Um estudo de Redes Complexas utilizando o DataSet 'Higgs Twitter'

Bruno Gabriel J. Santos, Maria Vitória R. Mendes

1

Resumo. *O então trabalho tem por objetivo o estudo do tema de redes complexas, uma importante estrutura de dados. A principal proposta é a análise da Rede "Higgs Twitter", a qual é composta por twitters envolvendo o tema do Boson de Higgs. Desta forma, utilizou-se a linguagem de programação Python e a biblioteca Networkx para a obtenção das informações propostas*

1. Introdução: Conceitos importantes

As medidas de centralidade nos ajudam a mapear poder, influência, controle e status em uma rede de relacionamentos.
O número de arestas (edges) é um indicativo da memória necessária para armazenar a rede.
O grau de centralidade (degree centrality) indica os nós que têm mais links com outros nós na rede, e partir dele é possível fazer classificação de popularidade. A avaliação a partir dessa medida é individual e direta, ou seja, nos mostra as influências diretas de um nó, pois obtemos quantas conexões chegam (in-degree) e saem (out-degree) do nó.
A medida de betweenness centrality indica quais nós podem agir como ponte entre os nós, pois nós com alta betweenness centrality aparecem mais nos shortest-paths. A fim de obter os valores, são calculados os menores caminhos e conta-se a frequência de cada nó nesses shortest paths. Por causa disso, essa medida pode indicar quem tem mais controle sobre o fluxo de informação e, por consequência, desbalancear o fluxo caso seja removido.
A medida de closeness é obtida a partir do cálculo da proximidade de um nó com outros nós, que é obtida a partir de uma pontuação dada a cada nó a partir da soma de seus menores caminhos. Em uma rede fortemente conectada, a métrica de closeness é semelhante entre os nós. Esse dado é útil para saber quem consegue obter informações de outros nós mais eficientemente, e também espalhar informações rapidamente, além de encontrar o nó mais influente em um cluster.
A medida de eigenvector centrality é útil na análise macro da rede a fim de encontrar quem tem uma influência ampla, já que são avaliados os links do nó, das conexões do nó e assim por diante. Portanto, permite avaliar a influência e sua distribuição.
A medida de page rank também ajuda a avaliar a influência indireta de um nó ao identificar a significância de um nó pontuando-o de acordo com o in-degree. É relevante para avaliar citações e autoridade sobre a informação.
O coeficiente de agrupamento é uma métrica de suma importância na análise de redes complexas, visto que ele fornece o grau com que os nós estão agrupados. Tal efeito pode ser denominado como efeito de small world. Desta forma, na análise do processo, o coeficiente refere-se ao círculo social estabelecido com os usuários do Twitter.
A média do coeficiente estabelece, de certo modo, o quanto os usuários estão conectados.
O diâmetro é definido como o maior caminho entre dois pares de vértices A e B.

O número de componentes conexos fortes representa um subgrafo presente na rede que possui uma alta correlação entre seus vértices internos e uma relação fraca com o restante da rede. Já o número de componentes conexos fracos refere-se aos componentes que estão conectados por no mínimo uma aresta. Em um grafo de rede social, podemos entender que um triangulo, sendo 3 vértices conectadas, representam os usuários que possuem uma forte relação entre si. Então é provavél que ocorra uma maior quantidade de twittes/compartilhamento e interações entre os mesmos.Logo, uma rede com muitos triângulos significaria uma densidade de usuários intimamente relacionados entre si.

2. Fase 1

Utilizando o recurso do Google Colab Pro, foi possível calcular algumas medidas utilizando a biblioteca Networkx. Seguem abaixo:

Número de nós: 456.626 Número de arestas: 14.855.842
Densidade: 7.124870092245855e-05

DEGREE

Max degree: ('1503', 51388)

Top 20 biggest: [('1503', 51388), ('206', 48475), ('88', 45349), ('138', 44190), ('1062', 40160), ('677', 39870), ('352', 39586), ('220', 39251), ('317', 37855), ('301', 37844), ('383', 35277), ('8', 32160), ('1274', 30086), ('15', 28945), ('6948', 28112), ('7533', 27903), ('1988', 27782), ('3549', 27641), ('6', 27189), ('2055', 26724)]

Min degree: ('456626', 1)

Top 20 smallest: [('456579', 1), ('456581', 1), ('456584', 1), ('456587', 1), ('456590', 1), ('456592', 1), ('456598', 1), ('456601', 1), ('456605', 1), ('456607', 1), ('456610', 1), ('456612', 1), ('456614', 1), ('456615', 1), ('456617', 1), ('456620', 1), ('456621', 1), ('456624', 1), ('456625', 1), ('456626', 1)]

Sum of degrees: 29711684
Mean of degrees: 65.06787611743528

IN DEGREE

Max in-degree: ('1503', 51386)
Min in-degree: ('456626', 0)

Top 20+: [('1503', 51386), ('206', 48414), ('88', 45221), ('138', 44188), ('1062', 40120), ('677', 39820), ('352', 39527), ('220', 39227), ('317', 37848), ('301', 37730), ('383', 35182), ('8', 32106), ('1274', 29995), ('15', 28844), ('6948', 28038), ('7533', 27810), ('3549', 27555), ('6', 27088), ('1988', 27065), ('2055', 26685)]

Top 20-: [('456607', 0), ('456608', 0), ('456609', 0), ('456610', 0), ('456611', 0), ('456612', 0), ('456613', 0), ('456614', 0), ('456615', 0), ('456616', 0), ('456617', 0), ('456618', 0), ('456619', 0), ('456620', 0), ('456621', 0), ('456622', 0), ('456623',

0), (’456624’, 0), (’456625’, 0), (’456626’, 0)]

Sum of in degrees: 14855842
Mean in degrees: 32.53393805871764

OUT DEGREE

Max out-degree: (’13115’, 1259)
Min out-degree: (’456626’, 1)

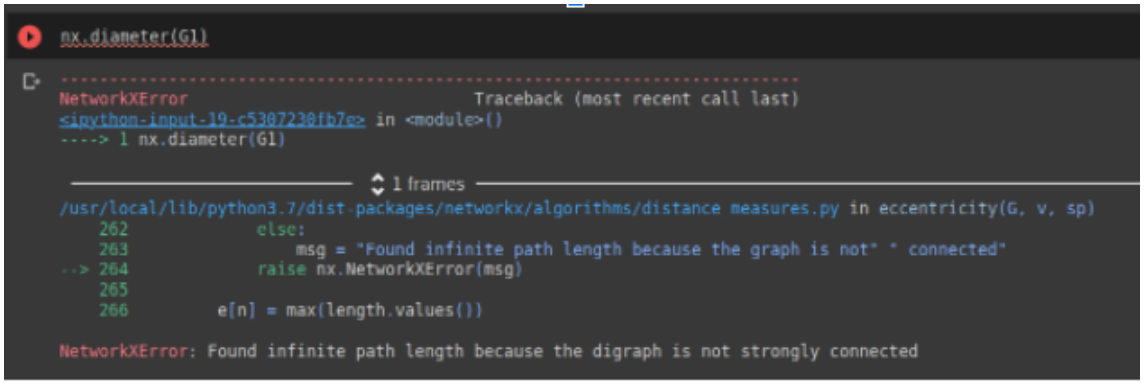
Top 20+: [(’13115’, 1259), (’49180’, 1155), (’50338’, 1097), (’1984’, 1093), (’3628’, 1027), (’16092’, 985), (’5226’, 924), (’26150’, 921), (’21386’, 899), (’7643’, 877), (’35673’, 846), (’6560’, 843), (’20379’, 833), (’4177’, 814), (’83217’, 808), (’502’, 798), (’40973’, 798), (’26095’, 796), (’2687’, 795), (’27290’, 787)]

Top 20-: [(’456579’, 1), (’456581’, 1), (’456584’, 1), (’456587’, 1), (’456590’, 1), (’456592’, 1), (’456598’, 1), (’456601’, 1), (’456605’, 1), (’456607’, 1), (’456610’, 1), (’456612’, 1), (’456614’, 1), (’456615’, 1), (’456617’, 1), (’456620’, 1), (’456621’, 1), (’456624’, 1), (’456625’, 1), (’456626’, 1)]

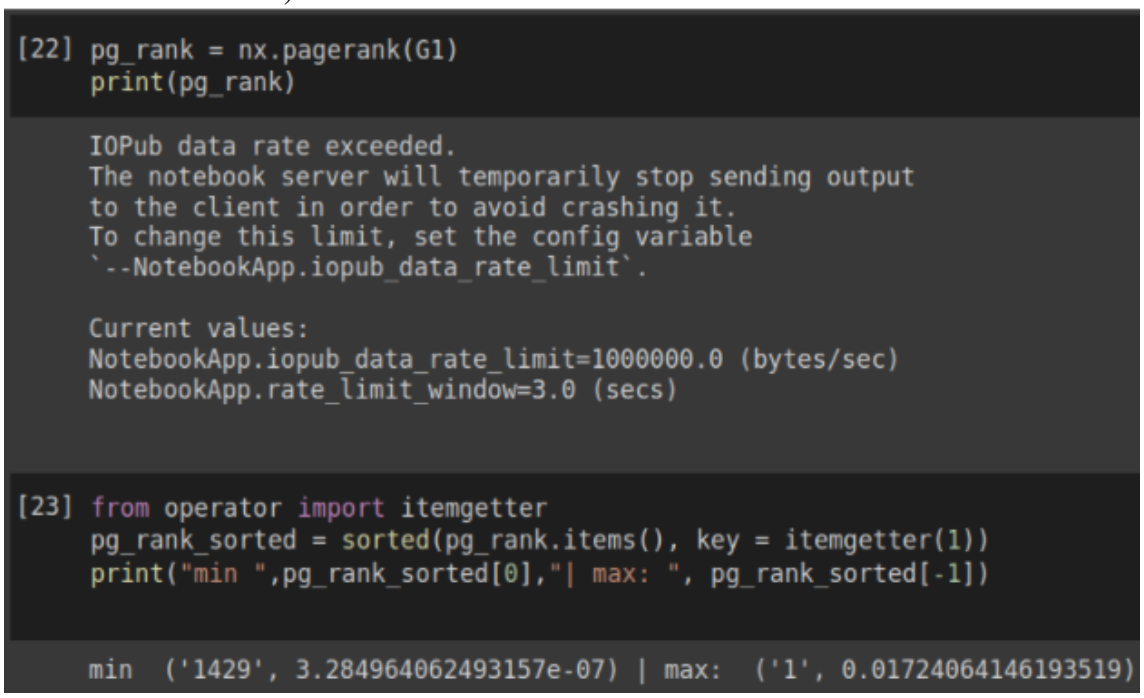
Sum out-degrees: 14855842
Mean out-degrees: 32.53393805871764

Diâmetro da rede: não foi possível calcular o diâmetro da rede dado o tamanho da rede.
Weakly connected: 156
Nodes size: 456290
Edges size: 14855466
Strongly connected: 91664

A rede tem muito mais componentes conexos fortes do que fracos, A partir desse ponto, não foi possível continuar a executar o código, visto que o Colab rodava por horas e acabava por exceder a memória RAM. O fato de os algoritmos percorrerem cada vértice e, para cada nó no vértice, buscar ligações com outros nós, entre outras abordagens do tipo, acaba tornando-os muito complexos. Juntando isso à quantidade de dados, entendemos que seria computacionalmente inviável, com nossos recursos, finalizar os cálculos.
Medidas que não foram calculadas por limitação do Google Colab: betweenness centrality, closeness centrality, cluster coefficient.



Eigenvector centrality: min ('1429', 3.023208089722967e-69) — max: ('104774', 0.1040845641840786)



Pagerank: min ('1429', 3.284964062493157e-07)
max: ('1', 0.01724064146193519)

Hits: hub (pontuação do nó baseada em outgoing links) e authority (pontuação do nó baseada em incoming links)

Hubs: min: ('409970', -9.4736264851754e-27) max: ('3013', 4.272150660026888e-05)

Authority: min: ('390880', -5.530966168936666e-24) max: ('1503', 0.0005568103177331846)

3. Fase 2

Mention graph: MultiDiGraph with 116419 nodes and 505311 edges
Reply graph: MultiDiGraph with 38928 nodes and 115889 edges

Retweet graph: MultiDiGraph with 256496 nodes and 1131744 edges
Usuários que mais retweetaram: nós com maior out-degree no grafo de retweet.

Mais retweet: ('1', 223408)

Top 10+: [('1', 223408), ('88', 14062), ('2', 10873), ('14454', 6190), ('677', 5624), ('1988', 4337), ('349', 2804), ('3', 2318), ('283', 2039), ('3571', 1982)]

Sum out-degrees: 1131744
Usuários mais retweetados: nós com maior in-degree no grafo de retweet.

Max in degree: ('1', 223408) Top 10+: [('1', 223408), ('88', 14062), ('2', 10873), ('14454', 6190), ('677', 5624), ('1988', 4337), ('349', 2804), ('3', 2318), ('283', 2039), ('3571', 1982)]

Usuários mais respondidos: nós com maior out-degree no grafo de reply

Top 5+: [('1', 25313), ('2', 2242), ('677', 1208), ('88', 1071), ('220', 470)]

Usuários mais mencionados: usuários com maior in-degree no grafo de mention

Max in degree: [('1', 98874), ('88', 11960), ('2', 7637), ('677', 3915), ('2417', 2538), ('3', 1704), ('59195', 1604), ('3998', 1594), ('7533', 1531), ('383', 1359)]

O nó 1 tem maior pagerank, que usualmente é usado para rankear páginas web. Considerando a atuação do nó 1 nas atividades de tweet, retweet, respostas e menções, isso dialoga com a alta pontuação dele, em comparação aos outros nós. Além disso, observa-se que os nós mais ativos se repetem entre as ações, sendo eles: nó 1,88, 2, 14454, 677, 3.

4. Apêndice

Os dados supracitados foram obtidos a partir de um programa desenvolvido em Python 3 pela dupla. A fase 1 e a fase 2 estão em programas separados.
Fase 1: https://colab.research.google.com/drive/1KYthjBPVkKI43ZgFKjbPIJA7q2rg4Wxj?usp=sharing
Fase 2: https://colab.research.google.com/drive/1B1yHVywjX1Q9jMT20DNkG3xsHg0li7gu?usp=sharing

5. Referências

1) Social network white paper, Cambridge Intelligence
2) cambridge-intelligence.com/keylines-faqs-social-network-analysis/
3) en.wikipedia.org/wiki/Directed$_graph$
4)$www.ime.usp.br/\ pf/algoritmos_para_grafos/aulas/strong-comps.html$
5)$Slidesdeaula$