Random Walks and Node Centrality on Hypergraphs

Candidate Number: 1041671[a]

[a]University of Oxford, Mathematics Institute

This manuscript was compiled on April 1, 2020

The application of modern network science and graph theory has been largely successful in modelling the dynamics and structures of many interacting systems in the real world. Typically, the areas in which the networks have found particular success are when the interactions within a system are pairwise, for example in the modelling of citation networks. For more complex systems such as social networks however, the interactions are typically of a higher-order nature rather than pairwise interactions. In this report, we discuss how hypergraphs can instead be used to capture the higher-order interactions of complex systems and networks. We also discuss how centrality measures of graphs can be generalised to hypergraphs and how they can be extended to encapsulate the higher-order information of hypergraphs. In addition, we expand on previous work regarding random walkers on hypergraphs, to construct a novel class of walkers that can be used to measure the centrality and relative importance of nodes within a hypergraph.

Hypergraph | Random Walk | PageRank | Centrality Measures

In the general application of network science, one looks at modelling the interaction of individuals within a system as a graph, where an interaction between individuals (modelled as nodes) occurs via a pairwise connected edge. Such a model of interactions may be relevant for certain applications, for example in the modelling of the spread of sexually transmitted diseases (1) or computer networks (2). The applications of graphs to model system interactions has been studied extensively, utilising theory from many different fields, such as mathematics, statistical mechanics and information theory. Within the literature, there have been a vast variety of techniques proposed to observe the structure and the dynamics of graphs. Among these, centrality and node ranking measures have been prominent in identifying which nodes of a graph are particularly important. Examples of commonly used measures are eigenvector centrality (3), the clustering coefficient (4) and PageRank (5). Centrality or node ranking measures can provide colour to the structure of a particular graph, as well as having practical uses. For example, centrality measures can be used to identity which members of a population are more likely to spread a disease to many other individuals, and thus find the best candidates for vaccination or quarantining (6). In the case of more complex systems, such as social (7) and biological networks (8), the pairwise model of interactions is not so accurate. Instead, these networks exhibit higher-order interactions which typically cannot be encompassed in a standard graph. The idea of a hypergraph is to extend the typical pairwise interaction model from graph theory to encompass more complex dynamics, such as higher-order interactions. After giving a formal introduction to hypergraphs, we will discuss methods to measure node centrality of hypergraphs before proposing a novel alteration to methods within current literature.

1. Hypergraphs

A hypergraph is a generalisation of a normal graph (an introduction for which can be found at (9)), in which a single edge of the graph can join more than 2 nodes. We let $\mathcal{H}(V, E)$ be a hypergraph, where $V = \{1, 2, ..., n\}$ is the set of n nodes, and $E = \{e_1, e_2, ..., e_m\}$ is the set of m hyperedges. A hyperedge e is a non-empty subset of V such that $\cup_{e \in E}\, e = V$. If there are only two nodes within a particular hyperedge e, that is $e = \{i, j\}$ for i, $j \in V$ (and $i \neq j$), then we see that the hyperedge e is just a the standard edge of a graph. If we have a hypergraph such that for all $e \in E$, $|e| = 2$ then the hypergraph $\mathcal{H}$ completely reduces to a graph. This is a useful property, as we will be able to see how certain generalised hypergraph node measures compare to their graph counterparts.

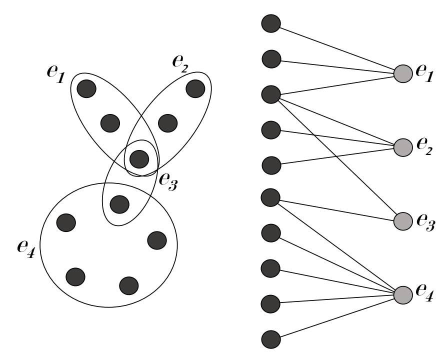


Fig. 1. Example hypergraph and its equivalent bipartite projection. The circles of left image denote hyperedge membership.

An equivalent way to display the information given by a hypergraph is through the use of a bipartite network, where one side of the network denotes the nodes of the hypergraph and

Significance Statement

Networks and graphs have been used to model a wide variety of real-world systems, namely in areas such as finance, epidemiology and transportation. However, in some real-world systems we see higher-order interactions, which cannot be accurately described by graphs. We instead model such systems with hypergraphs. In this report, we will discuss how one can use centrality measures to find the relative importance of nodes within a hypergraph. In addition, we also propose a new class of centrality measures that allows the user to observe certain structural biases within a hypergraph.

the other corresponds its the hyperedges. In this setting, an edge between the two sets of the bipartite network denotes membership of that particular node to the paired hyperedge. An example of projecting a hypergraph to a bipartite network is given in Fig. 1. To explicitly encode the membership of the nodes within a hypergraph to their respective hyperedges, we can define the *(hyper)incidence* matrix $\mathbf{I} \in \mathbb{R}^{n\times m}$, where

$$I_{i,\alpha} = \begin{cases} 1, & \text{if } i \in e_\alpha \\ 0, & \text{otherwise.} \end{cases}$$

This matrix is equivalent to the biadjacency matrix of the hypergraph's bipartite projection (9). Such a matrix also exists for standard graphs, except that the columns in the matrix would only be able to contain two non-zero elements. In addition to the hyperincidence matrix, one can also define the *hyperadjacency* matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, which is given by

$$\mathbf{A} = \mathbf{I}\,\mathbf{I}^T. \qquad [1]$$

The entry $A_{i,j}$ of the hyperadjacency matrix represents the number of hyperedges which contain both nodes i and j. Lastly, we can define the *hyperedges* matrix $\mathbf{C} \in \mathbb{R}^{m\times m}$

$$\mathbf{C} = \mathbf{I}^T\,\mathbf{I}, \qquad [2]$$

where the elements $C_{\alpha,\beta}$ denote the number of common nodes between hyperedges e_α and e_β, also given by $|e_\alpha \cap e_\beta|$. For each node i within a hypergraph, we can also specify its *node degree* k_i, which is the total number of hyperedges i belongs to (found by summing over the i^{th} row of the incidence matrix $\mathbf{I}$). Likewise, for each edge e_α, we can specify its *edge degree* $\delta(e_\alpha)$, the total number of nodes within hyperedge e (which is also given by $C_{\alpha,\alpha}$).

2. Clique Reduction

Clique reduction (10) is a technique that can be used to reduce a hypergraph to graph $\mathcal{G}(V, E_\mathcal{G})$ using the hyperadjacency matrix $\mathbf{A}$. It does so by generating a graph with an adjacency matrix $\mathbf{A}_\mathcal{G}$ equal to the hypergraph's hyperadjacency matrix $\mathbf{A}$. A graph formed via such a reduction replaces each hyperedge e with a clique of degree equal to $\delta(e)$; that is every node within a hyperedge is now connected to every other node via a standard edge. For example, a node $i \in V$ belonging to k_i hyperedges will be reduced to a node with

$$\sum_{e_\alpha \in E} (\delta(e_\alpha) - 1) I_{i,\alpha} \quad \text{standard edges.}$$

An example of clique reduction applied to the hypergraph within Fig. 1 is given in Fig. 7 of the Supplementary Material.

3. Clique Motif Eigenvector Centrality (CEC)

The CEC centrality measure is a basic measure of centrality that has been extended to hypergraphs from the original eigenvector centrality measure for graphs proposed by Bonacich in 1972 (11). As revealed by its name, to calculate the CEC centrality scores for a hypergraph $\mathcal{H}$, one first reduces it to a graph $\mathcal{G}$ via a clique reduction. The graph $\mathcal{G}$, which has an adjacency matrix given by the hyperadjacency matrix $\mathbf{A}$, can now be analysed using standard techniques from network science. By the standard definition of eigenvector centrality for a graph (see, for example (9)), we have that the centrality scores $\mathbf{c}$ are given by the eigenvector

$$\mathbf{A}\mathbf{c} = \lambda_1 \mathbf{c} \qquad [3]$$

where λ_1 is the largest real eigenvalue of $\mathbf{A}$. In this report, the eigenvector $\mathbf{c}$ is normalised so that $||\mathbf{c}||_1 = 1$; doing so allows us to compare the node ranking with measures calculated in latter sections.

A. Intuition behind CEC. There are several ways in which the eigenvalue centrality for graphs, and therefore the CEC method, can be interpreted as plausible measures for node ranking (described in (12)). One possible interpretation is via a low rank approximation. Since the matrix $\mathbf{A}$ is symmetric (as the number of common hyperedges between nodes i and j, is the same as j and i), and contains only real non-negative entries, then by the Perron-Frobenius Theorem, we have that the largest eigenvalue of $\mathbf{A}$, namely λ_1, is real and unique. The corresponding eigenvector to the largest eigenvalue can also be chosen to have non-negative elements. In addition, the eigenvector $\mathbf{c}$ is the principle left and right vector for the singular value decmposition of $\mathbf{A}$ (12). Therefore, by the Eckart-Young-Mirsky Theorem (also known as the Matrix Approximation Lemma), we have that $\mathbf{c}$ is the best rank-one approximation for the matrix $\mathbf{A}$ under the Frobenius matrix norm, that is

$$\mathbf{c} \propto \underset{\mathbf{x}\in\mathbb{R}^n}{\arg\min}\, ||\mathbf{A} - \mathbf{x}\mathbf{x}^T||_F. \qquad [4]$$

Thus, the centrality scores $\mathbf{c}$ given by CEC try to describe the structure of a hypergraph's clique reduced adjacency matrix, via a rank-one approximation. In Fig.2, we display the hyperadjacency matrix (or the adjacency matrix of the clique-reduced graph) for the example hypergraph in Fig.1 and the rank one approximation to the graph, given by the CEC scores.

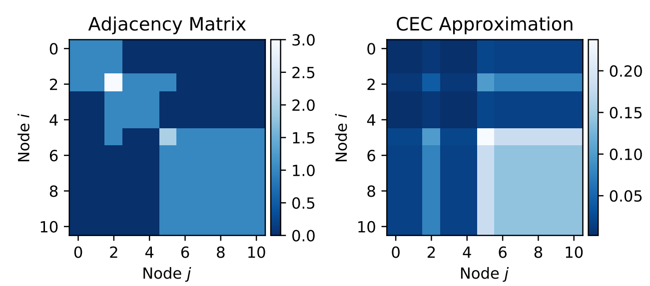


Fig. 2. (Left): Hyperadjacency matrix of the example hypergraph in Fig. 1. (Right): Rank one approximation of the hyperadjacency matrix, using the CEC centrality measure.

Despite its intuitive construction, the CEC measure does have some pitfalls, which we will briefly discuss. One common weakness is that the measure has no inherent scale, meaning that we have to impose some normalisation on $\mathbf{c}$ ourselves, so that we can compare rankings. In certain cases, such as the application of hypergraphs to machine-learning, the choice of normalisation can play an influence on the results; this is discussed further in (12). In addition, the CEC measure depends on the clique reduction preserving all the information of a hypergraph, which is often not the case (as discussed in (13)).

4. Random Walks on Hypergraphs

In addition to the centrality measure CEC, we will also discuss how centrality measures can be constructed from the behaviour of random walks on hypergraphs. First however, we have to introduce the idea of a random walk on a hypergraph, and how they are formulated. Random walks on graphs are well-studied, for example see (14), and have been used extensively to explore the dynamics and structure of many systems. In this section, we assume the reader is familiar with random-walks, and specifically random walks on graphs. After giving a simplistic introduction to random walks on hypergraphs, we propose a novel type of random walker for hypergraphs that is more generalisable.

A. Uniform Walk. To introduce the idea of random walks on hypergraphs, we discuss the most basic random walk which was first introduced by Zhou (15). Such a walker can be simply defined based on how it moves from one node to another. Suppose that we start a walker at node $i \in V$, then the walker chooses its next step according to the following rules:

- Pick a hyperedge e to which i belongs, at random.
- Randomly select from the hyperedge e, a node $j \in e$ such that $j \neq i$.
- Move to node j.

We introduce the name *uniform* (U) for a walk that abides by the above rules, so that when the we come to compare different walks in the latter sections of this report, it is explicitly clear to which type of walker we are referring to. We can mathematically formulate the walk by considering the transition probability $T^{(U)}_{i,j}$, that is the probability of the walk jumping from node i to node j. This is given by

$$T^{(U)}_{i,j} = \sum_{e_\alpha \in E} \frac{I_{i,\alpha}}{k_i} \frac{I_{j,\alpha}}{\delta(e_\alpha) - 1} \text{ for } i \neq j. \quad [5]$$

The first term in the summation gives the probability of selecting hyperedge e_α uniformly from those hyperedges containing node i, whilst the second term denotes the conditional probability of selecting node j, given you are choosing from hyperedge e. One might be wondering why we are only choosing from the nodes in an edge that are different from the current node. To explain this dynamic, we first recall how a random walker moves along a traditional graph. Once a walker has chosen which edge it will traverse, it jumps to the node at the end of the corresponding edge. Suppose that the walker on our hypergraph has selected a hyperedge e, where $|e| = 2$; for the dynamics of the walker on the hyperedge e to be equivalent to the the dynamics on a standard edge, we must have that the current node is removed from consideration.

An alternative way of writing Equation 5, is in terms of the hyperincidence matrix $\mathbf{I}$. By letting $\mathbf{D}_v = \text{diag}(k_1, k_2, ..., k_n)$ be a diagonal matrix of node degrees, we have that

$$\mathbf{T}^{(U)} = \mathbf{D}_v^{-1}\mathbf{I}(\mathbf{D}_e - I_m)^{-1}\mathbf{I}^T \quad [6]$$

where $\mathbf{D}_e = \text{diag}(\mathbf{C})$ and I_m is the $m \times m$ identity matrix. By recalling that the walker can't choose to stay at the current node (and assuming that there are no self-loop hyperedges), we set the the main diagonal of the matrix $\mathbf{T}^{(U)}$ equal to zero. The fact that the choice of each hyperedge is uniform means that the walker does not use any information when it chooses to explore the hypergraph. More complicated walking dynamics have recently been proposed so that the walker traverses the structure in a more informed manner.

B. Structured Walk. Recently, Carletti et al. (13) proposed a new class of random walkers for exploring hypergraphs in a way that utilises higher-order information of the nodes. It does so by allowing the walkers to choose the hyperedge non-uniformly, based on information about that hyperedge. In their paper, they propose a walk which targets nodes that have more in common with the current node. More formally, the transition probability of a jump from node i to j is proportional to the degree of the hyperedges (excluding node i) which both i and j belong. For example, if nodes i an j are both contained in many large hyperedges whilst i and k have only one small hyperedge in common, then the transition probability $T^{(S)}_{i,j}$ will be significantly higher than $T^{(S)}_{i,k}$. By considering the hyperedges which nodes i and j belong, the transition probability $\mathbf{T}^{(S)}$ can be defined as

$$T^{(S)}_{i,j} = \frac{\sum_{e_\alpha \in E} (\delta(e_\alpha) - 1) I_{i,\alpha} I_{j,\alpha}}{\sum_{l \in V} \sum_{e_\alpha \in E} (\delta(e_\alpha) - 1) I_{i,\alpha} I_{l,\alpha}} \text{ for } i \neq j. \quad [7]$$

As before, we consider simple random walks only, meaning that $T^{(S)}_{i,i} = 0$ for all $i \in V$. We call a walk with such a transition probability a *structured* walk (hence the label (S) on the transition matrix), as it uses structural information regarding hyperedges when computing the probability of a jump.

C. Adaptive Walk. In this section, we discuss a novel class of random walkers on hypergraphs that combine the theory and nuances of the two previous types of walkers. First however, we re-consider the transition probabilities for the uniform and structured walkers, given by Equations 5 and 7 respectively. By re-arranging the transition probability of a structured walk and letting

$$k^{(S)}_i = \sum_{l \in V} \sum_{e_\alpha \in E} (\delta(e_\alpha) - 1) I_{i,\alpha} I_{l,\alpha},$$

we have that

$$T^{(S)}_{i,j} = \sum_{e_\alpha \in E} \frac{I_{i,\alpha}}{k^{(S)}_i} \frac{I_{j,\alpha}}{(\delta(e_\alpha) - 1)^{-1}}. \quad [8]$$

This takes a similar form to Equation 5 for the uniform walker, except the $\delta(e) - 1$ term is now reciprocated to favour the nodes in larger hyperedges. This motivates the idea of a general, adaptive, non-linear transition probability, defined by

$$T^{(A)}_{i,j} = \sum_{e \in E} \frac{I_{i,\alpha} I_{j,\alpha}}{k^{(A)}_i} f(\delta(e) - 1), \quad [9]$$

where

$$k^{(A)}_i = \sum_{l \in V} \sum_{e_\alpha \in E} f(\delta(e_\alpha) - 1) I_{i,\alpha} I_{l,\alpha} \quad [10]$$

for some function f. In the case of the structured walker, the function $f(x) = x$ is used, whilst for the uniform walk, we have $f(x) = {}^1/_x$. In this report we propose the use of the function $f(x; \gamma) = x^\gamma$, where γ is a parameter of the walk that is used to control the bias of a walker. By substituting f, we obtain a walker with a transition probability $T_{i,j}^{(A)}$ is given by

$$T_{i,j}^{(A)} = \sum_{e \in E} \frac{I_{i,\alpha} I_{j,\alpha}}{k_i^{(A)}} (\delta(e) - 1)^{\gamma}. \qquad [11]$$

If $\gamma > 0$, then a node belonging to a large hyperedge has a higher transition probability then that of a node belonging to a small hyperedge. For $\gamma > 0$, the walker is therefore biased towards nodes that belong to larger hyperedges. Consequently, for $\gamma < 0$, we have a walker which is biased towards nodes belonging to smaller hyperedges. This means that, relatively speaking, a neighbouring node connected via a hyperedge with cardinality 2, will have a higher transition probability than that of a node connected via a hyperedge which contains 5 other nodes. This type of bias random walk, which we label an *adaptive* (A) walk, is more general than both the uniform and structured walkers suggested previously, as they can easily be generated with walk bias parameters $\gamma = -1$ and 1 respectively. We label this walk adaptive, due to the fact that the walker can be biased to explore different structures of a hypergraph. We can also write Equation 11 in matrix form, which is convenient for when we come to implement the walkers numerically, as we can utilise fast matrix multiplication compilers to compute the transition matrices. We first define a diagonal matrix $\bar{\mathbf{C}}_{\gamma} \in \mathbb{R}^{m \times m}$, given by

$$\bar{\mathbf{C}}_{\gamma} = \begin{bmatrix} (\delta(e_1) - 1)^{\gamma} & & \\ & \ddots & \\ & & (\delta(e_m) - 1)^{\gamma} \end{bmatrix}, \qquad [12]$$

where the edge degrees $\delta(e_{\alpha})$ can be computed from the diagonal of the hyperedge matrix $\mathbf{C}$. Using $\bar{\mathbf{C}}_{\gamma}$, we can rewrite the transition probability in Equation 11 in the form

$$T_{i,j}^{(A)} = \frac{(\mathbf{I}\,\bar{\mathbf{C}}_{\gamma}\mathbf{I}^T)_{i,j}}{\sum_{l \in V} (\mathbf{I}\,\bar{\mathbf{C}}_{\gamma}\mathbf{I}^T)_{i,l}}. \qquad [13]$$

The transition matrix of a walker $\mathbf{T}$ tells us the probability of a jump from one node to another, however currently tells us nothing about the how the dynamics of the walker develop over multiple jumps, or even the how the walker behaves in infinite time. To consider the temporal dynamics of the walker, we first let $\mathbf{p}(t) = (p_1(t), ..., p_n(t))$, where $p_i(t)$ denotes the probability that the walker is at node i after $t > 0$ steps, given some initial condition $\mathbf{p}(0)$. By considering how a walker transitions between nodes, $\mathbf{p}(t)$ evolves according to

$$p_i(t+1) = \sum_{j \in V} p_j(t) T_{j,i} \qquad [14]$$

which can be re-written more succinctly in matrix form, as

$$\mathbf{p}(t+1) = \mathbf{p}(t)\mathbf{T}. \qquad [15]$$

We can consider the long term dynamics by letting $t \to \infty$. If we let $\mathbf{p}^{(\infty)}$ denote the asymptotic probability distribution of the walker, then it must satisfy

$$\mathbf{p}^{(\infty)} = \mathbf{p}^{(\infty)}\mathbf{T}. \qquad [16]$$

This asymptotic distribution of a walker is therefore given by the left eigenvector associated to the eigenvalue $\lambda = 1$ of $\mathbf{T}$. A stationary distribution of the adaptive walker defined by Equation 13, is given analytically by

$$p_i^{(\infty)} = \frac{\sum_{l \in V} (\mathbf{I}\,\bar{\mathbf{C}}_{\gamma}\,\mathbf{I}^T)_{i,l}}{\sum_{m \in V} \sum_{l \in V} (\mathbf{I}\,\bar{\mathbf{C}}_{\gamma}\,\mathbf{I}^T)_{m,l}}. \qquad [17]$$

The derivation of this distribution is given in Section B of the Supplementary Information. Note that the eigenvalue $\lambda = 1$ may have more than one corresponding eigenvector, which is certainly the case where the hypergraph is not connected (as discussed in (14)). If a transition matrix has a unique stationary distribution, then its elements give the proportion of time a walker spends at each node, during an infinitely long random walk. Evidently, this could be used as a way to rank the importance of nodes, the more time a walker spends at a particular node, the more important it is to dynamics and of structure of the hypergraph. However, since a unique stationary distribution is not guaranteed for arbitrary hypergraphs, we consider implementing techniques from the PageRank measure of a graph (5).

5. PageRank Measure, $\mathrm{PR}_{(\gamma)}$

PageRank (5), is a type of node ranking measure for arbitrary networks that considers the asymptotic behaviour of a random walker. To deal with the fact that there are multiple possible stationary distributions when a graph has many absorbing states, PageRank introduces the idea of *teleportation* on graphs. The idea behind teleportation is to ensure there is a non-zero probability that a walker can travel between any nodes of the graph, even if they are not connected by an edge. To do so, PageRank assigns with probability $1 - \alpha_{\mathrm{PR}}$, that the walker will randomly select a node in the graph and teleport there. Otherwise, with probability α_{PR}, the walker's dynamics are as discussed previously, given by the transition matrix and Equation 15. Mathematically, the altered random walk can be formulated as

$$\mathbf{p}(t+1) = \alpha_{\mathrm{PR}} \underbrace{\mathbf{p}(t)\mathbf{T}}_{\text{Normal Step}} + (1 - \alpha_{\mathrm{PR}}) \underbrace{\frac{\mathbb{1}}{n}}_{\text{Teleportation Step}} \qquad [18]$$

where $\mathbb{1}$ denotes an n-dimensional vector of ones. Since there is a non-zero probability of any node jumping to another, the walker is now traversing a connected version of the original graph, and by implementing teleportation, graphs that once featured separated components are now connected. The left eigenvector corresponding to the eigenvalue $\lambda = 1$ of such a walk is now unique, as the graph is connected. Since there is now no ambiguity regarding the asymptotic distribution $\mathbf{p}^{(\infty)}$, PageRank can be used as a measure of node centrality.

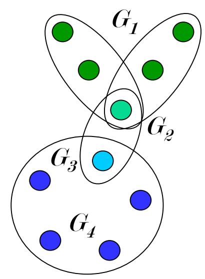
The idea of PageRank can be readily extended for random walks on hypergraphs by considering a process defined by Equation 18, where the transition matrix $\mathbf{T}$ is now the transition matrix of the walker on a hypergraph. As a result, for any hypergraph, we can obtain a unique stationary distribution which gives the relative amount of time a walker with transitions given by Equation 18 spends at each node. We let $\mathrm{PR}_{(\gamma)}$ denote the PageRank score of a walker with transition matrix $\mathbf{T}^{(A)}$ and walk bias γ.

6. Results

This section is split into two parts, first we provide more insight into how the dynamics of the novel adaptive walker depend of the bias parameter γ. Secondly, we analyse how the PageRank measure $\mathrm{PR}_{(\gamma)}$ compares with the standard CEC centrality measure given in Section 3 for large random hypergraphs.

A. Small Example. In this section we will explore the dynamics of the non-linear adaptive random walker defined by the transition matrix in Equation 13 on the hypergraph $\mathcal{H}$ displayed in Fig 1 of Section 1. This hypergraph has four types of nodes that are of interest, which we cluster into groups $G_i, i \in \{1, ..., 4\}$. These groups are displayed in Fig 3. We gather the nodes into groups based on their node degree and hyperedge membership, as nodes belonging to the same group have the same stationary distribution and therefore the same centrality measures. To confirm our apriori belief that the adaptive walker targets small or large hyperedges depending on the value of γ supplied, we compute the stationary distribution of $\mathbf{T}^{(A)}$ for different values of γ. One practical caveat which we ought to mention is that all of these methods, including the CEC measure of Section 3, deal solely with the unique dominant eigenvalue and associated eigenvector of certain matrices. Rather than calculating the entire eigenpair decomposition of a matrix, we can instead utilise the power method to efficiently obtain the eigenvector we are interested in; further details of the power method can be found at (16). By computing the unique stationary distribution of $\mathbf{T}^{(A)}$ (since the hypergraph is connected) for different values of γ, we obtain Fig 4 which shows how the centrality measures of the nodes in the various groups change based on γ. We see that when $\gamma \to \infty$, the centrality measure is dominated by the groups of nodes G_4 and G_3, which are the nodes belonging to the largest hyperedge e_4 (from Fig 1). The five nodes belonging to e_4 (nodes in groups G_3 and G_4) have centrality measure 0.2 each, while the other have centrality measures of 0. Likewise, when $\gamma \to -\infty$, we have that the only nodes of non-zero centrality are the nodes in groups G_2 and G_4, that is the nodes that belong to the smallest hyperedge e_3, of hyperedge degree 2. By considering the large limit behaviour of γ, it is evident that for $\gamma > 0$ we have a walker that is biased towards nodes of larger hyperedges, to the point in which, when γ is large enough, it will almost never leave the largest hyperedge. The same can be said for when $\gamma < 0$, the walker favours nodes of smaller hyperedges to the point in which, once entering the smallest hyperedge, it will not leave.

Fig. 3. Hypergraph of Fig. 1 displaying node groupings.



To fully observe the effect of the bias parameter γ, in Fig. 4, we considered the stationary distribution of the transition matrix given by $\mathbf{T}^{(A)}$ as it has a unique stationary distribution. In the generic application of random-walk centrality to hypergraphs however, connected-ness is not guaranteed and so one would instead consider the PageRank measure $\mathrm{PR}_{(\gamma)}$ of the hypergraph, given by the stationary distribution of the process in Equation 18. For completeness, we also observe how the PageRank node centrality $\mathrm{PR}_{(\gamma)}$ (with $\alpha = 0.85$ rather than $\alpha = 1$ as above) changes based on γ, this is displayed in

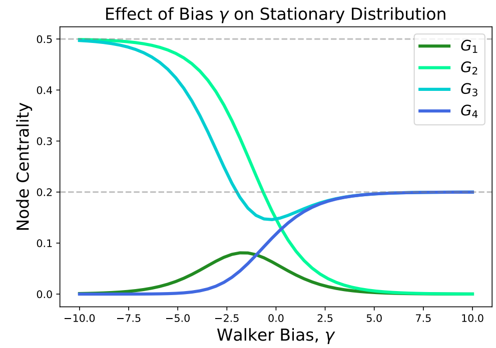


Fig. 4. Observing how the node centrality of the groups, given by Fig. 3 change based upon the bias of the random walker γ.

Fig. 8 of the Supplementary information, which the reader is encouraged to view. The dynamics of the PageRank for changing γ are very similar but slightly harder to interpret, hence the inclusion and discussion of Fig. 4 instead. Before moving onto discuss the application of these methods to more general graphs, we briefly discuss how the CEC centrality measure compares to $\mathrm{PR}_{(\gamma)}$, depending on the value of γ. By

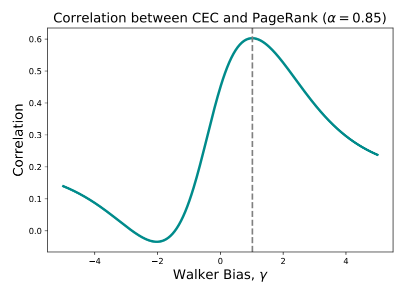


Fig. 5. Correlation between CEC centrality measure and the hypergraph PageRank measure $\mathrm{PR}_{(\gamma)}$ for different walk bias parameters, γ.

observing Fig. 5, we see that the PageRank centrality measure with $\alpha = 0.85$ is most correlated with CEC measure, when a bias parameter $\gamma = 1$ is used; this is the structured walker given previously, defined by Equation 7. In Fig. 9 of the Supplementary Information, we also display how this correlation changes based upon the the teleportation parameter α.

B. Large Graphs. In this section, we will apply and compare the presented theory to more general graphs, rather than a small example. In particular, we discuss the correlations of different centrality measures for random hypergraphs, generated by the Erdős-Rényi (E-R) model (17). To generate a random hypergraph, we first generate a random biparite graph, utilising the Python library NetworkX (18). As displayed in Fig. 1, any hypergraph can be projected to a biparite network and, provided one knows which side of the bipartite graph represents the hyperedges, one can map a biparite network back to a hypergraph. Therefore by constructing a random bipartite network, we can construct a random hypergraph with a pre-specified number of hyperedges and nodes. For more details on how we constructed random hypergraphs, the reader is directed to the Python notebook "`erdos_renyi_centrality_correlations.ipynb`" submitted with this report.

B.1. Erdos-Renyi Model. In this section, we observe how changing the size of the graph n affects the correlation between the centrality measures CEC and $\text{PR}_{(\gamma)}$. In our numerical examples, we generate a hypergraph of n nodes, with $\sqrt{n}$ hyperedges, and set the probability of a node belonging to any particular hyperedge as $p = 0.2$. Previously, we saw in our small example, given in Fig. 4, for different values of γ, the correlation with between the PageRank and the CEC scores varied greatly. Now, we wish to observe how the correlation between the different centrality measures change, as we increase the size of the graph. When considering the PageRank centrality measures, we solely use the teleportation parameter $\alpha = 0.85$, however we allow the walk bias parameter γ to vary. In Fig. 6, we display how the correlation between the scores CEC and $\text{PR}_{(\gamma)}$ change, based on the bias of the walker and the size of the graph it is exploring. We observe that as the

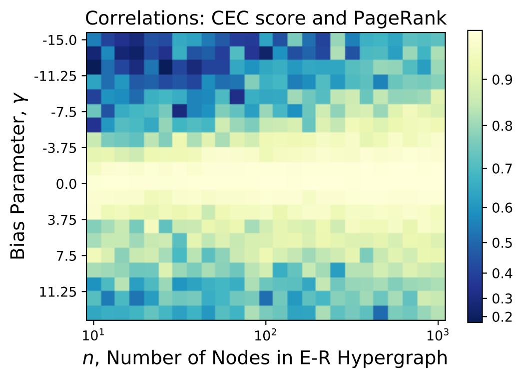


Fig. 6. Correlation between CEC node centrality and PageRank score for walkers of varying bias γ on hypergraphs produced using an Erdos-Renyi Model (p = 0.2). All PageRank scores used a teleportation parameter $\alpha = 0.85$.

size of the hypergraph increases, the correlation between the CEC score and PageRank for any bias increases. We also see that as we increase hypergraph size, we must consider more extreme values of γ to obtain centrality scores that differ greatly. To counteract this, we propose replacing γ with a function $\Gamma(\gamma, n) = \gamma \log(n)$. Doing so scales the bias of the walker so that more extreme values of γ are not needed as the size of the hypergraph increases. In Fig. 10 of the Supplementary Information, we display the equivalent of Fig. 6, but using a bias that is scaled with the hypergraph size.

7. Application and Discussion

From observing Fig. 6, we see that when the random hypergraphs get sufficiently large, the PageRank of the walkers will become very highly correlated with the CEC score. However if the bias γ is extreme enough or for small/medium sized graphs, there can be a significant difference between the measures. This difference between the two measures can capture valuable information about the nodes and the underlying structure of the hypergraph. Say for example that we wish to model how secrets or rumours spread within a population, then the PageRank measure $\text{PR}_{(\gamma)}$ with bias $\gamma < 0$ would be very suitable. Since one typically shares secrets with close friends rather than large groups, a secret would spread with some bias towards nodes of smaller frienship groups (hyperedges). A PageRank measure with $\gamma < 0$ would therefore capture these dynamics, unlike that of the CEC score. In addition to having to scale the bias parameter as discussed in Section 6B.1, another modelling issue the PageRank measure $\text{PR}_{(\gamma)}$ presents, is when it is applied to weighted hypergraphs. A weighted hypergraph is a hypergraph in which each hyperedge e has a corresponding weight w_e (13). A weighted hypergraph adds another level of complexity, and so in addition to our walker having a bias towards certain hyperedge structures (large or small hyperedges), it would also have to include a bias towards certain hyperedge weights. Random walks and and node ranking on weighted hypergraphs are considered by (13), but this issue is not discussed in the literature. For an application of the various centrality measures we have discussed in this report to a real hypergraph dataset, the reader is directed to the attached file "`real_data_example.ipynb`". For brevity, a discussion of those results are omitted.

In this report we have discussed two different methods of node centrality measures for hypergraphs, namely the CEC centrality measure and a PageRank-esque measure $\text{PR}_{(\gamma)}$ that utilises the theory surrounding random walks on hypergraphs. We proposed a novel class of walkers that can be used to construct centrality measures, and compare this novel class to the CEC measure on a simple example and to larger random hypergraphs. To extend the ideas in this report further, we would consider hypergraphs generated via algorithms that produce more structured models, rather than E-R model which is random in nature.

Materials and Methods

Please see the attached Python notebooks for numerical methods and figure replication. We note that in the practical implementation of the methods described in this paper, the majority of code has been self-developed, however we have utilised both the NetworkX (18) and the non-deprecated parts of the Halp (19) packages, where appropriate.

1. JJ Potterat, et al., Sexual network structure as an indicator of epidemic phase. *Sex. Transm. Infect.* **78**, 152–158 (2002).
2. J Webb, F Docemmilli, M Bonin, Graph theory applications in network security. *arXiv preprint arXiv:1511.04785* (2015).
3. P Bonacich, Power and centrality: A family of measures. *Am. journal sociology* **92**, 1170–1182 (1987).
4. K Börner, JT Maru, RL Goldstone, The simultaneous evolution of author and paper networks. *Proc. Natl. Acad. Sci.* **101**, 5266–5273 (2004).
5. L Page, S Brin, R Motwani, T Winograd, The pagerank citation ranking: Bringing order to the web., (Stanford InfoLab), Technical report (1999).
6. TW Valente, K Coronges, C Lakon, E Costenbader, How correlated are network centrality measures? *Connect. (Toronto, Ont.)* **28**, 16 (2008).
7. D Li, Z Xu, S Li, X Sun, Link prediction in social networks based on hypergraph in *Proceedings of the 22nd International Conference on World Wide Web.* pp. 41–42 (2013).
8. S Klamt, UU Haus, F Theis, Hypergraphs and cellular networks. *PLoS computational biology* **5** (2009).
9. M Newman, *Networks.* (Oxford university press), (2018).
10. J Rodriguez, Laplacian eigenvalues and partition problems in hypergraphs. *Appl. Math. Lett.* **22**, 916–921 (2009).
11. P Bonacich, Technique for analyzing overlapping memberships. *Sociol. methodology* **4**, 176–185 (1972).
12. AR Benson, Three hypergraph eigenvector centralities. *SIAM J. on Math. Data Sci.* **1**, 293–312 (2019).
13. T Carletti, F Battiston, G Cencetti, D Fanelli, Random walks on hypergraphs. *Phys. Rev. E* **101**, 022308 (2020).
14. F Göbel, A Jagers, Random walks on graphs. *Stoch. processes their applications* **2**, 311–336 (1974).
15. D Zhou, J Huang, B Schölkopf, Learning with hypergraphs: Clustering, classification, and embedding in *Advances in neural information processing systems.* pp. 1601–1608 (2007).
16. GH Golub, HA [van der Vorst], Eigenvalue computation in the 20th century. *J. Comput. Appl. Math.* **123**, 35 – 65 (2000) Numerical Analysis 2000. Vol. III: Linear Algebra.
17. JL Guillaume, M Latapy, Bipartite graphs as models of complex networks. *Phys. A: Stat. Mech. its Appl.* **371**, 795–813 (2006).
18. A Hagberg, et al., NetworkX, Python Package. *http://networkx. github.io/index.html* (2013).
19. B Avent, A Ritz, T Murali, J Cadena, Y Keneshloo, Halp: Hypergraph Algorithms Package, Python Package. *http://murali-group.github.io/halp/* (2014).

8. Supplementary Information

A. Clique Reduction Example. In Fig. 7, we display the hypergraph of Fig. 1 after being projected to a normal graph via a clique reduction. We see that every node in a hyperedge of degree k now forms a clique of size k.

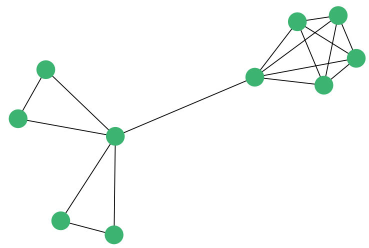


Fig. 7. Clique projected graph of the hypergraph displayed in Fig 1.

B. Stationary Distribution Proof. In this section, we verify that Equation 17 is indeed a stationary distribution of a random walker defined by the transition matrix $\mathbf{T}^{(A)}$, given by Equation 11. The stationary distribution is given by

$$p_i = \frac{\sum_{l \in V} (\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)_{i,l}}{\sum_{m \in V} \sum_{l \in V} (\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)_{m,l}}. \quad [19]$$

If p_i is the stationary distribution of the walker, then it satisfies

$$\sum_{j \in V} T_{j,i} p_j - p_i = 0.$$

By substituting Equations 17 and 11 in for p_i and $T_{j,i}$, we obtain

$$\left(\sum_{j \in V} \frac{(\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)_{j,i}}{\sum_{l \in V} (\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)_{j,l}} \sum_{l \in V} (\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)_{j,l} \right) - \sum_{l \in V} (\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)_{i,l} = 0.$$

Grouping the terms together, we have that

$$\sum_{j \in v} \left\{ \sum_{l \in V} (\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)_{j,l} \left[\frac{(\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)_{j,i}}{\sum_{l \in V} (\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)_{j,l}} - \delta_{i,j} \right] \right\} = 0$$

which simplifies to the following

$$\sum_{j \in V} (\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)_{j,i} - \sum_{l \in V} (\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)_{i,l} = 0$$
$$\implies \sum_{j \in V} (\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)_{j,i} - \sum_{j \in V} (\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)_{i,j} = 0.$$

Therefore Equation 17 gives the stationary distribution for the walker if the matrix $(\mathbf{I}\bar{\mathbf{C}}\mathbf{I}^T)$ is symmetric, which it is.

C. Effect of Bias on PageRank. We now display the figure that was briefly discussed in Section 6A of the main text, exploring how the bias parameter γ affects the PageRank measure of our sample hypergraph. We can observe that a similar pattern occurs between Fig. 6 and Fig. 8. In particular, we see that there is rank inversion taking place between $\gamma \in [-3, 3]$. The bias of the walker to nodes belonging to smaller/larger hyperedges evidently still present however the teleportation of PageRank means that the nodes are never completely irrelevant as was the case in Fig. 6.

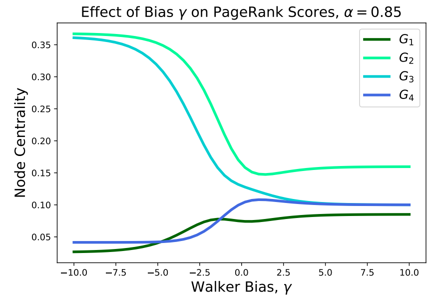


Fig. 8. We observe the effect the bias γ has on the $\mathrm{PR}_{(\gamma)}$ scores for the example hypergraph in Fig 1.

We also observe how the correlation between the centrality measures CEC and $\mathrm{PR}_{(\gamma)}$ change based on the teleportation parameter α of the PageRank walk. We see that as α approaches one, the walker bias has a more extreme effect on the correlation between measures.

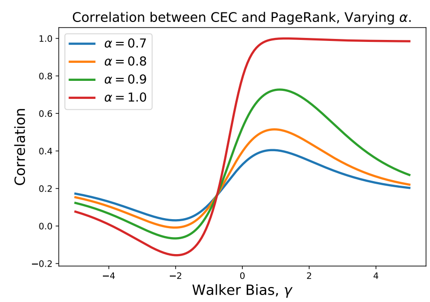


Fig. 9. We observe the effect that changing the walk bias γ and teleportation parameter α have on the correlation between the CEC centrality measure and $\mathrm{PG}_{(\gamma)}$ for the hypergraph in Fig. 1.

D. Scaled Bias, Large E-R Hypergraphs. In this section, we display the results of using a scaled bias

$$\Gamma(\gamma, n) = \gamma \log(n), \quad [20]$$

so that as the size of the graph increases, we do not need to consider more extreme values of γ. We replace γ of Equation 11 by Γ.

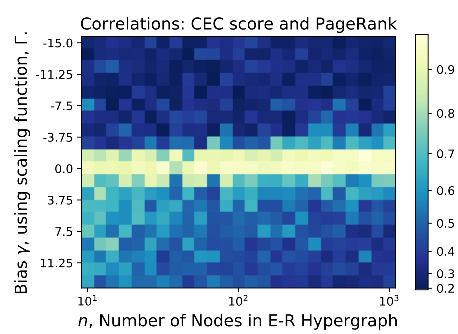


Fig. 10. Computing correlation between CEC and adaptive PageRank for randomly generated hypergraphs. The y-axis displays the value γ which is used in Equation 20. The bias is scaled based on the number of nodes in the hypergraph considered.

To further develop the theory presented here, one could consider applying the same centrality measures and techniques to hypergraphs constructed via preferential attachment. This is considered in the notebook "`preferential_attachment_model.ipynb`" of the attached material.