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].

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.

Nosso arquivo de input é no formato .tsv com duas colunas, onde cada coluna representa um vértice e cada linha representa uma aresta.

Temos um grafo $G_1 = (V, E)$ com $V = 2017$ e $E = 2929$. Vamos analisar algumas métricas:

Graus de G_1

- **Máximo ($n = 5$):** [91, 82, 81, 52, 46]
- **Mínimo ($n = 5$):** [1, 1, 1, 1, 1]
- **Média** ≈ 2.904
- **Mediana** $= 2.0$
- **Desvio padrão** ≈ 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 têm 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 $AA_{(i,j)}$ para um par de vértices (i, j) em G_1, por exemplo:

$AA_{(6,249)} \approx 0.71$

Seria interessante aplicar uma métrica de redes sociais em uma rede biológica. Se fossem proteínas de um patógeno, isso poderia ser usado para, por exemplo, 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. Ã~. (2010). BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC bioinformatics, 11(1), 213.

Nosso arquivo de input é um .tsv com duas colunas, onde cada coluna representa um vértice e cada linha representa uma aresta **direcionada**.

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** ≈ 11.167
- **Mediana** $= 6.0$
- **Desvio padrão** ≈ 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

$G_3 = (325729, 1117562)$

Graus de G_3

- **Máximo ($n = 5$):** [10721, 7622, 7026, 4317, 4282]
- **Mínimo ($n = 5$):** [1, 1, 1, 1, 1]
- **Média** ≈ 6.862
- **Mediana** $= 2.0$
- **Desvio padrão** ≈ 42.928

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 $PR(V_0) \approx 0.005 \mid V_0 \in G_3$.

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, Aug 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