Nome: Vinícius Silva Campos

DRE: 113023327

Relatório trabalho prático 1

1. Introdução:

Junto com a ampliação da capacidade de processamento e armazenamento de dados ocorreu um aumento na difusão de informação por meio da infraestrutura da internet. A partir da criação da World Wide Web na década de 1990, tivemos um aumento exponencial no número de sites e a análise de sua estrutura passou a ser utilizada por profissionais no ramo científico e na indústria.
Devido ao meu interesse pela área resolvi optar por utilizar grafos retirados da Web disponibilizados pela Universidade de Stanford, cujos links estão listados abaixo:

- http://snap.stanford.edu/data/web-Google.html
- http://snap.stanford.edu/data/web-NotreDame.html
- http://snap.stanford.edu/data/web-Stanford.html

Diversos são os caminhos que auxiliam no entendimento de grafos, seja por meio da análise de sua estrutura ou de fenômenos como epidemias que se propagam ao longo dessa malha, por exemplo. Para esse fim, diversas métricas são utilizadas, dentre elas podemos citar: grau, distância, tamanho das componentes conexas, clusterização, closeness, betweenness, page rank, etc. Nesse relatório optei por utilizar as seguintes métricas: grau, tamanho das componentes conexas, local clustering e page rank. A justificativa utilizada para optar pelas mesmas métricas para o estudo dos três grafos escolhidos é a de que como pertencem à mesma classe de redes devem compartilhar propriedades em comum.

2. Stanford web graph

Esse grafo representa a estrutura da rede do domínio da Universidade de Stanford (stanford.edu) em 2002. Os nós representam as páginas e as arestas (direcionadas) representam a existência de um *hiperlink* entre elas. O grafo possui 281903 nós e 2312497 arestas.
Conforme podemos observar pelos resultados obtidos sobre o tamanho das componentes conexas, a rede em questão não é fortemente conexa. Verificamos também que a maior componente conexa possui mais de 50% de todos os nós da rede e mais de 80% das componentes conexas possuem tamanho inferior a 5000. Isso explica os valores obtidos para o desvio padrão e a média do tamanho das componentes conexas. Além disso, temos indícios da existência super hubs na rede. Essa hipótese é sustentada ao observarmos que, apesar de o grau máximo e o desvio

padrão do grau de entrada serem altos a Figura 2 aponta que praticamente todos os vértices possuem grau de entrada inferior a 5000. Isso também se verifica nos resultados do *page rank*, pois ao realizarmos um passeio aleatório pelo grafo com essa estrutura esperaríamos ter uma probabilidade máxima de visitar um site alta. Considerando que nosso grafo possui 281903 nós o valor máximo obtido de 0.011 muito superior ao desvio padrão e à média evidencia a presença de pelo menos um super hub. Isso também é evidenciado pela Figura 4, onde vemos que quase a totalidade dos vértices possuem índice de clusterização próximo de 0. Além disso, podemos verificar a presença de um índice de clusterização média alto, ou seja, dado que um site A tenha relacionamento com B e C faz com que B e C se relacionem com uma probabilidade considerável.

2.1 Resultados:

	max	std-deviation	mean	min
Grau de entrada	38606	166.33	8.20	1
Grau de saída	255	11.31	8.20	0
Tamanho das Componentes	150532	871.54	9.42	1
Local Clustering	2	0.45	0.59	0
Page Rank	0.011	4.218e-05	3.538e-06	5.321e-07

Tabela 1. Resultados das métricas calculadas para a rede Stanford web graph

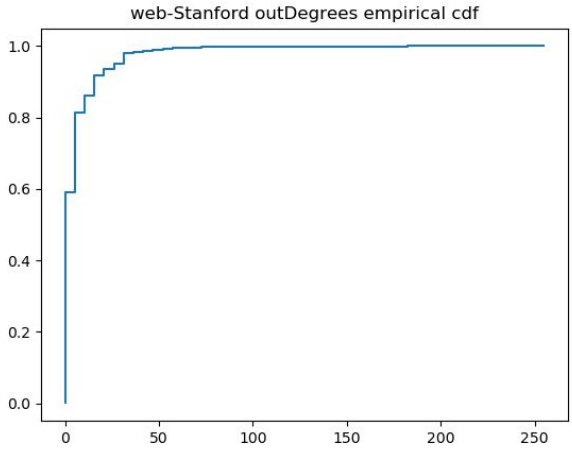

Figura 1 - ECDF dos graus de saída do grafo Stanford web graph

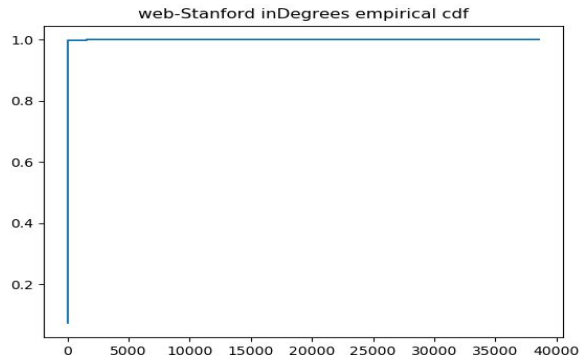

Figura 2 - ECDF dos graus de entrada do grafo Stanford web graph

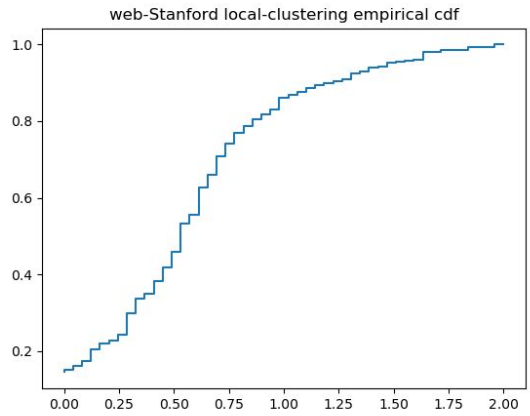

Figura 3 - ECDF dos coeficientes de clusterização locais do grafo Stanford web graph

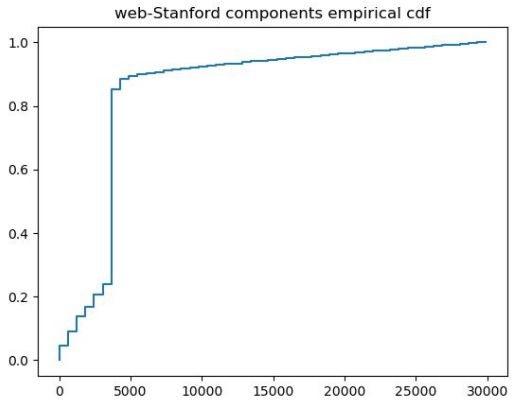

Figura 4 - ECDF do tamanho das componentes conexas do grafo Stanford web graph

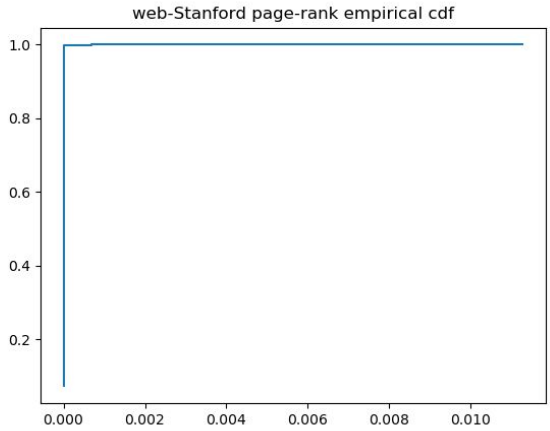

Figura 5 - ECDF page rank do grafo Stanford web graph

3. Notre Dame web graph

Esse grafo representa a estrutura da rede do domínio da Universidade de Notre Dame (nd.edu) em 1999. Os nós representam as páginas e as arestas (direcionadas) representam a existência de um *hiperlink* entre elas. Essa rede possui 325729 nós e 1497134 arestas.

A rede estudada tem diversas propriedades e resultados semelhantes com a anterior. Dentre as semelhanças podemos citar o fato de não ser fortemente conexa. Porém, podemos notar a partir da Figura 8 que possuímos um comportamento de aumento gradual do tamanho das componentes conexas o que pode indicar um número maior de componentes conexas em relação a anterior. Levando em consideração isso e os resultados relativos aos graus de entrada e saída dos vértices bem como suas respectivas ECDF's (Figuras 6 e 7) que as componentes conexas possuem um grau de densidade local maior evidenciado também pela Figura 8.

3.1 Resultados:

	max	std-deviation	mean	min
Grau de entrada	10721	39.05	4.60	1
Grau de saída	3445	21.48	4.60	0
Tamanho das Componentes	53968	123.96	1.60	1
Local Clustering	2.00	0.37	0.22	0.0
Page Rank	0.002	9.921e-06	1.160e-06	4.606e-07

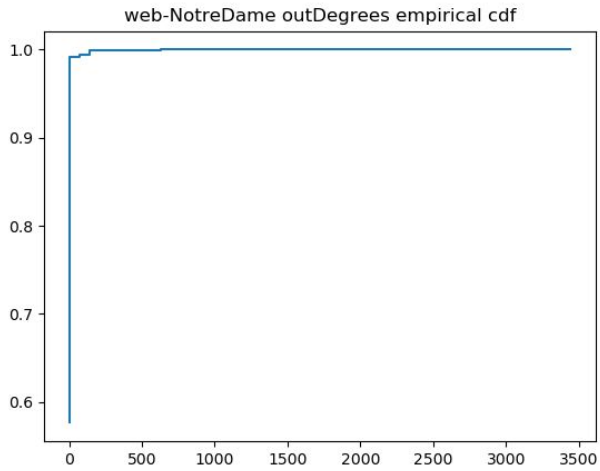

Figura 6 - ECDF dos graus de saída do grafo Notre Dame web graph

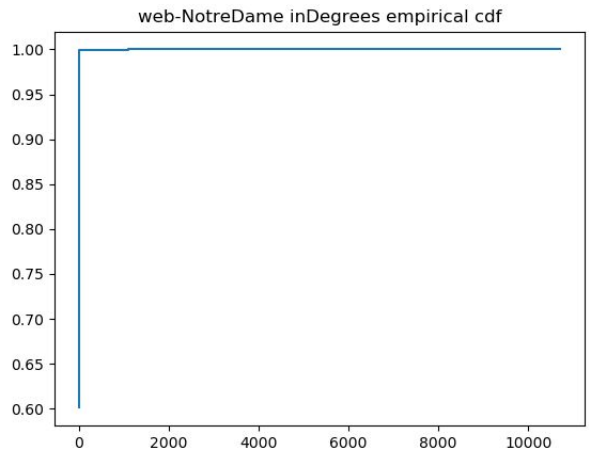

Figura 7 - ECDF dos graus de entrada do grafo Notre Dame web graph

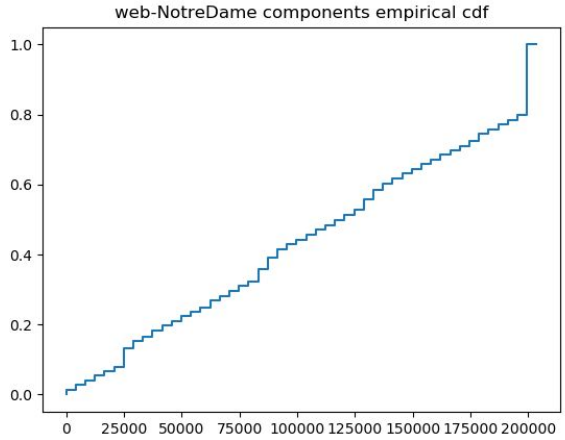

Figura 7 - ECDF dos tamanhos das componentes do grafo Notre Dame web graph

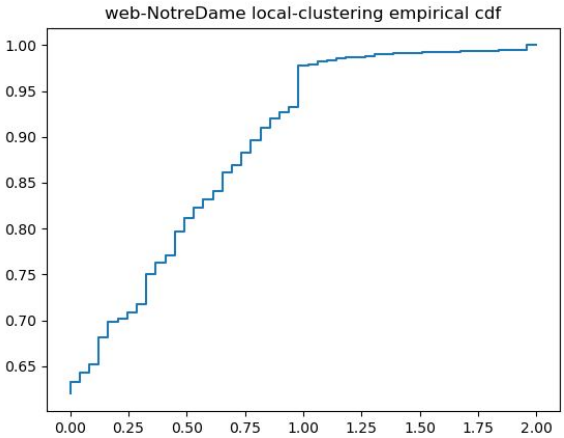

Figura 8 - ECDF dos coeficientes de clustering locais do grafo Notre Dame web graph

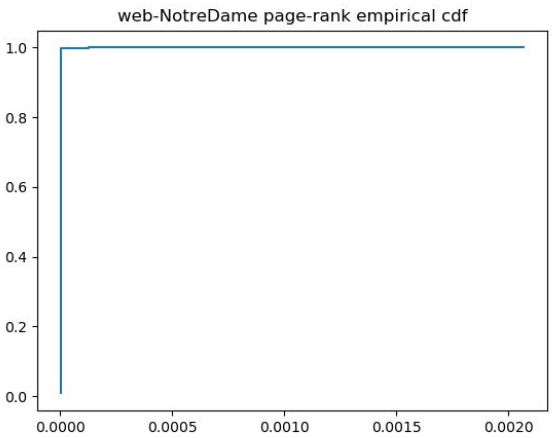

Figura 9 - ECDF do page rank do grafo Notre Dame web graph

4. Google web graph

Os nós dessa rede representam as páginas da web e as areastas direcionadas os hyperlinks entre elas. Esses dados foram divulgados em 2002 pela Google como parte do *Google Programming Contest*. A rede possui 875713 nós e 5105039 arestas direcionadas.
Essa rede possui uma estrutura mais parecida com a da primeira que com a da segunda. Baseio essa minha hipótese na ECDF do coeficiente local de clustering dela e da rede de Stanford (Figuras 3 e 13, respectivamente). Além disso, as distribuições dos graus de entrada, saída e do *page rank* parecem ser semelhantes. Porém, ao contrário da outra, a ECDF do tamanho das componentes conexas possui um degrau e posteriormente continua a subir (Figura 12) enquanto na Figura 4 após o degrau o valor da ECDF já está próximo de 1 o que deve indicar que essa rede possui um número maior de componentes conexas grandes.

4.1 Resultados:

	max	std-deviation	mean	min
Grau de entrada	6326	38.37	2.22	1
Grau de saída	456	6.56	5.57	0
Tamanho das Componentes	434818	677.03	2.22	1
Local Clustering	2.0	0.54	0.57	0.0
Page Rank	0.0006	4.1511e-06	7.34630e-07	1.63686e-07

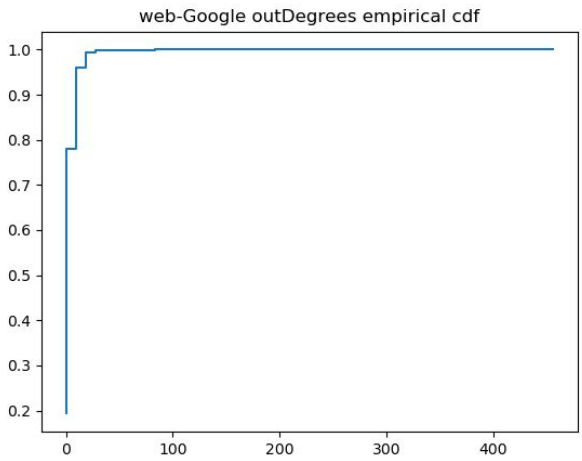

Figura 10 - ECDF do grau de saída do grafo Google web graph

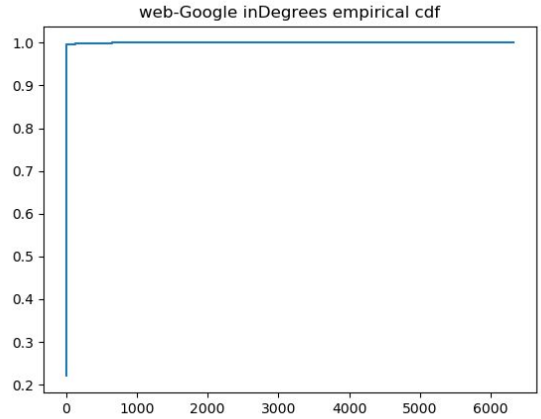

Figura 11 - ECDF do grau de entrada do grafo Google web graph

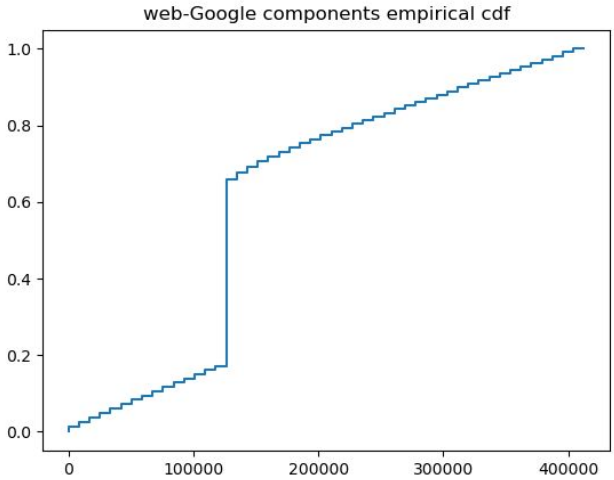

Figura 12 - ECDF do tamanho das componentes conexas do grafo Google web graph

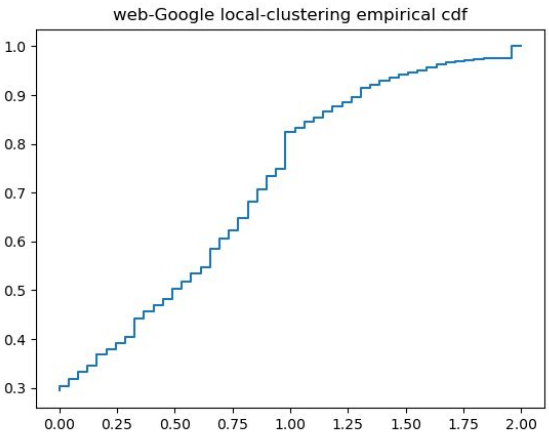

Figura 13 - ECDF do coeficiente de clustering local do grafo Google web graph

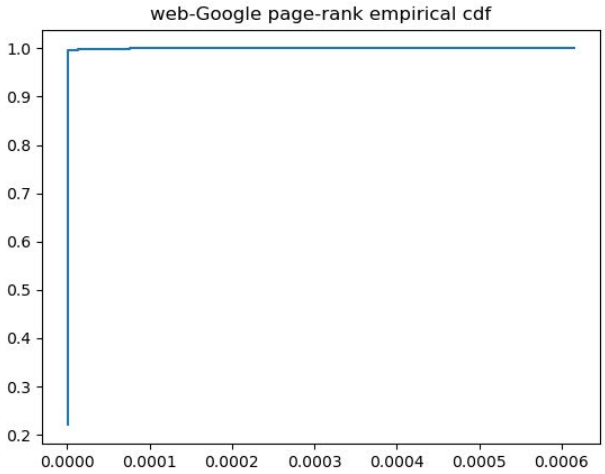

Figura 14 - ECDF do page rank do grafo Google web graph