# Relatório Prático

Disciplina de Redes Complexas - PESC - COPPE - UFRJ

*Vinícius W. Salazar, Prof. Daniel R. Figueiredo*

*Outubro de 2019*

### Introdução

Nesse relatório vamos explorar a análise de redes de alguns conjuntos de dados do livro "Network Science", do professor Albert-Laszló Barabási. Os dados podem ser encontrados **nesta página.** Para cada conjunto, vamos caracterizar métricas como grau, distância e tamanho de componentes conexas. Para isso, foram usadas as bibliotecas NetworkX v2.4[1], Pandas v0.25.4[2] e Matplotlib v3.1.1[3]. O notebook de análise foi escrito com Jupyter v1.0.0[4] e o repositório está **disponível aqui.**

### Conjunto 1 - proteínas

Uma rede representando interações proteína-proteína em leveduras. Cada vértice representa uma proteína, que estão conectadas se interagem fisicamento dentro da célula. Dados originais:

- Yu, H., Braun, P., Yildrim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., ... & Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network. Science, 322(5898), 104-110.

Temos um grafo $G_1 = (V, E)$ com $V = 2017$ e $E = 2929$

#### Graus de $G_1$

- **Máximo ($n = 5$):** [91, 82, 81, 52, 46]
- **Mínimo ($n = 5$):** [1, 1, 1, 1, 1]
- **Média** $\approx 2.904$
- **Mediana** $= 2.0$
- **Desvio padrão** $\approx 4.884$
  **Densidade** de $G_1 \approx 1.44 \times 10^{-}3$

#### Componentes conexas

$G_1$ possui **185 componentes conexas**, sendo que a maior componente conexa $CC_{G1} = (1646, 2681)$ enquanto a segunda e terceira menor possuem 6 e 5 vértices, com a **mediana = 2** para o número de vértices em cada componente conexa, demonstrando outro **padrão de cauda pesada**.

Como $G_1$ não é conectado, vamos medir o diâmetro e calcular e baricentro de $CC_{G1}$:

- $diam(CC_{G1}) = 14$
- $baricentro(CC_{G1}) = V_{1356}$

#### Adamic-Adar

Podemos computar o índice Adamic-Adar $AA_{(u,v)}$ para um par de vértices $(u, v)$ em $G_1$, expressado por:

$$\sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(w)|}$$

, onde $\Gamma(u)$ representa o conjunto de vizinhos de $u$.

Por exemplo:

$AA_{(2015,2016)} \approx 0.33$

Seria interessante aplicar uma métrica de redes sociais em uma rede biológica. Se os dados fossem proteínas de um patógeno, essa métrica poderia ser usada, por exemplo, para auxiliar no design de fármacos.

### Conjunto 2 - metabólitos

Uma rede representando reações metabólicas em bactérias *E. coli*. Cada vértice é um metabólito, e cada link direcionado $A \to B$ representa uma reação onde $A$ é um reagente e $B$ é um produto. Dados originais:

- Schellenberger, J., Park, J. O., Conrad, T. M., & Palsson, B. A. (2010). BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC bioinformatics, 11(1), 213.

Temos um grafo $G_2 = (V, E)$ com $V = 1039$ e $E = 5801$

### Graus de $G_2$

- **Máximo ($n = 5$):** [906, 522, 337, 289, 279]
- **Mínimo ($n = 5$):** [1, 1, 2, 2, 2]
- **Média** $\approx 11.167$
- **Mediana** $= 6.0$
- **Desvio padrão** $\approx 37.621$
  **Densidade** de $G_2 \approx 5.38 \times 10^{-}3$

### Componentes conexas

$G_2$ não é fortemente conexo, porém é fracamente conexo. Possui **147 componentes fortemente conexas**, cuja maior é representada por:

- $CC_{G2} = (893, 5435)$
- $diam(CC_{G2}) = 8$
- $baricentro(CC_{G2}) = V_{592}$

### Caminho mais curto

Podemos computar um caminho mais curto $S_{(i,j)}$ dentro de $CC_{G2}$ como por exemplo $S_{(991,749)} = [991, 735, 589, 1003, 750, 749]$. Note que o caminho inverso $S_{(j,i)}$ pode ser diferente ou nem existir. Nesse caso, é mais curto: $S_{(749,991)} = [749, 589, 992, 991]$. **Isso poderia ser um algoritmo útil para calcular vias metabólicas, desde as moléculas precursoras até o composto final.**

### Conjunto 3 - WWW

Vértices representam páginas web da Universidade de Notre Dame sob o domínio nd.edu, e links direcionados apresentam hyperlinks entre elas. Dados coletados em 1999. Referência:

- Albert, R., Jeong, H., & Barabási, A. L. (1999). Internet: Diameter of the world-wide web. Nature, 401(6749), 130-131

$G_3 = (325729, 1117562)$

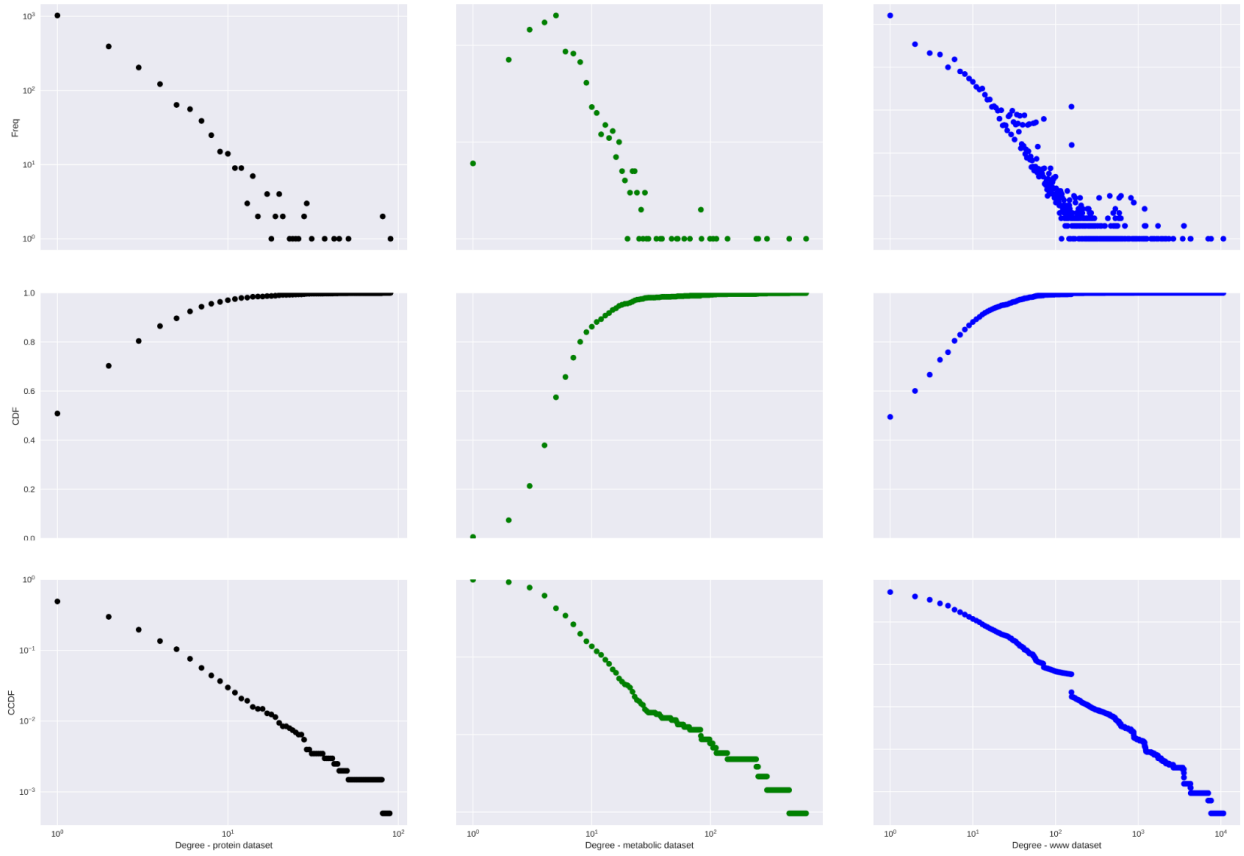


Figure 1: Frequência absoluta (primeira linha), distribuição cumulativa (segunda linha) e distribuição cumulativa complementar (terceira linha) dos graus para $G_1$ (preto), $G_2$ (verde) e $G_3$ (azul).

### Graus de $G_3$

- **Máximo** ($n = 5$)**:** [10721, 7622, 7026, 4317, 4282]
- **Mínimo** ($n = 5$)**:** [1, 1, 1, 1, 1]
- **Média** $\approx 6.862$
- **Mediana** $= 2.0$
- **Desvio padrão** $\approx 42.928$
  **Densidade** de $G_3 \approx 1.411 \times 10^{-}5$

### Componentes conexas

$G_2$ não é fortemente conexo, porém é fracamente conexo. Possui **203609 componentes fortemente conexas**, cuja maior, $CC_{G3}$ é descrita por $CC_{G3} = (53968, 304684)$.

### PageRank

Podemos calcular o PageRank $PR(A)$ aonde $A$ é um vértice de $G$. Em $CC_{G3}$, o vértice com maior índice de PageRank é $PR(V_0) \approx 0.0322 \mid V_0 \in CC_{G3}$ e o grau de $V_0 = 7633$. Se olharmos para $G_3$, isso muda, com o maior índice sendo o de $PR(V_{1963}) \approx 0.01$, e $PR(V_0) \approx 0.005 \mid V_0 \in G_3$.

**Conjunto 4 - Citations**

Uma rede de citações das revistas APS (Physical Review Letters, Physical Review, e Reviews of Modern Physics). Cada vértice representa um artigo, e existe um link direcionando ligando o vértice A ao vértice B, se A cita B. Dados originais:

- Redner, S. (2004). Citation statistics from more than a century of physical review. arXiv preprint physics/0407137.

$G_4 = (325729, 1117562)$

**Graus de $G_4$**

- **Máximo ($n = 5$):** [4767, 3717, 3248, 3108, 2652]
- **Mínimo ($n = 5$):** [1, 1, 1, 1, 1]
- **Média** $\approx 20.856$
- **Mediana** $= 15.0$
- **Desvio padrão** $\approx 31.948$ **Densidade** de $G_4 \approx 2.32 \times 10^{-}5$

**Componentes conexas**

$G_4$ não é fortemente conexo, e nem fracamente conexo. Possui 356584 componentes fortemente conexas, cuja maior, $CC_{G4}$ é descrita por $CC_{G4} = (90965, 743145)$. A maior componente fracamente conexa é descrita por $wkCC_{G4} = (448817, 4688894)$.

**Discussão**

Foi empregado o pacote NetworkX junto com bibliotecas de Python científico para análise de conjuntos de dados modelados como grafos. O NetworkX (NX) é bem intuitivo, provendo vários métodos, funções e algoritmos práticos para ciência das redes. Os conjuntos de dados do livro do Barabási também são bem interessantes. Algo que me fez optar pelos conjuntos desse livro foi que, embora só hajam duas colunas, ele explica o que cada uma significa. No entanto, futuramente quero explorar conjuntos mais elaborados, que tenham atributos para os vértices, e também explorar outras visualizações além das distribuições. As figuras de distribuição de grau permitiram a visualização do fenômeno de cauda pesada. Isso também foi observado no tamanho das componentes conexas, como em $G_1$ onde a maior componente conexa é de maior que a segunda e terceira em duas ordens de grandeza. Esse foi um relatório que eu achei bem interessante de fazer, no entanto o limite de páginas foi um pouco restritivo.

**Referências**


1. Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart. **Exploring network structure, dynamics, and function using NetworkX**, Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15. (2008)
2. Wes McKinney. **Data Structures for Statistical Computing in Python**, Proceedings of the 9th Python in Science Conference, 51-56 (2010)
3. John D. Hunter. **Matplotlib: A 2D Graphics Environment**, Computing in Science & Engineering, 9, 90-95 (2007), DOI: 10.1109/MCSE.2007.55
4. Kluyver, Thomas, et al. **Jupyter Notebooks-a publishing format for reproducible computational workflows.** ELPUB. (2016)