Beyond degree sequences: the descriptive power of walk sequences in graphs and biological networks

Tom Wyatt

May 2, 2012

Abstract

The analysis of topology in biological networks has been shown to reveal important biological information. Here, a novel tool for the analysis of network topology is proposed. We define the *walk matrix* as the matrix whose elements contain the number of closed walks which can be taken from different nodes in the network.

This matrix is shown to carry important topological information, above and beyond the information carried by the degree sequence.

We find that, for a general node, it is difficult to obtain information from the walk matrix which improves our ability to predict gene expression correlation, beyond simple predictions based on the existence or absence of a connection. However, in the case of network hubs, we show that the walk matrix can be used to predict correlation between the expression of a hub and its neighbours in the PPI network.

Contents

1 Introduction

Over the past decade, network biology has become an invaluable avenue of biological investigation [1]. The previous century was dominated by reductionist approaches, where individual biomolecules were investigated in minute detail, in isolation or as components of small systems [2]. It has become clear, however, that few biological processes can be entirely explained from a reductionist viewpoint due to the sheer complexity of biological organisms [3]. Network approaches address this problem by taking a systems level perspective.

A typical biological network represents biomolecules as *nodes* which may be connected to other nodes by *edges*, representing some form of pair-wise interaction or other relationship. Much biological detail is therefore removed, creating a simplified representation which allows for study on a broader scale. Crucially, biological networks are data driven models. They have been made possible by the emergence of high-throughput experimental techniques which obtain the data required to draw networks containing tens of thousands of nodes and edges. This use of data makes network approaches compare favourably with previous attempts to address complexity, such as chaos theory, whose models built from differential equations incorporated little data.

Here, a novel tool for the analysis of biological networks is introduced and explored. For any node in a network, a *closed walk* of length x is a sequence of x steps between connected nodes which starts and ends on the same node. Let $w_x(i)$ be the total number of possible closed walks of length x from node i. Now form the matrix $\mathbf{W}$ with elements $W_{ij} = w_{i+1}(j)$. The matrix $\mathbf{W}$ is called the *walk matrix* of the network and it is this matrix which is explored here.

The study of this matrix falls into two categories. The first is the mathematical characterisation of the matrix. An important question is precisely what information is contained in the walk matrix. Section 2 addresses this class of question. Firstly, relevant results from graph theory, the field of mathematics which underpins network theory, are reviewed. Then, in Section 2.2, results found here which fall into this category are summarised. This includes an informative decomposition of the walk matrix rows and an exploration of some networks which are uniquely defined by the matrix.

The second category is concerned with how the walk matrix can be used in biological networks. Section 3 addresses this topic. Firstly, a selection of results from network biology are reviewed. This review focusses specifically on what has be learnt from the topology of biological networks, since the walk matrix depends only on network topology. We then use data from a gene expression experiment along with a protein-protein interaction network and explore how the matrix can be applied usefully to this data. We find evidence that a closed walk based method can categorise network hubs into *party* and *date* hubs more effectively than previous topology based methods.

2 Mathematical Analysis

In mathematical parlance, the set of nodes, V, and edges, E, as described in the introduction, form a *graph*, $G(V, E)$. The edge connecting nodes i and j is denoted ij and nodes i and j are said to be *adjacent* if $ij \in E$. Graphs have been the subject of mathematical investigation since Leonhard Euler's 1736 paper on the seven bridges of Königsberg (see Figure 1) [4].

A representation of graphs which is frequently utilised in graph theory is the *adjacency matrix*,

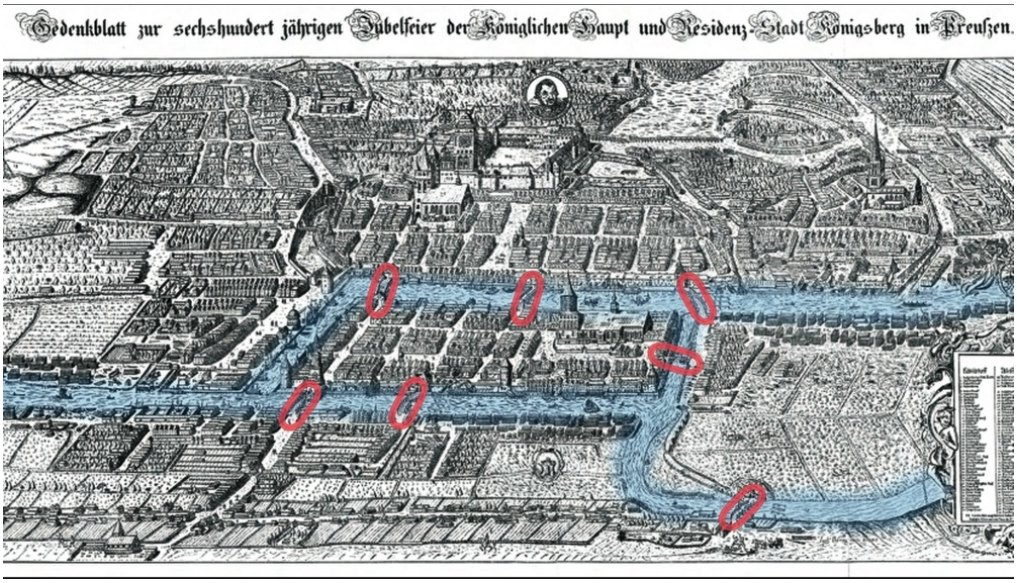

Figure 1: The seven bridges of Königsberg problem is the problem of finding a path which crosses each bridge in Königsberg (pictured here) once and once only. It can be framed as a graph theory problem where each separate land mass is represented by a node and each bridge is represented by an edge. Leonhard Euler proved in 1735 that it was not possible.

A. This is the matrix such that $A_{ij} = 1$ if nodes i and j are connected and $A_{ij} = 0$ otherwise. The adjacency matrix is of particular importance here since the walk matrix is simple to calculate via the relation [5]

$$w_x(i) = (\mathbf{A}^x)_{ii}. \tag{2.1}$$

A basic property of any node is its *degree*, which is the number of other nodes which the node connects to. The ordered list of the degrees of every node in a graph is called the *degree sequence*. It should be clear that the degree of a node is precisely the quantity $w_2(i)$ defined above and the degree sequence is then the first row of the walk matrix. The walk matrix, therefore, can be viewed as a generalisation of degree sequences. This makes the study of degree sequences highly relevant to this project and so the next section reviews some key results in this area.

2.1 Graphicality and unigraphs

An early study of degree sequences was from A. Cayley in 1874 [6]. Cayley was attempting to enumerate the distinct isomers of formula C_nH_{2n+2} and therefore wanted to enumerate the graphs with degree sequence $(4^x, 1^{2x+2})$ (where the index indicates the number of repetitions of that number). Although he did not quite succeed, this was the first serious application of graph theory to the sciences. The problem was not solved until 1927 when J. H. Redfield published the first version of the Pòlya enumeration theorem [7] which was later popularised by Pòlya [8].

Graphicality is a key topic of research in degree sequences. An ordered multiset of integers $\mathbf{w_2} = (w_2(1), w_2(2), \ldots, w_2(n))$ is called *graphic* if a graph can be found with degree sequence $\mathbf{w_2}$.

An important question, therefore, is as follows. Given a list $\mathbf{w_2}$, how can one test whether or not it is graphic? In 1960, Erdős and Gallai [9] proved a useful theorem addressing this question.

Theorem 2.1 *A given degree sequence* $\mathbf{w_2}$ *is graphic if and only if the sum of all degrees is even and the sequence obeys the property*

$$\sum_{i=1}^{r} w_2(i) \leq r\,(r-1) + \sum_{i=r+1}^{n} \min\left(r, w_2(i)\right) \tag{2.2}$$

for each integer $r \leq n-1$.

Further, it was shown in 2003 that this sequence need only be checked for as many r as there are distinct values of $w_2(i)$ [10].

Another useful characterisation of graphicality was found independently by both Havel in 1955 [11] and Hakimi in 1963 [12]. For any degree sequence $\mathbf{w_2}$, define the *reduced degree sequence*, $\bar{\mathbf{w}}_\mathbf{2}$, formed by deleting the node with largest degree and assuming it had been connected to the node set with the highest possible combined degree. The sequence $\bar{\mathbf{w}}_\mathbf{2}$ then is

$$\bar{\mathbf{w}}_\mathbf{2} : \bar{w}_2(2) \leq \bar{w}_2(3) \leq \ldots \leq \bar{w}_2(n), \tag{2.3}$$

which is the non-increasing rearrangement of the sequence

$$\left(w_2(2)-1, w_2(3)-1, \ldots, w_2(w_2(1)+1)-1, w_2(w_2(1)+2), w_2(w_2(1)+3), \ldots, w_2(n)\right). \tag{2.4}$$

Havel and Hakimi then both proved the following theorem.

Theorem 2.2 *The degree sequence* $\mathbf{w_2}$ *is graphic if and only if* $\bar{\mathbf{w}}_\mathbf{2}$ *is.*

This characterisation is convenient for use in recognition algorithms which iteratively reduce the sequence to simpler sequences.

A final important result on graphicality, by Behzad [13], is that no sequence where each value of degree occurs with multiplicity 1 can be graphical.

Another key problem regarding degree sequences is finding which graphs are defined exactly by their degree sequences. To formulate this question properly, first define an *isomorphism* between two graphs G and H as a bijection between the vertex sets of G and H

$$f : V(G) \to V(H) \tag{2.5}$$

such that any vertices i and j are adjacent in G if and only if $f(i)$ and $f(j)$ are adjacent in H. Additionally, an *automorphism* is an isomorphism of G onto itself. Clearly, if an isomorphism exists between two graphs, they are topologically identical and cannot be distinguished by their degree sequences. Returning to the question above, a graph is called a *unigraph* if and only if it is the only realisation of its degree sequence, up to isomorphism. The task then is to characterise exactly which degree sequences are *unigraphic*.

A complete description of the structure of unigraphs is given in the series of papers [14, 15, 16, 17, 18]. Among other things, the authors describe canonical decompositions of unigraphs, give a characterisation of the automorphism group of unigraphs and find algorithms for the generation and decomposition of the unigraphs. They also prove the inequality

$$(2.3)^{n-2} \leq u_n \leq (2.6)^n, \tag{2.6}$$

where u_n is the number of unigraphs with n vertices.

These papers, however, did not lead directly to an efficient unigraphic degree sequence recognition algorithm. This was first achieved by Li, who proved the following theorem [19].

Theorem 2.3 *The unigraphicality of a sequence of length n can be tested or reduced to the unigraphicality of a shorter sequence (or two sequences whose composite length is less than n). This can be done in $O(n)$ steps of additions and comparisons.*

The reductions mentioned are all simple transformations of the sequence which lead to a recognition algorithm which tests the unigraphicality of a sequence in $O(n^2)$ steps.

The earliest linear time unigraph recognition algorithm was by Kleitman [20] and made use of Ferrer diagrams. A characterisation of unigraphs, which led to another linear time recognition algorithm, was given by Borri et al. [21]. Their work is a continuation of, among others, the work of Johnson [22], who was the first to begin research into the structure of unigraphs. Their theorem completely characterises unigraphs in a straightforward manner, so is reproduced here. All required definitions can be found in Appendix A.

Theorem 2.4 *A graph G is a unigraph if and only if its vertex set can be partitioned into three disjoint sets V_K, V_S and V_C such that:*

1. *$V_K \cup V_S$ induces a split unigraph F in which K is the clique and S is the stable set;*
2. *V_C induces a crown H and either H or $\bar{H}$ is one of the following graphs:*

$$C_5, mK_2 (m \geq 2), U_2(m, s), U_3(m); \tag{2.7}$$

3. *the edges of G can be colored red and black so that:*
   a. *the red partial graph is the union of H and of vertex-disjoint pieces $P_i, i = 1, \ldots, z$. Each piece P_i (or $\bar{P}_i$, or P_i^I or $\bar{P}_i^I$) is one of the following graphs:*

$$K_1, S_2(p_1, q_1; \ldots; p_t, q_t), S_3(p, q_1; q_2), S_4(p, q), \tag{2.8}$$

   considered without the edges in the clique;
   b. *The linear ordering $P_1, \ldots, P_z$ is such that each vertex in V_K belonging to P_i is not linked to any vertex in V_S belonging to $P_j, j = 1, \ldots, i-1$, but is linked by a black edge to every vertex in V_S belonging to $P_j, j = i+1, \ldots, z$. Furthermore, any edge connecting either two vertices in V_K or a vertex in V_K and a vertex in V_C is black.*

For completely different characterisations of unigraphs, as well as extensions to digraphs, see [23] and [24].

2.2 Mathematical analysis of the walk matrix

The definition of the term graphic could be applied to a walk matrix, just as it was applied to degree sequences. We could then ask if there is any definitive test for the graphicality of a walk matrix and if any algorithm can implement the test efficiently. The concept of a unigraph could also be generalised. Starting from the first row of the walk matrix each additional row of the walk matrix has the potential to contain extra discriminatory information. A graph could be defined to be *nth row unigraphic* if n is the smallest integer such that the walk matrix up to row n is sufficient to uniquely define the graph, up to isomorphism. We could then ask if there is any complete characterisation of nth row unigraphs, analogous to Theorem 2.4.

These are all clearly important but difficult questions to answer. The work summarised here attempts to address these questions by beginning to explore what information is contained within the walk matrix.

2.2.1 Walk matrix decomposition

Let a *cycle* be a closed walk on a graph which does not visit any node more than once. A node is said to be *on an n-cycle* if it forms part of a cycle of lengh n. Any element in row n of the walk matrix then can be written as a function of the number of n-cycles which node is on, along with walk matrix elements with row numbers $r < n$.

To complete a closed walk of odd length, a cycle of odd length must be traversed at some point. There is therefore only one way to make closed walks of length 3 and that is by traversing 3-cycles, which are also referred to as *triangles*. If the number of n-cycles on node i is denoted $c_n(i)$, then the ith element of the 3rd row of the walk matrix can be written as

$$w_3(i) = 2c_3(i). \tag{2.9}$$

This means that the *clustering coefficient*, which will be defined in Section 3.1, can easily be calculated from the walk matrix.

Closed walks of length 4 must be divided into 3 classes. Class a walks traverse a 4-cycle which the node is on. There are $2c_4(i)$ such walks from node i. Class b walks step to a neighbour of i, step back, step to a neighbour again and finally step back again. There are $w_2(i)^2$ such walks. Finally, class c walks take 2 non-repeated steps from i, then return by exactly the same path. There are $\sum_j w_2(j)$ such walks, where the sum is over all nodes adjacent to node i. This covers every possible type of closed walk of length 4, so $w_4(i)$ can always be written

$$w_4(i) = 2c_4(i) + w_2(i)^2 + \sum_j (w_2(j) - 1). \tag{2.10}$$

This decomposition shows the first limit to the information contained in the walk matrix. Since $w_2(i)$ is known, the combination $2c_4(i) + \sum_j (w_2(j) - 1)$ can be found. Both $c_4(i)$ and $\sum_j (w_2(j) - 1)$ constitute important topological information (the number of 4-cycles on node i and the total degree of its neighbours) but the two can not, in general, be separated.

The total decomposition of closed walks of length 5 from node i is

$$w_5(i) = 2c_5(i) + 4c_3(i)w_2(i) + 2\sum_j (w_2(j) - 1)c_3(i,j) - c_3(i) + 2\sum_j c_3(j) - 4c_3(i), \qquad (2.11)$$

where each sum is over nodes adjacent to node i. For the derivation of this decomposition and the definition of $c_3(i,j)$ see Appendix B. Again, a limitation to the information carried in this row of the walk matrix is clear. Important topological information regarding the 5-cycles on node i and the triangles on its neighbours is contained in the row elements but it is not in general possible to disentangle this information.

As a final remark on this subject, a different approach to the decompositions described here would be to obtain the generating functions of the walk matrix columns. For more information on generating functions, see [25].

2.2.2 Row n unigraphs

I now present an example of a class of graphs which are not uniquely characterised by their degree sequence, but are indeed uniquely characterised when further rows of the walk matrix are included.

Let a *path graph* be a graph with no branches and no cycles (i.e. simply a line of connected nodes). Let a *balloon graph* be a graph made from any path graph plus one edge which connects one of the external nodes (i.e. a node from one of the ends of the path) to one of the internal nodes. Clearly, for a balloon graph of cardinality n, there are $n-3$ graphs you can make which are not isomorphic to each other. It should also be clear that all such graphs can not be distinguished by their degree sequence, since the degree sequence for them all is $(3, 2, 2, \ldots 2, 1)$. It can be shown, however, that the walk matrix will quickly distinguish them.

Proposition 2.1 *Any of the $n-3$ different balloon graphs of cardinality n can be distinguished from one another by the first $(n-1)/2$ rows of the walk matrix (round $(n-1)/2$ down for even n).*

For a proof of this proposition see Appendix C. For analysis on tree graphs which are characterised by the walk matrix, see Appendix D. For an analysis of the combinatorics of closed walks on paths see Appendix E.

2.2.3 The walk matrix as a surface

Degree distributions of real-world networks follow strict rules which will be discussed in Section 3. In the context of the walk matrix, the concept of a network's degree distribution is generalised to a *walk distribution surface*. Such a surface is pictured in Figure 2, for a random *scale-free* network created using the Barabasi-Albert model [41]. An interesting question is whether or not this surface abides by any particular rules. If the surface can be approximated by a function, the varying form of this function could be used to classify different types of network. This is an important area for further research.

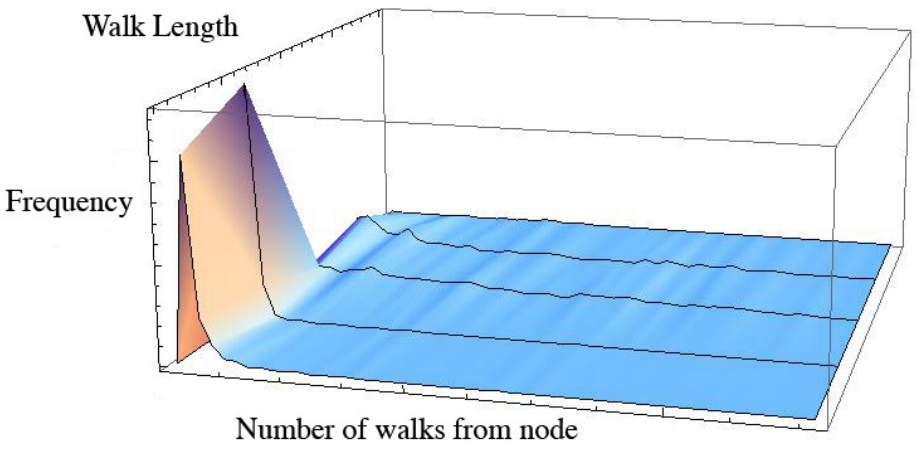

Figure 2: This surface is a visualisation of the walk matrix of a random scale-free network created using the Barabasi-Albert model [41]. It is essentially a 3D histogram with walk length on one axis and the number of walks of that length which can be taken from a node on the other.

3 Biological networks

Up to now, the treatment of graphs and the walk matrix has been entirely mathematical. The hope is, however, that since the walk matrix of a network encodes important topological information, that it may be useful in a biological context. Section 3.1 briefly reviews network biology with a focus on how topological analysis has been utilised. Section 3.2 then describes an attempt to use the walk matrix to make some predictions in a biological network.

3.1 Topology of biological networks

Networks arise in biology in a variety of scenarios. One of the most prominent is the protein-protein interaction (PPI) network [26] for which each node is a protein and each edge signifies an interaction. The PPI network is fundamental to biology since PPIs underpin all biological processes. Other biological networks include metabolic networks [27], transcription-regulatory networks [28], phosphorylation networks [29] and disease networks [30].

One of the most important discoveries in network science was that a huge range of real-world networks have *scale-free* degree distributions [3]. A distribution is called scale-free when it is characterised by a power law relationship and indeed, a large proportion of real-world networks have a degree distribution which is well approximated by $P(k) \sim k^{\gamma}$, with $2 \leq \gamma \leq 3$ and where $P(k)$ is the probability that a node has degree k.

Scale-free networks are in stark contrast to random graphs (such as the Erdős-Rényi random model [31]), which have Poissonian degree distributions where most degree values are close to the mean, $\langle k \rangle$. A scale-free network has, instead, divergent $\langle k \rangle$ and is characterised by a large number of nodes with small degree and a few with very large degree, called *hubs*.

Metabolic [32], PPI [33] and gene expression networks [34] have all been shown to be scale-free, with important biological consequence. It was shown that the hubs, characteristic of scale-free networks, are more likely to be lethal if removed from the network [33]. In all, however, a scale-free

topology makes a network more robust to random deletions, such as those which occur in nature [33]. It was also found that hubs in PPI are more frequently altered in cancer [35, 36].

In 1999, Barabasi et al. [37] proposed an evolutionary model explaining the prevalence of scale-free networks. The model consists of two components, *growth* and *preferential attachment*. Networks grow over evolutionary time scales as new nodes are added to the network. Preferential attachment is the assertion that new nodes are more likely to connect to nodes which have a large degree. Barabasi et al. showed this lead to a scale free network (with exponent $\gamma = 2.9 \pm 0.1$) by simulating a growing network where newly introduced nodes connect to any node i with a probability π_i given by

$$\pi_i = \frac{k_i}{\sum_j k_j}. \tag{3.1}$$

The best explanation for preferential attachment lies in gene duplication. When a gene is duplicated the network gains an extra node with exactly the same connections. A node with a high degree is more likely to be connected to a node which duplicates and so it is more likely that its degree will grow.

Another important topological measure is the *clustering coefficient*, $C(i)$, of a node i in a network [38]. It is defined as

$$C(i) = \frac{2c_3(i)}{w_2(i)(w_2(i)-1)}. \tag{3.2}$$

It is argued that the clustering coefficient can be used as a measure of modularity, i.e. the degree to which a network is segmented into separate modules with separate functions. The average clustering coefficient of biological networks has been found to be significantly higher than random in many cases, from the metabolic network of *Escherichia coli* [32] to the nervous system of *Caenorhabditis elegans* [39].

The average clustering coefficient of biological networks appears to be independent of the number of nodes in the network [40]. This is significant because for a random scale-free network, the average clustering coefficient scales as $N^{-0.75}$ [41]. Another property concerns the relationship between the clustering coefficient and the degree of a node. In a random graph, one would expect degree and clustering coefficient to be independent. In a large number of biological networks, however, it has been discovered that the relationship between the two is well approximated by $C(k) \sim k^{-1}$ [40]. This relationship was first noticed by Dorogovtsev et al. [42] in a class of scale-free graphs which were defined using a deterministic growing rule.

The combined properties of a power law degree distribution (described above) and a seemingly modular topology seem at first to be in contradiction. A power law degree distribution implies the presence of hubs which, since they connect many nodes, prohibit a truly modular topology. These conflicting ideas were reconciled by Ravasz et al. [40, 43] who introduced the idea of a *hierarchical network*. The model hierarchical network which they posit is built by a simple algorithm, described in Figure 3. The hierarchical network is easily shown to be scale free, with a high clustering coefficient which obeys $C(k) \sim k^{-1}$.

Han et al. [44] discovered that biological hubs could be divided into two categories, *party* hubs and *date* hubs. Party hubs are defined as the proteins in PPI networks for which gene expression levels

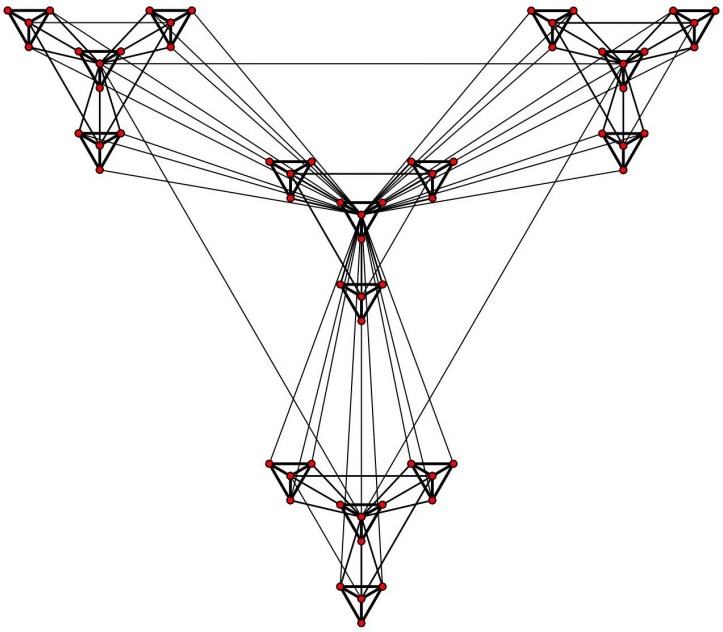

Figure 3: The model hierarchical network of Ravasz et al. [40, 43] is created by a simple algorithm. Start with a complete graph of, say, 3 nodes. Define one node to be central and the rest peripheral. Make 3 more exact copies of this graph and connect the peripheral nodes of each to the central node of the original. The resulting graph can then again be copied 3 times and similarly peripheral nodes are connected to the central node. This can be repeated indefinitely.

correlate strongly with neighbours. These proteins were found to interact mainly with proteins with similar functional roles, i.e. they interacted only within functional modules. Date hubs, on the other hand, were defined as having lower gene coexpression between their neighbours and themselves. It was found that date hubs act on a more global scale, organising the modules without playing important roles within them.

Taylor et al. [45] found that party hubs are more highly conserved between yeast and human PPI networks than date hubs. Previously this had been argued to be due to evolutionary restrictions caused by the macromolecular complexes that date hubs tend to form part of [46]. From a biochemical perspective, Taylor et al. also found that date hubs had, on average, significantly more numerous (but smaller is size) domains than party hubs.

Betweenness is a toplogical measure of centrality in networks which originated in communication theory [47] and was first quantified by Freeman [48]. For every pair of nodes, there is one or more shortest paths between the nodes. A node is said to have high betweenness if many such shortest paths pass through the node.

Betweenness can be calculated for any node as follows. If nodes i and j have g_{ij} paths between them, which are all shortest paths, and if node k lies on exactly $g_{ij}(k)$ of these, then the partial betweenness of node k is

$$bij(k) = \frac{g_{ij}(k)}{g_{ij}}. \tag{3.3}$$

The overall betweenness of node k is then given by

$$B(j) = \sum_{i<j} \sum_{j} b_{ij}(k). \tag{3.4}$$

Yu et al. [49] studied the role of betweenness in yeast PPI and expression networks. They found that betweenness is a much better predictor of essentiality than degree in regulatory networks, whereas degree is better in interaction networks. This they put down to a difference in the nature of the two networks. Regulatory networks, they argued, have a more well defined notion of information flow, since they have directed edges. They argued that betweenness is expected to be more important when information flows through a network because it is then that information will 'bottle-neck' at nodes with high betweenness.

Yu et al. [49] also used betweenness in a useful extension of the work of Han et al. [44], described above. They found that in hubs, betweenness correlates negatively with the average coexpression between a hub and its neighbours. This means that the party and date hubs described by Yu et al. can be identified on purely topological measures.

Taylor et al. [45] found that date hubs were altered significantly more frequently in cancer genes as defined by Online Mendelian Inheritance in Man [50]. They found 256 hubs which for which the average correlations between neighbours varied as a function of outcome and found this to be an effective prognostic tool, above and beyond other clinical data.

Yet another important tool applied to the analysis of network topology is the measurement of entropy. There are a large number of possible definitions of entropy which depend on topology only and are used for a variety of purposes. For a review see [51] and for a definition used by Demetrius et al. [52], along with their definition of network robustness, see appendix F.

In [51], Demetrius et al. show that the entropy they define correlates positively with *robustness*, where robustness is the ability of a network to be unaffected by random fluctuations. In a continuation of this work, published in [53], the authors show that there is strong correlation between the entropy of a node and the likelihood that deleting that node from the network will be lethal. Using a different definition of entropy, Teschendorff et al. [54] show that a metastatic cancer phenotype is characterised by an increase of this entropy.

3.2 Using the walk matrix

The fact that the walk matrix of a network contains important topological information was demonstrated in Section 2.2. It is therefore suggested that the walk matrix could be a useful tool for the analysis of biological networks. Here, the walk matrix is applied to a version of the human PPI network [55] with 10720 nodes and 120454 edges. A large database of human gene expression data, obtained in [56], is also used. This data was obtained using extensive mRNA microarrays which measured gene expression in 79 different human tissues and was used to analyse global trends in gene expression.

Since it is known that $w_2(i)$ follows a scale-free distribution, an interesting preliminary question is what distributions do w_x (for $x > 2$) follow. The first few decompositions in section 2.2.1 show that this question is complicated by the fact that $w_x(i)$ depends on $w_y(i)$, for $y < x$, in a complicated way.

For instance, the decompositions show that for even x, one component of $w_x(i)$ is always $w_2(i)^{x/2}$. Despite this complexity it is interesting to gain a qualitative view of the distribution and this is displayed in the surface in Figure 4 a) - d).

In 2001, Ge et al. [57] made an important observation, fundamental to the integration of gene expression and PPI networks. They showed that proteins in clusters in the *Saccharomyces cerevisiae* PPI network had, on average, significantly more highly correlated gene expression than proteins from different clusters. This important discovery demonstrates that gene expression information can be predicted solely from topology of the PPI network, which demonstrates that the two are intimately linked.

An interesting question then, is whether or not the walk matrix can be used to increase the power of the PPI network in predicting gene expression. To address this question, we first demonstrated a simple relationship between the two data sets. Firstly, the correlation between the expression of different genes was measured for each pair of genes in the gene expression data. This was done using the Pearson's correlation coefficient, ρ. We then compared ρ for genes whose proteins were connected or unconnected in the PPI network.

The distribution of ρ's for connected and unconnected genes are shown in Figure 5 a) and b), respectively. The two distributions are significantly different, with connected genes having a broader distribution of ρ's (standard deviations of 0.23 and 0.19, respectively) and a slightly higher average (means of 0.19 and 0.15, respectively). To demonstrate the significance of this result, we created a random scale-free network with the same number of nodes using the Barabasi-Albert model [41]. The nodes are randomly assigned a gene and Figure 5 c) shows the distribution of ρ for genes which are connected in this network. This distribution has a mean of 0.15 and and standard deviation 0.19. The higher mean in the distribution of connected nodes shows the relationship between the two networks, as expected.

We then attempted to use the walk matrix to obtain further information regarding the relationship between the two networks. An initial problem was that the number of closed walks from nodes with high degree was found to increase rapidly in the PPI network. In many cases there were $\sim 10^{10}$ closed walks of length 6 from a node. This limited the calculation of the walk matrix up to the 5th row, meaning that each node was characterised by 5 numbers and that these numbers were to be used to predict gene expression.

It turned out that, for general pairs of nodes, it was difficult to use the walk matrix to predict gene expression. No significant correlations could be found between correlation in gene expression and walk matrix elements.

An interesting result was discovered, however, regarding network hubs. As described in Section 3.1, Han et al. [44] showed it was useful to divide network hubs into party hubs and date hubs, based on the coexpression between the hub and its neighbours. Further, Yu et al. [49] showed that party hubs are characterised by significantly lower betweenness than date hubs.

We defined hubs as nodes with > 150 connections, of which there were 245 in our network. For each of these hubs, we then calculated the average ρ value for the gene expression correlation between the hub and its neighbours. This we will denote $\bar{\rho}$. As expected from the results of Yu et al. [49], there was significant negative correlation between $\bar{\rho}$ and the betweenness of the hubs. The Pearson correlation coefficient between these two variables was $\rho = -0.22$. We also found a correlation of

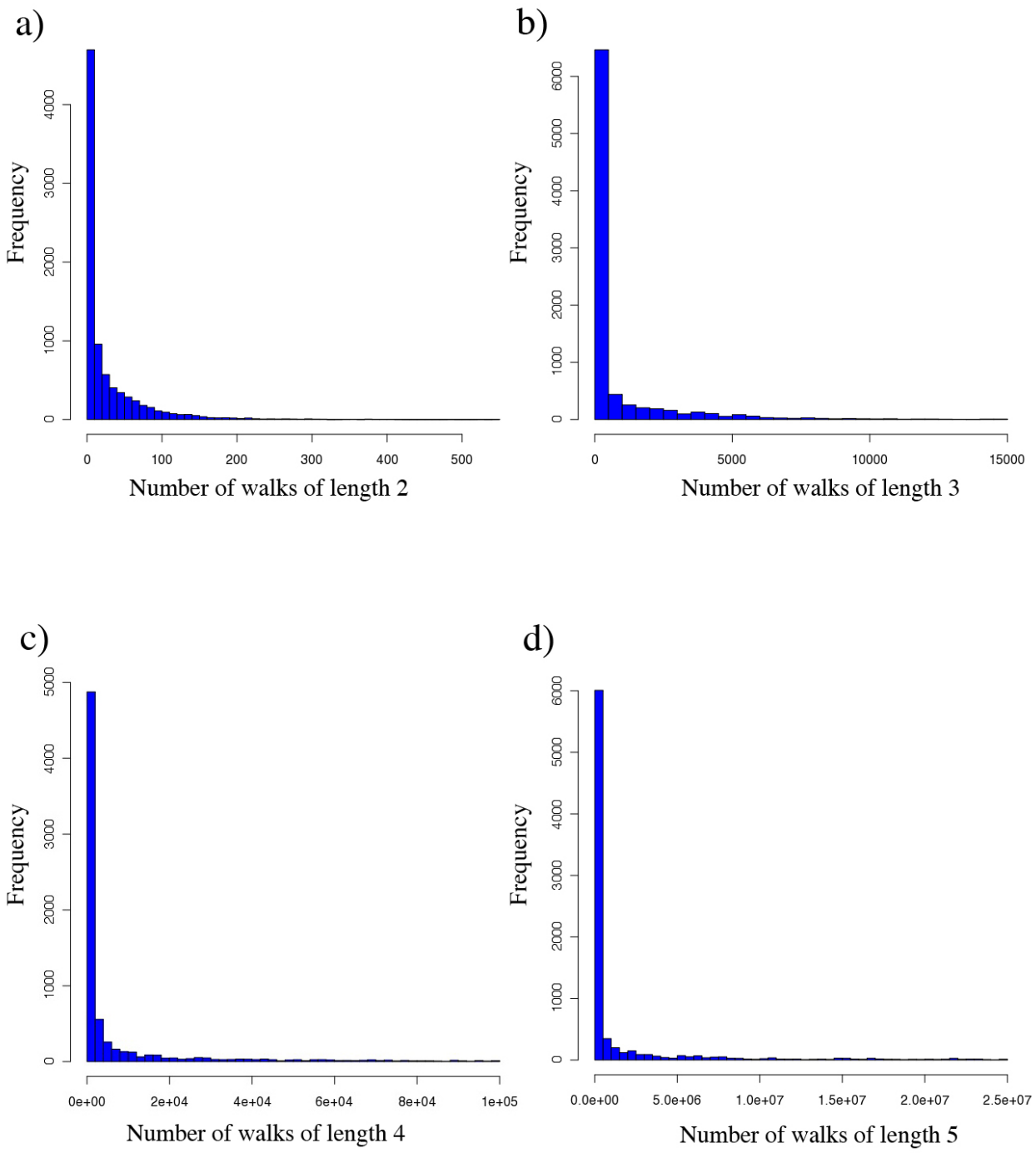

Figure 4: Figures a) - d) show histograms of the number of walks which can be made from a node, for different walk lengths, for the PPI network described in the text.

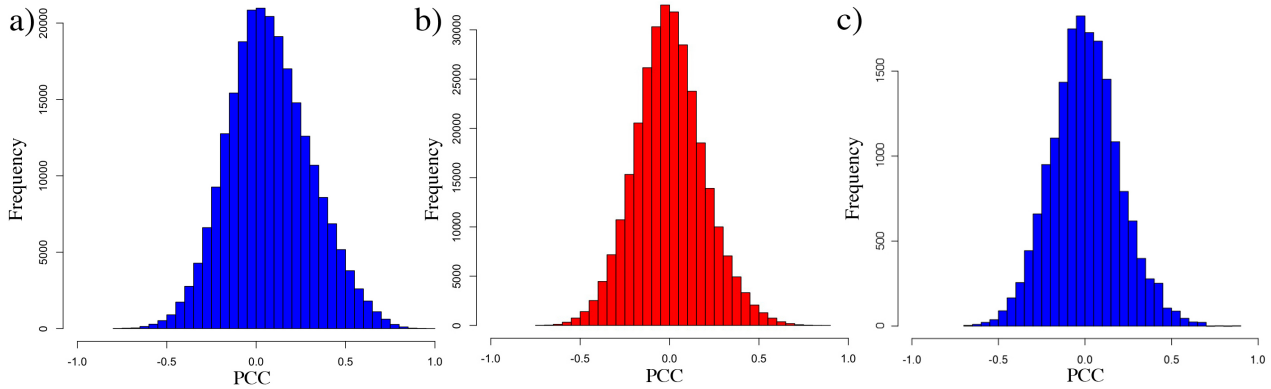

Figure 5: The ρ value histograms for a) the connected nodes in the PPI network, b) unconnected nodes in the PPI network and c) connected nodes in a random scale-free network.

$\rho = 0.22$ between $\bar{\rho}$ and the clustering coefficient of the node.

We hypothesised that the walk matrix could provided a measure which distinguishes between party and date hubs more strongly than betweenness. The reason for this is that party hubs are know to be more prominent in densely connected regions, which represent functional modules in the network. Date hubs, on the hand, connect to many nodes but are less likely to be in densely connected regions since they are known to regulate numerous *separate* functional modules. This is the reason that the clustering coefficient correlates with $\bar{\rho}$ for hubs. The clustering coefficient, however, measures the density of connections in too small a region around the hub, as it only considers neighbours of the hub.

We argued that considering the number of closed walks of length 6 from the node is more general. A high number of walks of length 6 shows that the node is in a highly connected region, taking into account nodes that are a distance of 2 and 3 from the hub. We proposed a new measure of clustering, $K(i)$, inspired by the walk matrix, given by

$$K(i) = \frac{w_6(i)}{w_2(i)^\alpha}. \tag{3.5}$$

The degree of the hub, $w_2(i)$, appears in this definition as a crude normalisation, for the same reasons that $w_2(i)$ appears in the definition of the clustering coefficient of Equation 3.2. Our hypothesis was immediately supported by a correlation of $\rho = 0.33$ between $K(i)$ and $\bar{\rho}$, when an optimum value of $\alpha = 1.5$ was found. This significantly stronger correlation suggests that this new measure of clustering may distinguish between party and date hubs significantly more effectively than betweenness. Since there are only 245 hubs in the network used here, it will be necessary to gather further evidence from other networks in order to confirm this finding.

4 Conclusion

The walk matrix may turn out to be an important tool for the analysis of topology in biological networks. Section 2.2 showed that the walk matrix contains important topological information but also that there is a limit to the information that it carries. Evidence has been found that, in

the specific case of network hubs, the walk matrix can be used as an effective predictor of gene coexpression between a hub and its neighbours.

There is much room for further work using the walk matrix. A full characterisation of the graphs which are row n unigraphic would be of great value, however, this is a highly nontrivial problem. The surfaces produced by plotting the distributions of walks of different lengths are another feature which should be explored. There are also many potentially useful ways in which the matrix could be used to define a measure of complexity for its associated network.

The evidence provided here, which suggests that the walk matrix can be used to predict the coexpression of genes adjacent to hubs, must be verified using separate data. It will be interesting to see whether the effect is specific to PPI networks or whether it can be found in other networks. Since certain classes of hubs are known to be highly essential, it would also be interesting to investigate what the walk matrix can be used to predict regarding essentiality.

Finally, the form of the clustering measure we propose in Equation 3.5 was chosen heuristically. Using the decompositions of Section 2.2.1, it will be possible to make a different but similar definition which is rooted more firmly in the theory.

References

[1] Albert-Laszlo Barabasi. The network takeover. *Nature Physics*, 8(1):14–16, 2012.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell.* Garland Science, 2008.

[3] Albert-Laszlo Barabasi and Zoltan N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101, 113 2004.

[4] N. Biggs, E. Lloyd, and R. Wilson. *Graph Theory, 1736-1936.* Oxford University Press, 1986.

[5] Lowell W. Beineke, Robin J. Wilson, and Peter J. Cameron. *Topics in Algebraic Graph Theory: Algebraic Graph Theory.* Cambridge University Press, 2004.

[6] A. Cayley. On the mathematical theory of isomers. *Philosophical Magazine*, 1874.

[7] J. Howard Redfield. The theory of group-reduced distributions. *American Journal of Mathematics*, 49(3):433–455, 1927.

[8] G. Pólya. Kombinatorische anzahlbestimmungen für gruppen, graphen und chemische verbindungen. *Acta Mathematica*, 68(1):145–254, 1937.

[9] P. Erdős and T. Gallai. Graphs with prescribed degrees of vertices (Hungarian). *Mat. Lapok*, 11:264–274, 1960.

[10] A. Tripathi and S. Vijay. A note on a theorem of erdős and gallai. *Discr. Math.*, 265:417–420, 2003.

[11] V. Havel. A remark on the existence of finite graphs. *Časopis Pest. Mat.*, 80:477–480, 1955.

[12] S. Hakimi. On the realizability of a set of integers as degrees of the vertices of a graph. *SIAM J. Appl. Math.*, 10:496–506, 1963.

[13] M. Behzad and G. Chartrand. No graph is perfect. *Amer. Math. Monthly*, 74:962–963, 1967.

[14] R. Tyshkevich and A. Chernyak. Unigraphs, i. *Vesti Akademii Navuk BSSR*, 5:5–11, 1978.

[15] R. Tyshkevich and A. Chernyak. Unigraphs, ii. *Vesti Akademii Navuk BSSR*, 1:5–12, 1979.

[16] R. Tyshkevich and A. Chernyak. Unigraphs, iii. *Vesti Akademii Navuk BSSR*, 2:5–11, 1979.

[17] R. Tyshkevich and A. Chernyak. Canonical decomposition of a unigraph. *Vesti Akademii Navuk BSSR*, 5:14–26, 1979.

[18] R. Tyshkevich. Decomposition of graphical sequences and unigraphs. *Discrete Mathematics*, 220:201–238, 2000.

[19] Shuo-Yen R Li. Graphic sequences with unique realization. *Journal of Combinatorial Theory, Series B*, 19(1):42 – 68, 1975.

[20] D. Kleitman and S.-Y. Li. A note on unigraphic sequences. *Studies in Applied Mathematics*, 4:283–287, 1975.

[21] Alessandro Borri, Tiziana Calamoneri, and Rossella Petreschi. Recognition of unigraphs through superposition of graphs (extended abstract). In Sandip Das and Ryuhei Uehara, editors, *WALCOM: Algorithms and Computation*, volume 5431 of *Lecture Notes in Computer Science*, pages 165–176. Springer Berlin / Heidelberg, 2009.

[22] R.H. Johnson. Simple separable graphs. *Pacific J. Math.*, 56(1):143–158, 1975.

[23] P. Das. Unidigraphic and unigraphic degree sequences through uniquely realizable integer-pair sequences. *Discrete Mathematics*, 45:45–59, 1983.

[24] M. Koren. Sequences with a unique realization by simple graphs. *J. Combin. Theory*, 21B:235–244, 1976.

[25] Richard P. Stanley. *Enumerative Combinatorics*. Cambridge University Press, 2000.

[26] Ulrich Stelzl and Erich E Wanker. The value of high quality protein protein interaction networks for systems biology. *Current Opinion in Chemical Biology*, 10(6):551 – 558, 2006.

[27] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 10 2000/10/05/print.

[28] Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet*, 31(1):64–68, 05 2002/05//print.

[29] Boris Macek, Matthias Mann, and Jesper V. Olsen. Global and site-specific quantitative phosphoproteomics: Principles and applications. *Annual Review of Pharmacology and Toxicology*, 49(1):199–221, 2009.

[30] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[31] P. Erdos and A. Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61, 1960.

[32] Andreas Wagner and David A. Fell. The small world inside large metabolic networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1478):1803–1810, 2001.

[33] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

[34] David E. Featherstone and Kendal Broadie. Wrestling with pleiotropy: Genomic and topological analysis of the yeast gene expression network. *BioEssays*, 24(3):267–274, 2002.

[35] Jorrit J. Hornberg, Frank J. Bruggeman, Hans V. Westerhoff, and Jan Lankelma. Cancer: A systems biology disease. *Biosystems*, 83(23):81 – 90, 2006.

[36] Pall F. Jonsson and Paul A. Bates. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291–2297, 2006.

[37] Barabási. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[38] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 06 1998.

[39] J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. The structure of the nervous system of the nematode caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 314(1165):1–340, 1986.

[40] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[41] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.

[42] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Pseudofractal scale-free web. *Phys. Rev. E*, 65:066122, Jun 2002.

[43] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67:026112, Feb 2003.

[44] Jing-Dong J. Han, Nicolas Bertin, Tong Hao, Debra S. Goldberg, Gabriel F. Berriz, Lan V. Zhang, Denis Dupuy, Albertha J. M. Walhout, Michael E. Cusick, Frederick P. Roth, and Marc Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93, 2004.

[45] Ian W Taylor, Rune Linding, David Warde-Farley, Yongmei Liu, Catia Pesquita, Daniel Faria, Shelley Bull, Tony Pawson, Quaid Morris, and Jeffrey L Wrana. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, 27(2):199–204, 2009.

[46] Hunter B Fraser. Modularity and evolutionary constraint on proteins. *Nat Genet*, 37(4):351–352, 04 2005.

[47] Alex Bavelas. A mathematical model for group structure. *Applied Anthropology*, 7:16–30, 1948.

[48] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):pp. 35–41, 1977.

[49] Haiyuan Yu, Philip M Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*, 3(4):e59, 04 2007.

[50] Victor A. McKusick. Mendelian inheritance in man and its online version, omim. *Am J Hum Genet.*, 80(4):588–604, 2007.

[51] Matthias Dehmer. Information theory of networks. *Symmetry*, 3(4):767–779, 2011.

[52] Lloyd Demetrius and Thomas Manke. Robustness and network evolution-an entropic principle. *Physica A: Statistical Mechanics and its Applications*, 346(3-4):682 – 696, 2005.

[53] Thomas Manke, Lloyd Demetrius, and Martin Vingron. Lethality and entropy of protein interaction networks. *Genome Informatics*, 16(1):159–163, 2005.

[54] Andrew E Teschendorff and Simone Severini. Increased entropy of signal transduction in the cancer metastasis phenotype. *BMC Systems Biology*, 4(104), 2010.

[55] E.G. Cerami, B.E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G.D. Bader, and C. Sander. Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(1):685–690, 2011.

[56] Andrew I. Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A. Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael P. Cooke, John R. Walker, and John B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067, 2004.

[57] Hui Ge, Zhihua Liu, George M. Church, and Marc Vidal. Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nat Genet*, 29(4):482–486, 12 2001.

A Definitions required for theorem 2.4

If $[A]^k$ denotes the set of all k-element subsets of A, then the *complement* of G is the graph $\bar{G}$ on V with edges $[V]^2 \setminus E$. A graph whose every node is connected to every other node is called *complete*. A complete graph with n nodes is denoted K_n. A subset of vertices U within G for which every vertex in U is connected with every other vertex, is called a *clique*. A graph which consists of n vertices of degree 2, which form a closed chain is called a *cycle graph* and is denoted C_n.

A graph $G(U, V, E)$ is called *bipartite* if and only if every edge in E makes a connection between, but not within, the disjoint vertex sets U and V. A *complete bipartite graph* is a bipartite graph where every vertex in U is connected to every vertex in V and is denoted $K_{n,m}$ where n and m are the cardinalities of U and V. A *star graph*, S_k, is the complete bipartite graph $K_{1,k}$.

A set of edges is called a *perfect matching of dimension h* of vertex set X onto vertex set Y if and only if X and Y have equal cardinality and each edge in E is connected to exactly one member of X and one of Y. A perfect matching of dimension m is denoted mK_2.

A vertex is *isolated*, *stable* or *independent* with respect to a set of other vertices if it makes no connections with any of those vertices. G is said to be *split* if there is a partition $V = V_K \cup V_S$ of its vertices such that the induced subgraphs K and S are complete and stable, respectively. If the vertex set of G can be partitioned into three sets, V_K, V_S and V_C such that every vertex in V_C is adjacent to every vertex in V_K but is adjacent to no vertex in V_S, then the subgraph induced by V_C is called a *crown*.

$U_2(m, s)$ will denote the disjoint union of the perfect matching mK_2 and the star $K_{1,s}$, for $m \geq 1, s \geq 2$. Define $U_3(m)$, for $m \geq 1$, as the graph made by the cycle C_4 and m copies of K_3, all sharing one of their vertices. $S_2(p_1, q_1; \ldots; p_t, q_t)$ is made by adding all the edges connecting the centers of l nonisomorphic arbitrary stars K_{1,p_i}, $i = 1, \ldots, t$, each one occurring q_i times, where $p_i, q_i, t \geq 1$, $q_1 + \ldots + q_t = l \geq 2$. $S_3(p, q_1; q_2)$ is made by taking a graph $S_2(p, q_1; p+1, q_2)$ where $p \geq 1, q_1 \geq 2$ and $q_2 \geq 1$. Add a new vertex v to the stable part of the graph and add the set of q_1 edges $((v, w) : w \in VK, deg_{V_S}(w) = p)$. Make $S_4(p, q)$ by taking the graph $S_3(p, 2; q), q \geq 1$ and adding a new vertex u to the clique part. Connect u to each vertex in the stable part except for v.

Finally, for a split graph $G(V_K \cup V_S, E)$, the authors define its *inverse* G^I as the graph obtained from G by deleting the set of edges $((a_1, a_2) : a_1, a_2 \in V_K)$ and adding the set of edges $((b_1, b_2) : b_1, b_2 \in V_S)$. i.e. the inverse of a split graph swaps the roles of the clique and the stable set.

B Decomposition of $w_5(i)$ derivation

Closed walks of length 5 are divided into 4 classes. Class a are the walks which traverse 5-cycles. There are $2c_5(i)$ such walks from node i. The remaining 3 classes must at some point traverse a triangle and use up 2 steps somewhere else. Class b walks traverse a triangle on node i but either start or end by stepping to and from a node adjacent to i. There are $4w_2(i)c_3(i)$ such walks. Class c walks also traverse a triangle on node i but take the extra 2 steps from either of the other 2 nodes of any triangle. To enumerate this class introduce the notation $c_n(i, j)$ for the number of n-cycles shared by nodes i and j. There are then $2\sum_j (w_2(j) - 1)c_3(i, j) - c_3(i)$ class c walks, where again the sum is over nodes adjacent to node i.

The final type of closed 5-walks, type d, traverses triangles which are attached to neighbours of node i, but not attached to node i itself. There are $2\sum_j c_3(j) - 4c_3(i)$ such walks.

C Balloon graph proof

Take a to be the path graph which starts at node x and extends indefinitely from that point. Graph a therefore is a *semi-infinite* path. Take graph b to be the node y with 3 semi-infinte paths extending from it. Calculate the walk matrix columns for nodes a and b up to $(n-1)/2$. Now calculate the walk matrix for the balloon graph up to row $(n-1)/2$. Compare the columns of this matrix which have $w_2 = 1$ and $w_2 = 3$ with the calculated columns of nodes x and y respectively. In graphs where the connection is made close to the node with degree 1, walks from this node will reach the node with degree 3 within $(n-1)/4$ steps. The node of degree 3 presents more options to a walk which reaches it so vector a and vector x will differ. If the connection is made on the half of the graph which is far from the node with degree 1 then the loop created will in general be small. This means that walks of length $(n-1)/2$ can always transverse the cycle, which in turn causes vector b to differ from vector x. Therefore, any balloon graph can be identified by comparing vector a with x or b with y.

D Tree graphs

Tree graphs are defined as graphs which contain no cycles. This makes them interesting here as the above decompositions of the walk matrix are greatly simplified by the absence of cycles. The decompositions are now given by

$$\begin{aligned} w_3(i) &= 0 \\ w_4(i) &= w_2(i)^2 + \textstyle\sum_j (w_2(j) - 1) \\ w_5(i) &= 0 \\ w_6(i) &= w_2(i)^3 + 2w_2(i)\textstyle\sum_j w_2(j) + \sum_j w_2(i)^2 + \sum_k w_2(k), \end{aligned} \tag{D.1}$$

where sums over j are over all nodes adjacent to node i and the sum over k is over all nodes which can be reached from i in 2 steps.

Only a small subset of trees are unigraphs. The above set of decompositions show that a larger subset of trees are row n unigraphs fro $n > 1$, since the walk matrix clearly contains additional discriminatory information, as compared to the degree sequence. The simplest example is that from $w_4(i)$, the total degree of and node's neighbours can easily be found. This information alone allows the walk matrix to discriminate between a large class of tree graphs which have identical degree sequences. A full characterisation of which tree graphs are row n unigraphs is an important goal for future work.

E Path graphs

Start from a node a distance n from the one of the ends of the path (assume the other end is further away). Consider closed walks from this node. If the maximum closed walk length you consider is 2n

then the path will "appear infinite". i.e. the number of closed paths does not distinguish this path from an infinite path since the furthest you can reach (before having to come back) is n steps. So here we obtain a rule for the number of closed walks of length x from a node on an infinite path.

The sequence for walks of length $n = (1, 2, 3, 4, 5, 6, 7, 8, ...)$ is $w = (2, 6, 20, 70, 252, 924, 3432, 12870, ...)$. From https://oeis.org/ it was found that this is the sequence of 'Central binomial coefficients' and is given by: $C(2n, n) = (2n)!/(n!)^2$ where C stands for the binomial coefficient. i.e. these are the numbers down the centre of pascals triangle.

Now start at a node at one of the ends of a path. Again we consider closed walks for which the path "appears infinite", but now it only appears infinite in one direction (since we start at one of the ends). Again we obtain how many closed walks there are for a given walk length.

The sequence, as given by the left most column above, for $n = (1, 2, 3, 4, 5, 6, 7, 8, ...)$, is $w = (1, 2, 5, 14, 42, 132, 429, 1430, ...)$. From https://oeis.org/ it was found that this is the sequence of 'Catalan numbers' and is given by $C(2n, n)/(n + 1) = (2n)!/(n!(n + 1)!)$. These are also called the Segner numbers. It is interesting that the number of walks on the infinite path is given by (n + 1) times the number of walks on the half-infinite path. Its not immediately obvious why this is. It would also be interesting to see whether or not this manner of generating the Catalan numbers features in Richard Staney's 'Catalan addendum' of 198 combinatorial interpretations of the Catalan numbers.

It may be useful to continue this work in order to produce a general function $F(w, x, y)$ for the number of walks of length w on a path from a node which starts x nodes away from one end and y nodes away from the other.

F Demetrius entropy

Demetrius et al. [52] define the entropy as follows. Let A be the adjacency matrix of the network and let λ and $\{v_i\}$ be the leading eigenvalue and eigenvector respectively. Now define a set of matrices $M_A = \{P\}$ with $P = \{p_{ij}\}$ such that $p_{ij} \geq 0$ and $\sum_j p_{ij} = 1$. Assert that P must satisfy the condition $a_{ij} = 0 \iff p_{ij} = 0$. Let the stationary vectors of P be π so that $\pi = \pi P$. It can be shown via a variational principle that

$$\log \lambda = \sup_{P \in M_A} [-\sum_{ij} \pi_i p_{ij} \log p_{ij} + \sum_{ij} \pi_i p_{ij} \log a_{ij}]. \tag{F.1}$$

The matrix which satisfies this supremum has the unique form $p_{ij} = a_{ij} v_j / \lambda v_i$. Choosing this as the matrix to be used, entropy is defined as

$$H(P) = \sum_{i=1}^{N} \pi_i H_i \text{ where } H_i = -\sum_j p_{ij} \log p_{ij} \tag{F.2}$$

which simplifies to $H = \log \lambda$. Note that although P has the form of a dynamic matrix representing a Markov process, since it must satisfy equation F.1, it becomes defined entirely through A and is therefore a solely topological property of the network.

The authors of [52] then define a measure of robustness and show prove a relation between robustness and the entropy defined above. Let $P_\epsilon(t)$ be the probability that some measure of a given

network observable deviates by more than ϵ at a time t after some perturbation. The robustness, R, is then defined as

$$R = \lim_{t \to \infty} [-\frac{1}{t} \log(P_\epsilon(t))]. \tag{F.3}$$

Via a fluctuation theorem, they prove that robustness must be positively correlated to network entropy by their definitions, i.e. that $\Delta H \Delta R \geq 0$.