein–protein interactions represent the chemical reactions and physical contacts between proteins [1]. Their statistical analysis can give insights into underlying cellular processes and the organism they govern. They are therefore used in various bioinformatics applications, such as, the reconstruction of phylogenetic trees, the prediction of proteins' biological functions and the identification of functional modules (for reviews see [1, 2]). One important application is the prediction of whether a gene that codes a certain protein is essential [3, 4]. Typically, a dataset of protein–protein interactions is represented as a binary undirected network, with proteins as nodes and edges representing interactions. For predicting essential proteins, one can for example, investigate the centralities of nodes [5] or combinations of multiple measures [6] in such a protein–protein interaction network.

In such studies, the interaction between the proteins are modelled as pairwise. More than half of all proteins, however, form *multiprotein complexes* that may consist of more than two proteins that are linked by non-covalent interactions [7, 8]. Protein complexes are crucial for most biological processes *ATP synthase*, for example, an enzyme that creates the energy storage molecule adenosine triphosphate (ATP), consists of up to eight different subunits, each a protein [9]. Proteins can be involved in different complexes or have additional activities, independent of the complex itself [10]. These biological observations indicate that mathematical objects that take multiprotein complex information as high-order interactions into account might be an appropriate way to study cellular systems in general and the prediction of essentiality, specifically. In this study, we use *hypergraphs* [11], which are one way to represent *polyadic interactions* (i.e., interactions of higher order than pairwise), to analyse a network of human multiprotein complexes.

In Fig. 1, we give three examples of multiprotein complexes and the representation of their high-order interactions as *hyperedges* in a hypergraph. The *exon junction complex* is a crucial molecular machine that influences the translation of mRNA molecules [12]. It consists of four different protein components CASC, Y14, MAGOH, and EIF4A3 and we therefore represent this interaction as a four-edge in a hypergraph. Two of these four proteins (Y14 and MAGOH) can form a separate complex (the Y14–MAGOH complex), which we represent as a two-edge [13]. The PYM protein can bind to this complex and we represent the formed complex which consists of three proteins as a three-edge. These three different complexes demonstrate only a small subset of the complex higher-order interactions that we observe in the human body. The PYM protein, for example, interacts with the 40S ribosomal subunit, which itself consists of thirty-three proteins (not shown). Higher-order interactions are common in cellular processes and in this study, we represent their complex interaction structure as a hypergraph.

Other mathematical structures are potentially also suited to represent high-order interactions. *Simplicial complexes*, for example, have been used to investigate time-series [14, 15], and many other systems [16, 17, 18]. For our purposes,

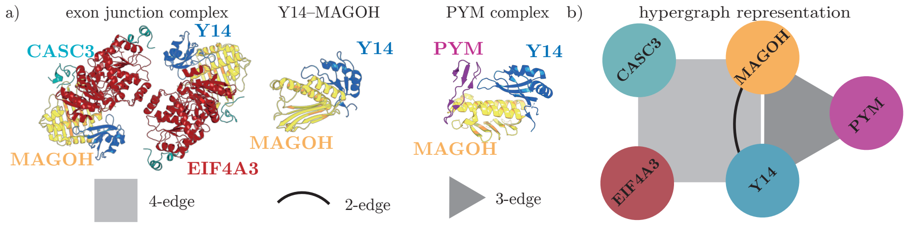

Figure 1: Protein may interact with each other and form complexes. (a) We show the exon junction complex, the Y14–MAGOH complex, and the PYM complex in cartoon representations. We can represent these interactions as hyperedges, whose cardinality is the number of different involved proteins. The exon junction complex, for example, consists of four different proteins (CASC3, Y14, MAGOH, and EIF4A3; shown in green, blue, yellow, and red, respectively) and thus we represent it as a four-edge. The two proteins Y14 and MAGOH can also interact with each other and we represent this interaction as a 2-edge. The interaction between PYM, Y14, and MAGOH is represented in a 3-edge. (b) We jointly represent these three interactions in a hypergraph, with the $N = 5$ nodes representing proteins and the $M = 3$ hyperedges representing multiprotein complexes.

we argue that hypergraphs are a more suited mathematical framework because simplicial complexes require *set inclusion*[1], which for our application implied that for every multiprotein complex, all subsets of constituent proteins would also form a multiprotein complex, which in general is not the case.

Hypergraphs have been identified as a framework to investigate metabolic pathways, which describe molecular reactions [20, 21, 22] and some methods for the statistical analysis of hypergraphs have been developed (e.g., centralities [23], local clustering coefficient [21], configuration models [24]). In this manuscript, we focus on the degree and clustering coefficient as node-measures because they are commonly used for predicting the lethality or essentiality of proteins [3, 5].

First, we assess whether the constructed hypergraph contains signal beyond the degree. To test this, we construct two random null models, one an existing hypergraph configuration model [24] and one a new Erdős–Rényi-type hypergraph model. We compare the hypergraph properties of these null models with the data hypergraph. Similarly to many empirical graphs, we find that assortativity, clustering, and number of connected components differ strongly from these random null models.

Often, one projects higher-order interactions to pairwise interactions to be able to use a broad selection of tools that have been developed for the analysis of graphs. To test which representation, graph or hypergraph, is more suitable

[1] The convex hull of any subset of the $n+1$ points that define a n-simplex is called a *face*. A simplicial complex $\mathcal{S}$ is a set of simplices in which every face of a simplex is also in $\mathcal{S}$ [19]. This property is called *set inclusion*.

for the analysis of the multiprotein complex data, we compare gene essentiality data from the *Online GEne Essentiality database* [25] with the degree in the hypergraph and the degree in the pairwise interaction graph. As the former is in stronger agreement, the hypergraph representation outperforms the pairwise graph representation for identifying essential genes using degrees.

Next, we show that using a pairwise interaction graph may lead to a spurious result for the commonly asserted *hierarchical organisation* of complex systems: together with the degree, the local clustering coefficient has been used to quantify the hierarchical organisation of pairwise complex networks [26, 27] in general and metabolic networks, specifically [28, 29]. Here, we demonstrate that projecting a hypergraph onto a graph can drastically increases the local clustering coefficient of many nodes. Furthermore, this projection may indicate a statistically significant hierarchical organisation of the graph that is not observed in its hypergraph form. As such a projection is common in many network studies —either explicitly or implicitly— one should be careful about the interpretation of such results.

Overall, we propose that a hypergraph representation for multiprotein complex data is a better approach to identify essential genes and that not using this representation may lead to a spurious hierarchical structure in the graph.

2 Methods

2.1 Hypergraph measures

A *graph* is an ordered pair $\mathcal{G} = \{V, E\}$, where V is a set of nodes and $E \subset V \times V$ a set of edges that connect the nodes pairwise; in this paper, edges are undirected and unweighted. Two nodes which are connected by an edge are called *neighbours* and the *neighbourhood* $\mathcal{N}(i)$ of a node i is the set of all its neighbours. A *hypergraph* is a generalisation of a graph that allows edges that connect more than a pair of nodes and are therefore called *hyperedges*. Formally, we define $\mathcal{H} = \{V, E\}$, where V is a set of nodes and the hyperedge set E is a subset of the *power set* $P(V)$, which is the set of all subsets of S, but excluding the empty set $\emptyset$. Therefore, a hyperedge may connect any set of nodes but not the empty set. The number $c = |e|$ of nodes that a hyperedge $e \in E$ connects to is called the hyperedge's *cardinality*. A hyperedge with cardinality c is also called a *c-edge*.

For graphs, the degree k_i of a node i is the number of edges it connects to. For simple graphs (i.e., graphs without parallel edges and without self-loops), the degree is identical to the number of neighbours this node has. In accordance with graphs, the degree $k_i^{(\text{hyp})} = \sum_{i \in E} 1$ of node i in a hypergraph is the number of hyperedges it connects to. In contrast to simple graphs, the degree of a node in a hypergraph is not necessarily equal to the number of its neighbours [30]. The maximum degree $\max(k)$ is the largest degree of any node in a hypergraph.

The *local clustering coefficient* C_i of a node i in a graph is

$$C_i = \begin{cases} \frac{2|(l,m)\in\mathcal{N}(i) \text{ with } (l,m)\in E|}{k_i(k_i-1)}\,, \text{ if } k_i > 0\,, \\ \text{not defined}\,, \text{ if } k_i = 0\,. \end{cases}$$

We choose a definition from [21] to generalise the *local clustering coefficient* but adapt it slightly for clarity. The local clustering coefficient $C_i^{(\text{hyp})}$ of a node i in a hypergraph is

$$C_i^{(\text{hyp})} = \begin{cases} \frac{1}{k_i^{(\text{hyp})}(k_i^{(\text{hyp})}-1)} \sum_{e=(i,j)\in E} \sum_{e'=(i,j)\in E} EO(e,e')\,, \text{ if } k_i^{(\text{hyp})} > 1\,, \\ 0\,, \text{ if } k_i^{(\text{hyp})} = 1\,, \\ \text{not defined}\,, \text{ if } k_i^{(\text{hyp})} = 0\,, \end{cases}$$

where the *extra overlap* $EO(e,e')$ between two intersecting hyperedges e and e' is defined as

$$EO(e,e') = \frac{|\mathcal{N}(D_{e,e'}) \cap D_{e',e}| + |\mathcal{N}(D_{e',e}) \cap D_{e,e'}|}{|D_{e,e'}| + |D_{e',e}|}\,,$$

with $D_{e,e'} = e - e'$, the (asymmetric) set difference between e and e'. We define $EO(e,e') = 0$ for $e = e'$. The neighbourhood $\mathcal{N}(S)$ of a set S of nodes is the union of the neighbourhoods of each node in the set, i.e. $\mathcal{N}(S) = \cup_{i\in S}\,(\mathcal{N}(i))$. For isolated nodes with $k_i^{(\text{hyp})} = 0$, the local clustering is not defined. Another variant of local clustering is suggested in [20] and a global clustering coefficient is discussed in [23]. The mean local clustering coefficient $\langle C_i \rangle$ is defined as $\langle C_i \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i$. The *assortativity* ρ of a hypergraph is a measure of the correlation between a node's degree and the degree of its neighbours. Following [24], the assortativity ρ of a hypergraph is the Pearson correlation between the nodes' degrees $k_i^{(\text{hyp})}$, $i \in V$ and the mean degree $\langle k_i^{(\text{hyp})} \rangle = \frac{1}{k_i^{(\text{hyp})}} \sum_{j\in N(i)} k_j^{(\text{hyp})}$ of all of its neighbours.

For graphs, the relationship between degree k_i and local clustering coefficient C_i has been approximately described through a power-law $C_i(k_i) \sim k_i^{-\beta}$ in which β is called the *hierarchical exponent* [26]. In practice, we estimate β by calculating the Pearson correlation between $\log_{10}(k_i)$ and $\log_{10}(C_i)$ for $i = 1,\ldots.N$. With the definitions of local clustering coefficient and degree, we can also compute the *hypergraph hierarchical exponent* $\beta^{(\text{hyp})}$ as Pearson correlation between $\log_{10}(k_i^{(\text{hyp})})$ and $\log_{10}(C_i^{(\text{hyp})})$ for $i = 1,\ldots.N$.

For a graph, a *component* is a subgraph in which any two nodes are connected to each other by paths. For hypergraphs, more nuanced definitions exists, e.g., 'j-component' are sets of vertices such that consecutive edges in paths intersect in at least j vertices [31]. We discuss here exclusively 1-components and call them 'components' for simplicity. The number n_{com} of components is the amount of components in a hypergraph. The size $S^{(m)}$ of a component is the number of nodes in it. We also compute the relative size $S_{\max}/N \in (0,1]$ of the largest connected component.

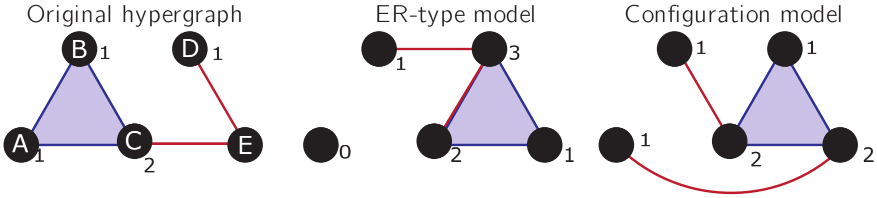

Figure 2: We discuss two null models. Both preserve the number N of nodes, the number M of hyperedges, and the cardinality of the hyper-edges, which in this example is $(3, 2, 2)$. We indicate the degree of each node as number next to it. In the *degree-preserving null model* the degree of each node is preserved. In the *ER-type null model* the degree is not preserved.

2.2 Data and preprocessing

We constructed a hypergraph from REACTOME version May 2019 [32]. The data set 'Human complexes with their participating protein molecules' consists of a total of $\sim 12,000$ complexes. The complexes include not only proteins but also other ligands, for example, small molecules (described by 'chebi' codes) and RNA molecules (described by 'ensemble' IDs). We ignored these entities and only kept entities that describe proteins with a uniprot ID. After deletion of duplicate entries, we obtained a hypergraph with $N = 8243$ nodes (representing proteins) and $M = 6688$ hyperedges (representing multiprotein complexes). This data combines obligate and non-obligate protein complexes, as well as, transient and stable protein complexes. We used gene-essentiality data from the *Online GEne Essentiality database* (OGEE) v2 [25]. We mapped genes to proteins with the Retrieve/ID mapping tool from UNIPROT [33].

2.3 Null models

Null models for hypergraphs have been developed in different domains. In uniform random hypergraphs all hyperedges have the same cardinality c [34, 35]. In this study, we use two different null models (see Fig. 2) that are non-uniform. For the construction of a *configuration model* of hypergraphs, we use definitions from [24]. We also define a novel null model, called Erdős–Rényi-type hypergraph model (ER-type hypergraph model), that does not fix the degrees of nodes in the hypergraph. For the configuration model, we use a randomisation algorithm. For the latter null model, in contrast, we construct hypergraphs directly.

Erdős–Rényi-type hypergraph model For a hypergraph with N nodes and M hyperedges, we define the degree sequence $\mathbf{k}$ as the N-vector in which the ith element is the degree $k_i^{(\text{hyp})}$ of node i. Similarly, we define the cardinality sequence $\mathbf{c}$ as the M-vector in which the ith element is the cardinality c_i of hyperedge i.

Let $\mathcal{H}(\mathbf{c})$ be the set of all hypergraphs with a fixed cardinality sequence $\mathbf{c}$. The random hyperedges hypergraph model is the uniform distribution on $\mathcal{H}(\mathbf{c})$ with self-loops and multiple hyperedges possible. We construct realisations of this model by connecting uniformly at random c_i nodes for every hyperedge i. In Algorithm 1, we show the procedure used to construct ER-type hypergraphs. For dense hypergraphs, this algorithm may construct hypergraphs with multiple hyperedges (i.e., hyperedges that connect the same set of nodes). For our examples, this was, however, not the case, because we do not have multiple multiprotein complexes that connect the identical proteins.

For graphs, (i.e., $c_i = 2$ for all edges) with low connection density, this random hypergraph model is identical to the $G(N, M)$ model by Erdős–Rényi [36]. Therefore we call it *Erdős–Rényi-type hypergraph model*. This model has some similarity with the *Poisson random hypergraph* in which the number of hyperedges between two nodes is Poisson distributed [37, 38].

Configuration hypergraph model Let $\mathcal{H}(\mathbf{k}, \mathbf{c})$ be the set of all hypergraphs with a fixed degree sequence $\mathbf{k}$ and a fixed cardinality sequence $\mathbf{c}$. The *vertex-labelled hypergraph configuration model* is then the uniform distribution on $\mathcal{H}(\mathbf{k}, \mathbf{c})$. To construct random hypergraphs, we use a pairwise reshuffeling algorithm [24], which preserves the degree sequence and the cardinality sequence. As the reshuffling algorithm is an irreducible, reversible, and aperiodic Markov Chain, it has an equilibrium distribution, which we call the configuration hypergraph model. In this model, self-loops and multiple hyperedges are possible.

2.4 Constructing graphs from hypergraphs

We construct a *representing graph* from a hypergraphs as follows. The *representing graph* $R(H) = (V', E')$ of a hypergraph $H = (V, E)$ is the graph with the same set $V' = E'$ of vertices as the hypergraph, and edges between all pairs of vertices contained in the same hyperedge (i.e, $(i, j) \in E'$ if there exists an edge $e \in E$ such that $(i, j) \subset e$). Thus, in the representing graph, hyperedges are translated into complete subgraphs of simple edges.

Algorithm 1: Constructing an Erdős–Rényi-type hypergraph H with N nodes and a cardinality sequence $\mathbf{c} = (c_1, c_2, \ldots, c_M)$

```
initialisation of empty hypergraph $H = (V, \emptyset)$;
initialisation of empty edge set $E = \emptyset$;
for $i \leftarrow 1$ to $M$ do
    1) Construct hyperedge $e$ by sampling $c_i$ nodes without replacement
       uniformly at random from $V$;
    2) Add the hyperedge to the edge list $E = E \cup e$
end
return $H(V, E)$
```

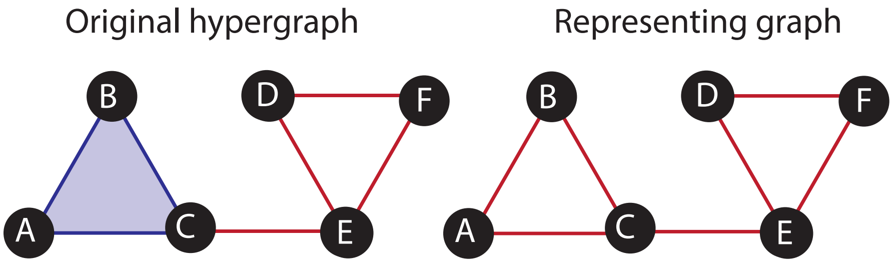

Figure 3: To construct the *representing graph* of a hypergaph, we replace every hyperedge of cardinality c with $c(c-1)/2$ 2-edges.

Alternatively, one could construct the *dual graph* from the hypergraphs. This method is discussed in SI A.

3 Results

3.1 Random topology null models

First, we explore whether the multiprotein complex hypergraph can be modelled (a) by the ER-type hypergraph model, which uses only its cardinality sequence $\mathbf{c}$ or (b) by the configuration hypergraph model, which uses only the hypergraph's degree sequence $\mathbf{k}$ and cardinality sequence $\mathbf{c}$ (see Subsection 2.3 for definitions of both models). For both null models, we construct 100 independent realisations. We focus on graph measures (maximum degree, degree-assortativity, number of components, relative size of the largest component, and the mean local clustering); Table 1 gives their mean and standard deviation in these simulated hypergraphs, and the values for the data hypergraph. In SI C, we illustrate the distributions of these measures for the simulated hypergraphs.

The maximum degree in the multiprotein hypergraph is $\max(k_i) = 283$; in the ER-like hypergraph the maximum degree $\max(k_i) \approx 27$ is far smaller. By contrast, the maximum degree $\max(k_i) = 283$ is by construction the same in the configuration model and the data. Computing the degree assortativity for the hypergraph yields $\rho_{\text{data}} \approx 0.44$. This indicates that proteins with high degree k_i tend to form complexes with other proteins with high degree [24]. For both null models, we find a assortativity close to zero. Performing a Monte-Carlo test yields a p-values $p < 0.01$ (see Fig. 12 in SI D), indicating that assortativity of the multiprotein hypergraph is significantly larger than expected by either null model.

For the mean local clustering $\langle C_i \rangle$, we find that the original hypergraph has a significantly smaller clustering than both, the ER-type model and the configuration model. This contrasts with pairwise protein interaction networks, in which the clustering is normally higher than in random null models [39].

We next investigated the components of the hypergraph. The hypergraph has $n_{\text{com}} = 253$ components of which the two largest consist of 7249 and 93 nodes, respectively. This means that almost 88 % of all nodes belong to the largest component. The smaller components range in size from 22 to 2. There are 131 components which have the minimum size 2 and thus are two proteins that are connected by a 2-edges and otherwise not involved in a multiprotein complex.

In both null models, the number n_{com} of connected components is much smaller than in the data hypergraph. For the ER-type model and the configuration model, we obtain $n_{\text{com}} = 1.01 \pm 0.01$ and $n_{\text{com}} = 3.69 \pm 1.5$, respectively. This indicates that if the edges were evenly distributed between the nodes, there would exist a path between almost all nodes. Fixing the degree distribution of the hypergraph leads to a slightly larger number of connected components. This occurs because of the larger number of nodes with degree $k_i = 1$ than in the ER-like model which has a mean degree of $\langle k_i \rangle \approx 1.6$. The relative size $S_{\max}/N$ of the largest connected component is 1 ± 0 for the ER-type model and 0.9993 ± 0.0004 for the configuration model. For both null models, the number n_{com} of connected components is significantly smaller and the size $S_{\max}/N$ of the largest connected component is significantly larger than for the protein data (p-values $p < 0.01$ shown in Fig. 13 in SI D).

	original hypergraph	ER-type model	configuration model
number of nodes	8243	8243	8243
number of hyperedges	6688	6688	6688
maximum degree	283	21.3 ± 1.4	283
degree-assortativity	0.44	0.000 ± 0.008	-0.005 ± 0.007
number of components	253	1.01 ± 0.01	3.69 ± 1.5
size of largest component	0.8794	1 ± 0	0.9993 ± 0.0004
mean^2 local clustering	0.079	0.39 ± 0.0008	0.48 ± 0.001

Table 1: Structural information about the protein hypergraph and the two investigated null models (ER-type model and configuration model). We constructed 100 null models and present the mean ± standard deviation. For definitions of hypergraphs measures see Subsection 2.1.

3.2 Degree distribution in graph and hypergraph

Next, we compare the degrees of the nodes in the hypergraph with the degree of nodes in the representing graph. To construct the representing graph, we replace each hyperedge of cardinality c with $c(c-1)/2$ simple edges. Therefore, the total number of edges in the representing graph is at least as large as the number of hyperedges in the hypergraph, and the degree distribution is wider for the representing graph (see Fig. 4). This indicates that replacing higher-order interactions with pairwise edges broadens the degree distribution. The

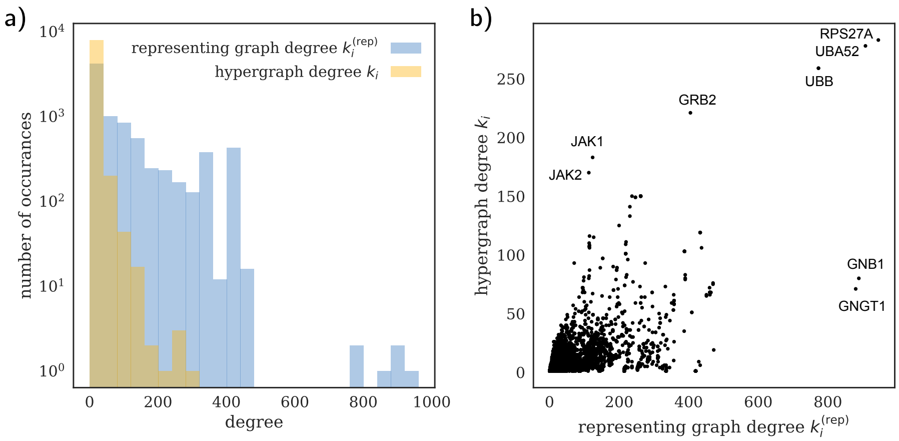

Figure 4: (a) Distribution of degrees in hypergraph and its representing graph. (b) Scatterplot of the hypergraph degree and the degree in the representing graph.

mean graph-degree $\langle k_i^{(\mathrm{rep})} \rangle \approx 92.99$ is an order of magnitude larger than that of the hypergraph $\langle k_i^{(\mathrm{hyp})} \rangle \approx 8.21$. In the right panel of Fig. 4, we plot the hypergraph-degree with the degree in the representing graph for each protein. The Spearman correlation between both is 0.34, which indicates a weak correlation between the two quantities. The genes with highest degree, RPS27A and UBA52, are the same in both structures. These genes encode the protein ubiquitin, which targets proteins to degrade them and is known to bind to many different proteins [40]. For the hypergraph degree $k_i^{(\mathrm{hyp})}$, Ubiquitin B (UBB) has the third highest degree. For the representing graph degree $k_i^{(\mathrm{rep})}$, GNB1 and GNGT1, two guanine nucleotide-binding proteins have the third and fourth highest degrees. Both proteins are membrane-bound proteins that form complexes consisting of a large number of proteins. In Fig. 5, we show a force-directed layout of the representing graph. The size of the nodes is proportional to their representing graph degree $k_i^{(\mathrm{rep})}$ and the colour indicates the hypergraph degree $k_i^{(\mathrm{hyp})}$. We observe cliques of nodes with high $k_i^{(\mathrm{rep})}$ and low $k_i^{(\mathrm{hyp})}$: these nodes represent proteins that participate in a large multiprotein complex but no other interactions. These observations indicate that representing multiprotein complex data as hypergraphs identifies some different proteins as 'hubs' than its representing graph. In Subsection 3.3, we compare the identified degrees with gene-essentiality information.

3.3 Identifying essential proteins and protein complexes

One of the prominent applications of protein–protein interaction networks is the identification of essential proteins (i.e., proteins without which an organism

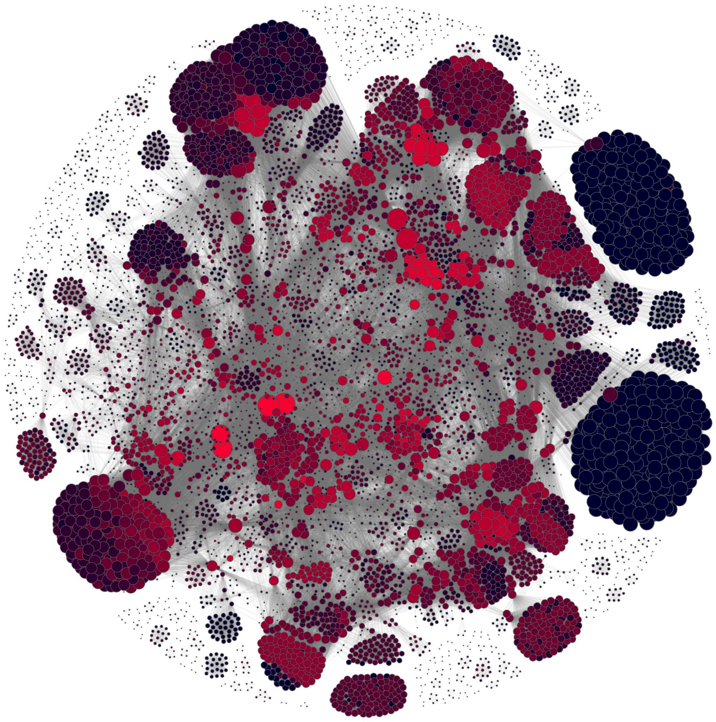

Figure 5: Force-directed layout of the representing graph. The size of nodes is in proportion to the representing graph degree $k_i^{(\mathrm{rep})}$ and the colour indicates the hypergraph degree $k_i^{(\mathrm{hyp})}$ from low (blue) to high (red). Accordingly, large, blue nodes indicate proteins with a high representing graph degree and a low hypergraph degree. Illustration created with Netwulf [41].

cannot survive). The degree of proteins has been suggested as a way to predict essentiality [5]. We now assess whether the degree in the hypergraph is also able to predict the essentiality of proteins.

For this task, we compare the average degree of proteins expressed by essential genes with the average degree of proteins expressed by genes that are not essential or conditionally essential (see Table 2). We observe that in the hypergraph essential proteins have a higher mean degree than non-essential proteins and conditionally essential proteins lie between those. This indicates that the more multiprotein complexes proteins participate in, the more functionally important they are. We use a χ^2-test to investigate the null hypothesis that essential and non-essential proteins belong to the same population. We obtain a $\chi^2 \approx 1459$ with one degree of freedom (p-value $< 10^{-5}$) and reject the null hypothesis, which is strong evidence that hypergraph degrees of proteins can predict essentaility of genes.

mean degree	hypergraph	representing graph
essential	14.06	85.48
conditional	9.87	77.45
non-essential	6.444	110.89

Table 2: The mean degree $\langle k_i \rangle$ of nodes which connect to essential proteins (i.e., proteins expressed by essential genes), to conditionally essential proteins, and to non-essential proteins in the hypergraph and its representing graph.

In the representing graph, we do not observe that essential proteins tend to have a higher degree than non-essential proteins. This occurs because essential proteins tend to be connected to hyperedges of low cardinality. Our results indicate that the hypergraph representation is more fruitful than the representing graph for this application.

A further advantage of the hypergraph in comparison with the representing graph is that we can associate a protein complex with each hyperedge. While there is no information available whether a certain protein complex is essential, we may infer whether a complex is potentially essential from protein-essentiality data by asserting that only a multiprotein complex that has at least one composing protein that is essential may also be essential. In total, 811 out of 6688 complexes have at least one essential protein associated with them and are therefore *potentially essential* protein complexes (see Fig. 6). The Spearman correlation between the number $c_{\text{essential}}$ of essential proteins connected by a hyperedge and the hyperedge's cardinality c is 0.35 ($p < 10^{-193}$). This is in accordance with earlier findings on yeast that larger complexes tend to be more essential [42].

To test which of the protein complexes consist of more essential proteins than expected by chance, we construct a random null model: For each essential hyperedge of cardinality c we fix one essential protein in the hyperedge and sample $c-1$ proteins. Out of a total of 8243, proteins 108 are essential, which gives a density of essential proteins of $\rho_{\text{essential}} \approx 0.013$. Assuming that we pick $c-1$ proteins at random with replacement, independently of each other, each pick would have probability $\rho_{\text{essential}} \approx 0.013$ of being essential. Therefore, under this model, the number of essential proteins in an essential complex would follow a shifted binomial distribution with an expectation value of $E(c) = 1 + (c-1)\rho_{\text{essential}}$, which we show as dashed line in the lower panel of Fig. 6. In our data set, 246 out of 6688 protein complexes are essential and shown as red disks in Fig. 6. When comparing the fraction $c_{\text{essential}}/c$ of essential proteins to the size c of the complex, we observe an anticorrelation (see Figure in SI B). This indicates that larger complexes tend to be more essential because they are larger and not because they have a higher density of essential proteins.

Overall, this analysis indicates that the hypergraph degree is in better agreement with gene-essentiality information than the representing graph. Additionally, the hypergraph allows us to statistically investigate the essentiality of protein complexes.

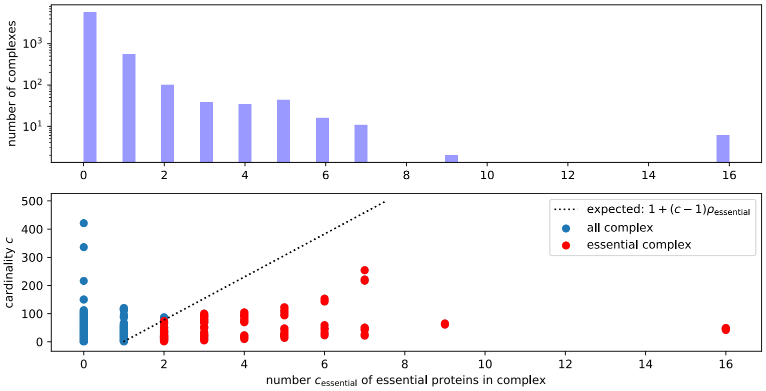

Figure 6: (Upper Panel) The distribution of the number $c_{\text{essential}}$ of essential proteins in multiprotein complexes. (Lower Panel) The number $c_{\text{essential}}$ of essential proteins versus the cardinality c of the complexes. We highlight complexes that have more essential proteins than expected under a random null model (dashed line) in red.

3.4 Hierarchy coefficient

The results above indicate that the hypergraph contains biological signal and conveys different information than the representing graph. Next, we find that the representing graph may have a hierarchical structure which arises solely from the translation of hyperedges into simple edges.

For many complex networks, it has been reported that the local clustering coefficient follows the degree in a power law, which has been interpreted as a sign of a hierarchical organisation [26]. First, we compute the local clustering for the hypergraph and for the representing graph. Fig. 7 shows that on average, the local clustering C_i is much larger in the graph representation than in the hypergraph. The average local clustering $\langle C_i \rangle_{\text{graph}} \approx 0.810$ in the graph is an order of magnitude larger than of the hypergraph $\langle C_i \rangle_{\text{hypergraph}} \approx 0.078$. We find that the local clustering between both graphs is anticorrelated (Spearman correlation -0.49). This is plausible because some high-cardinality hyperedges connect many different proteins, which leads to a local hypergraph clustering close to 0 but a local graph clustering close to 1 (see right panel of Fig. 7 for an extreme example). This indicates that constructing a representing graph from a hypergraph inflates the clustering coefficient for nodes incident to such high-cardinality hyperedges.

To explore the hierarchical organisation for hypergraphs, we investigate the relationship between clustering and degree for the hypergraph and the representing graph (see Fig. 8). For the hypergraph, we estimate the power-law exponent $\beta_{\text{hyp}} \approx 0.001$ (p-value 0.899) and thus do not observe a hierarchical

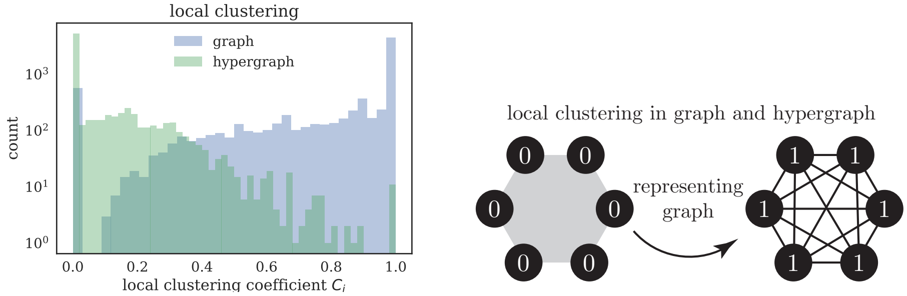

Figure 7: (Left Panel) The distribution of local clustering C_i differs strongly between hypergraph and graph. (Right Panel) An example showing that replacing hyperedges with high cardinality with a c-clique may inflate the local clustering from zero to one.

organisation. For the representing graph, however, we find $\beta_{\text{rep}} \approx 0.07$ (p-value 4×10^{-9}) and thus observe an organisation that appears to be 'hierarchical' in the sense of [26].

This indicates that hypergraphs that themselves do not show a statistically significant relationship between local clustering C_i and degree k_i can show a statistically significant relationship after being translated into their representing graph. This behaviour can be explained by the pairwise projection procedure in Fig. 7b: hyperedges of cardinality c that do not intersect with any other edges are replaced with a c-clique in the representing graph. This creates c nodes with degree $k_i = c-1$ and clustering coefficient $C_i = 1$. As we have many hyperedges with only small overlap with other hyperedges, we obtain many of such nodes in the representing graph, which creates an apparent hierarchical organisation.

4 Discussion

Multiprotein complexes are biological polyadic interactions between proteins that can be represented by hypergraphs. In this paper, we have used a hypergraph representation of the data and using two null models for hypergraphs have found that the data hypergraph contains signal beyond their cardinality sequence **c** and their degree sequence **k**. Similar results are well-established for protein interaction graphs but have not been tested for multiprotein complex hypergraphs [43]. By projecting the hypergraph into a graph representation, we illustrate that this simplification reveals different degree-rankings, which indicates that using both mathematical structures may reveal complementary information. In our test on human data, the hypergraph representation revealed a stronger correlation with gene-essentiality information than the representing graph. We then estimated the essentiality of protein complexes by comparing it with a null model and found that larger complexes tend to be more essential. In future work, one could investigate whether other hypergraph centralities (e.g.,

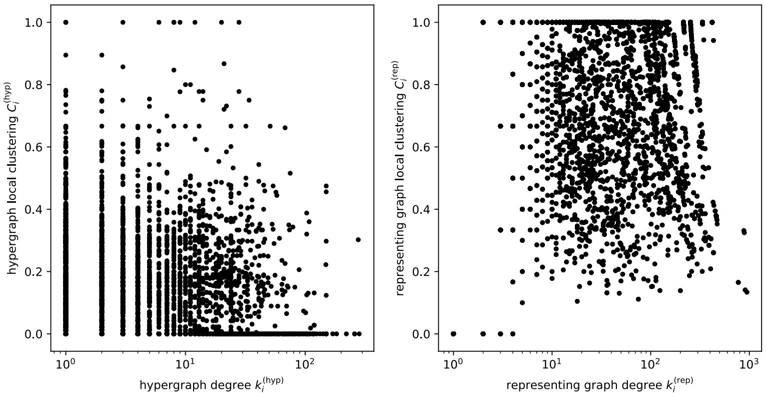

Figure 8: (Left Panel) Degree and local clustering for all nodes in the hypergraph. (Right Panel) Degree and local clustering for all nodes in the representing graph.

an eigenvector-based centrality [44]) are in even better agreement with essentiality.

Using an established definition of local clustering coefficient in hypergraphs, we defined the hierarchy coefficient for hypergraphs. We then showed that a pairwise graph may appear to show a hierarchical organisation while the hypergraph does not. As graphs are often constructed from polyadic interaction data, this finding reveals that such results might occur through the projection process and not the biological systems themselves.

In this study, we have demonstrated that hypergraphs are a fruitful representation of higher-order interactions between proteins. We did, however, ignore the stoichiometry (i.e., the number of proteins of a certain type that are involved in a complex). The investigation of a mathematical structure that incorporates such information might be a fruitful extension to our work. Furthermore, one could consider the different role (e.g., catalyst) that proteins have in chemical reactions and investigate them as *annotated hypergraphs* [45].

The formation of stable protein complexes investigated in this study is just one way in which proteins interact with each other. There exist many *transient* protein interactions that form and break on shorter time scales but are nevertheless of crucial biological importance [46]. An integrative analysis of pairwise protein interaction sources (e.g., BIOGRID) with multiprotein-complex data may reveal a more nuanced picture of the cellular processes than either data set on their own.

Hypergraphs and their null models might be used to analyse other data sets. Among them are association-based data (e.g., ingredient–product networks [47, 48], authorship networks [49], company–directorate networks [50]) or social networks in which higher-order interactions have been shown to be prominent [51, 52]. The tools in this manuscript could be used to investigate such data as hypergraphs and so reveal organisational principles beyond their pairwise interactions.

References

[1] Waqar Ali, Charlotte M Deane, and Gesine Reinert. Protein interaction networks and their statistical analysis. In Michael P. H. Stumpf, David J. Balding, and Mark Girolami, editors, *Handbook of Statistical Systems Biology*, pages 200–234. John Wiley & Sons, Ltd Chichester, UK, 2011.

[2] Marc Vidal, Michael E Cusick, and Albert-László Barabási. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011.

[3] Ernesto Estrada. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics*, 6(1):35–40, 2006.

[4] Xionglei He and Jianzhi Zhang. Why do hubs tend to be essential in protein networks? *PLoS Genetics*, 2(6), 2006.

[5] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41, 2001.

[6] Minoo Ashtiani, Ali Salehzadeh-Yazdi, Zahra Razaghi-Moghadam, Holger Hennig, Olaf Wolkenhauer, Mehdi Mirzaie, and Mohieddin Jafari. A systematic survey of centrality measures for protein-protein interaction networks. *BMC Systems Biology*, 12(1):80, 2018.

[7] Jose B Pereira-Leal, Emmanuel D Levy, and Sarah A Teichmann. The origins and evolution of functional modules: lessons from protein complexes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1467):507–517, 2006.

[8] Emmanuel D Levy, Jose B Pereira-Leal, Cyrus Chothia, and Sarah A Teichmann. 3D Complex: a structural classification of protein complexes. *PLoS Computational Biology*, 2(11):e155, 2006.

[9] Paul D Boyer. The ATP synthase—a splendid molecular machine. *Annual Review of Biochemistry*, 66(1):717–749, 1997.

[10] Or Matalon, Amnon Horovitz, and Emmanuel D Levy. Different subunits belonging to the same protein complex often exhibit discordant expression levels and evolutionary properties. *Current Opinion in Structural Biology*, 26:113–120, 2014.

[11] Claude Berge. *Hypergraphs: combinatorics of finite sets*, volume 45. Elsevier, 1984.

[12] Hervé Le Hir, David Gatfield, Elisa Izaurralde, and Melissa J Moore. The exon–exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *The EMBO Journal*, 20(17):4987–4997, 2001.

[13] Michael D Diem, Chia C Chan, Ihab Younis, and Gideon Dreyfuss. PYM binds the cytoplasmic exon-junction complex and ribosomes to enhance translation of spliced mRNAs. *Nature Structural & Molecular Biology*, 14(12):1173, 2007.

[14] Dane Taylor, Florian Klimm, Heather A Harrington, Miroslav Kramár, Konstantin Mischaikow, Mason A Porter, and Peter J Mucha. Topological data analysis of contagion maps for examining spreading processes on networks. *Nature Communications*, 6:7723, 2015.

[15] Chad Giusti, Robert Ghrist, and Danielle S Bassett. Two's company, three (or more) is a simplex. *Journal of Computational Neuroscience*, 41(1):1–14, 2016.

[16] Ginestra Bianconi and Robert M Ziff. Topological percolation on hyperbolic simplicial complexes. *Physical Review E*, 98(5):052308, 2018.

[17] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1):17, 2017.

[18] Michelle Feng and Mason A Porter. Spatial applications of topological data analysis: Cities, snowflakes, random structures, and spiders spinning under the influence. *arXiv preprint arXiv:2001.01872*, 2020.

[19] Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532, 2018.

[20] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS Computational Biology*, 5(5):e1000385, 2009.

[21] Wanding Zhou and Luay Nakhleh. Properties of metabolic graphs: biological organization or representation artifacts? *BMC Bioinformatics*, 12(1):132, 2011.

[22] Aziz Mithani, Gail M Preston, and Jotun Hein. Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison. *Bioinformatics*, 25(14):1831–1832, 2009.

[23] Ernesto Estrada and Juan A Rodríguez-Velázquez. Subgraph centrality and clustering in complex hyper-networks. *Physica A: Statistical Mechanics and its Applications*, 364:581–594, 2006.

[24] Philip S Chodrow. Configuration models of random hypergraphs and their applications. *arXiv preprint arXiv:1902.09302*, 2019.

[25] Wei-Hua Chen, Guanting Lu, Xiao Chen, Xing-Ming Zhao, and Peer Bork. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Research*, page gkw1013, 2016.

[26] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112, 2003.

[27] Florian Klimm, Danielle S Bassett, Jean M Carlson, and Peter J Mucha. Resolving structural variability in network models and the brain. *PLoS Computational Biology*, 10(3):e1003491, 2014.

[28] Soon-Hyung Yook, Zoltán N Oltvai, and Albert-László Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, 2004.

[29] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101, 2004.

[30] Eduardo López. The distribution of the number of node neighbors in random hypergraphs. *Journal of Physics A: Mathematical and Theoretical*, 46(30):305003, 2013.

[31] Oliver Cooley, Mihyun Kang, and Christoph Koch. The size of the giant high-order component in random hypergraphs. *Random Structures & Algorithms*, 53(2):238–288, 2018.

[32] David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(suppl_1):D691–D697, 2010.

[33] Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 32(suppl_1):D115–D119, 2004.

[34] Gourab Ghoshal, Vinko Zlatić, Guido Caldarelli, and Mark EJ Newman. Random hypergraphs and their applications. *Physical Review E*, 79(6):066118, 2009.

[35] Oliver Cooley, Wenjie Fang, Nicola Del Giudice, and Mihyun Kang. Subcritical random hypergraphs, high-order components, and hypertrees. In *2019 Proceedings of the Sixteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 111–118. SIAM, 2019.

[36] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1):17–60, 1960.

[37] Christina Goldschmidt and James Norris. Essential edges in poisson random hypergraphs. *Random Structures & Algorithms*, 24(4):381–396, 2004.

[38] Richard WR Darling, James R Norris, et al. Structure of large random hypergraphs. *The Annals of Applied Probability*, 15(1A):125–152, 2005.

[39] Caroline C Friedel and Ralf Zimmer. Inferring topology from clustering coefficients in protein-protein interaction networks. *BMC Bioinformatics*, 7(1):519, 2006.

[40] David Komander and Michael Rape. The ubiquitin code. *Annual Review of Biochemistry*, 81:203–229, 2012.

[41] Ulf Aslak and Benjamin F Maier. Netwulf: Interactive visualization of networks in python. *The Journal of Open Source Software*, 4, 2019.

[42] Haidong Wang, Boyko Kakaradov, Sean R Collins, Lena Karotki, Dorothea Fiedler, Michael Shales, Kevan M Shokat, Tobias C Walther, Nevan J Krogan, and Daphne Koller. A complex-based reconstruction of the saccharomyces cerevisiae interactome. *Molecular & Cellular Proteomics*, 8(6):1361–1381, 2009.

[43] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

[44] Austin R Benson. Three hypergraph eigenvector centralities. *SIAM Journal on Mathematics of Data Science*, 1(2):293–312, 2019.

[45] Philip Chodrow and Andrew Mellor. Annotated hypergraphs: Models and applications. *arXiv preprint arXiv:1911.01331*, 2019.

[46] James R Perkins, Ilhem Diboun, Benoit H Dessailly, Jon G Lees, and Christine Orengo. Transient protein-protein interactions: structural, functional, and network properties. *Structure*, 18(10):1233–1243, 2010.

[47] Vaiva Vasiliauskaite and Tim S Evans. Social success of perfumes. *PloS One*, 14(7):e0218664, 2019.

[48] Sarah M Griffin and Florian Klimm. Networks and museum collections. *Oxford Handbook of Archaeological Network Research*, 2020.

[49] Mark EJ Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, 101(suppl 1):5200–5205, 2004.

[50] Nial Friel, Riccardo Rastelli, Jason Wyse, and Adrian E Raftery. Interlocking directorates in irish companies using a latent space model for bipartite networks. *Proceedings of the National Academy of Sciences of the United States of America*, 113(24):6629–6634, 2016.

[51] Vedran Sekara, Arkadiusz Stopczynski, and Sune Lehmann. Fundamental structures of dynamic social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 113(36):9977–9982, 2016.

[52] Eduardo López. Weighted projected networks: mapping hypergraphs to networks. *Physical Review E*, 87(5):052813, 2013.

[53] Jeremy T Blitzer and Roel Nusse. A critical role for endocytosis in Wnt signaling. *BMC Cell Biology*, 7(1):28, 2006.

[54] Catriona Y Logan and Roel Nusse. The Wnt signaling pathway in development and disease. *Annual Review of Cell and Developmental Biology*, 20:781–810, 2004.

5 Data and code availability

We make the code and data available on GitHub under `https://github.com/floklimm/hypergraph`.

6 Funding

F.K. was supported by the Engineering and Physical Sciences Research Council (funding references EP/R513295/1 and EP/N014529/1). G.R. was supported in part by the EPSRC grant New Approaches to Data Science: Application Driven Topological Data Analysis EP/R018472/1.

7 Acknowledgements

We thank Christopher Koch, Christina Goldschmidt, and Phil Chodrow for fruitful discussions.

Supplementary Information

A Dual graph

A.1 Definition

The *dual graph* $D(\mathcal{H}) = (V', E')$ (also called *line graph*) of a hypergraph $\mathcal{H} = (V, E)$ is the graph whose vertex set V' is the set of hyperedges of the hypergraph with edges between them if the hyperedges have at least one node in common (i.e., $V' = E$ and $(e_i, e_j) \in E' \leftrightarrow e_i \cap e_j \neq \emptyset$). In Fig. 9, we show an example of the dual graph constructed from a hypergraph.

A.2 Results for the dual graph

In Fig. 10 we compare the hypergraph with its dual graph, keeping in mind that in the dual graph $D(\mathcal{H})$ the nodes represent the edges of the hypergraph $\mathcal{H}$. Therefore, we compare the degree $k_i^{(\mathrm{dual})}$ of dual graph nodes V' with the cardinality of the hyperedges E in the original hypergraph (see Fig. 10). There are 1717 hyperedges with minimum cardinality $c_{\min} = 2$. This is the cardinality that occurs most often. The mean cardinality is $\langle c_e \rangle \approx 10.12$. and the mean degree is $\langle k_i^{(\mathrm{dual})} \rangle \approx 8.21$.

In the dual graph $D(\mathcal{H})$, we also investigate the degrees of nodes nodes and compare it with the cardinality of the hyperedges E in the original hypergraph $\mathcal{H}$. The Pearson correlation is -0.03 and the Spearman correlation of -0.03, indicating that the size of the complex and the degree of its associated node in the dual graph are almost uncorrelated.

The node in $D(\mathcal{H})$ with the highest degree 960 represents a Wnt complex and has a cardinality of 119. We observe that there are multiple other nodes that have a slightly lower degree and similar cardinality. All of these nodes in $D(\mathcal{H})$ represent complexes that are also involved in the Wnt signaling pathway and are active in the clathrin-coated endocytic vesicle membrane, which plays a critical role in the Wnt signaling pathway [53]. This pathway itself is crucial for stem cell development and disease progression [54]. The complex with the highest cardinality of 421 is the 'Olfactory receptor–G protein trimer'. It has a degree of 83. The second highest cardinality has the 'KRAB-ZNF / KAP Complex' with a cardinality of 336 and a degree of 6.

This investigation illustrates that degree of the dual graph and cardinality of the protein complexes identify distinct protein as high ranked. Both approaches reveal protein complexes of crucial cellular function and are therefore fruitful strategies to investigate cellular hypergraphs.

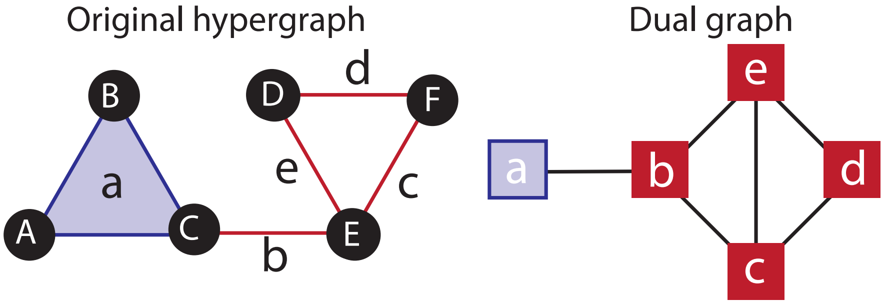

Figure 9: Example of a *dual graph* constructed from a hypergraph. Each hyperedge is represented by a node. These nodes are connected if the hyperedges share a node.

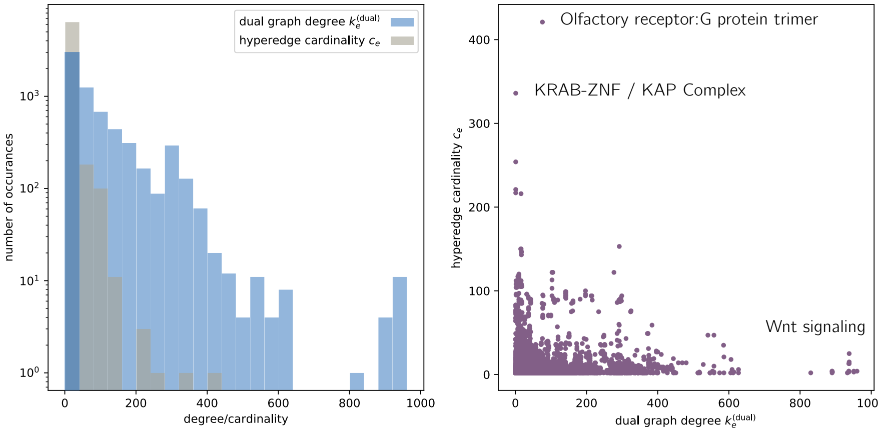

Figure 10: (Left) Distribution of c_e cardinality of hyperedges in H and the degrees of the associated nodes in the dual graph $D(H)$ (Right) Scatterplot of these two.

B Fraction of essential proteins in multiprotein complexes

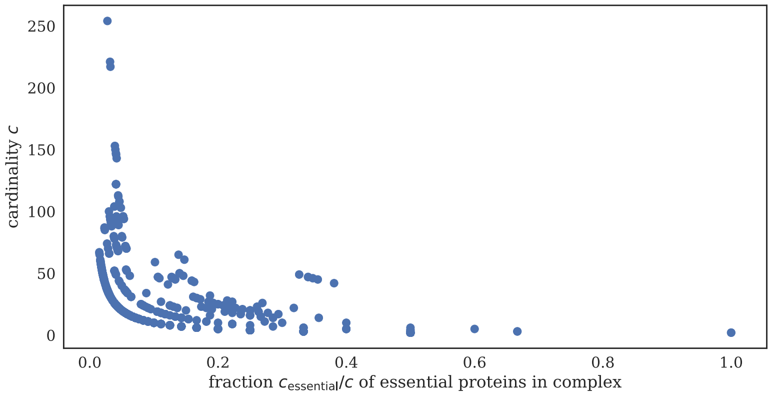

Figure 11: The fraction $c_{\text{essential}}/c$ of essential proteins in complexes, in dependence of the cardinality c. The two are anticorrelated, i.e., larger protein complexes tend to have a smaller fraction of essential proteins.

C Distribution of hypergraph measures for the null models

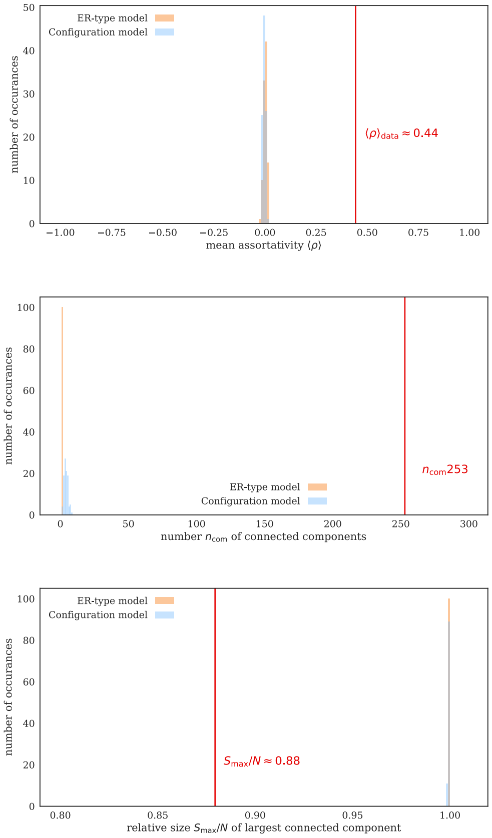

Figure 12: The distribution of mean assortativity $\langle\rho\rangle$, number n_{com} of components, and relative size S_{max}/N of the largest component for the ER-type model (orange) and the configuration model (blue) for 100 realisations. The mean assortativity $\langle\rho\rangle_{\mathrm{data}} \approx 0.44$ of the protein hypergraph (red vertical line) is clearly larger than for these null models. The number $n_{\mathrm{com}} = 253$ of components is also larger for the protein hypergraph. The relative size $S_{\mathrm{max}}/N \approx 0.88$ of the largest component is smaller.

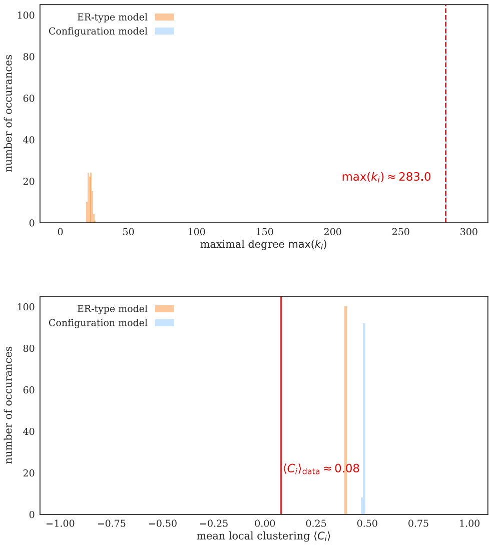

Figure 13: The distribution of the maximal degree $\max(k_i)$ and mean local clustering $\langle C_i \rangle$ for the ER-type model (orange) and the configuration model (blue) for 100 realisations. The maximal degree is (by construction) the same for the configuration model as for the protein hypergraph. The ER-type model has a much smaller maximum degree. The mean local clustering $\langle C_i \rangle \approx 0.8$ is smaller for the data than for the null models.