UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

José Ricardo Furlan Ronqui

Estudo e comparação da topologia de redes de interação de proteínas

São Carlos

2018

José Ricardo Furlan Ronqui

Estudo e comparação da topologia de redes de interação de proteínas

Tese apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Doutor em Ciências.

Área de concentração: Física Aplicada
Opção: Física Computacional

Orientador: Prof. Dr. Gonzalo Travieso

Versão corrigida
(Versão original disponível na Unidade que aloja o Programa)

São Carlos
2018

Ronqui, José Ricardo Furlan
Estudo e comparação da topologia de redes de interação de proteínas / José Ricardo Furlan Ronqui; orientador Gonzalo Travieso - versão corrigida -- São Carlos, 2018.
107 p.

Tese (Doutorado - Programa de Pós-Graduação em Física Aplicada Computacional) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2018.

1. Redes complexas. 2. Redes de interação entre proteínas. 3. Estrutura topológica de redes. I. Travieso, Gonzalo, orient. II. Título.

AGRADECIMENTOS

Aos meus pais Célia e Dari por sempre terem me apoiado e incentivado ao longo de toda essa jornada.

Ao professor Gonzalo Travieso por ter me orientado e aconselhado ao longo do desenvolvimento de toda a pesquisa e também durante os anos de graduação.

Aos meus amigos no Instituto de Física de São Carlos e de São Carlos por todos os cafés, conversas e apoio ao longo desses anos; certamente a jornada foi muito mais leve com a presença de vocês.

A todos os professores que tive até agora, pelo incentivo e dedicação de vocês.

Ao Instituto de Física de São Carlos e à Universidade de São Paulo por terem me permitido realizar minha formação como seu aluno.

A CAPES por ter financiado a pesquisa aqui realizada.

Finalmente, agradeço às comunidades do GNU/Linux, Python, NetworkX, Scikit-learn e Matplotlib pois graças aos seus projetos open-source muitas tarefas desenvolvidas foram facilitadas e aceleradas.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

"All we have to decide is what to do with the time that is given us."
Gandalf para Frodo em The Lord of the Rings: The Fellowship of the Ring
de J. R. R. Tolkien

RESUMO

RONQUI, J. R. F. **Estudo e comparação da topologia de redes de interação de proteínas**. 2018. 107p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2018.

Redes complexas são utilizadas para representar sistemas complexos, compostos de elementos que interagem uns com os outros. Uma das grandes vantagens de se empregar as redes é a possibilidade de se estudar a topologia presente nos mais diversos sistemas para obtermos informações sobre eles, entendê-los e compará-los. Devido à sua importância para a compreensão de processos intracelulares, desde início do desenvolvimento da área das redes complexas estudou-se a topologia da interação entre proteínas. Entretanto nos últimos anos com o desenvolvimento de novas técnicas de detecção o número de proteínas e interações reportadas cresceu de maneira muito acentuada; além disso, também existem alguns pontos sobre a sua topologia sobre os quais ainda não existe um consenso, como por exemplo qual a distribuição de graus desse tipo de rede. Neste trabalho estudamos as propriedades topológicas de redes de interação entre proteínas, utilizando as informações do banco de dados STRING, com ênfase no comportamento de suas medidas de centralidade e do espectro da matriz Laplaciana normalizada. Tanto a análise das medidas de centralidade e de suas correlações, quanto do espectro da matriz Laplaciana mostram que existem padrões topológicos que são conservados entre as redes dos organismos e que os mesmos também podem ser empregados para sua caracterização. Nossos resultados também mostram que as funções biológicas desempenhadas pelas proteínas podem ser identificadas pelas medidas de centralidade. Especificamente para a centralidade de autovetor, nossas análises indicam que ela está localizada nos maiores K-cores das redes consideradas. Os resultados aqui obtidos ressaltam que muitas informações relevantes podem ser extraídas da topologia das interações entre proteínas, além de indicarem a existência de possíveis estruturas conservadas; entretanto devido a incompletude dessas redes mais estudos precisam ser conduzidos para a avaliação de possíveis mudanças nos resultados aqui apresentados.

Palavras-chave: Redes complexas. Redes de interação entre proteínas. Estrutura topológica de redes.

ABSTRACT

RONQUI, J. R. F. **Topological studies of protein interaction networks**. 2018. 107p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2018.

Complex networks can be used to model complex systems, composed of main elements that interact with each other. The advantage of using this approach is the possibility to study the topology of a wide range of systems so that we can get more information, understand and compare them. Due to its importance on the understanding of the intracellular biological processes, since the early beginning of the development of the complex networks field protein-protein interaction topologies have been studied. However, new techniques for the detection of proteins and their interactions have been developed recently, which has significantly increased the availability and reliability of the corresponding data over the last few years; moreover, there still are some debate about the topology of protein-protein interaction networks such as the degree distribution of this type of network. Here we will study the topological properties of protein-protein interaction networks created using the information of the STRING database focusing on centrality measures of their nodes, the correlation between them, and the normalized Laplacian matrix spectrum. Our results show the existence of topological patterns conserved between the protein interaction networks of different organisms and that both the correlation of the centrality pairs and the spectrum of the Laplacian matrix can be used for network characterization. Another study indicates that the set of centrality measures of a protein can be used to identify clusters with well defined biological functions. A more detailed look at the eigenvector centrality behavior reveals that this measure is localized on the proteins of the highest k-cores for all networks. These results highlight the importance of the topology on the study of protein-protein interactions and that more studies can lead to a better a more complete understanding of such systems.

Keywords: Complex networks. Protein interaction networks. Network topology.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

B	Centralidade de interposição
BA	Barabási-Albert
C	Centralidade de proximidade
CFB	Centralidade de interposição utilizando o modelo de fluxo de corrente
CFC	Centralidade de proximidade utilizando o modelo de fluxo de corrente
D	Centralidade de grau
E	Centralidade de autovetor
ER	Erdős-Rényi
PCA	Análise de componentes principais
PPI	Interação proteína-proteína

SUMÁRIO

1 INTRODUÇÃO

Redes complexas[1–7] utilizam o conceito de grafos para representar sistemas do mundo real[8] que são compostos por partes que interagem entre si. Por volta dos anos 2000 uma série de estudos sobre estes sistemas mostraram que eles possuíam propriedades topológicas diferentes dos grafos estudados até então, e em muitos casos, algumas características que eram comuns a uma grande variedade de sistemas, como por exemplo distribuições de grau que seguem uma lei de potência.[9] Dessa forma, iniciaram-se os estudos na área que hoje chamamos de redes complexas: grafos que possuem uma topologia não trivial e que geralmente representam elementos do mundo real e suas interações, com uma grande quantidade de vértices — geralmente centenas ou milhares — e um número ainda maior de conexões e muitos avanços foram feitos tanto na área da topologia das redes complexas, — como por exemplo no desenvolvimento de medidas para sua caracterização[10] e entendimento da topologia das redes — quanto na parte de processos dinâmicos que ocorrem nessas redes — como por exemplo processos de sincronização[11] e propagação de epidemias,[12] — evidenciando a existência de uma relação íntima entre estrutura, função e processos que ocorrem nestes sistemas.

Devido à sua generalidade, redes complexas podem ser aplicadas a uma grande diversidade de sistemas, como por exemplo interações sociais,[13–15] redes de comunicação, páginas da Internet,[16] sistemas de transporte[17] e biológicos,[18–20] entre muitos outros. Dentro do área da biologia, existe um tipo de rede cujo o entendimento é particularmente importante — as redes de interação entre proteínas (PPI networks). Este tipo de sistema, é particularmente importante pois as proteínas são responsáveis por desempenhar uma grande quantidade de funções dentro dos organismos, como por exemplo: agindo como catalisadores de reações metabólicas, transcrição e replicação de DNA e transporte de moléculas. Além disso, dificilmente uma proteína desempenha sua função sem interagir com as outras proteínas,[21] o que torna muito importante levar-se em conta também as interações e com que elas ocorrem para um bom entendimento dos processos relacionados às redes PPI. Trabalhos recentes também destacam a importância de redes PPI para a descoberta de novos medicamentos[22] e no diagnóstico e desenvolvimento de estratégias terapêuticas de algumas doenças.[23–26]

Desde aproximadamente o início do desenvolvimento da área de redes complexas, estudos aplicando-as para entender a topologia da interação entre proteínas foram conduzidos[18,27–29]; entretanto a quantidade de dados sobre este tipo de rede vem crescendo de maneira exponencial[30] nos últimos anos, fazendo com que estudos sobre a topologia de redes PPI continuem sendo relevantes. Com o aumento da confiança sobre a existência das interações existentes e a inclusão de mais informações sobre novas proteínas e interações

às redes, algumas de suas propriedades topológicas ainda não estão muito claras. Apenas citando um exemplo, alguns artigos indicam que redes PPI possuem uma distribuição de graus livre de escala[18,31]; outros pesquisadores porém indicam que nem todas as redes possuem esta propriedade,[32] levantando a dúvida se organismos diferentes podem apresentar distribuições de graus distintas, ou se trata-se apenas de um problema de incompletude dos dados. Brader e Hogue também destacam[33] que dependendo da origem dos dados também podem existir diferenças com relação a conectividade observada para as redes PPI. Esses fatos demonstram que apesar da topologia dessas redes já ter sido previamente explorada na literatura, as inconsistências observadas justificam a continuidade de estudos nessa linha.

Além da caracterização da topologia das redes PPI, outro ponto que tem sido muito estudado é a comparação de redes de organismos distintos. Alguns trabalhos por exemplo buscam realizar um alinhamento de redes,[34] que consiste em mapear proteínas de redes distintas que estejam relacionadas de maneira funcional ou evolutiva e suas as interações entre as redes[35]; este tipo de comparação permite encontrar informações importantes como vias metabólicas conservadas ao longo do processo de evolução. Outros estudos[36–39] reportam que proteínas que possuem padrões similares de conexões tendem a apresentar funções similares entre organismos. Este tipo de análise permite a transferência de informação entre espécies como realizar experimentos com virus em um organismo como a mosca e utilizar o conhecimento adquirido para melhorar o entendimento da doença no ser humano, por exemplo,[40] ressaltando a importância da comparação entre redes PPI diferentes.

Devido a relevância da estrutura das redes PPI para o entendimento dos processos biológicos, o aumento na quantidade de proteínas conhecidas nas redes PPI e na quantidade de interações entre elas, utilizaremos dados recentes sobre as interações entre proteínas para realizar um estudo da topologia das redes que as representam. De maneira mais específica empregamos as medidas de centralidade,[1,10,41] — métricas que baseiam-se na estrutura das interações entre os elementos da rede para classificá-los de acordo com sua importância, — para analisar redes PPI de organismos distintos. Como encontramos poucos estudos[42] sobre a relação entre as centralidades nas redes PPI, exploramos o comportamento dessas medidas e como elas estão relacionadas umas com as outras. Para isso, realizamos comparações entre seu comportamento para os diferentes organismos e avaliamos se suas correlações podem ser utilizadas para caracterização das redes PPI. Outro ponto estudado sobre as centralidades é a existência de uma associação entre as medidas de um vértice e sua função biológica. Além disso, como os autovalores das matrizes de adjacências e Laplaciana também estão relacionados a propriedades topológicas das redes[43]; utilizamos o espectro da matriz Laplaciana normalizada dos organismos considerados, analisando suas distribuições como uma forma de comparação entre as redes, com o objetivo realizar uma caracterização destes organismos através de seus espectros.

Assim, os objetivos principais deste trabalho são:

- Estudar a topologia de redes PPI utilizando as medidas de centralidade e as propriedades espectrais — este tipo de análise é importante pois a topologia das redes PPI pode estar associada a funções biológicas, o que justifica a necessidade de uma boa caracterização de seus elementos. Por esta razão escolhemos utilizar as medidas de centralidade, já que elas baseiam-se em conceitos diversos para caracterizar os vértices de um grafo. Além disso a informação proveniente desta caracterização mais detalhada pode ser utilizada para o desenvolvimento de modelos mais realistas de redes PPI.

- Caracterização e comparação das redes, utilizando as propriedades topológicas estudadas — a comparação entre organismos pode ser utilizada para a identificação de proteínas e interações similares em organismos diferentes, já que organismos que sejam mais "semelhantes" provavelmente possuem estruturas de interação mais "parecidas".

No capítulo 2 discutimos alguns conceitos básicos e referências sobre grafos, medidas de centralidade e redes complexas utilizados ao longo das análises desenvolvidas neste documento, além disso também apresentamos algumas informações sobre o banco de dados STRING de onde retiramos as informações sobre as interações entre proteínas dos organismos utilizados neste trabalho. Em seguida no capítulo 3 apresentamos os resultados obtidos sobre o comportamento e correlação das medidas de centralidade para as redes PPI, a localização da centralidade de autovetor observada nessas medidas, a comparação entre redes utilizando o espectro da matriz Laplaciana normalizada e o agrupamento de proteínas utilizando as medidas de centralidade como características. Finalmente no capítulo 4 apresentamos as conclusões do trabalho e levantamos alguns pontos e perguntas que precisam ser analisados em trabalhos futuros e levantamos algumas ressalvas quanto às analises aqui conduzidas.

2 CONCEITOS DE REDES COMPLEXAS

2.1 Redes complexas

Redes complexas são uma maneira de representar sistemas compostos por partes principais que possuem algum tipo de interação entre si, como por exemplo pessoas (elementos principais) e suas amizades, relacionamentos ou vínculos de parentesco (como as interações). Para isso, as redes complexas utilizam o conceito dos grafos[44] — estruturas matemáticas compostas de um conjunto de vértices (também chamados de nós e geralmente representados graficamente como círculos) e arestas (ou ligações representadas por uma linha ou seta) conectando um par de vértices — para representar sistemas do mundo real; os vértices de um grafo representam os elementos principais do sistema sendo estudado e as arestas suas interações. Uma grande vantagem dos grafos é que eles permitem que ideias, métricas e conceitos baseados em seu formalismo sejam aplicados em uma grande diversidade de sistemas compostos por partes interagentes.

Embora não exista um consenso sobre o quais as diferenças entre um grafo e uma rede complexa, geralmente subentende-se que a rede representa um sistema real,[45] ou modelo que possui caraterísticas topológicas diferentes (ou mais raras) do que os grafos estudados anteriormente pelos matemáticos — como por exemplo as características *small-world*[46] e o coeficiente de aglomeração[47] — e um número mais elevado de vértices e arestas. Neste trabalho, entretanto, não seguiremos essa diferenciação de maneira tão rígida de modo que doravante os termos rede e grafo serão utilizados como sinônimos.

2.1.1 Origem da teoria de grafos

A primeira referência que conhecemos onde se utilizou o conceito de grafos é o artigo de 1736 do matemático suíço Leonard Euler onde ele desenvolve e utiliza o conceito de grafos para resolver o problema das pontes de Königsberg. O problema consistia no seguinte: o território da antiga cidade de Königsberg (figura 1) na Prússia (atualmente Kalingrado na Rússia) era dividido pelo rio Pregel, de modo que haviam construções da cidade em ambas as margens do rio, e em duas grandes ilhas centrais com sete pontes conectando as ilhas e as margens como ilustrado na figura 2a. Uma das perguntas da população na época é se seria possível fazer uma caminhada pela cidade passando por cada uma das pontes apenas uma vez. Inicialmente as pessoas tentaram responder essa pergunta realizando caminhadas pela cidade.

Utilizando uma simplificação semelhante àquela representada na figura 2b, Euler abstraiu as regiões da cidade e suas pontes como vértices e arestas de um grafo respectivamente. Euler utilizou o seguinte raciocínio para resolver o problema: para cruzar cada

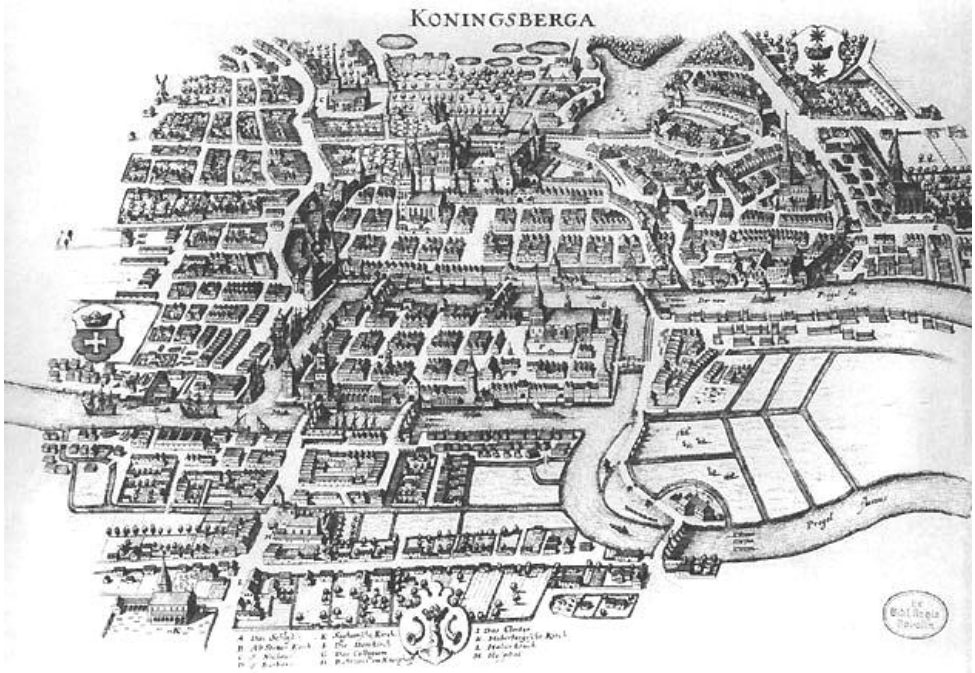


Figura 1 – Mapa da cidade de Königsberg na Prússia (século XVIII).

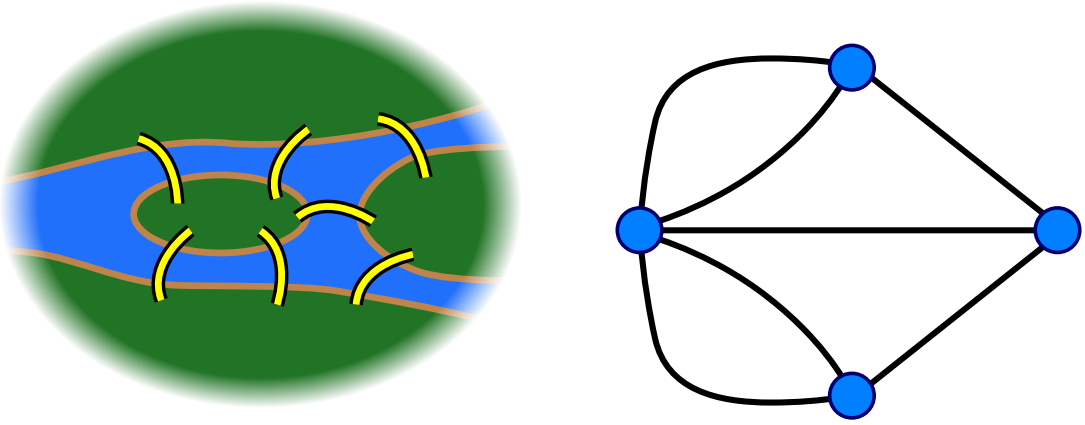
Fonte: MAP...[48]

ponte uma única vez, uma pessoa precisa chegar e sair de um vértice por arestas distintas, ou seja, com exceção do primeiro e do último vértice da caminhada (caso eles sejam vértices distintos), todos os demais precisam estar conectados a um número par de arestas. Devido a este problema, caminhadas que passam apenas uma vez por cada aresta de um grafo são chamados de caminhos Eulerianos (ou ciclos Eulerianos caso a caminhada termine no vértice inicial), ou seja, um grafo conectado (ver Seção 2.2) onde todos os vértices possuem número de conexões par possui um ciclo euleriano, e se ele possuir exatamente dois vértices com número ímpar de conexões então ele possui um caminho Euleriano. Desta maneira, utilizando o conceito de grafos, Euler criou um método geral de resolução muito mais simples que as buscas exaustivas realizadas anteriormente.

2.2 Conceitos básicos de grafos

Agora descreveremos brevemente e de maneira informal alguns conceitos básicos de grafos; essa etapa é importante pois esses termos aparecerão futuramente no texto. Para maiores detalhes sobre estes conceitos indicamos as referências[1,10,44,50]

O primeiro conceito que iremos descrever é o dos vértices $V \equiv \{n_1, n_2, \ldots n_N\}$ — muitas vezes também chamados de nós, ou pontos — eles são os elementos básicos do grafo; geralmente são representados visualmente por um círculo e contém algum tipo de índice ou identificador único que permite diferenciá-los. Em casos mais sofisticados os vértices da rede também podem possuir outras propriedades além do seu identificador que podem acrescentar informações dependendo da análise ou sistema que o grafo representa.



(a) Cidade de Königsberg (b) Abstração em um grafo

Figura 2 – Simplificação do mapa da cidade de Königsberg. Em (a) apresentamos uma visão simplificada da distribuição da cidade, suas pontes, e do rio Pregel. Em (b) a cidade é visualizada como um grafo.

Fonte: SEVEN...[49]

O próximo elemento de um grafo que descreveremos são as arestas $E \equiv \{e_1, e_2, \ldots, e_m\}$ — também conhecidas como *links* ou conexões. Elas representam algum tipo de relação entre os vértices e visualmente são representadas como uma linha conectando-os. Outra forma de se representar uma aresta é pela tupla dos vértices que ela conecta, como por exemplo $e_1 = (n_1, n_2)$ se a aresta e_1 conecta os nós n_1 e n_2. Dependendo do problema, as arestas podem ou não possuir direção, peso e ser ou não múltiplas — quando existe mais de uma aresta ligando o mesmo par de nós.

O terceiro conceito, que já utilizamos na seção 2.1.1, é o de caminhos, que são definidos como uma sequência de arestas conectadas existentes no grafo que conectam um vértice n_1 a um vértice n_2. Quando o vértice inicial é igual ao final ($n_1 = n_2$) o caminho é conhecido como fechado ou como um ciclo. Outra definição importante é a de caminho mínimo ou geodésico, que é a menor sequência de arestas que conecta dois vértices; caminhos mínimos não são necessariamente únicos e são utilizados nas definições de algumas medidas de centralidade (ver Seção 2.6) de grafos

O próximo conceito é o de componente conectada (ou apenas componente), que é definida como o maior subgrafo do grafo original onde para quaisquer dois vértices pertencentes a ela sempre existe pelo menos um caminho conectando-os, e nunca existe um caminho conectando dois vértices de componentes distintas. Um vértice sem nenhuma aresta é considerado uma componente.

Outro conceito relacionado a conectividade é o de grafo conectado — quando ele possui um único componente conectado.

K-core é o maior subconjunto de vértices tal que cada vértice é conectado a pelo

menos K outros vértices do subconjunto.[51] K-cores também indicam o quão externos ou internos os vértices são — vértices de K-core menor são mais periféricos enquanto que os de K-core mais elevados encontram-se mais no centro do grafo. O processo ocorre dessa maneira: inicialmente removemos todos os vértices da rede com grau 1; com a remoção deles alguns outros vértices também podem ficar com grau 1, nesse caso eles também são removidos; repete-se esse processo até que nenhum elemento com grau igual a 1 reste para ser removido; todos os nós removidos até essa etapa pertencem ao K_1-core, e em seguida repete-se o processo novamente para o grau 2 para o K_2-core e assim sucessivamente até que todos os vértices da rede possuam uma categoria de K-core.

2.3 Medidas de caracterização de redes complexas

Com o desenvolvimento das redes complexas, algumas medidas topológicas foram desenvolvidas com o intuito de se realizar uma melhor caracterização dos sistemas que elas representam. A primeira medida deste tipo que apresentaremos é o menor caminho médio,[1] que é o valor médio dos menores caminhos entre todos os pares de vértices da rede, ou seja:

$$\langle l \rangle = \frac{1}{N(N-1)} \sum_{s,t \in V} d(s,t), \tag{2.1}$$

onde $d(s,t)$ é o comprimento do menor caminho entre os vértices s e t e o fator $N(N-1)$ representa o número total de arestas que podem existir na rede.

Outro conceito que é bastante utilizado para a caracterização de redes é a assortatividade (*assortativity*) que mede se existe a tendência de que vértices com características em comum possuam um número maior de conexões com outros vértices com as mesmas características. Um tipo de medida muito importante e utilizada para a caracterização de redes é a assortatividade de grau.[52,53] Para as redes complexas a assortatividade de grau r é dada pelo coeficiente de correlação de Pearson[1] dos graus entre os pares de vértices conectados, que é equivalente a expressão[1]:

$$r = \frac{\sum_{ij}(A_{ij} - k_i k_j/2m)k_i k_j}{\sum_{ij}(k_i \delta_{ij} - k_i k_j/2m)k_i k_j}, \tag{2.2}$$

onde A_{ij} é a matriz de adjacências, m é o número total de arestas, k_i é o grau do vértice i e δ_{ij} é o delta de Kronecker.

Uma padrão de topologia muito utilizado para a caracterização de redes é a presença de triângulos — conjunto formado por três vértices e três arestas conectando todos os possíveis pares de vértices do conjunto, — que indicam a tendência de que dois vizinhos de um vértice da rede também estejam conectados entre si. Existem duas medidas que são mais comumente utilizadas para a caracterização da presença de triângulos na rede: transitividade[54] (T) e o coeficiente de aglomeração médio dos vértices.[47] A transitividade

de uma rede é dada por:

$$T = 3 \times \frac{(\text{número de triângulos})}{(\text{número de tríades})}, \tag{2.3}$$

onde o número de triângulos representa o total de triângulos presentes na rede e o número de tríades é a quantidade de arestas da rede que compartilham um mesmo vértice. O coeficiente de aglomeração médio ($\triangle$) da rede é dado pela média dos coeficientes de aglomeração individuais dos vértices, ou seja:

$$\triangle = \frac{1}{N} \sum_i \triangle_i, \tag{2.4}$$

onde $\triangle_i$ é o coeficiente de aglomeração do vértice i, e é dado por:

$$\triangle_i = \frac{\text{número de triângulos ligados ao vértice } i}{\text{número de tríades conectadas a } i} = \frac{2\sum_{k>j} A_{ij}A_{ik}A_{jk}}{k_i(k_i-1)}. \tag{2.5}$$

2.4 Representação computacional de uma rede complexa

Nesta seção discutiremos de maneira resumida algumas das formas de se representar computacionalmente um grafo ou rede complexa. Trataremos basicamente de três tipos de representação: matriz de adjacências, lista de arestas e lista de adjacências; maiores informações sobre cada uma dessas formas pode ser encontrada em.[55]

2.4.1 Matriz de adjacências

A primeira forma de representação de um grafo que apresentaremos é a matriz de adjacências. Nela, o grafo é representado por uma matriz $\mathbf{A}$ quadrada (e simétrica para grafos sem direção) com N linhas e colunas, onde N representa o número total de vértices presentes na rede. As entradas A_{ij} podem assumir os valores 1 caso exista uma aresta entre os nós i e j ou 0 caso contrário.

A vantagem desta representação é que para saber se existe uma ligação entre os vértices i e j basta verificar-se o elemento A_{ij} da matriz ($\mathcal{O}(1)$); sua desvantagem é que $\mathbf{A}$ necessita de $\mathcal{O}(N^2)$ espaço em memória o que pode tornar essa representação muito custosa para o caso de grafos esparsos por exemplo.

2.4.2 Lista de arestas

Quando representamos um grafo como uma lista de arestas, cada entrada da lista possui dois valores representando os identificadores dos nós que são conectados por aquela aresta. Caso as arestas possuam peso, pode-se adicionar um terceiro valor em cada tupla associado a ele. A grande vantagem deste tipo de representação é que para grafos esparsos ela ocupa menos espaço em memória ($\Theta(E)$), já que apenas as arestas existentes no grafo são guardadas. Por outro lado, a desvantagem deste método é que quando precisamos identificar se uma ligação existe no grafo é necessário percorrer a lista, o que pode tornar-se custoso quando trabalha-se com grafos muito grandes.

2.4.3 Lista de adjacências

A representação em lista de adjacências é uma combinação dos dois tipos apresentados anteriormente, já que ela guarda apenas informação sobre as conexões existentes como a lista de arestas mas apresentando uma visão dos vértices como a matriz de adjacências. A lista possui n entradas — geralmente ordenadas de acordo com o índice dos vértices — onde cada entrada é uma lista com os vértices adjacentes ao elemento referente àquela entrada. Para descobrir se os nós i e j estão conectados, basta procurar por j na entrada referente ao nó i, que leva um tempo $\Theta(k_i)$ onde k_i é o grau do vértice i. A representação em lista de adjacências apresenta vantagens e desvantagens com relação aos métodos anteriores: com relação ao tempo de busca ela é mais rápida que a lista de arestas e mais lenta que a matriz de adjacências; porém se o grafo for esparso, o espaço ocupado pela lista de adjacências é menor que o da matriz de adjacências e da mesma ordem que a lista de arestas.

2.5 A matriz Laplaciana

Outra matriz associada à estrutura da rede é a matriz Laplaciana,[43,50] que recebe este nome por estar associada ao operador Laplaciano. A matriz Laplaciana é definida como:

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \tag{2.6}$$

onde $\mathbf{A}$ é a matriz de adjacências definida na seção 2.4.1 e $\mathbf{D}$ é a matriz de graus que possui na diagonal d_{ii} o grau do nó i e os outros elementos iguais a zero. Os elementos l_{ij} de $\mathbf{L}$ podem assumir os valores:

$$l_{ij} = \begin{cases} k_i, & \text{se } i = j, \\ -1, & \text{se } i \neq j \text{ e existe uma aresta entre } i \text{ e } j, \\ 0, & \text{caso contrário.} \end{cases} \tag{2.7}$$

Os autovalores de $\mathbf{L}$ estão limitados no intervalo $0 \leq \lambda_j(\mathbf{L}) \leq 2k_{max}$.[43,56] A matriz Laplaciana é importante pois seu espectro está relacionado com a topologia da rede, por exemplo: a multiplicidade de autovalores $\lambda = 0$ está associada ao número de componentes conectadas da rede.

Existe também outra versão da matriz Laplaciana, que é chamada de matriz Laplaciana normalizada:[43,50]

$$\tilde{\mathbf{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}. \tag{2.8}$$

Os elementos de $\tilde{\mathbf{L}}$ são definidos como:

$$\tilde{l_{ij}} = \begin{cases} 1, & \text{se } i = j, \\ -(k_i k_j)^{-\frac{1}{2}}, & \text{se } i \neq j \text{ e existe uma aresta entre } i \text{ e } j, \\ 0, & \text{caso contrário.} \end{cases} \tag{2.9}$$

Uma característica interessante é que os autovalores da matriz Laplaciana normalizada estão limitados em: $0 \leq \lambda_j(\tilde{\mathbf{L}}) \leq 2$.[43,57]

2.6 Medidas de Centralidade

Quando analisamos redes complexas, uma pergunta que podemos fazer é: qual nó ou conjunto de nós é mais importante ou influente? A resposta para esta pergunta depende muito do problema e do tipo de sistema que está sendo analisado, já que um elemento pode ser considerado importante por diversos motivos e para isso foram desenvolvidas as *medidas de centralidade*. **Medidas de centralidade** são uma maneira de se mensurar a importância de um nó da rede baseando-se em algum conceito de importância e em sua topologia, ou seja, ideias diferentes de importância implicam em medidas distintas de centralidade e classificações diferentes dos nós. A seguir listaremos as medidas de centralidade utilizadas neste trabalho e explicaremos como elas são calculadas e qual o conceito de importância levado em consideração em cada uma delas.

2.6.1 Grau (*degree*)

O conceito da centralidade de grau é bastante simples: quanto maior a quantidade de conexões um elemento possuir, maior sua importância. Esta característica é bastante relevante em alguns sistemas como por exemplo redes sociais, onde quanto mais pessoas um indivíduo conhece, mais fácil é a divulgação e obtenção de informações do mesmo. Esta medida tem a vantagem de possuir um custo computacional baixo para ser calculada, permitindo classificação de nós em sistemas com centenas de milhares de elementos sem maiores problemas. Utilizando a representação de matriz de adjacências, a medida de grau para cada um dos vértices da rede é dada por:

$$k_i = \sum_{j=1}^{N} A_{ij}. \tag{2.10}$$

Nesta equação k_i é o valor da centralidade de grau para o nó i, N é o número total de vértices presentes na rede e A_{ij} é a matriz de adjacências representando a rede. Na figura 3 apresentamos o grafo estrela, onde o vértice central — de número 0 — possui uma quantidade de conexões muito maior do que os demais vértices do grafo. Quando deseja-se comparar a importância de nós presentes em redes distintas, é comum normalizar-se k_i por $(N-1)$ — número máximo de conexões que um nó pode possuir quando não existem auto ligações) — já que quanto maior a quantidade de elementos no sistema, maior são os valores médios obtidos para 2.10, ficando:

$$\tilde{k_i} = \frac{1}{(N-1)} \sum_{j=1}^{N} A_{ij}. \tag{2.11}$$

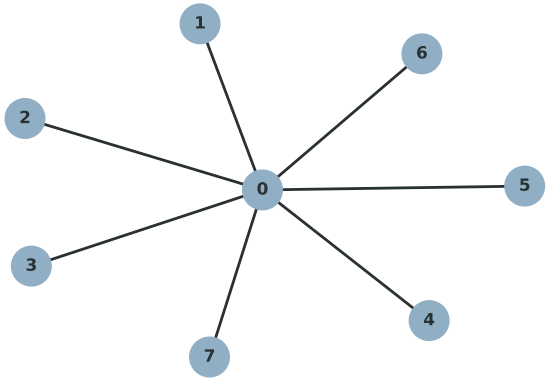


Figura 3 – Grafo estrela com 8 vértices e 7 arestas, nele o vértice central possui um número de conexões muito maior que os demais, sendo considerado mais importante pela centralidade de grau.

Fonte: Elaborada pelo autor

2.6.2 Autovetor (*eigenvector*)

Um vértice do grafo pode ser importante, não por possuir um grande número de conexões, mas sim por possuir apenas algumas com elementos que estão muito bem conectados. No caso de uma rede social, por exemplo, um indivíduo pode não estar muito conectado com os demais elementos da rede, entretanto se ele possuir algumas conexões com elementos importantes, ele também deve ser considerado importante. Essa é a ideia empregada na medida de autovetor, e trata-se de um extensão natural da medida de centralidade de grau.[1,58] Na figura 4 apresentamos um exemplo de grafo que contém um elemento — vértice 10 — que recebe valor elevado da centralidade de autovetor por estar ligado à dois outros vértices que possuem grau elevado.

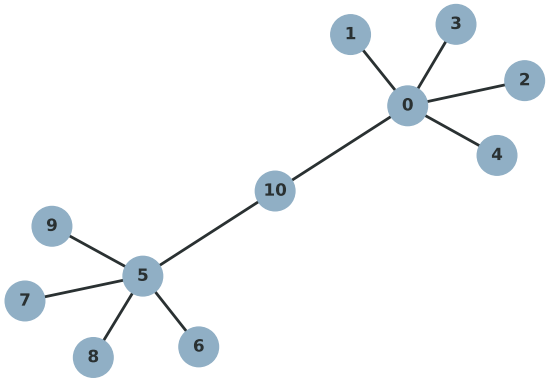


Figura 4 – O vértice 10 do grafo representa um elemento que é destacado como um dos mais importantes pela centralidade de autovetor.

Fonte: Elaborada pelo autor

Para calcular a medida de autovetor, vamos considerar que inicialmente (em t)

todos os vértices possuem centralidade $x_i = 1$. Assim, a centralidade $x_i^{'}$ na próxima interação ($t = t + 1$) é dada pela soma das centralidades de seus vizinhos, ou seja:

$$x_i^{'} = \frac{1}{\kappa_1} \sum_j A_{ij} x_j, \tag{2.12}$$

com A_{ij} sendo os elementos da matriz de adjacências $\mathbf{A}$ e κ_1 sendo o maior autovalor de $\mathbf{A}$. A equação 2.12 também pode ser escrita de forma matricial como $\kappa_1 \mathbf{x}^{'} = \mathbf{A}\mathbf{x}$ onde $\mathbf{x}$ é o vetor com os elementos x_i no tempo t e $\mathbf{x}^{'}$ é o vetor com os elementos x_i em $t + 1$. Repetindo-se o processo da equação 2.12 t vezes, ficamos com:

$$\kappa_1^t \mathbf{x(t)} = \mathbf{A^t x}, \tag{2.13}$$

com $\mathbf{x(t)}$ representando o vetor de centralidades após t iterações. Também podemos escrever o vetor $\mathbf{x}$ com os valores iniciais das centralidades na base dos autovetores $\mathbf{v_i}$ da matriz de adjacências, ou seja:

$$\mathbf{x} = \sum_i c_i \mathbf{v_i}. \tag{2.14}$$

Substituindo-se a equação 2.14 na 2.13 obtemos:

$$\kappa_1^t \mathbf{x(t)} = \mathbf{A^t} \sum_i c_i \mathbf{v_i} = \sum_i c_i \kappa_i^t \mathbf{v_i} = \kappa_1^t \sum_i c_i \left[\frac{\kappa i}{\kappa_1}\right]^t \mathbf{v_i}, \tag{2.15}$$

onde κ_i representa os autovalores da matriz de adjacências $\mathbf{A}$. Quando o número de iterações for grande o suficiente, os valores de $\mathbf{x(t)}$ na equação 2.15 serão proporcionais a $\mathbf{v_1}$, ou seja, na equação 2.15 quando $t \to \infty$ obtemos que $\mathbf{x(t)} \to c_1 \mathbf{v_1}$. Isto implica que o valor da centralidade de autovetor para o estado estacionário é dada por:

$$\mathbf{A}\mathbf{x} = \kappa_1 \mathbf{x}. \tag{2.16}$$

Logo a centralidade de autovetor do vértice i é proporcional a i-ésima componenete de $\mathbf{v_1}$, que é a centralidade proposta por Bonacich em 1987[58]; escrevendo a equação 2.16 de maneira explícita podemos observar que a centralidade do vértice i depende das centralidades de seus vizinhos como desejava-se inicialmente:

$$x_i^{'} = \kappa_1^{-1} \sum_j A_{ij} x_j. \tag{2.17}$$

2.6.3 Proximidade (*closeness*)

A centralidade de proximidade considera que os nós que estão mais próximos (em média) dos demais, são mais importantes para a rede. Podemos imaginar o caso em que deseja-se espalhar uma informação para todos os nós da rede no menor tempo possível, como por exemplo quando deseja-se divulgar um produto ou ainda informações sobre métodos de prevenção ou sintomas de uma determinada doença. Assim, a escolha de um vértice (ou conjunto de vértices) que encontra-se mais próximo de todos os demais como

fonte inicial para difundir a informação é fundamental para esta tarefa. A centralidade de proximidade pode ser definida[59] como:

$$c_i = \frac{1}{l_i} \text{ onde } l_i = \frac{1}{N-1} \sum_{j(\neq i)}^{N} d_{ij}, \tag{2.18}$$

sendo que N é o número total de nós da rede e d_{ij} é o comprimento da menor distância entre os nós i e j. A menor distância entre os vértices i e j é o comprimento do caminho mínimo que conecta esses dois vértices.

Podemos observar que l_i na equação 2.18 é a média das distâncias entre o vértice i e todos os demais vértices da rede. Entretanto devemos ressaltar que para calcularmos o valor da equação 2.18 é necessário que a rede seja conectada. A medida de proximidade c_i é definida como o inverso de l_i para manter-se o padrão das demais medidas, onde os vértices com maiores valores de centralidade são considerados mais importantes.

No grafo da figura 5 o vértice 9 recebe um grau de importância elevado pela centralidade de proximidade por possuir a menor distância média com relação aos demais elementos; utilizando-se o exemplo em que se deseja transmitir uma informação na rede, este vértice seria uma boa escolha para realizar esta tarefa.

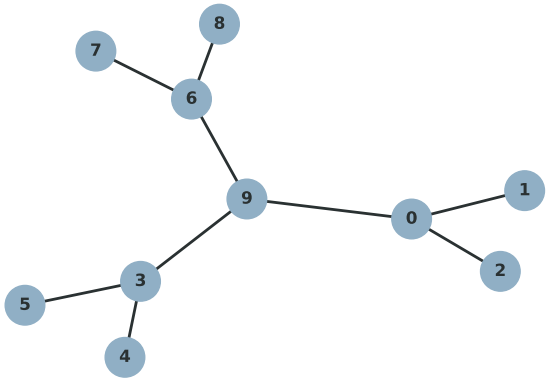


Figura 5 – Neste grafo o vértice 9 é considerado importante pela centralidade de proximidade por possuir a menor distância média com relação ao demais vértices do grafo.

Fonte: Elaborada pelo autor

2.6.4 Interposição (*betweenness*)

A centralidade de interposição[60,61] também utiliza o conceito de caminhos mínimos (ver seção 2.2). Ela considera mais importantes os vértices por onde passam uma grande fração dos caminhos mínimos da rede. Este conceito de importância pode ser utilizado, por exemplo, quando deseja-se realizar ataques maliciosos às redes pois a remoção de uma parte significativa dos nós com centralidade de interposição elevada faz com que o caminho mínimo médio da rede aumente, gerando implicações no tráfego de informações

entre os vértices. Outra importante aplicação onde a medida de interposição foi utilizada é na detecção de comunidades[62] — grupos de nós que são muito mais interconectados uns com os outros do que com os demais elementos da rede, — já que os vértices que conectam comunidades distintas participam de uma grande quantidade dos caminhos mínimos da rede. A figura 6 apresenta um grafo que foi composto por dois grafos completos — (0, 1, 2, 3, 4) e (5, 6, 7, 8, 9) — que foram conectados pelo vértice 10, que recebe um valor elevado da centralidade de interposição; note que os dois grafos completos poderiam ser estruturas mais complexas e que a remoção do elemento 10 transformaria a rede em dois grupos de elementos altamente conectados entre sí — ideia utilizada para a detecção de comunidades.

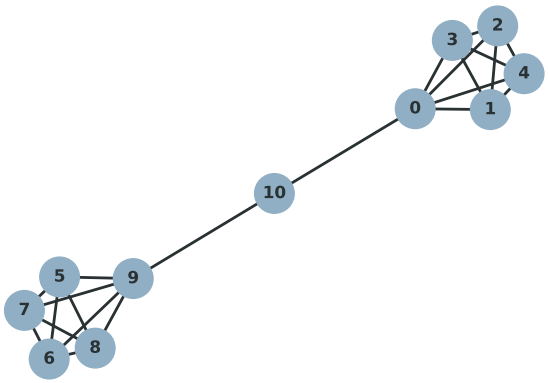


Figura 6 – O vértice 10 recebe o maior valor da centralidade de interposição no grafo abaixo, já que todos os menores caminhos conectando os elementos (5, 6, 7, 8, 9) a (0, 1, 2, 3, 4) passam por ele.

Fonte: Elaborada pelo autor

Matematicamente, a centralidade de interposição é definida como:

$$b_i = \sum_{st} \frac{n_{st}^i}{g_{st}}, \tag{2.19}$$

onde n_{st}^i é o número de caminhos mínimos entre os vértices s e t que passam pelo vértice i, e g_{st} é o número total de caminhos mínimos entre os vértices s e t. No caso de uma rede com mais de uma componente, a soma da equação 2.19 considera apenas os vértices pertencentes à mesma componente do nó i.

2.6.5 Interposição utilizando o modelo de fluxo de corrente (*current-flow betweenness*)

Um dos problemas presentes na medida de interposição, é que ela considera apenas os menores caminhos, sendo que caminhos ligeiramente maiores também são importantes. Imaginemos um cenário onde todas as informações trocadas entre os vértices da rede utilizam apenas os menores caminhos; isso geraria um congestionamento nessas arestas sendo que outras opções um pouco maiores permaneceriam livres. Na prática, em muitos

sistemas reais muitas vezes não temos informação sobre qual o menor caminho entre dois vértices da rede, já que este cálculo pode ser muito custoso.

Para resolver este problema, foi proposta[63,64] a medida de interposição baseada em fluxo de corrente, também conhecida como interposição utilizando o modelo de caminhada aleatória. Nesta medida, imaginamos que o grafo representa uma rede elétrica[44]: as arestas são resistores e os vértices são pontos de junção entre eles. Aqui utilizaremos a mesma abordagem feita por Newman[63] onde todos os resistores são iguais e de resistência unitária (a referência[64] apresenta o caso geral).

Nesta rede, escolhemos dois nós s e t ($s \neq t$) de modo que uma unidade de corrente é injetada em s e retirada em t a cada instante de tempo (como se existisse uma bateria ligada nos vértices s e t). Dessa forma a centralidade de interposição é a corrente líquida que passa por um vértice da rede, e para calculá-la faremos uso das leis de Kirchhoff.

Primeiramente precisamos lembrar que para um resistor, a lei de Ohm diz que $V = RI$, onde V é a diferença de potencial no resistor, R é o valor de sua resistência e I é a corrente passando por ele. Esta mesma expressão pode ser escrita como: $I = \frac{V}{R}$ ou ainda, $I = CV$, onde $C = \frac{1}{R}$ é a condutância do resistor. E como consideramos que o valor da resistência é unitário para essa rede, isso implica que $R = 1$, $C = 1$ e que $V = I$ para cada um dos resistores considerados. Escrevendo a lei de Kirchhoff das correntes diz que a soma das correntes que chegam em um nó do circuito é igual à soma das correntes que o deixam. Assim, para qualquer vértice i da rede, temos que:

$$\sum_j A_{ij}(V_i - V_j) = \delta_{is} - \delta_{it}, \tag{2.20}$$

onde A_{ij} são os elementos da matriz de adjacências, e os δ_{ij} são os delta de Kronecker, definidos como:

$$\delta_{ij} = \begin{cases} 1, & \text{se } i = j, \\ 0, & \text{se } i \neq j. \end{cases} \tag{2.21}$$

Desenvolvendo os termos da equação 2.20, vemos que ela pode ser escrita como:

$$\sum_j A_{ij}(V_i - V_j) = V_i \sum_j A_{ij} - \sum_j A_{ij}V_j = k_i V_i - \sum_j A_{ij}V_j = \delta_{is} - \delta_{it}. \tag{2.22}$$

Na segunda expressão da equação 2.22 utilizamos a definição da centralidade de grau (equação 2.10) para obter k_i.

Podemos escrever a equação 2.22 de forma matricial; para isso primeiramente precisamos definir o vetor $\mathbf{V}$ que é o vetor cujas componentes V_i representam os potenciais de cada um dos vértices i da rede. Utilizando as definições de $\mathbf{V}$ e $\mathbf{D}$ (seção 2.5), podemos escrever a equação 2.22 de forma matricial:

$$k_i V_i - \sum_j A_{ij}V_j = \delta_{is} - \delta_{it} = (\mathbf{D}_{(i)})\mathbf{V} - (\mathbf{A}_{(i)})\mathbf{V} = s_i = (\mathbf{D} - \mathbf{A})\mathbf{V} = \mathbf{s}, \tag{2.23}$$

onde $\mathbf{D}_{(i)}$ e $\mathbf{A}_{(i)}$ representam a linha i das matrizes $\mathbf{D}$ e $\mathbf{A}$ respectivamente, e $\mathbf{s}$ é o vetor composto pelos elementos s_i definidos como:

$$s_i = \begin{cases} +1, & \text{se } i = s, \\ -1, & \text{se } i = t, \\ 0, & \text{se } i \neq s, t. \end{cases} \tag{2.24}$$

O termo $(\mathbf{D} - \mathbf{A})$ presente na equação 2.23 é a definição da matriz laplaciana $\mathbf{L}$ como discutido na seção 2.5. Na seção 2.5 também discutimos que o menor autovalor da matriz Laplaciana tem valor 0 e possui o autovetor unitário $\mathbf{1} = (1, 1, 1, \dots)$ associado a ele. Como o determinante de uma matriz pode ser escrito como o produto de seus autovalores, isso implica que a matriz $\mathbf{L}$ possui determinante nulo e por esta razão $\mathbf{L}$ não possui inversa. Ou seja, para obtermos os valores de $\mathbf{V}$ necessários para o cálculo da interposição de fluxo de corrente não podemos simplesmente inverter a matriz $\mathbf{L}$.

Matematicamente o autovalor nulo de $\mathbf{L}$ nos indica que uma das equações do sistema é redundante, e fisicamente ele mostra que a corrente é conservada. Uma forma de contornarmos o problema, é escolher uma das equações do sistema e removê-la, ou de maneira mais prática, escolhermos um vértice v do sistema e medir os potenciais com relação a ele. Removendo-se a v-ésima linha e v-ésima coluna da matriz $\mathbf{L}$, ficamos com a matriz $\mathbf{L}_v = (\mathbf{D}_v - \mathbf{A}_v)$ que possui dimensão $(N-1) \times (N-1)$ e é invertível.

Dessa forma, a equação 2.23 fica:

$$\mathbf{V}_v = {\mathbf{L}_v}^{-1} \mathbf{s}_v. \tag{2.25}$$

Como o vértice v foi escolhido como referência, o potencial com relação a ele (V_v) é nulo. Dessa forma definimos a matriz $\mathbf{T}$ que é a matriz $\mathbf{L}_v$ acrescida novamente das informações sobre o vértice v, que tem potencial nulo, ou seja adicionamos uma linha e uma coluna com todos os elementos nulos na matriz $\mathbf{L}_v$. Utilizando a matriz $\mathbf{T}$ na equação 2.23, a voltagem $V_i^{(st)}$ do vértice i quando a bateria está ligada nos nós s e t é:

$$V_i^{(st)} = T_{is} - T_{it}. \tag{2.26}$$

A lei de Kirchhoff nos diz que a soma das correntes na junção de um circuito é nula; assim se considerarmos apenas o valor absoluto, estamos contabilizando duas vezes a corrente que passa por aquele elemento, ou seja, a corrente fluindo pelo i-ésimo vértice da rede, é dada pela metade da soma do valor absoluto das correntes fluindo através das ligações incidentes naquele vértice:

$$\begin{aligned} I_i^{(st)} &= \frac{1}{2} \sum_j A_{ij} |V_i^{(st)} - V_j^{(st)}| = \\ &= \frac{1}{2} \sum_j A_{ij} |T_{is} - T_{it} - T_{js} + T_{jt}|, \text{se } i \neq s, t. \end{aligned} \tag{2.27}$$

Para os vértices s e t que possuem fluxo de corrente unitário, o valor absoluto das correntes fica:

$$I_s^{(st)} = 1, I_t^{(st)} = 1. \tag{2.28}$$

Assim a centralidade de interposição baseada no fluxo de corrente de um vértice i da rede é definida como:

$$\omega_i = \frac{\sum_{s<t} I_i^{(st)}}{\left(\frac{1}{2}\right) N(N-1)}. \tag{2.29}$$

Diferentemente da medida tradicional de interposição, a interposição baseada no fluxo de corrente leva em consideração todos os caminhos existentes entre os nós s e t, com os caminhos menores sendo privilegiados (pois possuem uma resistência menor).

Existe também uma equivalência entre o modelo de fluxo de corrente e caminhadas aleatórias para o cálculo da medida de interposição; maiores detalhes podem ser obtidos na referência.[63]

2.6.6 Proximidade utilizando o modelo de fluxo de corrente (*current-flow closeness*)

A centralidade de proximidade utilizando o modelo do fluxo de corrente[64] é equivalente a outra medida chamada de information centrality,[65] que não é muito utilizada por ser pouco intuitiva. Segundo os autores a ideia de proximidade de fluxo de corrente é mais natural, fazendo com que seu uso seja mais frequente.

Para a centralidade de proximidade utilizando o fluxo de corrente, utilizamos as ideias da rede de resistores e suas junções apresentada na seção 2.6.5 e da medida de proximidade 2.6.3. A diferença é que ao invés de considerarmos apenas a distância dos menores caminhos entre os pares de vértices para o cálculo da média, consideramos a distância resistiva entre os elementos da rede. A distância resistiva entre dois vértices (i e j) da rede é dada pela diferença de potencial entre os mesmos quando uma unidade de corrente é injetada em i e removida em j. Desta forma, a distância resistiva entre os vértices i e j é dada por $R_{ij} = V_i^{(ij)} - V_j^{(ij)}$, onde o potencial $V_i^{(ij)}$ pode ser obtido através da equação 2.26. Sendo assim, a medida de centralidade de proximidade utilizando o modelo de fluxo de corrente de um vértice i do grafo é dada por:

$$\rho_i = \frac{(N-1)}{\sum_j R_{ij}}, \tag{2.30}$$

onde N é o número total de vértices na rede e R_{ij} é a distância resistiva entre os vértices i e j.

2.7 Modelos de redes complexas

No estudo de redes complexas, muitas vezes faz-se necessário o uso de modelos redes — redes geradas de maneira aleatória, seguindo uma regra de criação e que pos-

suem características específicas. Em muitos casos essas características tentam representar propriedades presentes em sistemas reais.

Dessa forma os modelos de rede permitem a criação de redes artificiais com propriedades específicas, o que nem sempre é possível para sistemas reais; ou seja para um conjunto de parâmetros de um modelo, podemos obter uma grande quantiade de redes com as mesmas propriedades, que devem possuir características semelhantes. Esse conjunto de redes com propriedades similares é importante pois possibilita a análise estatística dos fenômenos estudados. Além disso, outra vantagem dos modelos é que eles também possibilitam o isolamento de apenas alguns dos fatores que se deseja estudar, facilitando a associação de causa e consequência.

Nas seções 2.7.1 e 2.7.2 descrevemos os modelos de Erdős-Rényi e Barabási-Albert que foram utilizados neste trabalho, com algumas informações sobre suas propriedades. Escolhemos utilizar esses modelos por eles serem amplamente utilizados na literatura e por possuírem várias características já conhecidas; além disso o modelo ER gera redes em que todos os vértices são muito similares possuindo aproximadamente o mesmo grau médio, já o modelos BA gera redes heterogêneas com relação aos graus dos vértices e com a presença de *hubs*.

2.7.1 Modelo de Erdős-Rényi

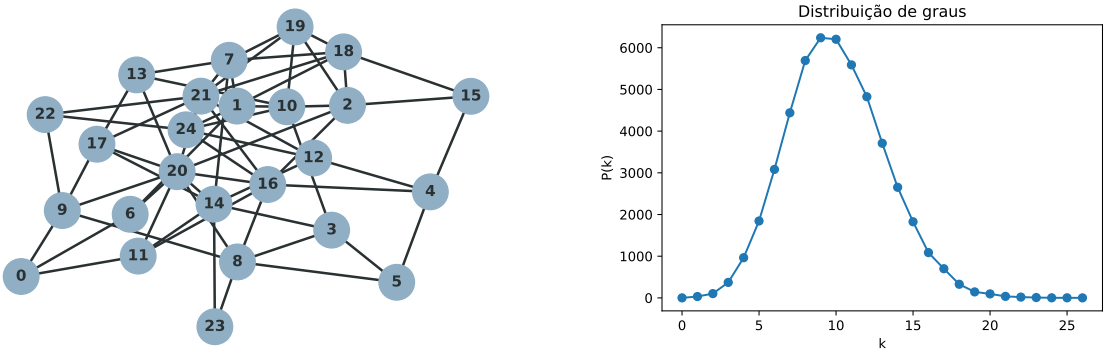
O modelo de redes aleatórias de Paul Erdős e Alfréd Rényi[66,67] (ER) é um dos mais conhecidos e utilizados no estudo de redes complexas. O modelo recebe dois parâmetros: N e p, onde N é o número de vértices que o grafo irá possuir e p é uma probabilidade — devido a esse fato este modelo é muitas vezes chamado de G(N, p). Inicialmente todos os N vértices encontram-se desconectados, em seguida, para todos os $N(N-1)/2$ possíveis pares de vértices é feita uma conexão com probabilidade p. No caso de $p = 0$ obtêm-se um grafo sem nenhuma aresta e quando $p = 1$ o resultado é o grafo completamente conectado. Devido à aleatoriedade do modelo para valores de $p \neq 1$ é possível encontrar a presença de vértices sem nenhuma ligação. Um exemplo de grafo gerado com este modelo é apresentado na figura 7a. Como mostrado na equação 2.31, o grau médio $\langle k \rangle$ do modelo ER pode ser calculado utilizando-se os valores de seus parâmetros N e p para redes suficientemente grandes.

$$\langle k \rangle = p(N-1). \tag{2.31}$$

Além disso os graus dos vértices de uma rede ER seguem a distribuição binomial, e para o limite de $N \to \infty$, $p \to 0$ mantendo-se $\langle k \rangle$ constante, sua distribuição de graus segue uma distribuição de Poisson[?] (equação 2.32), como apresentado na figura 7b, nela podemos observar também que o valor mais comum de grau acontece próximo do valor 10,

que é o esperado segundo a equação 2.31.

$$P(k) = \frac{e^{-\langle k \rangle} \langle k \rangle^k}{k!}. \tag{2.32}$$



(a) Grafo criado com o modelo ER

(b) Distribuição de graus do modelo ER

Figura 7 – Exemplo de grafo criado com modelo ER. Em (a) apresentamos um grafo pequeno (N=25 e p=0.2) gerado com o modelo. A distribuição de graus de um grafo maior (N=50000, p=0.0002) é apresentada na figura (b).

Fonte: Elaborada pelo autor

2.7.2 Modelo de Barabási-Albert

O modelo de Barabási-Albert[68] (BA) é um método de criação de redes que utiliza a ligação preferencial para adição de novos vértices na rede. No instante t_0 o grafo possui m_0 vértices cada um deles está conectado a um número de arestas ≥ 1. Em seguida a cada passo no tempo adiciona-se um novo vértice na rede, que irá se conectar a $m \leq m_0$ vértices preexistentes a ele — sendo que m é um número inteiro. A probabilidade de uma dessas m conexões ser feita com um vértice i é dada por:

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}. \tag{2.33}$$

Dessa maneira um vértice com grau 4 possui o dobro de probabilidade do que um com $k_i = 2$ de receber uma nova conexão, fazendo com que os elementos de maior grau na rede recebam mais conexões a cada instante no tempo, fenômeno conhecido como *rich-gets-richer*. Este modelo gera redes com uma distribuição de graus que segue uma lei de potência (Figura 8b):

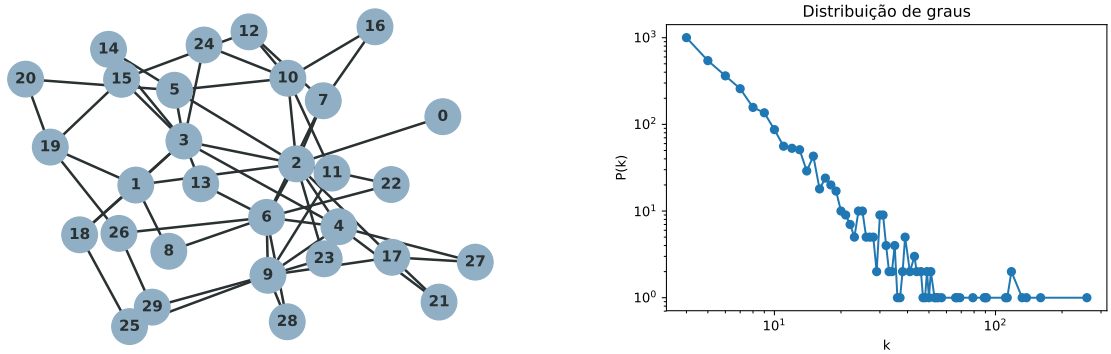
$$p(k) \propto k^{-\gamma}, \text{ onde } \gamma = 3. \tag{2.34}$$

Devido à forma como a rede é gerada, com exceção dos primeiros m_0 vértices da rede, todos os demais farão m novas conexões. Assim, para redes suficientemente

grandes, o número médio de conexões de uma rede gerada com o modelo BA será dado pela equação 2.35.

$$\langle k \rangle = 2m. \tag{2.35}$$

Na Figura 8a apresentamos um exemplo de rede gerada com este modelo. Na Figura da direita 8b apresentamos a distribuição de graus de uma realização do modelo, com N=3000 e m=4; podemos observar que a distribuição apresenta flutuações para graus mais altos.Para amenizar este fato para distribuições com cauda longa, geralmente utilizamos um gráfico da distribuição cumulativa $P(K \geq k)$ ao invés do gráfico de $P(k)$, onde $P(K \geq k)$ representa a probabilidade de um vértice da rede possuir grau maior ou igual a k.



(a) Grafo criado com o modelo BA

(b) Distribuição de graus do modelo BA

Figura 8 – Exemplo de grafo criado com modelo BA. Em (a) apresentamos um grafo pequeno (N=30 e m=2) gerado com o modelo. A distribuição de graus de um grafo maior (N=3000, p=4) é apresentada na figura (b).

Fonte: Elaborada pelo autor

2.8 O Banco de dados STRING

Neste trabalho, utilizaremos as redes de interação de proteínas (abreviadas como PPI — protein-protein interaction networks) de alguns organismos. As redes consideradas foram criadas utilizando a informação presente no banco de dados STRING.[69–78] Nele estão presentes informações sobre interações físicas (também chamadas de diretas) — quando duas ou mais proteínas interagem umas com as outras por contato físico para realizar suas funções biológicas — e funcionais (ou indiretas) — quando duas proteínas regulam suas quantidades mutuamente através da transcrição ou ainda quando elas participam de reações metabólicas subsequentes na mesma via metabólica — de proteínas.[79] Segundo a referência,[79] considerar os dois tipos de interação (diretas e indiretas) faz mais sentido do ponto de vista de função biológica e além disso mostrou-se que muitas das ligações que no passado aceitavam-se como físicas eram de fato funcionais.

O banco de dados apresenta informações sobre interações que já são conhecidas e também contém informações sobre interações previstas, dessa forma a cada uma delas é associado uma pontuação de confiabilidade de sua existência que pode assumir valores entre 0 e 1; quanto mais próximo de 1 for a pontuação, maior a certeza da existência da interação. Valores de pontuação menores que 0.4 são considerados de baixa confiabilidade, entre 0.4 e 0.7 intermediária e valores maiores 0.7 são considerados de altamente confiáveis.[71,80] Essas interações provém de cinco fontes principais: *predições de contexto genômico* — técnicas[81] que utilizam dados sobre os genes como *insights* sobre a interação entre proteínas como o método de perfil filogenético[82] que verifica que proteínas que evoluem de maneira correlacionada tendem a interagir, — *experimentos de laboratório de alto throughput* — experimentos auxiliados por computadores que identificam uma grande quantidade de interações de uma única vez, — *co-expressão conservada*[83] — predição da interação entre proteínas através da análise de genes que se expressam de maneira correlacionada após o processo de especiação ou duplicação, — *mineração de textos automatizada* — algorítimos de aprendizado de máquina aplicados em artigos científicos para buscar a confirmação da existência das interações — e *informações previamente conhecidas obtidas de outros bancos de dados de interações entre proteínas* — quando a informação também é reportada em outros bancos de dados amplamente utilizados pela comunidade científica. O score final S de uma interação é calculado[71] utilizando-se a equação 2.36, onde S_i representa o score individual de cada uma das fontes citadas acima.

$$S = 1 - \prod_i (1 - S_i). \tag{2.36}$$

Devido à forma como o STRING é construído, ele contém um número de interações e de organismos maior do que os demais bancos de dados, além das informações sobre as interações entre as proteínas e os métodos utilizados para identificá-las serem atualizados com mais frequência que os demais bancos de dados PPI. Ele também é muito utilizado na literatura, possuindo um grande número de citações, possui uma ferramenta de visualização do grafo de interações entre proteínas (para um número pequeno de elementos) e também possui ferramentas de análise de funções biológicas exercidas por um subconjunto de proteínas indicando se existem funções biológicas associadas aos elementos do subconjunto. Esses motivos e a facilidade para obtenção dos dados colaboraram para nossa escolha em utilizar o STRING como fonte de informações sobre as interações entre proteínas. Em nossa análises nós consideramos um valor de *threshold* para o score das interações, ou seja, apenas interações com score maior que o *threshold* foram consideradas verdadeiras, e as demais foram removidas. Todos os dados utilizados nesse trabalho foram retirados da versão 10.0 do STRING que contém um total de 2031 organismos, 9.6 milhões de proteínas e 184 milhões de interações. Escolhemos dois *thresholds* para a criação das redes: 0.5 e 0.7, utilizando a maior componente conectada e modelando-a como um grafo sem direção.

Utilizamos esses dois valores de *threshold* pois eles fornecem confiabilidades intermediária e elevada[71,80] para as interações presentes nas redes.

3 RESULTADOS E DISCUSSÃO

Neste capítulo iremos apresentar informações sobre os métodos utilizados em cada uma de nossas análises e discutir os resultados obtidos.

Na seção 3.1 apresentamos uma caracterização inicial das redes PPI, utilizando algumas medidas tradicionais para análise da topologia de redes; em seguida na seção 3.2 analisamos o comportamento e a relação entre as medidas de centralidade para as redes PPI de organismos distintos e como elas são alteradas pelas mudanças do *threshold* para a geração das redes e como esse comportamento difere do apresentado pelos modelos de rede utilizados.

Em seguida devido aos resultados obtidos pela comparação entre as medidas de centralidade, investigamos se a centralidade de autovetor está localizada paras as redes consideradas (seção 3.3) e qual o tipo de localização associada a essas redes.

Nas seções 3.4 e 3.5 buscamos caracterizar e comparar as redes PPI utilizando a correlação entre as medidas de centralidade e o espectro da matriz Laplaciana normalizada respectivamente.

Finalmente, na seção 3.6 realizamos um agrupamento das proteínas utilizando as medidas de centralidade e analisamos se é possível encontrar grupos de proteínas com funções biológicas definidas através dessa comparação.

3.1 Organismos escolhidos e caracterização inicial das redes de interação de proteínas

Como descrito na seção 2.8, modelamos as redes PPI de alguns organismos como redes sem direção. Escolhemos 8 organismos para nossos estudos: B. taurus (boi), C. elegans (verme), D. melanogaster (mosca-da-fruta ou mosca-do-vinagre), D. rerio (peixe-zebra), E. coli K12 W3110 (bactéria), H. sapiens, P. troglodytes (chimpanzé), S. cerevisae (levedura). O motivo de termos selecionado esses organismos é que eles são considerados organismos modelos e todos eles são amplamente estudados na literatura; além de representar uma faixa ampla de tipos diferentes de organismos. Em seguida nós modelamos suas PPI como redes e calculamos algumas medidas de grafos para comparação; os valores obtidos estão apresentados na Tabela 1. As colunas presentes nela referem-se ao número total de vértices presentes na rede (N), o número de arestas (m), o maior grau existente na rede (k_{max}), o valor médio da distribuição de graus ($\langle k \rangle$), o segundo momento da distribuição de graus ($\langle k^2 \rangle$), o maior autovalor da matriz de adjacências (λ_1), o caminho mínimo médio da rede ($\langle l \rangle$), a assortatividade (r), a aglomeração média dos vértices ($\triangle$) e o índice de transitividade (T).

Observando os dados da tabela notamos a existência de algumas características interessantes: primeiro ao diminuir o *threshold* aumentamos um pouco a quantidade de vértices em cada uma das redes, entretanto este acréscimo é muito mais significativo na quantidade de interações das redes, existindo casos em que elas apresentam aumento de mais de 100% — como para os organismos C. elegans e D. rerio por exemplo. Isto nos indica que a redução do score implica em mudanças significativas na estrutura topológica das redes de interações dos organismos estudados. A diminuição do *threshold* também implica em um grau máximo maior para todas as redes, indicando que algumas das novas conexões também são realizadas com os vértices de maior grau; destacamos os organismos B. taurus e D. rerio que apresentam um aumento significativo.

Tabela 1 – Características topológicas das redes PPI consideradas. O número em parênteses na frente do nome do organismo indica o valor de *threshold* daquela rede. Os símbolos significam: N: número de vértices, m: número de arestas, k_{max}: maior grau da rede, $\langle k \rangle$: grau médio, $\langle k^2 \rangle$: segundo momento da distribuição de graus, $\langle k^2 \rangle / \langle k \rangle$: o coeficiente de heterogeneidade (razão entre o segundo e o primeiro momento da distribuição de graus), λ_1: maior autovalor da matriz de adjacências, $\langle l \rangle$: caminho mínimo médio entre os vértices, r: grau de assortatividade, $\triangle$: aglomeração (clustering) média dos vértices e T: índice de transitividade.

Organismo	N	m	k_{max}	$\langle k \rangle$	$\langle k^2 \rangle$	$\langle k^2 \rangle / \langle k \rangle$	λ_1	$\langle l \rangle$	r	$\triangle$	T
B. taurus (0.7)	14582	338265	892	46.39	10600.05	228.50	449.10	3.92	0.77	0.37	0.79
B. taurus (0.5)	16787	540303	2320	64.37	15147.01	235.31	449.12	3.26	0.30	0.31	0.55
C. elegans (0.7)	10317	202489	1180	39.25	6270.14	159.06	266.86	3.87	0.31	0.36	0.48
C. elegans (0.5)	13391	477511	1748	71.32	18644.54	261.42	418.64	3.34	0.29	0.33	0.45
D. melanogaster (0.7)	10382	168275	1016	32.42	3698.88	114.09	170.91	3.93	0.21	0.39	0.42
D. melanogaster (0.5)	11834	328453	1528	55.51	9526.15	171.61	257.82	3.26	0.15	0.34	0.34
D. rerio (0.7)	15791	285630	881	36.18	4543.77	125.59	194.97	3.88	0.17	0.35	0.41
D. rerio (0.5)	20506	653449	1974	63.73	15011.04	235.54	251.80	3.30	-0.06	0.28	0.20
E. coli K12 W3110 (0.7)	3595	26902	166	14.97	476.20	31.81	63.87	4.40	0.43	0.41	0.42
E. coli K12 W3110 (0.5)	4095	53657	242	26.21	1370.82	52.30	78.30	3.45	0.34	0.35	0.31
H. sapiens (0.7)	15240	320926	4379	42.12	8101.95	192.35	342.86	3.38	0.02	0.35	0.52
H. sapiens (0.5)	18413	534778	6309	58.09	12988.25	223.59	342.87	3.10	-0.014	0.27	0.33
P. troglodytes (0.7)	11545	129690	1751	22.47	2093.36	93.16	151.94	3.69	0.01	0.27	0.28
P. troglodytes (0.5)	15295	320634	2756	41.93	6935.75	165.41	204.35	3.40	0.01	0.27	0.21
S. cerevisiae (0.7)	6035	137295	2182	45.50	6119.22	134.49	197.34	3.06	0.05	0.34	0.39
S. cerevisiae (0.5)	6220	239038	3045	76.86	14312.76	186.22	236.60	2.60	0.03	0.29	0.30

Fonte: Elaborada pelo autor

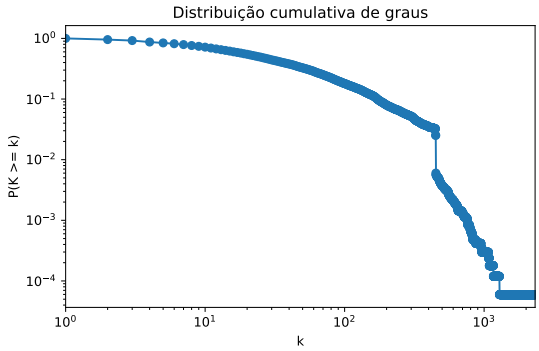
Tanto o grau médio quanto o segundo momento da distribuição de graus aumenta com a diminuição do *threshold*, como era esperado; entretanto olhando para o maior autovalor da matriz de adjacências observamos um fenômeno curioso — para todos os organismos o autovetor aumenta significativamente com a diminuição do *threshold* com exceção de dos organismos B. taurus e H. sapiens, onde a variação do autovetor é muito pequena. O caminho médio entre vértices tende a diminuir quando utilizamos um *threshold* menor, como era de se esperar devido ao número de arestas aumentar. A assortatividade de grau das redes possui valores próximos de 0 para a maioria dos organismos, com algumas exceções onde esse valor chega próximo de 0.4. A única exceção acontece para a rede *B. taurus* que apresentou o de 0.7 para o *threshold* mais elevado. A aglomeração média (média do *clustering* local) das redes não mostrou grandes variações entre organismos e nem mesmo com relação à mudança de *threshold*, ficando em torno de 0.3. Diferentemente da medida anterior a transitividade (*clustering* global) apresentou uma variação maior entre os organismos sendo que geralmente ela possui valores mais elevados pra o *threshold* maior, e tanto para o *B. taurus* quanto para o *H. sapiens* esse aumento é mais significativo do que nos demais organismos.

Além das medidas apresentadas na tabela 1, outra característica muito utilizada para se comparar redes é sua distribuição de graus. Na literatura ainda não existe um consenso sobre a distribuição de grau das redes PPI[84]; porém em alguns artigos elas são reportadas como possuindo uma distribuição de graus que segue uma lei de potência. Nas Figuras 9 e 10 apresentamos as distribuições cumulativas de grau das redes PPI com *thresholds* de 0.5 e 0.7 respectivamente. Comparando os gráficos das duas Figuras para um mesmo organismo, notamos que eles não apresentam mudanças significativas com a variação do *threshold* escolhido. As distribuições de graus são claramente heterogêneas — fato que também pode ser verificado pelo coeficiente de heterogeneidade ($\langle k^2 \rangle / \langle k \rangle$) também apresentado na tabela 1, entretanto por não comportarem-se como uma linha reta em escala log-log elas não apresentam características de uma distribuição de graus livre de escala.

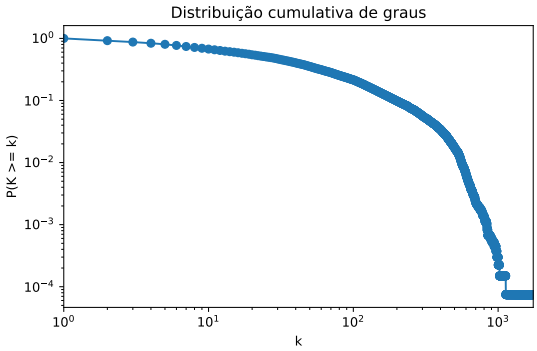
Também podemos observar que as distribuições de grau dos organismos *B. taurus* (Figuras 9a e 10a) e *H. sapiens* (Figuras 9f e 10f) apresentam uma região onde ocorre uma queda abrupta na curva da distribuição cumulativa de graus.

3.2 Correlação entre as medidas de centralidade das redes de interação de proteínas

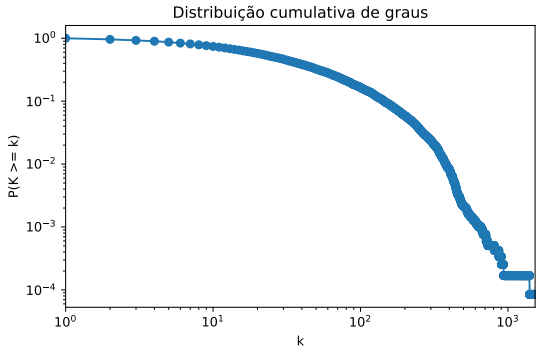
Após a caracterização inicial das redes PPI realizada na seção 3.1 nós daremos prosseguimento às análises de nossos resultados. Nesta seção estudaremos o comportamento das medidas de centralidade — mais especificamente sua correlação — para as redes PPI, comparando-as com os modelos ER e BA. Essa análise baseia-se em um trabalho anterior[85]



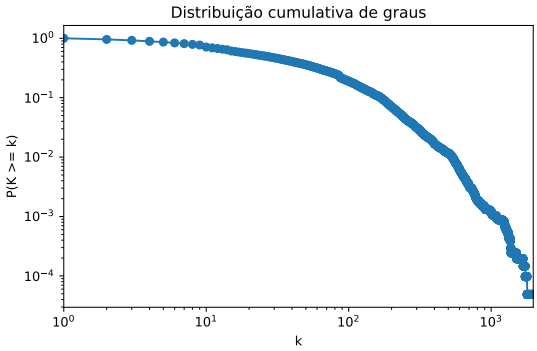
(a) B. taurus



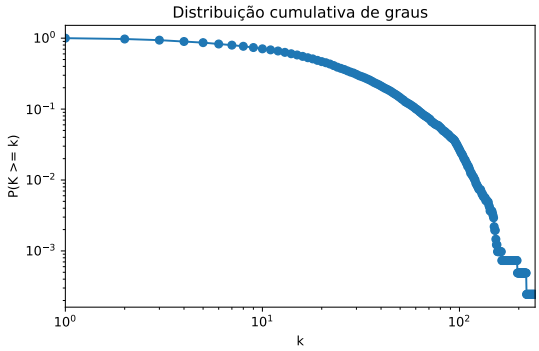
(b) C. elegans



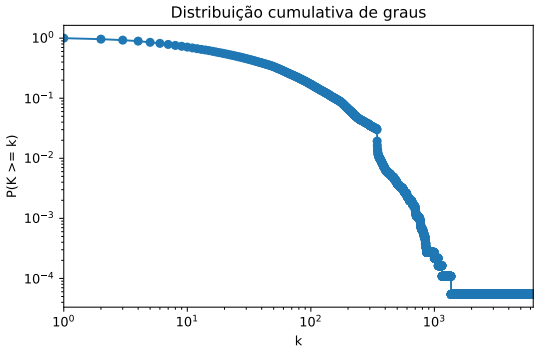
(c) D. melanogaster



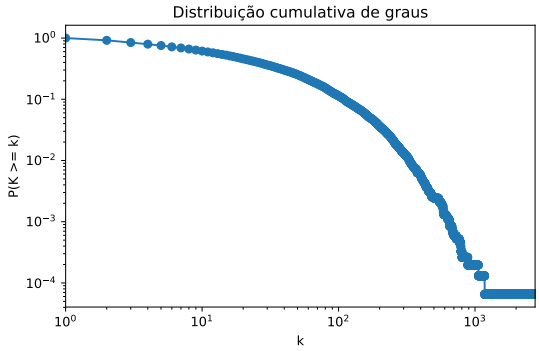
(d) D. rerio



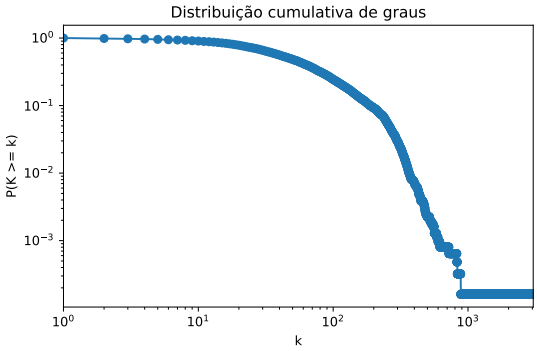
(e) E. coli



(f) H. sapiens



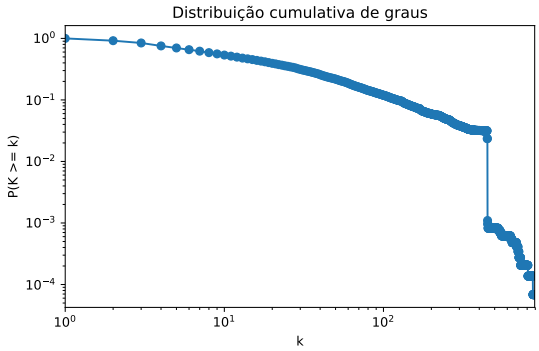
(g) P. troglodytes



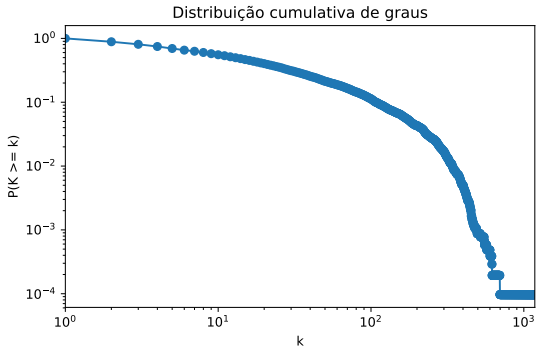
(h) S. cerevisiae

Figura 9 – Distribuições cumulativas de grau para as redes PPI com *threshold* de 0.5

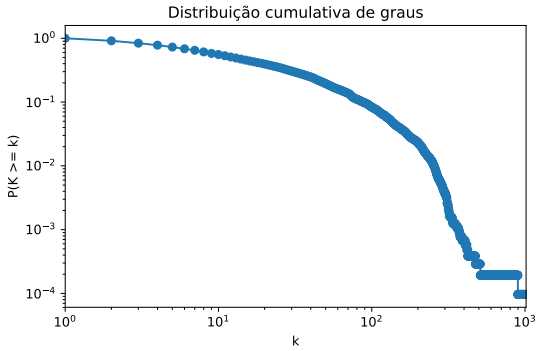
Fonte: Elaborada pelo autor



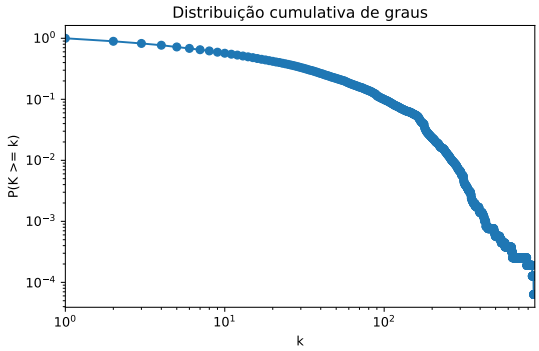
(a) B. taurus



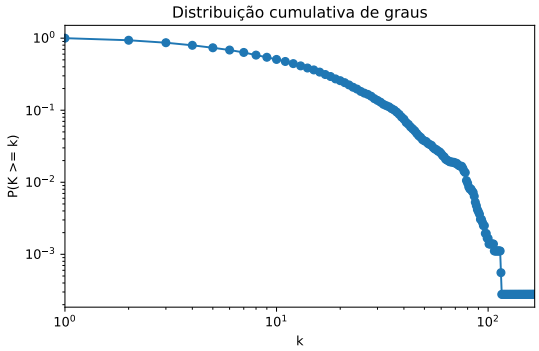
(b) C. elegans



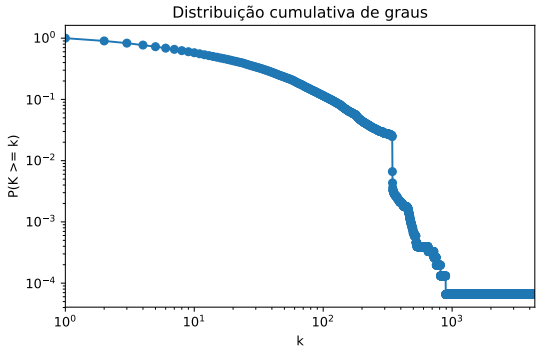
(c) D. melanogaster



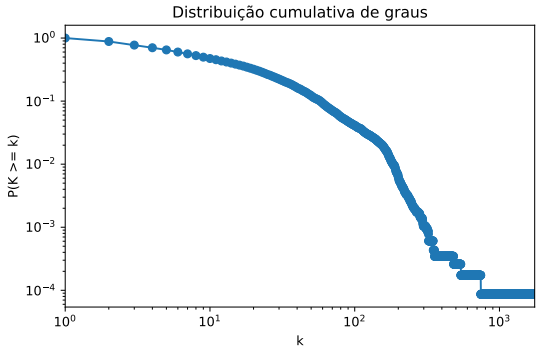
(d) D. rerio



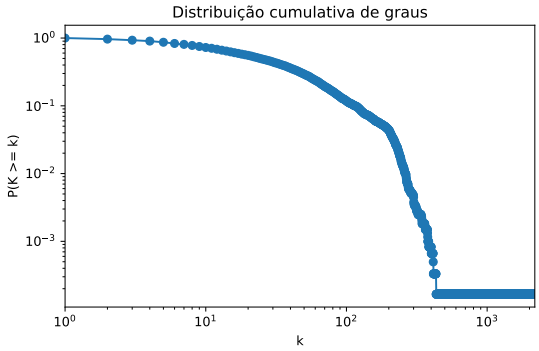
(e) E. coli



(f) H. sapiens



(g) P. troglodytes



(h) S. cerevisiae

Figura 10 – Distribuições cumulativas de grau para as redes PPI com *threshold* de 0.7

Fonte: Elaborada pelo autor

onde os autores demonstraram que é possível utilizar a correlação entre medidas de centralidade como *features* para comparar redes de tipos distintos. Além disso, outro trabalho[86] também argumenta que as correlações entre medidas de centralidade estão intimamente relacionadas com as propriedades estruturais das redes para alguns tipos de grafos. Aqui realizaremos o estudo das correlações citado acima considerando apenas as redes PPI e os modelos ER e BA.

Nessa análise calculamos as medidas de centralidade descritas na seção 2.6 para todos os organismos considerados e em seguida geramos *scatter-plots* entre cada par delas. Nas Figuras 11 a 18 apresentamos os gráficos para as redes com *threshold* de 0.5 e nas Figuras 19 a 26 para 0.7. Como podemos observar nelas a distribuição dos pontos não apresenta um comportamento linear, e devido a este fato, utilizamos o coeficiente de correlação de Kendall[87] para calcular o valor numérico da correlação entre as medidas de centralidade (apresentadas como a letra t no título de cada *scatter-plot*). Escolhemos a correlação de Kendall (também chamada de τ de Kendall) pois a forma como as medidas estão correlacionas não é linear e por ela apresentar vantagens[88] com relação a correlação de Spearman.[89] Os valores do coeficiente de correlação de Kendall para todos os pares de medidas dos organismos considerados também são apresentados na Tabela 2. Além disso, observando as Figuras notamos que a forma da correlação entre as centralidades parece ser pouco influenciada pelo *threshold* escolhido, e que seu comportamento é bastante semelhante entre organismos diferentes havendo poucas variações presentes entre eles. As únicas exceções acontecem para a medida de autovetor nas redes *B. taurus* e *H. sapiens*. Os resultados sugerem que os padrões observados podem estar relacionados com as funções desempenhadas pelas redes PPI.

Na Tabela 2, podemos observar que para o *threshold* de 0.5 os valores da correlação tendem a ser um pouco maiores, como por exemplo nas correlações entre grau e proximidade (D/C) ou ainda proximidade e proximidade de fluxo de corrente (C/CFB). Outra característica que podemos observar é que no geral, independentemente do *threshold* escolhido, as correlações tendem a dar valores médio-altos de concordância ($\geqslant 0.5$) entre as centralidades, com as exceções estando relacionadas aos pares formados pela medida de autovetor com as medidas de interposição (E/B) e interposição de fluxo de corrente (E/CFB). Além disso os dados da Tabela 2 nos mostram que para uma dada escolha de *threshold* os valores de correlação para um par de medidas é bastante próximo entre todos os organismos.

Um padrão bastante recorrente nos gráficos é o formato de "joelho" ou "cotovelo" que ocorre quando uma medida apresenta uma grande variação em uma região na qual a outra medida praticamente não exibe mudanças. Esse padrão está presente para os pares de medidas: grau e proximidade (D/C), grau e proximidade de fluxo de corrente (D/CFC), interposição e autovetor (B/E), interposição e proximidade de fluxo de corrente

(B/CFC), autovetor e interposição de fluxo de corrente (E/CFB), autovetor e proximidade de fluxo de corrente (E/CFC) e interposição de fluxo de corrente e proximidade de fluxo de corrente (CFB/CFC). É importante ressaltar quando esse padrão de "joelho" ocorre, isso indica que as duas medidas são boas para caracterizar vértices distintos, ou seja, uma ou outra medida pode ser mais ou menos discriminatória para caracterização dependendo do conjunto de vértices analisados. Existe uma grande quantidade de vértices que possuem centralidade de grau próxima de 0, que podem ser diferenciados através das medidas de proximidade ou proximidade de fluxo de corrente e para valores ligeiramente maiores de grau ambas as medidas de proximidade tornam-se correlacionadas com o grau. A mesma ideia também pode ser aplicada para a medida de interposição, onde as medidas de autovetor e proximidade de fluxo de corrente podem ser utilizadas para diferenciar os vértices que recebem valores de interposição próximos de 0. O mesmo tipo de raciocínio também pode ser aplicado às demais medidas citadas no início desse parágrafo, mostrando que apenas uma centralidade não é suficiente para uma caracterização completa dos vértices das redes PPI e que utilizar um conjunto de medidas é capaz de extrair mais informação do que apenas uma delas.

Outro padrão frequente é o que ocorre para os pares de medidas de grau e interposição (D/B), proximidade e interposição (C/B) e proximidade e interposição de fluxo de corrente (C/CFB). Nesse segundo padrão observamos duas regiões: a primeira — associada aos vértices com *ranking* de importância menor — onde as duas medidas não apresentam muita correlação e a segunda — associada aos vértices com importância médio-alta — onde ambas as medidas apresentam alguma correlação. No caso do par grau e interposição, por exemplo, observamos que vértices com valores de interposição próximos a 0 possuem valores de baixos a médios da centralidade de grau (que pode ser utilizada para diferenciá-los), enquanto que proteínas com valores mais elevados da centralidade de interposição recebem uma classificação semelhante da medida de grau. O mesmo também acontece para os demais pares de centralidades onde esse padrão está presente.

O terceiro padrão de correlação entre as medidas que discutiremos é o presente entre os pares grau e autovetor (D/E) e proximidade e autovetor (C/E). Nele existem claramente dois grupos de proteínas, aquelas nas quais ambas as medidas estão correlacionadas e as que não possuem correlação entre as medidas. Para o par grau e autovetor existem vértices que recebem valores baixos de importância pela medida de autovetor, mas que recebem as mais variadas classificações da medida de grau, enquanto que ao mesmo tempo existe um conjunto de elementos que recebem graus de importância parecidos por ambas as medidas — o grupo correlacionado — que possui representantes desde os menores valores das duas medidas até os maiores valores. Para as centralidades de proximidade e autovetor, o grupo descorrelacionado apresenta um comportamento parecido com o caso do par grau e autovetor onde para vértices com centralidade de autovetor próxima de 0 recebem os mais diversos valores de centralidade de proximidade; porém o grupo correlacionado possui um

comportamento um pouco diferente pois a correlação existe para proteínas com rankings de proximidade intermediários e termina nas proteínas com classificação médio-altas da medida de proximidade.

Os demais pares de centralidades são caracterizados por padrões diferenciados que ocorrem apenas para um par de medidas, como no caso do par grau e interposição de fluxo de corrente (D/CFB) onde a maioria dos vértices recebem *rankings* médio-baixos de importância por ambas as medidas. O valor da correlação de Kendall (Tabela 2) nos mostra que existe uma relação não trivial de correlação entre os elementos na região dos valores médio-baixos. E existe uma pequena quantidade de proteínas que recebe valores elevados por ambas as medidas, entretanto não existe um padrão claro de como as duas centralidades se relacionam para esses elementos mais importantes.

Os vértices classificados pelas medidas de proximidade e proximidade de fluxo de corrente (C/CFC) distribuem-se como um "S" nos scatter-plots de todos os organismos. Isto nos indica que as duas medidas de proximidade concordam com relação aos elementos de maior e menor importância das redes consideradas; entretanto existe alguma discordância entre os elementos com *rankings* intermediários. Apesar disso, os valores da correlação de Kendall para esse par de medidas nos indica que mesmo para as proteínas que recebem esses valores intermediários ambas as centralidades tendem a concordar.

O último par de medidas que apresenta um comportamento único é o formado por interposição e interposição de fluxo de corrente. De maneira similar ao par grau e interposição de fluxo de corrente, os vértices ficam concentrados na região dos rankings baixo-médios; entretanto eles são mais correlacionados para essa região que no caso anterior, fato que pode ser verificado tanto visualmente quanto pelo valor da correlação de Kendall.

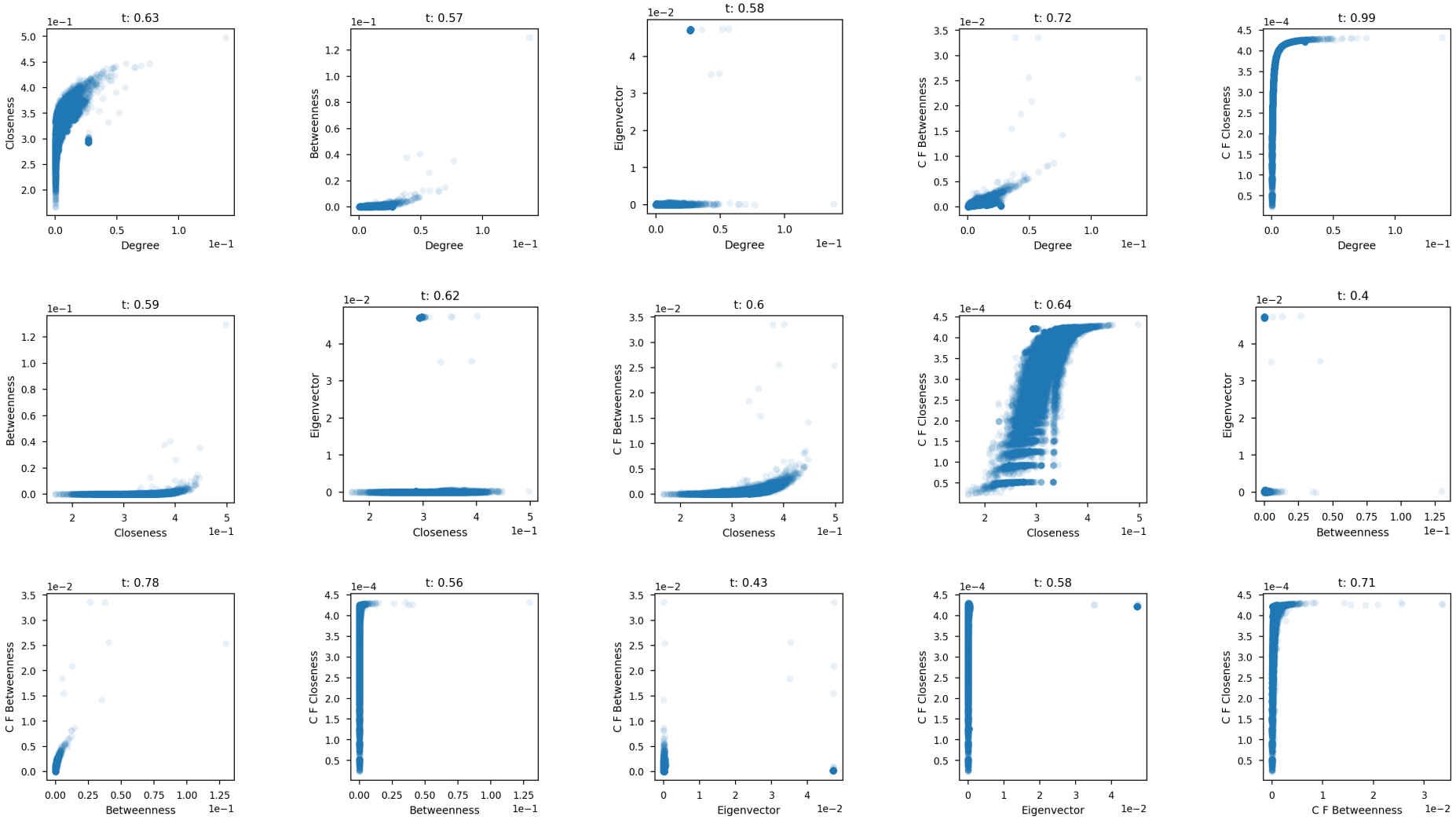


Figura 11 – Scatter-plots entre os pares de centralidade para o organismo *B. taurus* com *threshold* de 0.5

Fonte: Elaborada pelo autor

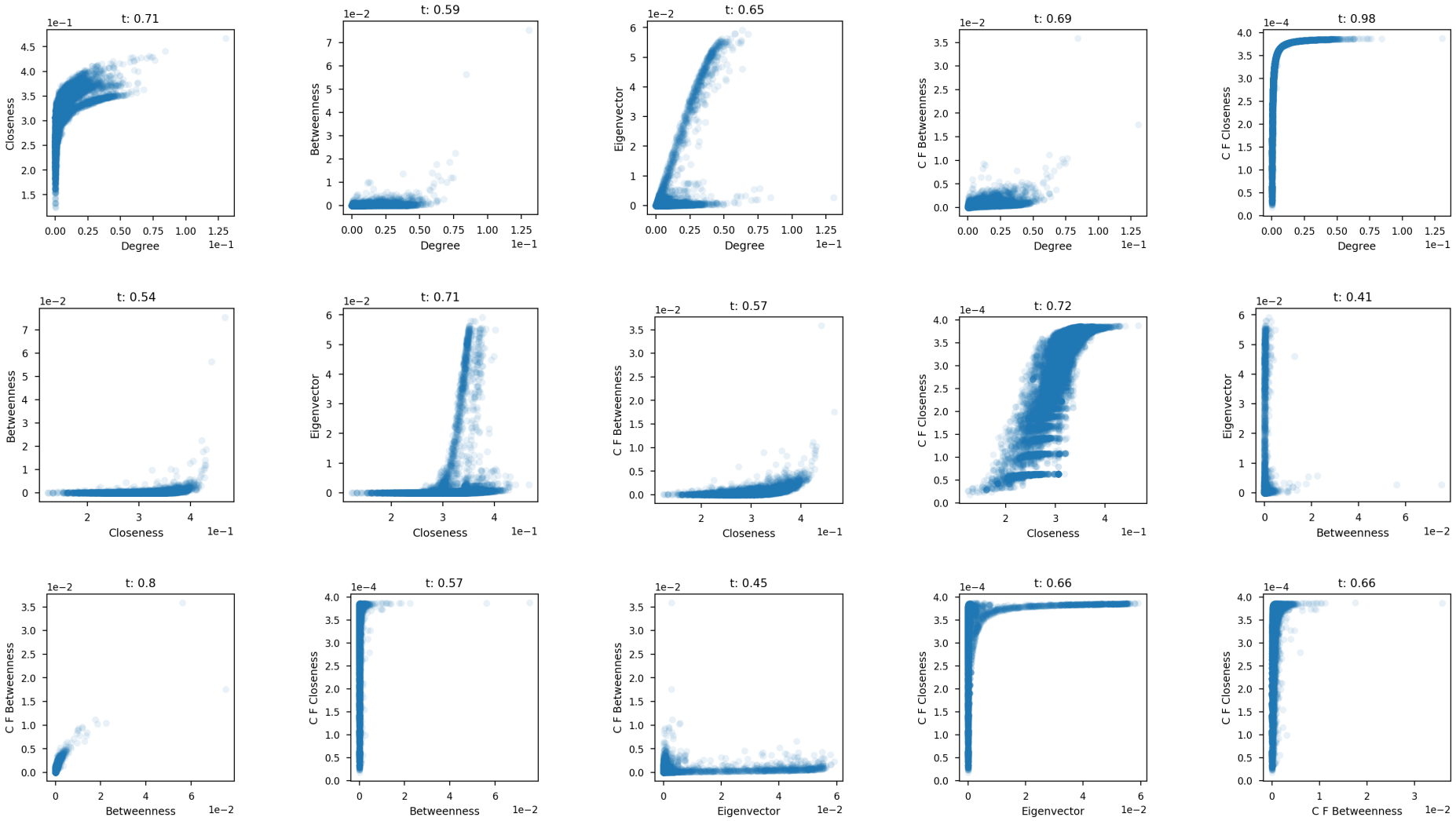


Figura 12 – Scatter-plots entre os pares de centralidade para o organismo *C. elegans* com *threshold* de 0.5

Fonte: Elaborada pelo autor

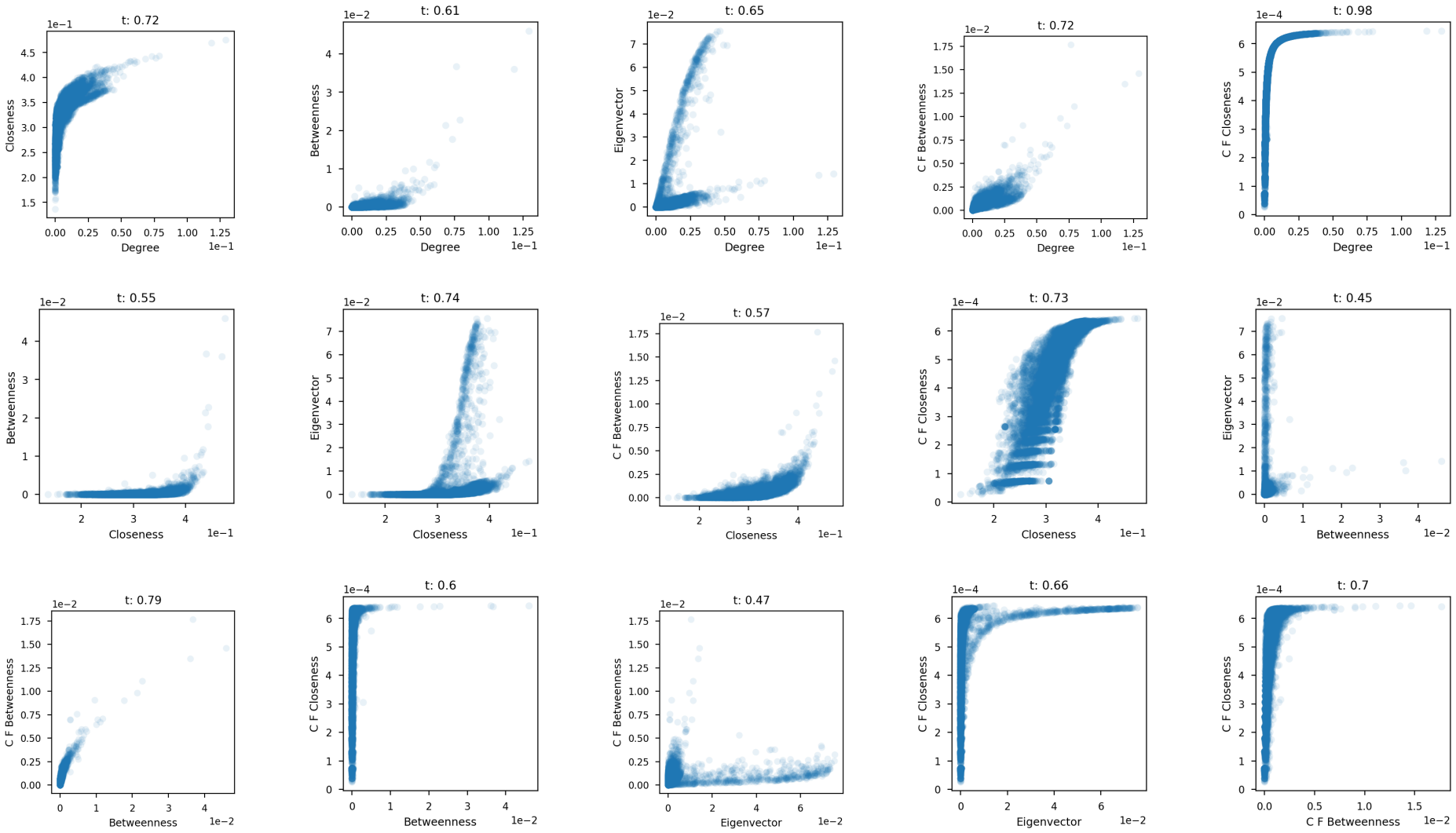


Figura 13 – Scatter-plots entre os pares de centralidade para o organismo *D. melanogaster* com *threshold* de 0.5

Fonte: Elaborada pelo autor

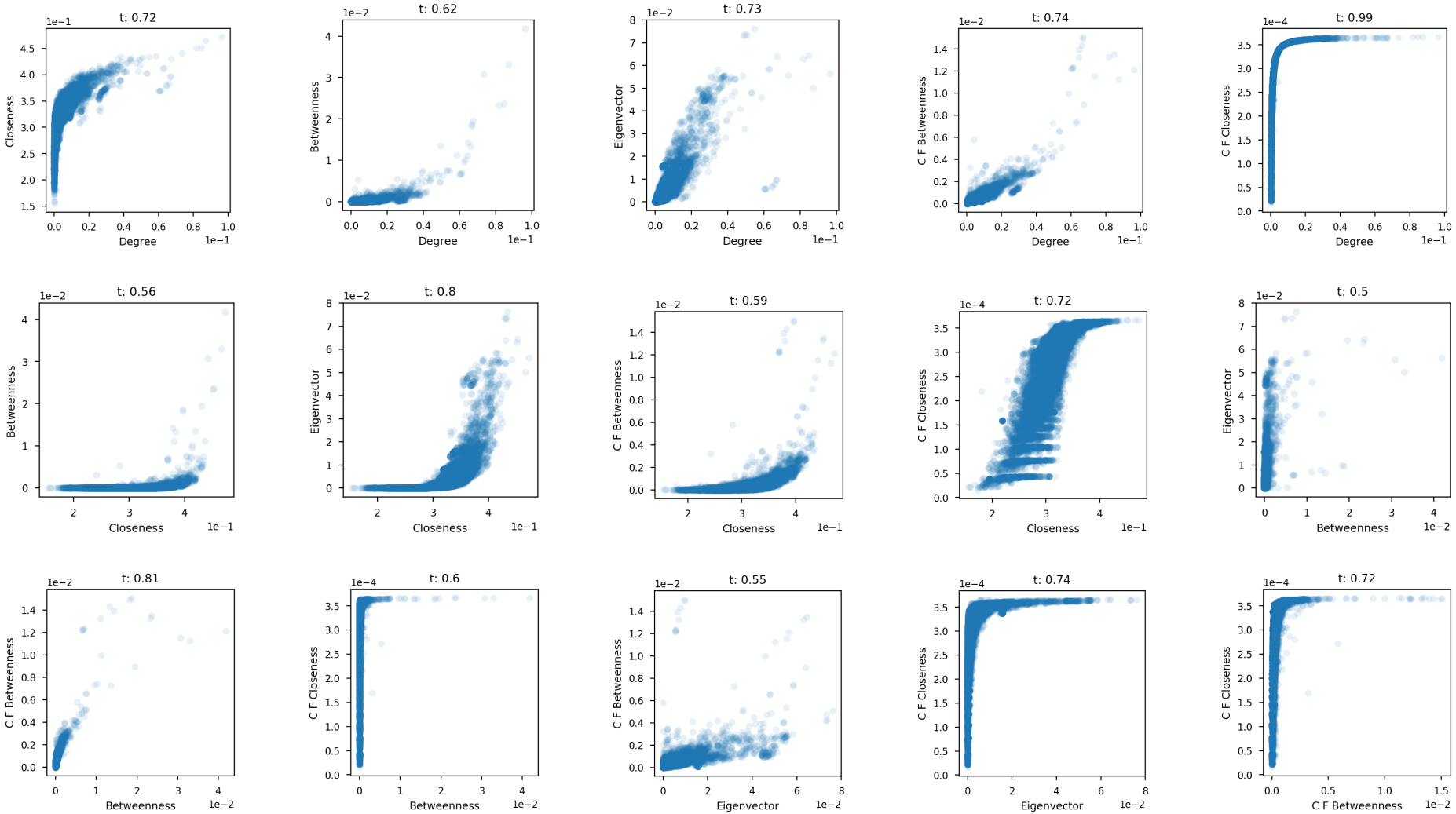


Figura 14 – Scatter-plots entre os pares de centralidade para o organismo *D. rerio* com *threshold* de 0.5

Fonte: Elaborada pelo autor

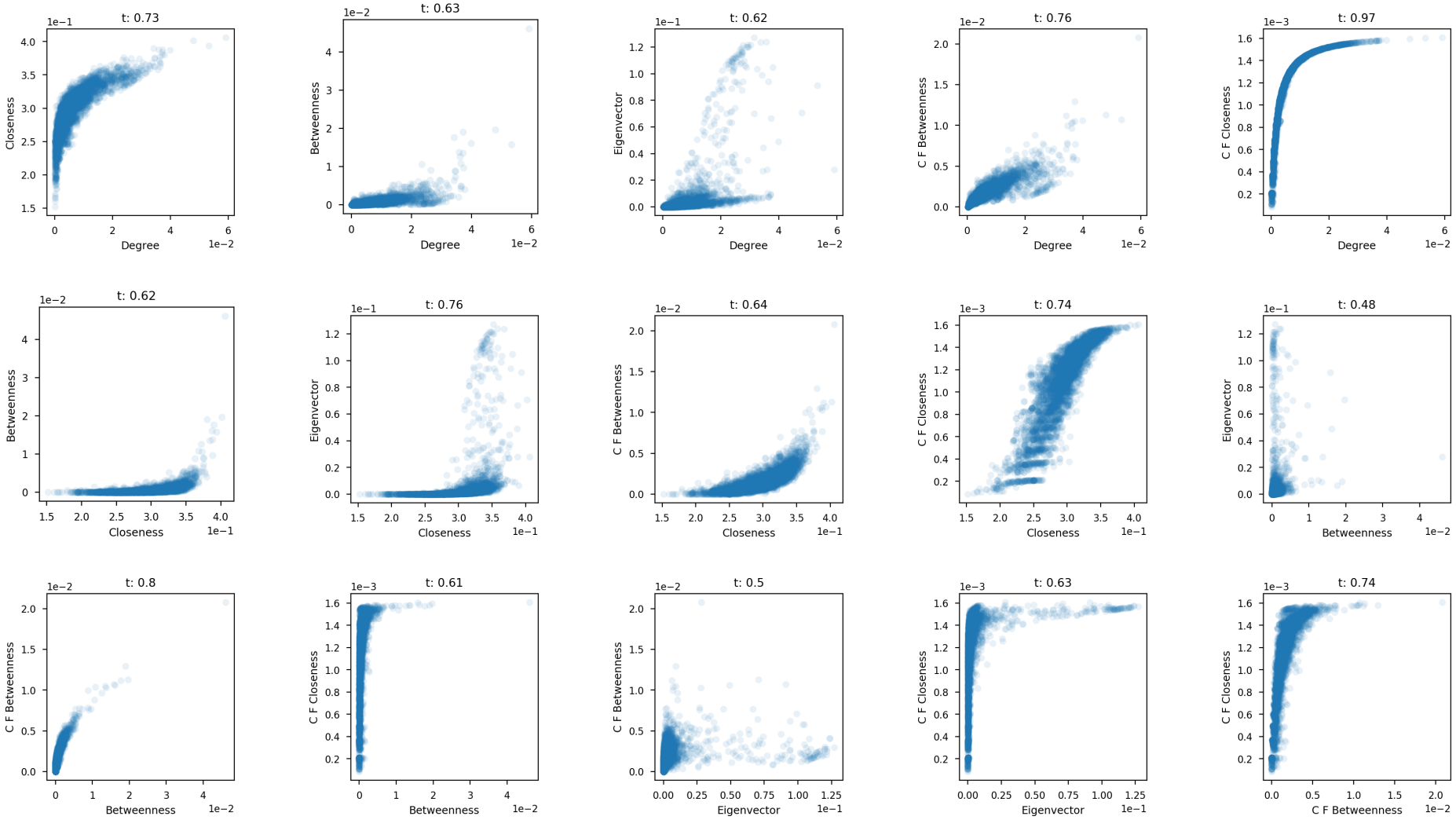


Figura 15 – Scatter-plots entre os pares de centralidade para o organismo *E. coli K12 W3110* com *threshold* de 0.5

Fonte: Elaborada pelo autor

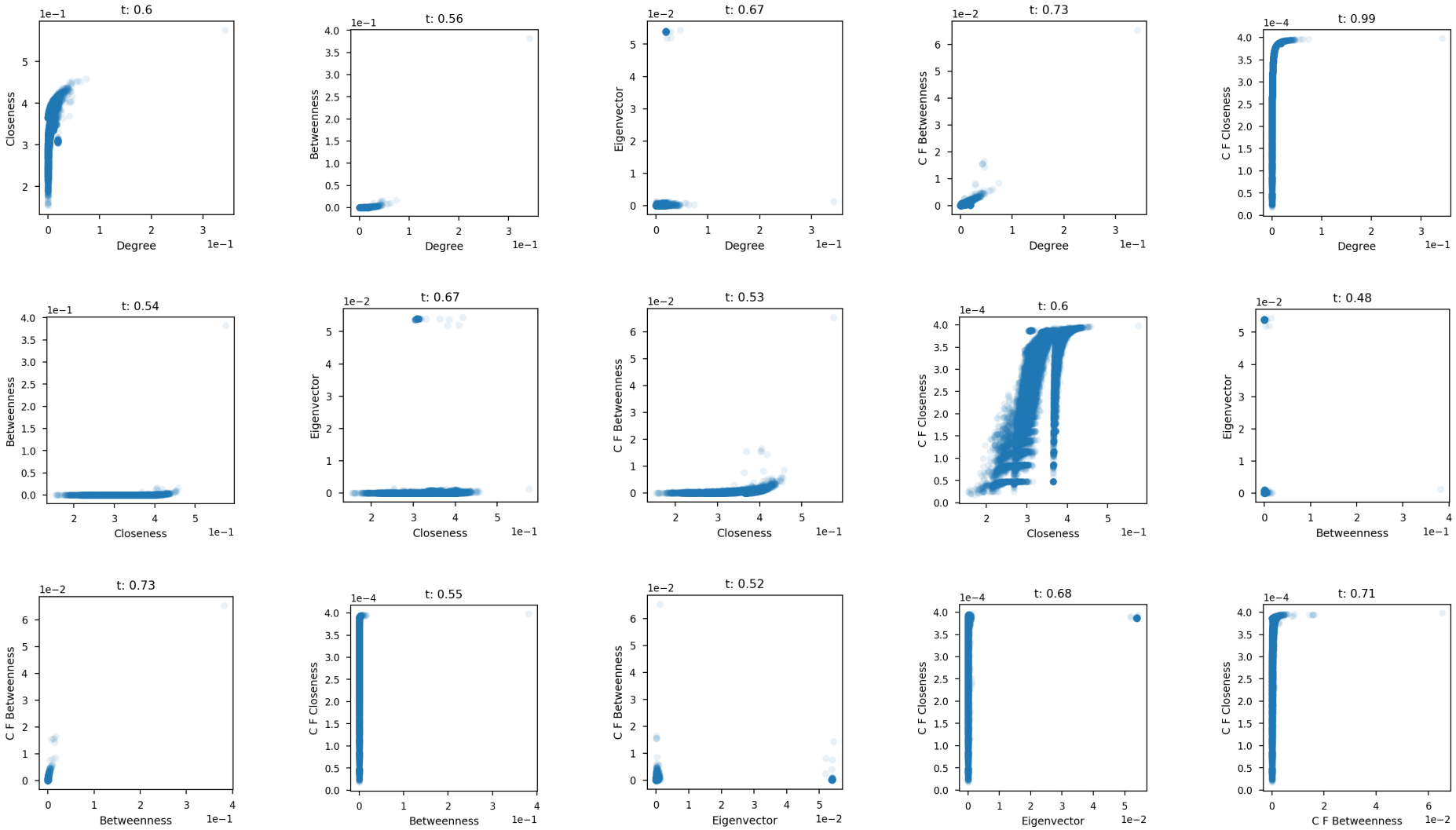


Figura 16 – Scatter-plots entre os pares de centralidade para o organismo *H. sapiens* com *threshold* de 0.5

Fonte: Elaborada pelo autor

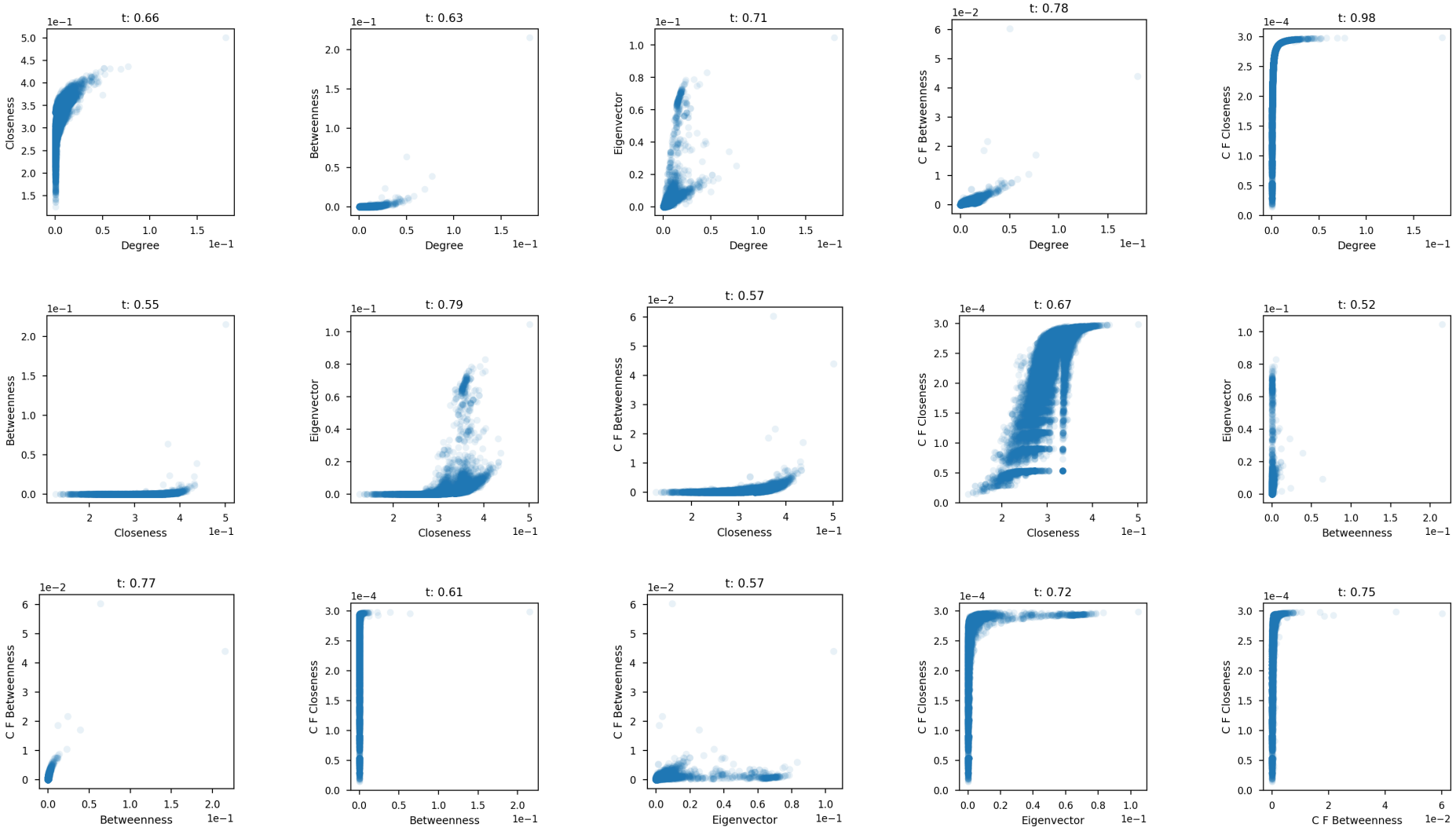


Figura 17 – Scatter-plots entre os pares de centralidade para o organismo *P. troglodytes* com *threshold* de 0.5

Fonte: Elaborada pelo autor

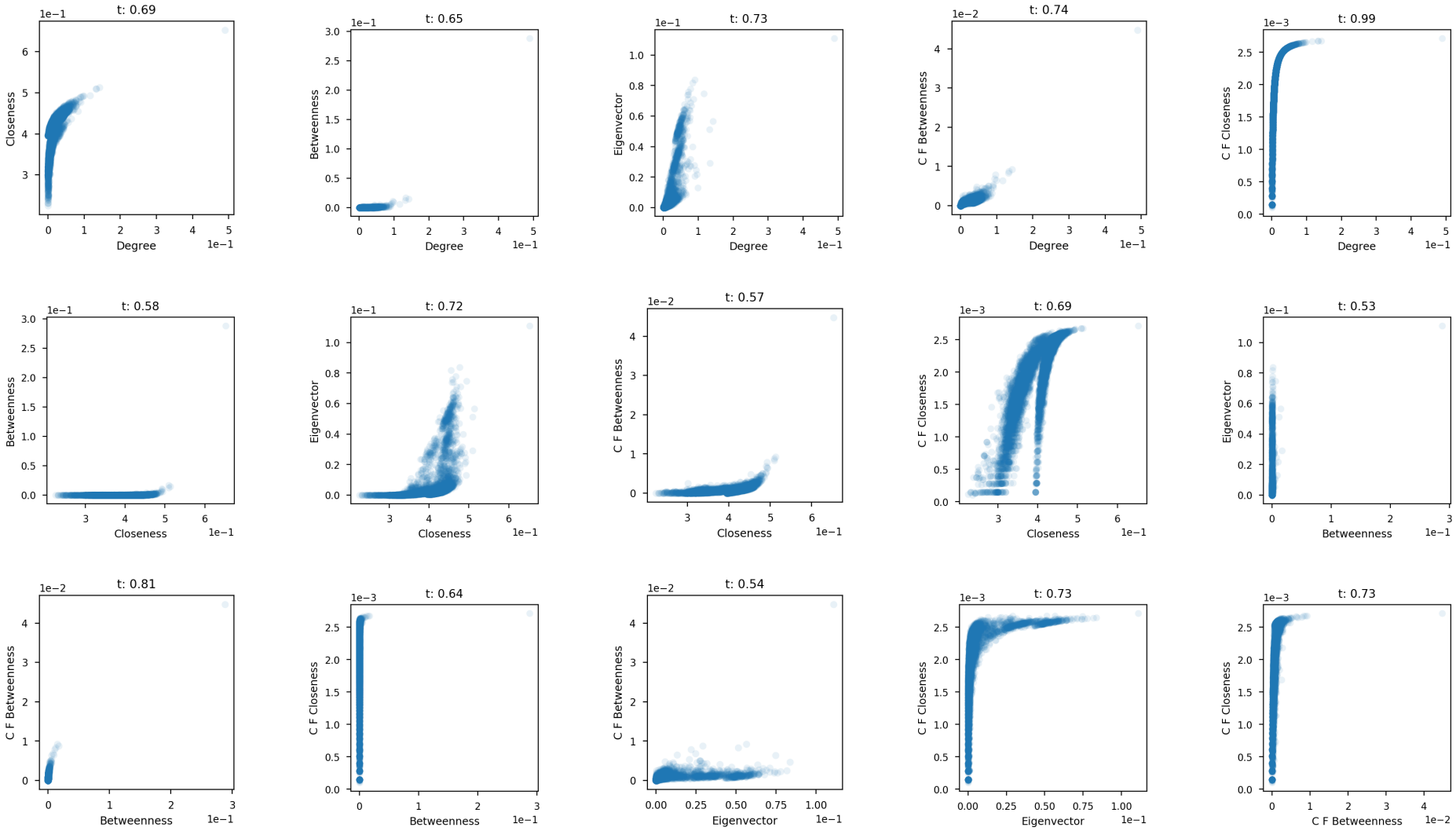


Figura 18 – Scatter-plots entre os pares de centralidade para o organismo *S. cerevisiae* com *threshold* de 0.5

Fonte: Elaborada pelo autor

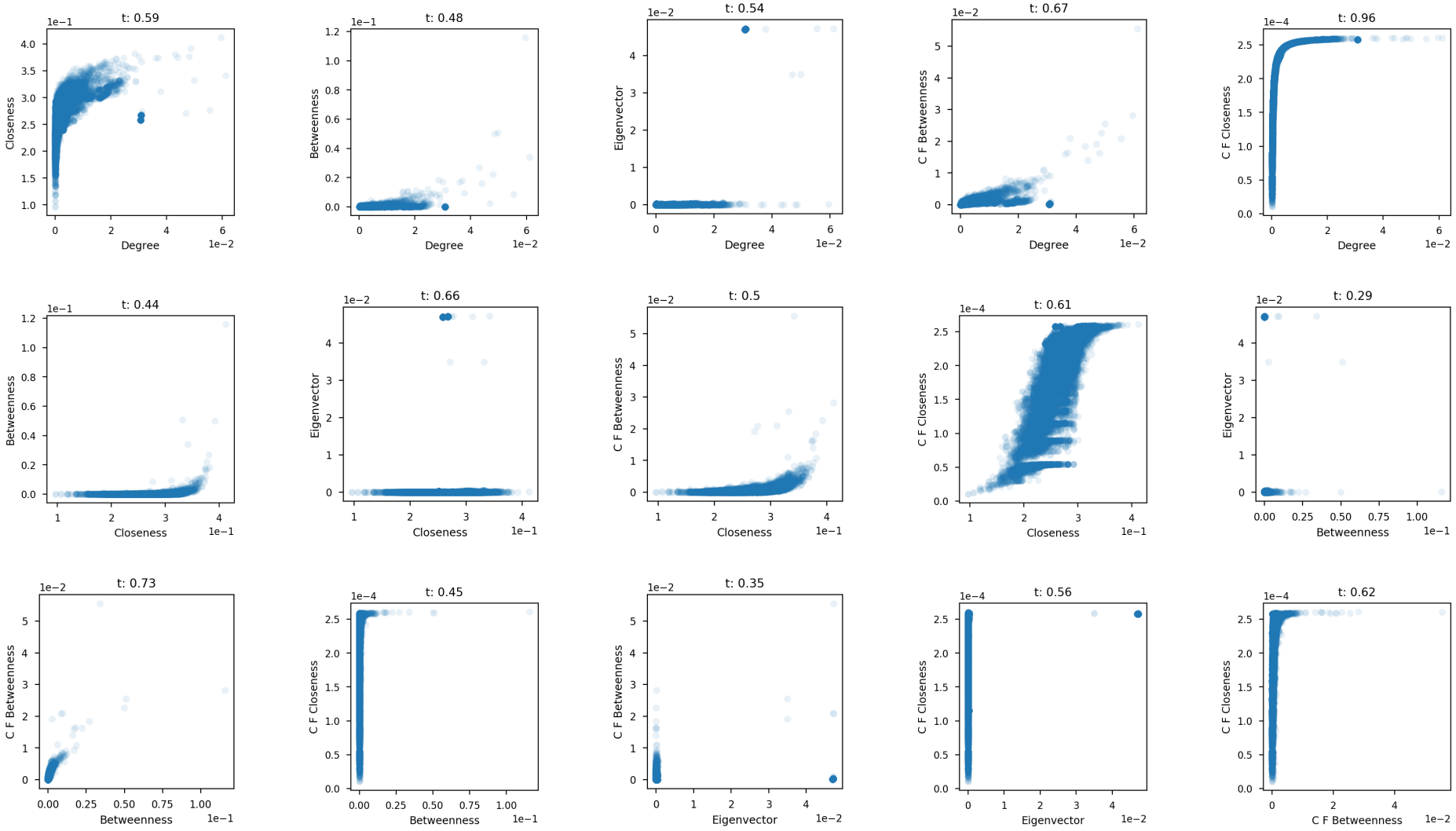


Figura 19 – Scatter-plots entre os pares de centralidade para o organismo *B. taurus* com *threshold* de 0.7

Fonte: Elaborada pelo autor

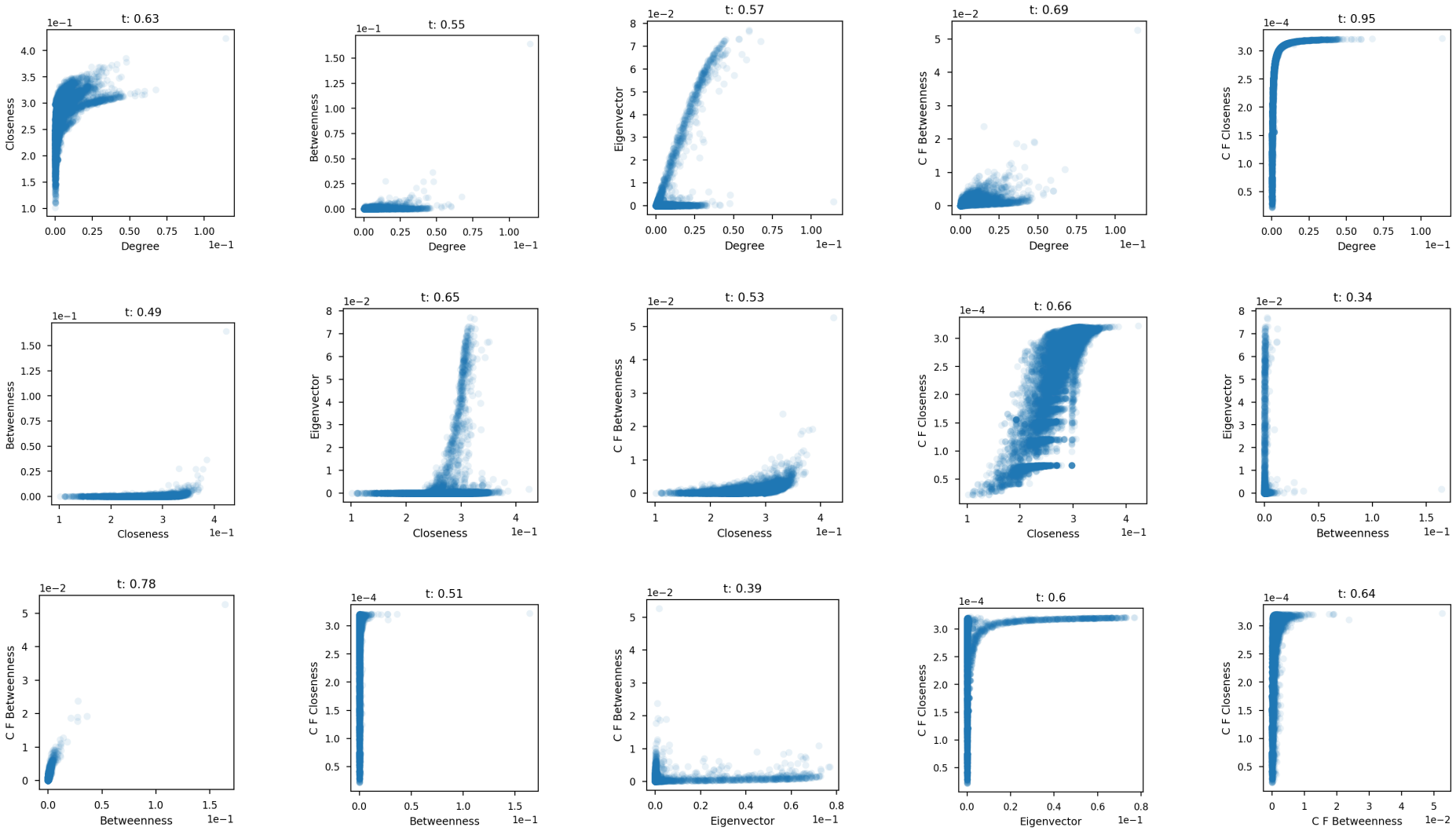


Figura 20 – Scatter-plots entre os pares de centralidade para o organismo *C. elegans* com *threshold* de 0.7

Fonte: Elaborada pelo autor

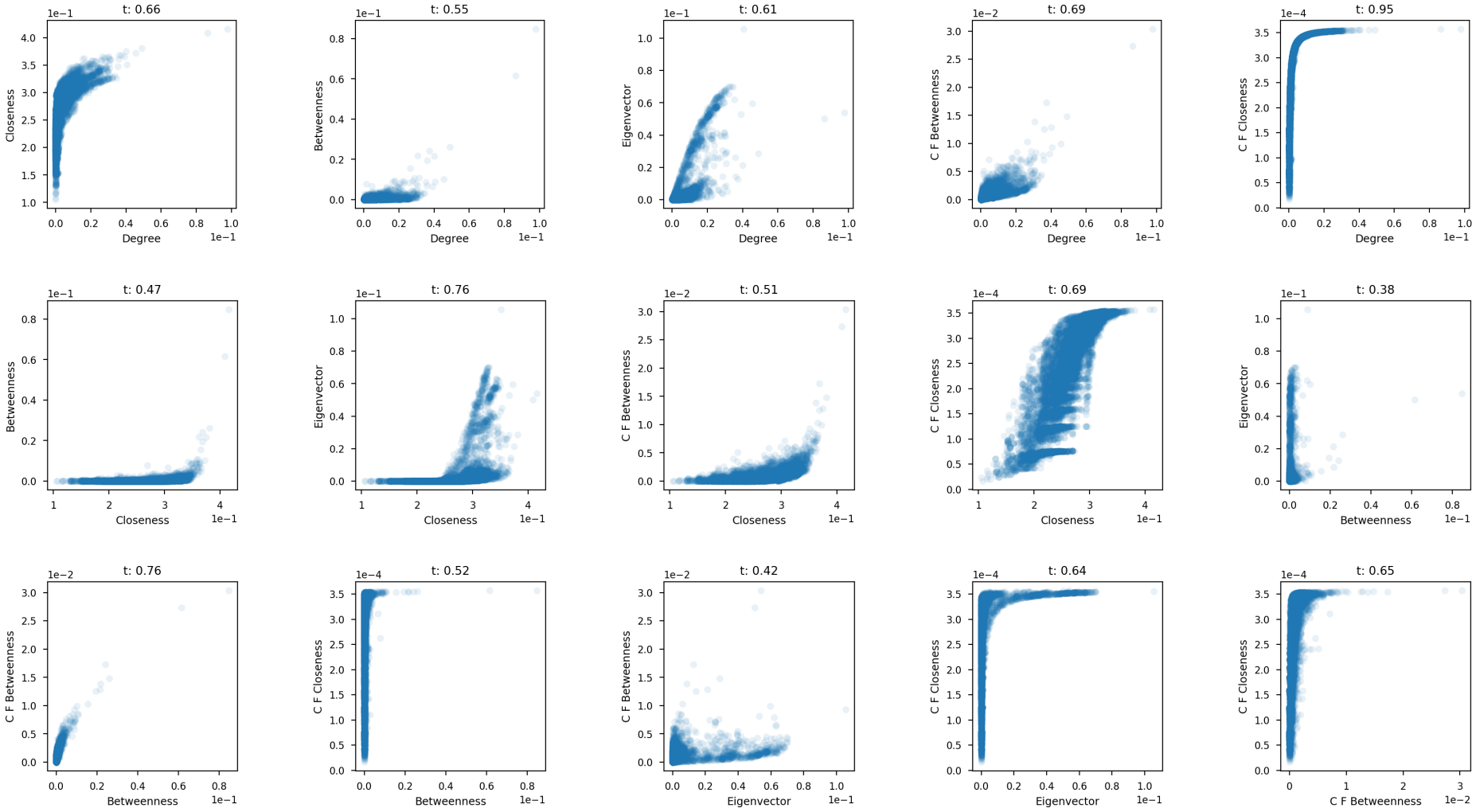


Figura 21 – Scatter-plots entre os pares de centralidade para o organismo *D. melanogaster* com *threshold* de 0.7

Fonte: Elaborada pelo autor

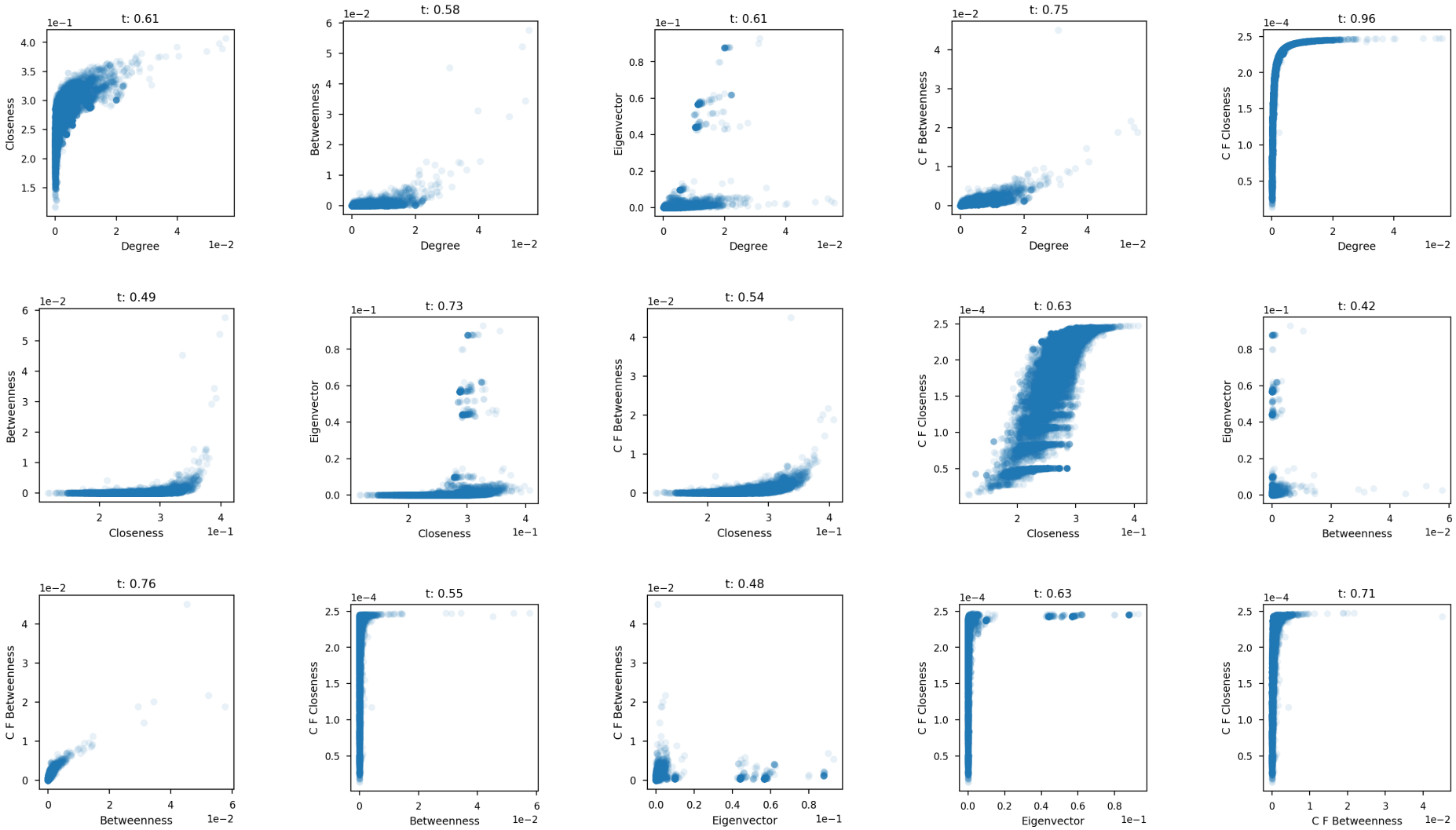


Figura 22 – Scatter-plots entre os pares de centralidade para o organismo *D. rerio* com *threshold* de 0.7

Fonte: Elaborada pelo autor

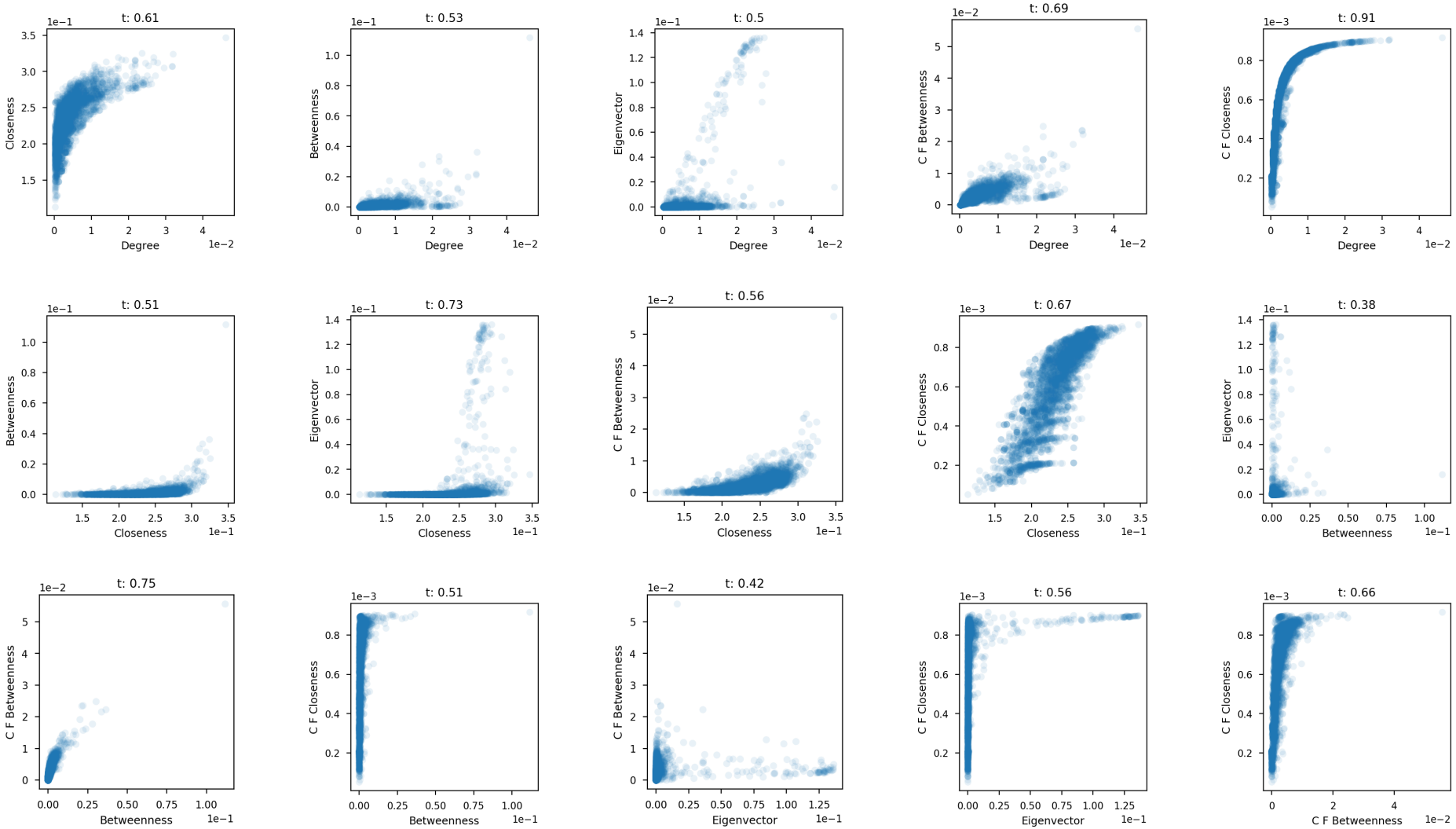


Figura 23 – Scatter-plots entre os pares de centralidade para o organismo *E. coli K12 W3110* com *threshold* de 0.7

Fonte: Elaborada pelo autor

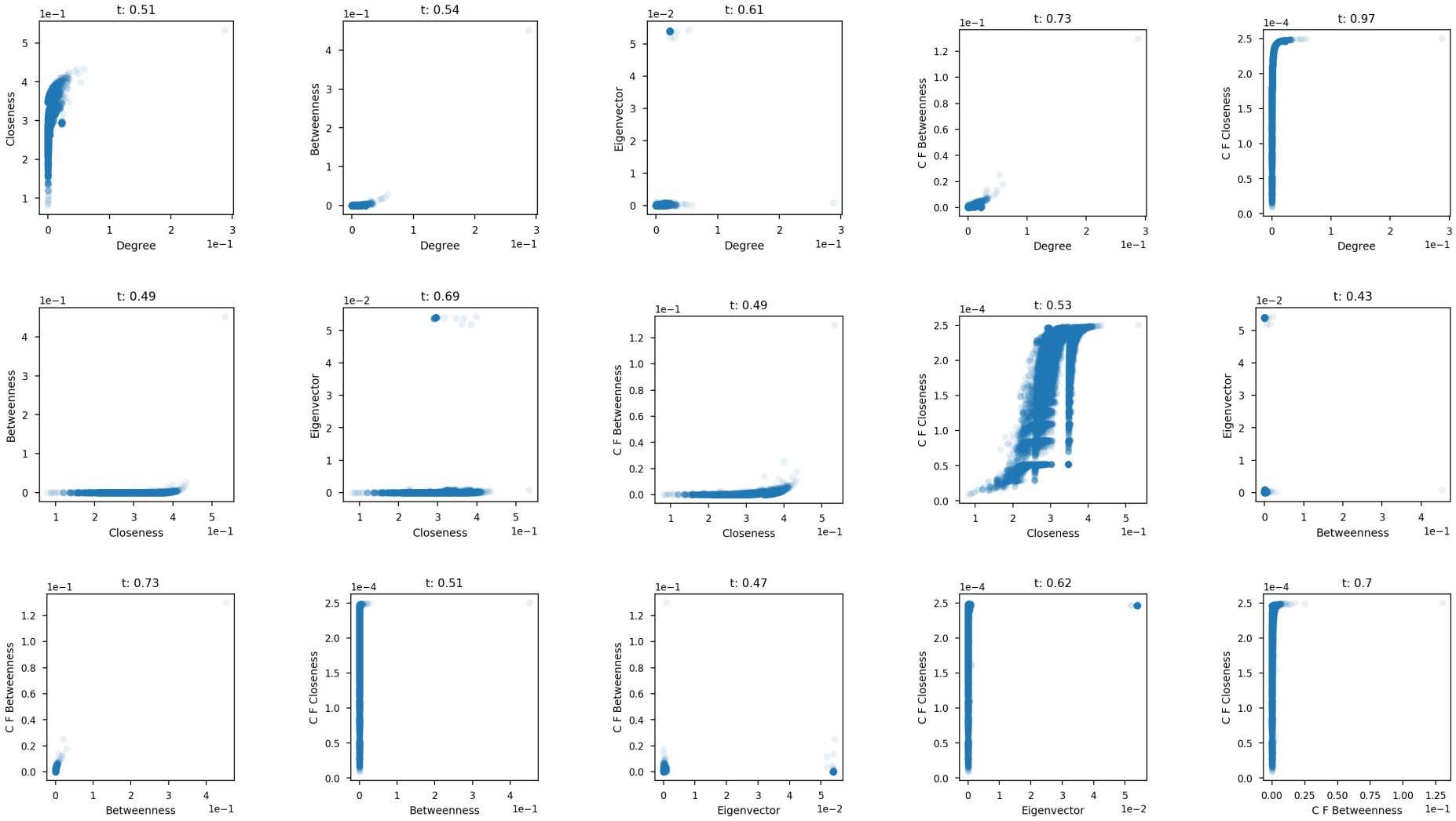


Figura 24 – Scatter-plots entre os pares de centralidade para o organismo *H. sapiens* com *threshold* de 0.7

Fonte: Elaborada pelo autor

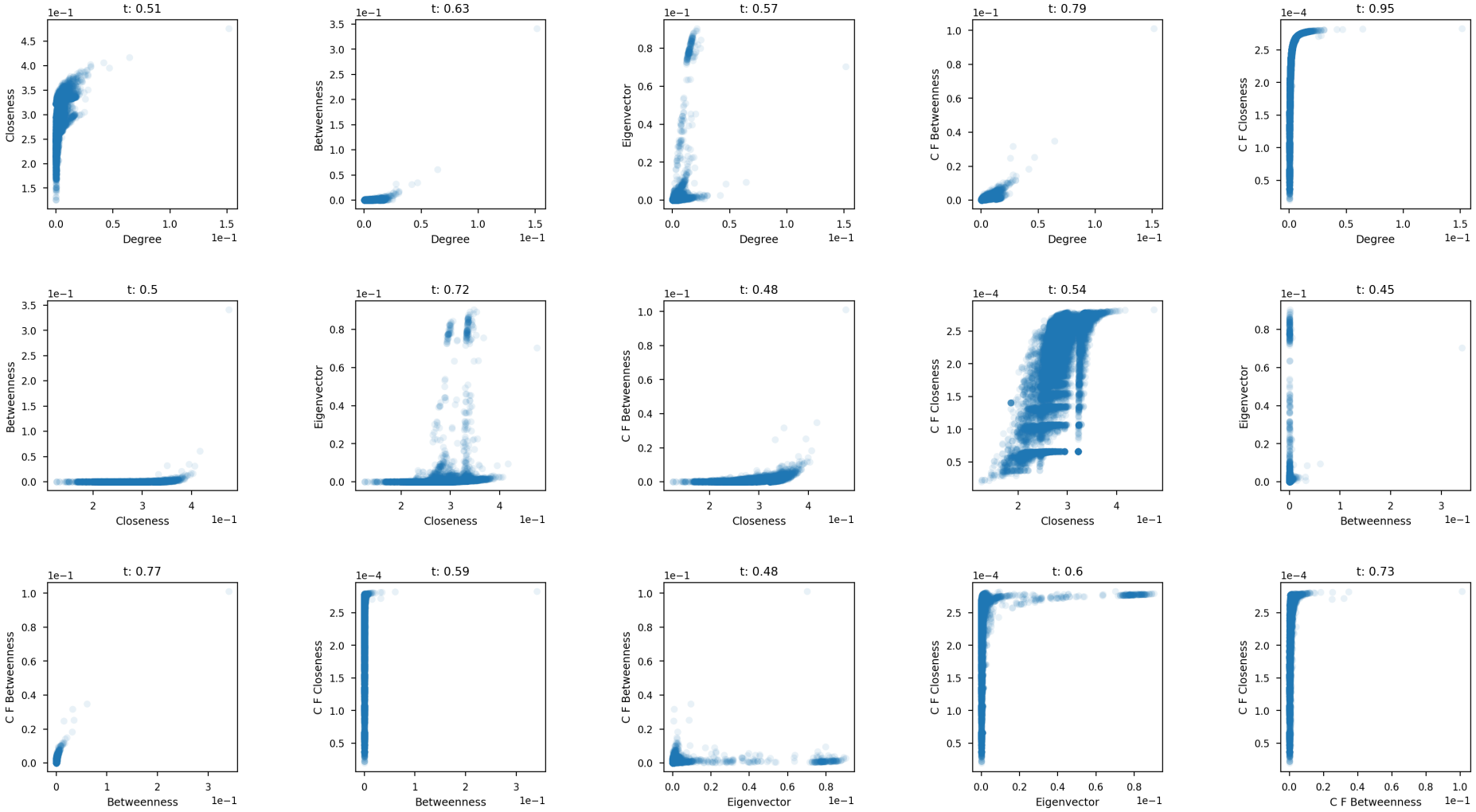


Figura 25 – Scatter-plots entre os pares de centralidade para o organismo *P. troglodytes* com *threshold* de 0.7

Fonte: Elaborada pelo autor

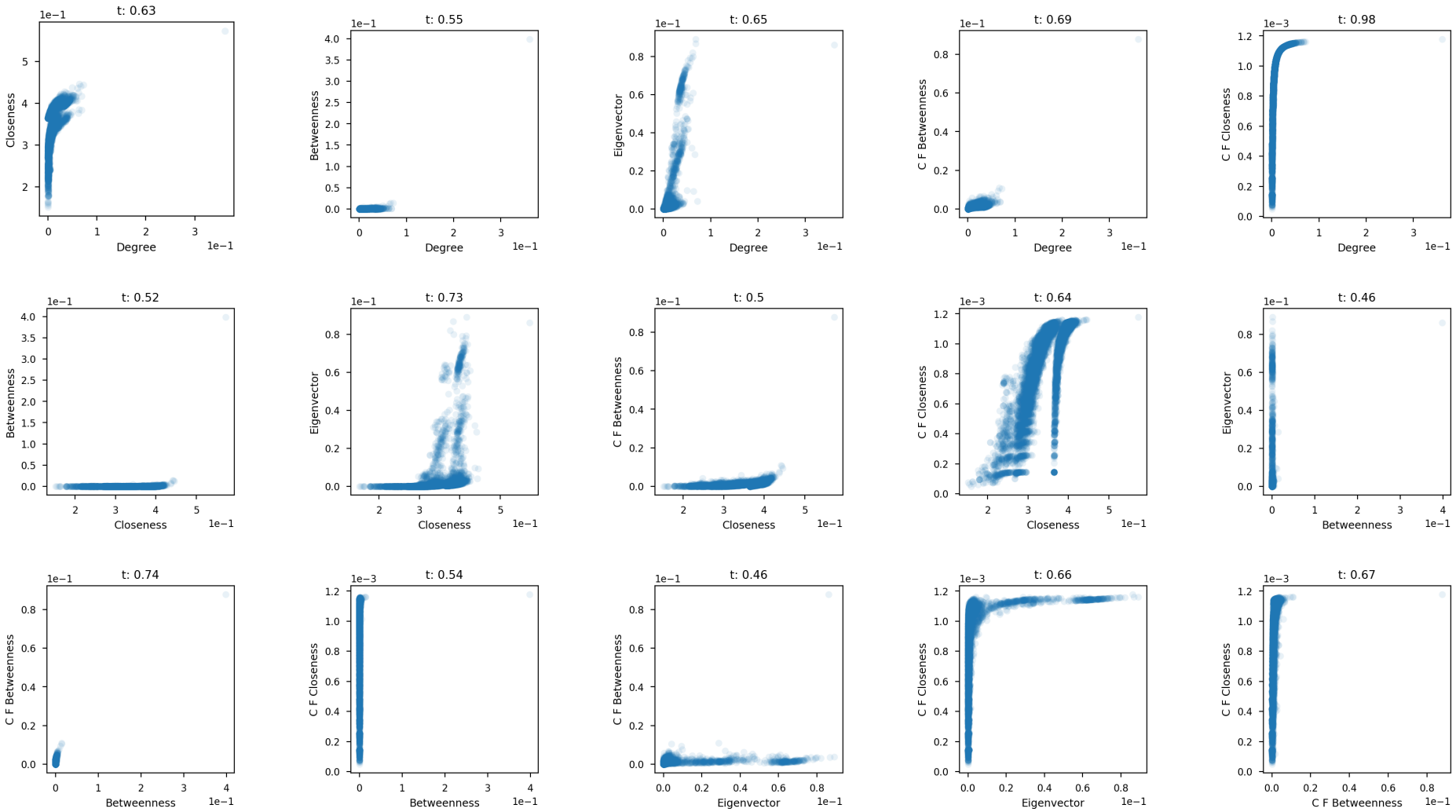


Figura 26 – Scatter-plots entre os pares de centralidade para o organismo *S. cerevisiae* com *threshold* de 0.7

Fonte: Elaborada pelo autor

Tabela 2 – Correlação de Kendall para todos os pares de medidas de centralidade. As colunas representam as redes PPI dos organismos. Em cada coluna o primeiro valor é refente ao *threshold* de 0.7 e o segundo (dentro dos parênteses) é para o *threshold* de 0.5. As abreviações para a coluna de centralidades significam: grau (D), proximidade (C), interposição (B), autovetor (E), interposição de fluxo de corrente (CFB) e proximidade de fluxo de corrente (CFC).

Centralidades	B. taurus	C. elegans	D. melanogaster	D. rerio	E. coli K12 W3110	H. sapiens	P. troglodytes	S. cerevisiae
D/C	0.59 (0.63)	0.63 (0.71)	0.66 (0.72)	0.61 (0.72)	0.61 (0.73)	0.51 (0.60)	0.51 (0.66)	0.63 (0.69)
D/B	0.48 (0.57)	0.55 (0.59)	0.55 (0.61)	0.58 (0.62)	0.53 (0.63)	0.54 (0.56)	0.63 (0.63)	0.55 (0.65)
D/E	0.54 (0.58)	0.57 (0.65)	0.61 (0.65)	0.61 (0.73)	0.50 (0.62)	0.61 (0.67)	0.57 (0.71)	0.65 (0.73)
D/CFB	0.67 (0.72)	0.69 (0.69)	0.69 (0.72)	0.75 (0.74)	0.69 (0.76)	0.73 (0.73)	0.79 (0.78)	0.69 (0.74)
D/CFC	0.96 (0.99)	0.95 (0.98)	0.95 (0.98)	0.96 (0.99)	0.91 (0.97)	0.97 (0.99)	0.95 (0.98)	0.98 (0.99)
C/B	0.44 (0.59)	0.49 (0.54)	0.47 (0.55)	0.49 (0.56)	0.51 (0.62)	0.49 (0.54)	0.50 (0.55)	0.52 (0.58)
C/E	0.66 (0.62)	0.65 (0.71)	0.76 (0.74)	0.73 (0.80)	0.73 (0.76)	0.69 (0.67)	0.72 (0.79)	0.73 (0.72)
C/CFB	0.50 (0.60)	0.53 (0.57)	0.51 (0.57)	0.54 (0.59)	0.56 (0.64)	0.49 (0.53)	0.48 (0.57)	0.50 (0.57)
C/CFC	0.61 (0.64)	0.66 (0.72)	0.69 (0.73)	0.63 (0.72)	0.67 (0.74)	0.53 (0.60)	0.54 (0.67)	0.64 (0.69)
B/E	0.29 (0.40)	0.34 (0.41)	0.38 (0.45)	0.42 (0.50)	0.38 (0.48)	0.43 (0.48)	0.45 (0.52)	0.46 (0.53)
B/CFB	0.73 (0.78)	0.78 (0.80)	0.76 (0.79)	0.76 (0.81)	0.75 (0.80)	0.73 (0.73)	0.77 (0.77)	0.74 (0.81)
B/CFC	0.45 (0.56)	0.51 (0.57)	0.52 (0.60)	0.55 (0.60)	0.51 (0.61)	0.51 (0.55)	0.59 (0.61)	0.54 (0.64)
E/CFB	0.35 (0.43)	0.39 (0.45)	0.42 (0.47)	0.48 (0.55)	0.42 (0.50)	0.47 (0.52)	0.48 (0.57)	0.46 (0.54)
E/CFC	0.56 (0.58)	0.60 (0.66)	0.64 (0.66)	0.63 (0.74)	0.56 (0.63)	0.62 (0.68)	0.60 (0.72)	0.66 (0.73)
CFB/CFC	0.62 (0.71)	0.64 (0.66)	0.65 (0.70)	0.71 (0.72)	0.66 (0.74)	0.70 (0.71)	0.73 (0.75)	0.67 (0.73)

Fonte: Elaborada pelo autor

3.2.1 Comparação com os modelos de Erdős-Rényi e Barabási-Albert

Nós também investigamos o comportamento das medidas de centralidade para redes geradas com os modelos de Erdős-Rényi e Barabási-Albert a fim de compará-las com os resultados obtidos para as redes PPI estudadas. Para as redes geradas com os modelos, utilizamos como parâmetro o organismo *C. elegans* com o *threshold* de 0.7, ou seja, as redes geradas possuem número de vértices e valor de grau médio aproximadamente iguais ao do organismo modelo. Para cada modelo foram geradas 10 redes e para cada uma delas calculamos as mesmas medidas de centralidade utilizadas paras os organismos e geramos os gráficos do tipo *scatter-plot*.

Os resultados obtidos para cada modelo são apresentados nas Figuras 27 e 28. Observando as Figuras, notamos que para as redes geradas pelos modelos não existe grande variação na maneira como os pontos se distribuem, e que o padrão do espalhamento dos pontos para um modelo específico é bastante característico. O segundo ponto que destacamos é que existem diferenças claras na maneira como as centralidades se correlacionam entre os dois modelos considerados. De maneira similar, comparando a forma como os pontos se distribuem para os pares de centralidades modelos com os organismos, observamos que o formato obtido para os dois também é distinto daqueles observados para as redes PPI. Entre os dois modelos o que mais se aproximou das redes de interação entre proteínas é o BA — por exemplo compare o par grau e proximidade de fluxo de corrente, — mas apesar disso ainda existem diferenças claras entre o modelo e os organismos — como no caso das centralidades de autovetor e interposição de fluxo de corrente. Apesar disso precisamos ter em mente que ambos os modelos utilizados aqui não apresentam assortatividade de grau, e esses resultados podem ser diferentes para modelos que possuam esta característica. Ambos os modelos apresentam valores de correlação entre as centralidades mais elevados do que os obtidos para as redes PPI. Apesar disso as correlações para o modelo BA também estão mais próximas do valores das redes reais do que as do modelo ER; mesmo assim ainda existem casos onde a diferença é grande como para o par formado por interposição e autovetor. Isso nos indica que os padrões observados nas correlações entre medidas são típicos de redes PPI, não se tratando apenas de correlações ordinárias entre as centralidades.

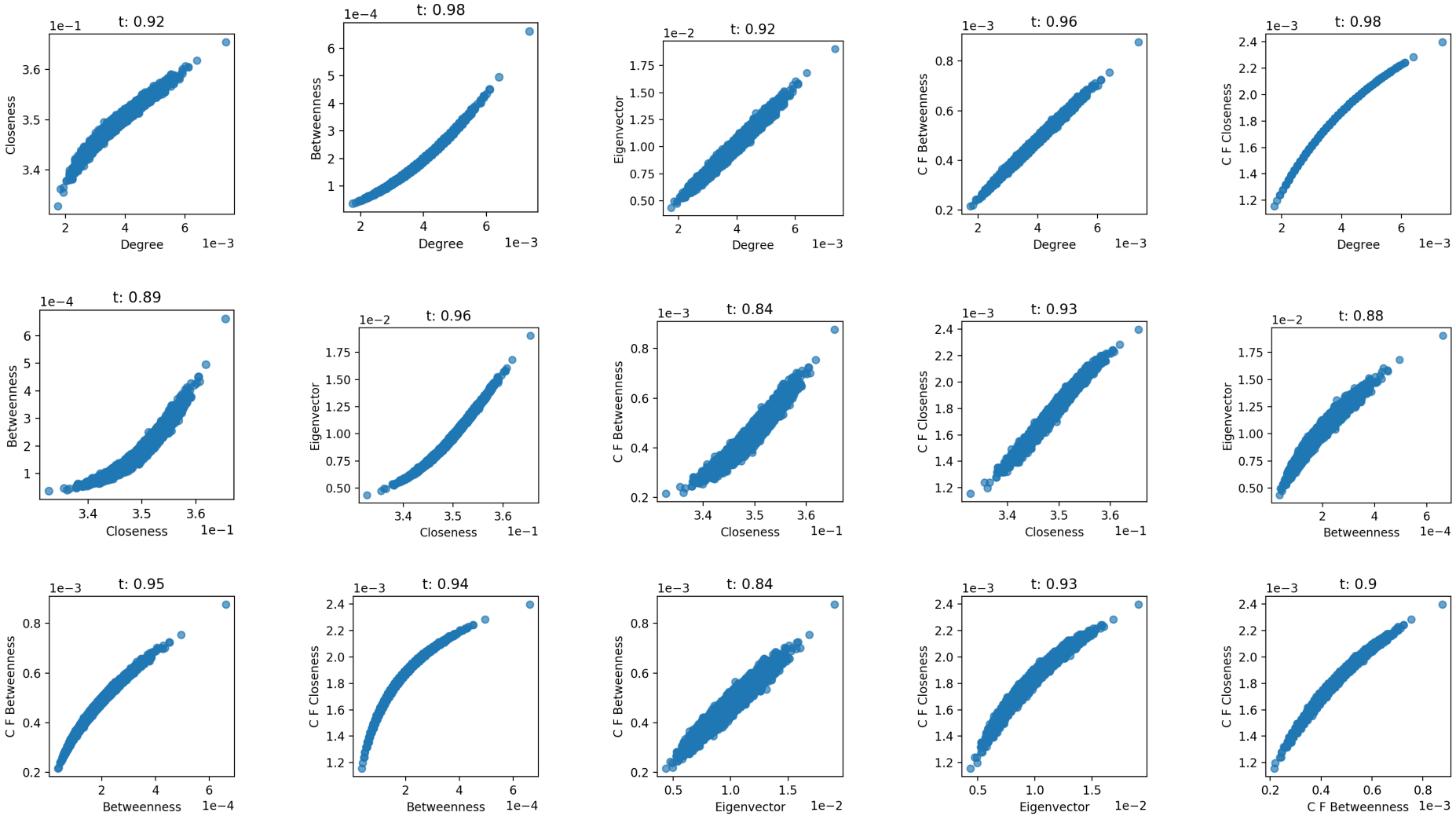


Figura 27 – Scatter-plots entre os pares de centralidade para 10 redes geradas com o modelo de Erdős-Rényi

Fonte: Elaborada pelo autor

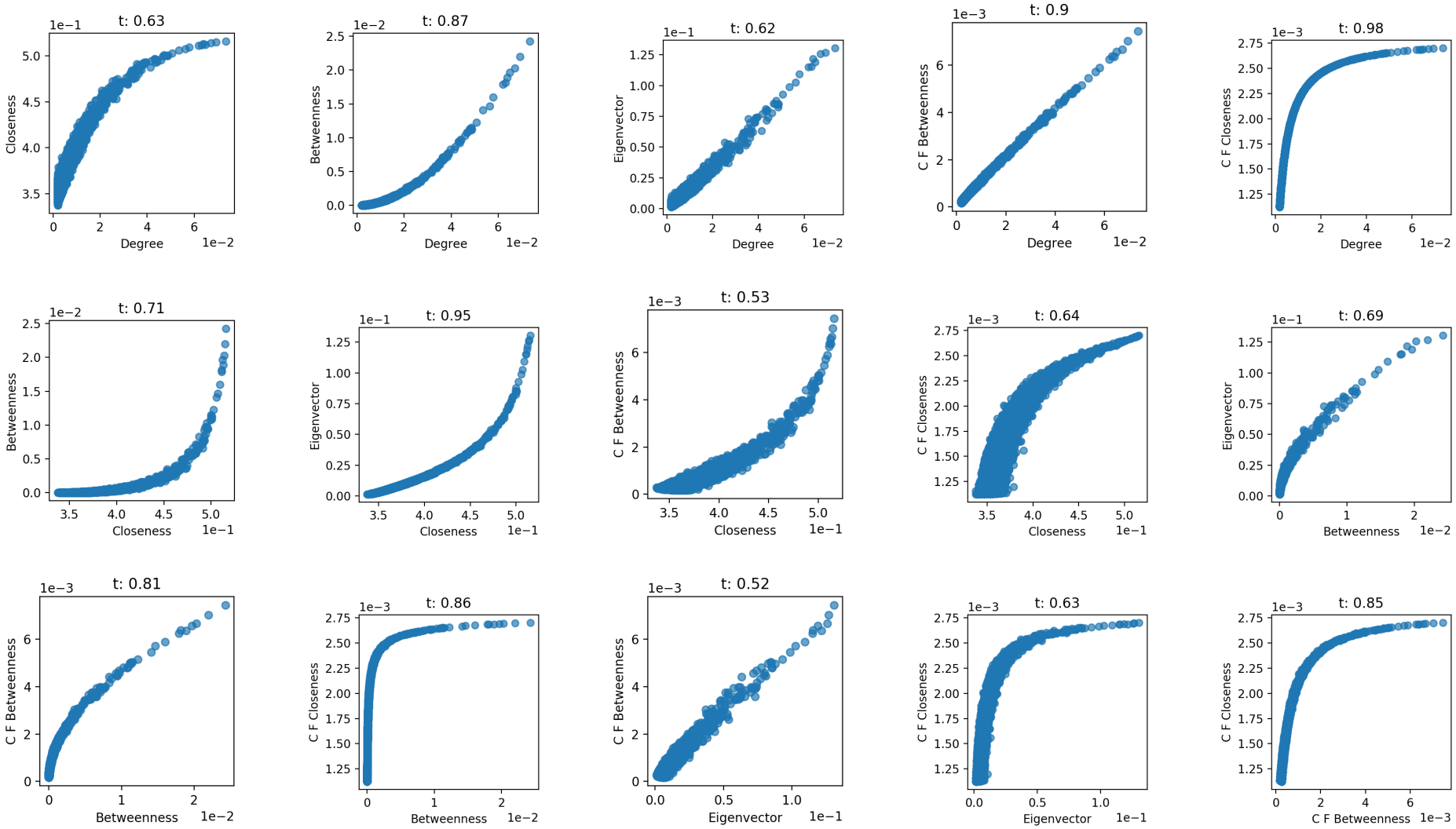


Figura 28 – Scatter-plots entre os pares de centralidade para 10 redes geradas com o modelo de Barabási-Albert

Fonte: Elaborada pelo autor

3.3 Medida de centralidade para *B. taurus* e *H. sapiens* — indícios da localização de autovetor

Desde o início dos estudos em redes complexas muita pesquisa foi desenvolvida com intuito de entender-se as topologia das redes através das propriedades espectrais das matrizes de adjacências e Laplaciana,[43] já que elas mostraram estar associadas as propriedades topológicas da rede. Alguns trabalhos enfatizam o entendimento das propriedades espectrais de redes heterogêneas.[90–93] Além de sua importância no entendimento da topologia da rede e na categorização de seus vértices, o autovetor principal (AP) — autovetor associado ao maior autovalor da matriz de adjacências — também está associado a processos dinâmicos em redes complexas como por exemplo sincronização[94] e no estudo de propagação de epidemias.[95,96]

Como podemos observar na seção 3.2, a centralidade de autovetor apresenta um comportamento curioso para *B. taurus* e *H. sapiens* que mostram distribuições similares. Nos gráficos apresentados nas Figuras 11, 16, 19 e 24 podemos verificar que a medida de autovetor classifica apenas alguns vértices como importantes, com todos os demais recebendo um *ranking* baixo. Considerando por exemplo o par grau e autovetor (D/E), notamos que apenas alguns poucos vértices receberam valores elevados da medida de autovetor — e curiosamente esses vértices não são os que recebem maior grau. Essa característica é diferente do que acontece nos demais organismos onde o par grau e autovetor (D/E) apresenta dois conjuntos de vértices — um onde grau e autovetor possuem uma correlação linear e outro onde as duas medidas não são correlacionadas. Esse comportamento diferenciado nos levou a uma análise mais detalhada da medida de autovetor para os organismos considerados.

O fenômeno que acontece nos organismos *B. taurus* e *H. sapiens*, onde todo o peso da medida de autovetor fica concentrado em apenas um subconjunto dos vértices da rede, parece ser um caso de localização de centralidade de autovetor[97] que pode ocorrer tanto no maior *hub* da rede,[98,99] quanto no maior K-core[100] ou uma combinação de ambas as condições anteriores.[101] Quando ocorre o fenômeno de localização isto indica que utilizar unicamente centralidade de autovetor para caracterização dos vértices da rede não é uma boa escolha, pois todo o peso da centralidade fica localizado em um subconjunto dos vértices da rede, tornando a comparação com os demais injusta. Pastor-Satorras e Castellano[100,101] mostraram que para redes que seguem uma distribuição de graus do tipo lei de potência, a localização acontece no maior *hub* da rede se a heterogeneidade for média ($\gamma > 5/2$) como previsto por trabalhos anteriores,[97,98] porém quando a rede possui distribuições de grau altamente heterogêneas ($\gamma < 5/2$) os autores mostram que a localização acontece no maior K-core da rede.

A definição de localização da centralidade de autovetor para modelos de redes[97,100] depende de *ensembles* de redes com quantidades distintas de vértices e da medida conhecida

como razão de participação inversa (IPR) que é definida[97,99] como:

$$Y_\lambda = \sum_i {f_i}^4(\lambda) \tag{3.1}$$

onde Y_λ é a razão de participação inversa do autovalor λ e ${f_i}^4(\lambda)$ são as componentes i do autovetor associado ao autovalor λ elevadas a quarta portência. Para analisar se o autovetor é localizado em um subconjunto de vértices da rede, podemos analisar se o comportamento do IPR com o tamanho do sistema segue uma lei de potência:

$$Y_\lambda(N) \approx N^{-\alpha} \tag{3.2}$$

Se α é igual a 1 isso indica que não existe localização do autovetor, se α for menor que 1 isso é indicativo que algum tipo de localização está acontecendo.[100]

Para redes reais, onde não é possível gerar *ensembles* de redes com tamanhos diferentes, Satorras analisa o comportamento da medida de autovetor com relação ao K-core ao qual o vértice pertence. Outra característica citada pelo autor — apesar de ele não conseguir observá-la nas redes reais estudadas — é a relação de linearidade entre os valores do autovetor e o grau dos vértices para redes geradas com o modelo configuracional descorralacionado[102] (uncorrelated configurational model) — fenômeno observado para alguns vértices nas redes de todos organismos com exceção de *B. taurus* e *H. sapiens*. De maneira similar aos artigos que analisam redes reais criamos *scatter-plots* de todas as redes estudadas. Os gráficos são apresentados nas Figuras 29 e 30. Em cada gráfico do tipo *scatter-plot* dessas Figuras, os pontos representam os vértices da rede, no eixo horizontal apresentamos o K-core aos quais os mesmos pertencem e no eixo vertical o quadrado do valor da centralidade de autovetor dos vértices.

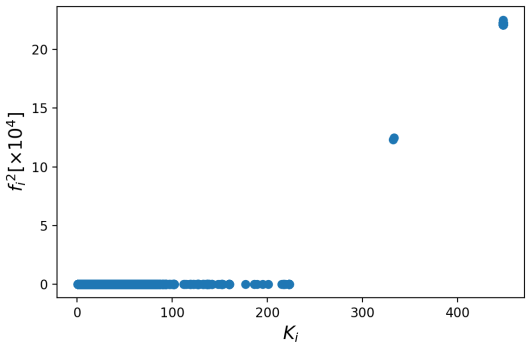
Inicialmente podemos realizar a comparação das redes do mesmo organismo para os diferentes valores de *threshold*. Como podemos observar, alguns organismos são mais afetados pela mudança de threshold do que os outros, como por exemplo *D. melanogaster*, *D. rerio*, *E. coli*, *P. troglodytes* e *S. cerevisiae*. Para o organismo *D. melanogaster*, por exemplo, observamos uma mudança no comportamento dos vértices de K-core médio-altos, onde para um threshold menor, o valor da componente do maior autovetor cresce de maneira mais acentuada do que para o maior *threshold* — ressaltando que nos dois casos os maiores valores estão claramente concentrados no maior K-core. Para o *D. rerio*, o comportamento muda bastante nos dois casos: enquanto que na rede de *threshold* menor os maiores valores de componentes de autovetor estão concentrados próximos do K-core 100, a situação para o caso de *threshold* 0.7 tende a concordar com o descrito por Satorras, com os maiores valores ficando concentrados nos dois maiores K-cores da rede. Para a *E. coli*, em ambos os casos os maiores valores das componentes do autovetor estão concentrados no maior K-core da rede, o que muda é a dispersão dos pontos na região dos K-cores mais elevados que é menor para a rede de maior *threshold*. No organismo *P. troglodytes*

os maiores valores estão concentrados no maior K-core da rede independentemente do *threshold*, a diferença é a maneira como os valores aumentam para os K-cores médio-altos, que é mais acentuada para o *threshold* maior. O organismo *S. cerevisiae* apresentou um comportamento diferente dos demais: para ambos os *thresholds* um boa parte dos vértices associados aos maiores valores da centralidade de autovetor estão associados ao maior K-core da rede, entretanto existe um segundo valor de K-core elevado, — próximo de 137 para o *threshold* menor e de 108 no caso do *threshold* maior — que também apresenta valores elevados das componentes do maior autovetor. *C. elegans* aparentemente não apresentou mudanças significativas no comportamento de seus gráficos com a mudança de *threshold*. Para os organismos *B. taurus* e *H. sapiens* é claro que independentemente do *threshold* escolhido o peso da medida de autovetor está concentrado nos vértices pertencentes aos dois maiores K-cores da rede. Para a maioria dos casos analisados os resultados concordam com os obtidos por Satorras para os casos de alta heterogeneidade das redes complexas, especialmente nos casos dos maiores *thresholds* onde a certeza da estrutura das redes dos organismos é maior.

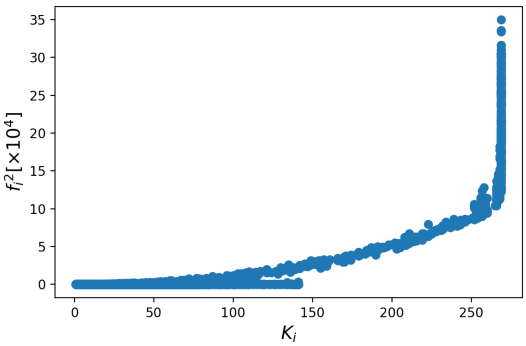
Goltsev[97] também mostrou que o maior autovalor da matriz de ajacências é dado pelo máximo entre $\sqrt{k_{max}}$ e $\langle k^2 \rangle / \langle k \rangle$ para redes que possuem uma distribuição de graus que segue uma lei de potência $P(k) \sim k^{-\gamma}$ e que neste caso, $\lambda_1 \sim \sqrt{k_{max}}$ se $\gamma > 5/2$ e que $\lambda_1 \sim \langle k^2 \rangle / \langle k \rangle$ quando $\gamma < 5/2$. Assim, outra análise que também é realizada por Satorras para redes reais[100] é a comparação do maior autovalor da matriz de adjacências com os dois valores citados anteriormente; devido a correlações não triviais de grau existentes nas redes reais uma análise da distância ou proximidade de λ_1 com os valores anteriores é mais adequada.

Realizamos esta mesma comparação para as redes dos organismos considerados. Os resultados obtidos são apresentados na Tabela 3. Primeiro podemos observar que $\sqrt{k_{max}}$ e a razão entre o segundo momento da distribuição de graus e o grau médio da rede possuem valores significativamente distintos, indicando que deve ser relativamente fácil diferenciar os dois casos possíveis. Em seguida, notamos que o maior autovalor das matrizes de adjacências para todos os organismos e independentemente do *threshold* escolhido encontra-se muito mais próximo da razão entre $\langle k^2 \rangle$ e $\langle k \rangle$ do que do valor de $\sqrt{k_{max}}$. Esses resultados concordam com o os gráficos das Figuras 29 e 30 que nos indicam que a localização que acontece nas redes de interação de proteínas acontece no maior K-core da rede e não no maior hub das redes, característica de redes altamente heterogêneas.

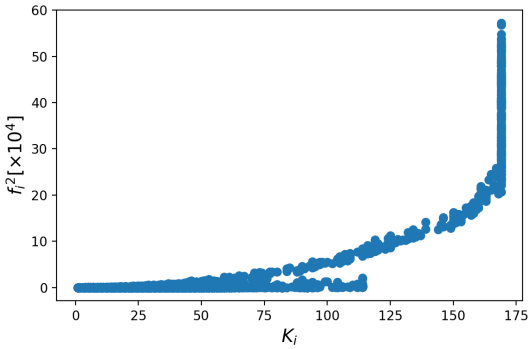
Na sequência, decidimos explorar em maiores detalhes o que acontece para as redes de *B. taurus* e *H. sapiens* que apresentam um comportamento mais acentuado de localização. O maior K-core de redes com grau médio elevado como é o caso das redes anterior deve ser algo muito similar a um grafo completo — onde sempre existe uma conexão entre quaisquer dois vértices pertencentes ao grafo. Verificando os dois maiores



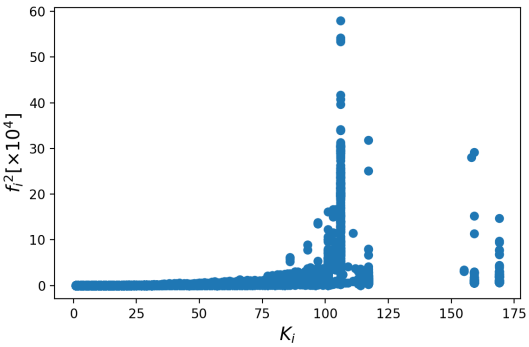
(a) B. taurus



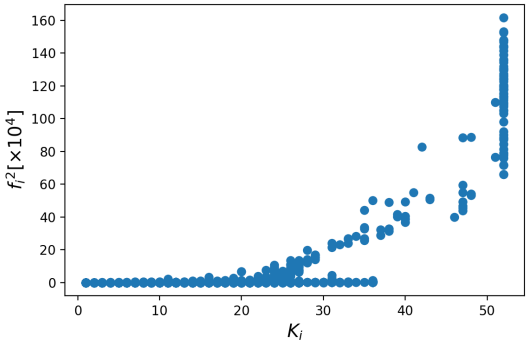
(b) C. elegans



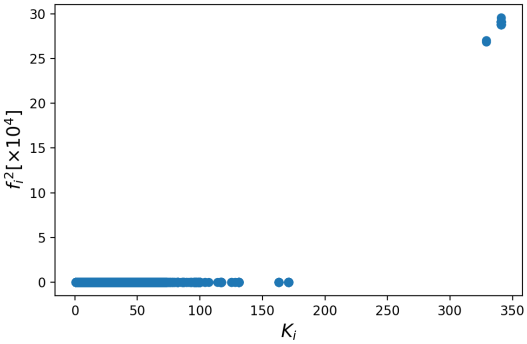
(c) D. melanogaster



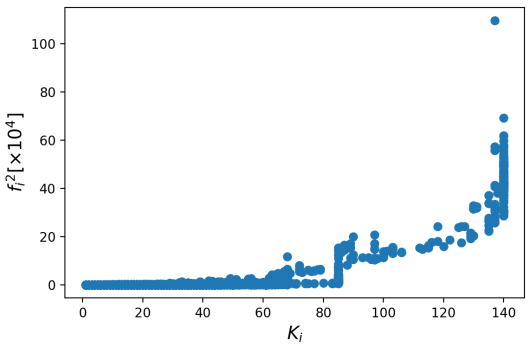
(d) D. rerio



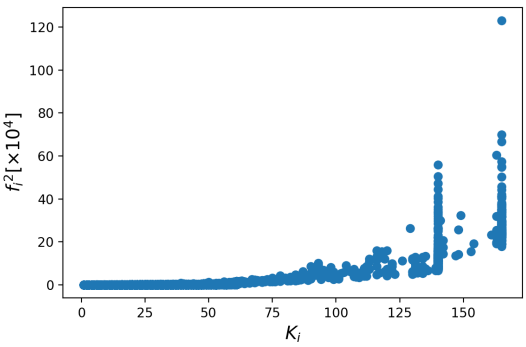
(e) E. coli



(f) H. sapiens



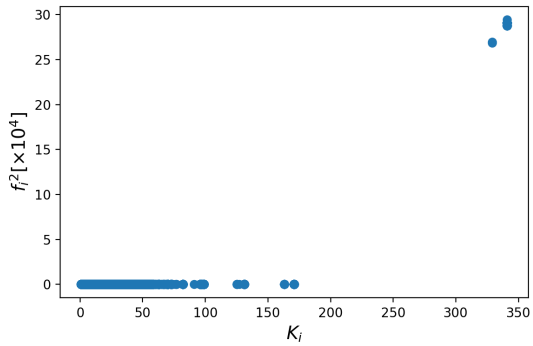
(g) P. troglodytes



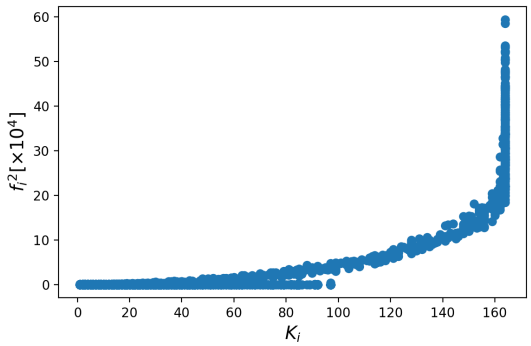
(h) S. cerevisiae

Figura 29 – Scatter-plots dos valores de K-core e autovetor para os vértices das redes com *threshold* de 0.5

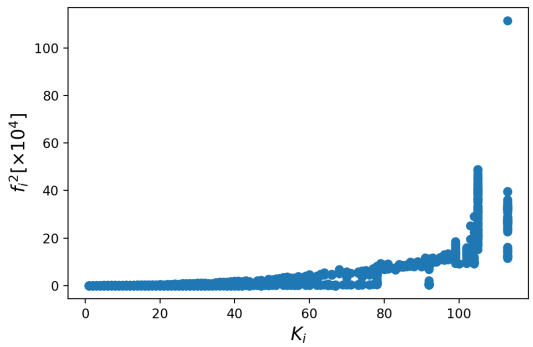
Fonte: Elaborada pelo autor



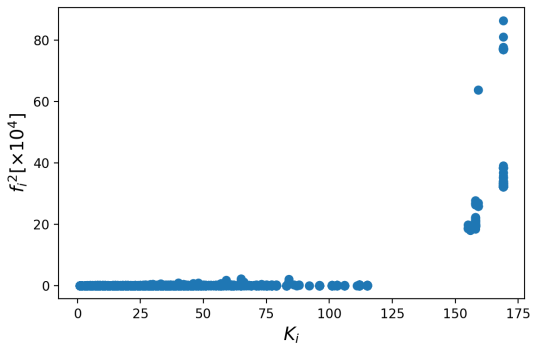
(a) B. taurus



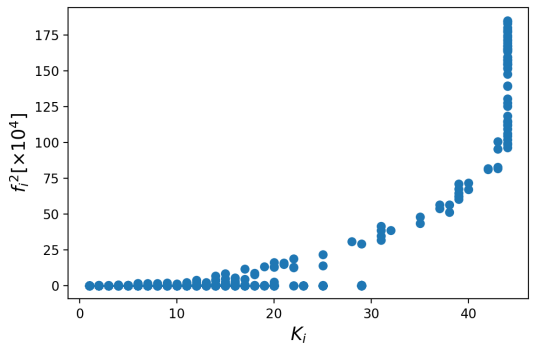
(b) C. elegans



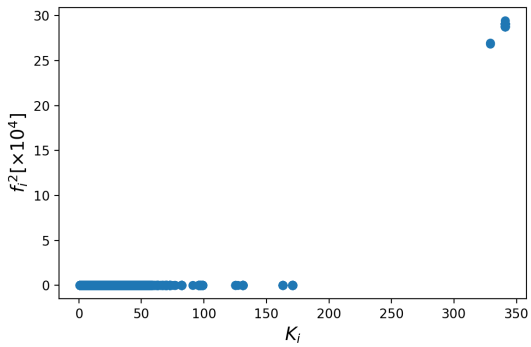
(c) D. melanogaster



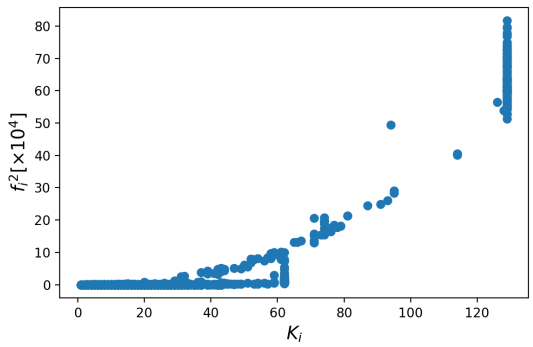
(d) D. rerio



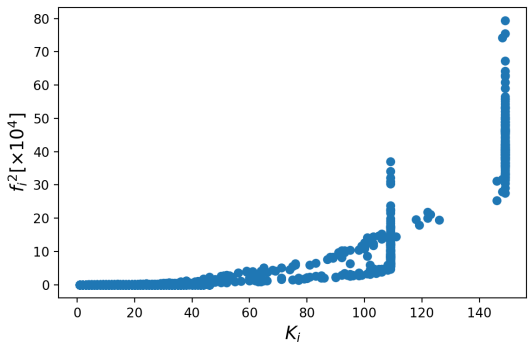
(e) E. coli



(f) H. sapiens



(g) P. troglodytes



(h) S. cerevisiae

Figura 30 – Scatter-plots dos valores de K-core e autovetor para os vértices das redes com *threshold* de 0.7

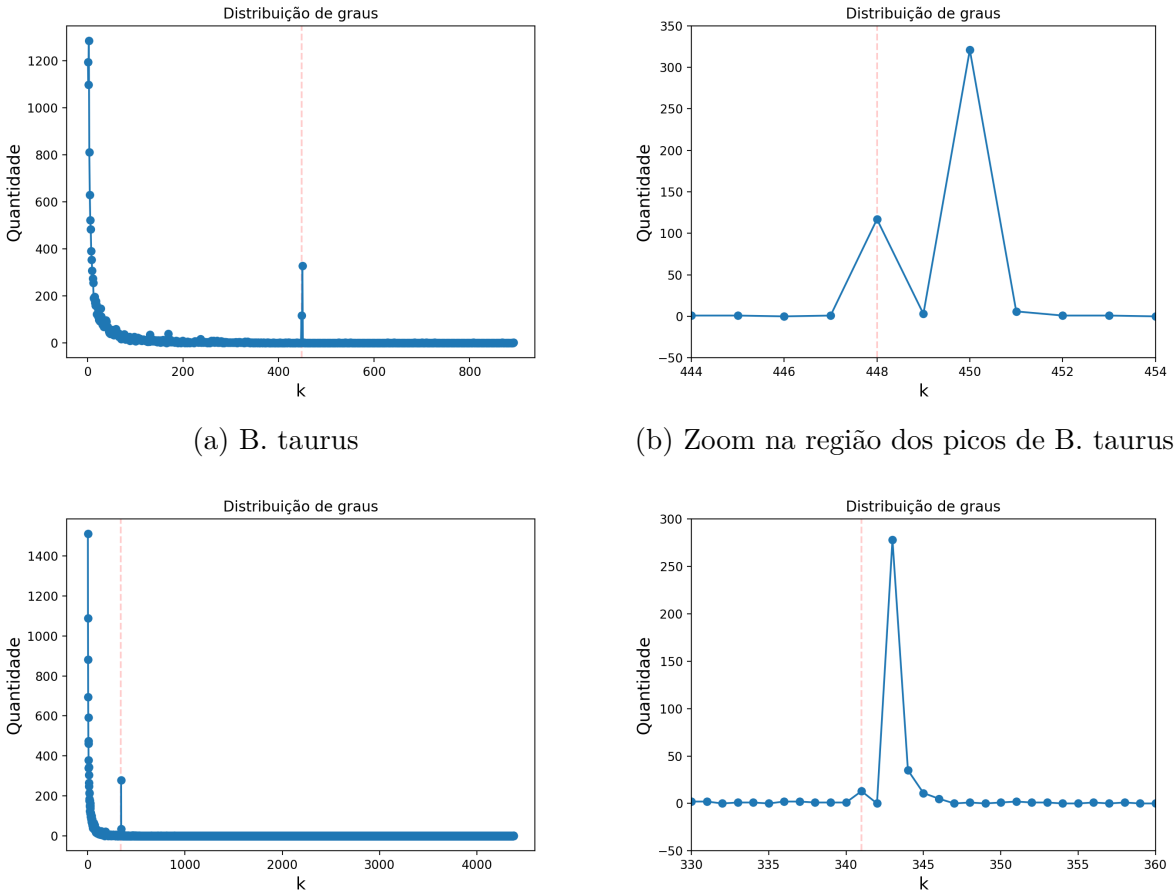
Fonte: Elaborada pelo autor

Tabela 3 – Valores do maior autovalor da matriz de adjacências, raiz quadrada do maior grau da rede e razão entre $\langle k^2 \rangle / \langle k \rangle$ para todos os organismos considerados

Rede	λ_1	$\sqrt{k_{max}}$	$\langle k^2 \rangle / \langle k \rangle$
B. taurus (0.7)	449.10	29.87	228.50
B. taurus (0.5)	449.12	48.17	235.31
C. elegans (0.7)	266.86	34.35	159.06
C. elegans (0.5)	418.64	41.81	261.42
D. melanogaster (0.7)	170.91	31.87	114.09
D. melanogaster (0.5)	257.82	39.10	171.61
D. rerio (0.7)	194.97	26.68	125.59
D. rerio (0.5)	251.80	44.43	235.54
E. coli K12 W3110 (0.7)	63.87	12.88	31.81
E. coli K12 W3110 (0.5)	78.30	15.56	52.30
H. sapiens (0.7)	342.86	66.17	192.35
H. sapiens (0.5)	342.87	79.43	223.59
P. troglodytes (0.7)	151.94	41.84	93.16
P. troglodytes (0.5)	204.35	52.50	165.41
S. cerevisiae (0.7)	197.34	46.71	134.49
S. cerevisiae (0.5)	236.60	55.18	186.22

Fonte: Elaborada pelo autor

K-cores das redes com *threshold* 0.7, observamos que eles são de fato grafos completos: para *H. sapiens* o maior K-core é composto de 342 proteínas todas com grau 341; e no caso de *B. taurus* 449 proteínas com grau 448. Além disso, se observarmos as distribuições cumulativas de graus das Figuras 9 e 10, podemos observar que para esses dois organismos existe uma região onde ocorre uma queda abrupta do valor da probabilidade de encontrar um vértices com grau maior ou igual a k para valores próximos aos citados acima. Para facilitar a visualização na Figura 31 geramos os gráficos das distribuições de graus; a linha vermelha representa os graus dos maiores K-cores para os dois organismos citados acima. Como podemos observar existe um número grande de proteínas na região do grau selecionado (apresentando um comportamento distinto da distribuição geral). O zoom na região dos picos (Figuras 31b e 31d) mostra que na realidade o grau dessas proteínas é ligeiramente maior que o grau do maior K-core, mas é preciso ter em mente que o processo de remoção do K-core irá remover proteínas de grau ligeiramente menor que podem estar conectadas ao vértices dos maiores K-cores. Acreditamos que a maioria dessas proteínas que aparecem em grande quantidade são as que compõe o maior K-core de cada rede. Também podemos notar que o para *B. taurus* existe um segundo pico próximo ao maior (provavelmente o segundo maior K-core da rede), e que uma parte significativa da localização acaba caindo nele também, como pode ser observado na Figura 30a.



(a) B. taurus

(b) Zoom na região dos picos de B. taurus

(c) H. sapiens

(d) Zoom na região dos picos de H. sapiens

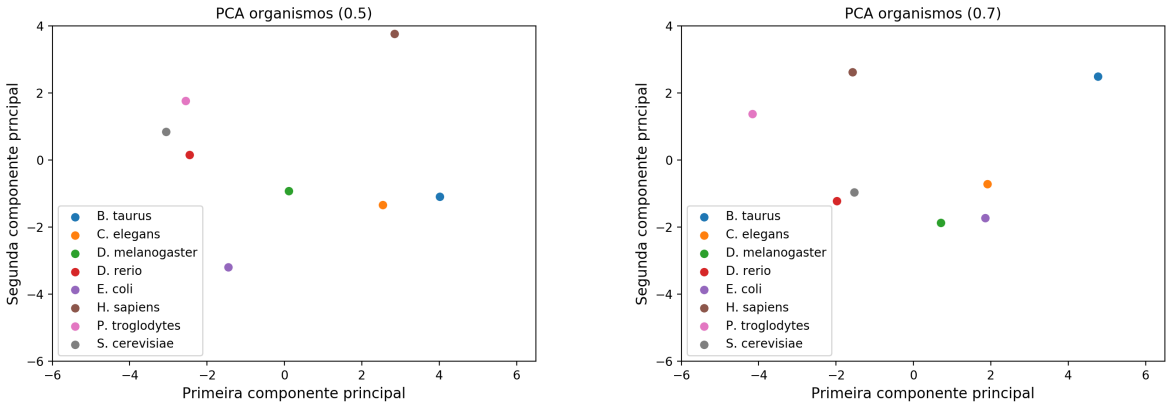
Figura 31 – Distribuições de graus das redes PPI com *threshold* de 0.7 dos organismos *B. taurus* e *H. sapiens*. A linha vermelha representa o grau do maior K-core de cada rede, que é um grafo completo

Fonte: Elaborada pelo autor

3.4 Comparação das redes utilizando as correlações de Kendall

Como vimos na seção 3.2, para todos os organismos considerados as centralidades apresentaram padrões de comportamento entre os pares de medidas apresentando algumas pequenas variações com as mudanças de organismo e de *threshold*. Além disso, os valores das correlações de Kendall apresentados na Tabela 2 também refletem esse fato, onde a maioria das medidas está próxima de um determinado valor (com algumas variações). Dessa forma investigamos se é possível utilizar os valores das correlações entre as medidas como características das redes e utilizá-las de forma a comparar os organismos utilizando sua estrutura. Resultados anteriores[85] indicam que é possível utilizar as correlações entre as medidas para comparar redes reais de tipos distintos, porém essa ideia ainda não foi bastante explorada para redes de um mesmo tipo.

Dessa forma exploraremos a possibilidade de se utilizar as correlações de Kendall[87]



(a) *Threshold* 0.5 (b) *Threshold* 0.7

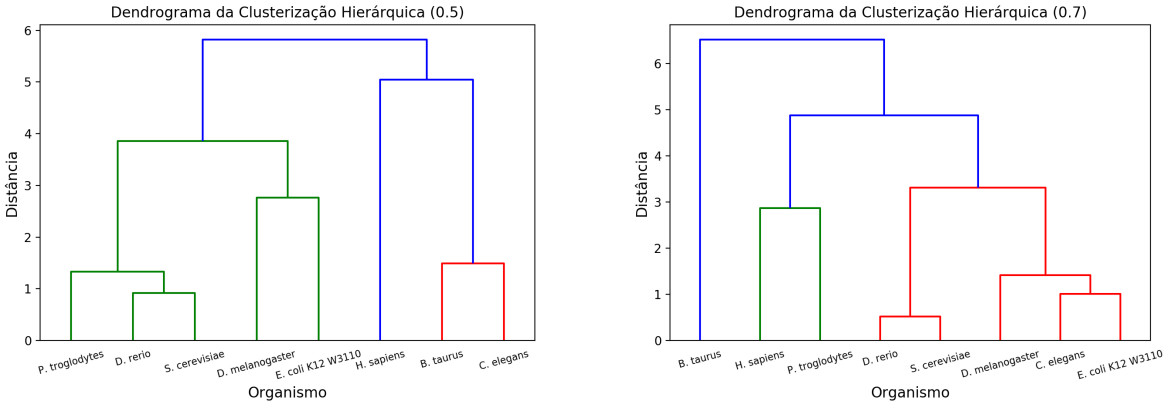
Figura 32 – Análise de componentes principais empregando as correlações de Kendall entre as medidas de centralidade dos organismos como características das redes

Fonte: Elaborada pelo autor

entre os pares de medidas de centralidade como características das redes PPI, e buscaremos realizar uma comparação entre elas. Para cada rede calculamos as centralidades de seus vértices e para todos os pares de medidas calculamos a correlação de Kendall, associando cada um dos valores como uma característica da rede. Em seguida nós realizamos uma análise de componentes principais como uma forma de visualização dos resultados obtidos através da redução de dimensionalidade.

Nas Figuras 32 e 33 apresentamos as projeções dos organismos nas duas maiores componentes principais para os organismos com os *thresholds* de 0.5 e 0.7 e sua clusterização hierárquica; nestas análises de componentes principais 72% da informação é explicada pelas duas componentes principais mais significantes para o *threshold* menor e 67% para o valor maior e a clusterização feita utilizou a distância euclideana e o método *average* de *linkage* — para os dendrogramas utilizamos apenas as duas maiores componentes para calcular a distância. Podemos notar todos os organismos foram diferenciados pelo conjunto de correlações, não havendo dois organismos sobrepostos independentemente do *threshold* escolhido. Também observamos que com a mudança de *threshold* existem alterações no posicionamento de alguns organismos, como por exemplo: *B. taurus* e *C. elegans* que estão mais próximos no *threshold* 0.5 do que no de 0.7; um comportamento semelhante acontece entre os organismos *P. troglodytes*, *S. cerevisiae* e *D. rerio* onde o primeiro organismo se afasta mais dos outros dois no caso do *threshold* maior. Por outro lado, também houveram organismos que apresentaram comportamentos semelhantes independentemente do *threshold* escolhido, como o pare *D. rerio* e *S. cerevisiae* que estão próximos nos dois gráficos da Figura 32.

Utilizando a mesma metodologia anterior, também realizamos uma comparação entre as redes com os dois *thresholds* do organismos *S. cerevisiae* e redes geradas com os



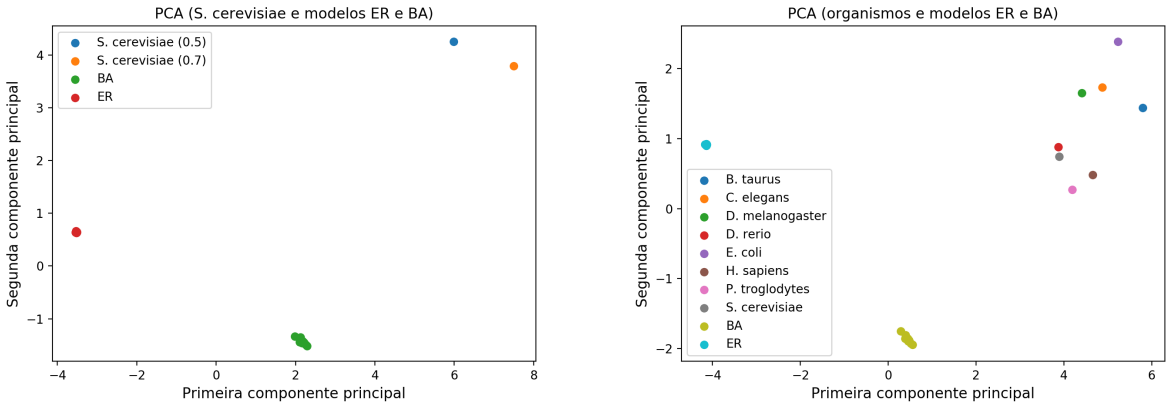
(a) *Threshold* 0.5 (b) *Threshold* 0.7

Figura 33 – Clusterização hierárquica da análise de componentes principais realizada

Fonte: Elaborada pelo autor

modelos ER e BA, que utilizaram o número de vértices e grau médio desse organismo para serem gerados. A comparação entre eles é apresentada na Figura 34a, e 92% da informação dessa PCA é explicada por essas duas variáveis. Na Figura observamos que os dois modelos e os dois *thresholds* do organismo foram bem caracterizados, não havendo regiões de superposição entre eles. Também podemos observar, a distância entre entre os modelo ER e BA é maior do que entre o das duas de *S. cerevisiae* e que os dois modelos estão bastante distantes das redes do organismo. Finalmente também observamos que é possível observar uma dispersão maior nas redes do tipo BA, do que nas do tipo ER, mostrando que existe uma maior variação nos valores das correlações da rede BA — que também pode ser observado na comparação dos *scatter-plots* das figuras 27 e 28. Na Figura 34b realizamos a mesma comparação entre os modelos, porém neste caso consideramos as redes PPI de todos os organismos para o *threshold* de 0.7 e verificamos que o comportamento observado é similar àquele apresentado na Figura 34a, onde podemos notar uma divisão bastante clara entre as redes dos organismos e a dos modelos considerados.

Ali et al. desenvolveram o método NetDis[103] que utiliza apenas informações sobre a topologia das redes para a comparação de redes PPI. O método desenvolvido associa a cada rede um vetor de contagem de subgrafos induzidos também chamados de *graphlets*[104] de tamanhos 2 a 5 extraídos das ego-redes[105] de dois passos (*two-steps ego-networks*) extraídas de todos os vértices de cada rede. Com ele, os autores mostraram que o método foi capaz de agrupar de maneira hierárquica redes geradas utilizando modelos, redes PPI de alguns organismos e um subconjunto de 151 redes sem peso e sem direção comparadas por Onnella et al. em outro trabalho.[106] Para as redes PPI os autores consideraram os organismos *H. sapiens*, *D. melanogaster*, *S. cerevisiae* (Yeast), *H. pylori* e *E. coli*. Os dois organismos mais próximos segundo o método NetDis são o *H. sapiens* e *D. melanogaster* que formam o primeiro grupo; em seguida os organismos *H. pylory* e *E. coli* são agrupados em um



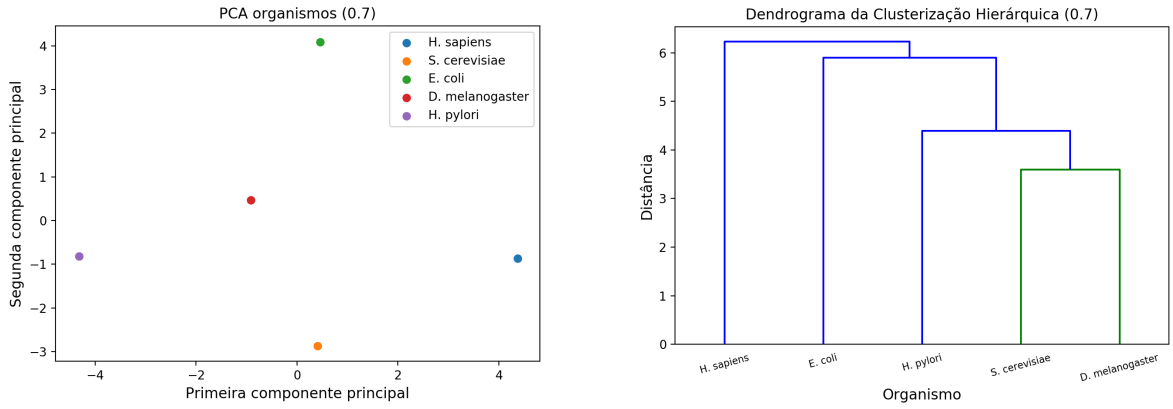
(a) (b)

Figura 34 – Comparação entre as redes geradas pelo modelos ER e BA e a dos organismos. Em 34a utilizamos apenas os modelos e o organismo *S. cerevisiae*, na Figura 34b apresentamos a comparação com todos os organismos para o *threshold* de 0.7

Fonte: Elaborada pelo autor

segundo *cluster* apresentando aproximadamente a mesma distância que os dois organismos anteriores; em seguida o organismo *S. cerevisiae* é agrupado ao conjunto de *H. sapiens* e *D. melanogaster* pois está mais próximo desses dois do que dos demais; e finalmente o grupo das bactérias e dos outros três organismos é unido por apresentar a maior distância. Segundo os autores, o agrupamento do método NetDis retorna o dendrograma que representa a árvore filogenética atualmente aceita como a correta, separando um ramo com as bactérias e outro com os organismos *H. sapiens*, *D. melanogaster* e *S. cerevisiae*; eles sugerem que a topologia das redes também possui alguma informação sobre o processo de evolução mas que o método não deve ser utilizado para isso, já que existem técnicas mais adequadas para esta tarefa. Para comparação nós realizamos um agrupamento hierárquico considerando somente os mesmos organismos do trabalho, já que os dois métodos consideram apenas as características topológicas das redes para realizar a comparação. Para isso montamos a rede do organismo *H. pylori* (*threshold* 0.7) da mesma forma que os demais organismos considerados até agora, calculando as correlações entre os pares de suas medidas de centralidade. Na sequência realizamos o PCA utilizando a densidade de probabilidade dos autovalores como características das redes consideradas (da mesma forma que o realizado anteriormente), realizando um agrupamento hierárquico na sequência. Para o agrupamento realizado, utilizamos um método hierárquico *bottom-up*, empregando o método *average linkage* para juntar os grupos e a distância Euclidiana entre os pares de organismos.

Os resultados obtidos são apresentados na Figura 35 onde apresentamos a projeção do PCA e o dendrograma obtido. Pelo PCA podemos observar que para este conjunto de organismos as redes continuam sendo diferenciadas pelas correlações entre as medidas. O resultado do agrupamento entretanto é bastante diferente do que foi obtido pelo



(a) (b)

Figura 35 – Agrupamento dos organismos *E. coli K12 W3110*, *H. pylori*, *H. sapiens*, *D. melanogaster* (mosca) e *S. cerevisiae* (Yeast)

Fonte: Elaborada pelo autor

método NetDis: o organismo *H. sapiens* deveria estar mais próximo dos organismos *D. melanogaster* e *S. cerevisiae* respectivamente e *H. pylori* deveria estar mais próximo de *E. coli*. É importante ressaltar aqui que o algoritmo NetDis possui uma "granularidade mais fina" do que o método das correlações proposto que analisa o comportamento geral das medidas; essa diferença poderia ser uma possível justificativa para a diferença nos resultados.

3.5 Comparação das redes utilizando o espectro da matriz Laplaciana normalizada

Como discutimos na seção 2.2, muitas propriedades topológicas das redes complexas podem ser extraídas do espectro da matriz Laplaciana 2.5. Com este fato em mente, uma pergunta que podemos nos realizar é se existe a possibilidade de caracterizar ou diferenciar redes complexas através do espectro da matriz Laplaciana normalizada.

Escolhemos trabalhar com a Laplaciana normalizada pois ela é definida para qualquer rede sem direção, com um ou mais componentes conectados e seus autovalores são limitados no intervalo [0, 2], o que facilita a comparação entre redes que possuem diferentes quantidades de vértices e arestas. Para a comparação nós calculamos todos os autovalores associados à matriz Laplaciana dos grafo das redes de proteínas de todos os organismos considerados. Em seguida montamos histograma normalizado para cada conjunto de autovalores associado a uma rede PPI. Todos os histogramas gerados possuíam 100 caixas igualmente espaçadas no intervalo [0, 2]. As densidades de frequência obtidas foram utilizadas *features* das redes para caracterizá-las.

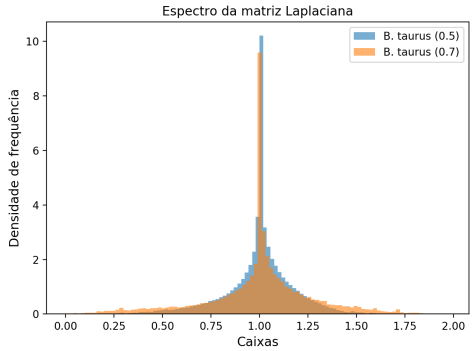
Na Figura 36 apresentamos a densidade de frequência dos autovalores dos espectros

para todos os organismos considerados. Como podemos observar na Figura, independentemente do *threshold* escolhido, existe um pico em torno do valor 1 para todas as redes PPI. Além disso mudanças no *threshold* implicam em pequenas variações no espectro da matriz Laplaciana, e para o valor 0.5 os autovalores parecem estar mais concentrados em torno do centro (valor 1) do que para o *threshold* maior de 0.7 que apresenta valores mais dispersos. Podemos observar que é comum o espectro assumir o formato de um "cone" ou "triângulo" ao redor do valor 1. Também observamos que a maioria dos autovalores encontra-se no intervalo entre 0.25 e 1.75, e que raramente encontramos autovalores nas caixas fora desse intervalo, com uma exceção sendo o organismo *E. coli K12 W3110* que apresenta uma quantidade de autovalores de valor baixo um pouco maior que os demais organismos. Comparando-se organismos distintos, podemos observar que apesar de existir um formato geral de "cone" entre os organismos existem também diferenças nos espectros, que pode ser observada nos valores das densidades de frequência dos organismos.

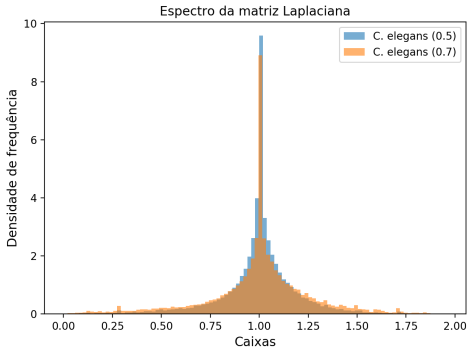
Na Figura 37 apresentamos os espectros das matrizes Laplacianas para o organismo *S. cerevisiae*. Em 37a apresentamos o espectro de *S. cerevisiae* para ambos os *thresholds*; na Figura 37b apresentamos o espectro das redes baseadas em nos modelos ER e BA comparadas com o espectro da rede PPI *S. cerevisiae* com *threshold* 0.7 — que foi a rede da qual tiramos os parâmetros de números de vértices e grau médio utilizados nos dois modelos.

Na Figura 37b observamos que os espectros dos dois modelos é bastante distinto da rede PPI, porém não tão distintos entre os dois modelos, com a diferença entre eles aparecendo mais na região central do intervalo (em torno do valor 1.0), onde existe um número maior de autovalores para o modelo BA do que o modelo ER, que apresenta uma quantidade um pouco maior de autovalores nas extremidades do que o BA. Aparentemente existe maior diferença entre os espectros das redes PPI com relação ao *threshold* do que entre os modelos ER e BA.

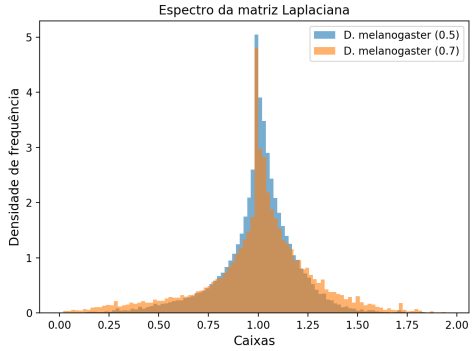
Utilizamos as densidades de frequência dos autovalores apresentadas na Figura 37 como características de cada uma das redes. Na sequência realizamos uma análise de componentes principais[107] (PCA) para redução de dimensionalidade e aumento da variância; projetando cada uma das redes em um espaço 2-dimensonal. É importante ressaltar que mais de 99% da variância é explicada por essas duas componentes. Os resultados obtidos são apresentados na Figura 38. No gráfico 38a apresentamos os 4 tipos de redes juntas (dois thresholds da rede PPI e 10 realizações de cada modelo). Como podemos ver a distância entre os dois *thresholds* de *S. cerevisiae* realmente é maior do que entre os dois modelos, e ambas as redes PPI estão significativamente distantes dos modelos. Na Figura 38b mostramos um zoom na região dos modelos mostrando que a diferença observada entre os espectros dos modelos também é suficiente para diferenciá-los, e que apesar de existir um espalhamento nas realizações dos modelos cada um deles possui uma região muito bem



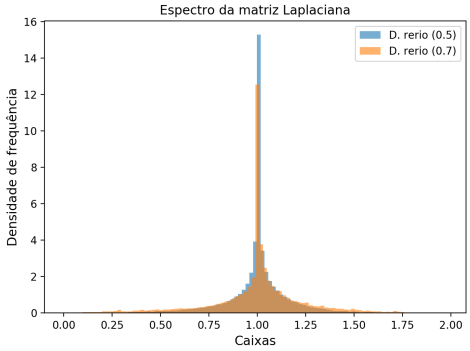
(a) B. taurus



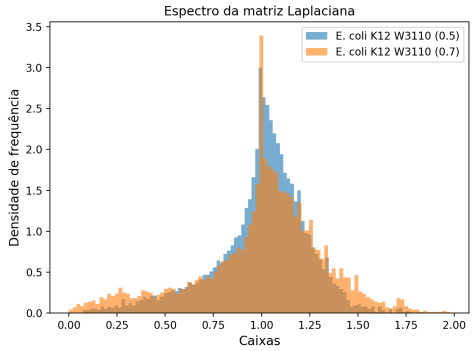
(b) C. elegans



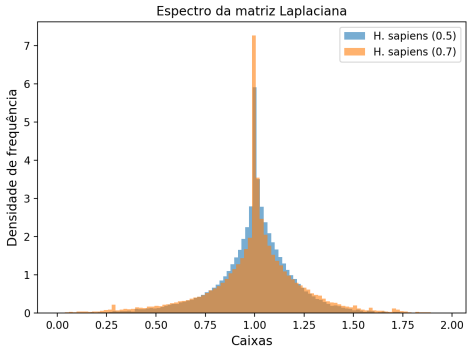
(c) D. melanogaster



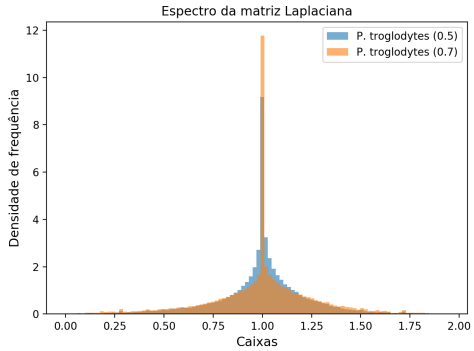
(d) D. rerio



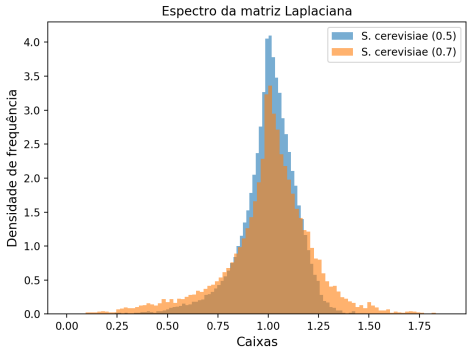
(e) E. coli



(f) H. sapiens



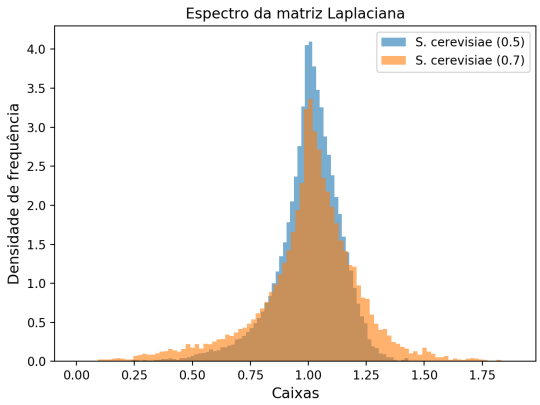
(g) P. troglodytes



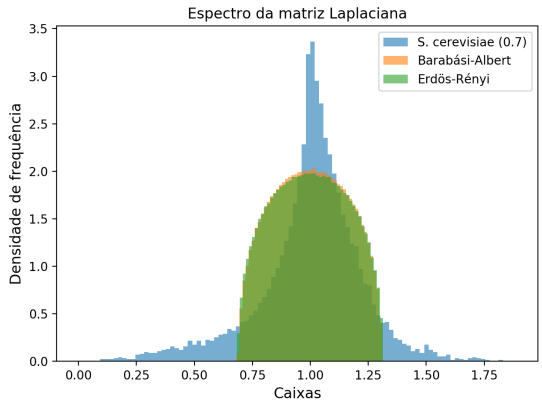
(h) S. cerevisiae

Figura 36 – Espectros das redes PPI para todos os organismos considerados

Fonte: Elaborada pelo autor



(a) Comparação dos espectros das matrizes Laplacianas do organismo *S. cerevisiae* para os dois valores de *threshold*

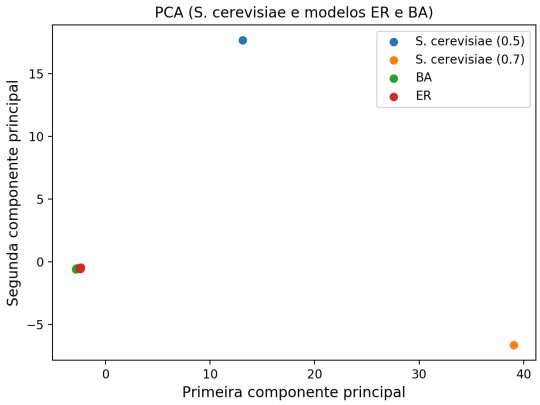


(b) Comparação dos espectros das matrizes Laplacianas dos modelos ER, BA e do organismo *S. cerevisiae*(0.7)

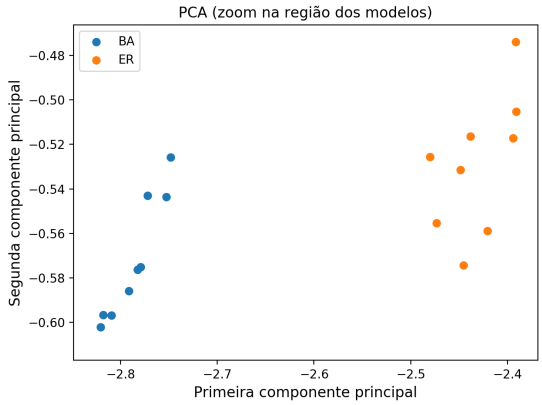
Figura 37 – Comparação dos espectros das matrizes Laplacianas. Em 37a comparamos os espectros do organismo S. cerevisiae com a mudança de *threshold*. Na Figura 37b apresentamos a comparação dos espectros dos modelos ER, BA com a a da redes PPI de *S. cerevisiae* (0.7)

Fonte: Elaborada pelo autor

definida. Também podemos observar que existe maior dispersão nas realizações do modelo ER do que para as do BA.



(a)



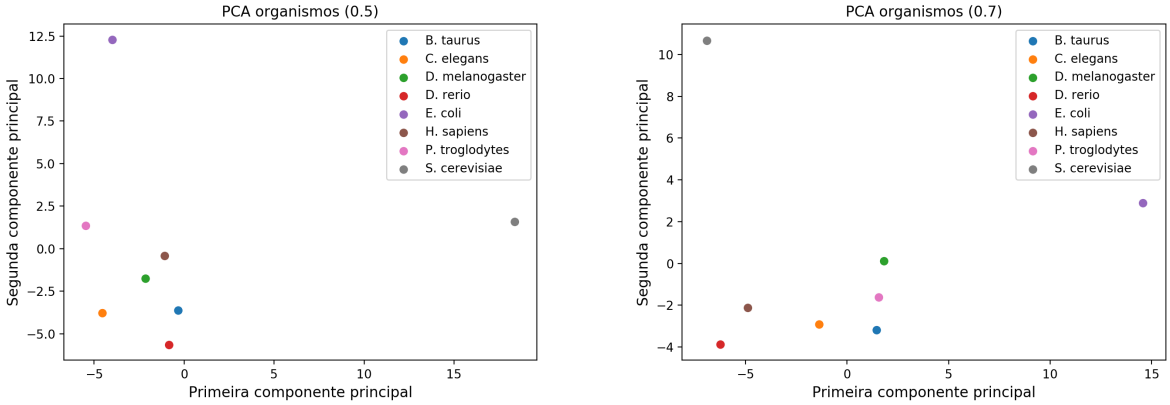
(b)

Figura 38 – Análise de componentes principais utilizando as densidades de frequência da figura 37 como características das redes

Fonte: Elaborada pelo autor

De modo similar ao que foi desenvolvido na Figura 38, também realizamos uma análise de componentes principais considerando apenas as redes PPI; os resultados para os *trheshold* considerados são apresentados na Figura 39, e os dendrogramas resultantes desse PCA são apresentados na Figura 40. Para o PCA do *threshold* 0.5 78% da variância é

explicada pelas duas componentes principais (88% para 3 componentes principais); já para o *threshold* 0.7 61% da variância está presente nas duas componentes principais (75% para 3 componentes) — isto nos indica que para 0.7 é interessante levar em consideração as 3 maiores componentes principais. Apesar disso, podemos observar que tanto na Figura 39a quanto na 39b cada organismo ficou bem caracterizado em uma região do espaço, não havendo nenhuma região onde dois organismos estão sobrepostos. Além disso, mesmo existindo diferença entre os resultados com a mudança do *threshold*, podemos observar que nos dois gráficos da Figura 39 os *E. coli K12 W3110* e *S. cerevisiae* estão mais distantes dos demais organismos; e que *B. taurus*, *P. troglodytes* e *D. melanogaster* estão próximos uns dos outros.

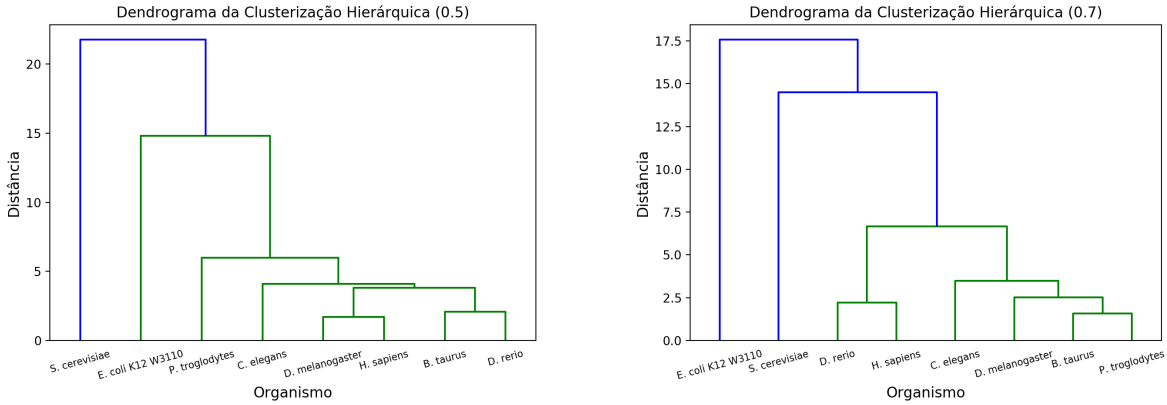


(a) *Threshold* 0.5 (b) *Threshold* 0.7

Figura 39 – Análise de componentes principais utilizando as densidades de frequência dos espectros da matriz Laplaciana dos organismos como características das redes

Fonte: Elaborada pelo autor

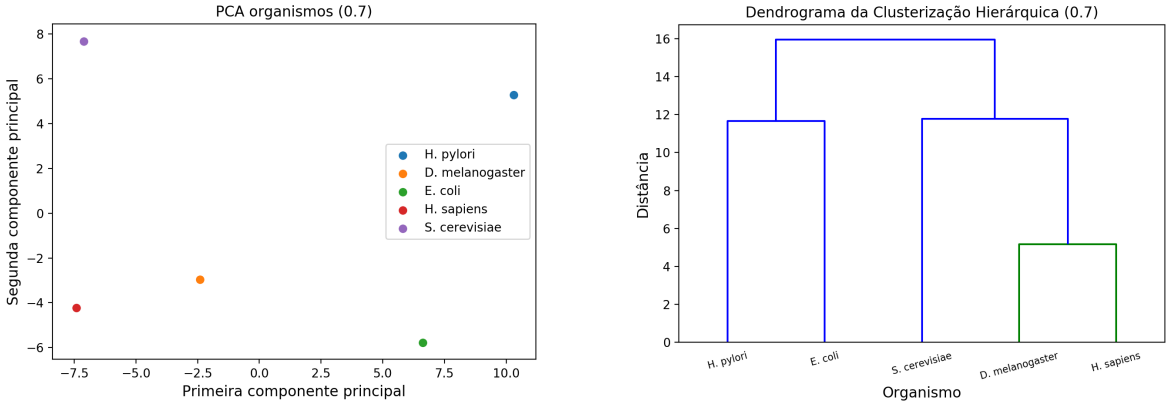
De maneira similar com o que foi desenvolvido na seção 3.4 aqui também comparamos os resultados obtidos com o método NetDis. O resultado é apresentado na Figura 41, onde o PCA contendo apenas os organismos da referência[103] é apresentado na Figura 41a. o resultado obtido pelo agrupamento hierárquico (Figura 41b) é o mesmo que o obtido pelos autores do artigo com relação aos agrupamentos e na ordem de junção dos organismos, com a diferença ocorrendo nas distâncias entre os dois métodos (o NetDis o agrupamento de *H. pylori* e *E. coli* é aproximadamente a mesma de *H. sapiens* e *D. melanogaster*). Também realizamos o mesmo agrupamento entre esses 5 organismos sem realizar a redução de dimensionalidade do PCA na densidade de probabilidade do espectro da matriz Laplaciana, utilizando apenas a standardização para garantir que todas as caixas do histograma do espectro possuem a mesma importância na comparação, e obtivemos o mesmo resultado entre os pares agrupados.



(a) *Threshold* 0.5 (b) *Threshold* 0.7

Figura 40 – Dendrograma das análises de componentes principais de todos os organismos para os dois *thresholds* utilizados, utilizando o espectro da matriz Laplaciana normalizada

Fonte: Elaborada pelo autor



(a) (b)

Figura 41 – Agrupamento dos organismos *E. coli K12 W3110*, *H. pylori*, *H. sapiens*, *D. melanogaster* (mosca) e *S. cerevisiae* (Yeast)

Fonte: Elaborada pelo autor

3.6 Agrupamento de proteínas utilizando as medidas de centralidade

Com o desenvolvimento dos experimentos de alto *throughtput*,[108] — que permitem a detecção de um número elevado de interações entre proteínas de uma única vez — o número de novas interações entre proteínas reportadas aumentou significativamente. Dessa forma um ponto que gostaríamos de analisar é se existe alguma associação entre as funções biológicas desempenhadas pelas proteínas e a topologia das redes. Caso esta associação exista, poderíamos utilizar a estrutura das redes para auxiliar no entendimento de novas proteínas, interações descobertas e nos processos biológicos a elas associados.

Devido ao fato observado na seção 3.2 de que dependendo da centralidade utilizada conjuntos distintos de vértices da rede são mais ou menos caracterizados, nós também investigamos a possibilidade de associação entre as estruturas topológicas das redes PPI e as funções biológicas desempenhadas por elas. Mais especificamente, iremos utilizar as medidas de centralidade das proteínas como suas características e em seguida realizaremos um agrupamento; finalmente verificaremos se existem funções biológicas que se destaquem para as proteínas dentro de um mesmo *cluster*. Para verificar a função biológica, utilizaremos uma ferramenta de análise também presente o banco de dados STRING, onde um conjunto de proteínas é fornecido para o banco de dados e ele nos retorna se existem funções que se destaquem dentro do conjunto fornecido.

Nesta análise, utilizamos a rede PPI de *H. sapiens* com *threshold* de 0.7 — fizemos esta escolha pois este é um dos organismos mais amplamente estudados com relação as vias metabólicas e a escolha do *threshold* mais elevado é devida a maior confiança da existência nas interações. Em seguida, utilizamos o conjunto de medidas de centralidade descrito na seção 2.6 como características de cada uma das proteínas estudadas.

Como cada medida de centralidade apresenta valores variando em escalas diferentes — por exemplo a medida de proximidade tende a atribuir valores com variações pequenas entre os vértices da rede — realizamos a reescala dos valores das medidas; e pelo fato das distribuições dos valores serem do tipo cauda longa, utilizamos um método não linear de reescala dos valores das medidas.[109] A transformação realizada é do tipo *deterministic rank-based inverse normal transformation*.[109]Em seguida, realizamos um agrupamento hierárquico (*hierarchical clustering*) das proteínas utilizando o método de *linkage* ward, a distância Euclidiana e 20 *clusters* de proteínas. A quantidade de proteínas presentes em cada *cluster* é apresentada na tabela 4.

Na Figura 42 apresentamos as duas componentes mais relevantes do PCA dos 20 grupos de proteínas encontrados (representados por cores diferentes), as cores associadas ao grupos são apresentadas na barra de cores ao lado do gráfico. Como podemos ver em duas componentes existem 6 grupos de proteínas que são visualmente separados e, em alguns deles, existem subgrupos divididos pelo algoritmo de aglomeração; é importante ressaltar que os subgrupos encontrados pela aglomeração podem apenas parecer estar juntos pois esta é uma visão contendo apenas duas dimensões. Em seguida utilizamos os grupos selecionados pelo algoritmo de clusterização para verificar a existência de funções biológicas associadas a eles.

Como citado anteriormente utilizamos o STRING para vertificar a existência de funções biológicas associadas às proteínas; o site reporta funções utilizado vias metabólicas de ontologia de genes[110] — que divide as funções em três categorias: função molecular, componente celular e processo biológico. O STRING também inclui informações de classificações reportadas nos bancos de dados KEGG,[111,112] PFAM[113] e INTERPRO.[114]

Tabela 4 – Quantidade de proteínas presentes em cada cluster encontrado e categorias de funções presentes. As categorias são: não analisado devido a quantidade de proteínas (*), nenhuma categoria presente (-), ontologia de genes - processo biológico (GOBP), ontologia de genes - função molecular (GOMF), ontologia de genes - componente celular (GOCC), via metabólica do KEGG (K), domínio de proteínas PFAM (P) e domínio de proteínas e características INTERPRO (I)

Cluster	Quantidade de proteínas	Categorias encontradas
0	2077	*
1	739	GOBP, GOMF, GOCC, K, P, I
2	620	GOBP, GOMF, GOCC
3	1070	GOBP, GOMF, GOCC, K, P, I
4	1146	-
5	439	GOBP, GOMF, GOCC
6	615	GOBP, GOMF, GOCC
7	1060	GOBP, GOMF, GOCC, P
8	979	GOBP, GOMF, GOCC, K, P, I
9	853	GOBP, GOMF, GOCC, K
10	817	GOBP, GOMF, GOCC, K, P, I
11	230	GOBP, GOMF, GOCC, K, P, I
12	797	GOBP, GOMF, GOCC, K, P, I
13	541	GOBP, GOMF, GOCC, K, P, I
14	365	-
15	634	GOBP, GOMF, GOCC
16	681	GOBP, GOMF, GOCC, K, P, I
17	321	GOBP, GOMF, GOCC, K, I
18	567	GOBP, GOMF, GOCC
19	689	GOBP, GOMF, GOCC, K, P, I

Fonte: Elaborada pelo autor

Os tipos de funções observados em cada *clsuter* também são apresentados na tabela 4. Em seguida listaremos algumas das funções (as 5 mais relevantes para cada categoria segundo a ferramenta) reportadas para cada grupo obtido; utilizaremos o nome das funções em inglês da mesma forma como é reportado pelo STRING.

Cluster 0 — não foi possível obter resultados para ele pois o STRING consegue lidar apenas com grupos de até 2000 proteínas.

Cluster 1 — na categoria de ontologia de genes este grupo apresentou as seguintes funções de *protein binding*, *binding*, *enzyme binding*, *receptor binding* e *carbohydrate derivative binding* na categoria de função molecular; *Pathways in cancer*, *HTLV-I infection*, *PI3K-Akt signaling pathway*, *Epstein-Barr virus infection*, *Rap1 signaling pathway* com relação ao banco de dados KEGG; *protein kinease domain*, *SH3 domain*, *Protein tyrosine kinease*, *SH2 domain* e *14-3-3 protein* pelo banco de dados PFAM; e com relação ao

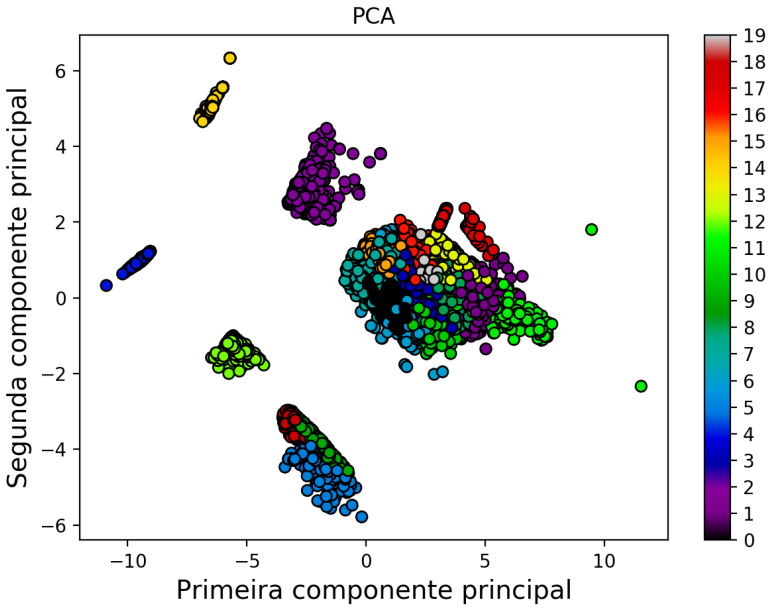


Figura 42 – Projeção dos 20 clusters de proteínas encontrados nas duas componentes principais mais significantes do PCA

Fonte: Elaborada pelo autor

banco de dados INTERPRO algumas das categorias reportadas foram *protein kinease-like domain*, *protein kinease domain*, *protein kinease, ATP binding site*, *serine/threonine/dual specificity protein kinease, catalytic domain* e *serine/threonine-protein kinease, active site*.

Cluster 2 — este grupo apresentou apenas funções no conjunto de ontologia de genes. Da categoria processo biológico apenas a função *positive regulation of GTPase activity* aparece; na categoria função molecular se destacam *GTPase activator activity*, *potassium channel activity*, *molecular function regulator*, *voltage-gated potassium channel activity* e *enzyme activator activity*.

Cluster 3 — com relação a categoria ontologia de genes (função molecular) as funções listadas foram *catalytic activity*, *nucleotide binding*, *small molecule binding*, *ligase activity* e *RNA binding*; do banco de dados PFAM a única função destacada foi *Histidine phosphatase superfamily (branch 1)*; com relação ao KEGG se destacaram as funções *metabolic pathways*, *amino sugar and nucleotide sugar metabolism*, *biosynthesis of amino acids*, *fructose and mannose metabolism* e *cysteine and methionine metabolism*; e do banco de dados INTERPRO se detacaram *P-loop containing nucleoside triphosphate hydrolase*, *histidine phosphatase superfamily, clade-1*, *histidine phophatase superfamily* e *phosphoglycerate/bisphosphoglycerate mutase, active site*.

Cluster 4 — não apresentou funções que se destaquem em nenhum tipo de categoria.

Cluster 5 — este grupo apresentou funções com destaque apenas na categoria de ontologia de genes: *anion transmembrane transport* da categoria processo biológico;

secondary active transmembrane transporter activity, *anion transmembrane transporter activity*, *active transmembrane transporter activity*, *neutral amino acid transmembrane transporter activity* e *transmembrane transporter activity* na categoria função molecular.

Cluster 6 — este grupo também apresentou funções da categoria ontologia de genes, sendo as do tipo processos biológicos *single-organism developmental process*, *developmental process*, *anatomical structure morphogenesis*, *system development*, *primary metabolic process*; já do tipo função molecular as funções destacadas para este grupo foram *sequence-specific DNA binding*, *protein binding*, *molecular function*, *binding*, *transcription regulatory region DNA binding.*

Cluster 7 — este grupo apresentou as funções *regulation of biological process*, *biological regulation*, *regulation of cellular process*, *negative regulation of biological process*, *protein dephosphorylation* na categoria processo biológico da ontologia de genes; na categoria função molecular também da ontologia de genes as funções foram *protein tyrosine phosphatase activity*, *phosphoprotein phosphatase activity*, *molecular function regulator*, *enzyme regulator activity* e *intracellular cAMP activated cation channel activity*; na categoria PFAM protein domains a função *leucine rich repeat* foi destacada.

Cluster 8 — este grupo apresentou funções em todas as categorias. Com relação a ontologia de genes destacamos *viral processs*, *symbiosis, encompassing mutualism threough parasitism*, *cellular component organization or biogenesis*, *cellular macromolecule catabolic process*, *cellular component organization* na categoria processo biológico; *RNA binding*, *poly(A) RNA binding*, *heterocyclic compound binding*, *organic cyclic compound binding* e *nucleic binding* na categoria função molecular. Na categoria KEGG pathways se destacaram as funções *ribosome*, *proteasome*, *spliceosome*, *metabolic pathways*, *ubiquitin mediated proteolysis*; na categoria PFAM *proteasome subunit*, *RNA recognition motif. (a.k.a RRM, RBD, or RNP domain)*, *LSM domain*, *PCI domain*, *ubiquitin-conjugating enzyme*; na categoria INTERPRO as funções *proteasome, subunit alpha/betha*, *proteasome B-type subunit*, *proteasome beta-type subunit, conserved site*, *nucleophile aminohydrolases, N-terminal*, *nucleotide-binding alpha-beta plait domain.*

Cluster 9 — com relação a processos biológicos este grupo apresentou as funções *single-organism metabolic process*, *carbohydrate derivative metabolic process*, *glycoprotein biosynthetic process*, *carbohydrate derivative biosynthetic process* e *post-translacional protein modification.* Em função molecular destacamos *transferase activity, transferring hexosyl groups*, *catalytic activity*, *transferase activity, transferring glycosyl groups*, *UDP-glycosyltransferase activity* e *transferase activity.* Apareceram também as seguintes funções da categoria KEGG pathways *metabolic pathways*, *glycosylphosphatidylinositol(GPI)-anchor biosynthesis*, *mucin type O-glycan biosynthesis*, *glycosphingolipid biosynthesis - lacto and neolacto series* e *glycosphingolipid biosynthesis - globo series.*

Cluster 10 — este grupo também apresentou funções em todas as categorias. Das

de ontologia de genes, na subcategoria processo biológico destacamos *single-organism metabolic process*, *organic substance metabolic process*, *nitrogen compound metabolic process*, *cellular metabolic process*, *metabolic process*; em função molecular *protein binding*, *binding*, *organic cyclic compound binding*, *heterocyclic compound binding* e *transcription factor activity, transcription factor binding.* Da categoria KEGG pathways observamos *matabolic pathways*, *drug metabolism — cytochrome P450*, *retinol metabolism*, *chemical carcinogenesis* e *metabolism of xenobiotics by cytochrome P450.* Na categoria PFAM protein domains destacam-se as funções *ligand-binding domain of nuclear hormone receptor*, *activin types I and II receptor domain*, *zinc finger, C4 type (two domains)*, *troponin.* E na categoria INTERPRO protein domains and features *Ser/Thr protein kinase, TGFB receptor*, *nuclear hormone receptor*, *activin types I and II receptor domain*, *nuclear hormone receptor, ligand-binding domain*, *zinc finger, NHR/GATA-type.*

Cluster 11 — a categoria processo biológico de ontologia de genes apresentou as funções *cellular response to endogenou stimulus*, *positive regulation of macromolecule metabolic process*, *enzyme linked receptor protein sugnaling pathway*, *cellular response to growth factor stimulus* e *positive regulation of cellular metabolic process.* A categoria função molecular de ontologia de genes apresentou as funções *protein binding*, *enzyme binding*, *organic cyclic compound binding*, *heterocyclic compund binding* e *binding.* Na categoria KEGG pathways as seguintes funções foram destacadas *pathways in cancer*, *protoglycans in cancer*, *PI3K-Akt signaling pathway*, *hepatitis B*, *Epstein-Barr virus infection.* Do banco de dados PFAM se destacaram as funções *protein kinase domain*, *protein tyrosine kinase*, *SH2 domain*, *Ras family*, *MH2 domain.* Com relação ao banco de dados INTERPRO o STRING mostra as seguintes funções: *protein kinase, ATP binding site*, *protein kinase-like domain*, *protein kinase domain*, *serine/threonine/dual specificity protein kinase, catalytic domain*, *serine/threonin-protein kinase, active site.*

Cluster 12 — este grupo apresentou poucas funções. Na categoria processo biológico de ontologia de genes a função *biological process* se destaca. Da categoria KEGG se destaca *mucin type O-Glycan biosynthesis.* Do banco de dados PFAM se destacam *glycosyl transferase family 2* e *ricin-type beta-trefoil lectin domain.* Da categoria INTERPRO se aparece a função *glycosyltransferase 2-like.*

Cluster 13 — as funções *G-protein coupled receptor signaling pathway*, *signal transduction*, *single organism signaling*, *cell communication* e *G-protein coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger* foram ressaltadas na categoria processo biológico da ontologia de genes. Da função molecular de ontologia de genes se destacam *signal transducer activity*, *G-protein coupled receptor activity*, *transmembrane signaling receptor activity*, *molecular transducer activity*, *signaling receptor activity.* As funções *neuroactive ligand-receptor interaction*, *calcium signaling pathway*, *cytokine-cytokine receptor interaction*, *chemokine signaling pathway*, *PI3K-Akt signaling pathway*

da categoria KEGG. Do banco de dados PFAM foram destacadas as funções *7 transmembrane receptor (rhodopsin family)*, *small cytokines (intecrine/chemokine), interleukin-8 like*, *GGL domain*, *RhoGAP domain*, *hormone receptor domain*. Das funções da categoria INTERPRO destacamos *G protein-coupled receptor, rhodopsin-like*, *GPCR, rhodopsin-like, 7TM*, *chemokine interleukin-8-like domain*, *chemokine receptor family*, *CXC chemokine*.

Cluster 14 — este grupo não apresentou funções que se destaquem em nenhuma das categorias.

Cluster 15 — o conjunto de proteínas apresentou apenas funções relacionadas à ontologia de genes. Na subcategoria processo biológico destacam-se *biological process*, *protein ubiquitination*, *protein modification by small protein conjugation or removal*, *cellular process* e *protein modification by small conjugation*. A subcategoria função molecular apresentou as características *poly(A) RNA binding*, *RNA binding*, *ligase activity*, *ubiquitin-protein transferase activity* e *zinc ion binding*.

Cluster 16 — este conjunto de proteínas apresentou funções em todas as categorias de ontologia de genes, sendo elas *signaling*, *single organism signaling*, *cell communication*, *signal transduction*, *cellular response to stimulus* na categoria processo biológico; *protein binding*, *receptor binding*, *molecular function regulator*, *nucleoside-triphosphatase regulator activity* e *GTPase regulator activity* na categoria função molecular. Na categoria KEGG pathways as funções presentes são *cytokine-cytokine receptor interaction*, *phosphatidylinositol signaling system*, *Jak-STAT signaling pathway*, *axon guidance* e *Ras signaling pathway*. Da categoria PFAM protein domains as funções associadas a esse conjunto são *ligand-binding domain of nuclear hormone receptor*, *zinc finger, C4 type (two domains)*, *PH domain*, *SH2 domain*, *diacylglycerol kinase catalytic domain*. Na categoria INTERPRO as seguintes funções foram destacadas *nuclear hormone receptor, ligand-binding domain*, *zinc finger, nuclear hormone receptor-type*, *nuclear hormone receptor*, *zinc finger, NHR/GATA-type* e *diacylglycerol kinase, catalytic domain*.

Cluster 17 — algumas das funções que este conjunto de proteínas apresentou foram *G-protein coupled signaling pathway*, *sensory perception of smell*, *neurological system process*, *detection of chemical stimulus involved in sensory perception of smell* e *response to chemical* associados à processo biológico; *transmembrane signaling receptor activity*, *G-protein coupled receptor activity*, *olfactory receptor activity* e *odorant binding* na categoria função molecular; *olfactoy transduction* na categoria KEGG pathways e *olfactory receptor* na categoria INTERPRO.

Cluster 18 — este conjunto de proteínas apresentou as funções *anion transmembrane transport*, *anion transport*, *organic anion transport*, *ion transmembrane transport* e *inorganic anion transmembrane transport* na categoria processo biológico; *anion transmembrane transporter activity*, *organic anion transmembrane transporter activity*, *amino acid transmembrane transporter activity*, *secondary active transmembrane transporter activity* e

symporter activity na categoria função molecular.

Cluster 19 — este grupo de proteínas apresentou as funções *miotic cell cycle*, *cellular macromolecule metabolic process*, *cell cycle*, *cellular component organization or biogenesis* e *macromolecule metabolic process* na categoria processo biológico; *RNA binding*, *heterocyclic compound binding*, *organic cyclic compound binding*, *nucleic acid binding* e *protein binding* em função molecular; *systemic lupus erythematosus*, *alcoholism*, *RNA transport*, *DNA replication* e *RNA degradation* na categoria KEGG pathways; *core histone H2A/H2B/H3/H4*, *RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)*, *RanBP1 domain* do banco de dados PFAM; *histone-fold*, *histone H2A/H2B/H3*, *histone H4*, *TATA box binding protein associated factor (TAF)* na categoria INTERPRO.

Como podemos observar (tabela 4), a maioria dos grupos encontrados apresentaram algumas funções biológicas, sendo que as mais frequentes são as que estão associadas as categorias definidas pela ontologia. De todos os grupos encontrados, apenas dois não apresentaram nenhuma função de destaque. Apenas destacando algumas das funções interessantes observadas: o cluster 1 está relacionado com vias metabólicas envolvidas com câncer, infecção do vírus Epstein-Barr (também relacionado a alguns tipos de câncer), infecção HTLV-I dentre outras; o cluster 3 está associado ao metabolismo de açúcares, biossíntese de aminoácidos e metabolismo de frutose; o grupo 10 está associado com o metabolismo de drogas; o cluster 11 também parece estar associado a vias de câncer, vírus Epstein-Barr e Hepatite B; o grupo 17 está associado a transdução olfatória (neste caso encontramos todas as proteínas associadas com esta função). Os *clusters* obtidos também apresentam algumas características em comum como por exemplo, apenas algumas centenas de proteínas presentes neles, e um número elevado de conexões entre elas. Como a seleção das proteínas foi feita utilizando unicamente as centralidades dos vértices das redes, esses resultados nos indicam a existência de uma associação entre as medidas de centralidade e as funções biológicas das redes PPI.

4 CONCLUSÃO

Redes complexas podem ser utilizadas para representar diversos tipos de sistemas compostos por partes interagentes. Dentre eles, na área da biologia, a interação entre proteínas é de vital importância para o entendimento do funcionamento dos processos biológicos. Neste trabalho nós analisamos as propriedades estruturais de redes de interação de proteínas criadas a partir das informações fornecidas pelo banco de dados STRING. Este banco de dados possui grande diversidade de organismos e de volume de informações, possivelmente sendo o que possui maior número de proteínas e interações, é amplamente utilizado pela comunidade científica e vem sendo atualizado com grande frequência. Quanto às análises estruturais nos focamos no comportamento das medidas de centralidade e suas correlações, análise do espectro da matriz Laplaciana, localização do centralidade de autovetor e do maior K-core das redes.

Em nossa primeira análise, avaliamos o comportamento das medidas de centralidade dos vértices da rede, com relação a organismos diferentes e a mudanças dos *thresholds* escolhidos para a criação da rede de um mesmo organismo. Observamos que dado um par de medidas de centralidade, a correlação entre elas tende a seguir um padrão para a grande maioria dos organismos, com a única exceção sendo a medida de autovetor que apresentou um comportamento distinto para alguns casos. Esse fato sugere que a topologia da rede pode estar associada com as funções biológicas desempenhadas pelas proteínas. Além disso, também observamos que a utilização de um conjunto de medidas de centralidade para caracterizar os vértices das redes PPI é necessária, já que existem grupos de vértices que são mais discriminados por uma medida do que pelas demais. Comparando este comportamento com o de redes criadas utilizando-se os modelos ER e BA, observamos que os padrões de espalhamento das redes PPI são claramente diferentes daqueles obtidos para os modelos, apresentando uma forma geral distinta e maiores variações entre redes de um mesmo tipo.

O comportamento diferenciado da medida de autovetor nos levou a analisá-la em mais detalhes, mostrando que para todos os organismos grande parte de seu peso fica concentrado em um subconjunto dos vértices, fenômeno conhecido como localização. Esta localização apresenta características de redes altamente heterogêneas, ficando concentrada no maior K-core da rede para a maioria dos organismos. Esta é a primeira indicação que este fenômeno ocorre para redes PPI. Entretanto ainda nos falta entender o motivo de em alguns casos uma parte dessa localização estar presente no segundo maior K-core ou nos K-cores menores da rede por exemplo. Outro ponto interessante envolve um melhor entendimento do que acontece nos organismos *B. taurus* e *H. sapiens*, o motivo de eles apresentarem um grafo completo como maior K-core da rede que apresenta uma localização

mais acentuada do que os demais organismos. Finalmente, nos falta ainda entender melhor o motivo de em alguns casos a localização ocorrer nos K-cores medianos das redes, e se existe algum tipo de topologia específica associada a esses casos.

Utilizando os valores das correlações de Kendall entre as centralidades como características das redes, realizamos uma comparação entre elas. Os resultados mostraram que as correlações podem ser aplicadas como características para diferenciar as redes PPI de organismos distintos e que as correlações também capturam mudanças no *thresold* escolhido. Também observamos que as correlações são suficientes para distinguir redes baseadas nos modelos ER e BA da redes PPI representando organismos onde cada modelo e redes reais apresentam conjuntos de valores de correlação diferentes.

Também analisamos o espectro da matriz Laplaciana normalizada das redes PPI, mostramos que para este caso ele possui um formato de cone para todos os organismos e apresenta algumas variações entre redes e *thresholds* diferentes. Comparamos os espectros das redes PPI consideradas com o dos modelos e observamos que o padrão é bastante diferente entre essas duas categorias, sendo clara a diferença entre uma rede PPI e o modelo ER ou BA. Os resultados também mostraram que para um *threshold* fixado, existem diferenças entre os espectros das matrizes Laplacianas entre organismos distintos. Este fato nos permitiu realizar o agrupamento de redes PPI, onde reproduzimos um resultado bastante similar ao de outro trabalho que também compara redes PPI utilizando apenas suas propriedades topológicas; o agrupamento obtido também é o que era esperado segundo a árvore filogenética de classificação dos organismos. Assim acreditamos que o espectro da matriz Laplaciana pode ser utilizado como uma característica das redes. Maiores estudos precisam ser conduzidos utilizando um número maior de organismos para verificar se as distâncias obtidas entre os organismos realmente fazem sentido e se essa semelhança com o agrupamento filogenético correto entre os organismos também se mantém. Outras formas de comparação utilizando o espectro também podem ser utilizadas, como por exemplo medidas de distância entre distribuições.

Finalmente realizamos um agrupamento de proteínas para o organismo *H. sapiens* utilizando suas medidas de centralidade como suas características. Esse método nos permitiu agrupá-las utilizando algoritmos de clusterização hierárquico e, utilizando apenas as características estruturais da rede, nos permitiu identificar conjuntos de proteínas que possuem funcionalidades biológicas bem definidas. Além de também indicar a existência de uma associação entre topologia e função, esse método pode facilitar o entendimento dos processos biológicos já existentes, bem como facilitar a categorização ou sugestão de funções para novas proteínas que forem adicionadas nas redes dos organismos, baseando-se em suas conexões.

Os resultados obtidos até agora indicam que muitas informações importantes podem ser obtidas através das propriedades topológicas presentes nas redes de interação

de proteínas e nos levantam várias perguntas importantes. Um ponto que ainda nos falta entender, por exemplo, é a razão das medidas de centralidade se comportarem dessa forma para um conjunto de organismos tão diferentes? Seria esse um indicativo de que independentemente do organismo considerado, existe algum tipo de estrutura que é conservada entres eles com relação a como essas proteínas interagem? Em caso afirmativo, como e quanto é esperado de variação entre essas estruturas? Além disso, outro ponto que pode ser levantado é o motivo dos padrões observados, ou seja, qual a relação entre as propriedades topológicas aqui observadas e a função biológica desempenhada entre as proteínas.

Com relação à localização da centralidade de autovetor, ainda é necessário um maior entendimento sobre o motivo de *H. sapiens* e *B. taurus* possuírem um comportamento distinto dos demais organismos, e por que o K-core que se aproxima de um grafo completo está presente apenas neles — no caso específico de *B. taurus* mais de um grafo diferenciado aparece; um estudo considerando mais organismos também pode ser conduzido para verificar se o maior K-core com topologia de grafo completo apareçe em outros organismos, buscando entender-se o motivo deste tipo de estrutura aparecer. Além disso, também observamos casos em que a localização está presente em um K-core intermediário presente na rede — como no caso de *S. cerevisiae*. O que nos leva a perguntar se o que observamos aqui é de fato a localização previamente reportada na literatura de redes altamente heterogêneas, sofrendo efeito de propriedades estruturais que não estão presentes nos modelos ou se trata-se de algum outro tipo de fenômeno; estudos considerando a introdução de um grafo completo que não seja o maior K-core da rede na topologia também podem ser realizados, buscando avaliar o impacto que este tipo de estrutura pode causar nas redes.

A comparação de redes utilizando o espectro da matriz Laplaciana normalizada também precisa ser testada com um número maior de organismos e considerando-se outros tipos de rede para verificação de sua validade. Métodos diferentes de comparação entre redes PPI ou mesmo formas distintas de comparação entre as distribuições dos autovalores, como por exemplo comparação entre as distâncias de distribuições, também podem ser utilizados. O padrão de "cone" conservado entre organismos (com algumas variações) também sugere uma estrutura semelhante entre todas as redes consideradas.

Finalmente, a associação entre medidas de centralidade e funções biológicas pode apresentar uma grande oportunidade para os estudos desse tipo de rede, porém outras análises ainda precisam ser conduzidos com relação a outras formas de agrupamento, ou ainda considerando outros tipos de informações além da estrutura de conexões, como por exemplo a semelhança entre a estrutura de proteínas. A quantidade ideal de clusters também precisa ser estudada em maiores detalhes, bem como outros métodos de agrupamento que não foram explorados neste trabalho.

Apesar de todos os resultados obtidos aqui, é importante sempre termos em mente

que essas redes ainda estão em construção, todos os pontos ressaltados até o momento podem mudar com a inclusão de proteínas, interações e mudanças na confiança associada àquelas já existentes, indicando a importância de se continuar estudando esse tipo de rede até que se tenha mais certeza das estruturas topológicas associadas a elas. Finalmente esperamos que os resultados apresentados aqui possam ajudar no entendimento das funções desempenhadas e na descoberta de novas proteínas, interações e processos biológicos e no desenvolvimento de modelos, além de incentivar ainda mais o estudo de redes de interação entre proteínas pela comunidade de redes complexas, já que este tipo de estudo pode auxiliar no desenvolvimento de medicamentos, entendimento de algumas doenças e evolução do entendimento das redes complexas no geral.

REFERÊNCIAS

1 NEWMAN, M. **Networks:** an introduction. Oxford: Oxford University Press, 2010.

2 _____. The structure and function of complex networks. **SIAM Review**, v. 45, n. 2, p. 167–256, 2003.

3 BOCCALETTI, S. et al. Complex networks: structure and dynamics. **Physics Reports**, v. 424, n. 4-5, p. 175–308, 2006.

4 ALBERT, R.; BARABASI, A. Statistical mechanics of complex networks. **Reviews of Modern Physics**, v. 74, n. 1, p. 47–97, 2002.

5 DOROGOVTSEV, S.; MENDES, J. Evolution of networks. **Advances in Physics**, v. 51, n. 4, p. 1079–1187, 2002.

6 DOROGOVTSEV, S. N.; GOLTSEV, A. V.; MENDES, J. F. F. Critical phenomena in complex networks. **Reviews of Modern Physics**, v. 80, n. 4, p. 1275–1335, 2008.

7 STROGATZ, S. Exploring complex networks. **Nature**, v. 410, n. 6825, p. 268–276, 2001.

8 COSTA, L. d. F. et al. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. **Advances in Physics**, v. 60, n. 3, p. 329–412, 2011.

9 RAVASZ, E.; BARABASI, A. Hierarchical organization in complex networks. **Physical Review E**, v. 67, n. 2, 2, 2003.

10 COSTA, L. d. F. et al. Characterization of complex networks: a survey of measurements. **Advances in Physics**, v. 56, n. 1, p. 167–242, 2007.

11 ARENAS, A. et al. Synchronization in complex networks. **Physics Reports**, v. 469, n. 3, p. 93–153, 2008.

12 PASTOR-SATORRAS, R. et al. Epidemic processes in complex networks. **Reviews of Modern Physics**, v. 87, n. 3, p. 925–979, 2015.

13 LILJEROS, F. et al. The web of human sexual contacts. **Nature**, v. 411, n. 6840, p. 907–908, 2001.

14 DODDS, P.; MUHAMAD, R.; WATTS, D. An experimental study of search in global social networks. **Science**, v. 301, n. 5634, p. 827–829, 2003.

15 LUSSEAU, D. The emergent properties of a dolphin social network. **Proceedings of The Royal Society B:** biological sciences, v. 270, n. 2, p. S186–S188, 2003.

16 FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the internet topology. In: CONFERENCE: APPLICATIONS, TECHNOLOGIES, ARCHITECTURES, AND PROTOCOLS FOR COMPUTER COMMUNICATIONS SIGCOMM, 99, 1999, Cambridge, USA. **Proceedings ...** New York: ACM, 1999, p. 251-262.

17 GUIMERA, R. et al. The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 22, p. 7794–7799, 2005.

18 JEONG, H. et al. Lethality and centrality in protein networks. **Nature**, v. 411, n. 6833, p. 41–42, 2001.

19 LEE, T. et al. Transcriptional regulatory networks in saccharomyces cerevisiae. **Science**, v. 298, n. 5594, p. 799–804, 2002.

20 DUNNE, J.; WILLIAMS, R.; MARTINEZ, N. Network structure and biodiversity loss in food webs: robustness increases with connectance. **Ecology Letters**, v. 5, n. 4, p. 558–567, 2002.

21 MOSCA, R. et al. Towards a detailed atlas of protein-protein interactions. **Current Opinion in Structural Biology**, v. 23, n. 6, p. 929–940, 2013.

22 ATHANASIOS, A. et al. Protein-protein interaction (ppi) network: recent advances in drug discovery. **Current Drug Metabolism**, v. 18, n. 1, p. 5–10, 2017.

23 GHIASSIAN, S. D.; MENCHE, J.; BARABASI, A.-L. A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. **Plos Computational Biology**, v. 11, n. 4, 2015. doi:10.1371/journal.pcbi.1004120.

24 SHARMA, A. et al. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. **Human Molecular Genetics**, v. 24, n. 11, p. 3005–3020, 2015.

25 TAYLOR, I. W.; WRANA, J. L. Protein interaction networks in medicine and disease. **Proteomics**, v. 12, n. 10, p. 1706–1716, 2012.

26 JAEGER, S.; ALOY, P. From protein interaction networks to novel therapeutic strategies. **Iubmb Life**, v. 64, n. 6, p. 529–537, 2012.

27 RAIN, J. et al. The protein-protein interaction map of helicobacter pylori. **Nature**, v. 409, n. 6817, p. 211–215, 2001.

28 GIOT, L. et al. A protein interaction map of drosophila melanogaster. **Science**, v. 302, n. 5651, p. 1727–1736, 2003.

29 RAVASZ, E.; BARABASI, A. Hierarchical organization in complex networks. **Physical Review E**, v. 67, n. 2, p. 026112, 2003.

30 CEOL, A. et al. Linking entries in protein interaction database to structured text: the febs letters experiment. **FEBS Letters**, v. 582, n. 8, p. 1171–1177, 2008.

31 BARABASI, A.; OLTVAI, Z. Network biology: understanding the cell's functional organization. **Nature Reviews Genetics**, v. 5, n. 2, p. 101–U15, 2004.

32 TANAKA, R.; YI, T.; DOYLE, J. Some protein interaction data do not exhibit power law statistics. **FEBS Letters**, v. 579, n. 23, p. 5140–5144, 2005.

33 BADER, G.; HOGUE, C. Analyzing yeast protein-protein interaction data obtained from different sources. **Nature Biotechnology**, v. 20, n. 10, p. 991–997, 2002.

34 ELMSALLATI, A.; CLARK, C.; KALITA, J. Global alignment of protein-protein interaction networks: A survey. **IEEE-ACM Transactions on Computational Biology and Bioinformatics**, v. 13, n. 4, p. 689–705, 2016.

35 MALOD-DOGNIN, N.; PRZULJ, N. L-graal: Lagrangian graphlet-based network aligner. **Bioinformatics**, v. 31, n. 13, p. 2182–2189, 2015.

36 VAZQUEZ, A. et al. Global protein function prediction from protein-protein interaction networks. **Nature Biotechnology**, v. 21, n. 6, p. 697–700, 2003.

37 SAMANTA, M.; LIANG, S. Predicting protein functions from redundancies in large-scale protein interaction networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 100, n. 22, p. 12579–12583, 2003.

38 NABIEVA, E. et al. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. **Bioinformatics**, v. 21, n. 1, p. I302–I310, 2005.

39 DAVIS, D. et al. Topology-function conservation in protein-protein interaction networks. **Bioinformatics**, v. 31, n. 10, p. 1632–1639, 2015.

40 UETZ, P. et al. Herpesviral protein networks and their interaction with the human proteome. **Science**, v. 311, n. 5758, p. 239–242, 2006.

41 LANDHERR, A.; FRIEDL, B.; HEIDEMANN, J. A critical review of centrality measures in social networks. **Business & Information Systems Engineering**, v. 2, n. 6, p. 371–385, 2010.

42 KOSCHÜTZKI, D.; SCHREIBER, F. Comparison of centralities for biological networks. 2004. Disponível em: <https://subs.emis.de/LNI/Proceedings/Proceedings53/GI-Proceedings.53-25.pdf>. Acesso em: 23 jan. 2019.

43 MIEGHEM, P. V. **Graph spectra for complex networks**. New York, NY, USA: Cambridge University Press, 2011.

44 BOLLOBÁS, B. **Modern graph theory**. Heidelberg: Springer, 1998. (Graduate texts in mathematics).

45 ALBERT, R.; BARABASI, A. Statistical mechanics of complex networks. **Reviews of Modern Physics**, v. 74, n. 1, p. 47–97, 2002.

46 NEWMAN, M. Models of the small world. **Journal of Statistical Physics**, v. 101, n. 3-4, p. 819–841, 2000.

47 WATTS, D.; STROGATZ, S. Collective dynamics of 'small-world' networks. **Nature**, v. 393, n. 6684, p. 440–442, 1998.

48 MAP of Königsberg from 1651. Disponível em: <https://en.wikipedia.org/wiki/K%C3%B6nigsberg#/media/File:Image-Koenigsberg,_Map_by_Merian-Erben_1652.jpg>. Acesso em: 26 set. 2018.

49 SEVEN Bridges of Königsberg. Disponível em: <https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg>. Acesso em: 26 set. 2018.

50 ESTRADA, E. **The structure of complex networks:** theory and applications. New York, NY, USA: Oxford University Press, Inc., 2011.

51 SEIDMAN, S. Network structure and minimum degree. **Social Networks**, v. 5, n. 3, p. 269–287, 1983.

52 NEWMAN, M. Mixing patterns in networks. **Physical Review E**, v. 67, n. 2, 2, 2003.

53 FOSTER, J. G. et al. Edge direction and the structure of networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 107, n. 24, p. 10815–10820, 2010.

54 NEWMAN, M. The structure of scientific collaboration networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 98, n. 2, p. 404–409, 2001.

55 CORMEN, T. H.; LEISERSON, C. E.; RIVEST, R. L.; STEIN, C. **Introduction to Algorithms**. 3r. ed. Cambridge: The MIT Press, 2009.

56 MERRIS, R. A note on laplacian graph eigenvalues. **Linear Algebra and Its Applications**, v. 285, n. 1-3, p. 33–35, 1998.

57 LI, J.; GUO, J.-M.; SHIU, W. C. Bounds on normalized laplacian eigenvalues of graphs. **Journal of Inequalities and Applications**, v. 2014, n. 316, 2014. doi:10.1186/1029-242X-2014-316.

58 BONACICH, P. Power and centrality: a family of measures. **American Journal of Sociology**, v. 92, n. 5, p. 1170–1182, 1987. Disponível em: <http://www.jstor.org/stable/2780000>. Acesso em: 26 set. 2018.

59 BEAUCHAMP, M. A. An improved index of centrality. **Behavioral Science**, v. 10, n. 2, p. 161–163, 1965.

60 ANTHONISSE, J. M. **The rush in a directed graph**. Amsterdam, 1971. Disponível em:<http://oai.cwi.nl/oai/asset/9791/9791A.pdf>. Acesso em: 07 set. 2018.

61 FREEMAN, L. C. A set of measures of centrality based on betweenness. **Sociometry**, v. 4, n. 1, p. 35–41, 1977.

62 NEWMAN, M.; GIRVAN, M. Finding and evaluating community structure in networks. **Physical Review E**, v. 69, n. 2, p. 026113, 2004.

63 NEWMAN, M. E. A measure of betweenness centrality based on random walks. **Social Networks**, v. 27, n. 1, p. 39–54, 2005.

64 BRANDES, U.; FLEISCHER, D. **Centrality measures based on current flow**. Berlin Heidelberg: Springer, 2005.

65 STEPHENSON, K.; ZELEN, M. Rethinking centrality - methods and examples. **Social Networks**, v. 11, n. 1, p. 1–37, 1989.

66 ERDŐS, P.; RÉNYI, A. On random graphs i. **Publicationes Mathematicae Debrecen**, v. 6, p. 290–297, 1959.

67 _____. On the evolution of random graphs. **Publications of the Mathematical Institute of the Hungarian Academy of Sciences**, v. 5, n. 1, p. 17–60, 1960.

68 BARABÁSI, A.; ALBERT, R. Emergence of scaling in random networks. **Science**, v. 286, n. 5439, p. 509–512, 1999.

69 SNEL, B. et al. String: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. **Nucleic Acids Research**, v. 28, n. 18, p. 3442–3444, 2000.

70 MERING, C. von et al. String: a database of predicted functional associations between proteins. **Nucleic Acids Research**, v. 31, n. 1, p. 258–261, 2003.

71 _____. String: known and predicted protein-protein associations, integrated and transferred across organisms. **Nucleic Acids Research**, v. 33, n. SI, p. D433–D437, 2005.

72 _____. String 7 - recent developments in the integration and prediction of protein interactions. **Nucleic Acids Research**, v. 35, n. SI, p. D358–D362, 2007.

73 JENSEN, L. J. et al. String 8-a global view on proteins and their functional interactions in 630 organisms. **Nucleic Acids Research**, v. 37, p. D412–D416, 2009.

74 SZKLARCZYK, D. et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. **Nucleic Acids Research**, v. 39, n. 1, p. D561–D568, 2011.

75 FRANCESCHINI, A. et al. String v9.1: protein-protein interaction networks, with increased coverage and integration. **Nucleic Acids Research**, v. 41, n. D1, p. D808–D815, 2013.

76 _____. Svd-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. **Bioinformatics**, v. 32, n. 7, p. 1085–1087, 2016.

77 SZKLARCZYK, D. et al. String v10: protein-protein interaction networks, integrated over the tree of life. **Nucleic Acids Research**, v. 43, n. D1, p. D447–D452, 2015.

78 _____. The string database in 2017: quality-controlled protein-protein association networks, made broadly accessible. **Nucleic Acids Research**, v. 45, n. D1, p. D362–D368, 2017.

79 MERING, C. von. Protein–protein interaction networks: assembly and analysis. In: _____. **Bioinformatics**. Singapore: World Scientific, 2012. p. 197–217. Disponível em: <http://www.worldscientific.com/doi/abs/10.1142/9789812838780_0008>. Acesso em: 07 set. 2018.

80 STRING INFO - Scores. Disponível em: <https://string-db.org/cgi/info.pl?UserId=M7TlcVi4nAjt&sessionId=2VNkNu3R5UGF&footer_active_subpage=scores>. Acesso em 05 mar. 2018.

81 SUN, J. et al. Inpreppi: an integrated evaluation method based on genomic context for predicting protein-protein interactions in prokaryotic genomes. **BMC Bioinformatics**, v. 8, 2007. doi:10.1186/1471-2105-8-414.

82 PELLEGRINI, M. et al. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. **Proceedings of the National Academy of Sciences of the United States of America**, v. 96, n. 8, p. 4285–4288, 1999.

83 NOORT, V. van; SNEL, B.; HUYNEN, M. Predicting gene function by conserved co-expression. **Trends in Genetics**, v. 19, n. 5, p. 238–242, 2003.

84 LIMA-MENDEZ, G.; HELDEN, J. van. The powerful law of the power law and other myths in network biology. **Molecular BioSystems**, v. 5, p. 1482–1493, 2009. Disponível em: <http://dx.doi.org/10.1039/B908681A>. Acesso em: 07 set. 2018.

85 RONQUI, J. R. F.; TRAVIESO, G. Analyzing complex networks through correlations in centrality measurements. **Journal of Statistical Mechanics:** theory and experiment, v. 2015, n. 5, p. P05030, 2015. Disponível em: <http://stacks.iop.org/1742-5468/2015/i=5/a=P05030>. Acesso em 29 set. 2018.

86 SCHOCH, D.; VALENTE, T. W.; BRANDES, U. Correlations among centrality indices and a class of uniquely ranked graphs. **Social Networks**, v. 50, p. 46 – 54, 2017. ISSN 0378-8733. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378873316303690>. Acesso em 29 set. 2018.

87 KENDALL, M. A new measure of rank correlation. **Biometrika**, v. 30, n. 1/2, p. 81–93, 1938.

88 CROUX, C.; DEHON, C. Influence functions of the spearman and kendall correlation measures. **Statistichal Methods and Applications**, v. 19, n. 4, p. 497–515, 2010.

89 SPEARMAN, C. The proof and measurement of association between two things. **American Journal of Psychology**, v. 15, p. 72–101, 1904.

90 FARKAS, I. et al. Spectra of "real-world" graphs: Beyond the semicircle law. **Physical Review E**, v. 64, n. 2, 2001. doi:10.1103/PhysRevE.64.026704.

91 CHUNG, F.; LU, L.; VU, V. Spectra of random graphs with given expected degrees. **Proceedings of the National Academy of Sciences of the United States of America**, v. 100, n. 11, p. 6313–6318, 2003.

92 DOROGOVTSEV, S. et al. Spectra of complex networks. **Physical Review E**, v. 68, n. 4, p. 046109, 2003.

93 KUEHN, R. Spectra of sparse random matrices. **Journal of Physics A:** mathematical and theoretical, v. 41, n. 29, p. 295002, 2008.

94 RESTREPO, J.; OTT, E.; HUNT, B. Onset of synchronization in large networks of coupled oscillators. **Physical Review E**, v. 71, n. 3, 2005. doi:10.1103/PhysRevE.71.036151.

95 CHAKRABARTI, D. et al. Epidemic thresholds in real networks. **ACM Transactions on Information and System Security**, v. 10, n. 4, 2008. doi:10.1145/1284680.1284681.

96 CASTELLANO, C.; PASTOR-SATORRAS, R. Thresholds for epidemic spreading in networks. **Physical Review Letters**, v. 105, n. 21, 2010. doi:10.1103/PhysRevLett.105.218701.

97 GOLTSEV, A. V. et al. Localization and spreading of diseases in complex networks. **Physical Review Letters**, v. 109, n. 12, p. 128702, 2012.

98 NADAKUDITI, R. R.; NEWMAN, M. E. J. Spectra of random graphs with arbitrary expected degrees. **Physical Review E**, v. 87, n. 1, 2013. doi:10.1103/PhysRevE.87.012803.

99 MARTIN, T.; ZHANG, X.; NEWMAN, M. E. J. Localization and centrality in networks. **Physical Review E**, v. 90, n. 5, 2014. doi:10.1103/PhysRevE.90.052808.

100 PASTOR-SATORRAS, R.; CASTELLANO, C. Distinct types of eigenvector localization in networks. **Scientific Reports**, v. 6, 2016. doi:10.1038/srep18847.

101 _____. Eigenvector localization in real networks and its implications for epidemic spreading. **Journal of Statistical Physics**, 2018. doi:10.1007/s10955-018-1970-8.

102 CATANZARO, M.; NÁ, M. B.; PASTOR-SATORRAS, R. Generation of uncorrelated random scale-free networks. **Physical Review E**, v. 71, p. 027103, 2005. doi:10.1103/PhysRevE.71.027103.

103 ALI, W. et al. Alignment-free protein interaction network comparison. **Bioinformatics**, v. 30, n. 17, p. I430–I437, 2014.

104 PRZULJ, N. Biological network comparison using graphlet degree distribution. **Bioinformatics**, v. 23, n. 2, p. E177–E183, 2007.

105 ERDMAN, J. Algebraic models for social networks - pattison,p. **International Journal of Comparative Sociology**, v. 37, n. 3-4, p. 319–321, 1996.

106 ONNELA, J.-P. et al. Taxonomies of networks from community structure. **Physical Review E**, v. 86, n. 3, 2012.

107 JOLLIFFE, I. **Principal component analysis**. New York: Springer Verlarg, 2002.

108 PENG, X. et al. Protein-protein interactions: detection, reliability assessment and applications. **Briefings in Bioinformatics**, v. 18, n. 5, p. 798–819, 2017.

109 BEASLEY, T. M.; ERICKSON, S.; ALLISON, D. B. Rank-based inverse normal transformations are increasingly used, but are they merited? **Behavior Genetics**, v. 39, n. 5, p. 580–595, 2009.

110 ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. **Nature Genetics**, v. 25, n. 1, p. 25–29, 2000.

111 KANEHISA, M.; GOTO, S. Kegg: Kyoto encyclopedia of genes and genomes. **Nucleic Acids Research**, v. 28, n. 1, p. 27–30, 2000.

112 KANEHISA, M. et al. Kegg: new perspectives on genomes, pathways, diseases and drugs. **Nucleic Acids Research**, v. 45, n. D1, p. D353–D361, 2017.

113 FINN, R. D. et al. The pfam protein families database: towards a more sustainable future. **Nucleic Acids Research**, v. 44, n. D1, p. D279–D285, 2016.

114 _____. Interpro in 2017-beyond protein family and domain annotations. **Nucleic Acids Research**, v. 45, n. D1, p. D190–D199, 2017.