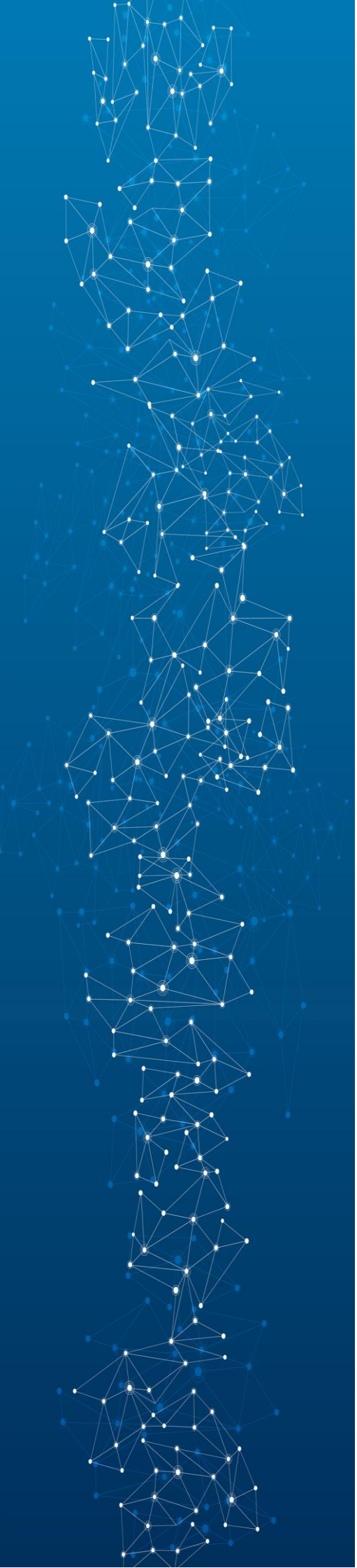


Mathematical analysis of the *Saccharomyces cerevisiae* polarity network

Spotting the differences between existing and non-existing networks in budding yeast

by Maaike Kempeneers

In partial fulfilment of the requirements for the degree of Bachelor of Science in Nanobiology

Date	September 2020 – January 2021
Student number	4466462
Supervisor	ir. E. Kingma
Thesis committee	Dr. ir. L. Laan Dr. J.L.A. Dubbeldam

Abstract

Cell polarity is an essential process for proliferation in *Saccharomyces cerevisiae*. This asymmetric organisation of cellular components is established by a group of proteins, in which Cdc42 plays a key role. This small GTPase acts as a molecular switch regulated by GTPase activating proteins (GAPs) and guanine nucleotide exchange factors (GEFs). The polarity establishing network is partially conserved amongst different strains and species of yeast, resulting in functional networks consisting of many possible combinations of these polarity proteins. However, there are even more combinations that do not exist in nature. The question we try to answer is whether we can distinguish the existing from the non-existing networks based on their mathematical properties. To answer this question, we first determined the physical interactions between the polarity proteins. After this, we constructed non-existing networks in different ways and determined the properties of those networks, both on the global-network-scale and on the individual-protein-scale. When non-existing networks were created using a protein distribution resembling the real distribution, no large differences could be found. More deviation in this distribution resulted in more deviation in the measured parameters. Furthermore, we found that Cdc42 is a critical node from a mathematical perspective and we found large differences in the networks which were constructed by removing Cdc42. Another finding is that Rac1 is a very important node as well, based on its similarities to Cdc42.

Table of contents

1. Introduction

1.1 Proliferation in *S. cerevisiae*

Saccharomyces cerevisiae, more commonly known as budding yeast or baker's yeast, is a fungal species widely used as a model organism. Reasons for this not only include its small size, short generation time and easy manipulation, but also the possibility to translate knowledge to other, more complex species. This is enabled by the eukaryotic nature of yeast, which creates a better representation of molecular mechanisms in other eukaryotic species than when a prokaryotic model organism is used.

S. cerevisiae reproduces asymmetrically in a process called budding (Figure 1). This process comprises of the determination of the budding site, consecutive growth of the bud, nuclear duplication and ultimately dissociation of the daughter cell. The initiation of this reproduction cycle requires a break of symmetry in the cellular organisation that determines the exact location of bud formation. This break of symmetry is called polarisation.

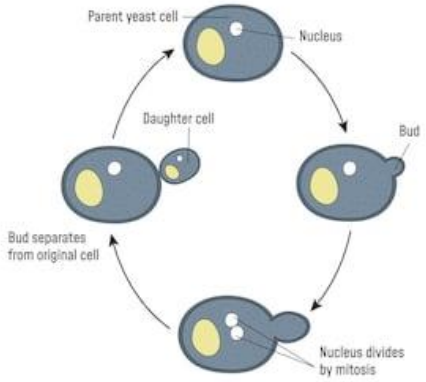


Figure 1 – Proliferation cycle of budding yeast. Polarity establishment is indicated by the upper-right arrow.

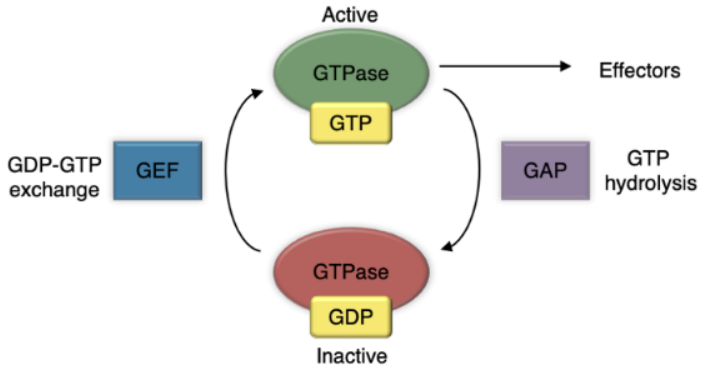


Figure 2 – Schematic overview of the interactions between GTPases, GAPs and GEFs (Antunez, 2019) .

1.2 Polarity establishing proteins

Cdc42 plays a key role in polarity establishment (Johnson, 1999). Together with other proteins, this small protein localises at the bud formation site. Cdc42 is essential for the assembly of actin cables, which on their turn transport other polarity proteins. The resulting asymmetric distribution of those proteins establishes the polarity of the cell (Irazoqui & Lew, 2004). Amongst this collection of polarity proteins are three different classes of proteins: GTPases, GTPase activating proteins (GAPs) and guanine nucleotide exchange factors (GEFs).

GTPases are a family of signal transducing proteins. A GTPase can be active or inactive, depending on whether it is bound to GTP or GDP respectively. The transition from active to inactive GTPases happens via the hydrolysis of GTP into GDP, a reaction which the GTPase itself performs. With doing so, it deactivates its own signal transduction. GAPs stimulate this hydrolytic activity of GTPases. Their binding results in a shortened lifespan of active GTPases. GEFs have the opposing effect. They regulate the release of GDP bound to the GTPase, after which the GTPase is able to bind to GTP again and is activated (Cherfils & Zeghouf, 2013). Figure 2 summarises the interaction between GTPases, GAPs and GEFs; Table 1 gives an overview of the GTPases, GAPs and GEFs present amongst the polarity proteins.

GTPase	GAP	GEF
Cdc42	Bem2	Bud3
Rac1	Bem3	Cdc24
Ras2	Msb3	Lte1

Rho3	Msb4	Scd1
Rsr1	Rga1	
Sec4	Rga2	

Table 1 – GTPases, GAPs and GEFs present in the polarity establishment network (SGD, 2012).

1.3 Polarity network in different fungal species

Diepeveen et al. (2018) selected 42 proteins related to polarity establishment, based on their physical and genetic interactions with GTPase Cdc42 and on their described functions in the polarity network. The authors found that the polarity network is partially conserved, as different species and even different strains within the same species contain a different subset of those 42 proteins. 149 unique networks were observed in 298 species and strains, in which a fundamental network of 23 proteins was conserved best: 95% of all examined species and strains express at least 14 of these fundamental proteins. Non of the fundamental proteins is conserved across all species considered in the study. The interactions between the proteins, however, are well-conserved across fungal species (Schoenrock, et al., 2017). This leads to diverse networks consisting of a different subset of proteins, but with constant interactions between the present proteins, as shown in the example in Figure 3. These proteins and their interactions can be represented in the form of a graph, in which the nodes represent proteins and the edges represent physical interactions between the proteins.

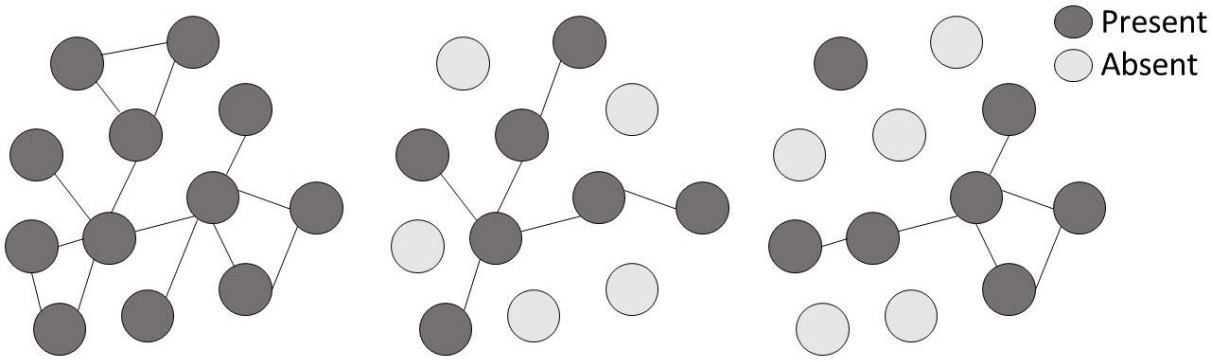


Figure 3 – Different networks constructed from the same proteins with equal interactions. When a different set of nodes is present, a different network is established.

1.4 Research question

Why do we find polarity establishment in yeast interesting? The answer lies in the link between cell polarity and cancer. Cdc42 has a human ortholog, which plays a role in the establishment of epithelial apical/basal polarity (NCBI, 2021), or in other words, makes it possible that epithelial cells can determine their orientation. When this polarity is disturbed, orientation is lost and cancer cells can invade their environment and metastasise (Wodarz & Näthke, 2007). This polarity establishment network is investigated in yeast due to its essential role in this species and the established knowledge on its genes. Better understanding of the yeast polarity network could eventually lead to a better understanding of its human counterpart.

When considering the set of proteins regarded to be part of the polarity module in this study, we observe that only a very small fraction ($149/2^{42} \approx 10^{-11}$) of the possible protein combinations is observed in nature (Diepeveen, Gehrmann, Pourquié, Abeel, & Laan, 2018). The absence of the other combinations and the fact that polarity establishment is an essential process for proliferation in *Saccharomyces cerevisiae*, leads to the assumption that those protein networks are either non-functional or competitively disadvantageous. This leaves us to question why the observed networks function properly and why the non-observed networks do not. Here, we ask whether we can

distinguish the networks observed in nature from those that are not observed, based on their mathematical network properties. Are there differences in the global characteristics of the networks? Are there differences in the individual characteristics of proteins? Which role does Cdc42 play in the network? These questions will be addressed from a mathematical perspective, with global network and individual node characteristics as our toolbox.

Why do we focus on the mathematical characteristics of the yeast polarity network? Our project contributes to the eventual goal of predicting evolutionary trajectories of gene networks and specifically to the prediction of their outcomes. In other words, we try to make a prediction which combinations of proteins are effective and which are not. Most often, a biological view is used for answering this question. Our mathematical network perspective is a yet unattempted approach and could give us different insights.

2. Materials & Methods

2.1 Defining interactions

Based on the networks of 298 different fungal species, we consider the polarity module to be comprised of 42 different proteins. The selection of these proteins was based on their described functions in the polarity network and on their interactions with GTPase Cdc42 (Diepeveen, Gehrmann, Pourquié, Abeel, & Laan, 2018). Of those 42 proteins, 37 are present in *S. cerevisiae*. Two of the absent proteins (For3 & Scd1) are present only in the fungal species *Schizosaccharomyces pombe* and the other three (Rac1, SepA & Dia) only in *Ustilago maydis* (Table 2). To construct one complete network, we decided to translate the proteins that are absent in *S. cerevisiae* to the *S. cerevisiae* protein network with the underlying reasons that most of the proteins of interest are present in this species and that information on interactions and orthologs of this species is well-curated (Yu, et al., 2008).

Axl2	*Boi2*	*Gic1*	*Msb4*	*Rsr1*	*Ste20*
Bem1	*Bud3*	*Gic2*	*Nrp1*	*Rac1 (U.m.)*	*Swi4*
Bem2	*Cdc24*	*Iqg1*	*Ras2*	*Sec15*	*Scd1 (S.p.)*
Bem3	*Cdc42*	*Lte1*	*Rdi1*	*Sec3*	*SepA (U.m.)*
Bem4	*Cla4*	*Msb1*	*Rga1*	*Sec4*	*Dia (U.m.)*
Bni1	*Don1*	*Msb2*	*Rga2*	*Skm1*	*Tea1*
Boi1	*For3 (S.p.)*	*Msb3*	*Rho3*	*Spa2*	*Ubi4*

Table 2 – 42 proteins of interest in the yeast polarity network. Between brackets is the fungal species in which the protein is present: Schizosaccharomyces pombe (S.p.) or Ustilago maydis (U.m.). No species between brackets indicates presence in Saccharomyces cerevisiae.

The interactions of absent proteins were determined as follows. In the absence of direct orthologs in *S. cerevisiae* of these genes, the assumption was made that the absent gene would interact with the orthologs of its interactors (Zhong, et al., 2016) as is depicted in Figure 4. Pombase (Lock, et al., 2018) was used to find the physical interactors of For3 and Scd1 and to retrieve their orthologs. These were used as the interactors of For3 and Scd1. The interactors of Rac1, SepA and Dia were determined with STRING (Active interaction sources: experiments, databases. Minimum required interaction score: medium confidence) (Szklarczyk, et al., 2019). Their protein sequences were then blasted with NCBI BLAST (Protein blast. Database: nr. Organism: taxid 4932. Algorithm: blastp. Parameters: default. No threshold used) (Altschul, Gish, Miller, Myers, & Lipman, 1990) and their results used as interactors. The results of both methods are displayed in Table 3.

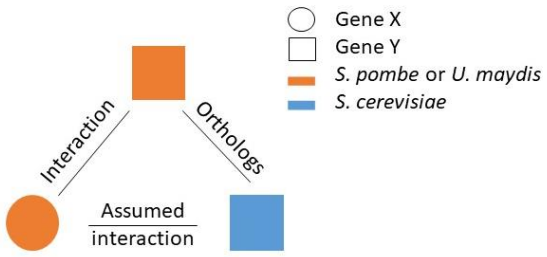


Figure 4 – Overview of the determination of the physical interactions between proteins absent from S. cerevisiae.

Absent proteins	Predicted interactors
For3	Boi1, Boi2, Cdc42, Rho3, Rac1, Sec3, SepA, Dia
Scd1	Bem1, Cdc42, Ras2, Rho3, Rac1
Rac1	Bem1, Bem2, Bem3, Bni1, Cdc24, Cdc42, Cla4, For3, Lte1, Ras2, Rdi1, Rga1, Rga2, Rho3, Rsr1, Sec4, Skm1, Ste20, Scd1, SepA, Dia
SepA	Cdc42, For3, Ras2, Rho3, Rsr1, Rac1, Sec4, Spa2

Dia	Cdc42, For3, Ras2, Rho3, Rsr1, Rac1, Sec4

Table 3 – Predicted interactions of proteins absent from S. cerevisiae with proteins present in S. cerevisiae.

The interactions between proteins can either be physical or genetic. A physical interaction occurs when two proteins bind to each other, whereas a genetic interaction indicates that the phenotype of a double mutant is not equal to the expectation based on both single mutants (e.g. synthetic lethals). As both types of interactions are defined differently and occur independently, they cannot be combined into one network. We chose to incorporate only the physical interactions, as the interpretation of the network becomes simpler. YeastMine (YeastMine, 2012) was used to find the interactions between *S. cerevisiae* proteins.

The network of physical interactions with both the present and absent proteins in *S. cerevisiae* is visualised with a graph in Figure 5. Interactions that occur between two of the same proteins (e.g. Cdc24 with Cdc24) were removed from the network as their resulting loops within nodes disturb the network analysis.

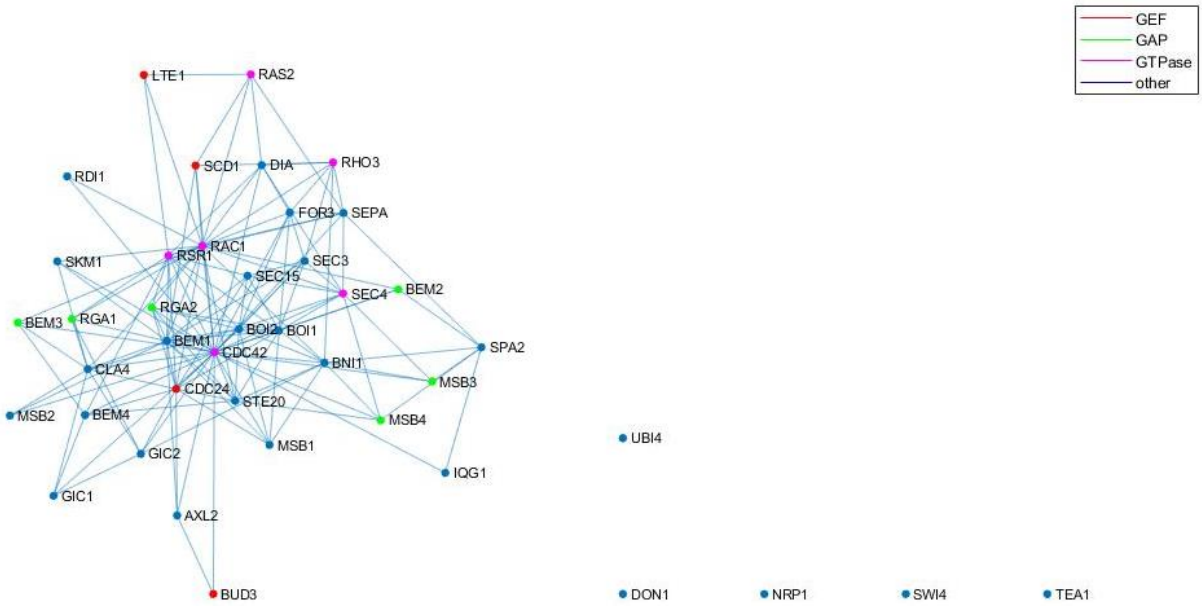


Figure 5 – Network of the 42 core proteins and their physical interactions. Cdc42 has a central position in this network and interacts physically with 31 of the 41 other proteins.

2.2 Building existing networks

We used three methods to build and examine the existing networks. As a first, we examined the 149 core networks containing only the 42 polarity proteins as described by Diepeveen et al. (2018). Those compositions observed in nature ranged in size from none to 36 of the 42 proteins present. Because observations of 0 or 1 proteins (both originating from the Microsporidia, which have a very compact genome due to their parasitic nature (Peyretaillade, et al., 2011)) cannot be regarded as networks, they were left out in further analyses, leaving in total 147 core networks. Secondly, we created a situation more representative of the in vivo network, by including proteins with a direct physical interaction with at least one of the 42 proteins. These first interactors were found with YeastMine. This resulted in an additional 953 proteins. Lastly, we zoomed in on the smaller networks, which are a subset of the core networks of maximum size 42. This way we could investigate whether the specific protein composition weighs more heavily in smaller networks than in regular sized networks. We determined the cut-off at a maximum of 23 proteins, as there is a large increase between networks containing 23 or 24 proteins (respectively 6 and 16 out of 147 networks, see Figure 6), leaving 41 small networks.

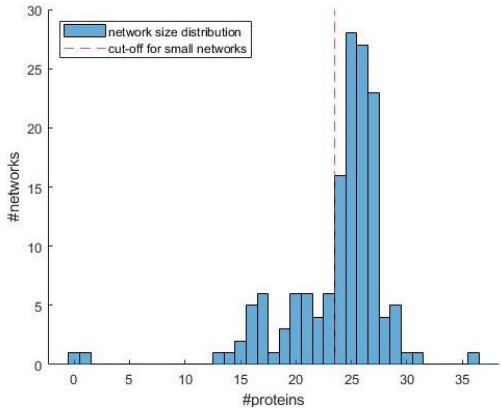


Figure 6 – Network size distribution and the indicated cut-off size for the small networks.

2.3 Building non-existing networks

To be able to compare the existing networks to non-existing networks, a subset of the non-existing networks was constructed, as there are too many non-existing networks ($\approx 10^{12}$) to analyse. One restriction we used is that the non-existing networks were required to have a size distribution representative of the existing networks. The second restriction was the probability distribution that was used to build the non-existing networks. We built non-existing networks with proteins randomly selected under different conditions:

- Using the probability distribution of all existing networks;
- Using the probability distribution of small existing networks (≤ 23 proteins);
- Using a uniform probability distribution;
- With exactly the same proteins as the existing networks, but without Cdc42.

We take the first condition as an example. The probability of occurrence of a protein is calculated among the 147 networks, which results in a number between 0 and 1. This number is then also used as probability of occurrence in non-existing networks. Using the same procedure with all proteins results in networks with the roughly same size distribution as the input networks. We built as many non-existing as existing networks, so either 147 or 41. The distributions of the first three ways of constructing non-existing networks are given in Figure 7.

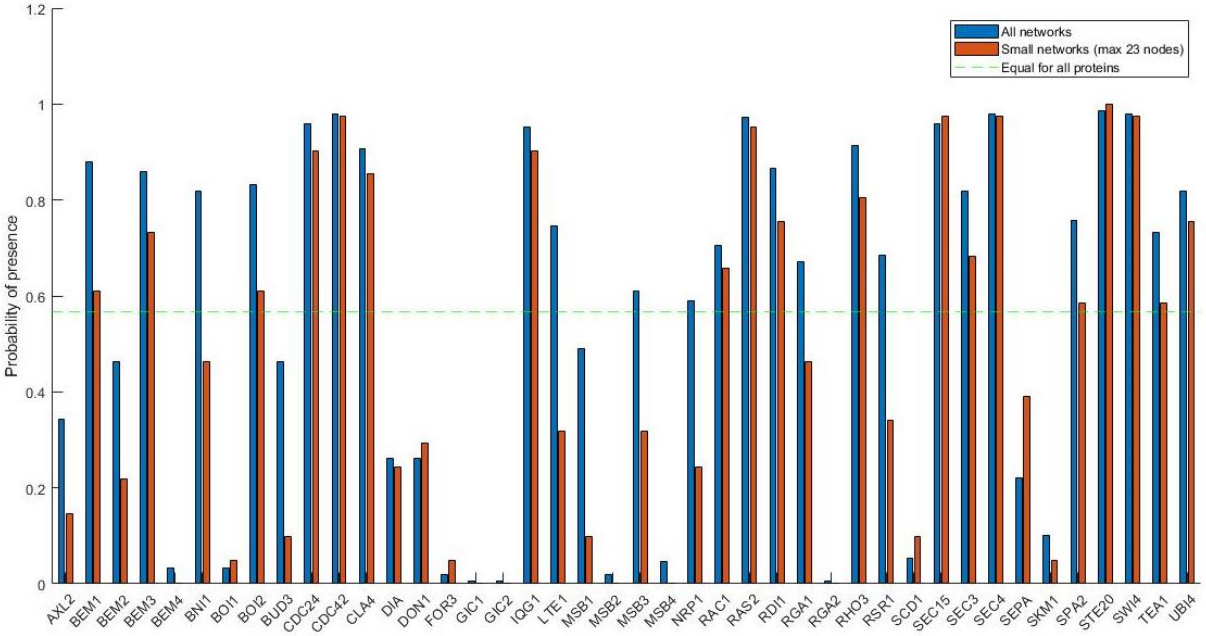


Figure 7 – Probability distributions of the presence of proteins measured in existing networks. These distributions were used as inputs for creating non-existing networks.

2.4 Determining mathematical characteristics

There are multiple ways of mathematically analysing a network. One can look at the characteristics of the network, or at characteristics of individual nodes within the network. Here we provide the definition of the network and node characteristics that were used in this study.

2.4.1 Global network characteristics

The size of a network can be described by the number of nodes and edges it contains. The ratio between those numbers sketches how well-connected the network is. We call this the network density and this property is calculated as

$$density = \frac{E}{\max E} = \frac{2 * E}{N * (N - 1)}$$

in which *E* presents the number of edges and *N* the number of nodes. It tells us how many edges there are with respect to the number of nodes, but it contains no information about how these edges are distributed over the nodes. For this we introduce the degree distribution. This distribution gives the probability *P* that a node has exactly *k* edges, or has a degree of *k*. Different classes of networks can be distinguished based on their distribution. For example, randomly generated networks (Figure 8, left) follow a Poisson distribution, as all nodes approximately have an equal number of edges. Most biological networks, however, are claimed to be scale-free networks (Barabási & Oltvai, 2004) due to their natural way of growing: nodes with more connections have a higher probability of obtaining more connections than nodes with few connection (Barabási & Albert, 1999). In this way, networks consisting of many nodes with few edges and a few so-called hubs with many edges (Figure 8, right) originate. This results in their degree distribution following a power-law

$$P(k) \sim k^{-\alpha}$$

with $2 < \alpha < 3$ typically for scale-free networks.

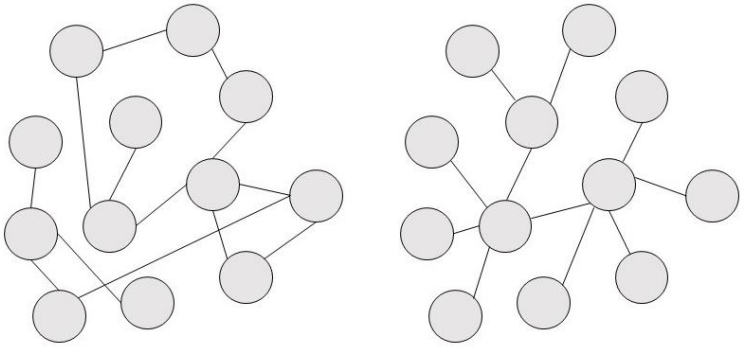


Figure 8 – A randomly distributed network (left) and a scale-free network (right).

Another perspective for looking at a network is its navigability, measured by the shortest path between two nodes in the network. The shortest path is the path connecting two nodes with the least nodes traversed.

2.4.2 Individual node characteristics

Within a network, different nodes have different positions, which creates a divergence in their importance. This importance can be described by various centralities, which are quantitative measures of the role of a node (Koutrouli, Karatzas, Paez-Espino, & Pavlopoulos, 2020).

A high betweenness centrality expresses that a node functions as a bridge between groups of nodes (Freeman, 1977). It is defined as the number of shortest paths between all nodes x and y that pass through node i, symbolised by $\sigma_{xy}(i)$.

$$C_{betweenness}(i) = \sigma_{xy}(i)$$

How directly nodes can communicate with other nodes is expressed by the closeness centrality (Sabidussi, 1966). This centrality is calculated as the inverse of the sum of the distance between node i and all other nodes.

$$C_{closeness}(i) = \frac{1}{\sum distance_{i,j}}$$

The most commonly used centrality is the degree centrality. It expresses how connected node i is by measuring the number of edges and thus connections to other nodes it has (Bonacich, 1987). Thereby it distinguishes nodes with many neighbours.

$$C_{degree}(i) = degree(i)$$

The eccentricity centrality tells us how accessible node i is from the rest of the network (Hage & Harary, 1995). It is the inverse of the maximum distance between node i and all other nodes. A high eccentricity centrality thus implies that a node is very accessible.

$$C_{eccentricity}(i) = \frac{1}{\max\left(distance_{i,j}\right)}$$

Another way of describing the importance of a node is with the clustering coefficient. It measures the extent to which it belongs to a highly connected community and it is defined as the number of edges between the neighbours of node i divided by the number of possible connections between those neighbours (Watts & Strogatz, 1998).

$$C_{clustering}(i) = \frac{2 * edges\ between\ neighbours(i)}{degree(i) * (degree(i) - 1)}$$

An example of the calculation of the various centralities and the clustering coefficients is shown in Figure 9 (left).

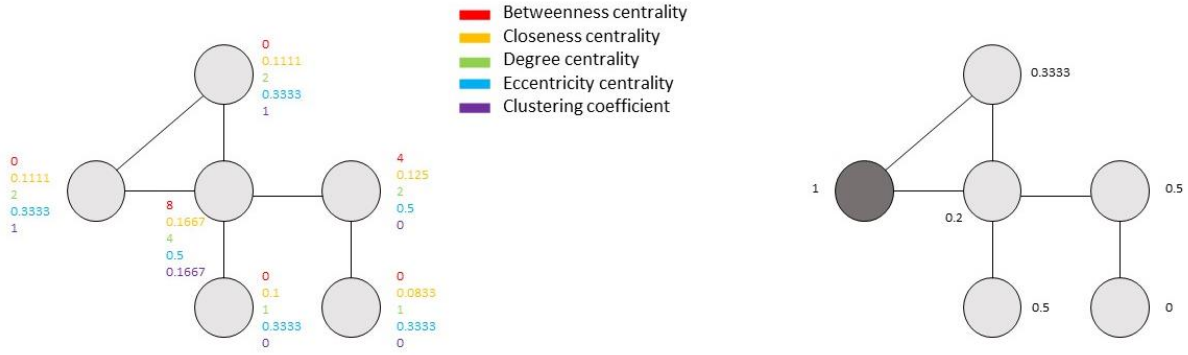


Figure 9 – Example of the calculation of the different centralities and the clustering coefficient (left) and the matching index from the perspective of the highlighted node (right).

Lastly, the matching index describes the functional similarity between two nodes based on the fraction of their neighbours that are shared between the two nodes (Figure 9, right).

$$M = \frac{\sum distinct\ common\ neighbours}{\sum total\ number\ of\ neighbours}$$

An overview of the used network characteristics can be found in Table 4. Most characteristics were calculated with a self-made MATLAB script. Exceptions were the (average) clustering coefficient (Tun, 2009) and the characteristics related to the power-law fit and thus the scale-freeness: alpha and the p-value (Clauset, Shalizi, & Newman, 2009). Alpha represents the scaling parameter of the probability distribution $P(k) \sim k^{-\alpha}$ and the p-value gives an estimation of how well this power-law fits the distribution. The adopted approach of Clauset et al. combines a maximum-likelihood fitting method to determine alpha and a goodness-of-fit test based on the Kolmogorov-Smirnov (KS) statistic to determine the p-value. A higher p-value indicates a more plausible fit to the power-law distribution.

Global network characteristics	Individual protein characteristics
Number of nodes	Betweenness centrality
Number of edges	Closeness centrality
Density	Degree centrality
Average degree	Eccentricity centrality
Average pathlength	Clustering coefficient
Average clustering coefficient	Matching index
Alpha	
P-value	

Table 4 – Characteristics used for network analysis. Alpha and the p-value are characteristics referring to the scale-freeness.

2.5 Determining biological characteristics

2.5.1 Morphological characteristics

Next to mathematical characteristics, biological properties of the proteins were also examined. The *Saccharomyces cerevisiae* Morphological Database 2 (SCMD2, 2018) contains structured information on the morphological properties of nearly 6000 mutants, with each mutant containing one gene deletion. Knock-outs of all 37 polarity proteins from *S. cerevisiae* are present, except for Bud3. The properties examined were the area of both the mother and the bud, the elliptical approximation of the mother cell, the bud to mother ratio and the no bud ratio. These were chosen to gain insight on

deviation in size, shape and bud-forming capacity, which are properties likely to be influenced by a non-functioning polarity network. All properties were normalised by dividing the data of the knock-outs by the average value of the corresponding wild-type data, which were different for the essential and non-essential genes.

2.5.2 Molecular functions

To gain insight in the biological roles of the proteins in the network, Gene Annotation (GO) terms of the nodes were extracted from YeastMine. GTPases, GAPs and GEFs were selected with the terms 'GTPase activity', 'GTPase activator activity' and 'guanyl-nucleotide exchange factor activity' respectively. The computationally annotated terms were left out.

3. Results

3.1 Differences in global network properties of existing and non-existing networks

First, we tried to distinguish existing from non-existing networks based on global network characteristics. We looked at their differences at three different levels: the polarity networks, the polarity networks extended with the first interactors and the polarity networks containing a small number of nodes (≤ 23 proteins). For the core networks we used the core and the uniform distribution, for the networks with first interactors we used the core distribution and for the small networks we used the small, the core and the uniform distribution to build non-existing networks.

3.1.1 Core networks

We started off with the core networks and its corresponding distribution and found that there was no clearly visible separation between the existing and non-existing networks based on the network parameters (full image in Appendix A, selection in Figure 10). This implies that the exact composition of the core networks does not influence the measured parameters. Notable is that there are no strong correlations between variables, except for the logical positive correlation between the number of nodes and the number of edges, and the number of edges and the average degree.

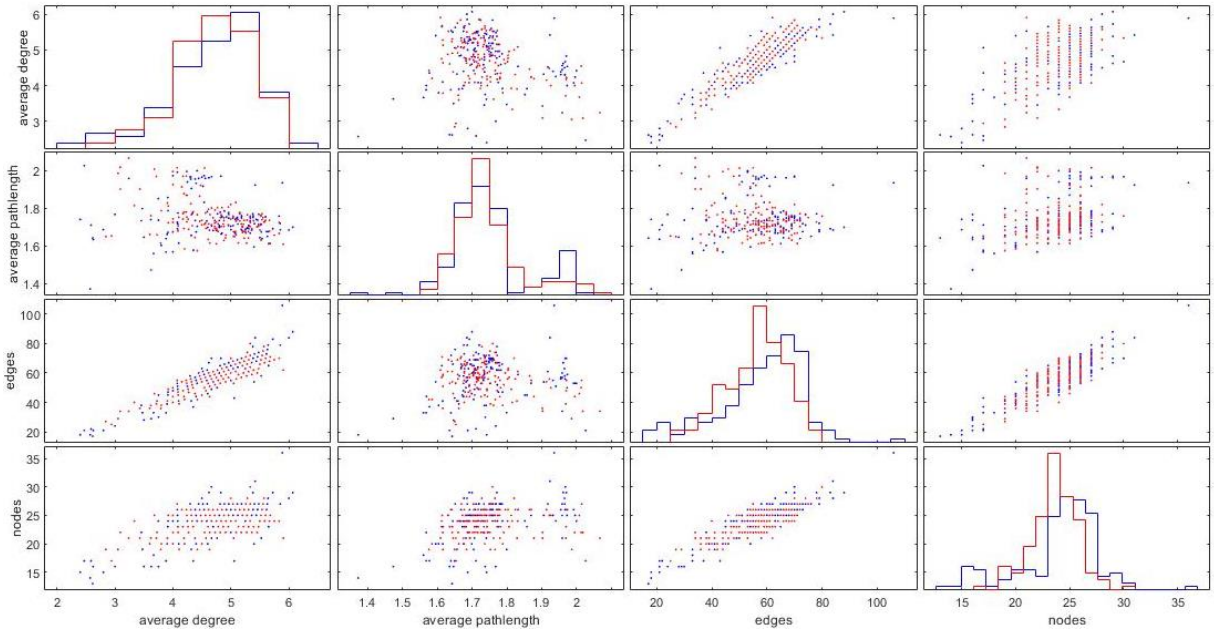


Figure 10 – Matrix scatter plot of the global network characteristics for existing (blue) and non-existing (red) core networks, with the core distribution used to build the non-existing networks. The diagonal shows the distribution of the variable of that row. The positive correlations between the number of nodes and edges and between the number of edges and the average degree are visible. Average pathlength and other (non-displayed) parameters show no correlation.

To see whether the used protein distribution would have an impact on the network characteristics, a uniform distribution was implemented to create the non-existing networks. This resulted in networks with a lower average clustering coefficient, average degree and density, a slightly lower number of edges and a higher average pathlength (full image in Appendix B, selection in Figure 11). Furthermore, the non-existing networks seemed to obey the power-law less certainly on average, which shows in a p-value that has been shifted to the left. Those visible distinctions imply that at least some of the non-existing networks deviate from existing networks and could thus be distinguished, and that those networks appear less stable than the existing ones.

Compared to the non-uniform distribution, a stronger correlation is observed between the clustering coefficient and other variables (average degree, density and pathlength). This is caused by the non-

existing networks occupying for example lower clustering coefficient and density values, resulting in a shift in their plot, while the values of existing networks naturally remain constant as their composition is fixed.

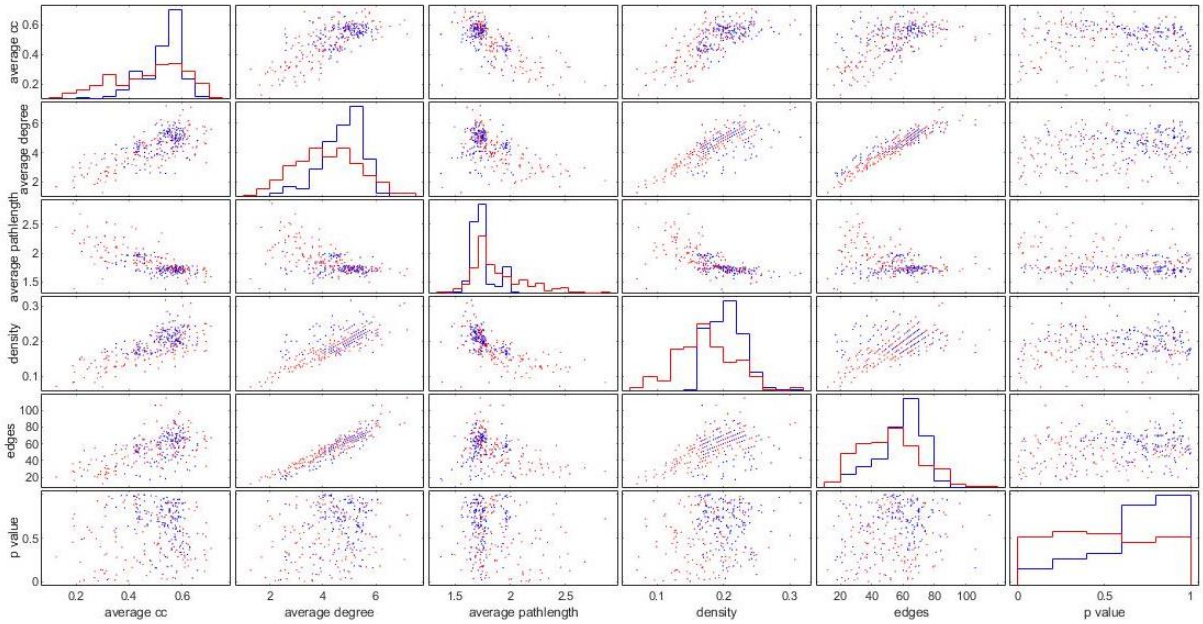


Figure 11 - Matrix scatter plot of the global network characteristics for existing (blue) and non-existing (red) core networks, with the uniform distribution used to build the non-existing networks. The diagonal shows the distribution of the variable of that row.

3.1.2 Networks with first interactors

The networks with all the physical first interactors of the core proteins were used to create a more realistic representation of the polarity network, as the core proteins also interact with numerous other proteins in vivo. This perspective makes the characteristics connected to the scale-freeness more interesting. Those parameters are more reliable for larger datasets, as the bias for different values of alpha decays faster than the statistical error of alpha, with n ≈ 50 (number of proteins in the networks) given as a lower bound for reliable estimates (Clauset, Shalizi, & Newman, 2009). Interestingly, the extended existing and extended non-existing networks do not show differences in either the estimator of the exponent, alpha, or the probability that the distribution follows a power-law, the p-value (full image in Appendix C, selection in Figure 12). In other characteristics they do not show clear differences either. However, the values of the parameters vary less (e.g. the clustering coefficient ranges from 0.332 to 0.340 instead of 0.2-0.6, as seen in the networks without the first interactors). This is as expected, because the extended networks are more similar than the core networks and the unvarying network of first interactors has a large influence. In this smaller range, correlations between different parameters (clustering coefficient, degree, pathlength and density) are still visible, as expected based on the results above.

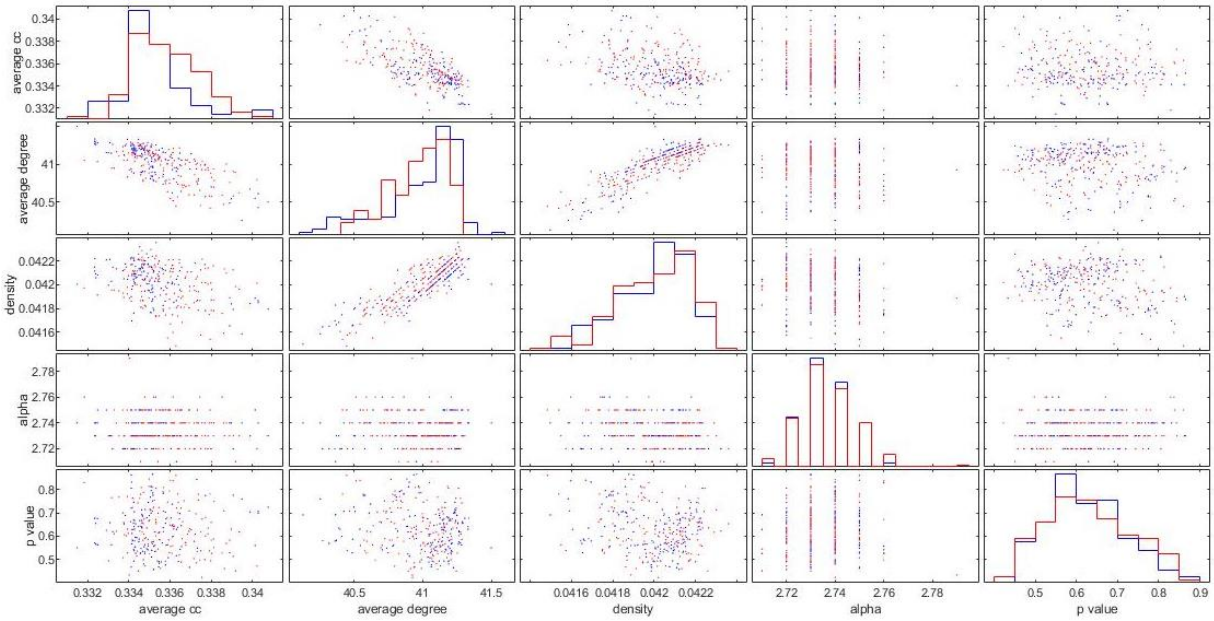


Figure 12 – Matrix scatter plot of the global network characteristics for existing (blue) and non-existing (red) networks with first interactors included, with the core distribution used to build the non-existing networks. The diagonal shows the distribution of the variable of that row.

3.1.3 Small networks

We looked at the small networks (≤ 23 proteins, see Figure 8) to see whether their composition would have more influence on the measured parameters than the composition of the larger core networks. This was based on the assumption that the choice of a single node has more influence amongst a smaller group of nodes with respect to a larger group.

When small networks were constructed with the same protein distribution as the existing small networks, not many differences could be found in their characteristics (Figure 13). These differences also did not appear when the boundary of 23 proteins was lowered even further (data not shown). It is interesting to see, however, that the average pathlength in existing networks is sharply peaked around 1.6, which was the lower pathlengths of all core networks.

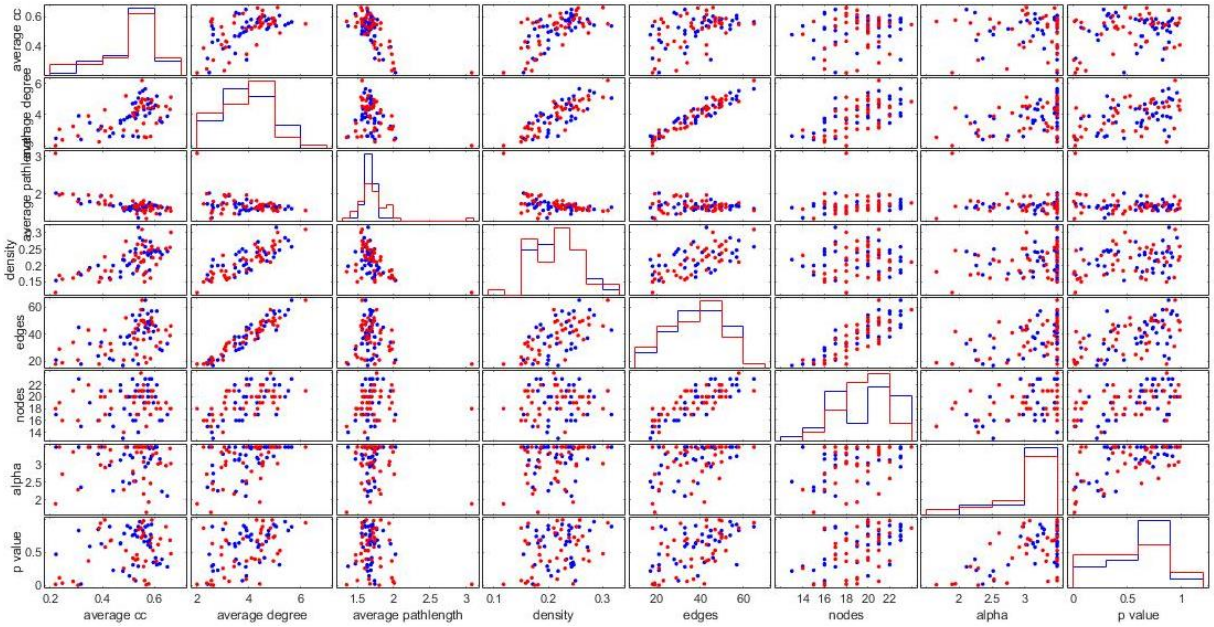


Figure 13 - Matrix scatter plot of the global network characteristics for existing (blue) and non-existing (red) small networks (≤ 23 proteins), with the small distribution used to build the non-existing networks. The diagonal shows the distribution of the variable of that row.

When the protein distributions of all core networks were used to build small networks, we found little more deviation in the average clustering coefficient, pathlength and density (Figure 14). These deviations were better visible when the uniform protein distribution was used (Figure 15). These observations substantiate the observations in the core networks, where the non-existing networks could be distinguished better when the used protein distribution deviated more from the distribution of the existing networks.

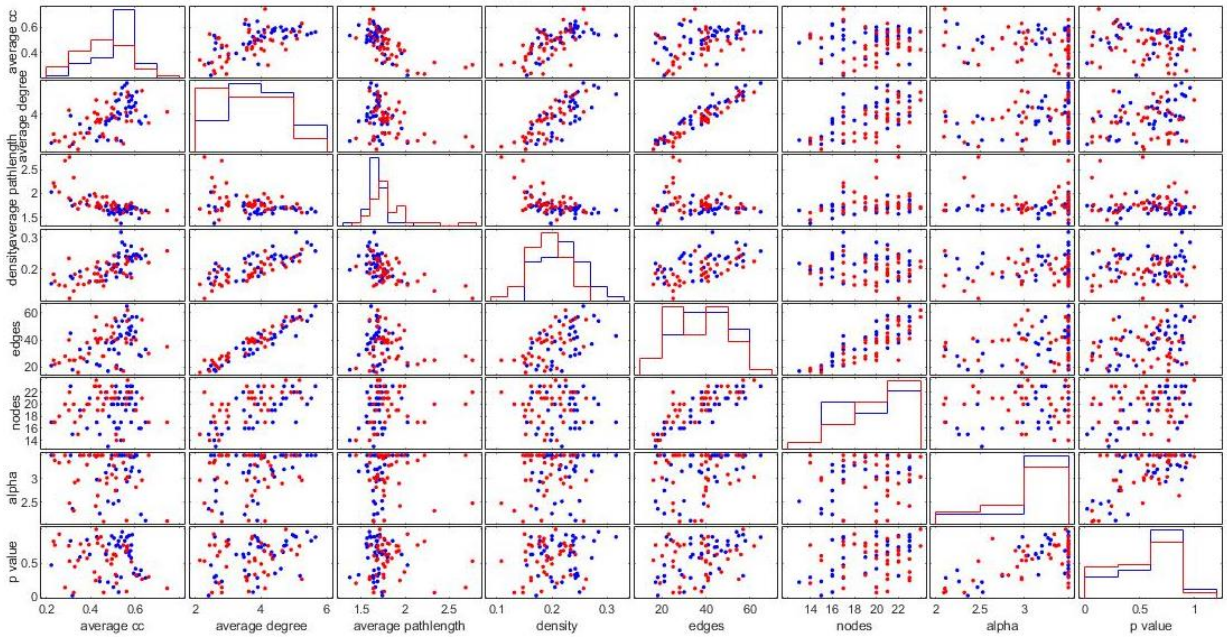


Figure 14 - Matrix scatter plot of the global network characteristics for existing (blue) and non-existing (red) small networks (≤ 23 proteins), with the core distribution used to build the non-existing networks. The diagonal shows the distribution of the variable of that row.

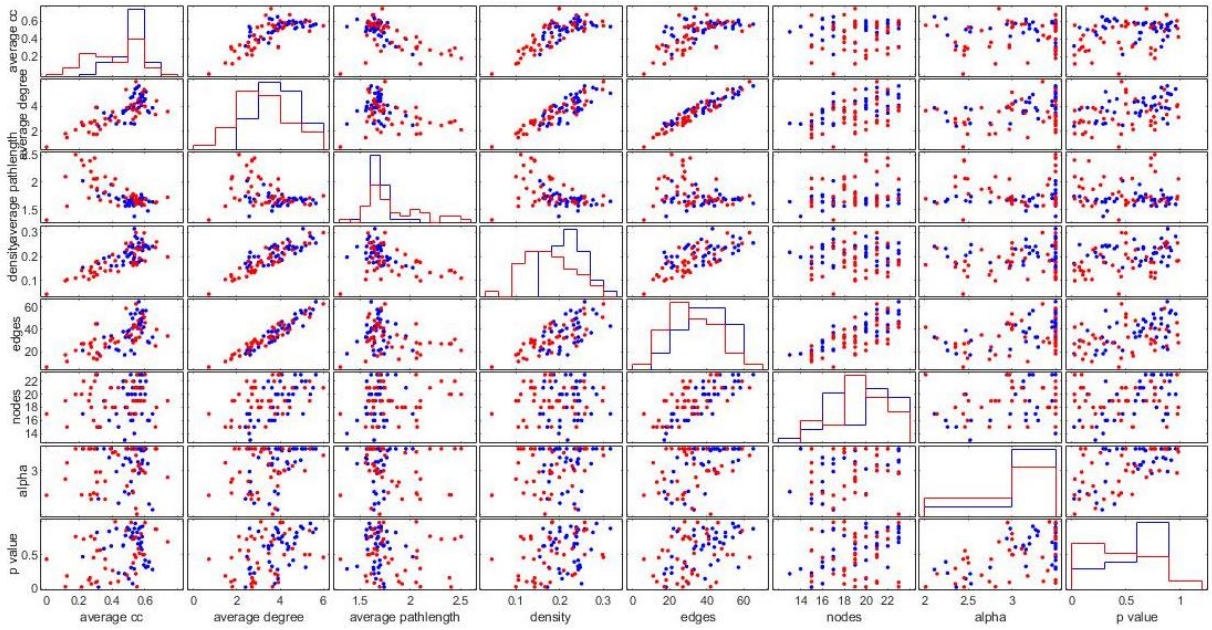


Figure 15 - Matrix scatter plot of the global network characteristics for existing (blue) and non-existing (red) small networks (≤ 23 proteins), with the uniform distribution used to build the non-existing networks. The diagonal shows the distribution of the variable of that row.

3.2 Mathematical and morphological characteristics of individual proteins

Secondly, we investigated the roles of individual proteins, to see whether we could find patterns on this level. We looked at individual mathematical characteristics and morphological characteristics.

3.2.1 Mathematical characteristics

We made a scatterplot of the five node-specific networks measures (the betweenness, closeness, degree, eccentricity centrality and the clustering coefficient) per protein, averaged over the core networks (Figure 16). Overall, the highest centrality value and thus the most important role is occupied by two GTPases, Cdc42 and Rac1 (indicated in the top-left panel), with the clustering coefficient as an exception. In the first three parameters (betweenness, closeness and degree centrality), their values are segregated from the other proteins, indicating their importance. For the clustering coefficient, GAPs seem to score relatively high.

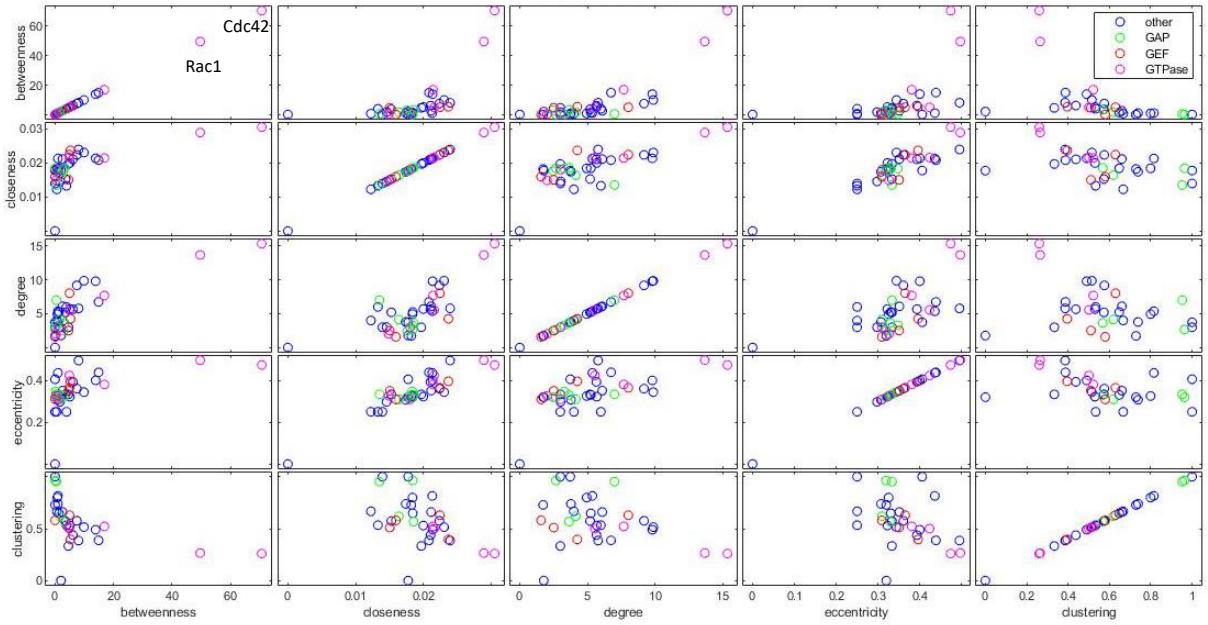


Figure 16 – Matrix scatter plot of average centralities for the core networks.

Taking only the small networks did not have a large influence on the outcomes (Figure 17), also when the maximum of 23 nodes was lowered (data not shown). Cdc42 and Rac1 seem to behave equally and the only difference is a shifted scale in betweenness and degree. Both are logical consequences of viewing only the smallest networks.

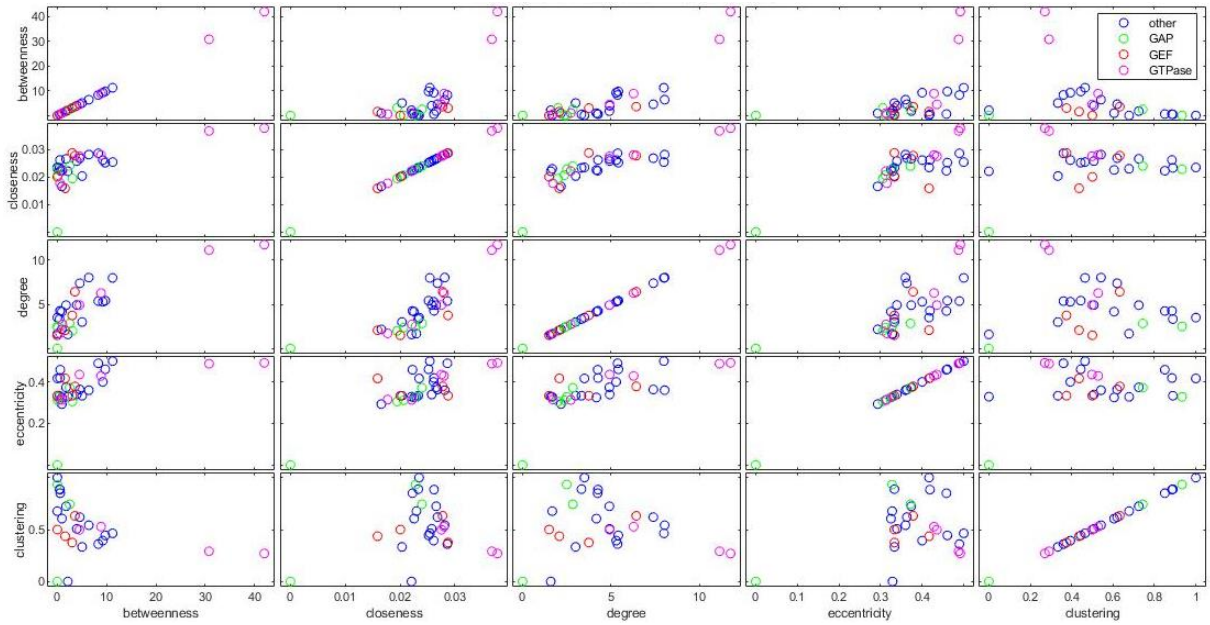


Figure 17 – Matrix scatter plot of average centralities for small networks (≤ 23 proteins).

3.2.2 Morphological characteristics

We calculated the influences of the deletion of all *Saccharomyces cerevisiae* genes (except Bud3) for five morphological parameters: the influence on the area of the mother cell, area of the bud, elliptical approximation of the mother cell, bud to mother ratio and no bud ratio. We plotted these values of influence against the mathematical characteristics (Figure 18). These plots were also constructed with the absolute values of the morphological influence (data not shown). We found that there were no correlations between the morphological and mathematical properties and neither between the

morphological properties and the molecular functions. This implies that the mathematical importance of a node cannot be traced back to its morphological importance and vice versa.

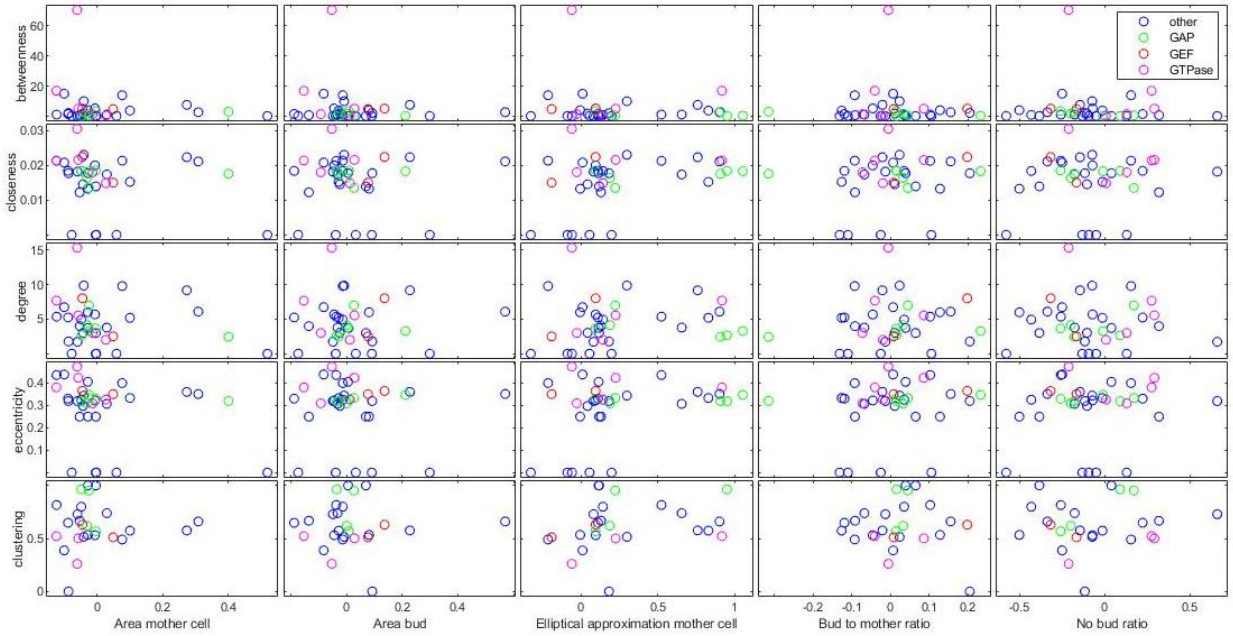


Figure 18 – Matrix scatter plot of centralities (y-axis) against morphological properties (x-axis). A positive value for the morphological properties indicates an increase of this parameter and a negative value a decrease. A larger score for order of elliptical approximation of the mother cell means more deviation from an ellipse.

3.3 Importance of the role of Cdc42

That Cdc42 is an important link in the polarity establishment chain is well-known (Johnson, 1999) and is emphasised by the fact that the protein is present in 146 of the 147 networks. Here we try to answer the question how Cdc42 contributes to our networks mathematically.

3.3.1 Natural network without Cdc42

The graph of the only network without Cdc42 is shown in Figure 19. It is interesting to note that this network consists of only 14 proteins, that GAPs Bem2 and Rga1 have no physical interactions within the network and that the alternative GTPases Rho3 and Sec4 do not replace the central position of Cdc42 in the network, but that this position is taken by Boi2.

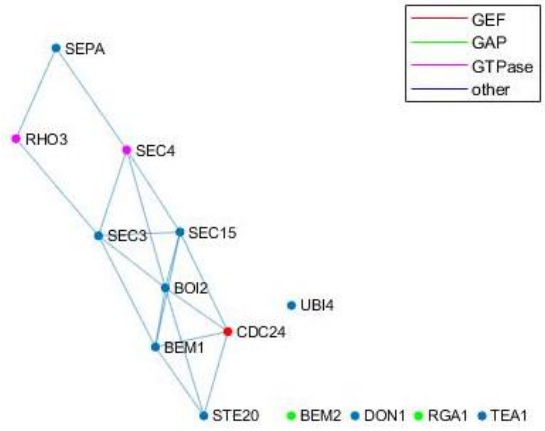


Figure 19 – Only existing network in which Cdc42 is absent.

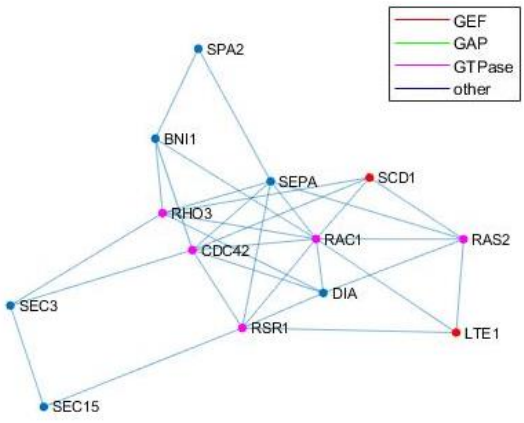


Figure 20 – Network with all proteins that do not interact with Cdc42 and proteins that interact at least twice with those proteins. Cdc42 still plays a central role.

3.3.2 Removal of central position

The biological importance of Cdc42 is reflected by the importance of its node in the network. A network with Cdc42 playing a more peripheral role can barely be constructed, as Cdc42 physically interacts with 31 out of the 36 other interacting proteins, where Lte1, Rho3, Sec15 and Spa2 are the exceptions. In order to build such a network, all these proteins were included, in combination with proteins that have at least two interactions with this group (Bni1, Rsr1, Rac1, Sec3, Scd1, SepA and Dia). The resulting network consists of 13 proteins (as large as the smallest existing network) and is shown in Figure 20. Cdc42 still plays a central role and is the second most connected node in the network. This proves that when Cdc42 is present in a network, it automatically claims an important role.

3.3.3 Removal of Cdc42

Another way of showing the importance of Cdc42 is to remove it completely from the networks containing the protein (full image in Appendix D, selection in Figure 21). This has a large impact on the average clustering coefficient, degree, pathlength, density and the number of edges of the networks, of which the clearest segregation can be seen in the average pathlength. It is interesting to note that the p-value does not decrease substantially after the removal of such an important node. But overall, the removal of Cdc42 has the most noticeable influence on the network parameters.

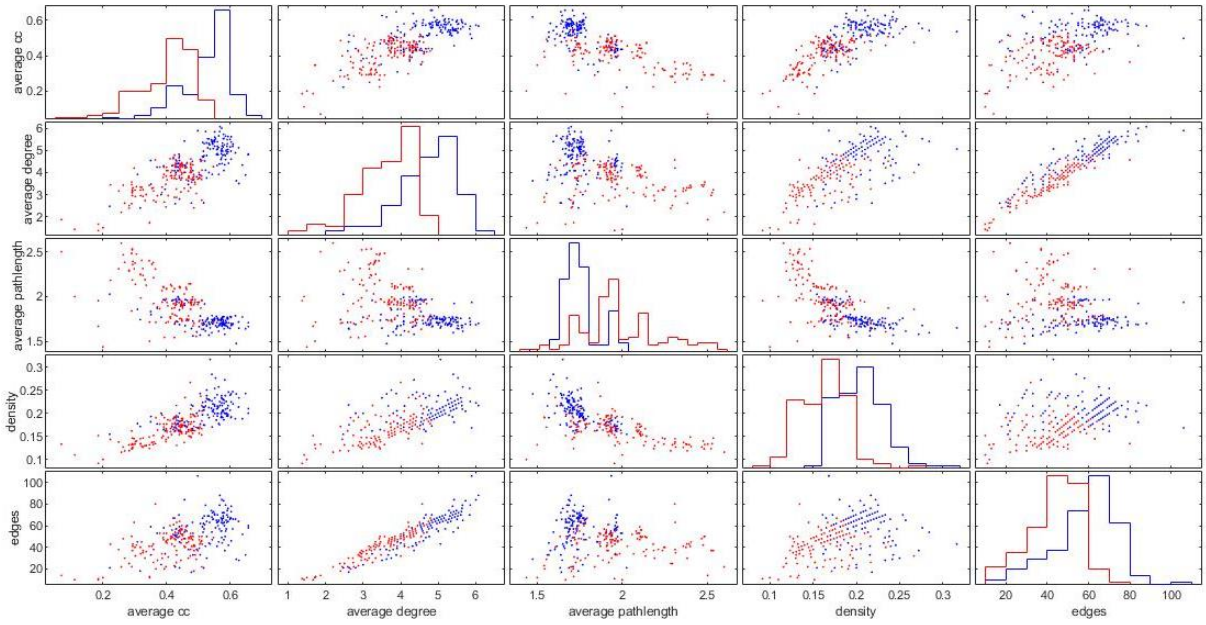


Figure 21 – Matrix scatter plot of the global network characteristics for core networks with Cdc42 (blue) and without Cdc42 (red). The diagonal shows the distribution of the variable of that row.

As shown before in Figure 16, Cdc42 is the most important protein in the network measured by the various centralities, so removing this protein logically affects all individual parameters (Figure 22). This is explicitly seen in the overall increase in betweenness and decrease in closeness, degree and eccentricity centrality. The reason that the betweenness centrality is the only parameter that increases, is because it depends on how many of the shortest paths from one node to another cross the specific node. It is thus increased by the removal of an important node. Furthermore, no large effects can be seen in the clustering coefficient, which can be explained by Cdc42 not playing an important role for this characteristic.

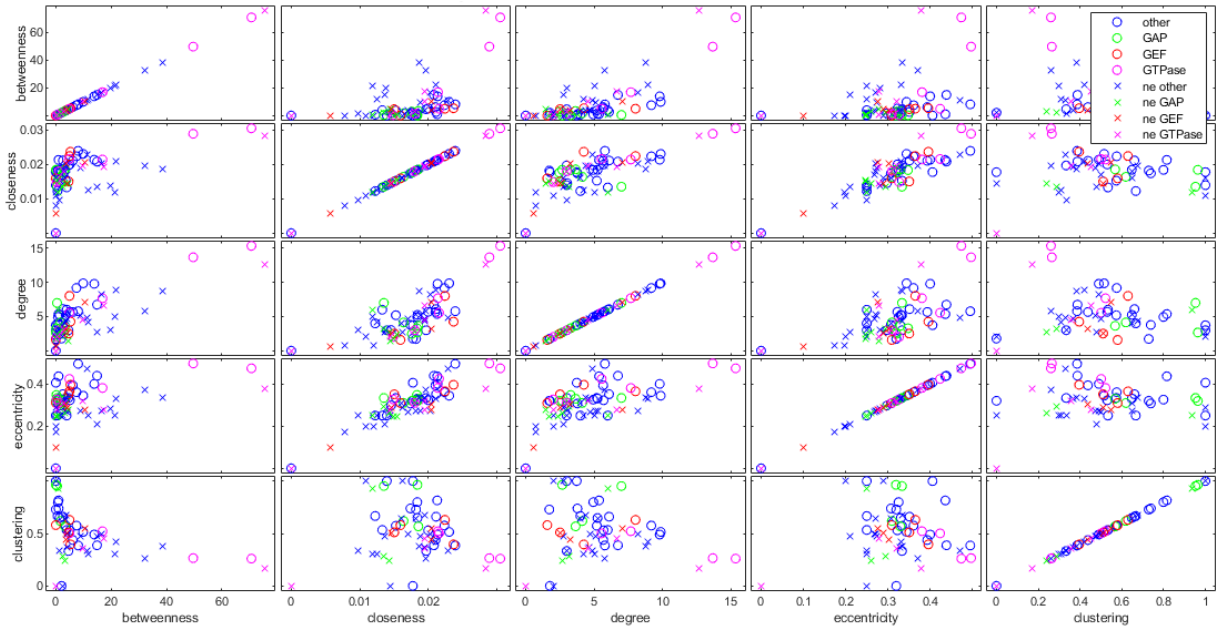


Figure 22 – Matrix scatter plot of average centralities for networks with (o) and without Cdc42 (x).

3.3.4 Proteins similar to Cdc42

The matching index allows us to compare the functionality of different nodes with each other. We identified which proteins have the most comparable network role to Cdc42 by calculating the matching index for all proteins in all existing networks. In almost two thirds of the networks, GTPase Rac1 had the highest matching index and in nearly one thirds it was Boi2. They had average matching indices of 0.5230 and 0.4437 respectively. Rac1 was absent whereas Boi2 was present in the only network where Cdc42 was absent, in which Boi2 was the most-connected node.

4. Discussion

Our main question was whether we could distinguish existing from non-existing physical interaction networks in *Saccharomyces cerevisiae* based on their mathematical properties. We divided this main question into the following questions: What are the differences between the existing and non-existing networks? Are there differences in the individual characteristics of proteins? And which role does Cdc42 play in the polarity network?

4.1 Differences in global network properties of existing and non-existing networks

The protein distribution that is used as an input to construct non-existing networks influences the measured network parameters. We see that when the used distribution deviates more from the actual protein distribution, e.g. when all proteins have an equal chance of being present in the network, the measured parameters of the non-existing networks deviate more. This difference is better visible in smaller networks, as changes in composition have a larger effect in those networks. This dependence on distribution is caused by the fact that all proteins function differently, which is in this case shown mathematically, but which also holds biologically. Different proteins perform different functions within a network and can be missed or replaced with a different capacity. A different distribution could thus create networks with less bounded values of our parameters.

The global network parameters affected by changing the protein distribution were the average clustering coefficient, average degree, average pathlength and density, with the most visible increase in average pathlength. Despite not seeing an exact separation, we can determine their boundaries (average clustering coefficient: 0.22-0.66, average degree: 2.40-6.07, average pathlength: 1.37-2.03, density: 0.15-0.32) and deduct which networks are non-existing when those boundaries are crossed. These networks resemble existing networks less.

4.2 Mathematical and morphological characteristics of individual proteins

When analysing the individual properties, we found that Cdc42 had the highest value of most characteristics (all centralities, but not the clustering coefficient). This was conserved in the non-existing networks and in the small networks. There was no clear distinction in importance of the different centralities, but we can conclude that the clustering coefficient did not function well as an estimator of the importance of a protein, based on the previous knowledge of the importance of Cdc42. Furthermore, we could not correlate any morphological parameter to a mathematical one, which shows the difficulty of uniting mathematical and biological knowledge of protein networks.

4.3 Importance of the role of Cdc42

Although the used distribution has an influence on the measured parameters, the largest impact was created by removing Cdc42 from the existing networks. We chose to remove this protein as it is, with one exception, an essential polarity protein. Networks without Cdc42 are thus not likely to function and interesting to analyse. The removal of Cdc42 increased the average pathlength as most of the proteins are interacting with Cdc42. This is not a surprising outcome, as the polarity proteins were partially selected based on their physical and genetic interactions with Cdc42. However, removing Cdc42 does not affect the scale-freeness of the networks. This, together with the observation that the power-law is also followed in non-existing networks, shows that the scale-freeness cannot be used as a predictor of existence.

GTPase Cdc42 has a very important role as initiator of the polarity establishment reaction. Rac1 has a similar role to Cdc42, which can be concluded from both the centrality plots and the matching index. This similarity could originate from Rac1 also being a member of the Rho GTPase family (Bustelo,

Sauzeau, & Berenjeno, 2007). The explanation of the relatively high matching index of Boi2 remains unclear, as there is no functional overlap between this protein and Cdc42.

4.4 Recommendations

The goal to predict whether a polarity establishment network is functional with machine learning is not achieved yet, but steps have been made in the right direction. We would like to make a few recommendations for future research in order to approach this goal.

First of all, we focused on the physical interactions and disregarded the genetic interactions in order to give all the edges in the network an equal definition. The network of genetic interactions contains a different type of information and this is where an opportunity rises. It tells us what influence a deletion of one protein has on all other proteins. Possibly, changes in the genetic network are more meaningful and could be a better predictor of existing networks.

Secondly, the prediction of the interactions of non-*Saccharomyces cerevisiae* proteins could be improved. Our approach did not take different domains into account and blasted the whole protein sequences instead of the specific interaction domains. This possibly resulted in a difference between the real and predicted interactions. We would recommend using the specific interaction domains for the blast to predict those interactions more accurately and to confirm the importance of Rac1.

Next to this, there are other interesting biological parameters to investigate. An example is the localisation of the proteins, as a corresponding position in the cell is required for a physical interaction. Which are located in the nucleus and in the cytoplasm? Does this influence how we view our networks? Another could be the link between the existing networks and the lineages and species where they were found. Is there a match between their lifestyle and network? Why are some networks more prevalent than others? These are amongst the parameters that are still unanalysed, but maybe contain useful information to incorporate in a prediction model.

Lastly, the question rises how universal the conclusions are. Do other protein networks in *S. cerevisiae* behave equally? And how does this translate to polarity establishment networks in other species? Are there rules that are always obeyed? The first question can be answered by investigating e.g. the well-known metabolism network in yeast and the other two can be started by answering the question whether the same interactions occur in e.g. *S. pombe* or *C. elegans,* with both a well-characterised protein map, and eventually the human polarity network.

5. Acknowledgements

First and most importantly, I would like to thank Enzo for teaching me so much during the past months. You patiently provided me all the time, help and support I needed. I also want to thank Liedewij for making it possible to do my project in this amazing lab. I admire you for your endless enthusiasm and insightful feedback. And indeed, I was your Secret Santa. Furthermore, I want to thank all the lab members for all the tea breaks, lunches and most importantly the numerous table tennis matches. I surely have learnt a lot from this too. And lastly, special thanks to Ramon for your never-ending stories and your even more never-ending chocolate supply. It was a pleasure to share an office with you, but I still want my marbles back.

6. References

Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990). *Basic Local Alignment Search Tool*. Retrieved October 2020, from NCBI: https://blast.ncbi.nlm.nih.gov/

Antunez, C. J. (2019, August 16). *Towards a GTPase/GEF assay. Background.* Retrieved from Open lab notebooks: https://openlabnotebooks.org/towards-a-gtpase-gef-assay/

Barabási, A., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 509-512.

Barabási, A., & Oltvai, Z. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 101-113.

Bonacich, P. (1987). Power and centrality: a family of measures. *American Journal of Sociology*, 1170-1182.

Bustelo, X., Sauzeau, V., & Berenjeno, I. (2007). GTP-binding proteins of the Rho/Rac family: regulation, effectors and functions in vivo. *BioEssays*, 356-370.

Cherfils, J., & Zeghouf, M. (2013). Regulation of Small GTPases by GEFs GAPs, and GDIs. *Physiological Reviews*, 269-309.

Clauset, A., Shalizi, C., & Newman, M. (2009). Power-law Distributions in Emperical Data. *SIAM Review*, 661-703.

Diepeveen, E., Gehrmann, T., Pourquié, V., Abeel, T., & Laan, L. (2018). Patterns of Conservation and Diversification in the Fungal Polarization Network. *Genome Biology and Evolution*, 1765-1782.

Freeman, L. (1977). A set of Measures of Centrality based on Betweenness. *Sociometry*, 35-41.

Hage, P., & Harary, F. (1995). Eccentricity and Centrality in Networks. *Social Networks*, 57-63.

Irazoqui, J., & Lew, D. (2004). Polarity Establishment in Yeast. *Journal of Cell Science*, 2169-2171.

Johnson, D. (1999). An essential Rho-type GTPase controlling eukaryotic cell polarity. *Microbiology and Molelucal Biology Reviews*, 54-105.

Koutrouli, M., Karatzas, E., Paez-Espino, D., & Pavlopoulos, G. (2020). A Guide to Conquer the Biological Network Era Using Graph Theory. *Frontiers in Bioengineering and Biotechnology*.

Lock, A., Rutherford, K., Harris, M., Hayles, J., Oliver, S., Bähler, J., & Wood, V. (2018). PomBase 2018: user-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Research*. Retrieved October 2020, from pombase.org

NCBI. (2021, January). Retrieved from National Center for Biotechnology Information: www.ncbi.nlm.nih.gov/gene/998

Peyretaillade, E., Alaoui, E. E., Diogon, M., Polonais, V., Parisot, N., Biron, D., . . . Delbac, F. (2011). Extreme reduction and compaction of microsporidian genomes. *Research in Microbiology*, 598-606.

Sabidussi, G. (1966). The Centrality Index of a Graph. *Psychometrika*, 581-603.

Schoenrock, A., Burnside, D., Moteshareie, H., Pitre, S., Hooshyar, M., Green, J., . . . Wong, A. (2017). Evolution of protein-protein interaction networks in yeast. *PLOS One*.

SCMD2. (2018). Retrieved October 2020, from Saccharomyces Cerevisiae Morphological Database 2: yeast.ib.k.u-tokyo.ac.jp/SCMD/datasheet.php

SGD. (2012). Retrieved October 2020, from Saccharomyces Genome Database: yeastgenome.org

Szklarczyk, D., Gable, A., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., . . . Mering, C. v. (2019, January). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 607-613. Retrieved October 2020, from STRING: string-db.org

Tun, K. (2009, June 6). *Graph package.* Retrieved October 2020, from MATLAB Central File Exchange: mathworks.com/matlabcentral/fileexchange/12648-graph-package

Watts, D., & Strogatz, S. (1998). Collective Dynamics of 'Small-world' Networks. *Nature*, 440-442.

Wodarz, A., & Näthke, I. (2007). Cell polarity in development and cancer. *Nature Cell Biology*, 1016-1024.

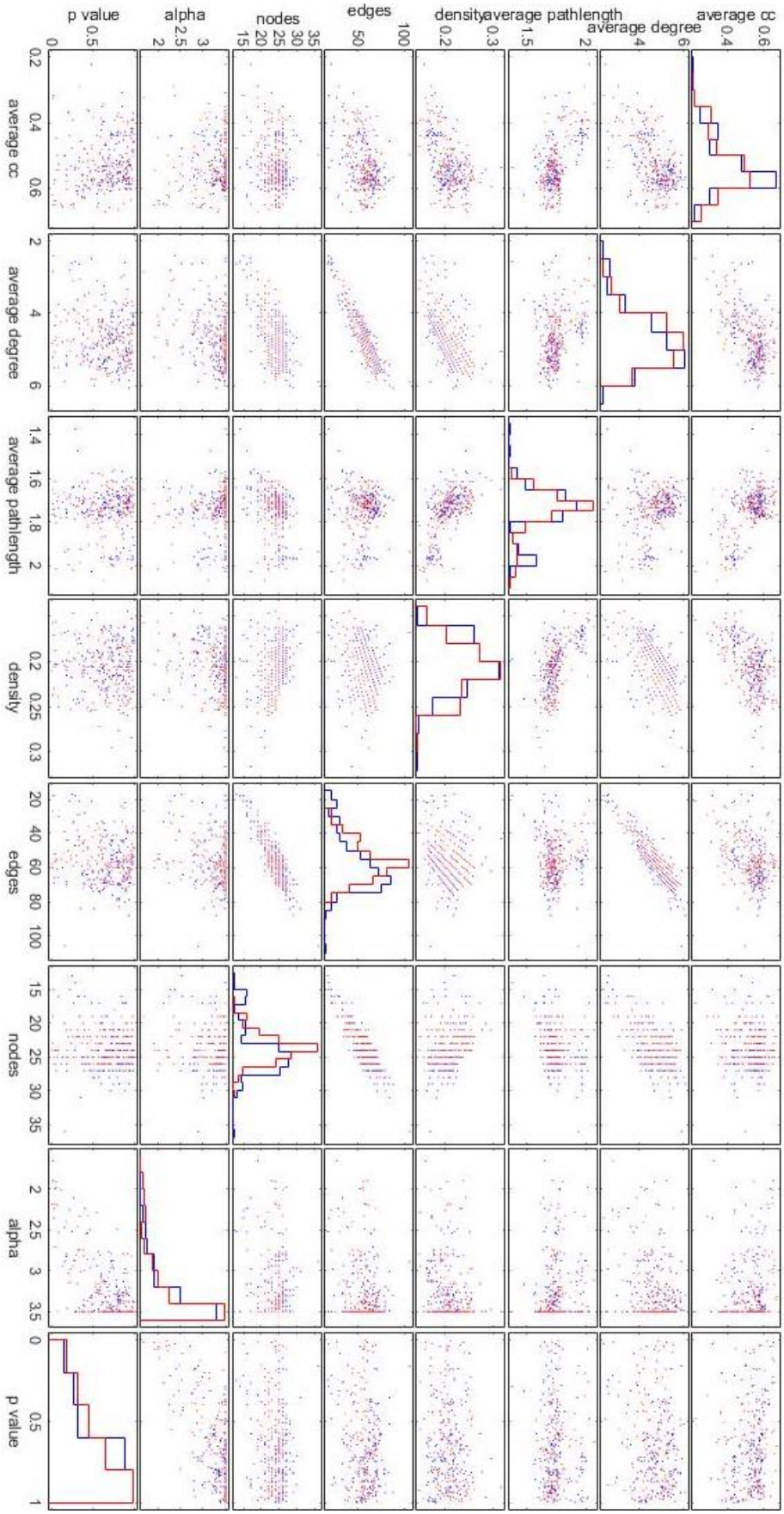
YeastMine. (2012). Retrieved October 2020, from yeastmine.yeastgenome.org

Yu, H., Braun, P., Yıldırım, M., Lemmens, I., Venkatesan, K., Sahalie, J., . . . Vidal, M. (2008). High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science*, 104-110.

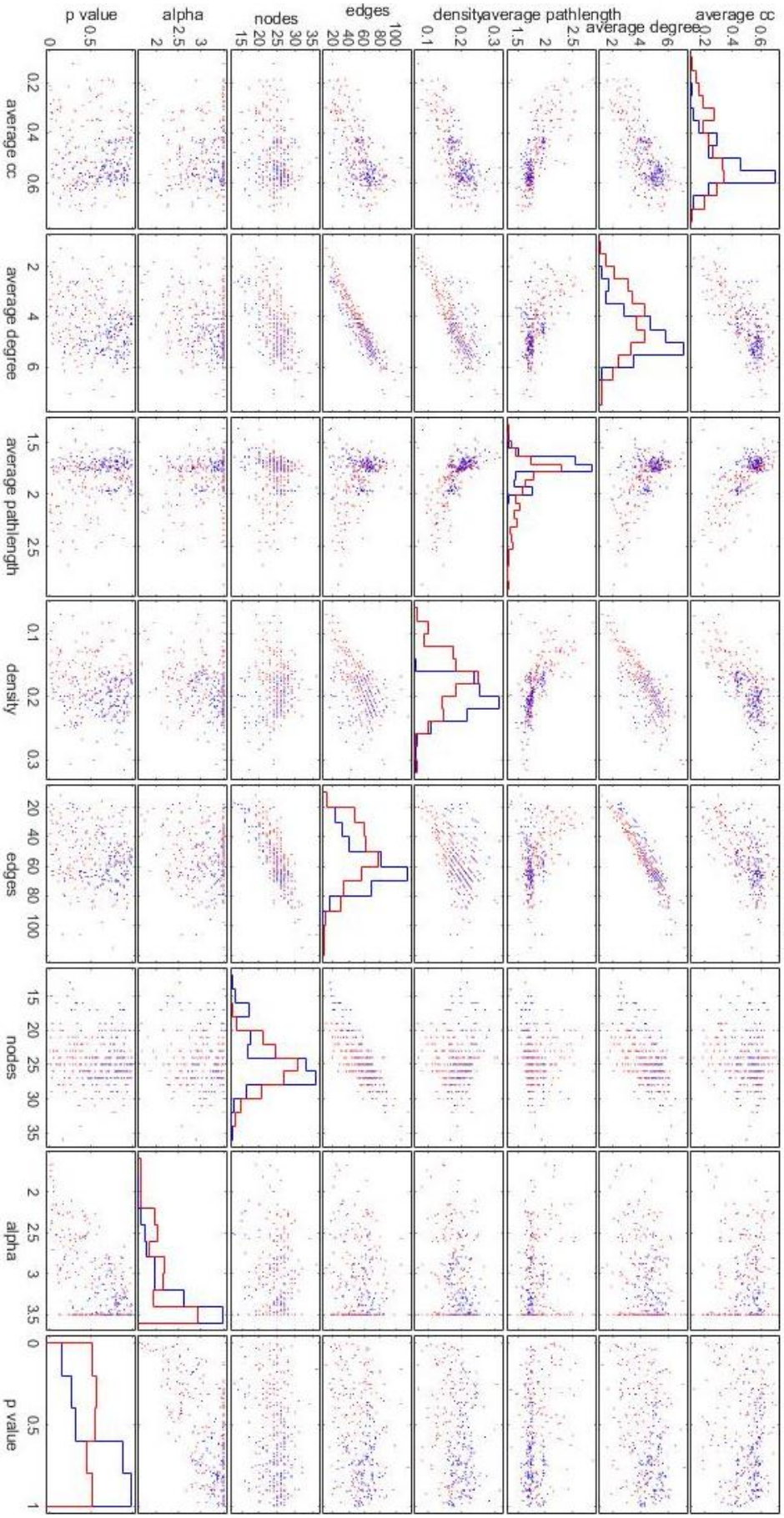
Zhong, Q., Pevzner, S., Hao, T., Wang, Y., Mosca, R., Menche, J., . . . Vidal, M. (2016). An Inter-Species Protein-Protein Interaction Nertwork across vast Evolutionairy Distance. *Molecular Systems Biology*, 865.

Appendices

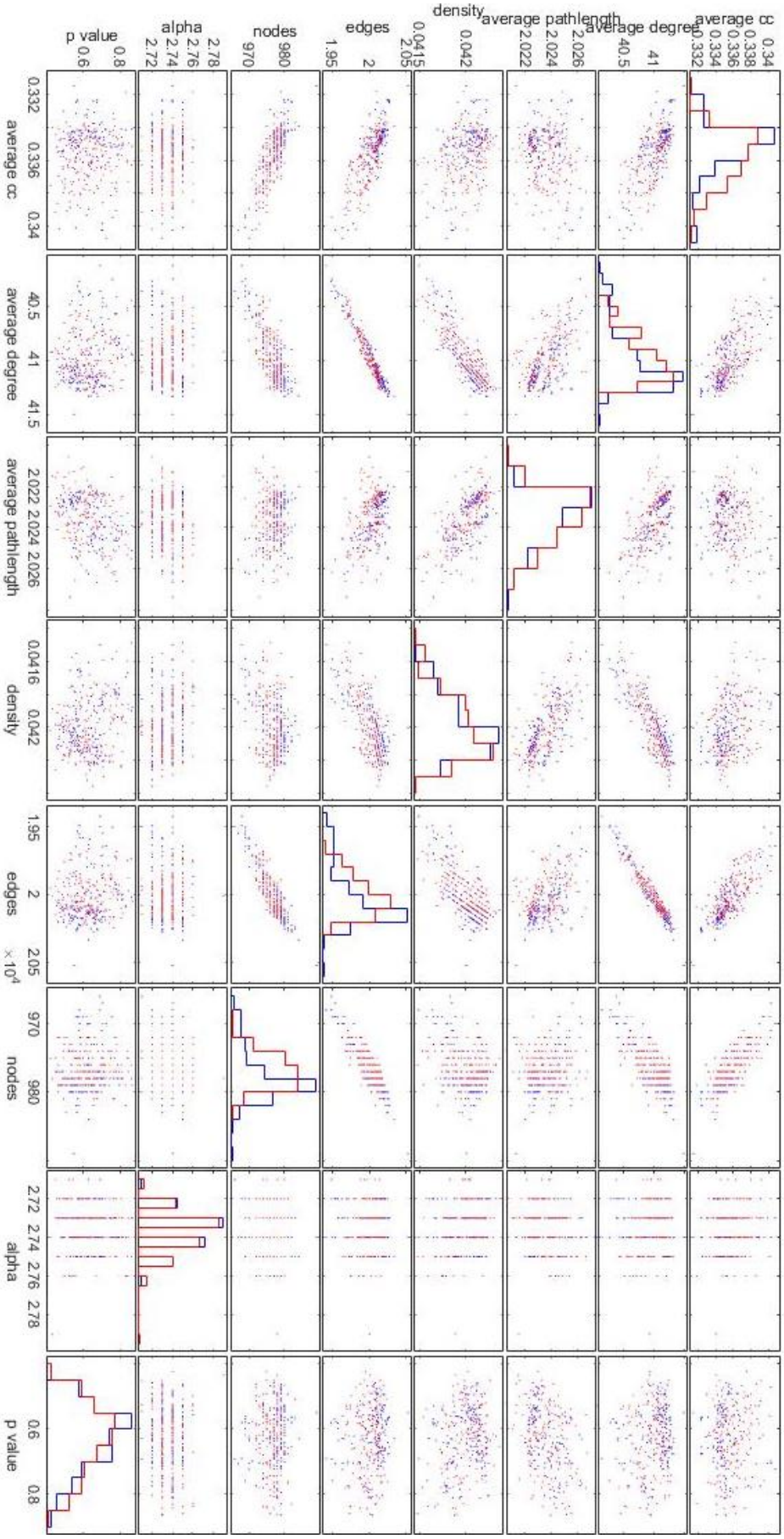
A. Matrix scatter plot of core networks with a core distribution



B. Matrix scatter plot of core networks with a uniform distribution



C. Matrix scatter plot of extended networks with a core distribution



D. Matrix scatter plot of networks with and without Cdc42