TRABALHO #3 – ESTRUTURAS DE DADOS II

EVANDRO E. S. RUIZ

As redes complexas são representações de iterações naturais entre as mais variadas entidades. Essas redes apropriam-se do tipo abstrato de dado grafo para caracterizar estruturas muito complexas tanto na sua estrutura como na sua dinâmica. Este projeto tem como objetivo principal expandir o conhecimento sobre estruturas de redes complexas por meio da aplicação dos algoritmos que ajudam a definir estas estruturas.

1. Motivação

Este semestre faremos vários trabalhos exploratórios sobre redes complexas. Teremos vários conjuntos de dados representativos de várias redes complexas com dois objetivos: 1) caracterizar estas redes, e; 2) desvendar informações sobre estas redes.
Para dinamizar este desafio, os vários *datasets* serão sorteados entre os grupos.

2. Enunciado

Os dados. Segue abaixo uma lista de vários *datasets*. Cada grupo se encarregará de tratar apenas um destes conjuntos de dados objetivando cumprir os dois objetivos acima. São estes os conjuntos:

(1) `musae-github` Uma grande rede social de desenvolvedores do GitHub, coletada da API pública em junho de 2019, com mais de 37 mil vértices.
`https://snap.stanford.edu/data/github-social.html`

(2) `feather-deezer-social` Uma rede social de usuários do Deezer, coletada da API pública em março de 2020, com mais de 28 mil vértices.
`https://snap.stanford.edu/data/feather-deezer-social.html`

(3) `com-DBLP` DBLP é um repositório bibliográfico de Ciência da Computação hospedado na *Universität Trier*, na Alemanha. Esta é uma rede de co-autoria com mais de 300 mil nós.
`https://snap.stanford.edu/data/com-DBLP.html`

(4) **Six Degrees of Francis Bacon** A famosa lista que é é uma reconstrução digital da rede social moderna com mais de 15 mil atores.
`https://www.kaggle.com/datasets/rtatman/six-degrees-of-francis-bacon`

(5) **Trade Network** Transações comerciais entre 163 países nos anos de 2.000, 2.005, 2.010, 2.015 e 2.018.
`https://www.kaggle.com/datasets/yasirtariq/tradenetwork`

(6) **Python Dependency Network** Redes de dependência de pacotes do Python. Os nós da rede são pacotes Python registrados no PyPI e as arestas são dependências entre pacotes. Uma rede com mais de 58 mil nós.
`https://icon.colorado.edu/#!/networks`

(7) **DBpedia work-genre network** Uma rede bipartida de afiliações entre artistas e suas obras, por um lado, e classificações de gênero, por outro, extraída da Wikipédia pelo projeto DBpedia. Mais de 260 mil nós.
`https://icon.colorado.edu/#!/networks`

(8) **Reuters Sept 11 news (2001)** Uma sequência de redes diárias de palavras que apareceram nas notícias da Reuters publicadas em cada um dos 66 dias consecutivos após

os ataques terroristas de 11 de setembro de 2001. Nós são palavras, existe uma aresta se duas palavras aparecerem na mesma frase, e o peso da aresta representa a frequência dessa co-ocorrência. Cerca de 13 mil vértices.
`https://icon.colorado.edu/#!/networks`

(9) **p2p-Gnutella24** Uma sequência de instantâneos da rede de compartilhamento de arquivos peer-to-peer Gnutella de agosto de 2002. Mis de 26 mil nós.
`https://snap.stanford.edu/data/p2p-Gnutella24.html`

(10) **ego-Twitter** Este conjunto de dados consiste em 'círculos' (ou 'listas') do Twitter. O conjunto de dados inclui recursos de nós (perfis), círculos e redes de ego. Mais de 80 mil vértices.
`https://snap.stanford.edu/data/ego-Twitter.html`

Qual conjunto de dados usar. O estudante deverá usar o conjunto de dados referente o último dígito de seu número USP. Por exemplo, para o estudante com número USP 12345 o conjunto de dados usados deverá ser o de número 5, **Trade Network**. Caso o dígito seja '0' (zero), o tema será o 10. Caso seja um trabalho com dois autores, o tema deve seguir o menor dígito entre os últimos dígitos do par. Por exemplo, para dois números USP, 12345 e 67890, o tema será o 10, referente ao dígito '0'.
Observação: Os alunos que vieram na aula no primeiro dia de dezembro puderam escolher o conjunto de dados que irão trabalhar.

Os desafios. Solicita-se assim um trabalho a ser desenvolvido sobre 3 (três) desafios, que são:

Desafio 1. Caracterização da rede complexa no âmbito dos vértices (nós da rede). Solicita-se:

(1) Elaborar um histograma dos graus dos nós (para grafos direcionados somar k^{in} e k^{out};
(2) Descobrir e mostrar os vértices com grau máximo e mínimo como também mostrar o grau médio entre os nós;
(3) Fazer uma histograma da força dos nós. Elencar os 5 nós 'mais fortes';
(4) Elaborar uma comparação entre o grau dos nós e a sua força;
(5) Elucidar os 5 vértices de maior centralidade de intermediação (*betweenness centrality*); e
(6) Calcular o PageRank[1] dos três vértices mais citados em cada grafo.

Desafio 2. Caracterizar as medidas globais da rede. Solicita-se:

(1) Descrever o número de nós e arestas da rede;
(2) Mostrar densidade da rede;
(3) Descrever o número de componentes conexos (fortes e fracos), além do número de nós e arestas para o maior destes componentes;;
(4) Baseado na estatística de distribuição dos nós, explique qual modelo esta rede segue;
(5) Mostrar o diâmetro da rede e o *average path length*; e
(6) Calcular a transitividade da rede.

Desafio 3. Este desafio consiste em retirar o máximo de informação conclusiva ou comparativa da rede estudada. Por exemplo, mostrar a resiliência da rede quando sob ataques aleatórios ou direcionados. Vamos exemplificar com algumas informações esperadas para alguns tipos específicos de redes:

Trade network: Mostrar a evolução da rede comercial mundial através da análise comparativa de todos os dados anuais.

Six Degrees of Francis Bacon: Grupos diferentes têm graus diferentes de conectividade?

Reuters Sept 11 news (2001): Mais citadas são substantivos ou adjetivos? Podemos recuperar expressões muito utilizadas desta lista?

[1]O PageRank calcula uma classificação dos nós de um grafo com base na estrutura dos links de entrada. Ele foi originalmente projetado como um algoritmo para classificar páginas da web.

DBpedia work-genre network: Podemos afirmar qual o artista mais 'conectado' dadas suas afiliações?

3. Avaliação

A avaliação deste trabalho será composta pela avaliação dos três desafios, sendo 40% da nota para os Desafios 1 e 2 e 20% da nota para o Desafio 3.

4. Importante

- Sugiro o uso do pacote NetworkX, `https://networkx.org/`, para o tratamento dos arquivos e execução das tarefas propostas;
- Os trabalhos são individuais, mas admite-se a possibilidade de ser produzido por, no máximo, dois autores;
- Caso o trabalho tenha dois autores, há a necessidade de explicar e detalhar qual a contribuição de cada um;
- Os trabalhos deverão estar na versão *notebook* do Google Colab e compartilhados com evandro@usp.br;
- Os trabalhos serão entregues via email (evandro@usp.br), incluindo o link do compartilhamento, até o dia 8 de dezembro às 23h59;
- Para que o seu email seja localizado utilize o seguinte assunto no email: `Trab-3 AED2`.

Bom trabalho!

Departamento de Computação e Matemática, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo – USP
E-mail address: `evandro@usp.br`