SHAPE OF WAR **Lucas Saraiva Ferreira**
Trabalho Pratico 1

1 INTRODUÇÃO

A quantidade de dados online existente hoje em dia é imensurável. Extrair visualizações, informações e perspectivas interessantes destes dados é uma tarefa interessante e que demanda bastante criatividade. Uma das grandes fontes de dados e informação online é a *Wikipedia* [9] que atualmente tem 17 anos de existência. Todos os dados deste trabalho foram coletados da *Wikipedia*. A coleta foi centrada em conflitos internacionais que ocorreram entre 1500 e 2018. Os dados foram extraídos, filtrados, organizados e analisados utilizando conceitos de redes complexas. A rede obtida tem um tamanho razoável, com um total de 166 nós e 3378 arestas. Cada métrica ajudou a identificar diferentes aspectos da rede e entender melhor algumas relações, para algumas métricas foram necessárias modificações no grafo. Um dos resultados encontrados foi que a distribuição dos graus não segue uma *Power Law* [10], que é muito presente em vários conjuntos de dados.

O trabalho se organiza da seguinte forma: Seção 1 introdução e explicações gerais, seção 2 metodologias da coleta e processamento, seção 3 analises e demonstração do grafo, seção 4 contendo conclusão e resultados e a seção 5 contendo a resposta da pergunta número 2 do trabalho.

1.1 Motivação

O mundo a todo tempo está em conflito, seja este um conflito interno ou externo. Ao longo dos anos, a humanidade passou por diversos episódios de conflitos armados, alguns mais marcantes que outros, mas todos impactantes para alguma parcela da população mundial. De acordo com o site *"War and Peace"*, existem cerca de 3708 conflitos registrados desde 1400. Com base nestes números, surgiu a ideia de analisar os conflitos internacionais (Conflitos estritamente entre países) estes conflitos montando uma rede complexa utilizando dados da *Wikipedia*. As condições dos dados analisados encaixam perfeitamente nas redes sinalizadas descritas por *Kleinberg* em seu livro "*Networks, Crowds, and Markets: Reasoning About a Highly Connected World*" [1].

Associado a isto, existe o grande interessa do autor em história, principalmente de conflitos armados, que era inclusive uma segunda opção para curso superior. Neste contexto, decidiu-se trabalhar em cima deste tema, tendo como apoio diversas ferramentas computacionais e conteúdo online disponível de forma gratuita.

1.2 Grafo

Optou-se por realizar uma análise sobre uma rede de guerras envolvendo países, ou seja, somente conflitos internacionais, as guerras coletadas vão de 1500 a 2018. A rede é composta dos seguintes elementos:

- **Vértice** = Pais
- **Aresta** = No X tem uma relação com nó Y. As relações são estritamente entre países e podem ser positivas ou negativas. Um mesmo par de nós pode ter múltiplas arestas distintas. Cada aresta tem dois atributos: sinal da relação, ano do acontecimento e cor da aresta.

Cada aresta pode ter duas interpretações: se o pais X era aliado do pais Y eles recebem uma ligação positiva representada pelo sinal "+" no atributo "sinal da relação", o ano do ocorrido e

a cor verde; se o pais X era inimigo do pais Y eles recebem uma ligação negativa representada pelo sinal de "-", o ano do ocorrido e a cor vermelha.

Por fim, temos uma rede complexa multiarestas que traça relações de amizade ou conflito entre países.

2 METODOLOGIA

A primeira parte do trabalho foi identificar de onde as informações poderiam ser extraídas. A *Wikipedia* conta com diversas páginas descrevendo conflitos armados por datas. É possível encontrar desde conflitos internacionais (e.g Rússia vs USA) a conflitos internos (e.g USA vs confederados). Todos dados foram coletados, mas durante a etapa de filtragem, somente os conflitos internacionais permaneceram sendo todos outros descartados.

Os dados foram coletados da *wikipedia* através de um web *scrapper* escrito em *python* com a biblioteca *beautifulsoup* [4]. O grafo foi montado utilizando a biblioteca *networkx* também de python [7].

Os nomes dos países coletados correspondem ao nome da época, por exemplo: Império do Japão, atualmente conhecido somente Japão. Dado a diferenciação de território e a relevância do conflito que eles estavam envolvidos, entendeu-se que todos estes países mereciam um vértice separado. Por exemplo, um vértice específico para a União Soviética, atual Rússia, e um vértice para a Alemanha Nazista.

Após a coleta foi feita a limpeza e padronização dos dados. O grafo foi montado usando estes dados e exportado para diversos tipos, como CSV, GEPHML e GEFX. Por fim, foi implementado, em uma classe separada, diversas métricas para serem aplicadas ao grafo. Os resultados das métricas se encontram na seção 3 deste trabalho.

O código fonte e pode ser encontrado no github: https://github.com/saraiva3/Shape-of-War

2.1 Arquivos

O codigo está organizado da seguinte forma:

- Pasta MultiEdge: Crawler que gera multigrafo (main.py, graph_builder.py e country.py) e arquivo com todas métricas possíveis em grafos multi arestas (analytics.py)
- Pasta MultiEdgeNotAllowed: Crawler que gera um grafo comum, sem multi arestas (main.py, GrafoNormal.py e country.py) e arquivos usados nas analises (brigde.py -pertencente a networkx e analytics.py)
- Triangles: Código com a solução da questão 2 (triangle.py).

Exceto o arquivo triangles.py, que foi feito em python 2.7, todo resto foi feito com python 3.

3 ANÁLISES

Diversas métricas foram calculadas com a ajuda da networkx e implementações de algoritmos. Cada analise tem uma subseção para maior organização e detalhamento:

3.1 Rede

Para criar uma representação visual da rede foi escolhido a ferramenta Gephi [2]. A ferramenta permite importar grafos de arquivos de diversos formatos, além de contar com diversos algoritmos de análise e de organização do grafo.

Após carregar a rede na ferramenta, diversos tratamentos foram necessários para se obter algo visualmente agradável. Primeiro os vértices foram separados por continentes, cada qual uma cor diferente, sendo elas:

- Azul Escuro: América do Norte e Central;
- Verde: América do Sul;
- Cinza: África;
- Amarelo: Europa;
- Marrom: Ásia;
- Azul claro: Oceania.

Após isso, as arestas foram separas em positivas e negativas baseadas na cor definida no atributo de cada qual. O algoritmo Atlas 2 do gephi foi aplicado para uma pequena agrupação dos nós, partindo deste agrupamento, os nós foram separados por continente e organizados em uma ordem semelhante ao do mapa-múndi. Por fim, foram filtradas as arestas positivas para se obter o snapshot da figura 1 que apresenta somente as arestas negativas.

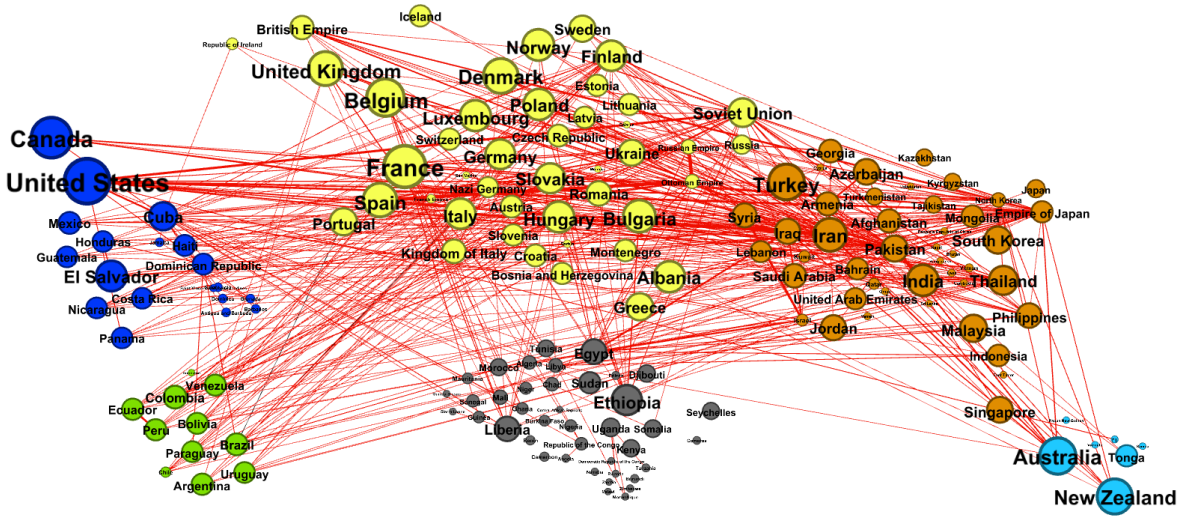


Figure 1: Snapshot contendo arestas negativas somente

A figura 2 mostra o mesmo grafo contendo somente as arestas positivas, ou seja, aliança entre países. O número de arestas de aliança é visivelmente maior que o de inimigos. De acordo com os dados, 86.63% das arestas consistem em alianças. Isso mostra que na maioria das vezes, as alianças são maiores que as desavenças. Outra forma de se ver esta informação é que diversas guerras tem um número grande de aliados para um número pequeno de inimigos, por exemplo: Na segunda guerra mundial um lado era composto por 16 países aliados entre si, enquanto o outro lado era composto de 56 aliados, o número de arestas positivas neste caso é maior que as negativas.

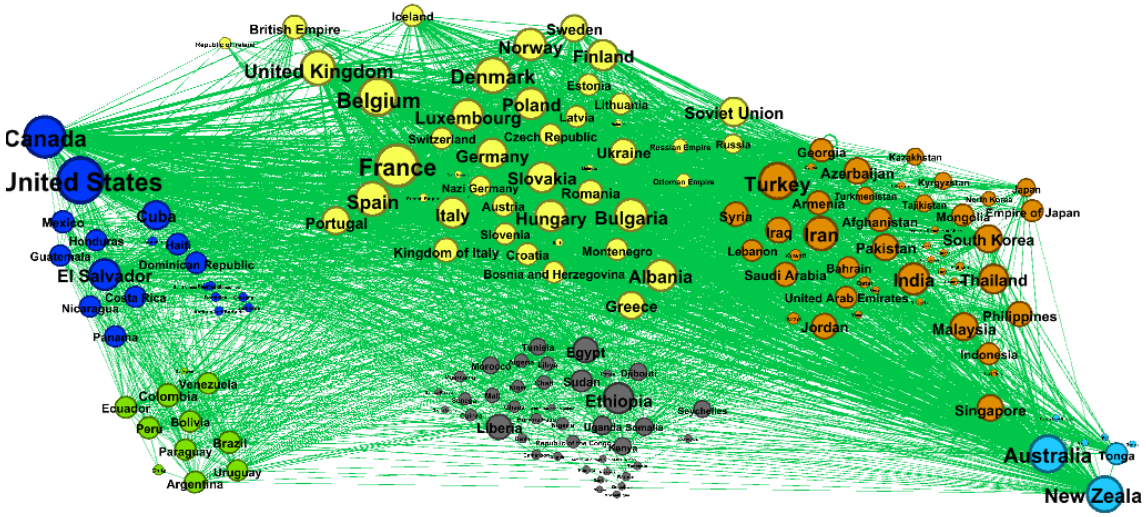


Figure 2: Snapshot contendo arestas positivas somente

3.2 Power Law

A figura 2 mostra a distribuição dos graus dos nós calculada para todas as arestas do grafo. Enquanto as figuras 3 e 4 apresentam, respectivamente, para as arestas negativas e as positivas. Uma vez que existem arestas de diferentes tipos, é interessante investigar a distribuição dos graus dos nós para cada uma delas. Nenhum dos gráficos segue uma *Power Law*. A explicação para isso pode ser dada por alguns fatores: A rede tem poucos nós e guerras não são uma tendência. As arestas negativas seguem uma pequena tendência, mas ainda assim não configura algo próximo de uma *Power Law*.

Neste caso o número de nós é limitado pelo número de países existentes, logo, o resultado não foi consequência de uma coleta "pequena".

O grau médio do grafo é de 86. Esse número elevado se dá a quantidade pequena de vértices e a quantidade elevada de arestas. O coeficiente alfa foi calculado usando a biblioteca PowerLaw

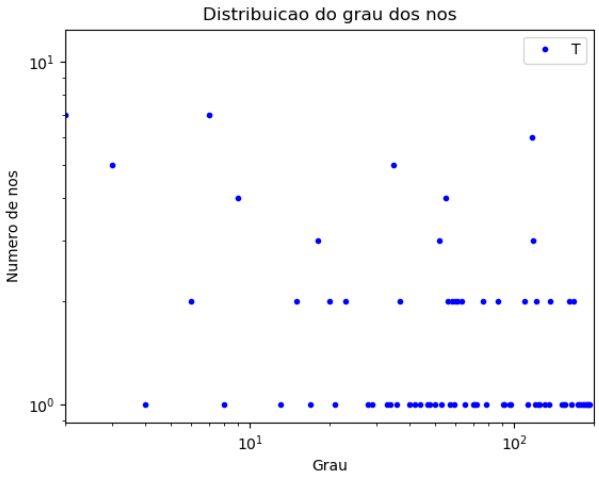
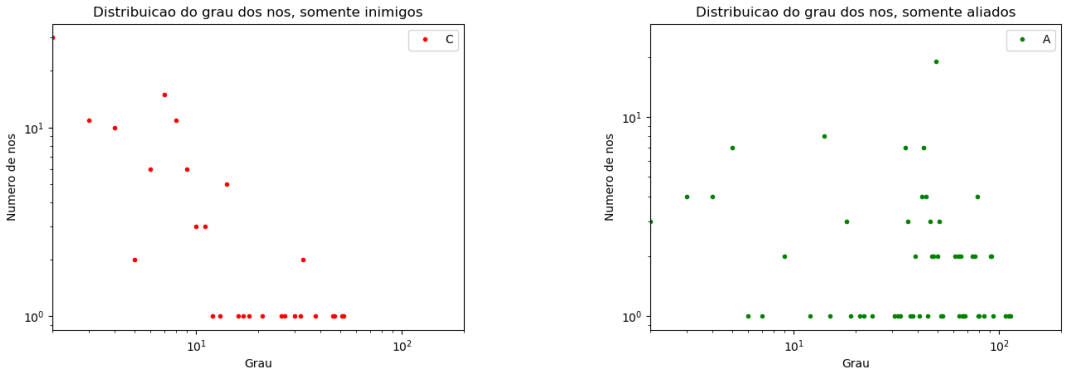


Figure 3: Distribuição dos graus de todos os nós



(a) Nos Negativos

(b) Nos Positivos

Figure 4: Distribuição dos graus dos nos separados

do python, o valor de alfa para a frequência dos graus foi 2.480270422530788 e para os graus foi 3.3277361144053152.

3.3 Componentes

O grafo conta com um único componente de 155 vértices. Para chegar neste resultado todos nós isolados foram removidos do grafo.

O cálculo foi realizado utilizando algoritmos da biblioteca networkx. A análise foi feita sobre todas as arestas, os componentes podem envolver arestas negativas e positivas.

3.4 Coeficiente de Clusterização

O grafo utilizado neste trabalho foi um multigrafo. Calcular o coeficiente de clusterização de cada nó em um multigrafo não é uma tarefa trivial [3]. Nenhum dos algoritmos de clusterização da networkx suporta multigrafos. Para realizar esta tarefa foi necessário modificar a rede, um merge de todas as arestas foi realizado e com este grafo simples foi calculado o coeficiente de clusterização.

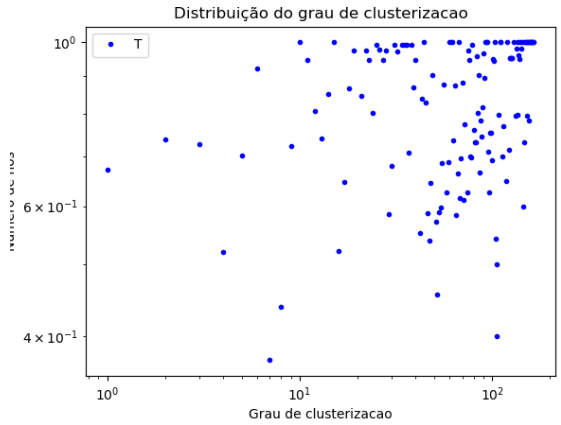


Figure 5: Coeficiente de clusterização de todos nós

Como é possível ver no gráfico da figura 5, os nós do gráfico modificado apresentam um elevado grau de clusterização. Como a rede tem poucos vértices e muitas arestas, este resultado era esperado. Além disso, as interações entre países acontecem muitas vezes entre seus vizinhos reais, facilitando a formação dos *clusters*.

O coeficiente de clusterização global calculado foi de: 0.7585725134589422.

3.5 Overlap da vizinhança

A distribuição do overlap da vizinhança é apresentado na figura 6. Uma parte dos nós tem um overlap bem baixo, e estes pontos podem talvez ser uma ponte local. O grafo não conta com muitas pontes e isso é atestado pela seção 3.8.

O overlap foi feito sobre o grafo completo, todas arestas inclusas.

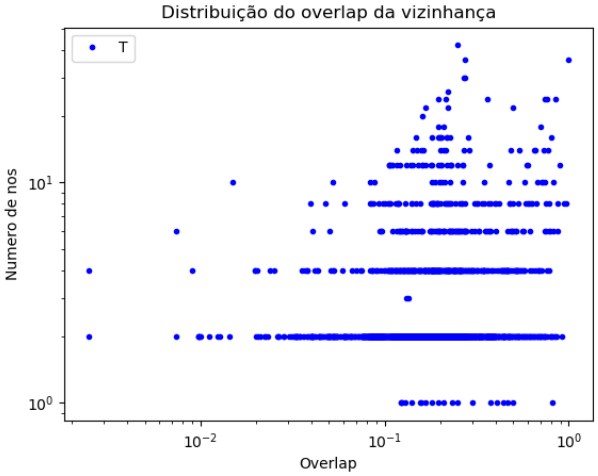


Figure 6: Overlap da vizinhança de todos nós

3.6 Distancia media

A distância média de cada par de nós é apresentada na figura 7, é possível ver que não existe uma grande variedade de tamanhos para os caminhos. Isso ocorre, pois, a rede tem muitas arestas, e todas arestas foram inclusas neste. O ponto indicando tamanho 0 no gráfico diz respeito aos vértices isolados do grafo. Pois não existe caminho deles para qualquer outro vértice.

A distância média de todos os nós encontrada foi de: 1.8740678676162548 (Para este cálculo, os nós isolados foram removidos).

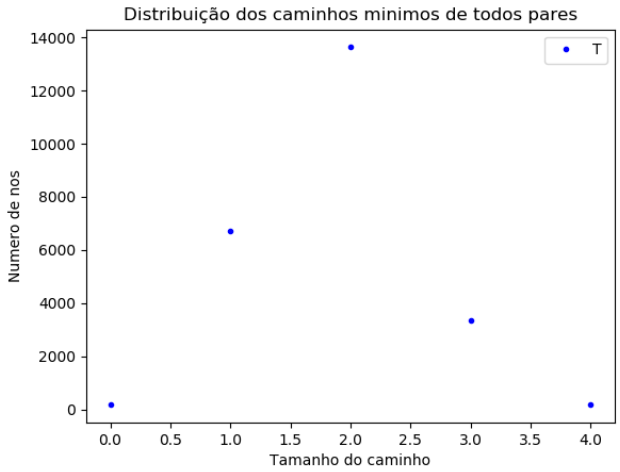


Figure 7: Caminhos minimos de todos pares

3.7 Betweenness

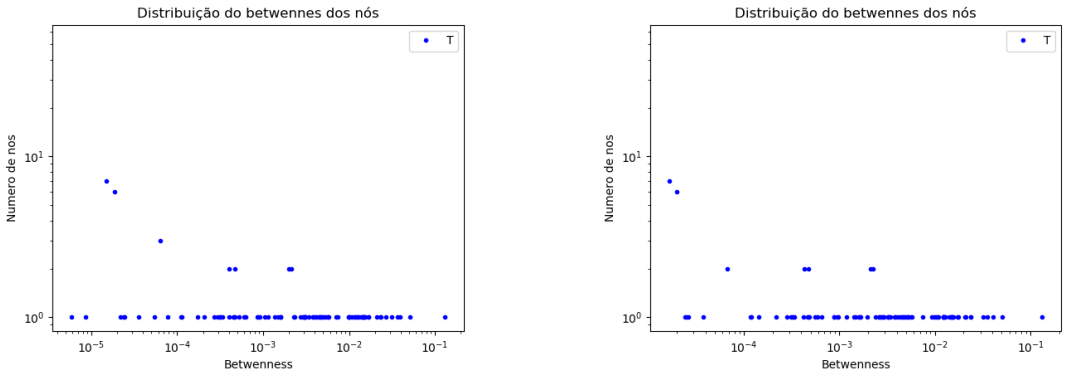
Foi calculado o betwenness dos nós e das arestas como pedido. A figura 7 mostra a distribuição do betwenness dos nós. Após a remoção dos 5 nós com maior betwenness, sendo estes: Império Frances, Catar, Kuwait, Venezuela e Haiti. O gráfico sofreu uma pequena mudança, o betwenness de alguns vértices aumentou, como podemos ver olhando para os pontinhos na base do gráfico da figura 8. Enquanto alguns vértices tiveram o betwenness reduzidos, como podemos ver os dois pontos mais a esquerda do gráfico.

O gráfico da figura 9 mostra a distribuição do betwenness para as arestas do gráfico.

3.8 Pontes

Para cálculo das pontes foi necessário a mesma modificação realizada para calcular o coeficiente de clusterização. Foram obtidas um total de 15 arestas que são pontes locais. As pontes obtidas pelo algoritmo, encontrado na biblioteca networkX, e o span respectivo de cada qual foram:

- Império Otomano e São Marinho (infinito)
- Estados Unidos e Fiji (infinito)
- França e Bielorrússia (infinito)
- Império Britânico e Butão (infinito)
- Afeganistão e Uzbequistão (infinito)



(a) Antes da remoção (b) Após a remoção

Figure 8: Betwenness dos nós antes e depois da remoção

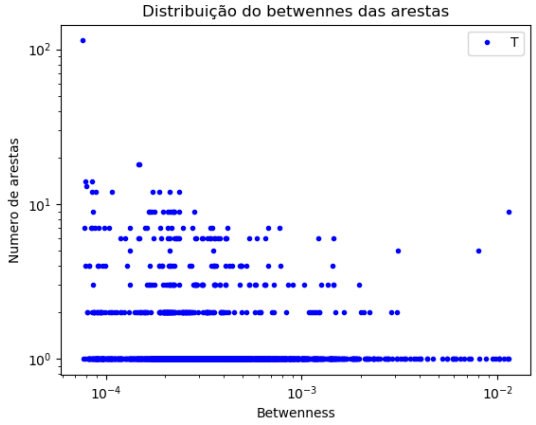


Figure 9: Betwenness das arestas

- Cuba e São Vicente-granadinas (3)
- Arábia Saudita e Iémen (3)
- Iémen e Eritreia (3)
- Libera e Serra leoa (infinito)
- Indonésia e Timor Leste (infinito)
- República do Congo e Uganda (infinito)
- Líbia e Moçambique (3)
- Uganda e Moçambique (3)
- Senegal e Guiné-Bissau (infinito)
- Granada e São Vicente-granadinas (3)

O span calculado médio é 3, tirando as pontes que constam infinito. Antes de calcular o span, esperava que o valor fosse 2, já que a distância média de todos nós é bem baixa e a

maior distância é 4. Logo, ao eliminar as arestas ligando estes vértices o caminho não aumenta drasticamente.

3.9 Assortatividade

A assortatividade do grafo e número de Pearson foram calculados utilizando os algoritmos da biblioteca networkx. Os seguintes valores foram encontrados:

1. Assortatividade: 0.11125046742711621
2. Coeficiente de Assortatividade: 0.10380811256195074
3. Pearson: 0.1112504674271155

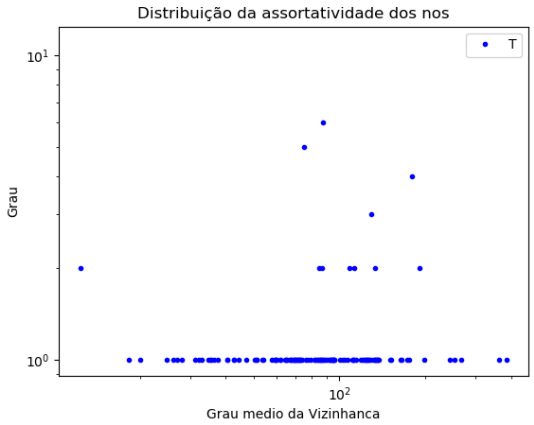


Figure 10: Assortatividade

4 Questão 2

As bases foram analisadas utilizando a biblioteca networkx [7].

Para encontrar os tipos de triângulos em uma rede direcionada foi utilizado o método: **triadic_censu**. Este método consegue retornar os tipos de triângulos da rede direcionada, baseado na nomenclatura da figura 10. Retirada do artigo "*Group Polarization: Connecting, Influence and Balance, a Simulation Study Based on Hopfield Modeling*" [6].

Os números abaixo são referentes a primeira base - soc-sign-epinions. Não foi inserido o resultado das outras 3 bases pois todos resultados dizem respeito aos mesmos tipos, mudando somente a quantidade encontrada em cada base:

- 003: 381729219743007
- 012: 76378799569
- 102: 17027513185
- 021D: 41490030
- 021U: 48493351

- 021C: 32569162
- 111D: 20362369
- 111U: 19115018
- 030T: 1479047
- 030C: 59899
- 201: 5433309
- 120D: 695450
- 120U: 926939
- 120C: 324586
- 210: 877116
- 300: 547039

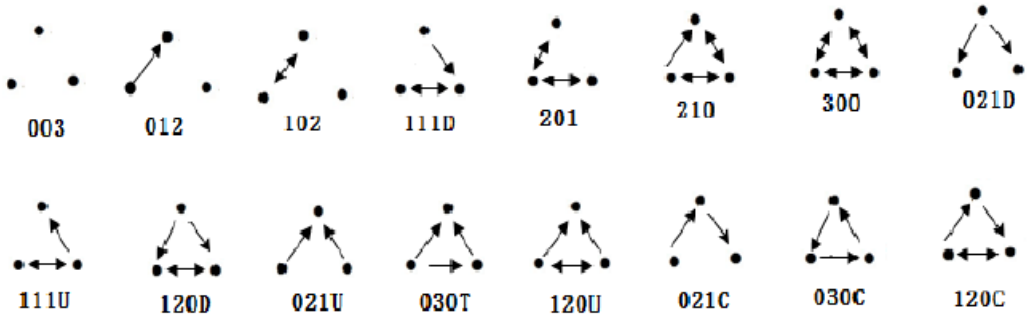


Figure 11: Tipos de triângulos [6]

Após a leitura do artigo referência de uma das bases: "Signed networks in social media" [5] cheguei à conclusão que usar a teoria de balanceamento não seria suficiente para dizer se os triângulos da rede são balanceados ou não. Para isso se faz necessário uma abordagem diferente, como a do artigo, usando status[5].

Para implementar a verificação de status utilizei a técnica descrita no capitulo 5 seção 5.14 do livro "Trust in social media" [8]. O código segue os seguintes passos:

1. Ler a base de dados e criar um grafo, sendo o atributo o sinal de cada aresta (1 ou -1);
2. Encontrar triângulos. A primeira iteração é sobre todos os nós. Após isso é gerado o conjunto de vizinhos deste nó. Para cada vizinho deste conjunto é criado um conjunto com todas interseções, ou seja, dado o vértice sendo analisado descubra os vizinhos que compartilham vizinhos com ele. Para cada interseção temos uma tripla a ser verificada;
3. Para cada tripla, verifica se ela já foi visitada, caso não, visite ela;
4. Para cada tripla sendo visita, inverter a direção das arestas com sinal negativo e marcar como visitada;

5. Verificar se o triangulo formado é um ciclo. Caso seja, está desbalanceado. Caso não, está balanceado [8];

6. Verificar se existem mais balanceados ou não balanceados.

Faltou otimização do código para se obter o resultado de forma mais rápida. O código feito se encontra no arquivo "***triangles.py***" e funciona normalmente, seguindo os passos listados acima. Sendo o único problema o tempo de execução. Os valores obtidos para cada base foram:

- soc-sign-Slashdot081106:
  - **Balanceados**: 525677
  - **Não balanceados**: 264077
- soc-sign-Slashdot090216:
  - **Balanceados**: 552263
  - **Não balanceados**: 261114
- soc-sign-Slashdot090221:
  - **Balanceados**: 559417
  - **Não balanceados**: 265524
- soc-sign-epinions:
  - **Balanceados**: 5119680
  - **Não balanceados**: 2131418

Podemos ver que para todas as bases o número de triângulos balanceados é quase o dobro dos não balanceados. Como citado no livro "Trust in social media" grande parte dos triângulos, seguindo a teoria do status, estão balanceados [8]. Dado a quantidade de triângulos e quantidade de cada tipo existente e olhando somente a tabela de tipos de triângulos não é possível deduzir se a rede é ou não balanceada. A teoria do status foi fundamental para chegar a esta conclusão no trabalho.

References

[1] Easley David and Kleinberg Jon. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

[2] GEPHI Team. Gephi, 2008 - 2017. [Online; accessed 03-May-2018].

[3] Madhav Jha, C Seshadhri, and Ali Pinar. When a graph is not so simple: Counting triangles in multigraph streams. 10 2013.

[4] Leonard Richardson. Beautifulsoup, 2004 - 2015. [Online; accessed 03-May-2018].

[5] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1361–1370. ACM, 2010.

[6] Zhenpeng Li and Xijin Tang. Group polarization: Connecting, influence and balance, a simulation study based on hopfield modeling. 09 2012.

[7] NetworkX Developers. Networkx, 2015. [Online; accessed 03-May-2018].

[8] J. Tang and H. Liu. *Trust in Social Media*. Synthesis Lectures on Information Security, Privacy & Trust. Morgan & Claypool Publishers, 2015.

[9] Wikipedia contributors. Wikipedia, 2004. [Online; accessed 03-May-2018].

[10] Xavier Gabaix. Power law, -. [Online; accessed 03-May-2018].