Article

Distances in Higher-Order Networks and the Metric Structure of Hypergraphs

Ekaterina Vasilyeva [1,2,*], Miguel Romance [3,4], Ivan Samoylenko [1,5], Kirill Kovalenko [6], Daniil Musatov [1,7,8], Andrey Mihailovich Raigorodskii [1,8,9,10] and Stefano Boccaletti [1,3,11]

1 The Phystech School of Applied Mathematics and Computer Science, Moscow Institute of Physics and Technology, Institutskiy per., 9, 141701 Dolgoprudny, Moscow Region, Russia
2 P.N. Lebedev Physical Institute of the Russian Academy of Sciences, Leninsky Prosp., 53, 119991 Moscow, Russia
3 Departamento de Matemática Aplicada, Ciencia e Ingeniería de los Materiales y Tecnología Electrónica, Universidad Rey Juan Carlos, Calle Tulipán s/n, Móstoles, 28933 Madrid, Spain
4 Mathematical Computation Laboratory on Complex Networks and Its Appliactions, Universidad Rey Juan Carlos, Calle Tulipán s/n, Móstoles, 28933 Madrid, Spain
5 Faculty of Mathematics, National Research University Higher School of Economics, Usacheva str., 6, 119048 Moscow, Russia
6 Scuola Superiore Meridionale, Largo S. Marcellino, 10, 80138 Napoli, NA, Italy
7 Institute of Economics, Mathematics and Information Technology, Russian Academy of National Economy and Public Administration, pr. Vernadskogo, 84, 119606 Moscow, Russia
8 Caucasus Mathematical Center, Adyghe State University, ul. Pervomaiskaya, 208, 385000 Maykop, The Republic of Adygea, Russia
9 Mechanics and Mathematics Faculty, Moscow State University, Leninskie Gory, 1, 119991 Moscow, Russia
10 Institute of Mathematics and Computer Science, Buryat State University, ul. Ranzhurova, 5, 670000 Ulan-Ude, The Republic of Buryatia, Russia
11 CNR—Institute of Complex Systems, Via Madonna del Piano 10, 50019 Sesto Fiorentino, FI, Italy
* Correspondence: serebryannikovaee@lebedev.ru

Abstract: We explore the metric structure of networks with higher-order interactions and introduce a novel definition of distance for hypergraphs that extends the classic methods reported in the literature. The new metric incorporates two critical factors: (1) the inter-node distance within each hyperedge, and (2) the distance between hyperedges in the network. As such, it involves the computation of distances in a weighted line graph of the hypergraph. The approach is illustrated with several ad hoc synthetic hypergraphs, where the structural information unveiled by the novel metric is highlighted. Moreover, the method's performance and effectiveness are shown through computations on large real-world hypergraphs, which indeed reveal new insights into the structural features of networks beyond pairwise interactions. Namely, using the new distance measure, we generalize the definitions of efficiency, closeness and betweenness centrality for the case of hypergraphs. Comparing the values of these generalized measures with their analogs calculated for the hypergraph clique projections, we show that our measures provide significantly different assessments on the characteristics (and roles) of the nodes from the information-transferability point of view. The difference is brighter for hypergraphs in which hyperedges of large sizes are frequent, and nodes relating to these hyperedges are rarely connected by other hyperedges of smaller sizes.

Keywords: hypergraphs; higher-order networks; distances on hypergraphs; linegraphs

1. Introduction

The modeling of social, biological, and technological systems as complex networks (or graphs) has gained significant attention across areas of research as diverse as computer science, statistical physics, biology, neuroscience and social science, among others. Network science has, in fact, become ubiquitous [1]. In particular, as it was found that there is a deep interplay between the structural and dynamical features of networked systems, the

analysis of the network topology therefore helps in understanding the network functional properties, such as robustness, efficiency in information transmission, and resiliency [1]. Metric parameters play a central role in deciphering information diffusion processes. Indeed, complex networks are renowned for their ability to efficiently transfer information or signal(s) from node to node; a proper way to quantitatively analyze the information flow is through the use of structural measures, such as the diameter, the characteristic path length, the closeness, the betweenness centralities, the efficiency, etc. [1–3].

The basis for the definition of all metric measures is the concept of *distance* between two nodes. If one deals with only dyadic node–node interactions, then there are no doubts as to how to define the distance between nodes, as the only way to construe a path from node s to node t is by means of a sequence of intersecting edges with the first one containing s and the last one containing t. Then, the corresponding path length is just the number of edges in the series, and the distance (or geodesic distance) between the two nodes is just the length of the shortest path.

However, the description of many processes in the real world requires to account for group interactions, i.e., interactions of more than two nodes, and it is, therefore, critical to refer to other mathematical objects that allow a better representation of such higher-order relationships [4,5]. These objects are called hypergraphs, and the quantification of efficiency of the spreading processes needs here the effort to extend the classical definitions of graph structural measures to such a hypergraph setting.

The simplest generalization to the higher-order case of any graph topological measure is just its computation in the hypergraph clique projection, a representation where each hyperedge is replaced by a clique of pairwise interactions among all its nodes. However, considering the clique projection, one obviously loses very important information about the structure of such higher-order interactions. This was brightly demonstrated in the series of recent studies, where it was shown that more complex generalizations of topological measures were needed in order to provide significant and meaningful information about the hypergraph structure [6–11]. In our study, we introduce a new concept of distance among nodes in a hypergraph.

Moreover, a series of fundamental macroscopic network topology descriptors (such as the betweenness centrality, the closeness centrality, or the efficiency) are explicitly based on the concept of metric structure or distance in the graph. As we will show further in more detail, the current literature contains some attempts to generalize the concept of distance [4,7,8,12–14], which, however, are not systematic, and are mostly only on one of the distinguishing features that higher-order structures display.

A first, naive extension for the term distance would be just replacing the word "edge" by the word "hyperedge" in the graph's path definition given above, and leaving the rest of the definition unchanged. This way, one would obtain the so-called "distance in the hypergraph clique projection". While such a definition frequently appears in the literature [15–22], it actually oversimplifies the rich structure of higher-order interactions. Indeed, in the case of a classic (unweighted) graph, all edges are roughly equivalent, as they all have the same cardinality (equal to two), and furthermore, all their intersections have the same cardinality (equal to one). This is not, however, the case of higher-order systems, where hyperedges may differ in their size and may intersect differently. Therefore, it seems reasonable that such extra information be somehow considered in the hypergraph distance definition.

To make an illustrative example, let us refer to the two panels of Figure 1. In both cases, the hypergraphs are composed of two communities (of arbitrary internal structure) and contain hyperedges bridging the communities. Let us consider the hypergraph of Figure 1a, and let us suppose that a random walker starts at node u in the first community. Then, the probabilities to get from node u to node v by using no more than two steps either through the path including only the hyperedge e_1 (p_1^a) or by using the path e_2, e_3 ($p_{2,3}^a$) are

$$p_1^a = \underbrace{\frac{1}{k_u} \cdot \frac{1}{|e_1|-1}}_{v \text{ is chosen randomly on the first step}} + \underbrace{\frac{1}{k_u} \cdot \frac{|e_1|-2}{|e_1|-1} \cdot \frac{1}{|e_1|-1}}_{\text{some other node from } e_1 \text{ is chosen on the first step}} = \frac{1}{k_u} \cdot \frac{2|e_1|-3}{(|e_1|-1)^2},$$

$$p_{2,3}^a = \frac{1}{k_u} \cdot \frac{1}{2},$$

where $|e_1|$ is the edge e_1's cardinality (in Figure 1a $|e_1| = 9$) and k_u is the number of hyperedges incident to node u. It is clear that p_1^a is significantly smaller than $p_{2,3}^a$, so it is natural to suppose that the length of the path $\{e_1\}$ should be larger than the length of the path $\{e_2, e_3\}$.

Now, if one considers instead the hypergraph of Figure 1b and, as before, compares the probabilities $p_{1,2}^b$ with $p_{3,4}^b$ of getting from node u to node v by using no more than two steps, then one has

$$p_{1,2}^b = \frac{1}{k_u} \cdot \frac{2}{5} \cdot \frac{1}{2} \cdot \frac{1}{5},$$

$$p_{3,4}^b = \frac{1}{k_u} \cdot \frac{1}{5} \cdot \frac{1}{2} \cdot \frac{1}{5}.$$

Now, one has $p_{1,2}^b = 2p_{3,4}^b$, and therefore the path $\{e_1, e_2\}$ should be assigned a distance shorter than that assigned to $\{e_3, e_4\}$.
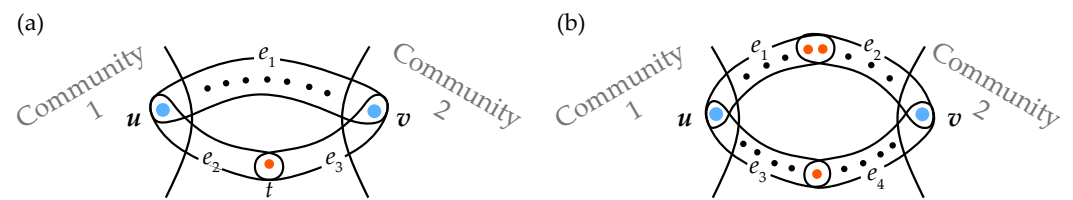
(a)   (b)

Figure 1. Illustrative sketch of two hypergraphs made of two communities of nodes, which are bridged by a series of hyperedges. Hypergraph in the panel (**a**) is used to provide explanation of the hyperedges' cardinality importance for the path length calculation, while the one in the panel (**b**) illustrates the impact of hyperedges intersection size (see explanation in the text).

As a consequence, the concept of *length of a path* in a hypergraph should not only take into account the number of hyperedges involved but also their size and the size of their intersection: the bigger the size of a hyperedge, the larger the associated distance should be, and the bigger the intersection between hyperedges, the closer the nodes should be. Such two distinguishing features (varying sizes of hyperedges and sizes of hyperedges intersections) are indeed *separately* discussed in the literature, with the following in particular:

(1) If one assigns weights to hyperedges as proper functions of their sizes, then the distance between a couple of nodes s, t is the sum of the weights of all hyperedges in the shortest path, i.e.,

$$d(s,t) = \min_{P \in \mathbf{P}_{st}} \left(\sum_{e \in P} f(|e|) \right), \tag{1}$$

where $\mathbf{P}_{st}$ is the set of paths between s and t and $f(|e|)$ is the weight of the hyperedge e of size $|e|$. The heuristics behind this definition come from social networks: if one assumes that the size of each hyperedge is the size of a company of friends, then when the company size grows, it becomes more difficult for the members of the network to keep in touch with each other. This idea was used in Refs. [7,14], where the authors define hypergraph random walks, arguing that "*the walkers may give more or less importance to hyperedges depending on their size*".

(2) The difference in the size of intersections between hyperedges was taken into account in the definition of *k*-walks introduced in Refs. [4,8,12,13]. A *k*-walk is a sequence of hyperedges such that each pair of successive hyperedges are adjacent, and they intersect in at least *k* vertices.

Despite the fact that both ingredients are separately considered in the structural analysis of hypergraphs, these two features have never, so far, been considered to be present together for the description of the metric structure of hypergraphs. The main goal of this paper is, therefore, to combine them in order to obtain a tailored metric structure for hypergraphs that would properly extend that of classic (dyadic) networks. Once this structure is defined (in the next section), network efficiency, closeness and betweenness centralities can also be defined as basic metric structural measures of the hypernetwork topology.

Our manuscript includes, therefore, three novel contributions to the literature. First, it introduces a new distance measure. Second, it gives the generalizations for the three crucial characteristics of network topology listed above. Third, it provides a series of examples showing the efficiency of the new concepts in uncovering meaningful information about higher-order networks. An illustration of such a novel approach is made with specific synthetic networks, where one immediately realizes that it provides important information on the properties of information transfer through the networks. Finally, our numerical testings on real-world higher-order networks reveal that the new metric measures uncover several important features of the hypergraphs properties. In particular, we show that our measures reveal different assessments of the nodes importance and of other network's features from the information-transferability point of view. Finally, we show that our measure is particularly suitable to describe the structure of hypergraphs, which are highly distinguishable from graphs, i.e., in which hyperedges of large size are frequent and nodes in these hyperedges are rarely connected by other hyperedges of smaller sizes.

2. Methods

The basic notation used here is the same as that of Ref. [5]. In particular, a (dyadic) *complex network* is a (undirected and unweighted) network $\mathcal{G} = (V, E)$, where $V = \{1, \cdots, n\}$ for some $n \in \mathbb{N}$ is the set of *nodes* (or vertices) and E is the set of *links* (or edges), i.e., a finite family of (unordered) pairs of nodes. Throughout this paper, only undirected and unweighted networks are considered, but all our results can be straightforwardly extended to the case of directed and weighted networks. A *hypergraph* is a pair $\mathcal{H} = (V, H)$, where $V = \{1, \cdots, n\}$ is a (finite) set of nodes and $H = \{e_1, \cdots, e_m\}$ is a (finite) family of (non-empty) subsets of V, i.e., $e_i \subseteq V$ for all $1 \leq i \leq m$. Each element of H is called a *hyperedge* of $\mathcal{H}$.

Notice that every complex network is a hypergraph, but the reverse is not true. Despite this latter fact, if one takes a hypergraph $\mathcal{H} = (V, H)$, one can define its complex network projection $\pi_2(\mathcal{H}) = (V, \pi_2(H))$, where each hyperedge $e_i \in H$ is replaced by the clique made of all the pairwise interactions among the hyperedge's nodes, i.e., $\pi_2(H) = \{\{i, j\}; \ i, j \in e_k \text{ for every } e_k \in H\}$. $\pi_2(\mathcal{H})$ is called the *clique projection* of $\mathcal{H}$.

In order to introduce a metric structure in $\mathcal{H} = (V, H)$ (i.e., a notion of distance between nodes of $\mathcal{H}$), we first need to establish the notion of paths in $\mathcal{H}$. If one takes a pair of nodes $i, j \in V$, a *path* from i to j is a (finite) sequence of hyperedges $(\tilde{e}_1, \cdots, \tilde{e}_k) \in H$ such that $i \in \tilde{e}_1$, $j \in \tilde{e}_k$ and $\tilde{e}_\ell \cap \tilde{e}_{\ell+1} \neq \varnothing$ for all $1 \leq \ell < k$. One can associate a (positive) real number to every path from i to j (the *length* of the path), and then the distance between i and j will be the minimal length among all the paths from i to j, so the critical point, as we will see in Section 2.1, is how to define the length for each path in a hypergraph.

The missing mathematical ingredient is the concept of the linegraph. Given a hypergraph $\mathcal{H} = (V, H)$, its (unweighted and undirected) *linegraph* $L(\mathcal{H}) = (H, L(H))$ is a graph, whose node set is the set of hyperedges of $\mathcal{H}$ and for which there is a link between $e_i, e_j \in H$ if $e_i \cap e_j \neq \varnothing$. Notice that the linegraph has self-loops in all its nodes since for every $e_i \in H$, one has that $e_i \cap e_i \neq \varnothing$. A weighted linegraph is introduced in Section 2.1 in

order to define the path length in a hypergraph by assigning a non-negative weight to each link in $L(\mathcal{H})$. Further properties and results about hypergraphs and their linegraphs can be found in Refs. [4,5].

2.1. Weighted Linegraphs

The use of weights in the links of the linegraph allows quantifying either the size of the hyperedges and that of the intersections among hyperedges. Given a hypergraph $\mathcal{H}$, there is not a unique way of assigning weights to the links of its linegraph, so a weight distribution mechanism must be included. As we will show momentarily, the weights of the self-loops (w_{ii}) depend only on the size of corresponding hyperedge e_i, while the weights of the edges in the line graph are calculated as proper functions of the corresponding hyperedges' sizes and the size of the intersection of these hyperedges.

For the sake of illustration of our definitions, let us initially refer to the hypergraph (made of 11 nodes and 3 hyperedges) depicted in Figure 2, and its associated linegraph.
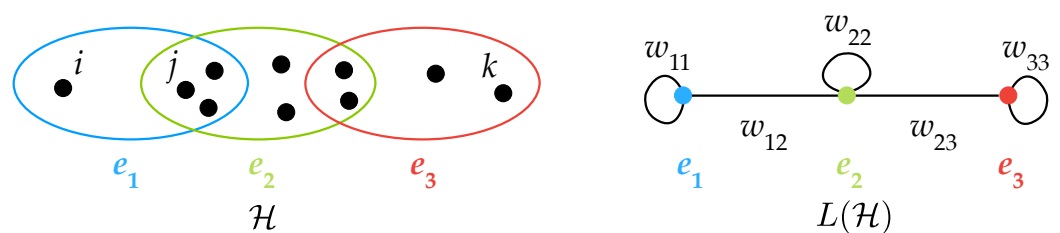

Figure 2. A 11-node and 3-hyperedge hypergraph $\mathcal{H}$ and its weighted line graph with self-loops $L(\mathcal{H})$

For every node $i \in V$ of a hypergraph $\mathcal{H} = (V, H)$, one can denote by E_i the set of hyperedges incident to the node i (in the example of Figure 2, $E_i = \{e_1\}$, $E_j = \{e_1, e_2\}$ and $E_k = \{e_3\}$). Then, one can define the *weighted distance* between the pair of nodes $i \neq k \in V$ as

$$d^w(i, k) = 1 + \min_{e \in E_i, \tilde{e} \in E_k} d^L(e, \tilde{e}), \tag{2}$$

where $d^L(e, \tilde{e})$ is the distance between the hyperedges e and $\tilde{e}$ ($e \neq \tilde{e}$) in the weighted linegraph of $\mathcal{H}$, while $d^L(e, e)$ is the weight of the self-loop $\{e, e\}$ in the weighted linegraph. The critical point is now defining the weights of the linegraph in order to give sense to the definition given above.

In particular, the weight function $w : L(H) \longrightarrow [0, +\infty)$ on the linegraph $L(\mathcal{H}) = (V, L(H))$ must have the following properties:

1. The distance between the nodes obtained from Formula (2) must be the same as the classic distance in a graph at all times that one considers a (dyadic) complex network $\mathcal{G} = (V, E)$ as a hypergraph.
2. The bigger the intersection of the hyperedges, the smaller the distance should be. For instance, if one considers the case illustrated in Figure 3, then the weighted distance between i and j in panel (a) should be smaller than that in panel (b). Therefore, the distance should be inversely proportional to the intersection size.
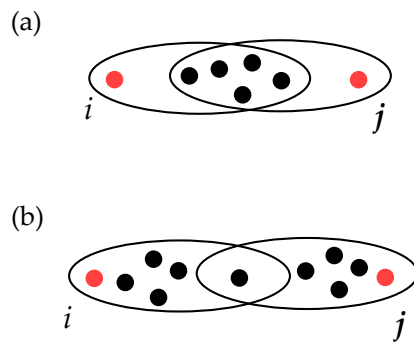
(a)

(b)

Figure 3. Two paths from nodes *i* to *j* in a hypergraph, each of them with 2 hyperedges of the same size but with different intersection size. The weighted path length in panel (**a**) should be smaller than that in panel (**b**).

3. The bigger the hyperedges involved, the larger the weighted distance should be. In particular, taking as an illustrative example the case of Figure 4, the weighted distance in panel (b) should be larger than that in panel (a) since the sizes of edges are bigger in case (b), while the size of the intersection is the same.
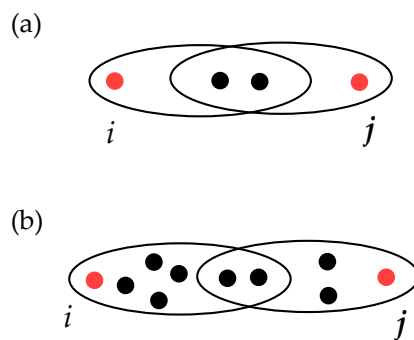
(a)

(b)

Figure 4. Two paths from nodes *i* to *j* in a hypergraph, each of them with two hyperedges with the same intersection size, but with different number of nodes in each hyperedge. The weighted path length in panel (**a**) should be smaller than that in panel (**b**).

4. Finally, the larger is the number of hyperedges involved in the path, the larger weighed distances one should obtain. In other words, the new metric structure should be sensitive to the number of hyperedges in the paths considered. In particular, if one takes two paths, one with only one hyperedge and another with two hyperedges but with the same number of nodes involved in both cases, then the path length should be smaller in the first case (see the illustration in Figure 5, where panel (b) must give a larger distance with respect to the case of panel (a)).
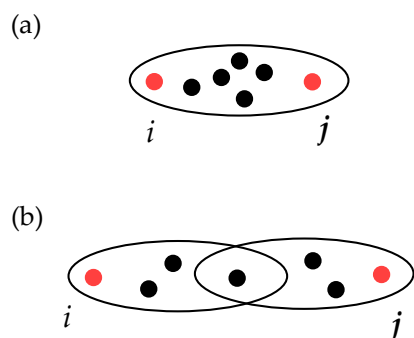
(a)

(b)

Figure 5. Two paths from nodes *i* to *j* in a hypergraph, one involving only one hyperedge and another with two hyperedges. Both paths include the same number of nodes. The weighted path length in panel (**a**) should be smaller than that in panel (**b**).

A possible choice for the weight function $w : L(H) \longrightarrow [0, +\infty)$ that verifies the four desired properties is to assign, for every $(e_i, e_j) \in L(H)$, a weight

$$w_{ij} = w_{ji} = w(|e_i \cup e_j|, |e_i \cap e_j|) = \frac{1}{3}\left(|e_i \cup e_j| + \frac{|e_i \cup e_j|}{|e_i \cap e_j|}\right) - 1. \tag{3}$$

It is easy, indeed, to verify that the function in Equation (3) accomplishes the following properties:

- As the Jaccard index [23] between e_i and e_j is given by $\mathcal{J}(e_i, e_j) = \frac{|e_i \cap e_j|}{|e_i \cup e_j|}$, then

$$w_{ij} = \frac{1}{3}\left(|e_i \cup e_j| + \frac{1}{\mathcal{J}(e_i, e_j)}\right) - 1.$$

- If one takes $e_i = e_j$, then $w_{ii} = \frac{1}{3}(|e_i| + 1) - 1$. Furthermore, if one starts from a (dyadic) network $\mathcal{G} = (V, E)$, then $w_{ii} = 0$ for every $e_i \in E$, and if $e_i \neq e_j$ (but e_i connected with e_j in L), then $|e_i \cup e_j| = 3$ and $|e_i \cap e_j| = 1$, which make that $w_{ij} = 1$. Hence, one has that for this choice of weight function, the distance between the nodes obtained from formula (2) is the same as the classic distance in a graph (first desired property).
- Properties 2–4 also hold for this choice of weight function.

Let us discuss explicitly the issue of the computational complexity associated with the calculation of our distance. It is reasonable to analyze separately two steps of the distance calculations: the construction of the weighted linegraph and the distance calculation per se. As for the construction of the weighted linegraph, the worst-case estimation of the complexity is, obviously, $O(|H|^2)$. Moreover, looking at Formula (2), estimating the distance $d^w(i, k)$ between nodes i and k needs the computation of the distances between all pairs of nodes $e \in E_i$ and $\tilde{e} \in E_k$ in the weighted linegraph $L(\mathcal{H})$. The complexity associated with computing $d^L(e, \tilde{e})$ is therefore given by the worst-case estimate for Dijkstra's algorithm, which is $O(|L(H)| + |H| \log |H|)$ [24]. The resulting estimate of the computational complexity is $O(|E_i| \cdot |E_k| \cdot (|L(H)| + |H| \log |H|))$.

To conclude this section, we remark that the choice of the weighting function in Equation (3) is not, in fact, unique, in the sense that other choices can be made that equally satisfy the four properties desired for a distance.

2.2. Some Structural Measures

The metric structure of a networked systems plays a central role in understanding the dynamics and processes that take place on them, including robustness [25], diffusion [1] and resilience [26]. There is a plethora of measures related to the metric structure of networks in the literature, and we will here limit ourselves to consider network efficiency, closeness and betweenness centrality, and to discuss their extension to the hypergraph setting by using the weighted distance introduced in Section 2.1.

The (global) efficiency of a graph, introduced in Ref. [27], measures the performance of the network information transfer. Formally, given a complex network $\mathcal{G} = (V, E)$ of n nodes, its efficiency is defined as

$$E(\mathcal{G}) = \frac{1}{n(n-1)} \sum_{i,j \in V, \, i \neq j} \frac{1}{d(i, j)}, \tag{4}$$

where $d(i, j)$ is the distance between nodes i and j.

It is straightforward to extend this measure for a hypergraph $\mathcal{H} = (V, H)$ of n nodes by simply considering the metric structure in $\mathcal{H}$. As we discussed already in the introduction, one could consider a simple approach and define a metric structure in $\mathcal{H}$ given by its clique projection network $\pi_2(\mathcal{H})$. By this metric, one could define the efficiency of $\mathcal{H}$ as

$E(\mathcal{H}) = E(\pi_2(\mathcal{H}))$. On the other hand, if one considers the weighted metric structure defined by Equation (2), then one can calculate the weighted efficiency of the hypergraph $\mathcal{H}$ as

$$E^w(\mathcal{H}) = \frac{1}{n(n-1)} \sum_{i,j \in V,\, i \neq j} \frac{1}{d^w(i,j)}. \tag{5}$$

Similarly, one can introduce a weighted hypergraph analogue of closeness centrality. If one takes again a network $\mathcal{G} = (V, E)$, then the closeness centrality [1–3] of node $i \in V$ is

$$C_i(\mathcal{G}) = \frac{1}{\sum_{j \in V, j \neq i} d(i,j)}. \tag{6}$$

Notice that the top nodes (according to the closeness centrality values) can be seen as the most aware nodes in an information/social network [1–3]. By using expression (6), it is easy to transfer this concept to the hypergraph setting, simply by computing

$$C_i^w(\mathcal{H}) = \frac{1}{\sum_{j \in V, j \neq i} d^w(i,j)}, \tag{7}$$

for every node i in a hypergraph $\mathcal{H} = (V, H)$. Additionally, in this case, if one considers the metric structure given by the clique projection $\pi_2(\mathcal{H})$, then one can define an alternative closeness centrality as $C_i(\mathcal{H}) = C_i(\pi_2(\mathcal{H}))$ for every node i in the hypergraph.

Finally, the weighted centrality of a node i in a hypergraph $\mathcal{H} = (V, H)$ can also be defined, but in this case, one should pay attention to some extra remarks. Let us remind indeed that, if one takes a network $\mathcal{G} = (V, E)$, the betweenness centrality of each node i is given by

$$b_i(\mathcal{G}) = \sum_{u,v \in V, u \neq v, u \neq i, v \neq i} \frac{|P(u,v,i)|}{|P(u,v)|}, \tag{8}$$

where $P(u,v)$ is the set of all shortest paths between u and v, whereas $P(u,v,i)$ is the set of shortest paths between u and v that pass through i and $|\cdot|$ is the set cardinality operator, i.e., it is the sum of fractions of shortest paths between all possible nodes pairs in the graph going through node i.

If one wants to extend this notion to a general hypergraph $\mathcal{H} = (V, H)$, one should specify first what it means that a path from u to v goes through node t when the path is a series of hyperedges. Following Ref. [18], we assume that if we have a path described by the sequence of hyperedges $\{e_1, \ldots, e_\ell\}$, where $u \in e_1$ and $v \in e_\ell$, then it provides us with a sequence of nodes $\{u, t_1, t_2, \ldots, t_{l-1}, v\}$, where $t_i \in e_i \cap e_{i+1}$ (nodes t_i, $i = 1, \ldots, l-1$ are called *bridging nodes* in Ref. [18]). Once we fix this notion, the weighted betweenness centrality of a node i in a hypergraph $\mathcal{H} = (V, H)$ is defined as

$$b_i^w(\mathcal{H}) = \sum_{u,v \in V, u \neq v, u \neq i, v \neq i} \frac{\sum_{p \in P_e^w(u,v,i)} \frac{1}{int_i(p)}}{|P_e^w(u,v)|}, \tag{9}$$

where $P_e^w(u,v)$ is the set of shortest paths between nodes u and v whose lengths are calculated by using the proposed weighted hypergraph measure, and the elements of the set are all possible sequences of hyperedges forming a path from node u to node v; $P_e^w(u,v,i)$ is the set of those shortest paths between nodes u and v, for which the intersection of one pair (or some pairs) of hyperedges includes node i; and $int_i(p)$ is the size of such a latter intersection. Once again, if one considers the clique projection $\pi_2(\mathcal{H})$, then one could define an alternative betweenness centrality as $b_i(\mathcal{H}) = b_i(\pi_2(\mathcal{H}))$ for every node i in the hypergraph.

3. Results

Let us compare the new weighted metric structure with the classic approach based on the clique projection with reference to a series of illustrative synthetic hypergraphs and with large-sized real-world higher-order networks.

3.1. Synthetic Examples

In the first example, we consider the hypernetwork $\mathcal{H}_1$ described in Figure 6. The hypergraph contains k nodes organized in the form of a ring, with an additional hyperedge grouping all nodes. Let us assume that hyperedges may be eliminated with a probability proportional to their size, an assumption which is natural in social networks since it is hard to maintain communication within huge teams. Considering this process, in our example, one has, therefore, a very high probability to obtain graph $\mathcal{H}_2$ (the right panel of Figure 6), which consists of just a ring structure without a central hyperedge.
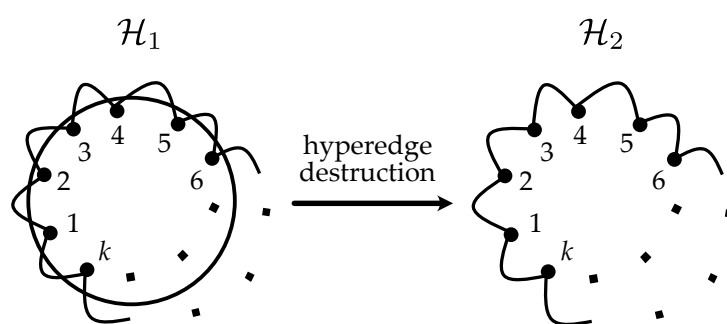
Figure 6. An illustrative example of a ring-like hypernetwork with an additional hyperedge grouping all nodes. If the central hyperedge is removed, the hypernetwork is transformed into a ring structure.

Table 1 shows the efficiency values calculated by using the proposed weighted hypergraph distance ($E^w(\cdot)$) and the traditional one, based on the clique projection ($E(\cdot)$). One immediately sees that the traditional measure overestimates the hypergraph robustness, whereas the fact that the central hyper-link is not reliable is correctly reflected in the case of our distance measure. This conclusion follows immediately from a comparison of the $E(\mathcal{H}_1)$ and $E(\mathcal{H}_2)$: in the case of the new distance definition, $\mathcal{H}_1$ and $\mathcal{H}_2$ have approximately the same efficiency, while in the case of the traditional measure, efficiency drops significantly when we destroy the central hyperedge.

Table 1. Efficiency values of $\mathcal{H}_1$ and $\mathcal{H}_2$ calculated using different distance definitions. $k = 20$.

	$E^w(\cdot)$	$E(\cdot)$
$\mathcal{H}_1$	0.311	1
$\mathcal{H}_2$	0.303	0.303

In the second example, we center our attention on the hypergraph of Figure 7. There, the assumption is made that one needs to transfer information from a source node i to a target node j, and in each intermediate node, information is copied and errors may appear. Therefore, the smaller the number of intermediate nodes, the less contaminated by errors the information will arrive at the target node. In addition, one assumes that the nodes of the hypernetwork can be attacked, and therefore some paths can be eliminated, which makes it such that the higher the intersection size is, the more reliable the communication.

Let us now consider two existing paths from node i to node j: (e_1, e_2) and (e_3, e_4). It is straightforward to check that these two paths are equivalent in the classic metric structure obtained from the clique projection. However, the first path is shorter than the second by using the weighted distance since the size of the edges union is the same for the both paths, while the intersection size is higher in the first case.
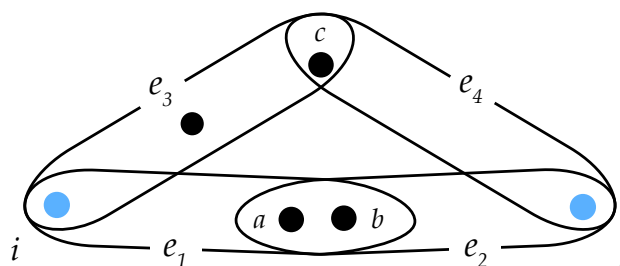
Figure 7. Illustrative example of a hypergraph, where information transfer occurs from a source node i to a target node j. The hypothesis is made that in each intermediate node, information is copied and errors may appear. In the example, there are two possible paths from node i to node j: (e_1, e_2) and (e_3, e_4). a, b, c are nodes forming the intersection between different hyperedges.

3.2. Real World Examples

We concentrate on the analysis of three real-world hypergraphs taken from https://www.cs.cornell.edu/~arb/data/ (accessed on 11 June 2023). The first is the Contact High School hypergraph [28,29], where nodes are students and hyperedges represent communications between them. The second hypernetwork is the Email Enron dataset based on lists of e-mails' senders and recipients (only hyperedges with a size up to 25 are considered, following Ref. [28]). The third hypergraph is reconstructed from Senate Committees data [10,30], where nodes are members of the US Senate, and hyperedges correspond to committee memberships.

The general characteristics and parameters of the three considered datasets are reported in Table 2, where it is shown that such hypergraphs have a comparable number of nodes, but they are significantly different in the properties of their hyperedge size distributions. In particular, the Contact High School hypergraph is characterized by the smallest size of the largest hyperedge. Furthermore, despite the fact that in the Email Enron data, one has significantly larger hyperlinks than in the Contact High School one, the median hyperedge size in both hypergraphs is equal to two, and the mean hyperedge size in Email Enron is also not far from the one characterizing the Contact High School case. The Senate Committees hypergraph differs from the other two in all the hyperedges distribution characteristics, and for this latter hypergraph, one has a relatively small number of hyperedges, which are, however, large.

Table 2. General characteristics of the real-world hypernetworks analyzed in our study.

	High School	Email Enron	Senate Committees
Number of nodes	327	143	282
Number of unique hyperedges	7818	1457	315
Maximal hyperedge size	5	18	31
Minimal hyperedge size	2	2	4
Mean hyperedge size	2.3	3.1	17.2
Median hyperedge size	2	2	19

In order to point out the additional information provided by the proposed weighted distance measure, one can calculate (and compare) the distances between all pairs of nodes in the considered hypergraphs by using our measure and the clique projection approach. The distributions of the differences between the results obtained with our measure and those obtained with the clique projection are presented in Figure 8. From these diagrams, one clearly sees that in the Contact High School data, the difference is not significant. Therefore, one can conclude that in this hypergraph, there is a large number of relatively short paths going only through edges of size two, alternative to the paths including edges of larger sizes. This conclusion is, however, incorrect for the other two hypernetworks,

especially for the Senate Committees one. One sees, indeed, that the distance in the clique projections of the hypergraphs is significantly smaller then the weighted one. Therefore, one expects a higher vulnerability of these latter graphs with respect to large edge removal.

If one wants to give a more quantitative ground to the last remark, one can calculate for each of the hypergraphs two efficiency measures: the one using the proposed weighted distance ($E^w(\cdot)$) and the other based on traditional distance measure ($E(\cdot)$). The results are shown in Table 3, where it is shown that the estimates have close values for the Contact High School hypergraph, whereas the values are significantly different for Email Enron and Senate Committees. If one compares the hypernetwork efficiencies using the traditional measure, one should conclude that the most efficient hypergraph is the Senate Committees one. On the contrary, by using the proposed weighted distance, one actually is led to the opposite conclusion. Notice that if our hypothesis that in social networks larger groups are less stable is correct, then only the $E^w(\cdot)$ measure provides a correct comparison of the hypergraph efficiencies.
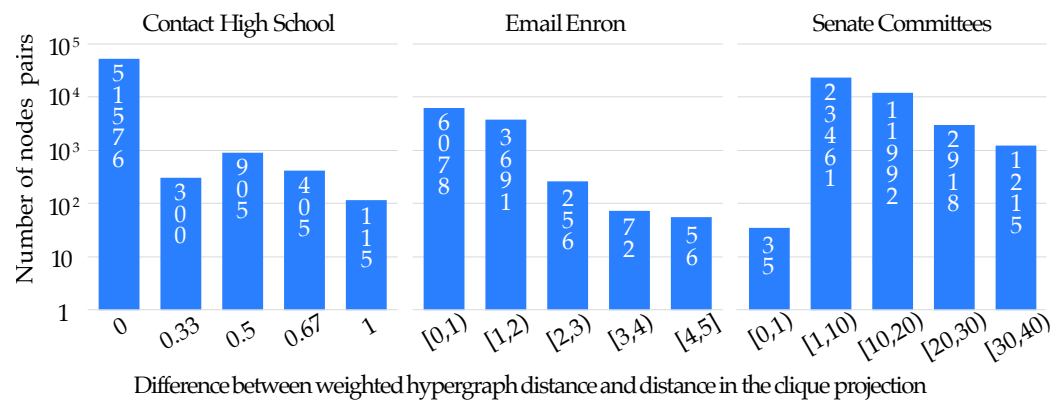

Figure 8. Distributions of differences between the proposed weighted hypergraph distance measure and the one calculated in the clique projection approach, for the real world hypergraphs analyzed in our study. In the first histogram, the number of pair of nodes is reported for which the difference between the two calculated distances takes the values specified in the horizontal axis. In the second and third histograms, we report instead the number of pairs of nodes, for which the difference between the two distances lies within the intervals indicated in the horizontal axis.

Table 3. Efficiency of the real-world hypergraphs computed using the proposed weighted hypergraph distance measure ($E^w(\cdot)$) and the traditional distance measure ($E(\cdot)$).

	Contact High School	**Email Enron**	**Senate Committees**
$E^w(\cdot)$	0.505	0.443	0.106
$E(\cdot)$	0.510	0.546	0.670

Besides efficiency, one can analyze node rankings based on the closeness and betweenness centralities by using either the traditional or the weighted hypergraph distances, as described in Section 2.2. In order to compare the rankings, one can use the Kendall rank correlation (KRC) and a measure μ that quantifies correlation given by

$$\mu_N(\mathbf{x}, \mathbf{y}) = \frac{|\text{top}_N(\mathbf{x}) \cap \text{top}_N(\mathbf{y})|}{N}, \tag{10}$$

where N is the number of top nodes considered, $\mathbf{x}$ and $\mathbf{y}$ are the two rankings, and $\text{top}_N(\mathbf{x})$ is the set of the N top nodes in the ranking $\mathbf{x}$. In particular, we compare rankings $\mathbf{b} = (b_1(\mathcal{H}), \dots, b_N(\mathcal{H}))'$ and $\mathbf{b}^w = (b_1^w(\mathcal{H}), \dots, b_N^w(\mathcal{H}))'$ obtained with the betweenness centrality, and rankings $\mathbf{C} = (C_1(\mathcal{H}), \dots, C_N(\mathcal{H}))'$ and $\mathbf{C}^w = (C_1^w(\mathcal{H}), \dots, C_N^w(\mathcal{H}))'$ obtained with the closeness centrality for the three real hypergraphs.

The results are presented in Figure 9, and the following conclusions can be drawn:

1. In the Contact High School hypergraph, the difference in rankings is significantly less considerable than that occurring in the other two hypergraphs, especially for the case of closeness centrality;
2. The sets of the first 50 top nodes for the cases of the Email Enron and Senate Committees hypergraphs are significantly different when one uses the traditional distance measure and our measure for both betweenness and closeness centralities (see Figure 9b,d);
3. The KRC coefficient is low in the case of betweenness centralities, even for the case of the Contact High School hypergraph (see Figure 9a);
4. The KRC coefficient for the case of closeness centrality is low for the Email Enron and Senate Committees hypergraphs (see Figure 9c), being even negative for the case of the Email Enron graph.

Summarizing the above evidence, one can say that the proposed distance measure provides significantly different assessments on the characteristics (and roles) of the nodes from the information-transferability point of view. The difference is brighter for the hypergraphs which are principally distinguished from graphs, i.e., in which hyperedges of large size are frequent and nodes in these hyperedges are rarely connected by other hyperedges of smaller sizes.
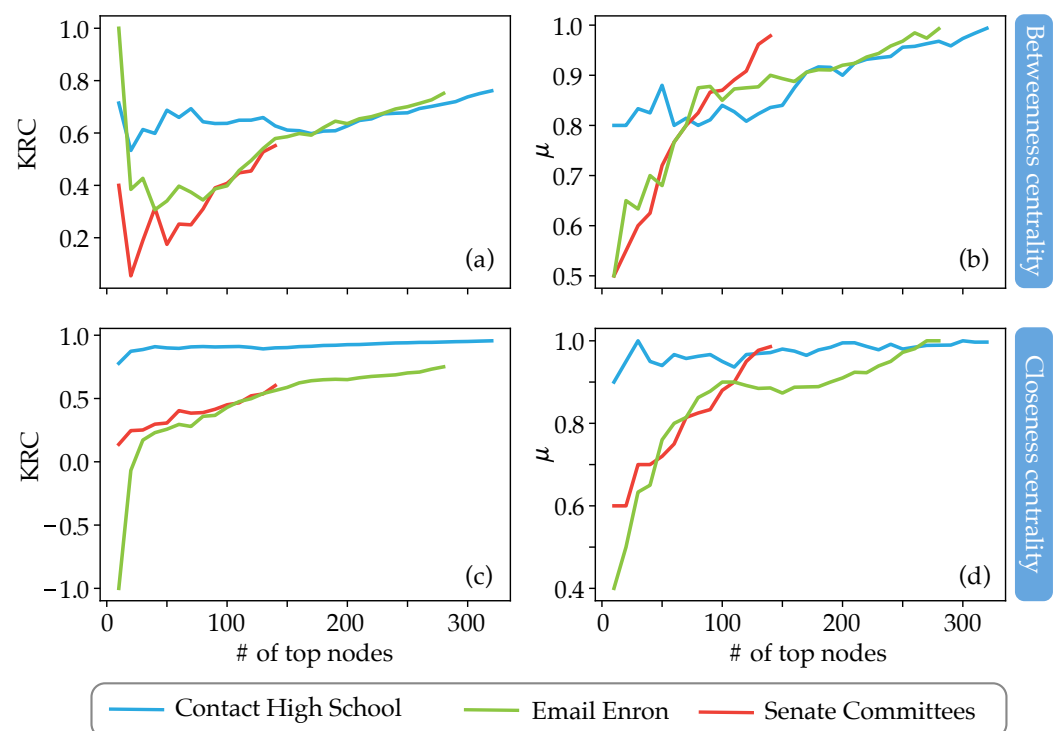

Figure 9. Correlations between the rankings based on betweenness (panels (**a**,**b**)) and closeness (panels (**c**,**d**)) centralities, calculated using the traditional distance measure and the weighted hypergraph one. (**a**,**c**) Kendall rank correlation (KRC) coefficients between the top nodes rankings. (**b**,**d**) The μ measure values (see Equation (10) for the definition of the μ measure). In all panels, the horizontal axis reports the number of top nodes considered in the rankings. The color code for identifying the curves plotted in all panels is reported in the horizontal bar at the bottom of the figure.

4. Conclusions

In conclusion, we here introduced a novel definition of distance for hypergraphs that extends the classic methods reported in the literature. In particular, the new distance takes explicitly into account two critical factors (the inter-node distance within each hyperedge and the distance between hyperedges in the network) which have never, so far, been considered together. The consequence is that the computation of distances is made in

a properly weighted linegraph of the original hypergraph. We illustrated the benefit of adopting such a metric structure with reference to a series of small synthetic hypernetworks, and we then applied our approach to large real-world hypergraphs, revealing that the obtained information about the hypergraphs structure is far different from that which is acquired by computing distances with the standard approach of clique projection. In particular, the difference is significant for hypergraphs in which hyperedges of large sizes are frequent and nodes are rarely connected by hyperedges of small sizes. This latter evidence points to the fact that our measure may be of great help in various circumstances, especially when a correct assessment of the roles of nodes is needed from the information-transferability point of view.

Possible follow-ups of our study include the use of our novel definition of distance for the measure of other topological properties of (or the analysis of processes taking places in) hypergraphs. This is the case, for instance, for random walks, local efficiency, clustering coefficients, community structure features, modularity, etc. Furthermore, a similar approach can be adopted also in the case of directed and annotated hypergraphs.

References

1. Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D.U. Complex networks: Structure and dynamics. *Phys. Rep.* **2006**, *424*, 175–308. [CrossRef]
2. Newman, M.E. The structure and function of complex networks. *SIAM Rev.* **2003**, *45*, 167–256. [CrossRef]
3. Estrada, E. *The Structure of Complex Networks: Theory and Applications*; Oxford University Press: Oxford, UK, 2012.
4. Battiston, F.; Cencetti, G.; Iacopini, I.; Latora, V.; Lucas, M.; Patania, A.; Young, J.G.; Petri, G. Networks beyond pairwise interactions: Structure and dynamics. *Phys. Rep.* **2020**, *874*, 1–92. [CrossRef]
5. Boccaletti, S.; De Lellis, P.; del Genio, C.I.; Alfaro-Bittner, K.; Criado, R.; Jalan, S.; Romance, M. The structure and dynamics of networks with higher order interactions. *Phys. Rep.* **2023**, *1018*, 1–64. [CrossRef]
6. Benson, A.R. Three hypergraph eigenvector centralities. *SIAM J. Math. Data Sci.* **2019**, *1*, 293–312. [CrossRef]
7. Carletti, T.; Battiston, F.; Cencetti, G.; Fanelli, D. Random walks on hypergraphs. *Phys. Rev. E* **2020**, *101*, 022308. [CrossRef] [PubMed]
8. Aksoy, S.G.; Joslyn, C.; Marrero, C.O.; Praggastis, B.; Purvine, E. Hypernetwork science via high-order hypergraph walks. *EPJ Data Sci.* **2020**, *9*, 16. [CrossRef]
9. Tudisco, F.; Higham, D.J. Node and edge nonlinear eigenvector centrality for hypergraphs. *Commun. Phys.* **2021**, *4*, 201. [CrossRef]
10. Chodrow, P.S.; Veldt, N.; Benson, A.R. Generative hypergraph clustering: From blockmodels to modularity *Sci. Adv.* **2021**, *7*, eabh1303. [CrossRef]
11. Kovalenko, K.; Romance, M.; Vasilyeva, E.; Aleja, D.; Criado, R.; Musatov, D.; Raigorodskii, A.M.; Flores, J.; Samoylenko, I.; Alfaro-Bittner, K.; et al. Vector centrality in hypergraphs. *Chaos Solitons Fractals* **2022**, *162*, 112397. [CrossRef]
12. Maletić, S.; Rajković, M.; Vasiljević, D. Simplicial Complexes of Networks and Their Statistical Properties. In Proceedings of the Computational Science—ICCS 2008, Kraków, Poland, 23–25 June 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 568–575.
13. Lu, L.; Peng, X. High-ordered random walks and generalized Laplacians on hypergraphs. In Proceedings of the International Workshop on Algorithms and Models for the Web-Graph, Atlanta, GA, USA, 27–29 May; Springer: Berlin/Heidelberg, Germany, 2011; pp. 14–25.

14. Carletti, T.; Fanelli, D.; Lambiotte, R. Random walks and community detection in hypergraphs. *J. Phys. Complex.* **2021**, *2*, 015011. [CrossRef]

15. Konstantinova, E.V.; Skorobogatov, V.A. Application of hypergraph theory in chemistry. *Discret. Math.* **2001**, *235*, 365–383. [CrossRef]

16. Estrada, E.; Rodríguez-Velázquez, J.A. Subgraph centrality and clustering in complex hyper-networks. *Phys. A Stat. Mech. Its Appl.* **2006**, *364*, 581–594. [CrossRef]

17. Zhou, D.; Huang, J.; Schölkopf, B. Learning with hypergraphs: Clustering, classification, and embedding. *Adv. Neural Inf. Process. Syst.* **2006**, *19*, 1601–1608.

18. Puzis, R.; Purohit, M.; Subrahmanian, V. Betweenness computation in the single graph representation of hypergraphs. *Soc. Netw.* **2013**, *35*, 561–572. [CrossRef]

19. Gao, J.; Zhao, Q.; Ren, W.; Swami, A.; Ramanathan, R.; Bar-Noy, A. Dynamic shortest path algorithms for hypergraphs. *IEEE/ACM Trans. Netw.* **2014**, *23*, 1805–1817. [CrossRef]

20. Shun, J. Practical parallel hypergraph algorithms. In Proceedings of the 25th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, San Diego, CA, USA, 22–26 February 2020; pp. 232–249.

21. Lee, J.; Lee, Y.; Oh, S.M.; Kahng, B. Betweenness centrality of teams in social networks. *Chaos Interdiscip. J. Nonlinear Sci.* **2021**, *31*, 061108. [CrossRef]

22. Behague, N.C.; Bonato, A.; Huggan, M.A.; Malik, R.; Marbach, T.G. The iterated local transitivity model for hypergraphs. *arXiv* **2021**, arXiv:2101.12560.

23. Jaccard, P. The distribution of the flora in the alpine zone. 1. *New Phytol.* **1912**, *11*, 37–50. [CrossRef]

24. Fredman, M.L.; Tarjan, R.E. Fibonacci Heaps and Their Uses in Improved Network Optimization Algorithms. *J. ACM* **1987**, *34*, 596–615. [CrossRef]

25. Cohen, R.; Havlin, S. *Complex Networks: Structure, Robustness and Function*; Cambridge University Press: Cambridge, UK, 2010.

26. Casablanca, R.M.; Criado, R.; Mesa, J.A.; Romance, M. A comprehensive approach for discrete resilience of complex networks. *Chaos Interdiscip. J. Nonlinear Sci.* **2023**, *33*, 013111. [CrossRef]

27. Latora, V.; Marchiori, M. Efficient Behavior of Small-World Networks. *Phys. Rev. Lett.* **2001**, *87*, 198701. [CrossRef] [PubMed]

28. Benson, A.R.; Abebe, R.; Schaub, M.T.; Jadbabaie, A.; Kleinberg, J. Simplicial closure and higher-order link prediction. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E11221–E11230. [CrossRef] [PubMed]

29. Mastrandrea, R.; Fournet, J.; Barrat, A. Contact Patterns in a High School: A Comparison between Data Collected Using Wearable Sensors, Contact Diaries and Friendship Surveys. *PLoS ONE* **2015**, *10*, e0136497. [CrossRef] [PubMed]

30. Stewart, C., III; Woon, J. Congressional Committee Assignments, 103rd to 114th Congresses, 1993–2017: Senate. Available online: http://web.mit.edu/17.251/www/data_page.html (accessed on 11 June 2023)