TRABALHO DE GRAFOS
Universidade Federal de Lavras
Nome: Felipe Mateus Maximiniano e Silva Ribeiro
Turma: 10A
Matrícula: 202220179
Curso: Ciências da Computação
Disciplina: Algoritmos em Grafos - GCC218
Professor: Vinícius Vitor

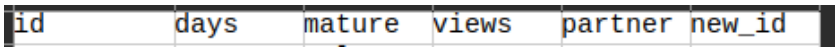
Apresentação da base de dados:

Os conjuntos de dados foram coletados com o propósito de realizar a classificação de usuários em redes sociais do Twitch, uma plataforma de transmissão de jogos. Cada rede representa gamers que fazem streaming em uma língua específica, onde os usuários são representados como "nós" e as conexões entre eles indicam amizades mútuas. As características individuais dos usuários, como os jogos jogados, preferências, localização e hábitos de transmissão, foram extraídas para formar as características dos "vértices".
É relevante destacar que todos os conjuntos de dados compartilham o mesmo conjunto de características dos usuários. Essas redes sociais foram coletadas em maio de 2018, fornecendo uma visão específica desse ecossistema naquele período.

O grafo é não direcionado, possui grau médio de 32.73953974895397, não possui loops de usuarios com eles mesmos e não existe mais de uma ligação de um mesmo vertice com outro mesmo vertice. Informaçoes adicionais encontram-se na imagem abaixo:

Dataset statistics

	DE	EN	ES	FR	PT	RU
Nodes	9,498	7,126	4,648	6,549	1,912	4,385
Edges	153,138	35,324	59,382	112,666	31,299	37,304
Density	0.003	0.002	0.006	0.005	0.017	0.004
Transitvity	0.047	0.042	0.084	0.054	0.131	0.049



As bases de dados possuem estes atributos a cima e sendo id o id real do usuário da Twitch, days a quantidade de dias da existencia da conta, mature significa se o conteúdo é classificado para maiores de idade ou não, views são as visualizações que o streamer possui, partner significa se o usuário tem contrato com a twitch ou não e new_id é o novo id do usuário utilizado nos grafos gerados.
Vale ressaltar que a base de dados se chama Musae_twitch, ela possui 3 arquivos para cada país presente, sendo um representando o grafos(edges) outro representando features para aprendizagem de maquina(não foi utilizado e seu formato é um json) e outro chamado de alvo(target) contêm os atributos do país selecionado. A base escolhida para ser utilizada foi a da pasta PTBR, “musae_PTBR_target” ‘’musae_PTBR_edges”.

Definição do problema:

O objetivo desta investigação reside na identificação e compreensão das comunidades formadas por streamers brasileiros na plataforma Twitch, com o intuito de desvelar suas representações e entender seu impacto na diversidade ou não de conteudos transmitidos. Este desafio implica não apenas na identificação das comunidades, mas também na interpretação do significado por trás dessas interações digitais. O entendimento dessas nuances é crucial para oferecer insights valiosos sobre as motivações dos streamers e entender as estratégias mais eficazes na criação e promoção de conteúdo na plataforma Twitch.
Dessa forma utilizamos como entrada o grafo e seus atributos que estao em formato csv e como saída temos um arquivo de texto contendo informações como as comunidades encontradas, modularidade, hubs, pontes entre comunidades e um arquivo de imagem contendo uma representação visual das comunidades.

Exemplos:

Tomemos o Streamer A como um caso de destaque, cuja comunidade apresenta números significativos de visualizações. Em contraste, uma segunda comunidade apresenta poucos streamers em destaque e, consequentemente, registros modestos de visualizações.

Podemos utilizar os dados encontrados para identificar esses streamers, observar os jogos trasmitidos, o linguajar utilizados nas transmissoes por exemplo. Desta forma entenderemos melhor o porque dessa disparidade das comunidades.
Podemos também analisar pontes entre comundiades e o que aconteceria e quais seriam as consequencias para as comundiades de streamers se essas relaçoes fossem removidas. Podemos identificar os usarios reais que representam essas pontes e assim compreender melhor o que o rompimento desta significa.
Podemos identificar os usúarios mais fomosos das comunidades e tambem os mais famosos da twitch e entender que tipo de conteúdo eles trasmitem, se estão na mesma comunidade, se tem releção diretamente.

Solução:

Procedeu-se à transformação de um grafo, inicialmente representado em formato CSV junto a seus atributos, para uma lista de adjacências, dada a sua baixa densidade. Em seguida, implementou-se o algoritmo de Louvain, uma abordagem eficaz também em grafos desconexos.

O algoritmo de Louvain destaca-se por sua capacidade de identificar comunidades em grafos de maneira eficiente a partir de um vertice arbitrário.Vale ressaltar que o algoritmo gera algumas diferentes comunidades dependendo do ponto de partida. Sua aplicação compreende uma estratégia iterativa de otimização da modularidade, um indicador que avalia a qualidade da divisão do grafo em comunidades. A modularidade é calculada comparando a densidade de arestas dentro das comunidades identificadas com a densidade esperada em um grafo aleatório com a mesma distribuição de nós.
Formula da modularidade:

$$\mathrm{Q}(\mathcal{C}) = \sum_{C \in \mathcal{C}} \left[\frac{|E(C)|}{m} - \left(\frac{\sum_{v \in C} \deg(v)}{2m} \right)^2 \right]$$

O procedimento iterativo do algoritmo de Louvain ocorre em duas fases principais: a fase de otimização local e a fase de agregação global. Na primeira fase, os nós são realocados para comunidades vizinhas, visando maximizar a modularidade local. Na segunda fase, a estrutura do grafo é simplificada pela agregação de comunidades em um único nó, facilitando a análise de níveis mais amplos de organização.

Nesse sentido esse algoritmo se torna eficaz também para grafos desconexos pois opera independentemente em cada componente conectado, permitindo uma análise modular eficaz mesmo em estruturas mais fragmentadas.
Além da implementação do algoritmo de Louvain, foram incorporadas análises adicionais ao grafo. Em particular, foram realizados cálculos relacionados à identificação de hubs e à quantificação de pontes entre comunidades.

O cálculo de hubs foi efetuado por meio da métrica de centralidade de grau (degree centrality) utilizando a biblioteca NetworkX. Essa métrica avalia a importância relativa de um nó com base na quantidade de conexões que possui no grafo. Nós com uma centralidade de grau elevada são considerados hubs, indicando que estão fortemente conectados com outros nós na rede. O limiar (valorhub) estabelecido permite definir quais nós serão considerados hubs, sendo possível ajustá-lo conforme necessário.Além disso, a quantificação de pontes entre comunidades visou identificar nós que conectam diferentes grupos no grafo. Essas pontes desempenham um papel crucial na integração de informações entre comunidades distintas, revelando pontos de transição ou interação significativa. Por fim os resultados são gerados e gravados em um arquivo de imagem grafo.png e resultados.txt.

Sobre a ferramenta:

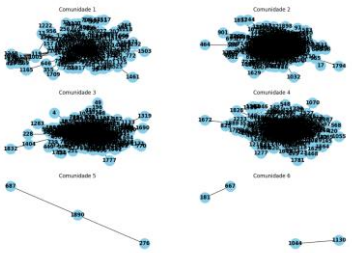
O projeto utiliza a linguagem Python e consiste em três códigos: `r2.py`, `r1.py` e `interface.py`. O `r1.py` é uma versão adaptada do `r2.py` para interagir com a interface gráfica `interface.py`. Consta também um backgrond.mp4 que interage com a interface.py

O gerenciamento do projeto é feito por um arquivo `makefile` que oferece funcionalidades como:

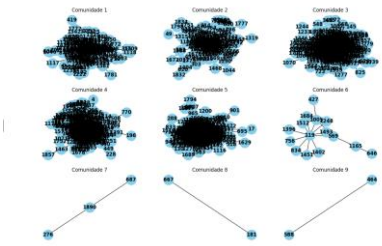
- `make`: compilação/ instalação do projeto e suas dependencias.
- `make run`: execução do `r2.py` com configurações padrão.
- `make runmanual`: execução com entrada manual do usuário.
- `make runwithinterface`: execução da versão `r1.py` com interação da interface gráfica.

Resultados:

No decorrer da execução dos algoritmos, observou-se que, em algumas iterações, comunidades específicas apresentavam repetição consistente. Uma análise mais aprofundada foi conduzida, focando exclusivamente em identificadores reais usando o site (https://streamscharts.com/tools/convert-username?login=1702603500) para converter o id para nome, o streamer 36772976,Tecnosh, jogador de PUBG,foi detectado como um hub,o que se confirma, pois é um streamer muito reconhecido no Brasil pela comunidade de PUBG. Nesse sentido, analisando mais ids da comunidade de Tecnosh, tem-se 36264712, identificado como SilvioSantosdoCs, o qual trasmite jogos de tiro também e partilha do linguajar mature = false, característica marcante da comunidade '1', comunidade de Tecnosh. Por fim observando e consultando alguns ids como(21919955, 99695943, 32351776, 160986323...) entende-se que os todos esses streamers jogam jogos de ação com Tiro,e se relacionam por meio de jogatinas conjuntas ou porque jogam os mesmos games. Além disso as comunidades isoladas representaram a de jogos indies(jogos independentes, muitas vezes pouco famosos).



Algo interessante, foi o encontro das pontes entre comunidades:[(197, 181)], que quando removidas da base de dados para efetuação de destes obtivemos mais comunidades, dividindo ainda mais os Streamers de jogos indies.



Limitações:

Sobre o metodo escolhido**,** algoritmo Louvain, utilizado para detectar comunidades em redes complexas, enfrenta limitações em grafos densos devido a problemas de eficiência. Além disso, sua sensibilidade à inicialização significa que diferentes configurações iniciais podem levar a resultados divergentes, influenciando a qualidade da solução.