

Analizando a semelhança entre álbuns por meio de grafos

Lucas Guido Ciliberti Portugues

RA: 11202321391

Universidade Federal do ABC

Santo André, SP, Brasil

lucas.guido@aluno.ufabc.edu.br

Luis Felipe Ruiz Perez

RA: 11202321409

Universidade Federal do ABC

Santo André, SP, Brasil

luis.perez@aluno.ufabc.edu.br

Nicolas Gomes Greco

RA: 11202321465

Universidade Federal do ABC

Santo André, SP, Brasil

nicolas.greco@aluno.ufabc.edu.br

Resumo—Com o crescente acesso à música por meio de plataformas digitais, entender as preferências musicais se tornou fundamental para aprimorar sistemas de recomendação. O presente projeto se propõe a explorar as similaridades entre os 1000 álbuns mais bem avaliados na lista “Top albums of all time” do RateYourMusic, uma comunidade e base de dados focada em música. A partir da construção e análise de uma rede de álbuns, que considera critérios como gênero, gênero secundário, artista, década e *descriptor*, buscamos identificar tendências e padrões que influenciam as avaliações dos usuários usando métricas de centralidade e da identificação de comunidades. Além disso, busca-se desenvolver um sistema simples de recomendação de álbuns que, além de considerar aspectos musicais, seja capaz de refletir as preferências e vieses da comunidade do RateYourMusic. Com isso, pretendemos não apenas destacar álbuns semelhantes, mas também entender como certos estilos e características influenciam a percepção e avaliação musical dentro da plataforma.

Palavras-chave—música, similaridade entre músicas, algoritmo de recomendação, redes, grafo

I. INTRODUÇÃO

A música é uma forma de arte profundamente enraizada na cultura, desempenhando um papel significativo na construção de identidade, memória e na evocação de emoções [1]. Assim como as músicas individuais, os álbuns também refletem essa prática cultural, funcionando como um repositório coletivo das experiências e preferências musicais de uma comunidade. Com base nisso, este estudo explora como os álbuns mais bem avaliados em uma das maiores comunidades online de música, RateYourMusic¹, se conectam, focando nos 1000 álbuns mais bem posicionados na lista “Top albums of all time”² [10].

Nas últimas décadas, o avanço da tecnologia tem desempenhado um papel crucial na transformação da indústria da música, alterando radicalmente a maneira como o conteúdo musical é distribuído e consumido. As plataformas de streaming permitiram o consumo em tempo real de músicas, eliminando a necessidade de downloads de arquivos [2]. De acordo com a IFPI (International Federation of the Phonographic Industry), cinquenta por cento da receita da indústria fonográfica em 2016 foi proveniente dessas plataformas [3], o que evidencia a sua influência crescente. Além da democratização

do acesso à música, surgiram sites de avaliação, como o RateYourMusic, onde os usuários podem avaliar álbuns e artistas, influenciando a percepção coletiva sobre a música.

No RateYourMusic, os usuários podem não apenas avaliar álbuns com base em suas preferências pessoais, mas também registrar metadados que incluem gênero principal, gênero secundário, artista, década de lançamento, descrições (também chamados de *descriptors*) e outras informações relevantes. A coleta e o armazenamento desse tipo de dado são essenciais para a análise de padrões e tendências, especialmente no contexto da popularização dessas plataformas. Essa riqueza de informações abre caminho para uma variedade de estudos, incluindo a recomendação de álbuns dentro de uma comunidade específica, explorando as conexões entre diferentes álbuns através de técnicas de ciência de redes [5]. Essas plataformas podem ser facilmente caracterizadas como redes, o que permite a aplicação de métodos avançados de análise estrutural.

Desta forma, foram usadas informações de álbuns retiradas da lista citada para conectá-los com base em similaridades de artista, gênero, década de lançamento e descrições do álbum. Isto gerou uma rede que associa álbuns que possuem semelhanças, permitindo entender como se relacionam. A partir disso, é possível entender o viés da comunidade do site usando métricas de centralidade, avaliando álbuns mais centrais, bem como comunidades, estudando possíveis formas de conexão que não são percebidas à primeira vista.

Após a compreensão da relação entre os álbuns, foi desenvolvido um sistema simples de recomendação de músicas com base na similaridade entre elas. A cada álbum recomendado, o *feedback* do usuário é coletado para aprimorar o sistema, aproximando as recomendações de álbuns que o usuário gostou.

Este relatório está organizado em 6 seções: A Seção 1 (esta) introduz o trabalho e a metodologia. A Seção 2 apresenta trabalhos relacionado e o estado da arte em sistemas de recomendação. A Seção 3 descreve o processo de coleta de dados e construção da rede. A Seção 4 apresenta o estudo e os resultados das análises da rede. A Seção 5 detalha o funcionamento do algoritmo de recomendação. Por fim, a Seção 6 conclui o trabalho e a Seção 7 descreve as contribuições de cada integrante do grupo.

¹<https://rateyourmusic.com/>

²Acessível em: <https://rateyourmusic.com/charts/top/album/all-time/>

II. ESTADO DA ARTE E TRABALHOS RELACIONADOS

Em geral, algoritmos de recomendação se encaixam em diferentes categorias de acordo com a sua especialização. Alguns sistemas são desenvolvidos para recomendar músicas em sequência (em tempo real), funcionando diferentemente de algoritmos para *playlists*, que são mais convencionais. Estes podem usar abordagens como: baseada em metadados, conteúdo de áudio, recomendação contextual e sistemas híbridos. Já que este trabalho é focado em gerar conexões a partir de metadados (i.e. artista, gênero e ano de lançamento), será enfatizada a revisão de trabalhos envolvendo a primeira abordagem.

Por exemplo, um artigo de Bogdanov et al. usa a base de dados do site *Discogs* para conectar artistas a partir de lançamentos usando vetores ponderados. Paralelamente, artigos como os de Pazzani e Billsus e de Green et al. usam palavras-chave sobre o artista para basear as recomendações a partir dos artistas [11]. Estas abordagens diferem da nossa, que considera os álbuns diretamente.

Apesar de muitas vezes usarem características acústicas complementadas por metadados para determinar a similaridade, sistemas mais avançados para gerar *playlists* automáticas empregam grafos de músicas e rodam algoritmos para encontrar menores caminhos para navegar a partir de um álbum inicial. Também é possível relacionar o caminho da lista ao problema do caixeiro viajante, e procurar o menor caminho nesta situação [11]. Optou-se por adotar uma abordagem mais simples usando apenas os metadados.

O modelo de Shao et al. se aproxima do utilizado neste projeto, gerando um grafo de relações entre músicas, mas relacionando-as a partir de um modelo híbrido, combinando descrições do áudio com *feedback* do usuário. O algoritmo propaga as avaliações do usuário para prever avaliações futuras [11]. Nosso modelo considerará avaliações para aumentar ou diminuir a distância entre álbuns comparando-o com outros já avaliados. Existem também outras abordagens que utilizam grafos, combinando-os com aprendizado de máquina, para prever as próximas músicas a serem tocadas a partir de sequências em *playlists* já existentes [11].

Algoritmos de aplicativos de *streaming* tendem a combinar diversas técnicas de recomendação, balanceando similaridade com músicas já ouvidas, novos lançamentos, estímulo a continuar escutando e relevância do conteúdo com base na popularidade. Para isso, no aplicativo *Spotify*, são usadas características como metadados e descrições sobre sensações provocadas no ouvinte ao escutar a música e, conseqüentemente, em que momento ela deve ser ouvida, além de estatísticas [12]. Este é um sistema bem mais elaborado, e que, apesar de não terem disponibilizado exatamente seu funcionamento, se apresenta como o estado da arte e sustenta uma das maiores plataformas de *streaming* do mundo.

III. CONJUNTO DE DADOS E CONSTRUÇÃO DA REDE

A. Coleta de Dados

O Rate Your Music (RYM) é uma plataforma online, lançada em 2000, que combina uma extensa base de da-

dos musical com uma comunidade ativa de usuários que catalogam, revisam e avaliam álbuns, EPs, singles e outras formas de lançamentos musicais. A plataforma atualmente abriga aproximadamente 23 milhões de músicas, 2 milhões de artistas e acumula mais de 140 milhões de avaliações de usuários. Esses números refletem a relevância do RYM como uma das principais fontes de dados para a análise musical, sendo uma ferramenta valiosa para estudos de tendências, popularidade e similaridade entre obras musicais. Assim, para obter as informações do banco de dados dos “Top Albums of All Time” do Rate Your Music, foi utilizado um *scraper* em Python, que acessou as páginas da lista para extrair informações diretamente do HTML das páginas da lista. Foram coletados os dados dos 1000 álbuns mais bem avaliados de todos os tempos, distribuídos nas primeiras 25 páginas da lista, e colocados em um arquivo CSV. Após o tratamento, as informações obtidas incluíam: posição na lista, nome do álbum, artista, ano de lançamento, gêneros principais, gêneros secundários, *descriptors* (sensações associadas às músicas), média de avaliações da comunidade e total de avaliações da comunidade.

B. Construção da Rede de Similaridade

A partir dos dados coletados, foi possível construir um grafo que relaciona os álbuns. A modelagem foi feita definindo os vértices como os álbuns e as arestas do grafo atuam como conectores de similaridade entre os álbuns, com pesos atribuídos proporcionalmente ao grau de similaridade entre eles. O pesos são calculados a partir dos seguintes critérios:

- Artista: peso 15
- Década: peso 5
- Gênero: peso 6
- Gênero secundário: peso 1
- *Descriptor*: peso 0.3

Cada critério foi ponderado de acordo com sua importância relativa na determinação da similaridade entre os álbuns, de acordo com critérios estabelecidos pelos autores. Para calcular o peso final da aresta, soma-se o peso de cada critério semelhante entre os álbuns ao peso total. Para critérios com possibilidade de várias semelhanças, a soma pode ser repetida de acordo com a quantidade de elementos semelhantes.

Para realizar esse processo, foi feito um *script* que roda o cálculo de peso descrito acima para todos os álbuns selecionados dois-a-dois. Para pares cujo peso total é zero, a aresta não é criada. Então, o *script* gera um arquivo CSV seguindo o formato: vértice1, vértice2, peso, peso invertido. Esta última coluna será usada para alimentar algoritmos nos quais o peso é considerado distância, como os para calcular o menor caminho, de modo que as arestas de maior peso (semelhança) também sejam as de menor distância.

Essa abordagem permite não apenas a visualização das relações entre os álbuns na rede, mas também a identificação de padrões e tendências baseados nas similaridades musicais, facilitando, assim, o desenvolvimento de sistemas de recomendação e o aprofundamento do entendimento das

conexões entre os álbuns mais aclamados na comunidade RateYourMusic.

Após a coleta dos dados da lista “Top Albums of All Time”, foi pensado em coletar dados de duração das músicas através da API do Spotify. Entretanto, devido a certas limitações como a ausência de alguns álbuns na plataforma e a existência de múltiplas versões de um mesmo álbum, por exemplo versões normais e *deluxe* (o que dificulta medir corretamente a duração do álbum), foi decidido abandonar a busca por essa informação, de modo que esta coleta não foi concluída.

O arquivo gerado a partir do cálculo dos pesos foi alimentado ao NetworkX – uma biblioteca para o Python que permite trabalharmos com grafos – que gerou o grafo. Usando a mesma biblioteca, calculamos as seguintes métricas: grau, menor caminho médio, menor caminho médio ponderado, centralidade de *closeness*, centralidade de *betweenness* e clusterização. Também foram geradas comunidades, usando o método de Louvain, bem como a modularidade dessa partição do grafo. Para medidas que calculam o caminho, foi usado o peso invertido, pelo motivo mencionado acima.

Para gerar a visualização, apesar de a ideia inicial ser colorir os vértices de acordo com a comunidade, foram encontrados problemas para fazer deste modo, então cada vértice foi colorido aleatoriamente. Depois, foram removidas todas as arestas com peso abaixo da média do peso das conexões. Após isso, foi usado o algoritmo *Spring* para determinar as posições dos vértices e a função *curved_edges*³ para adicionar mais possibilidades de customização às arestas.

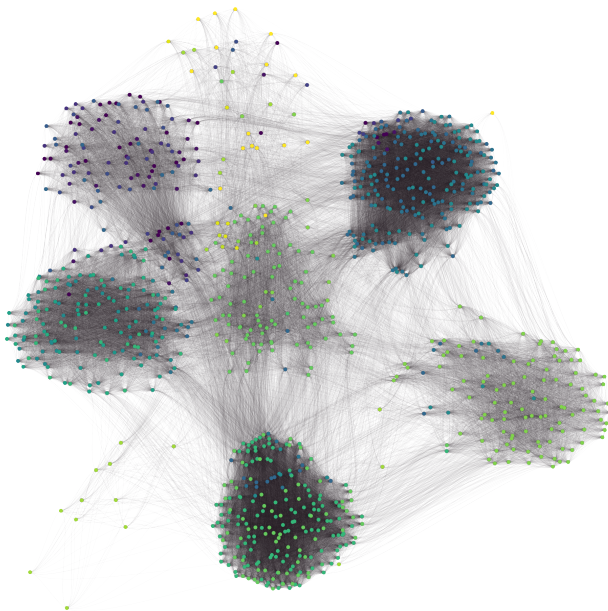


Fig. 1. Grafo da similaridade entre álbuns da lista “Top Rated of All time”

³Disponível em: <https://github.com/beyondbeneath/bezier-curved-edges-networkx>

Todos scripts utilizados para a coleta e processamento dos dados estão disponíveis publicamente no repositório do GitHub deste projeto⁴. Este repositório contém o código-fonte do *scraper* desenvolvido em Python, que foi empregado para extrair os dados diretamente do RateYourMusic, bem como os scripts para análise dos dados e geração e análise do grafo.

Além dos scripts, o repositório também inclui as tabelas (em formato CSV geradas a partir da análise dos dados coletados dos 1000 álbuns mais bem avaliados, detalhando aspectos como a distribuição dos gêneros musicais, *descriptors* e década de lançamento. Essas tabelas fornecem uma visão quantitativa preliminar que servirá de base para as etapas subsequentes de análise. Por fim, também inclui tabelas com métricas geradas a partir do grafo (bem como versões ordenadas de cada uma delas), e um arquivo com métricas gerais e com as comunidades.

O acesso ao repositório permite a reprodução dos resultados apresentados, bem como o uso dos dados e scripts para fins de pesquisa ou aplicação em outros contextos relacionados ao estudo da similaridade entre músicas.

IV. ANÁLISE DOS RESULTADOS

A. Análise das estatísticas do site

A partir dos dados preliminares coletados, foi possível realizar uma análise quantitativa que revelou padrões significativos nas três categorias principais: Gênero, *Descriptors* e Década. A divisão dessas categorias permitiu uma compreensão inicial das tendências e padrões nos álbuns analisados, bem como nos ajudou a refinar os pesos para as arestas.

TABELA I
QUANTIDADE DE ÁLBUNS POR GÊNERO

Gênero	Quantidade de álbuns
Singer-Songwriter	92
Art Rock	50
Progressive Rock	44
Alternative Rock	37
East Coast Hip Hop	37

A análise dos gêneros musicais destaca o gênero *Singer-Songwriter* como predominante, seguido por *Art Rock*, *Progressive Rock* e *East Coast Hip Hop*, sugerindo a formação de comunidades distintas no grafo impulsionadas por eles.

TABELA II
QUANTIDADE DE ÁLBUNS POR DESCRIPTOR

Descriptor	Quantidade de álbuns
Passionate	298
Melodic	277
Energetic	261
Atmospheric	208
Rhythmic	203

A análise dos descritores das músicas indica que termos como *passionate*, *melodic* e *rhythmic* ajudarão a criar comunidades. Como estão muito presentes nos álbuns, seu peso

⁴<https://github.com/grecoww/Scripts-CR-UFABC-2024>

foi reduzido para não se sobressaírem aos outros critérios na criação das arestas.

TABELA III
QUANTIDADE DE ÁLBUNS POR DÉCADA

Década	Quantidade de álbuns
1990	220
1970	219
2000	159
2010	121
1980	118

Além disso, as décadas de 1990 e 1970 têm maior representação na lista dos "Top Albums of All Time", com a década de 2000 também sendo significativa, enquanto décadas anteriores são menos representadas.

B. Análise das métricas Grafo

Feita a análise preliminar, estudaremos o grafo. Primeiramente, consideremos algumas métricas gerais do grafo. O caminho sem ponderação médio tem medida 1.386, indicando que muitos álbuns estão conectados diretamente. Combinando essa informação com o diâmetro e a excentricidade, conclui-se que a maior distância entre dois vértices é 2 e, com ainda mais detalhe, todos os vértices possuem excentricidade 2, de modo que os álbuns possuem ao menos um outro ao qual eles não conseguem se conectar diretamente. Isto faz sentido, pois os critérios para uma conexão ser estabelecida, mesmo que fraca, são lenientes, de modo que espera-se que a rede seja muito conectada, mas com muitas conexões de força média.

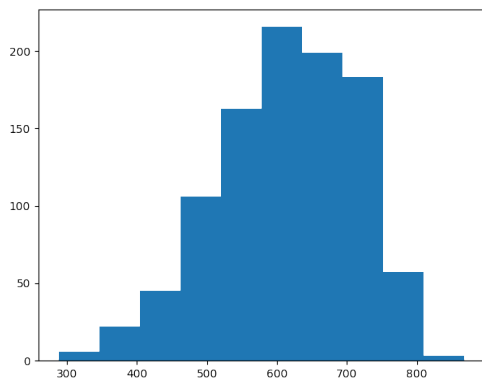


Fig. 2. Histograma de grau sem ponderação do grafo

O fenômeno citado pode ser observado ao compararmos os histogramas das Figuras 2 e 3, que tratam do grau sem ponderação e ponderado, respectivamente. Neles, é notável que os vértices de graus mais extremos estão em menor número. Com isso, é possível concluir que a grande maioria dos vértices está bem conectada: pelo histograma da Figura 2, é perceptível uma grande quantidade de nós com grau entre 500 e 700. A densidade de 0.614, que indica que o grafo possui mais que 60% de todas as arestas possíveis, corrobora a análise feita. Pode-se perceber, também, que o grau do vértice menos conectado é 288, que ainda é muito alto, de modo que todos

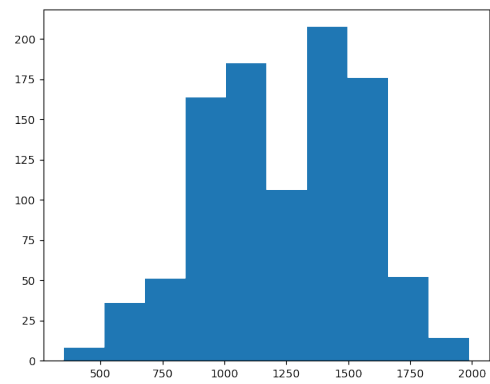


Fig. 3. Histograma de grau ponderado do grafo

os álbuns possuem alguma conexão na lista. A partir disso, é possível reparar algumas tendências na distribuição: a maior parte dos álbuns de música erudita possuem baixo grau, pois não são muitos álbuns que entram na lista e por vezes eles ainda não compartilham mesmos gêneros, o que contribui com as poucas conexões. Álbuns que possuem poucos gêneros e *descriptors* listados também possuem graus mais baixos, a exemplo do álbum "XXX", de Danny Brown, que só possui 1 gênero principal e 2 secundários listados e é o álbum de menor grau, apesar de ser de gêneros relativamente populares.

Já nos álbuns de maior grau, percebemos vários da década de 1970 e 1990, pois já possuem mais de 200 conexões garantidas, bem como álbuns do gênero *Singer-Songwriter*. Ambas essas percepções permitem entender alguns álbuns que parecem não se encaixar em altos graus, como é o caso do álbum "Complete / Sämtliche / Les 21 nocturnes" de Claudio Arrau, que possui poucos gêneros e é um álbum de música erudita, mas é da década de 1970 e possui *descriptors* muito comuns, como *passionate* e *atmospheric*. A capacidade da presença de múltiplos gêneros e *descriptors* influenciar a quantidade de conexões é uma característica do modelo que pode o tornar mais tendencioso a certos álbuns.

Considerando agora o grau ponderado dos vértices, vemos que os álbuns de maior grau são todos das décadas de 1970 e 1990, e de gêneros como *Singer-Songwriter*, *Folk Rock*, *Art Rock*, e outras variações do Rock, como Rock Alternativo, além de vertentes mais experimentais do Pop. Na categoria de artistas, é possível notar que David Bowie, Bob Dylan, George Harrison, Caetano Veloso e Neil Young que, juntamente com Crazy Horse, pois possuem muitos álbuns em conjunto, recebem maior destaque, por possuírem diversos álbuns entre os mais centrais se considerarmos o grau. Assim, é possível identificar que, como possuem maior conectividade entre os álbuns da lista, essas características e artistas citados são mais presentes e exercem maior influência no conjunto de dados, o que nos leva a considerar eles como parte do viés da plataforma.

Analisando outras métricas de centralidade, podemos chegar a novas conclusões. Como o diâmetro do grafo é 2, um maior grau sem ponderação também resulta em uma maior

centralidade de *closeness* (proximidade), pois o vértice possuirá mais conexões diretas, de modo que os álbuns de maior grau, também serão os de maior *closeness*. A centralidade de *betweenness* (intermediação) nos mostra que alguns álbuns de música erudita são mais centrais considerando essa medida. Isto pode ser explicado pela falta de conexão deles a outros álbuns, enquanto que estão internamente conectados, de modo que eles intermedeiam a conexão com outros nós. Apesar disso, os valores de *betweenness* são muito baixos, posto que a rede é muito interconectada, reduzindo a necessidade de vértices que conectam dois outros.

Para a centralidade *eigenvector* (de autovetor), podemos observar que muitos dos álbuns que ficaram entre os primeiros para o grau ponderado continuaram mantendo sua posição, destacando os de Neil Young e “Artaud”, de Pescado Rabioso. Como esta medida de centralidade usa o grau dos vizinhos para calcular a centralidade do vértice, é esperado que álbuns conectados à outros com alta centralidade de grau possuam *eigenvector* alta. Assim, fica evidente que os álbuns de Neil Young apareceriam no topo desta centralidade também.

Por fim, podemos estudar como os vértices se conectam a outros. Uma das métricas que permite isso é a assortatividade, que calcula a tendência de vértices se conectarem a outros semelhantes usando algum critério para comparação, como o grau. Como o grafo possui assortatividade 0.073, pode-se concluir que os nós não possuem preferência de grau ao se conectar a outros. Dada essa aparente aleatoriedade das conexões, também podemos analisar a clusterização do grafo; o coeficiente de clusterização médio é de 0.027, que é baixo se compararmos com redes de mundo pequeno, por exemplo. Isso nos indica que o grafo não segue esse modelo de rede e que não há tendência de clusterização entre os vértices, que se conectam de forma mais aleatória e que não formam tantos núcleos.

C. Análise das comunidades

Ao rodar o algoritmo de Louvain, fixando uma *seed* para garantir que os resultados não mudassem durante o estudo e usando resolução 3, que favorece a formação de comunidades menores, foram geradas 22 comunidades, com tamanhos bastante variados – de 2 a 157 álbuns. Essa partição resultou em uma modularidade de 0.37 que, apesar de não ser muito alta, em decorrência da quantidade de arestas, resultando em muitos vértices de comunidades conectados ao exterior, indica que existem conexões mais fortes entre determinados vértices que em um grafo aleatório.

Essas comunidades, que serão analisadas no restante desta sub-seção, servem para identificar grupos de músicas similares, possibilitando a identificação de um viés, bem como a aplicação no algoritmo de recomendação descrito na seção seguinte. Vamos, então, examinar as comunidades, focando nas que possuem um tamanho razoável.

Logo na primeira comunidade, é possível encontrar alguns dos maiores nomes do Rock Progressivo, como Genesis, Yes, Jethro Tull, King Crimson, Frank Zappa, entre outros. Aparecem também gêneros como *Jazz-Rock*, *Art Rock* e variações do

Rock Progressivo, que são todos gêneros tangenciados pelas bandas deste gênero [8]. Outra informação é de que todos os álbuns são da década de 1970, de modo que podemos ver que, além da comunidade representar uma região de maior interconexão no grafo, ela também apresenta sentido musical, representando similaridade entre estes álbuns.

A segunda comunidade contém álbuns de *Heavy Metal* e *Hard Rock* de bandas como Black Sabbath, Judas Priest e Deep Purple, todos lançados na década de 1970. A quarta engloba álbuns da mesma década, mas que tem, principalmente, o gênero *Singer-Songwriter* e subgêneros do *Folk*, apesar de também incluir subgêneros do *Rock* e do *Jazz*, além de MPB, mas que ainda conversam com os outros citados.

A quinta comunidade contém artistas como Brian Eno, Paul McCartney, David Bowie e Kate Bush, e gêneros como *Soft Rock*, *Art Pop*, *Glam Rock*, *Progressive Pop* e *Ambient*. Essa comunidade apresenta artistas ativos na década de 1970, muitos dos quais se conheciam e já haviam trabalhado juntos. Um exemplo disso é que Kate Bush, em uma entrevista para a revista Mojo4Music, expressou sua admiração pela música de David Bowie [6]. Passando para a sétima, temos uma comunidade centrada em músicas dos gêneros *Art Rock*, *Noise Rock*, *Emo*, *Shoegaze* e *Alternative Rock* e de artistas como R.E.M., Soda Stereo, Radiohead, Sonic Youth, entre outros, caracterizadas como parte do movimento Emo [9].

As comunidades maiores são consideravelmente mais difíceis de serem explicadas, mas uma tendência presente em todas as comunidades, à exceção das 10, 21 e 22, é que possuem álbuns da mesma década, por mais que esse peso seja relativamente pequeno ao compararmos com os do gênero e do artista. Isto pode indicar que, por pertencerem a um mesmo período da música, eles possuem características de gênero, artista e *descriptor* que intensificam a conexão entre eles. Este é um dos principais fenômenos percebidos pela distribuição das comunidades.

Apesar disso, é possível identificar comunidades nas mesmas décadas, mas com artistas e gêneros diferentes, indicando que existiam vários movimentos na música ao mesmo tempo, e que são mais conectados entre si. Também é possível notar mesmo gêneros em décadas diferentes, e com conexões um pouco diferentes, sendo possível entender que, ao longo do tempo, os artistas desses gêneros mudaram algumas características, de maneira que passaram a ter maiores conexões com outros grupos. Um exemplo disso é o Rock Progressivo, que aparece de gênero principal em combinações diferentes a depender da década: além da primeira comunidade, já comentada, a décima quinta comunidade apresenta álbuns desse gênero, mas que são da década de 1990 e se relacionam com diversos gêneros, incluindo vários subgêneros do Rock e da música experimental. Enquanto isso, a décima sexta comunidade apresenta também álbuns de Rock Progressivo, mas desta vez apenas da banda King Crimson, indicando que esses álbuns estão muito mais conectados entre si do que com outros álbuns de Rock Progressivo da mesma época.

Duas das comunidades que não seguem essa divisão delimitada, as 21 e 22, são exclusivamente de música erudita e Jazz,

que são gêneros que aparecem de forma mais esporádica na lista, assim, possivelmente estão mais conectados entre si do que com outros gêneros.

Em geral, pode-se concluir que, apesar de a década não aparentar ter tanto peso para a aresta, ela acaba se tornando um fator decisivo para formar as comunidades. Isso pode ser indicar que álbuns da mesma década possuem características específicas de gênero, artista e *descriptor*, aumentando a similaridade ainda mais.

V. ALGORITMO DE RECOMENDAÇÃO

Com o grafo processado e analisado, pode-se realizar a implementação do algoritmo de recomendação. Interpretando o peso invertido da conexão entre álbuns como as distâncias entre dois vértices, utilizou-se o algoritmo de Dijkstra para o cálculo das menores distâncias de um dado vértice para todos os demais vértices. Dessa forma, usaremos as distâncias obtidas para determinar a similaridade entre dois álbuns no grafo; quanto menor a distância entre dois álbuns, maior a similaridade entre eles. O algoritmo baseia-se em três pilares para sugerir álbuns, que são: avaliação, cálculo de distância agregada e sugestão.

Primeiro, os álbuns já ouvidos são classificados em curtidos e não curtidos pelo usuário. Após isso, calcula-se os álbuns mais similares a todos os álbuns já ouvidos (com o método descrito acima). Agora, basta calcular a distância agregada de cada álbum, isto é, a distância de cada álbum para o gosto musical do usuário; a distância agregada de cada álbum será a distância atual (no começo da execução todos os álbuns tem distância zero do gosto do usuário) somada ao valor da distância para os álbuns não curtidos e subtraído em módulo da distância para os álbuns curtidos, o que fará com que álbuns similares a álbuns curtidos, tenham distância agregada menor que álbuns similares a álbuns não curtidos. Ademais, as distâncias agregadas ganham um fator multiplicativo de 1.5x caso o álbum sugerido esteja na mesma comunidade que um álbum já ouvido, aumentando essa disparidade.

Finalmente, as distâncias agregadas de cada álbum são organizadas por ordem crescente, e o algoritmo retorna o álbum com menor distância agregada para o usuário, ou seja, o álbum mais similar com seu gosto. Este algoritmo de recomendação junto à uma simples interface para facilitar o uso deste, pode ser encontrado no repositório do GitHub deste projeto.

VI. CONCLUSÃO

Em resumo, foram coletados dados do RateYourMusic para criar uma rede de álbuns, conectando-os por similaridade seguindo diversos critérios. Esta rede foi estudada para entendermos como os álbuns interagem, se distribuem e conectam. Com base nisso, criamos um algoritmo de recomendação que usa o peso entre os álbuns e as comunidades geradas no grafo para encontrar álbuns mais similares.

Usando as métricas do grafo, foi possível identificar que a rede é extremamente densa, por causa da forma como o modelo foi definido. Apesar disso, o peso das arestas exerce

papel fundamental na interpretação da rede, de modo que muitas das métricas são dependentes dele. Por causa das muitas conexões, o grafo possui um diâmetro muito baixo, de modo que os álbuns estão, no máximo, a 2 passos de distância de todos os outros. Além disso, usando a assortatividade e a clusterização, foi possível concluir que as conexões se distribuem aleatoriamente e não seguem padrões de *cluster* ou de similaridade de grau.

Estudando as métricas de centralidade, percebe-se que os álbuns do gênero *Singer-Songwriter* aparecem como alguns dos mais centrais pelo grau ponderado (uma das métricas que julgamos ser extremamente importante, considerando o modelo desenvolvido). Além disso, nota-se que álbuns de subgêneros do Rock, como *Folk Rock* e *Art Rock*, bem como das décadas de 1970 e 1990 também estiveram presentes. A centralidade de intermediação apresentou os resultados mais distintos, pois álbuns que se conectam a outros de pouco grau na rede servem de intermediários entre esses e vértices mais centrais. Isto fica evidente ao considerar que os álbuns mais centrais por *betweenness* são de música erudita, que é um dos estilos menos presentes no conjunto de dados.

Estudando a partição do grafo, notamos que a década exerce grande influência na formação das comunidades, indicando que a década também influencia a distribuição de gêneros e de artistas, que se conectam mais com outros da mesma época. As comunidades mostram que é possível identificar grupos mais interconectados no grafo, indicando que são mais semelhantes entre si. Tais grupos permitiram estudar possíveis divisões entre o viés da comunidade do RateYourMusic. Alguns exemplos disso foram a comunidade com álbuns de Rock Progressivo, bem como a comunidade que engloba principalmente gêneros musicais predominantes no movimento *Emo*.

Usando o conhecimento sobre as comunidades e sobre o peso da conexão entre dois álbuns, foi possível desenvolver um algoritmo de recomendação que leva em consideração álbuns que o ouvinte gostou ou não para gerar uma pontuação para cada álbum. Um outro critério utilizado foi a presença de cada disco em comunidades, diminuindo a distância para comunidades que contém álbuns gostados e aumentando-a para o caso contrário. O algoritmo atua de forma simplificada, mas ilustra uma possível implementação de um sistema de recomendação.

Em um sistema real, os critérios para atribuir uma pontuação de recomendação possivelmente seriam mais complexos e haveriam mais álbuns na base de dados, permitindo maior refinamento das recomendações e maior gama de possibilidades de álbuns.

VII. PARTICIPAÇÃO DOS ALUNOS

Lucas: atuou na programação da geração, análise e visualização do grafo, criação do *scraper* para obter as informações gerais dos álbuns diretamente do rate your music, idealização do algoritmo de recomendação e escrita do relatório.

Luis: fez as análises do que foi coletado a partir das APIs e ajudou na escrita e detalhamento do relatório. Além disso,

coletou informações e ajudou na criação dos parâmetros utilizados para fazer o grafo.

Nicolas: atuou na programação da geração, análise e visualização do grafo, interação com a API do Spotify para obtenção de dados, programação do algoritmo de recomendação e escrita do relatório.

REFERÊNCIAS

- [1] T. DeNora, *Music in everyday life*. Cambridge: Cambridge University Press, 2000.
- [2] IFPI, 2017, IFPI facts and stats, International Federation of the Phonographic Industry. [Online]. Available: <http://www.ifpi.org/facts-and-stats.php>
- [3] T. Trefzger, M. Rose, C. Baccarella, and K.I. Voigt. Streaming Killed the Download Star! How the Business Model of Streaming Services Revolutionizes Music Distribution. *Journal of Organizational Advancement, Strategic and Institutional Studies*, Vol. VII, No. 1, 2015. [Online]. Available at SSRN: <https://ssrn.com/abstract=2587176>.
- [4] M. Mondelli, L. Gadelha Jr., and A. Ziviani. “O Que os Países Escutam: Analisando a Rede de Gêneros Musicais ao Redor do Mundo”, in *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*, Natal, 2018.
- [5] A.-L. Barabási and M. Pósfai, *Network science*. Cambridge: Cambridge University Press, 2016. <http://barabasi.com/networksciencebook/> (acessado 14 de Ago, 2024)
- [6] “Kate Bush On David Bowie: ‘He was just the right amount of weird...’” Mojo, 2024. <https://www.mojo4music.com/articles/stories/kate-bush-pays-tribute-to-david-bowie/> (acessado 27 de Ago, 2024).
- [7] A. A. Hagberg, D. A. Schult and P. J. Swart, “Exploring network structure, dynamics, and function using NetworkX”, in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008
- [8] E. Macan, “Rocking the classics”. Editorial: S.L.: Oxford University Press, 1997.
- [9] D. Buckingham, “Emo: The Paradox of Contemporary Youth Culture”. David Buningham. <https://davidbuckingham.net/wp-content/uploads/2017/10/emo.pdf> (acessado 27 de Ago, 2024).
- [10] “Best albums of all time”. Rate Your Music. <https://rateyourmusic.com/charts/top/album/all-time/> (acessado 21 de Jul, 2024).
- [11] M. Schedl, P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas, “Music Recommender Systems,” *Recommender Systems Handbook*, pp. 453–492, 2015, doi: https://doi.org/10.1007/978-1-4899-7637-6_13.
- [12] M. DEMARCO e G. dos Santos, “Mecanismos de Recomendação no Spotify: uma análise dos elementos que configuram a sugestão de conteúdos musicais na atual interface da plataforma”, *CM*, vol. 14, p. 18, nov. 2021.