Universidade Federal do Ceará – UFC
Centro de Ciências – CC
Mestrado e Doutorado em Ciências da Computação - MDCC
Estruturas de Dados

Exercício: Representando Redes Sociais como Grafos (Neo4J)
Objetivos: Exercitar os conceitos referente à análise de grafos.
Data da Entrega: 18/11/2024

OBS 1: Exercício Individual.
OBS 2: A entrega da lista deverá ser executada utilizando-se o SIGAA.

NOME: _____MATRÍCULA: _____

Questão 1

Crie um arquivo Jupyter Notebook e realize as seguintes operações:

a) Realizar o "restore" do arquivo (dump) denominado fd_whatsapp_0911_2023.zip no PostgreSQL. Esse arquivo está disponível no link a seguir: https://drive.google.com/drive/folders/1kEEnmZUVJEgYTynZjU6qMICbEVd8wKca?usp=sharing
b) Remova os trava-zaps.
c) Remover as linhas repetidas (duplicadas).
d) Remover textos com menos de 5 palavras.

e) Monte um grafo no Neo4J para modelar (representar) as relações entre postagens, usuários e grupos do WhatsApp.

Modelar as relações entre os usuários na forma de um grafo pode fornecer informações relevantes sobre padrões de comportamento. Porém, diferente de redes sociais como Twitter ou Facebook, onde existem conexões bem definidas entre os usuários pela relação de seguir (Twitter) ou de amizade (Facebook), no WhatsApp essas conexões não são explícitas.

Assim, propomos uma modelagem das relações entre usuários do WhatsApp na forma de grafos direcionados e valorados, considerando o envio de mensagens em grupos. Nessa modelagem, cada nó representa um usuário e podemos considerar um grafo para cada tipo de mensagem: mensagem em geral, mensagem viral e mensagem com desinformação.

Considerando o grafo de mensagens gerais, onde cada nó representa um usuário, existe uma aresta direcionada entre o usuário i e o usuário j se o usuário i enviou uma mensagem para um grupo do qual o usuário j faz parte. O peso dessa aresta é a quantidade de mensagens enviadas pelo usuário i para aquele grupo.

Um raciocínio análogo pode ser aplicado para criar um grafo apenas de mensagens virais: existe uma aresta direcionada entre o usuário i e o usuário j se o usuário i enviou uma mensagem viral para um grupo do qual o usuário j faz parte e o peso dessa aresta é quantidade de mensagens virais enviadas pelo usuário i para aquele grupo.

O mesmo procedimento pode ser utilizado para criar o grafo de desinformação: existe uma aresta direcionada entre o usuário i e o usuário j se o usuário i enviou uma mensagem contendo desinformação para um grupo do qual o usuário j faz parte e o peso dessa aresta é quantidade de mensagens virais enviadas pelo usuário i para aquele grupo.

Percebe-se que nos três grafos, a quantidade de nós é a mesma, variando a quantidade de arestas.

Monte uma tabela contendo a quantidade de nós e a quantidade de arestas para cada grafo (mensagens gerais, mensagens virais e mensagens com desinformação).

Verifique se existem grupos isolados e clusters de usuários fortemente conectados. Isso pode ocorrer caso existam usuários engajados que possuem participação ativa em diversos grupos.

Através dessa modelagem dos usuários, podemos obter algumas métricas básicas de nós em redes complexas (WEI et al., 2013), que chamamos aqui de atributos de rede dos usuários, similar ao que foi feito por Benevenuto et al. (2008) no contexto do YouTube. São eles:

• Grau de centralidade geral: quantidade de arestas de mensagens gerais saindo do nó. Ou seja, para um dado usuário, mensura o número de usuários que receberam ao menos uma mensagem dele. Quanto maior, com mais usuários esse usuário teve contato;
• Força geral: somatório dos pesos de todas as arestas de mensagens gerais saindo do nó. É correlacionado com o grau de centralidade geral, porém a quantidade de mensagens enviadas também é levada em conta para calcular essa métrica. Um valor alto indica que o usuário enviou muitas mensagens que atingiram muitos usuários;
• Grau de centralidade viral: análogo ao grau de centralidade geral, mas relativo somente às mensagens virais, sendo a quantidade de arestas de mensagens virais saindo do nó. Quanto maior, mais usuários receberam mensagens virais desse usuário;
• Força viral: análogo à força geral, é somatório dos pesos de todas as arestas de mensagens virais saindo do nó. Um valor alto indica que o usuário enviou muitas mensagens virais que atingiram muitos usuários;
• Grau de centralidade de desinformação: análogo ao grau de centralidade geral, mas relativo somente às mensagens de desinformação, sendo a quantidade de arestas de desinformação saindo do nó. Ou seja, a quantidade de usuários com quem esse usuário compartilhou desinformação. Quanto maior, mais usuários receberam desinformação desse usuário;
• Força de desinformação: análogo à força geral, é o somatório dos pesos de todas as arestas de desinformação saindo do nó. Um valor alto indica que o usuário enviou muita desinformação e que atingiu muitos usuários;

Percebe-se que as duas últimas podem ser calculadas devido aos rótulos manuais, enquanto as outras podem ser calculadas a partir de dados não rotulados. O conjunto de dados possui um atributo chamado misinformation_score (ou score de desinformação). Podemos definir um limiar (threshold) e atribuir um rótulo de desinformação automaticamente. Por exemplo, mensagens com misinformation_score maiores ou igual a 0.7 podem ser consideradas como textos que possuem desinformação.

Compute os valores dos atributos de rede para cada usuário.

Verifique se os atributos de rede seguem uma distribuição de cauda longa, possuindo valores baixos nesses atributos e alguns raros usuários com valores muito altos, indicando um comportamento anormal desses usuários que merece investigação.

Identifique os 5 usuários mais ativos.

Identifique os 5 usuários que mais espalham desinformação.

Identifique os 5 usuários mais influentes.

Identifique os 5 usuários mais conectados.

> “Ensinar inexiste sem aprender e vice-versa e foi aprendendo socialmente que, historicamente, mulheres e homens descobriram que era possível ensinar”.
>
> **Paulo Freire**