ESPAÇOS DE SIMILARIDADE DE CONTEÚDOS DE MÍDIA GERADOS A PARTIR DE DADOS DE USUÁRIOS EM REDES SOCIAIS ONLINE

PEDRO HENRIQUE FERNANDES DE HOLANDA

ESPAÇOS DE SIMILARIDADE DE CONTEÚDOS DE MÍDIA GERADOS A PARTIR DE DADOS DE USUÁRIOS EM REDES SOCIAIS ONLINE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Ana Paula Couto da Silva
Coorientador: Olga Nikolaevna Goussevskaia

Belo Horizonte
Julho de 2016

Holanda, Pedro Henrique Fernandes de

H722e Espaços de similaridade de conteúdos de mídia gerados a partir de dados de usuários em redes sociais online / Pedro Henrique Fernandes de Holanda. — Belo Horizonte, 2016
xxiv, 79 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais

Orientadora: Ana Paula Couto da Silva
Coorientadora: Olga Nikolaevna Goussevskaia

1. Computação – Teses. 2. Redes sociais on-line. 3. Redução de dimensionalidade. 4. Embedding de grafos. I. Orientadora. II. Coorientadora. III. Título.

CDU 519.6*75(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Espaços de similaridade de conteúdos de mídia gerados a partir de dados de usuários em redes sociais online

PEDRO HENRIQUE FERNANDES DE HOLANDA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. ANA PAULA COUTO DA SILVA - Orientadora
Departamento de Ciência da Computação - UFMG

PROFA. OLGA NIKOLAEVNA GOUSSEVSKAIA - COORIENTADORA
Departamento de Ciência da Computação - UFMG

PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES
Departamento de Ciência da Computação - UFMG

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 26 de julho de 2016.

A Darci Holanda, Angela, Carlos, Luiz e Ana Carolina.

Agradecimentos

Agradeço primeiramente às minhas orientadores, Ana Paula e Olga, por me oferecerem total apoio durante este mestrado. Suas orientações e ensinamentos foram fundamentais não só para este trabalho, mas para mim como ser humano e estudante. À Glívia e ao Ismael por darem o primeiro empurrão e despertado em mim o interesse na área acadêmica. Aos meus colegas de pesquisa Bruno, Fujii e João Paulo pelas horas trabalhando juntos em artigos e gráficos. Aos meus pais, Darci Holanda e Angela, por todo o amor, carinho e dedicação ao longo de toda a minha vida. Aos meus irmãos, Carlos, Luiz e Rodra, pela ajuda sempre que precisei. À minha parceira e companheira Carol pelo amor, por andar ao meu lado e sempre me dar aquela força a mais, necessária para que eu não interrompa minha caminhada.

Agradeço também a meu avô e avó, minhas tias, Branca, Dedê e Ieié. Aos meus amigos, dentre eles Marcus e Hudson pelas conversas e ideias de pesquisa. À família da Carol por permitirem meus devaneios em sua casa. Por fim, agradeço a todos os professores e funcionários do DCC/UFMG.

"Caminante, son tus huellas
el camino y nada más;
caminante, no hay camino,
se hace camino al andar.
Al andar se hace camino
y al volver la vista atrás
se ve la senda que nunca
se ha de volver a pisar.
Caminante no hay camino
sino estelas en la mar."
(Antonio Machado)

Resumo

A maneira como as pessoas assistem aos programas de TV, filmes e/ou escutam músicas tem mudado dramaticamente nos últimos anos. Atualmente as pessoas escutam músicas ou assistem aos seus programas preferidos através da internet com conexões de alta velocidade e utilizando várias fontes diferentes (p.ex, Youtube, Netflix e Last.fm). Além dos computadores, as pessoas acessam estes conteúdos a partir de smartphones, tablets e smartTVs. Um outro ponto interessante é que a quantidade e a diversidade do conteúdo disponível aumenta cada vez mais, uma vez que as tecnologias disponíveis para a produção de vídeo e música estão cada vez mais acessíveis.

Na maioria das vezes esta navegação é realizada através de estruturas pouco intuitivas, como listas sequenciais, organizadas em ordem alfabética ou de forma hierárquica. Este trabalho tem como objetivo principal propor estruturas mais amigáveis para a navegação. Para alcançar este objetivo, este trabalho está dividido em duas partes principais: (1) caracterização do comportamento dos usuários em relação a programas de TV, música e cinema, através de redes sociais online; (2) construção e avaliação de uma estrutura de dados que permita novas formas de navegação, visualização e análise de coleções de mídia pessoais, tais como *playlists* musicais e guias de programas de TV.

A partir dos dados coletados de redes sociais online, a estrutura de dados gerada é um espaço Euclidiano multidimensional, no qual cada item é representado por um conjunto de coordenadas e a distância entre dois itens representa a similaridade, onde a distância é curta para itens muito similares e longa caso contrário. Para a geração desta estrutura são utilizadas e comparadas técnicas de *embedding* de grafos.

As avaliações da qualidade das estruturas geradas se baseiam em métricas quantitativas e qualitativas. As métricas quantitativas avaliam se o espaço gerado é capaz de preservar as similaridades dos itens e se itens com o mesmo gênero e/ou artista estão próximos em termos de distância Euclidiana. A avaliação qualitativa é feita através da análise manual das vizinhanças de alguns pontos do espaço.

Palavras-chave: Redes Sociais Online, Caracterização de Dados, Redução de Dimensionalidade, Embedding de Grafos.

Abstract

The way people watch TV shows, movies and/or listen to music has changed dramatically in recent years. Currently people listen to music or watch their favorite programs over the internet with high-speed connections and using several different sources (e.g., Youtube, Netflix and Last.fm). In addition to computers, people access this content from smartphones, tablets and SmartTVs. Another interesting point is that the amount and diversity of available content increases more and more as the available technologies for video and music production are increasing.

Considering also that most of the time navigating this content is performed through sequential lists, in alphabetical or hierarchical orders, this work focuses on two main points: (1) characterize the behavior of users in relation to TV show, music and film through online social networks; (2) construction and evaluation of a data structure that allows new forms of navigation, visualization and analysis of personal media collections, such as music playlists and TV program guides.

The generated data structure is a multi-dimensional Euclidean space, where each item is represented by a set of coordinates and the distance between two items is the similarity, where the distance is short to similar items, and long otherwise. Graph embedding techniques are used and compared to generate this structure, given a similarity graph constructed from user data in online social networks.

The evaluation of the structure is made quantitative, since items with the same genre, artist, album and other attributes specific to each area should be close to the generated structure; and qualitative, since the quality of this structure is subjective.

Keywords: Online Social Networks, Data Characterization, Dimensionality Reduction, Graph Embedding.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A evolução da computação e o advento da Internet causaram uma importante revolução nas formas de armazenar e distribuir conteúdos de mídia nos últimos anos [Torrez-Riley, 2011]. Os atuais formatos de mídia (p.ex, mp3, mp4) combinados com a capacidade de armazenamento dos dispositivos eletrônicos permitem às pessoas possuírem milhares de músicas e vídeos nestes dispositivos.

Além da alta capacidade de armazenamento de conteúdo de mídia, os serviços de *streaming* (p.ex, Youtube[1], Netflix[2] e Last.fm[3]) fazem com que seus usuários escutem músicas ou assistam vídeos em uma escala muito maior e nos mais variados dispositivos. Todo esse conteúdo pode ser acessado em computadores pessoais, smartphones, tablets, smart TV entre outros.

Juntamente com o surgimento dos serviços especializados em distribuição de mídias e com o advento das redes sociais online, foram inseridas novas formas de compartilhar e interagir com conteúdos de mídia. Uma pessoa pode compartilhar seu gosto musical, qual programa de televisão está assistindo no momento ou qual vídeo deseja visualizar. A grande quantidade e qualidade dos dados gerados pela interação dos usuários com estas redes e entre si propiciam um conjunto de informações extremamente valioso que pode ser utilizado para impulsionar novos artistas, geração de novos conteúdos, popularidade e consequentemente, gerar lucro para diversos setores da economia.

Apesar de toda essa quantidade de músicas e vídeos disponíveis, que podem ser acessados pelos mais variados dispositivos, o acesso a estes itens normalmente é realizado através de listas sequenciais, organizadas em ordem alfabética ou hierárquica.

[1] www.youtube.com
[2] www.netflix.com
[3] www.last.fm

Essa forma de organização tende a dificultar a visualização de coleções de mídia, principalmente quando relacionada a milhares de itens. Além disso, muitas das metodologias de acesso utilizadas atualmente não consideram (ou consideram de forma limitada) a riqueza de informações geradas pelas redes sociais online, sendo que estas podem ser utilizadas para tornar mais intuitivo o acesso aos conteúdos de diversas mídias. Assim, esta dissertação tem como principais objetivos entender o comportamento de usuários em redes sociais online que focam em programas de TV e músicas, bem como definir espaços de similaridades destas mídias utilizando os dados provenientes dessas redes sociais.

Os espaços apresentados neste trabalho podem ser utilizados, por exemplo, em aplicações em que estes conteúdos podem ser acessados de maneira mais amigável e intuitiva, como a aplicação Mixtape[4], proposta por Cardoso et al. [2016].

1.1 Motivação

Poucos trabalhos na literatura propõem estratégias mais elaboradas para a navegação através de conteúdos de diferentes mídias. Podemos destacar algumas abordagens voltadas para o domínio da música e de programas de televisão [Knees et al., 2006; Neumayer et al., 2005; Goussevskaia et al., 2008]. Goussevskaia et al. [2008] apresentam uma abordagem de navegação em coleções de músicas baseada em uma estrutura que consiste de um espaço Euclidiano multidimensional, onde cada item é representado por um conjunto de coordenadas e a distância entre dois itens representa a não-similaridade entre esses itens. Nesta dissertação essa abordagem é estendida, aprimorando a metodologia através da avaliação de diferentes algoritmos e técnicas de pré-processamento para a construção desse espaço de similaridades entre itens.

A estrutura utilizada neste trabalho é um espaço Euclidiano de baixa dimensionalidade, onde cada item (música ou filme) é representado por um conjunto de coordenadas. Este espaço, definido como espaço de similaridades, deve ser capaz de preservar itens similares próximos e itens diferentes distantes. As principais vantagens do uso desse espaço são: o possível uso em visualizações de coleções de mídia, quando este espaço Euclidiano possui uma quantidade de dimensões visualizável (três dimensões ou menos); propriedades do espaço Euclidiano, como volume e direção, que permitem a implementação de diferentes formas de navegação; e a baixa dimensionalidade desse espaço também permite ganhos de armazenamento dessa estrutura.

[4] www.projectmixtape.org

Dentre as formas de navegação utilizando propriedades do espaço Euclidiano, podemos destacar [Cardoso et al., 2016], um trabalho anterior onde um dos espaços aqui apresentados é utilizado para propôr uma navegação baseada em vetores se locomovendo por este espaço, simulando um usuário "andando" por uma coleção de músicas. Com isso o senso de direção neste espaço foi utilizado para a geração de *playlists*, ou listas de músicas, com transições mais suaves entre os gêneros e artistas.

Uma vez que dispositivos móveis tem uma capacidade menor de armazenamento e podem nem sempre estar conectados à Internet, a baixa complexidade de espaço da estrutura pode ser utilizada. Dispositivos móveis podem manter em sua memória apenas as coordenadas de suas coleções locais de mídia para que diferentes formas de navegação sejam implementadas. Assim, em uma aplicação cliente-servidor, os dispositivos clientes precisam manter apenas as coordenadas de suas coleções locais atualizadas, não necessitando de toda a estrutura para uma navegação *offline*.

Consultas como *"Eu gostaria de um suspense mais leve."* ou *"Quero uma música que se pareça com essas outras duas"*, podem ser respondidas ao se sugerir um filme que esteja entre os filmes de suspense, mas na direção do centro de massa de filmes de comédia, ou sugerindo uma música que esteja entre outras duas músicas no espaço de similaridades.

1.2 Principais contribuições

As principais contribuições dessa dissertação são:

1. Uma análise aprofundada de redes sociais online focadas em conteúdos de mídia, principalmente da rede social tvtag. A caracterização apresentada nesta dissertação contribui para a melhor compreensão dos efeitos de fenômenos como a TV Social, ou o uso de redes sociais para compartilhar dados sobre programas de televisão e filmes, e o comportamento *second-screen*, que é o uso de múltiplos dispositivos ou telas em conjunto como ao compartilhar algo que está sendo visualizado na televisão ao mesmo tempo [Basapur et al., 2012]. Também destaca possíveis usos desses dados para a predição da popularidade de programas de televisão e filmes antes de suas estreias.

   Além de apresentarem contribuições científicas em relação ao estudo do comportamento dos usuários desses sistemas, as caracterizações apresentadas nesta dissertação também são utilizadas na compreensão dos espaços de similaridades criados, já que esses espaços são gerados através desses dados.

2. Criação de um espaço de similaridades de mídia utilizando dados de redes sociais online com o objetivo de prover uma estrutura de dados que permita diferentes meios de navegação, visualização e análise de coleções de mídia. Nesta dissertação, são focados os dados relacionados a programas de TV e filmes (coletados a partir do tvtag) e de músicas (coletados a partir do Last.fm).

3. Comparação entre técnicas de redução de dimensionalidade e técnicas específicas para *embedding* de grafos em espaços Euclidianos na geração do espaço de similaridades. Propomos diferentes métricas qualitativas para a avaliação dos espaços gerados.

A partir dos resultados alcançados nesta dissertação, os seguintes trabalhos foram publicados:

1. TV Goes Social: Characterizing User Interaction in an Online Social Network for TV Fans. Publicado em 15th International Conference on Web Engineering (ICWE), 2015. Pelos autores Pedro H. F. Holanda, Bruno Guilherme, Ana Paula Couto da Silva, Olga Goussevskaia. Este trabalho possui os principais resultados da caracterização dos dados relacionados a filmes e programas de televisão;

2. Mapeando o universo da mídia usando dados gerados por usuários em redes sociais online. Publicado em XXXIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC), 2015. Pelos autores Pedro H. F. Holanda, Bruno Guilherme, João Paulo V. Cardoso, Ana Paula Couto da Silva, Olga Goussevskaia. Neste trabalho é demonstrada a construção e análise de um espaço de similaridades de mídia, baseado nos dados relacionados a filmes e programas de televisão utilizados nesta dissertação;

3. Mixtape: Using real-time user feedback to navigate large media collections. A ser publicado em 17th International Society for Music Information Retrieval Conference (ISMIR), 2016. Pelos autores João Paulo V. Cardoso, Luciana Fujii Pontello, Pedro H. F. Holanda, Bruno Guilherme, Olga Goussevskaia, Ana Paula Couto da Silva. Neste trabalho são propostas e analisadas técnicas de navegação geradas a partir do espaço de similaridades de mídia gerado nesta dissertação.

1.3 Organização

Este trabalho é organizado da seguinte maneira. O capítulo 2 apresenta os trabalhos relacionados. Os capítulos 3 e 4 apresentam as caracterizações das redes sociais tvtag e

Last.fm, respectivamente. O capítulo 5 detalha os algoritmos utilizados na construção do espaço de similaridades. O capítulo 6 apresenta as métricas e análises experimentais entre os métodos apresentados para geração do espaço de similaridades. Finalmente, o capítulo 7 apresenta as conclusões e trabalhos futuros.

Capítulo 2

Trabalhos relacionados

Este capítulo revisa a literatura relevante relacionada à dissertação. A seção 2.1 apresenta artigos relacionados à caracterização de redes sociais online, com ênfase em redes sociais online de programas de TV, filmes e música. As técnicas encontradas na literatura para definição de similaridade de conteúdo são descritas na seção 2.2. A seção 2.3 apresenta os principais algoritmos de redução de dimensionalidade e *embedding* de grafos para geração de espaços de similaridade de conteúdos de diferentes mídias.

2.1 Redes sociais online

A importância das redes sociais online tem motivado vários pesquisadores a caracterizar diferentes aspectos das redes mais populares. Entender como as pessoas se comportam nestas redes é uma ferramenta poderosa para prever se estas redes irão sobreviver ou não [Ribeiro, 2014]. Além disso, resultados da caracterização destas redes permitem entender como as pessoas interagem entre si e como as mesmas assimilam a grande quantidade de informação compartilhada diariamente. No contexto deste trabalho a similaridade entre dois itens é gerada a partir de dados de redes sociais online e, portanto, os trabalhos apresentados nesta seção demonstram como os usuários dessas redes se comportam.

Os autores em [Backstrom et al., 2012] apresentam resultados sobre o cálculo das distâncias no grafo de amizades do Facebook a partir de toda a rede de usuários ($\approx$ 721 milhões de usuários, $\approx$ 69 bilhões de arestas). A distância média observada foi de 4,74, correspondendo a 3,74 nós intermediários. Em [Backstrom & Kleinberg, 2014], os autores focam na identificação das pessoas mais importantes na rede, ou seja, aqueles que estão conectados por laços sociais fortes. A questão principal a ser respondida pelo estudo era se o conhecimento de todas as conexões de amizade na

rede poderia ajudar a reconhecer os parceiros românticos inseridos na rede. Através dos resultados apresentados no trabalho, foi possível mostrar que é possível, com alta acurácia, estabelecer os parceiros na rede através da sua estrutura topológica. Ugander et al. [2011] analisaram o grafo social dos usuários ativos nesta rede social, focando no cálculo de diversas métricas topológicas, como distribuição do grau e coeficiente de clusterização.

Cha et al. [2010] analisam a influência dos usuários no Twitter utilizando três diferentes métricas: grau de entrada, *retweets* e menções. Os resultados mostraram que a métrica de grau de entrada representa popularidade do usuário, no entanto, não está correlacionada com outras noções importantes de popularidade como o engajamento das pessoas (que pode ser representado pelo número de *retweets* e menções). Os autores em [Kwak et al., 2010] coletaram a rede completa do Twitter, obtendo informações de 41,7 milhões de perfis de usuários, 1,47 bilhões de relações sociais, 4.262 *trending topics*, e 106 milhões de *tweets*. A rede seguidores-seguidos possui uma topologia com diâmetro pequeno, baixa reciprocidade e uma distribuição diferente da distribuição *power-law*.

A rede social do Google+ foi analisada por Gonzalez et al. [2013], através de uma caracterização detalhada baseada em medições de larga escala. Neste trabalho foram identificados os principais componentes da estrutura da rede, bem como caracterizados os principais atributos dos usuários e a sua evolução temporal. Um dos resultados apresentados é que a atividade dos usuários está diminuindo gradualmente, e somente uma pequena fração dos usuários é ativa na rede.

2.1.1 Redes sociais focadas em TV

Os hábitos em torno dos programas de TV e consumo de música mudaram nos últimos anos. Atualmente, o compartilhamento da experiência e opiniões em relação aos programas de TV e músicas rompeu a barreira de nossas casas e essas experiências passaram a ser compartilhadas e discutidas, por indivíduos ou grupos, através de diferentes dispositivos [Bondad-Brown et al., 2012]. Por exemplo, os autores em [Narasimhan & Vasudevan, 2012] estudam a viabilidade do uso de informações de atividades sociais como um mecanismo de detecção e caracterização do comportamento dos telespectadores. No entanto, os autores exploram dados do Twitter e Topsy[1], sem a caracterização de uma rede social online voltada para conteúdos de TV e música. Acreditamos que redes sociais online específicas fornecem um melhor entendimento de como as pessoas se relacionam com o conteúdo televisão e de música.

[1] www.topsy.com

Torrez-Riley [2011] apresenta uma perspectiva histórica sobre o papel da televisão na interação social e analisa como uma nova era da fragmentação do consumo do conteúdo pelos usuários modificou este papel. No entanto, não foram apresentados resultados quantitativos. Basapur et al. [2012] descrevem o desenvolvimento e aplicação da experiência através de multitelas ou *second-screen*, usando um protótipo chamado FanFeeds, que permite gerar e consumir conteúdo relacionados com diferentes séries de TV. Os participantes do experimento revelaram que o protótipo permitiu uma melhor conexão entre a TV e que através dele foi possível enriquecer contatos sociais, a partir das discussões geradas em torno destes programas. Geerts et al. [2011] ressaltam a importância de assistir e comentar programas de TV com os amigos a partir de redes sociais online e como esta atividade influencia as relações entre pessoas.

Os autores em [Mukherjee & Jansen, 2014] investigaram as interações dos usuários com outros dispositivos durante a transmissão de programas de TV ao vivo e gravados. Eles também exploraram o papel de diferentes dispositivos como smartphones e tablets na interação com multitelas através da análise de mais de 418.000 tweets para três programas populares de TV.

Nesta dissertação (capítulo 3) apresentamos uma caracterização detalhada da rede social online tvtag, focada em fãs de filmes e programas de TV. Pelo nosso conhecimento, uma caracterização profunda deste tipo de rede não tinha sido ainda proposta na literatura. Os resultados desta caracterização foram publicados em [Holanda et al., 2015b,a].

2.1.2 Redes sociais focadas em música

Uma das principais redes sociais direcionadas aos fãs de músicas é a Last.fm. Diversos trabalhos da literatura visam estudar o comportamento social dos usuários desta rede. Em [Pálovics & Benczúr, 2013] e [Moore et al., 2013] a dinâmica temporal dentro desta rede foi analisada. Em [Zhong et al., 2014], estruturas das comunidades, como conectividade, reciprocidade, clustering e influência foram analisadas e modeladas. Em [Levy & Bosteels, 2010] foi analisado como a popularidade dos conteúdos pode influenciar em sistemas de recomendação de músicas.

Nesta dissertação uma caracterização detalhada dos dados coletados entre Novembro de 2014 e Julho de 2015 através do Last.fm é apresentada (capítulo 5). Estes dados serão utilizados para a geração de um espaco de similaridade de músicas (capítulo 6), e são base da aplicação Mixtape.

2.2 Técnicas de definição de similaridade de conteúdos

Existem três estratégias principais para obter informação de similaridade: (1) análise de conteúdo de áudio [Knees et al., 2006; Logan, 2002; Neumayer et al., 2005; Pampalk et al., 2003, 2005a] (somente para o universo da música), (2) análise de metadados [Aucouturier & Pachet, 2002; Pampalk et al., 2005b; Platt, 2004; Platt et al., 2002] e (3) filtragem colaborativa [David Gleich & Lang, 2005; Ragno et al., 2005].

No contexto desta dissertação são utilizadas medidas de similaridades baseadas em filtragem colaborativa, que tipicamente exploram informações disponíveis ao público e, portanto, são mais escaláveis. Mais especificamente para o universo da música, em [Ragno et al., 2005], foi proposta uma medida de similaridade baseada em coocorrência de músicas apresentadas em transmissões de estações de rádio. Os valores finais de similaridade são calculados em tempo real diretamente de um grafo, que, em contraste com a abordagem apresentada neste trabalho de um espaço de similaridade de baixa dimensionalidade, não é uma técnica adequada para dispositivos de hardware limitados.

2.3 Técnicas de redução de dimensionalidade

Esta seção apresenta alguns trabalhos que têm como objetivo transformar dados de alta dimensionalidade em uma representação significativa de baixa dimensionalidade. Esta transformação é chamada de redução de dimensionalidade. Através de técnicas de redução de dimensionalidade os dados podem ser visualizados e analisados mais facilmente, já que a alta dimensionalidade possui diversos efeitos indesejáveis, como a maldição da dimensionalidade, dentre outros discutidos em [Jimenez & Langrebe, 1998]. Vale ressaltar que a lista apresentada não é exaustiva, e o objetivo principal é citar alguns dos métodos mais relevantes considerando as classes principais em que são divididos (lineares e não-lineares) e o tipo de entrada para a transformação dos dados (matriz de coordenadas, matriz de distâncias ou grafo).

As técnicas lineares têm como característica principal representar dados de alta dimensionalidade em um espaço de menor dimensionalidade assumindo que estes dados possuem uma relação linear. Já as técnicas não-lineares, também chamadas de aprendizado de variedades (*manifold learning*), assumem que esta relação entre os dados é não-linear. Uma variedade topológica ou *manifold* é um espaço topológico que localmente é similar a um espaço Euclidiano. As técnicas não-lineares buscam "aprender"

o *manifold* não-linear que represente os dados.

Algumas das técnicas lineares mais conhecidas são: Principal Component Analysis (PCA) [Shlens, 2005] e Multidimensional Scaling (MDS) [Cox & Cox, 2000]. Dentre as técnicas não-lineares, destacam-se: IsoMap [Tenenbaum et al., 2000], Locally Linear Embedding (LLE) [Roweis & Saul, 2000], Kernel PCA (KPCA) [Schölkopf et al., 1998], Laplacian Eigenmaps [Belkin & Niyogi, 2003], Large-scale Information Network Embedding (LINE) [Tang et al., 2015] e t-Distributed Stochastic Neighbor Embedding (t-SNE) [Van der Maaten & Hinton, 2008]. A seguir, apresentamos resumidamente cada uma destas técnicas.

Técnicas lineares: O PCA tem como entrada uma matriz de coordenadas e seu objetivo é encontrar projeções lineares para os dados tal que as dimensões maximizem a variância desses dados. Já o MDS clássico (cMDS) tem como entrada uma matriz de distâncias e seu objetivo é gerar a matriz de coordenadas que melhor preserva essas distâncias. Quando a matriz de entrada é uma matriz de distâncias Euclidianas o cMDS é equivalente ao PCA.

As técnicas lineares não são muito apropriadas para mapeamento de dados que se encontram em, ou próximos de, um *manifold* não-linear, já que elas assumem apenas uma correlação linear entre os dados. Entretanto dados de alta dimensionalidade se encontram frequentemente em *manifolds* não-lineares [Van der Maaten & Hinton, 2008; Shaw & Jebara, 2009; Tenenbaum et al., 2000].

Técnicas não-lineares: A técnica IsoMap é uma extensão do cMDS, onde os autores em [Tenenbaum et al., 2000] propõem a construção de um grafo a partir de uma matriz de coordenadas como entrada. Este grafo contém arestas apenas entre os k vizinhos mais próximos de cada ponto. A partir do grafo gerado a matriz de distâncias par-a-par completa é calculada e então esta matriz é utilizada como entrada para o método cMDS.

A técnica LLE [Roweis & Saul, 2000] também é baseada na construção de um grafo a partir de uma matriz de coordenadas inicial, porém ela busca preservar apenas as vizinhanças locais, ao contrário da IsoMap que foca em manter todas as distâncias globais geradas. Embora esta técnica utilize uma matriz de coordenadas como entrada, existe uma adaptação proposta por Roweis & Saul [2000] para serem utilizadas matrizes de distâncias.

Tanto a técnica IsoMap quanto a LLE, apesar de possuírem aproximações como L-ISOMAP [de Silva & Tenenbaum, 2004] e FastMap [Faloutsos & Lin, 1995], são pelo menos quadráticas em relação ao número de vértices, o que dificulta uma utilização em dados do mundo real com milhões de vértices [Tang et al., 2015].

A técnica t-SNE, proposta em [Van der Maaten & Hinton, 2008], busca preservar

as propriedades locais e globais dos dados e é uma variação da técnica Stochastic Neighbor Embedding. Seu foco principal é na geração de um espaço visualizável (2D/3D) e apresenta uma maior capacidade de lidar com dados do mundo real, envolvendo milhões de pontos. Esta técnica possui variações para utilizar como entrada matrizes de distâncias ou de coordenadas.

Dentre as técnicas que consideram um grafo como entrada podem ser destacadas na literatura mais recente LINE [Tang et al., 2015] e LargeVis [Tang et al., 2016], que possuem objetivos distintos. LargeVis, assim como t-SNE possui foco em gerar visualizações desses espaços de alta dimensionalidade. Os demais algoritmos podem ser usados como pré-processamento para essas técnicas de visualização [Tang et al., 2016]. A técnica LINE pode ser aplicada em diversos tipos de grafos: ponderados ou não, dirigidos ou não. Este método busca preservar as estruturas locais e globais da rede. Segundo os autores em [Tang et al., 2015], a técnica LINE é a mais eficiente para os casos onde o dado de entrada é um grafo.

2.4 Resumo do capítulo

Nesta dissertação iremos analisar a eficiência das seguintes técnicas de redução de dimensionalidade para a geração de espaços de similaridade para conteúdo de programas de TV e músicas: IsoMap/L-IsoMap, LLE e LINE. Estas técnicas foram selecionadas pelo fato de a IsoMap/L-IsoMap representar uma técnica com foco em preservar as distâncias entre todos os pontos do mapa e que assume como entrada uma matriz de distâncias. A técnica LLE é uma técnica com foco em preservar as vizinhanças locais entre os pontos, diferente da IsoMap o que pode ser um atributo interessante caso o espaço de similaridades gerado seja utilizado em uma navegação baseada nos vizinhos mais próximos. E por último, foi analisada a técnica LINE por ser uma técnica mais recente e que utiliza diretamente um grafo de similaridades como entrada, além de ter superado outras técnicas nas comparações em [Tang et al., 2015]. O capítulo 5 apresenta uma descrição mais detalhada destas técnicas.

Capítulo 3

Coleta e caracterização do tvtag

No capítulo 2 descrevemos alguns trabalhos encontrados na literatura que visam entender diversas redes sociais online. No entanto, trabalhos direcionados a entender o comportamento dos usuários em redes sociais online focadas em conteúdo de filmes e programas de TV são escassos. Desta forma, este capítulo apresenta uma ampla caracterização da rede social online tvtag [1]. Esta rede é formada por usuários que consomem conteúdo relacionado à programas de televisão e filmes. O principal objetivo é caracterizar o comportamento destes usuários, bem como o tipo de conteúdo discutido pelos mesmos. Estes dados serão utilizados para a construção de um espaço de similaridades específico para programas de TV e filmes (capítulo 6).

As seções deste capítulo buscam responder as seguintes questões:

- Como é distribuída a atividade dos usuários entre os conteúdos, outros usuários, gêneros de programas e filmes, e no tempo (seção 3.2)?
- Os usuários tendem a ser mais positivos ou negativos (seção 3.3)?
- O tvtag pode ser considerado um sistema em tempo real (seção 3.5)?
- Existe uma correlação entre a atividade dos usuários antes da estreia de um programa e a popularidade deste programa após (seção 3.6)?

[1] Durante o desenvolvimento desta dissertação a rede tvtag foi fechada e reconstruída como Telfie (`telfie.com`) e portanto não possível mais acessá-la. No entanto, os dados coletados foram capazes de forneces informações valiosas sobre o comportamento desses usuários.

3.1 Visão geral do tvtag

História: Lançado originalmente com o nome de GetGlue (2010), o tvtag (2013) é uma rede social online focada em entretenimento, onde os usuários podem fazer *check-in* em programas de TV e filmes, além de informar se gostam ou não desses programas através das funcionalidades *like* e *dislike*, respectivamente. Além de um aplicativo móvel, integrado com outras redes sociais, o tvtag tem parceria com produtores importantes de entretenimento, como 20th Century Fox, ESPN, HBO, Discovery Channel, Sony Pictures, e Warner Bros.

Sua base de usuários cresceu de 30.000 para aproximadamente 4,5 milhões entre 2010 e 2013. Entre *check-ins*, *likes* e revisões, o total de atividades desses usuários chegou a 500 milhões [Holanda et al., 2015b]. No início de 2015 a rede social tvtag foi fechada, sem informar aos usuários a razão, porém nesta mesma época o GetGlue foi adquirido por uma plataforma de TV social chamada Voice of TV e recriada como Telfie no final de 2015[2].

Redes de Interesse e Social: Duas estruturas principais podem ser destacadas no tvtag: um grafo de interesses de usuários e outro grafo social. O grafo de interesses representa as interações entre usuários e conteúdos: usuários que fazem *check-ins* nos mesmos *shows*, dão *likes* e *dislikes* ou deixam comentários. O grafo social é gerado através das interações entre usuários, em que eles podem seguir ou serem seguidos por outros usuários.

Interface do sistema: Após se conectar pela primeira vez ao *website* ou aplicativo móvel, o usuário é convidado a marcar uma lista de programas que gosta ou não para criar seu perfil. Após a criação do perfil o usuário pode seguir outros usuários e também fazer *check-ins* em programas de TV e filmes que tem interesse ou que está assistindo naquele momento. A Figura 3.1 exibe a tela onde o usuário pode pesquisar o programa ao qual está assistindo e fazer o *check-in* nele. Todo o conteúdo no tvtag, como filmes e artistas possuem sua própria página com uma lista de *check-ins* e *likes* de outros usuários, além de comentários.

3.2 Coleta e caracterização inicial

Fontes dos dados: Para a construção da base de dados de filmes e programas de TV foram utilizadas duas fontes de dados: a rede social online tvtag e uma base de dados aberta chamada TMDB (The Open Movie Database[3]). A partir do tvtag

[2] www.blog.getglue.com/getglue-becomes-telfie

[3] `www.themoviedb.org`

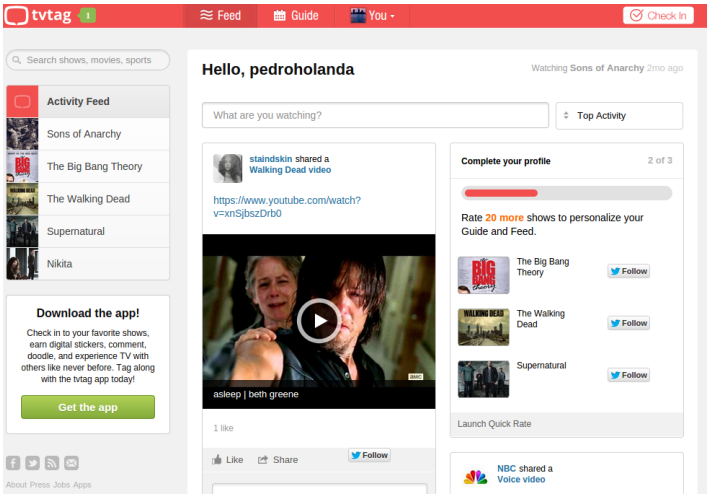

Figura 3.1: Interface do tvtag.

Tabela 3.1: Base de dados de filmes e programas de TV.

Usuários	1.745.000
Usuários c/ relações sociais	1.226.000
Filmes	9.300
Programas de TV	5.000
Gêneros (TMDB)	26
Itens (Filmes e Programas de TV) c/ gênero	9.977
Programas de TV c/ gênero	3.050
Check-ins	92.077.000
Likes	52.776.000
Dislikes	3.033.000

foram coletadas informações sobre preferências e atividades dos usuários em relação a conteúdos de TV e filmes, já do TMDB foram coletados os metadados, como data de estreia, atores, diretores e gêneros dos programas de TV e filmes (Holanda et al. [2015b,a]).

Coleta de dados: Para a coleta de dados, foram implementados dois *crawlers* Web. O primeiro *crawler* coletou as atividades dos usuários (*check-ins*, *likes* e *dislikes*) no tvtag. O conjunto de dados representa uma boa cobertura da rede do tvtag no período entre 2011 e 2012, e consiste no total de 29M *check-ins*, 21M *likes* e 1M *dislikes*. O segundo *crawler* coletou os metadados disponibilizados pelo TMDB. Foram coletados 100% dos dados de programas de TV e filmes disponíveis no TMDB. Após o cruzamento dos dois conjuntos de dados, foi possível obter tanto a atividade dos usuários quanto os metadados de aproximadamente 3.000 programas de TV e 7.000 filmes. As estatísticas relacionadas aos dados coletados estão descritas na Tabela 3.1.

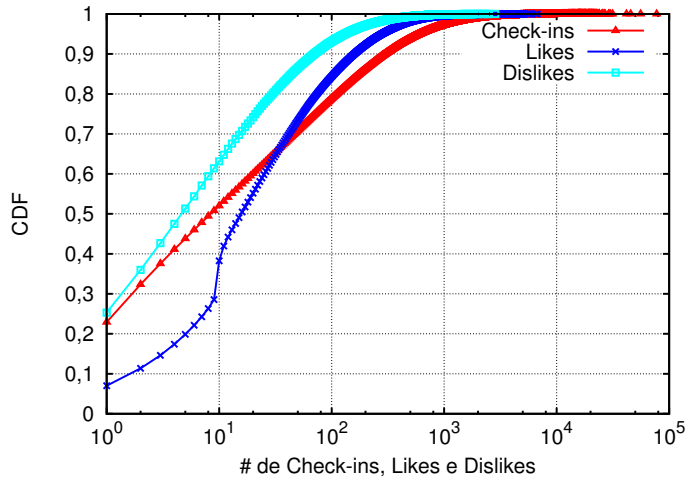

Figura 3.2: CDF por usuário (excluindo os usuários com 0 check-ins (1M), likes (900K) e dislikes (1,6M)).

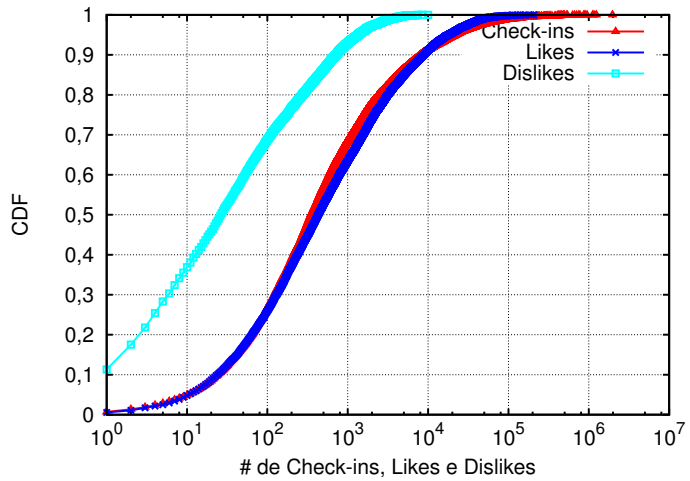

Figura 3.3: CDF por programa de TV (excluindo os programas com 0 check-ins (76), likes (6) e dislikes (910)).

Análise inicial: As Figuras 3.2 e 3.3 mostram a CDF (Função de Distribuição Acumulada) do total de *check-ins*, *likes* e *dislikes* realizados por cada usuário e em cada programa de TV. Para facilitar a visualização dos resultados, foram removidos dos gráficos os usuários e programas com zero *check-ins*, *likes* e *dislikes*. Observa-se que quase 52% dos usuários fizeram até 10 *check-ins*. Aproximadamente 50% dos programas receberam até 350 *check-ins*. 20% dos usuários são altamente engajados na rede social, realizando entre 100 e 1.000 *check-ins* no período coletado e 3% dos usuários realizaram mais de 1.000 *check-ins*.

Considerando o conjunto de dados utilizado, uma fração considerável de programas de TV (9%) recebe mais de 10.000 check-ins. As Tabelas 3.2, 3.3 e 3.4 mostram os

Tabela 3.2: Top 10 programas de TV em números de check-ins.

TV show	Check-ins	Likes	Dislikes
Bing Bang Theory	1.972.968	206.769	4.935
True Blood	1.238.715	143.888	5.253
Walking Dead	1.159.322	159.704	2.444
Supernatural	1.117.981	109.901	5.089
Glee	949.049	153.192	8.403
Fringe	903.700	89.435	4.510
Once Upon a Time	875.683	82.481	1.630
Vampire Diaries	865.218	94.770	5.964
Game of Thrones	776.313	99.821	1.693
Dexter	749.978	139.011	3.598
Pretty Little Liars	666.265	77.870	4.806

Tabela 3.3: Top 10 programas de TV em números de likes.

TV show	Likes	Check-ins	Dislikes
Big Bang Theory	206.769	1.972.968	4.935
Family Guy	206.458	459.463	6.428
Simpsons	192.276	378.779	4.449
House	180.631	405.951	3.380
Walking Dead	159.704	159.322	2.444
Glee	153.192	949.049	8.403
HIMYM	143.998	537.257	4.095
True Blood	143.888	1.238.715	5.253
South Park	139.220	153.792	5.216
Dexter	139.011	749.978	3.598

10 programas de TV com maior número de check-ins, likes e dislikes, respectivamente. Vale ressaltar que estes valores podem ser usados como uma estimativa razoável de audiência destes programas.

As Figuras 3.2 e 3.3 também mostram como as funcionalidades de likes/dislikes são utilizadas. Os usuários tendem a ser mais positivos do que negativos em relação a programas de TV. A distribuição de dislikes tem um maior decaimento quando comparada à distribuição de check-ins ou likes. Mais de 60% dos usuários atribuíram mais de 10 likes contra 35% de usuários que apresentaram a mesma quantidade de dislikes. Consequentemente, programas de TV recebem mais likes que dislikes: 75% deles receberam mais de 10 likes, 32% receberam mais de 100 dislikes. Estes resultados podem sugerir que os usuários do tvtag tendem a ser mais positivos em relação aos programas de TV nos quais expressam suas opiniões.

Tabela 3.4: Top 10 programas de TV em números de dislikes.

TV show	Dislikes	Check-ins	Likes
Two and Half Man	10.001	182.460	130.373
Glee	8.403	949.049	153.192
Jersey Shore	7.497	240.328	65.139
American Idol	74.62	414.212	84.919
Sex and City	7.379	33.000	51.010
Gossip Girl	7.211	417.960	70.649
CSI Miami	7.085	56.169	80.404
Greys Anatomy	7.039	532.299	112.149
Beavis Butt Head	6.513	38.523	54.770
Desperate Housewives	6.434	202.307	68.259

Tabela 3.5: Top 10 gêneros em número de programas de TV no TMDB e no tvtag.

TMDB		**tvtag**	
Gênero	% de programas	Gênero	% de programas
Drama	27.19	Comédia	39.11
Comédia	23.54	Drama	38.23
Documentário	14.65	Animação	18.20
Animação	14.02	Documentário	15.97
Ação/Aventura	6.23	Ação/Aventura	14.03
Sci-Fi/Fantasia	4.34	*Sci-Fi*/Fantasia	9.11
Notícias	3.43	Mistério	4.92
Mistério	1.73	Notícias	4.46
Família	0.91	Velho Oeste	1.34
Velho Oeste	0.82	Família	1.28

3.3 Caracterização de gêneros

Distribuição de gêneros: Os programas de TV com dados disponibilizados no TMDB (3.050) foram categorizados em 26 gêneros diferentes. 63% destes programas foram classificados com um gênero; 24% com 2 gêneros e 13% com 3 ou mais gêneros. A Tabela 3.5 compara os 10 gêneros com maior número de programas de TV disponíveis no TMDB e no tvtag. Os 10 primeiros gêneros nas duas listas são os mesmos, apesar de a ordem e a porcentagem de programas serem diferentes. Existe uma maior discrepância em relação aos gêneros de Animação, Ficção Científica(*Sci-Fi*)/Fantasia e Ação/Aventura.

Popularidade de gêneros: Uma maneira simples de medir a popularidade de um programa de TV é contar o número de check-ins e likes que eles receberam através

de uma Rede Social Online. Esta medida é muito utilizada na literatura para verificar popularidade [Vasconcelos et al., 2014]. A Figura 3.4 mostra como a popularidade entre os gêneros muda de acordo com o data de coleta. O número de check-ins para todos os gêneros tem picos durante os meses de estreia de temporadas nos EUA (Setembro/Outubro) e após as férias de início de ano (Fevereiro). A maior quantidade de check-ins ocorre, aproximadamente, durante o período de férias (Novembro/Dezembro e Junho/Julho). Drama e Comédia são os gêneros mais populares dentre os dados coletados.

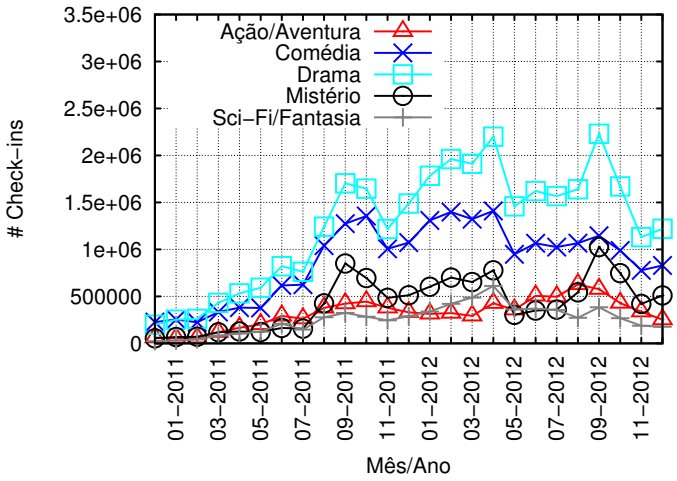

Figura 3.4: Evolução temporal do número de check-ins.

Positividade dos gêneros: Para verificar quais os gêneros que mais agradam os usuários do tvtag foram utilizadas duas medidas de positividade: (1) Positividade do usuário, definida como a média de likes que um usuário dá aos programas de um gênero, e (2) Positividade do programa, definida como a relação entre o número de likes e dislikes dentre os programas de um gênero.

A Figura 3.5 representa a quantidade total de likes dividida pelo total de usuários únicos que deram check-ins por gênero, considerando todos os 26 gêneros dos programas de TV e filmes. Os gêneros mais populares são os que tem maior número de likes em relação a usuários, como Comédia, Drama e Animação.

A seção 3.2 apresenta uma discussão sobre as preferências dos usuários em relação aos filmes e programas de TV: os resultados mostram que o número de likes fornecidos pelos usuários é maior do que o número de dislikes em toda a amostra, revelando que os usuários tendem a ser mais positivos em suas opiniões. A Figura 3.6 corrobora esta informação. A mesma mostra a razão entre o total de likes e dislikes que os programas receberam por gênero, ao longo do tempo. Apesar de gêneros mais populares, como

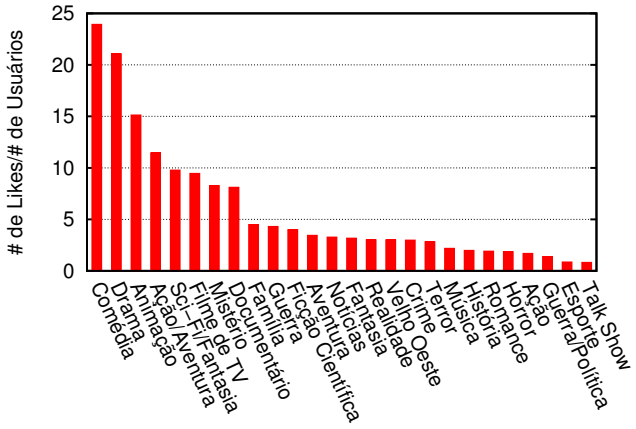

Figura 3.5: Positividade do usuário por gênero.

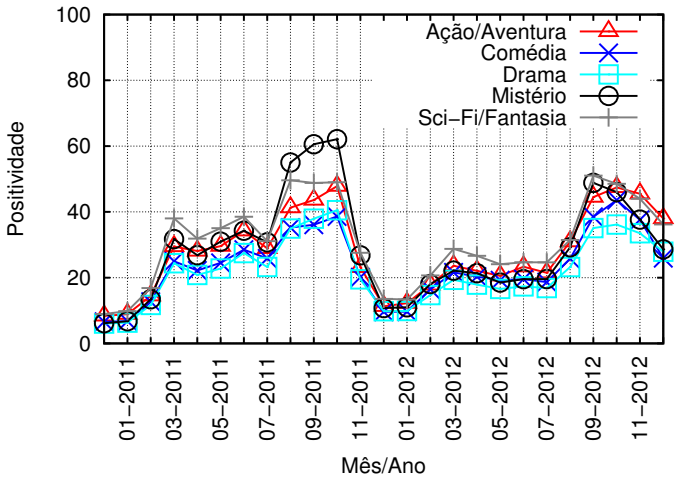

Figura 3.6: Positividade do programa por gênero ao longo do tempo (razão entre like-dislike).

Drama e Comédia, receberem maior quantidade de likes por usuário, estes gêneros também atraem mais dislikes. O gênero Mistério, embora menos popular, tem uma maior positividade principalmente entre Agosto e Outubro de 2011.

3.4 Vínculos sociais

A rede social dos usuários do tvtag, como no Twitter[4], é formada por seguidores: um usuário pode seguir ou ser seguido por outros usuários. Foram coletados os seguidores (*followers*) e os seguidos (*followee*) de aproximadamente 1.2M de usuários. A Figura 3.7

[4]http:twitter.com

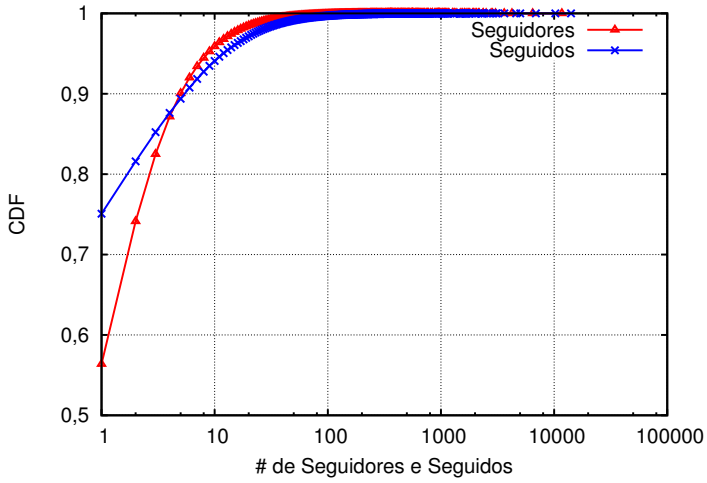

Figura 3.7: CDF dos seguidores e seguidos por usuário.

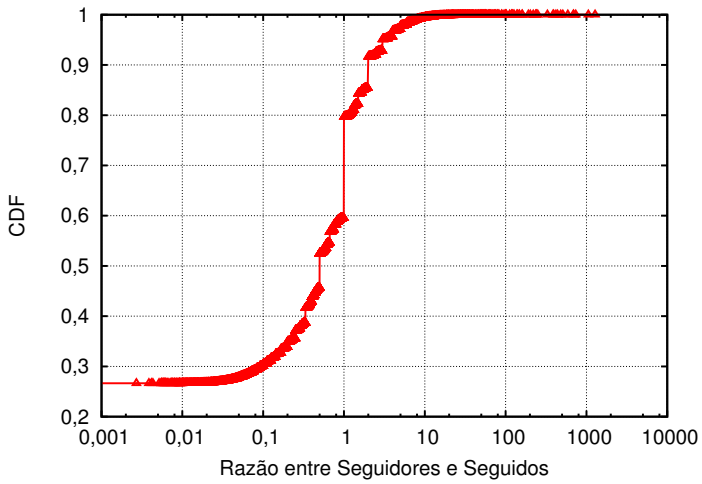

Figura 3.8: CDF da razão entre seguidores e seguidos por usuário.

representa a CDF do número de seguidores e seguidos por usuário. Uma pequena porção dos usuários(<1%) tem mais de 100 seguidores, enquanto a grande maioria (95%) tem menos de 10 seguidores. Apenas 1% dos usuários seguem 45 ou mais usuários e 94% seguem no máximo 10 usuários.

A CDF da razão entre seguidores/seguidos por usuários pode ser observada na Figura 3.8. Aproximadamente 60% dos usuários seguem mais pessoas do que são seguidas, e apenas 20% dos usuários tem mais seguidores do que seguem. Este é um resultado esperado, dado que poucos usuários atraem muitos seguidores (semelhante ao que ocorre no Twitter e em redes sociais online de amizades).

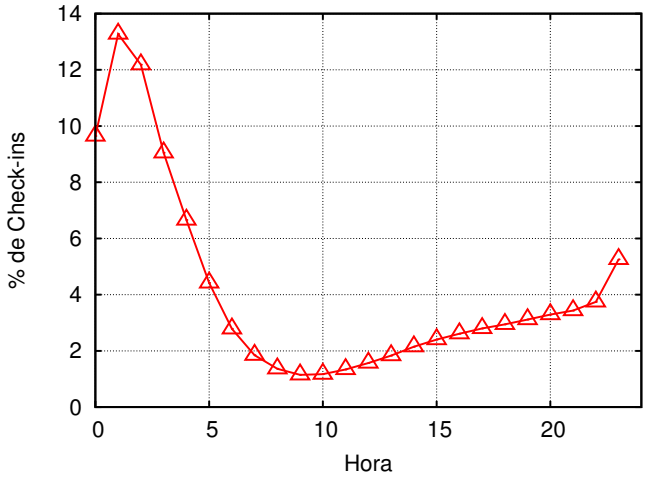

Figura 3.9: Distribuição de check-ins ao longo do dia (UTC).

3.5 Comportamento *second-screen*

Esta seção verifica se o tvtag pode ser visto como um sistema de tempo real (ou seja, os usuários fazem *check-ins* enquanto os programas estão sendo transmitidos), ou se as pessoas tendem a dar check-in em programas de TV independentemente do horário que cada programa vai ao ar na televisão.

O tvtag fornece a data e hora de check-in, enquanto o TMDB fornece a data e hora que o último episódio de um programa foi transmitido. Os horários de check-in fornecidos pelo tvtag são armazenados no servidor de acordo com o horário do servidor em UTC, e não de acordo com o horário local do usuário. A Figura 3.9 exibe a distribuição de todos os check-ins em programas de TV ao longo do dia, considerando o fuso horário UTC. É possível notar que as maiores concentrações de check-ins estão entre 23:00 e 5:00 UTC, o que corresponde aproximadamente ao horário nobre de televisão entre os fusos horários dos Estados Unidos *Pacific Standard Time* (UTC-8) e *Eastern Standard Time* (UTC-5).

Para verificar se os usuários aumentam suas atividades no dia que um programa vai ao ar, a Figura 3.10 apresenta os dias da semana em que esses check-ins ocorrem. Foram considerados os 5% e 10% programas de TV mais populares em número de check-ins[5]. Como os horários disponíveis na amostra estão apenas em UTC, foram subtraídas 9 horas de cada check-in, pois os check-ins poderiam estar até 9 horas adiantados em relação ao horário nobre. Assim, a Figura 3.10 apresenta o dia em que o programa de TV foi ao ar pela última vez e os dias da semana precedentes e posteriores à exibição

[5]Os 5% e 10% da amostra contém mais de 700 e 1400 programas, respectivamente, ambos pertencentes a uma variedade de gêneros.

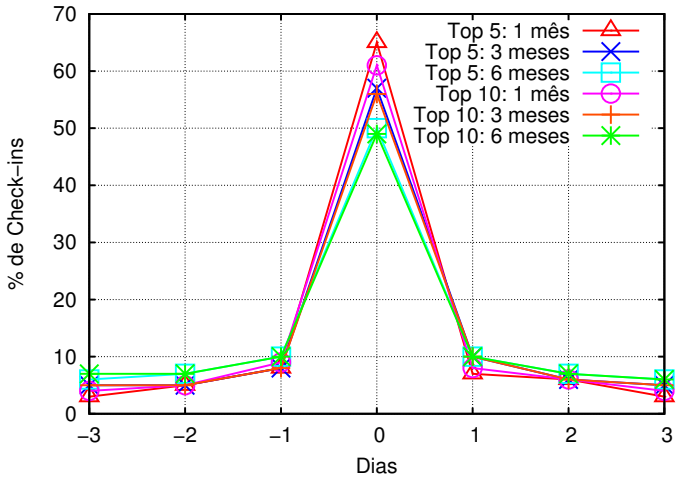

Figura 3.10: Comportamento *second-screen* nos 5% e 10% programas de TV mais populares: porcentagem de check-ins durante o mesmo dia da semana em que o último episódio foi ao ar considerando os 1, 3 e 6 meses precedentes.

desse programa. O eixo X é 0 quando o dia do check-in coincide com o dia da semana em que esse episódio foi ao ar e positivo ou negativo, quando o check-in foi após ou anterior ao dia do episódio, respectivamente. Em todas as análises mais de 50% dos check-ins ocorreram durante o dia em que o último episódio do programa foi ao ar. A porcentagem de check-ins em dias diferentes do dia 0 é $\leq 10\%$.

Claramente é possível notar que os picos de uso do sistema são em sua maior parte durante os dias da semana em que os programas são transmitidos. Adicionalmente, estes *check-ins* ocorrem aproximadamente no horário nobre, indicando que o tvtag comporta-se como uma aplicação em tempo real, ou *second-screen*, principalmente considerando os programas de maior popularidade.

3.6 Previsão da audiência inicial

A popularidade é um dos principais focos de programas de TV. Produtores, diretores e atores esperam conquistar grande audiência desde a estreia de seus programas. Aumentar a audiência indica maior possibilidade do aumento de arrecadação financeira e, consequentemente, maior chances de ter sequência deste programa aprovada. A possibilidade de prever a audiência inicial de um programa através de modelos simples é uma importante ferramenta para estratégias de marketing, dado que a popularidade dos programas atrai taxas de anúncio mais caras, gerando, possivelmente, o aumento as vendas dos produtos anunciados durante a exibição do programa.

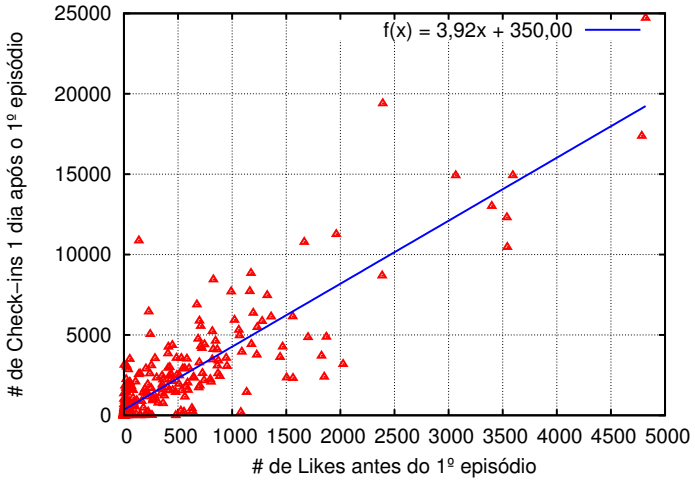

Figura 3.11: Correlação entre a quantidade de likes antes e o número de check-ins *um dia* após a estreia.

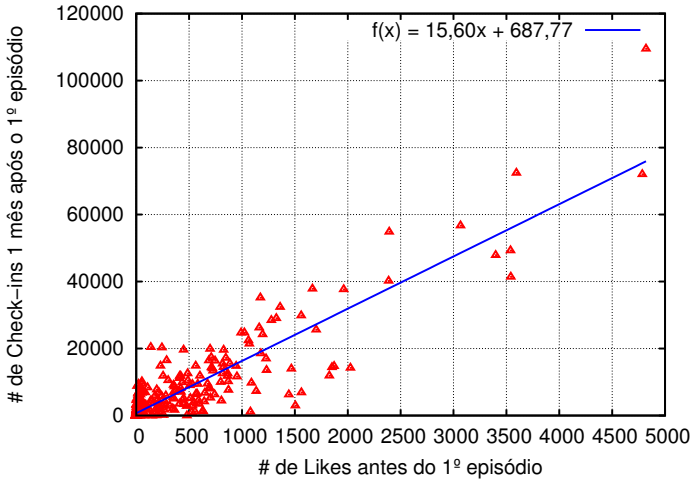

Figura 3.12: Correlação entre a quantidade de likes antes e o número de check-ins *um mês* após a estreia.

Uma metodologia que pode ser explorada para prever a audiência inicial de um programa de TV é analisar as interações de usuários de uma rede social antes da estreia deste programa. Assim, é possível verificar se existe engajamento e interesse das pessoas (que podem ser medidos através de comentários e atividades de likes e dislikes) antes da estréia dos programas.

Para realizar esta análise, foi considerado um subconjunto de 525 programas de TV que foram lançados entre os anos 2011 e 2012. As Figuras 3.11 e 3.12 mostram a correlação entre a quantidade de likes recebidas antes da estreia desses programas e a quantidade de check-ins após um dia e após um mês de exibição, respectivamente.

Através de uma regressão linear é possível observar o impacto do número de likes antes das estreias de um programa e sua consequência na popularidade após a estreia. Foram obtidos coeficientes de correlação de Pearson $\rho = 0.8750$ e $\rho = 0.9040$ para as regressões considerando os intervalos de tempo de um dia e um mês, respectivamente. Portanto estes resultados evidenciam uma alta correlação linear entre os likes antes de uma estreia e os check-ins realizados após esta estreia.

Para implementar o método de previsão, os dados foram divididos em dois subconjuntos: o conjunto de treino Υ, para calcular os parâmetros e o conjunto de testes Ω, para analisar a acurácia do modelo gerado. As amostras foram selecionadas aleatoriamente e os parâmetros lineares foram calculados para cada um dos 1000 pares de treinamento e teste gerados. As Tabelas 3.6 e 3.7 mostram a (coeficiente angular), b (coeficiente linear) e o coeficiente de determinação R^2, com intervalos de confiança de 95% gerado através das 1000 combinações aleatórias entre os conjuntos de treinamento e teste. Para todos os casos, R^2 tem valores médios entre 0.71 e 0.79. Como esperado, se forem utilizados mais programas de televisão no treinamento, o intervalo de confiança é menor.

Como um exemplo de aplicação do método de previsão, foi utilizado o seriado Gotham, que foi lançado em 2014 (aproximadamente 2 anos após as amostras de treinamento do modelo). Foram coletados 317 likes antes da estreia do programa e seus check-ins um dia $(1,934)$ e um mês $(7,901)$ após a data da primeira exibição. Através deste modelo simples, utilizando 50% do conjunto de treinamento, foram obtidos os valores de $1,591$ check-ins um dia após a estreia (18% de erro) e $5,623$ um mês após a estreia (29% de erro). Este erro de subestimação pode ser causado pela mudança do público do tvtag durante 2 anos.

O modelo gerado, embora simples, apresenta um primeiro passo na melhor compreensão de como os dados obtidos a partir de uma rede social online focada em programas de televisão podem ser utilizados para promover estes programas e aumentar sua audiência futura. Demonstrando ser útil o investimento das emissoras de TV em publicidade pré-lançamento através de redes sociais.

3.7 Resumo do capítulo

Neste capítulo apresentamos uma caracterização detalhada da rede social online tvtag. Estes resultados foram publicados em [Holanda et al., 2015b,a]. Como principais conclusões, podemos citar que a rede tvtag pode ser classificada como uma aplicação *second-screen*, onde os usuários podem fazer comentários paralelamente à exibição de

Tabela 3.6: Previsão da audiência inicial um dia após a estreia de um programa utilizando o número de likes antes da estreia.

Υ	Ω	**Um dia**		
		a	b	R^2
10%	90%	3.8500 ± 0.0435	368.2821 ± 10.085	0.7111 ± 0.0050
20%	80%	3.8801 ± 0.0148	352.7846 ± 3.5794	0.7342 ± 0.0018
30%	70%	3.9073 ± 0.0074	351.7603 ± 1.8233	0.7449 ± 0.0009
40%	60%	3.8999 ± 0.0042	351.9690 ± 1.0880	0.7480 ± 0.0006
50%	50%	3.9013 ± 0.0027	354.8540 ± 0.6946	0.7493 ± 0.0006
60%	40%	3.9083 ± 0.0018	352.7203 ± 0.4857	0.7469 ± 0.0005

Tabela 3.7: Previsão da audiência inicial um mês após a estreia de um programa utilizando o número de likes antes da estreia.

Υ	Ω	**Um mês**		
		a	b	R^2
10%	90%	24.1259 ± 4.3691	805.1159 ± 172.169	0.7489 ± 0.0345
20%	80%	15.1780 ± 0.5046	767.7531 ± 91.551	0.7735 ± 0.0153
30%	70%	15.4168 ± 0.3132	721.7714 ± 59.980	0.7894 ± 0.0082
40%	60%	15.4826 ± 0.2199	706.7614 ± 41.831	0.7938 ± 0.0064
50%	50%	15.5068 ± 0.1647	707.9705 ± 31.509	0.7973 ± 0.0055
60%	40%	15.5516 ± 0.1261	698.3868 ± 23.946	0.7982 ± 0.0052

programas de TV. Além da caracterização das atividades dos usuários (total de likes, dislikes, positividade em relação ao conteúdo discutido na rede), mostramos que estas atividades, quando realizadas antes da estréia de um programa de TV, são bons indicadores de engajamento inicial dos usuários em relação ao conteúdo. Propomos um modelo de regressão linear simples que prevê o engajamento dos usuários no primeiro dia e no primeiro mês após a estréia de um programa, tendo como base as atividades feitas pelos usuários antes da estréia do mesmo.

Capítulo 4

Coleta e caracterização do Last.FM

Similarmente à caracterização da rede social online tvtag, apresentada no capítulo 3, este capítulo apresenta a caracterização dos dados coletados a partir do Last.Fm[1], uma rede social online para fãs de música. Caracterizações desta rede são encontradas em diversos trabalhos na literatura. O nosso principal objetivo é analisar os dados que serão utilizados na definição de um espaço de similaridades cujos itens são músicas lançadas por diversos artistas.

A caracterização apresentada neste capítulo responde as seguintes questões:

- Qual é o perfil dos usuários presentes na nossa coleta (gênero, idade, localização geográfica) (seção 4.2)?
- Como a atividade dos usuários é dividida entre os conteúdos(seção 4.2)?
- Qual é a distribuição dos tipos de tags associadas as músicas(seção 4.4)?

4.1 Conjunto de dados

A coleta de dados do Last.fm pode ser feita através da API pública disponibilizada pelo site. Desta forma, informações sobre músicas, artistas, albuns e tags compartilhadas por milhões de usuários podem ser caracterizadas e utilizadas para entender melhor o comportamento das pessoas que são fãs do universo da música.

Os dados foram coletados em duas etapas: Primeiramente coletamos 0.28% das músicas que estão armazenadas no Last.fm (aproximadamente 100 mil canções) e os usuários que mais escutaram estas músicas. A partir deste conjunto inicial, passamos

[1] `www.last.fm`

a coletar as músicas mais escutadas por estes usuários, bem como a lista de amigos dos mesmos. As top-25 músicas mais escutadas por cada usuário foram coletadas.

A coleta ocorreu entre os meses de Novembro de 2014 e Julho de 2015, e possui 382.077 usuários (com as suas respectivas top-25 músicas mais escutadas), 2.224.227 canções e 374.402 artistas. Também foram coletadas 1.006.236 tags geradas pelos usuários e associadas as músicas. 47% das músicas tem pelo menos uma tag associada a ela.

Das 2.224.227 músicas pertencentes ao conjunto de dados, 983.010 possuem *MusicBrainz Identifiers* (MBID)[2]. Essa fonte de dados é importante para o aumento da confibilidade dos dados que serão utilizados na geração do espaço de similaridades, garantindo entradas únicas, evitando, por exemplo, erros de atribuição entre músicas e artistas e erros de digitação por parte dos usuários.

4.2 Caracterização dos usuários

O conjunto de dados possui a informação de 382.077 usuários originados de 237 países. A Tabela 4.1 apresenta o total de usuários divididos nos principais países representados nos dados coletados.

Países	Total de Usuários
US	54.248
BR	31.230
PL	25.882
RU	22.979
UK	22.515
DE	18.314
NL	7.237
CA	6.931
FI	6.356
FR	6.139

Tabela 4.1: Países com o maior número de usuários.

Muitos dos usuários não declaram o gênero. Dos usuários que declaram o gênero, 201.096 se declararam homens e 108.034 mulheres. A Tabela 4.2 mostra o número médio de músicas executadas por gênero. Para as mulheres, este número é maior do que para os homens. Usuários que não declaram o gênero possuem o menor número médio de músicas executadas.

[2]MBID é uma fonte confiável de dados que identifica unicamente uma música (`musicbrainz.org`).

Gênero	Média de Execuções
Mulher	19,9839
Homem	19,6071
Não informado	18,5338

Tabela 4.2: Média de execuções de músicas por gênero.

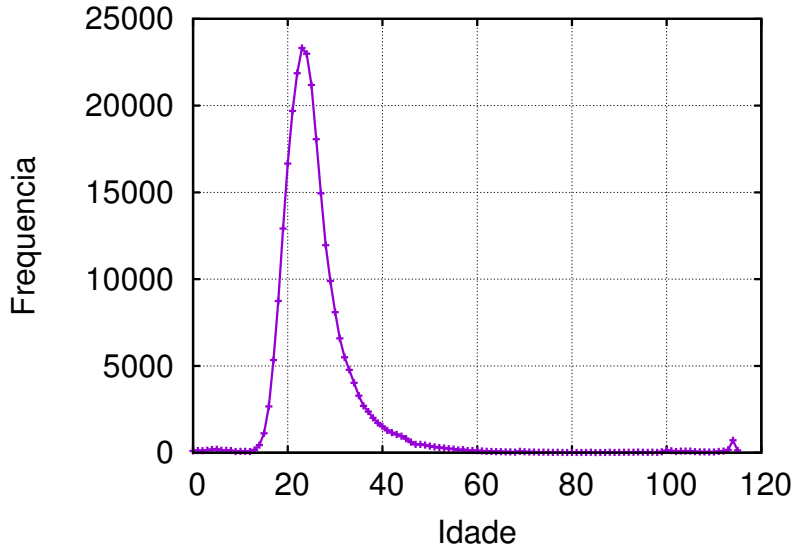

Figura 4.1: Last.fm: Distribuição da idade dos usuários.

A Figura 4.1 mostra a distribuição das idades dos usuários. A maior parte dos usuários estão entre os 18 e 30 anos. Aproximadamente 113.000 usuários não reportaram a idade.

A Figura 4.2 mostra que aproximadamente 27% dos usuários escutaram menos que 10.000 canções, enquanto 62% dos usuários escutaram entre 10.000 e 100.000 músicas. Usuários tendem a escutar uma grande quantidade de músicas, armazenando um histórico grande de músicas tocadas, mostrando o engajamento com a rede social.

A Figura 4.3 mostra o CDF dos amigos dos usuários. Aproximadamente 5% dos usuários não tem amigos no Last.fm, e 50% dos usuários tem no máximo 25 amigos. Como foram coletados somente os 50 amigos de cada usuários, a maior quantidade de amigos é de 50, sendo que 30% dos usuários possui pelo menos 50 amigos.

4.3 Músicas e artistas

Nesta seção apresentamos a caracterização das músicas que possuem o *MusicBrainz ID*, evitando problemas causados por erros feitos pelos usuários (por exemplo, erros de

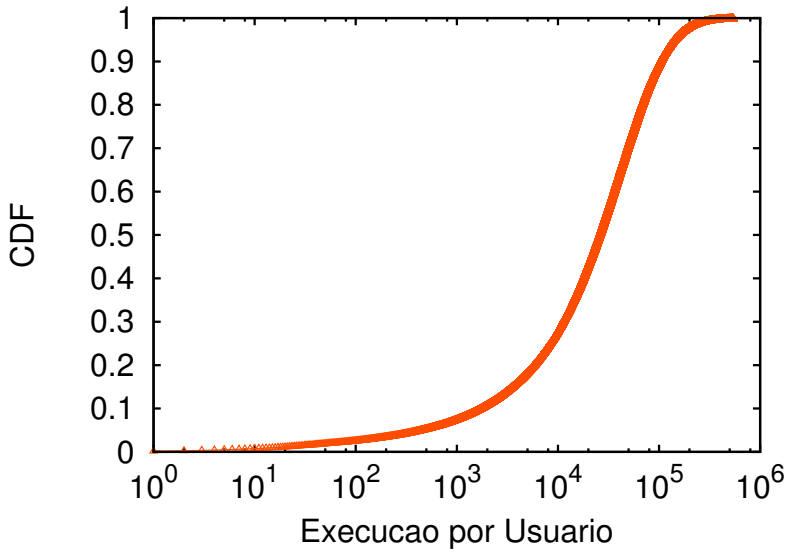

Figura 4.2: Last.fm: CDF do número de execução das músicas por usuário.

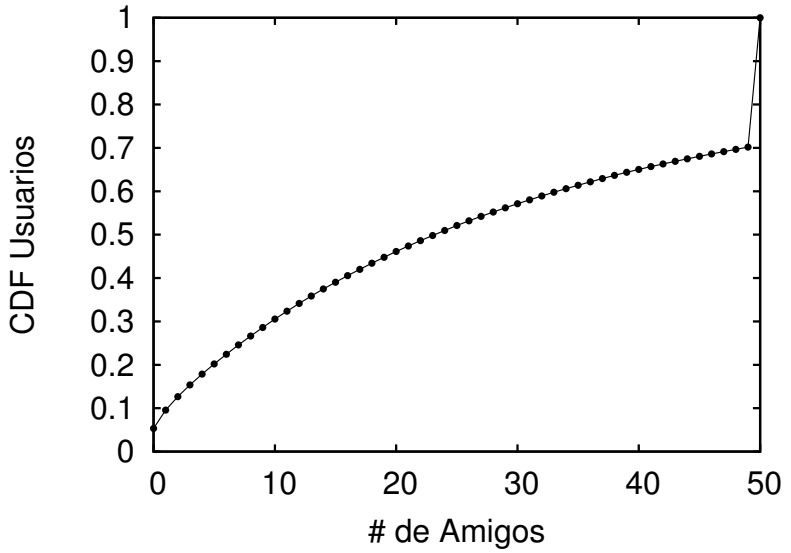

Figura 4.3: Last.fm: CDF do total de amigos dos usuários.

digitação).

As Figuras 4.4 e 4.5 mostram as CDFs da popularidade das músicas, considerando o total de usuários que as escutaram e o total de vezes que foram executadas. Podemos observar que, no nosso conjunto de dados, 50% das canções foram escutadas por no máximo 1.280 usuários únicos (0.3%); 10% atraíram a atenção de até 22.685 usuários únicos (5.9%). As músicas mais populares (top 1%) foram escutadas por mais de 155.468 usuáros (41%). As top 10 músicas considerando o número de usuários que

escutaram estão listadas na Tabela 4.3, cada uma delas escutadas por mais de 1.4M usuários.

Na Figura 4.5 observamos que 10% das trilhas sonoras foram executadas 1.000 vezes ou menos e mais de 50% foram executadas acima de 10.000 vezes.

Música	Artista	Número de usuários
Smells Like Teen Spirit	Nirvana	1.806.180
Mr. Brightside	The Killers	1.716.969
Wonderwall	Oasis	1.685.703
Come as You Are	Nirvana	1.597.611
Clocks	Coldplay	1.507.981
Somebody Told Me	The Killers	1.490.787
Take Me Out	Franz Ferdinand	1.462.621
Karma Police	Radiohead	1.431.055
Viva la Vida	Coldplay	1.431.034
The Scientist	Coldplay	1.404.877

Tabela 4.3: Top 10 músicas.

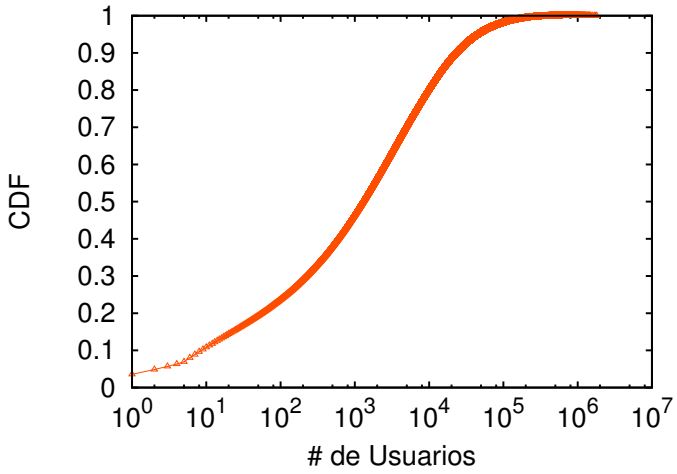

Figura 4.4: Last.fm: CDF da popularidade das músicas por número de usuários.

A Figura 4.6 mostra a CDF da popularidade dos artistas, considerando o total de usuários que escutaram uma de suas músicas. Os artistas mais populares (top 1% $\approx$ 3.577 artistas) foram escutados por mais de 764.000 usuários, enquanto aproximadamente 20% dos artistas foram escutados por somente 10 usuários.

Figura 4.7 mostra a CDF da popularidade dos artistas considerando os usuários que contém este artista em uma das primeiras 25 posições de sua lista de preferência. Mais de 35% dos artistas aparecem somente em uma das listas destes usuários, enquanto 1% aparece nas listas de mais de 600 usuários.

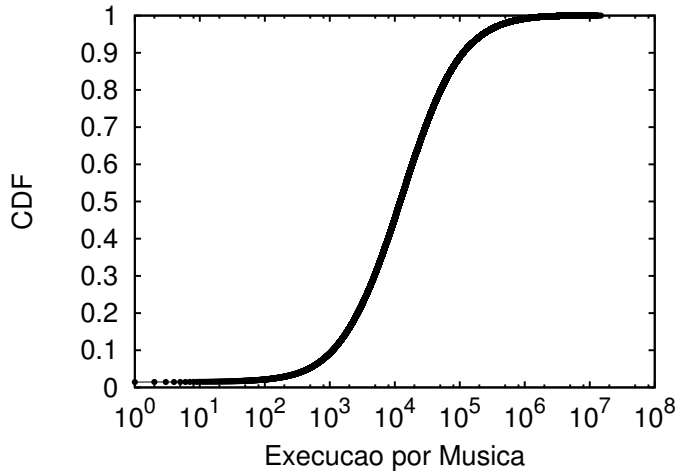

Figura 4.5: Last.fm: CDF da popularidade das músicas por total de execuções.

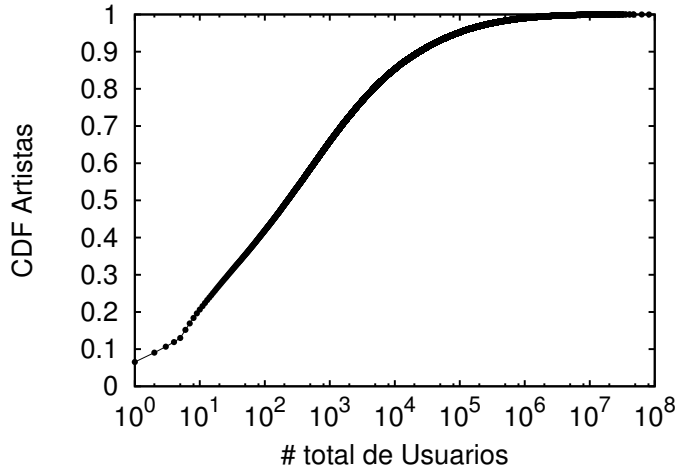

Figura 4.6: Last.fm: CDF da popularidade de artistas pelo número total de ouvintes.

A Figura 4.8 mostra a CDF da coocorrência entre as músicas (eliminando os casos de coocorrência igual a zero), considerando as listas de preferências com as top-25 músicas de cada usuário. Mais de 90% as coocorrências acontecem em somente uma vez, enquanto 1% coocorre 10 vezes ou mais. Se considerarmos todos os possíveis pares de músicas, a grande maioria não coocorre nas listas de preferências dos usuários. Assim, somente uma pequena porcentagem de similaridade poder ser obtida dos dados, sendo a grande maioria inferida pelas distâncias geradas pelos algoritmos de *embedding*.

A Figura 4.9 mostra a CDF de coocorrência entre artistas (eliminando os casos de coocorrência igual a zero). 75% dos artistas coocorrem somente uma vez, equanto aproximadamente 1% coocorre na lista de preferências de 22 usuários.

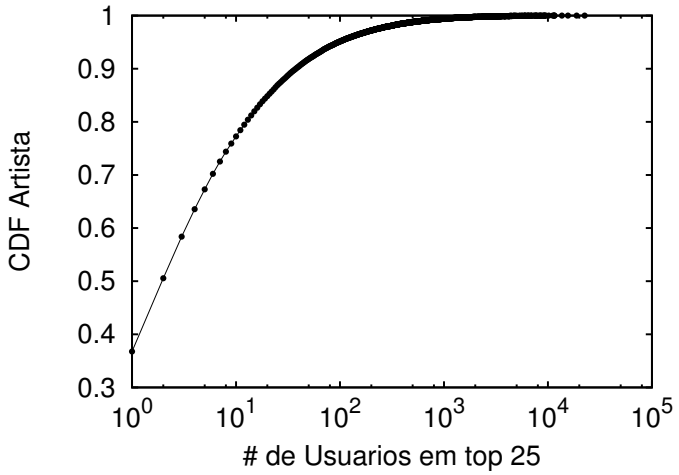

Figura 4.7: Last.fm: CDF da popularidade de artistas por top 25 de usuários.

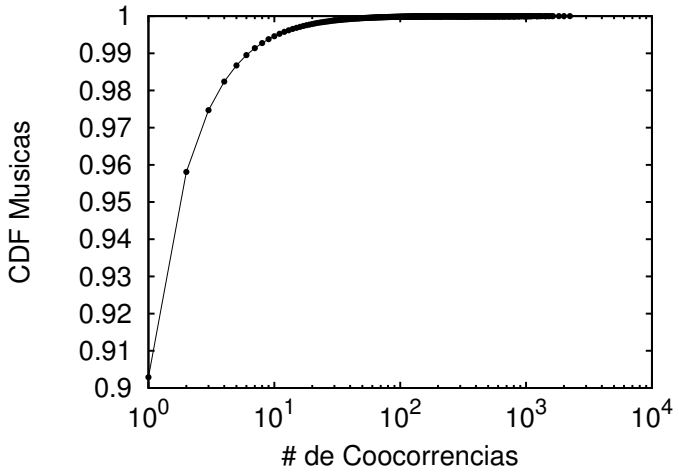

Figura 4.8: Last.fm: CDF da coocorrência de músicas.

4.4 Tags

Last.fm permite aos usuários criar e associar tags às músicas que são executadas. Uma determinada tag pode ser associada ate 100 vezes a uma música, permitindo ao usuário mostrar o nível de concordância de descrição da música pela tag. Coletamos aproximadamente 1M de tags geradas pelos usuários. 47% das canções tem pelo menos uma tag associada. Considerando somente músicas que possuem o MusicBrainz ID, 75% das músicas possuem pelo menos uma tag associada.

As top-5 mais populares tags (rock, alternative, pop, indie and electronic) foram associadas a 541,527 músicas. A Tabela 4.4 mostra as tags mais populares nos dados

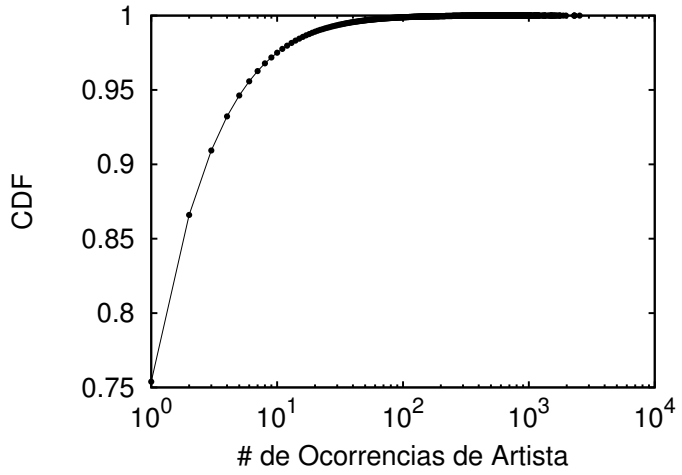

Figura 4.9: Last.fm: CDF da coocorrência de artistas.

do Last.fm e o total de músicas que tiveram as mesmas associadas.

Tag	Total de Músicas
rock	167.610
alternative	101.061
pop	96.654
indie	88.786
electronic	87.416
alternative rock	56.643
favorites	56.508
beautiful	51.870
love	50.918
awesome	42.364

Tabela 4.4: Top 10 tags

A Figura 4.10 mostra a CDF da popularidade das tags: 62% das tags foram associadas somente com uma música. As top-5 tags foram associadas com mais de 87.000 músicas cada.

4.5 Resumo do capítulo

A caracterização apresentada neste capítulo mostra detalhes da rede social online Last.fm, considerando os dados coletados entre os meses de Novembro de 2014 e Julho de 2015. Em particular, mostramos que os usuários são jovens, e na grande maioria localizados nos Estados Unidos. Adicionalmente, existe uma grande diferença no número de usuários que escutam músicas populares e aqueles que optam por músicas mais alternativas. Assim, a popularidade deve ser considerada para determinar similaridade

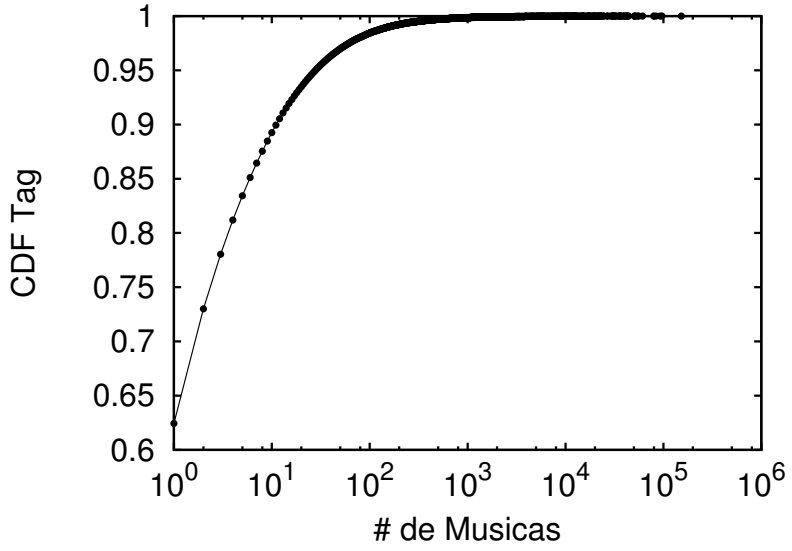

Figura 4.10: Last.fm: CDF da popularidade de tags.

entre músicas. Os usuários também exploram a idéia de tags para descrever as músicas. Esta informação é extremamente valiosa para definir similaridade entre itens, com o intuito de prover espaços de similaridades que podem ser usados para recomendação de novos itens ou melhor navegabilidade entre estes itens.

Capítulo 5

Geração dos espaços de similaridades de conteúdos de mídia

Este capítulo descreve os três pontos principais a serem considerados na geração de um espaço de similaridade de conteúdo de mídia e como os mesmos foram tratados nesta dissertação: (1) como inferir e medir similaridade entre conteúdos (seção 5.1); (2) como representar matematicamente a relação entre os conteúdos e as suas similaridades (seção 5.2) e; (3) como transformar a representação matemática em um espaço métrico, capaz de medir similaridade de pares de itens com custo computacional reduzido (seção 5.3).

5.1 Similaridade entre itens

Conforme descrito no capítulo 2, existem diversos métodos na literatura para a definição de similaridade entre um par de itens (filmes, programas de TV, músicas, livros, etc), baseados em dois princípios: análise do conteúdo e filtragem colaborativa. Nesta dissertação, a similaridade entre pares de itens é definida através de filtragem colaborativa, utilizando os dados coletados de redes sociais online de fãs de TV e de música.

Após definir o método de captura de similaridade entre conteúdos de mídia, o próximo passo é quantificar a força da similaridade entre pares de conteúdos (itens). Considere dois itens A e B que ocorrem em listas de preferências dos usuários destas redes, como, por exemplo, lista das músicas mais escutadas, lista dos programas de TV com maior número de *likes*, etc.

Coocorrência ($A \cap B$) é a ocorrência de dois itens na lista de preferências de um mesmo usuário. Portanto, $|A \cap B|$ é a quantidade de vezes em que os itens A e B ocorreram nas listas de diferentes usuários e $|A|$ a quantidade de vezes em que o item

A ocorreu nessas listas. Utilizando os dados coletados de redes sociais online, o nível de similaridade dos itens pode ser medido através de vários índices, como:

1. *Matching Coefficient:* Definido como $|A \cap B|$, esta métrica considera o total de vezes que os itens A e B coocorreram na lista de preferências do usuário, sem considerar total de ocorrências individuais de A e B.

2. *Dice Coefficient*: Este coeficiente normaliza o número de coocorrência de A e B pelo total de ocorrência individuais destes itens:

$$\frac{2|A \cap B|}{|A| + |B|}.$$

3. Jaccard: O coeficiente de Jaccard mede a similaridade entre dois conjuntos e é definido por:

$$\frac{|A \cap B|}{|A \cup B|}$$

4. Cosseno: Cosseno mede a similaridade entre dois vetores, expressando o cosseno do ângulo entre estes vetores:

$$\frac{|A \cap B|}{\sqrt{|A| \times |B|}}.$$

Neste contexto, a similaridade entre itens é calculada utilizando a similaridade cosseno, já que esta é uma medida normalizada e frequentemente utilizada em filtragem colaborativa. Como o vetor de ocorrências é um vetor positivo, o valor do cosseno está no intervalo $[0 : 1]$.

5.2 Grafo de similaridades e de distâncias

Após o cálculo das similaridades entre cada par de entidades (músicas ou filmes e programas de TV), definido na seção 5.1, é obtido um conjunto esparso de relações entre essas entidades e, portanto, só é possível analisar a similaridade entre itens que coocorreram na base de dados utilizada.

Uma estratégia para calcular a relação entre itens que não coocorreram na base de dados é através da construção de um grafo $G(V, E)$ que contém as relações E entre o conjunto de entidades V e utilizar este grafo para prever similaridades que não tem arestas diretamente.

A técnica de *embedding* de grafos LINE pode utilizar diretamente um grafo de similaridades G_{Sim} e convertê-lo em um espaço de coordenadas, com este espaço a similaridade entre dois itens é calculada através da distância Euclidiana entre dois pontos.

Outra forma de obter a similaridade entre estas duas entidades é converter o grafo G_{Sim} em um grafo de distâncias G_{Dis}, assim a similaridade entre itens é comparada através da soma do peso das arestas no menor caminho entre estas entidades.

Entretanto, muitas aplicações exigem que este cálculo de similaridades seja obtido rapidamente. Com um grande número de vértices este cálculo é computacionalmente caro[1]. Armazenar todas as distâncias par-a-par deste grafo em uma matriz $|V| \times |V|$ em memória também pode não ser possível para uma grande quantidade de vértices.

O grafo de distâncias também pode ser utilizado como entrada de técnicas de *embedding* de grafos, para as técnicas LLE e IsoMap também é necessário derivar uma matriz de distâncias, o que é descutido na seção 5.3.

O grafo de similaridades G_{Sim} gerado tem como peso das arestas sua similaridade, ou seja, a aresta entre dois vértices V_i e V_j é $Sim(V_i, V_j)$, em que $Sim(V_i, V_j) = cosseno(V_i, V_j)$.

Para gerar o grafo de distâncias G_{Dis} a similaridade deve ser convertida em distância e portanto é utilizada a equação

$$d_{acos}(i, j) = arccos(Sim(i, j)), \tag{5.1}$$

em que *arccos* é a função arco-cosseno e esta representa uma maneira de derivar uma distância métrica a partir de da similaridade cosseno [van Dongen & Enright, 2012].

A distância entre dois itens i e j é considerada métrica quando é considerada métrica quando obedece as seguintes propriedades:

- Positiva, $d(i, j) \geq 0$;
- Simétrica, $d(i, j) = d(j, i)$;
- $d(i, j) = 0$ se e somente se $i = j$;
- Obedece a desigualdade triangular, $d(i, k) \leq d(i, j) + d(j, k)$.

[1] A complexidade de tempo para calcular a menor distância entre dois vértices através do algoritmo Dijkstra com Fibonacci *heap* é $O(|E| + |V| \log |V|)$.

5.3 Técnicas de *embedding* de grafos

As técnicas de *embedding* de grafos são utilizadas para converter os dados de um grafo em um espaço Euclidiano, associando um vetor de coordenadas $\mathbf{x_i} = (x_1, x_2, x_3, \cdots, x_n)$ em um número finito de dimensões para cada nó i do grafo, preservando as distâncias entre pares de nós quaisquer. Assim, operações como cálculo de distâncias passa a ser uma tarefa computacionalmente trivial, dado que a distância Euclidiana entre dois pontos é calculada em tempo constante. Adicionalmente, técnicas de *embedding* proporcionam melhor visualização dos dados, bem como maneiras mais intuitivas de navegação através do conjunto de conteúdos de mídia [Cardoso et al., 2016].

As técnicas de *embedding* também devem ser capazes de reduzir a dimensionalidade dos dados, permitindo que essa matriz de coordenadas represente esses dados com poucas dimensões. Esta matriz de coordenadas geradas, portanto, pode ser armazenada em memória mais facilmente já que ela tem tamanho $N \times Q$ e não $N \times N$, em que $Q \ll N$. A redução de dimensionalidade é definida como a transformação de dados com alta dimensionalidade em uma representação significativa em uma dimensionalidade mais baixa. Esta transformação é feita com o objetivo de melhorar a visualização dos dados ou para evitar os efeitos indesejáveis da alta dimensionalidade dos mesmos [Jimenez & Langrebe, 1998].

Técnicas de redução de dimensionalidade podem ser aplicadas para mergulhar grafos em um espaço Euclidiano. Nesta dissertação, analisamos as seguintes técnicas: IsoMap [Tenenbaum et al., 2000], LLE [Roweis & Saul, 2000] e LINE [Tang et al., 2015]. Estas técnicas preservam as propriedades locais dos dados ou então, propriedades globais dos mesmos.

As técnicas locais (LLE, por exemplo) buscam preservar a geometria local dos dados; essencialmente, estas buscam mapear pontos próximos do *manifold*[2] em pontos próximos na representação de menor dimensão. As técnicas globais (IsoMap, por exemplo) buscam preservar esta geometria em todas as escalas, mapeando ponto próximos do *manifold* em pontos próximos na representação de menor dimensão, e pontos distantes em pontos distantes. Já a técnica LINE não representa uma técnica específica para redução de dimensionalidade, já que ela deve ser aplicada em grafos e não em uma matriz de distâncias ou de coordenadas, embora ela seja capaz de converter o grafo em um espaço de coordenadas de baixa dimensionalidade.

A seguinte notação será utilizada nas próximas seções. Matrizes serão denotadas por letra maiúscula em negrito, como por exemplo $\mathbf{D}$. Vetores por letras minúsculas em negrito ($\mathbf{x}$). Vetores ou matrizes transpostas serão denotadas por $\mathbf{x}^T$ e $\mathbf{D}^T$,

[2]Espaço topológico que localmente se assemelha a um espaço Euclidiano.

Tabela 5.1: Símbolos utilizados na definição dos métodos.

$\mathbf{D}$	Matriz de distâncias entre todos os pontos.
δ_{ij}	Distância entre os itens i e j na matriz de distâncias $\mathbf{D}$.
N	Quantidade de itens em $\mathbf{D}$.
Q	Número de dimensões desejadas no espaço euclidiano gerado.
$\mathbf{X}_Q$	Espaço euclidiano, com Q dimensões, gerado através dos métodos.
$\mathbf{x}_i$	Vetor de Q dimensões representando o item i no espaço $\mathbf{X}_Q$.
$\mathbf{B}$	Matriz de *Gram*, ou o produto interno de uma mesma matriz, como $\mathbf{B} = \mathbf{X}^T\mathbf{X}$.
$\mathbf{\Lambda}_Q$	Matriz diagonal dos Q maiores autovalores de uma matriz.
$\mathbf{V}_Q$	Matriz dos Q autovetores associados associados à $\mathbf{\Lambda}_Q$.

respectivamente. O produto interno entre dois vetores é denotado por $\mathbf{xy}$. A matriz identidade é denotada por $\mathbf{I}$. $\mathbf{1}$ é uma matriz completa composta por números reais em que todos os elementos são iguais a 1. Para facilitar a consulta e visualização, os símbolos utilizados na definição dos métodos estão sintetizados na Tabela 5.1.

5.3.1 Escalonamento Multidimensional Clássico (cMDS)

O termo escalonamento multidimensional (ou *multidimensional scaling* (MDS)) refere-se a uma família de técnicas de redução de dimensionalidade [Brandes & Pich, 2007], que mapeia relações par-a-par em coordenadas pertencentes a um espaço métrico. A primeira destas técnicas foi proposta por Torgerson [1965], referenciada como MDS Clássico (cMDS). Esta técnica linear possui como entrada uma matriz com as dissimilaridades entre todos os pares de itens e encontra as coordenadas para cada um dos pontos, preservando da melhor maneira possível as dissimilaridades destes pares.

Dada uma matriz $\mathbf{D} \in \mathbb{R}^{N\times N}$ de distâncias δ_{ij} (ou dissimilaridades) entre os itens $i, j \in \{1, ..., N\}$, o objetivo do método cMDS é encontrar uma matriz de coordenadas com dimensão reduzida $\mathbf{X}_Q \in \mathbb{R}^{N\times Q}$, com $Q \ll N$, onde a distância Euclidiana ao quadrado[3], $\|\mathbf{x_i} - \mathbf{x_j}\|^2$, dessa matriz de coordenadas seja aproximadamente igual a distância δ_{ij}. Portanto, as distâncias Euclidianas no espaço gerado pelas coordenadas da matriz $\mathbf{X}$ buscam preservar as distâncias δ_{ij}. As distâncias Euclidianas podem ser definidas através do produto interno entre os vetores $\mathbf{x_i}$ e $\mathbf{x_j}$, assumindo que $\mathbf{X}$ está centralizado [Torgerson, 1965], através da fórmula:

$$\|\mathbf{x_i} - \mathbf{x_j}\|^2 = (\mathbf{x_i} - \mathbf{x_j})^T(\mathbf{x_i} - \mathbf{x_j}) = \mathbf{x_i}^T\mathbf{x_i} - 2\mathbf{x_i}^T\mathbf{x_j} + \mathbf{x_j}^T\mathbf{x_j}. \tag{5.2}$$

Porém, como a matriz X não é conhecida, somente a matriz de distâncias $\mathbf{D}$,

[3]A distância ao quadrado é utilizada para associar mais peso aos objetos que estão mais distantes, de forma progressiva.

podemos encontrar a matriz de *Gram* $\mathbf{B} = \mathbf{X}^T\mathbf{X}$ [Lee & Verleysen, 2007], ou seja, a matriz de produtos internos de $\mathbf{X}$. Portanto:

$$\delta_{ij}^2 = b_{ii} - 2b_{ij} + b_{jj}, \tag{5.3}$$

assumindo que δ_{ij} é uma distância Euclidiana. Para calcular a matriz de *Gram* $\mathbf{B}$ é utilizado um método denominado "dupla centralização" (*double centering*) da matriz $\mathbf{D}$ [Torgerson, 1965]:

$$b_{ij} = -\frac{1}{2}\left(\delta_{ij}^2 - \frac{1}{N}\sum_{k=1}^{N}\delta_{ik}^2 - \frac{1}{N}\sum_{k=1}^{N}\delta_{kj}^2 + \frac{1}{N^2}\sum_{r=1}^{N}\sum_{s=1}^{N}\delta_{rs}^2\right), \tag{5.4}$$

ou

$$\mathbf{B} = -\frac{1}{2}\mathbf{J}\mathbf{D}^2\mathbf{J}, \tag{5.5}$$

em que $\mathbf{J} = \mathbf{I} - N^{-1}\mathbf{1}\mathbf{1}^T$. Após calcular a matriz $\mathbf{B}$, o *embedding* da matriz $\mathbf{D}$, em um espaço dimensional reduzido é obtido através da decomposição em autovalores e autovetores $B = \mathbf{V}_Q\mathbf{\Lambda}_Q\mathbf{V}_Q^T$ [Brandes & Pich, 2007], em que $\mathbf{\Lambda}_Q$ é a matriz diagonal dos Q maiores autovalores da matriz $\mathbf{B}$ e $\mathbf{V}_Q$ a matriz $N \times Q$ dos autovetores associados. O objetivo é considerar as Q linhas que melhor preservam as distâncias entre todas as possíveis reduções lineares de $\mathbf{X}$. Assim:

$$\mathbf{X}_Q = \mathbf{\Lambda}_Q^{\frac{1}{2}}\mathbf{V}_Q^T. \tag{5.6}$$

A decomposição de $\mathbf{B}$ em autovalores e autovetores possui uma complexidade $O(N^3)$ para calcular todos os autovetores. Tornando este método proibitivo para uso em dados reais, nos casos de uma grande quantidade de itens [de Silva & Tenenbaum, 2002]. Apesar de exigir como entrada a matriz de distâncias $\mathbf{D}$ completa, isto é, com todas as distâncias par-a-par calculadas, o cMDS possibilita interpretar e visualizar os dados em um espaço métrico, caso especialmente interessante para propostas de métodos de navegação em conjunto de mídias.

Como não existe uma matriz de distâncias $\mathbf{D}$ no contexto deste trabalho inicialmente, será utilizada uma extensão deste método denominada IsoMap.

5.3.2 IsoMap

Proposta por Tenenbaum et al. [2000], a técnica IsoMap estende a técnica cMDS, transformando-a em um método não-linear. Para tal, são incorporadas duas etapas

anteriores à aplicação da técnica cMDS:

1. Geração de um grafo G_{Iso} a partir de uma matriz de coordenadas $\mathbf{M} \in \mathbb{R}^{N \times P}$ do conjunto de dados, em que P é a quantidade de dimensões deste espaço. O grafo de vizinhanças deve incluir como vizinhos de um nó os k pontos mais próximos no espaço gerado pelas coordenadas em $\mathbf{M}$.

2. Geração de uma matriz de distâncias par-a-par completa $\mathbf{D_{Iso}}$ a partir desse grafo, calculando os menores caminhos entre todos os pares de dados.

Após estas duas etapas iniciais, o método cMDS é aplicado na matriz $\mathbf{D_{Iso}}$. O objetivo destes dois passos anteriores ao cMDS é capturar o *manifold* da matriz $\mathbf{M}$ ao montar o grafo de vizinhanças, assim os itens mais próximos em $\mathbf{M}$ permanecerão próximos e os distantes se distanciarão mais ainda. Com isso este grafo explora a linearidade local do *manifold*, construindo distâncias geodésicas entre os pontos [de Silva & Tenenbaum, 2002].

Esta técnica é aplicada no contexto deste trabalho com apenas uma modificação. Já que o grafo de distâncias é conhecido, construído como demonstrado na sessão 5.2, a etapa (1) não é executada ($G_{Iso} = G_{Dis}$), portanto, o método é executado a partir da etapa (2). Esta alteração na técnica para que ela seja executada diretamente a partir de um grafo já foi anteriormente explorada em [Goussevskaia et al., 2008] e [Platt, 2004]. A etapa (2) tem complexidade de tempo $O(|V||E| + |V|^2 log|V|)$ utilizando Dijkstra e Fibonacci *heap*, sendo $|V|$ e $|E|$ o total de nós e arestas do grafo, respectivamente, além de $O(N^3)$ para o cálculo do cMDS. A complexidade de espaço desta técnica é de $O(N^2)$, pois toda a matriz de distâncias é utilizada.

5.3.3 Landmark IsoMap

O Landmark IsoMap (L-IsoMap) é uma aproximação do IsoMap e tem como objetivo diminuir as duas maiores ineficiências do IsoMap [de Silva & Tenenbaum, 2002]: o cálculo dos caminhos mais curtos e a decomposição em autovalores e autovetores.

O primeiro passo do método é selecionar L *landmarks*, ou pontos de referência. Construir uma matriz de distâncias completa $\mathbf{D}_L \in \mathbb{R}^{L \times L}$ utilizando os caminhos mínimos entre estes L pontos e então aplicar o cMDS nesta matriz, gerando um espaço $\mathbf{X}_L = \mathbf{\Lambda}_L^{\frac{1}{2}} \mathbf{V}_L^T$, em que $\mathbf{\Lambda}_L$ é a matriz diagonal dos $|L|$ e $\mathbf{V}_L$ a matriz $N \times L$ dos autovetores associados, obtidos através do cMDS.

O segundo passo é fazer o *embedding* dos $N - |L|$ pontos restantes no espaço gerado. Dado que $\mathbf{D}_R$ é a distância entre os pontos restantes e os escolhidos como

landmarks. Os vetores $\mathbf{x}_R$ dos pontos restantes são calculados através da fórmula

$$\mathbf{x}_R = \frac{1}{2}(\mathbf{V}_L^T/\mathbf{\Lambda}_L^{\frac{1}{2}})(\mathbf{D}_R^2 - \overline{\mathbf{D}_L^2}), \tag{5.7}$$

em que $\overline{\mathbf{D}_L^2}$ é um vetor com a média de cada linha de $\mathbf{D}_L^2$.

O espaço final $\mathbf{X}_Q$ então é a união de $\mathbf{X}_L$ e $\mathbf{X}_R$. A complexidade de tempo para este algoritmo é de $O(|L||E| + |L||V| \log |V|)$ para calcular as distâncias entre todos os *landmarks* do grafo G, onde $Q < |L| \ll N$ e $O(|L|N^2)$ para o cálculo do Landmark MDS (LMDS) [de Silva & Tenenbaum, 2002]. Em comparação com o IsoMap, este método requer apenas $O(|L|N)$ de espaço.

5.3.4 Locally Linear Embedding

O Locally Linear Embedding (LLE), proposto primeiramente por Roweis & Saul [2000] também é um método baseado em grafos para recuperar um *manifold* não-linear, porém diferente do IsoMap seu foco é em preservar apenas a topologia dos vizinhos dos pontos sem considerar as distâncias dos outros pontos.

O algoritmo consiste em calcular os k vizinhos mais próximos de cada ponto, construir uma matriz de peso $\mathbf{W}$ que represente cada ponto baseado apenas em seus vizinhos, e calcular a matriz de coordenadas $\mathbf{X}_Q \in \mathbb{R}^{N \times Q}$ onde $Q \ll N$ utilizando essa matriz de peso.

O padrão do algoritmo tem como entrada uma matriz de coordenadas $\mathbf{M}$, entretanto uma alteração proposta por Roweis & Saul [2000] permite sua utilização em uma matriz de distâncias $\mathbf{D}$.

O objetivo final do algoritmo é minimizar o erro total

$$\mathcal{E}(W) = \sum_{i=1}^{N} \left\| \mathbf{m}_i - \sum_{j=1}^{N} \mathbf{w}_{ij}\mathbf{m}_j \right\|^2, \tag{5.8}$$

em que a matriz $\mathbf{W}$ representa o peso que cada vizinho $\mathbf{m}_j$ tem na reconstrução de um ponto $\mathbf{m}_i$. Os pesos $\mathbf{w}_{ij}$ definem o grau de contribuição que cada vizinho de i tem na sua reconstrução.

O primeiro passo do algoritmo é, portanto, calcular os k vizinhos mais próximos. Como neste trabalho a entrada do algoritmo é um grafo G, os k vizinhos mais próximos são definidos através do próprio grafo e então é gerada a matriz $\mathbf{D}_P$ representando as distâncias entre os pontos e seus k vizinhos mais próximos, portanto $\mathbf{D}_P$ representa uma matriz esparsa de distâncias.

Para calcular o peso $\mathbf{w}_{ij}$ de cada ponto $\mathbf{m}_i$, a função deve seguir duas restrições: todos os pontos que não vizinhos de $\mathbf{m}_i$ possuem peso igual zero $\mathbf{w}_{ij} = 0$ e o somatório de todos os pesos do ponto $\mathbf{m}_i$ deve ser igual a um $\sum_{j=1}^{N} \mathbf{w}_{ij} = 1$.

Considerando um ponto $\mathbf{m}_i$ seu erro é dado por

$$\mathcal{E}_i(\mathbf{W}) = \sum_{j,s=1}^{k} \mathbf{w}_{ij}\mathbf{w}_{is}\mathbf{c}_{js}, \tag{5.9}$$

em que $\mathbf{C} \in \mathbb{R}^{k\times k}$ representa a matriz local de covariâncias,

$$\mathbf{c}_{js} = (\mathbf{m}_i - \mathbf{e}_j)^T(\mathbf{m}_i - \mathbf{e}_s), \tag{5.10}$$

construída utilizando a submatriz $\mathbf{E} \in \mathbb{R}^{k\times k}$ que contém apenas os $\mathbf{m}$ vetores vizinhos de i.

Como para este trabalho é utilizada uma matriz de distâncias $\mathbf{D}$, e não a matriz de coordenadas $\mathbf{M}$, esta fórmula é alterada para

$$\mathcal{E}_i(\mathbf{W}) = \sum_{j,s=1}^{k} \mathbf{w}_{ij}\mathbf{w}_{is}\mathbf{b}_{js}, \tag{5.11}$$

de acordo com Roweis & Saul [2000]. Em que $\mathbf{B} \in \mathbb{R}^{k+1\times k+1}$ representa a matriz de Gram local. O cálculo de $\mathbf{B}$ é feito exatamente como na equação 5.3.1 do cMDS, porém a matriz utilizada no cálculo de $\mathbf{B}$ é apenas a submatriz completa $\mathbf{D}_E \in \mathbb{R}^{k+1\times k+1}$, que contém as distâncias entre i e seus k vizinhos mais próximos.

Os pesos $\mathbf{W}$ são calculados resolvendo o sistema de equações lineares $\sum_{j=1}^{k} \mathbf{b}_{js}\mathbf{w}_{ij} = 1$.

O último passo do algoritmo é encontrar os vetores $\mathbf{x}_i$ minimizando a função de custo, equação 5.3.4, considerando que os pesos $\mathbf{W}$ são fixos. Isto é feito com a decomposição em autovalores e autovetores de uma matriz $\mathbf{Y} = (\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})$. O espaço final $\mathbf{X}_Q$ então consiste dos $\mathbf{Q}+1$ menores autovetores com autovalores maiores que zero, descartando o primeiro autovetor pois ele contém apenas elementos iguais a 1.

O método LLE portanto busca preservar a topologia dos vizinhos de cada ponto, calculando os pesos que cada vizinho possui na reconstrução destes pontos. Sua complexidade de tempo é de $O(N^2k^3)$ para calcular a matriz de pesos e $O(QN^2)$ para calcular os menor autovetores dessa matriz.

5.3.5 Large-scale Information Network Embedding

O Large-scale Information Network Embedding (LINE) é um método mais recente, proposto por Tang et al. [2015] como um algoritmo específico para *embedding* de grafos e redes. Os autores destacam que os algoritmos IsoMap e LLE utilizam apenas a proximidade de primeira ordem, ou seja, apenas os vizinhos diretos de cada vértice de um grafo.

Além de utilizar a proximidade de primeira ordem, o LINE também pode utilizar a de segunda ordem, os vizinhos de vizinhos, já que eles também possuem informação da proximidade entre dois vértices. Vértices que compartilham vizinhos tendem a ser mais próximos que os que não compartilham, principalmente em grafos do mundo real [Tang et al., 2015].

Este método utiliza como entrada um grafo G de similaridades, ao contrário do LLE e do IsoMap que utilizam matrizes de distância, e este grafo pode ser direcionado, não-direcionado e com ou sem pesos nas arestas. Cada aresta $e \in E$ do grafo de similaridades $G(V, E)$ tem um peso w_{ij} entre os vértices i e j, como este é um grafo não direcionado $w_{ij} = w_{ji}$.

O algoritmo define uma função objetivo para o *embedding* que deve ser minimizada considerando os vizinhos de primeira ordem

$$\mathbf{O}_1 = -\sum_{(i,j)\in E} w_{ij} \log \mathbf{p}_1(v_i, v_j), \tag{5.12}$$

em que

$$\mathbf{p}_1(v_i, v_j) = \frac{1}{1 + \exp(-\mathbf{u}_i^T \mathbf{u}_j)}, \tag{5.13}$$

ao encontrar o vetor $\mathbf{u}_i$ que minimize este objetivo, cada vértice pode ser projeto no espaço Q-dimensional.

Para calcular o *embedding* utilizando a proximidade de segunda ordem, ou seja, vértices que compartilham muitos vizinhos devem ficar mais próximos a função objetivo para ser minimizada é

$$\mathbf{O}_2 = -\sum_{(i,j)\in E} \lambda_i w_{ij} \log \mathbf{p}_2(v_j|v_i), \tag{5.14}$$

onde

$$\mathbf{p}_2(v_j|v_i) = \frac{\exp(\mathbf{u'}_j^T)\mathbf{u}_i}{\sum_{k=1}^{|V|} \exp(\mathbf{u'}_k^T \mathbf{u}_i)}, \tag{5.15}$$

diferente da função de objetivo $\mathbf{O}_1$ esta função define o vetor $\mathbf{u'}_i$ que simboliza o

Tabela 5.2: Entrada e custo computacional por técnica de *embedding*.

Técnica	Entrada	Custo computacional (tempo)
IsoMap	Matriz completa de distâncias	$O(N^3) + O(\|V\|\|E\| + \|V\|^2 log\|V\|)$
L-IsoMap	Matriz esparsa de distâncias	$O(\|L\|N^2) + O(\|L\|\|E\| + \|L\|\|V\| \log \|V\|)$
LLE	Matriz esparsa de distâncias	$O(N^2k^3) + O(QN^2)$
LINE	Grafo de similaridades	$O(QK\|E\|)$

contexto específico no qual o vértice v_i está inserido. Este contexto representa neste caso as vizinhanças do vértice. Já λ_i é o peso que cada vértice tem na rede, o que é calculado utilizando o grau do vértice v_i.

O aprendizado de $\mathbf{u}_i$ e $\mathbf{u'}_i$ ao minimizar a função objetivo $\mathbf{O}_2$ permite a representação de cada vértice v_i com o vetor Q-dimensional $\mathbf{u}_i$.

Para otimizar estes objetivos Tang et al. [2015] utilizam uma estratégia de amostragem das arestas e então eles são calculados através de técnicas de gradiente descendente estocástico [Recht et al., 2011] a complexidade de tempo final do algoritmo é de $O(QK|E|)$, em que K representa o número de amostras negativas, com objetivo de minimizar o custo de comparar todas as arestas entre si.

A combinação das proximidades de primeira e segunda ordem é feita concatenando os *embeddings* treinados por cada método. Assim é obtido um espaço que otimize os dois objetivos.

5.4 Resumo do capítulo

A Tabela 5.2 apresenta um sumário dos custos computacionais e das entradas para cada técnica de *embedding* analisada. A técnica LINE apresenta o menor custo computacional, principalmente quando utilizada em grafos esparsos. O fato de usar diretamente um grafo de similaridades, além de diminuir o custo computacional, evita algum possível ruído nas estratégias de converter esse grafo em outras estruturas.

Já as outras técnicas utilizam como entrada matrizes de distância, pois elas são técnicas com foco apenas em redução de dimensionalidade e, portanto, exigem um pré-processamento do grafo de similaridades. A técnica IsoMap necessita de toda a matriz de distâncias o que pode ser impossível de armazenar em memória primária para entradas com muitos dados. A L-IsoMap é uma aproximação que com apenas cerca de 10% dessa matriz, consegue resultados muito próximos de sua técnica base. Assim como a L-IsoMap, a técnica LLE também permite o uso em matrizes esparsas.

O objetivo principal das técnicas Isomap/L-IsoMap é preservar as distâncias geodésicas globais da entrada, sendo indicada em aplicações que as distâncias da estrutura

final sejam linearmente preservadas em relação às distâncias do grafo inicial. A LLE busca manter a topologia das vizinhanças locais, ignorando os pontos mais distantes na reconstrução do espaço de menor dimensionalidade, o que pode causar uma distorção maior globalmente.

A técnica LINE, assim como a LLE, também tem como principal foco preservar a topologia das vizinhanças locais. Entretanto, ela também leva em conta vizinhanças de segunda ordem (vizinhos de vizinhos), essa estratégia pode ser bem útil em amostras com ruídos ou com poucas informações de vizinhanças já que ela complementa dados esparsos.

Capítulo 6

Análise Experimental

Neste capítulo as técnicas discutidas no capítulo 5 são comparadas e analisadas de acordo com experimentos. A seção 6.1 discute as estratégias para obter o grafo de similaridades e a matriz de distâncias, comparando o espaço de similaridades gerado por cada estratégia. A seção 6.2 apresenta as métricas utilizadas para comparação dos métodos de *embedding*. Na seção 6.3 são escolhidos e comparados os parâmetros de cada método. A performance dos métodos de *embedding* é comparada na seção 6.4, de acordo com as métricas definidas.

6.1 Obtenção do grafo de similaridade e da matriz de distâncias

Devido ao possível alto custo computacional para a execução do *embedding* do grafo gerado a partir dos dados coletados das redes sociais, esta seção apresenta diferentes metodologias para reduzir o grafo de similaridade, e consequentemente, a matriz de distâncias $\mathbf{D}$. Estas estruturas obtidas através dos cálculos apresentados na seção 5.2 e que serão utilizadas como dados de entrada dos algoritmos descritos na seção 5.3. No caso do algoritmo LINE, o grafo de similaridades é fornecido como entrada para o método. Para os demais algoritmos apresentados, a matriz de distâncias $\mathbf{D}$ deve ser fornecida. Devido à complexidade computacional do método LLE não foi possível executar este método para a base de dados de músicas nos servidores utilizados.

Como primeira etapa de filtragem dos dados, foram removidas arestas relacionadas a pares de itens que coocorreram somente uma vez na lista de preferências dos usuários. Para todos os algoritmos, somente a maior componente conexa do grafo de similaridades é utilizada. Denotaremos este grafo por grafo inicial, ou G_I. A partir

Tabela 6.1: Quantidade de vértices e arestas do grafo inicial

		Total	Coocorrência ≥ 2	Grafo G_I
Dados de filmes/TV	Vértices	14.358	14.206	14.206
	Arestas	74.866.374	65.088.065	65.088.065
Dados de música	Vértices	983.010	305.561	271.670
	Arestas	54.424.435	5.236.981	5.185.204

deste grafo inicial, a matriz de distâncias correspondente, $\mathbf{D_I}$, é definida.

A Tabela 6.1 mostra o tamanho do grafo ao se remover as arestas que coocorrem apenas uma vez. O grafo de coocorrência de músicas, considerando todos os dados possui 983.010 vértices e 54.424.435 arestas. Ao se remover as arestas que coocorrem apenas uma vez o grafo fica com 305.561 vértices 5.236.981 de arestas. O grafo G_I que contém apenas a maior componente conexa contém 271.670 de vértices e 5.185.204 de arestas (densidade igual a $0,00007$).

O grafo de filmes e programas de TV tem inicialmente 14.358 vértices, com 74.866.374 arestas. Considerando apenas os itens que coocorrem duas ou mais vezes resulta em 14.206 vértices e 65.088.065 (densidade igual a $0,32$), que também corresponde à maior componente conexa formando o grafo inicial G_I.

Mesmo com a redução inicial dos dados de entrada, o custo computacional dos métodos continua elevado. Por exemplo, apenas o custo de tempo para construir a matriz completa de distâncias é de aproximadamente $O(|V||E| + |V|^2 \log |V|) \approx$ 927.424.017.302 para o grafo de filmes e programas de TV, onde $|V|$ é o total de vértices e $|E|$ o total de arestas. Assim, propomos duas metodologias para descarte de arestas: (1) baseada no número mínimo de coocorrências e; (2) baseada no valor mínimo de cosseno. A metodologia (1) é simplesmente uma extensão da filtragem simples realizada no conjunto de dados inicial. A metodologia (2) se baseia na ideia de que ao se eliminar as arestas de maior distância (menor cosseno) os vizinhos mais próximos de cada vértice permanecerão próximos. No entanto, buscamos reduzir a quantidade de arestas a serem manipuladas pelos métodos sem que a componente conexa reduza muito de tamanho.

A Figura 6.1 mostra a distribuição das arestas de acordo com a coocorrência dos seus pares de itens, considerando os dois grupos de dados utilizados nesta dissertação: filmes e programas de TV (tvtag) e músicas (Last.fm). Aproximadamente 62% das arestas possuem coocorrência maior do que 10 para o grafo de similaridades de filmes e programas de TV e 6% para o grafo de músicas. Por exemplo, se decidirmos eliminar todas as arestas com limite mínimo de 20 coocorrências, é possível descartar 50% das arestas do grafo de similaridades do conteúdo de filmes/TV e $0,97\%$ das arestas do

Tabela 6.2: Grafos de filmes/TV após remoção de arestas

Grafo	Quantidade de vértices	Quantidade de arestas	Grau médio
Inicial	14.206	65.088.065	9.163 ± 62
Coocorrências ≥ 20	12.653	33.134.395	5.237 ± 61
Cosseno $\geq 2 \times 10^{-3}$	13.901	17.871.148	2.568 ± 41

Tabela 6.3: Grafos de música após remoção de arestas

Grafo	Quantidade de vértices	Quantidade de arestas	Grau médio
Inicial	271.670	5.185.204	$38,55 \pm 0,7$
Coocorrências ≥ 6	62.352	684.452	$21,95 \pm 0,5$
Cosseno $\geq 2 \times 10^{-3}$	259.150	1.583.161	$12,21 \pm 0,06$

grafo de similaridades do conteúdo de música.

A Figura 6.2 apresenta a distribuição das arestas de acordo com o cosseno. Cerca de 40% das arestas tem um cosseno de pelo menos 10^{-3} para o grafo de filmes e programas de TV e 20% para o grafo de músicas. Caso sejam eliminadas todas as arestas com cosseno de até 10^{-3}, 60% dessas arestas serão descartadas para o grafo de filmes e programas de TV e 80% para o grafo de músicas.

Para definir os limites dos valores de coocorrência e cosseno, arestas foram retiradas iterativamente e o tamanho da maior componente conexa foi calculada a cada eliminação de arestas. A Figura 6.3 mostra a evolução do tamanho da maior componente conexa para o grafo de filmes e programas de TV e de músicas, quando a métrica de cosseno é utilizada. Complementando os resultados, a Figura 6.4 mostra esta evolução, considerando a métrica de coocorrências.

Iremos considerar o valor de 2×10^{-3} para o valor mínimo de cosseno, que mantém 97% dos nós na componente conexa e remove aproximadamente 72% das arestas no grafo de filmes e programas de TV e remove cerca de 69% das arestas, mantendo 95% dos nós no grafo de músicas. As Tabelas 6.2 e 6.3 apresentam as características dos grafos reduzidos obtidos a partir do grafo inicial para os dados de filmes/progamas de TV e de música, respectivamente.

As Figuras 6.5 e 6.6 mostram o grau dos vértices dos grafos inicial e reduzidos, eliminando arestas através do valor mínimo de cosseno e de coocorrência. Mesmo com a estratégia de remoção de arestas o grafo de filmes e programas de TV é mais conexo que o de músicas, já que ele também tem mais usuários por item analisado.

Essa estratégia de remoção de arestas, assim como a diminuição dos graus dos vértices tem impacto direto, em termos de custo computacional, na execução dos métodos de *embedding*, bem como na qualidade do espaço de similaridades gerado. As

métricas de qualidade são discutidas na seção 6.2. Uma discussão mais detalhada sobre a influência da eliminação de arestas na qualidade dos espaços gerados é apresentada na seção 6.4.

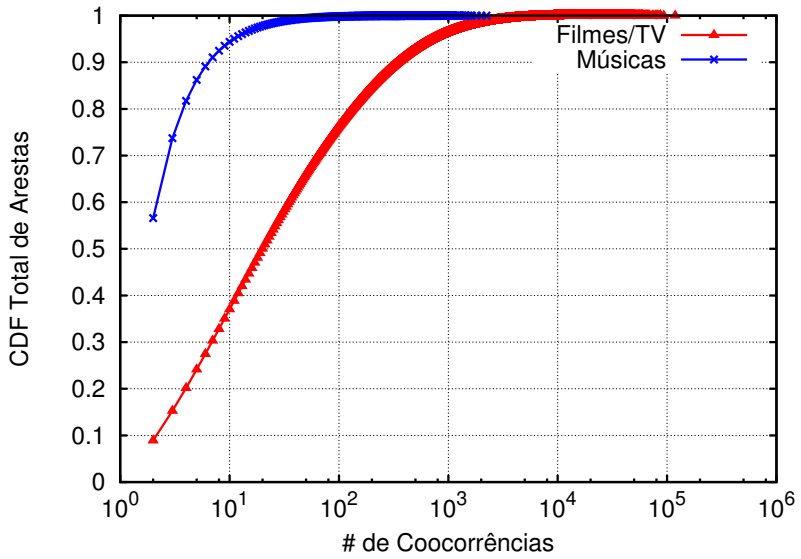

Figura 6.1: CDF do total de arestas por coocorrência.

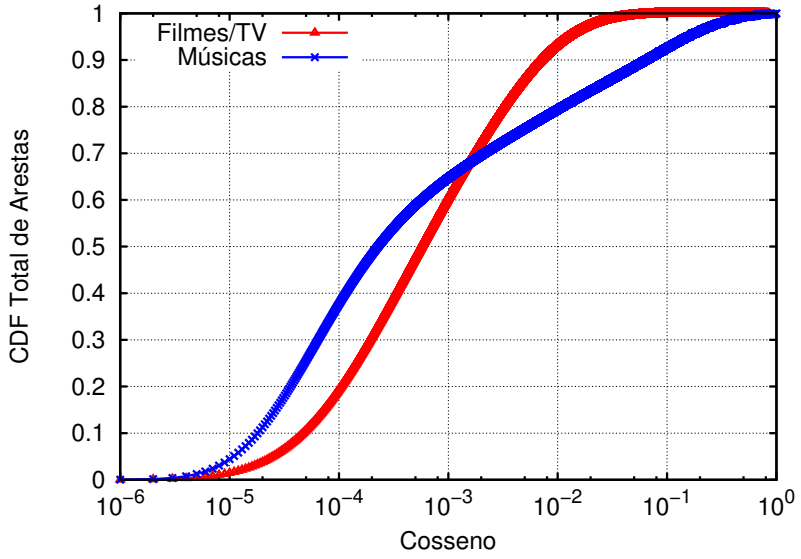

Figura 6.2: CDF do total de arestas por cosseno.

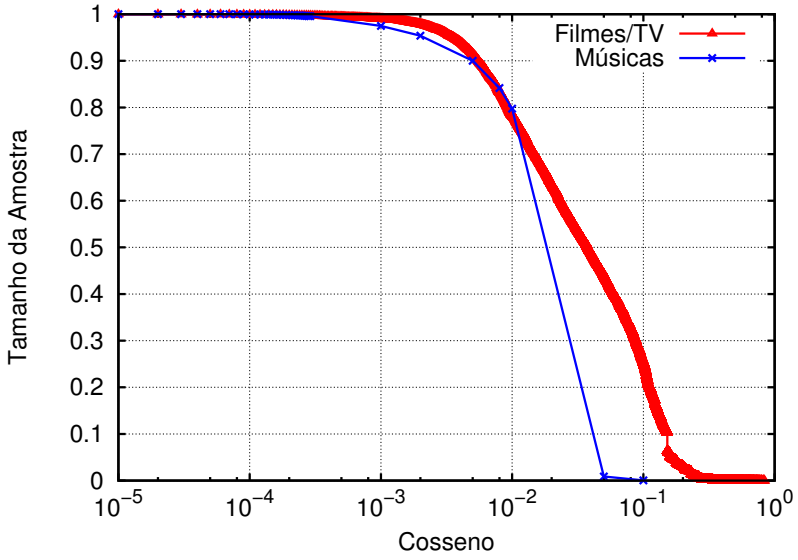

Figura 6.3: Tamanho da amostra por limite de cosseno.

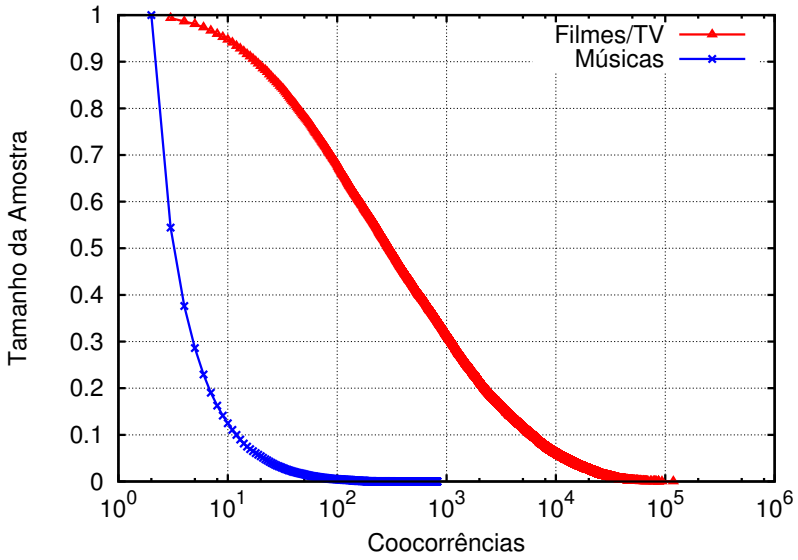

Figura 6.4: Tamanho da amostra por limite de co-corrências.

6.2 Métricas de qualidade

Diferentes técnicas de *embedding*, juntamente com a definição dos seus parâmetros, geram diferentes espaços de similaridades. Desta forma, métricas para a avaliar a qualidade destes espaços são utilizadas. Esta seção apresenta estas métricas. As métricas de variância residual (Tenenbaum et al. [2000]) e da ordem de global dos vizinhos (Kayo [2006]), são encontradas na literatura. As demais métricas, similaridades entre

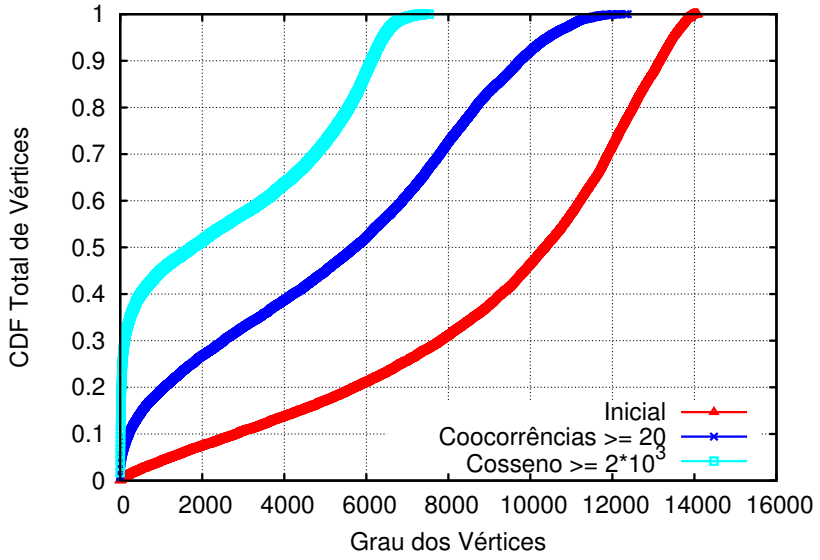

Figura 6.5: CDF de graus dos vértices para a amostra de filmes/TV.

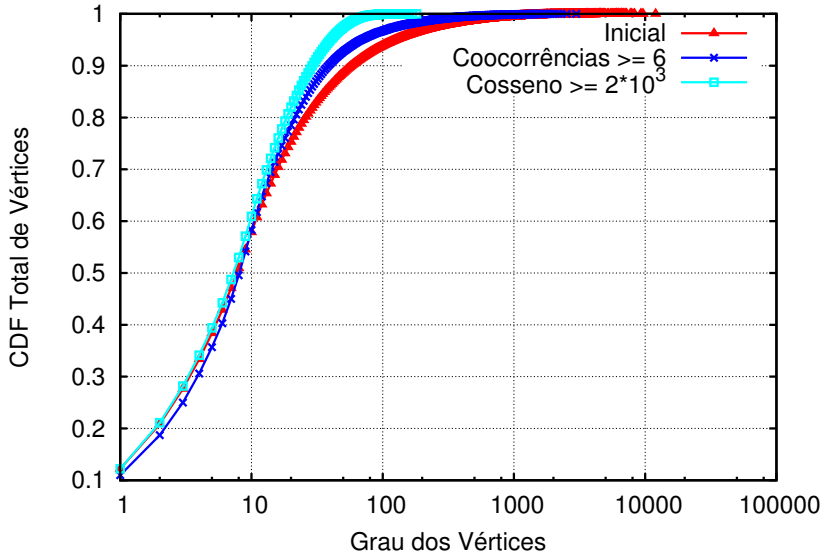

Figura 6.6: CDF de graus dos vértices para a amostra de músicas.

vizinhos, gêneros e a análise manual de vizinhanças, foram definidas nesta dissertação.

Variância residual: Baseada nas matrizes de distâncias, esta métrica é calculada pela fórmula: $1 - \rho^2(\mathbf{D}_E, \mathbf{D}_I)$, onde $\mathbf{D}_E$ representa as distâncias entre todos os pontos de um espaço gerado pela técnica de *embedding* E, e $\mathbf{D}_I$ são as distâncias entre os vértices do grafo inicial G_I utilizado na geração do espaço de similaridades. ρ é o coeficiente de correlação de Pearson. Esta métrica varia no intervalo $[0, 1]$ e representa a não-correlação linear entre as distâncias do espaço gerado por cada técnica e o grafo

inicial. Valores próximos a 0 significam que a técnica é capaz de preservar as distâncias do grafo inicial, mesmo após o *embedding* do grafo de similaridades.

Ordem global dos vizinhos: Esta métrica tem como objetivo analisar se a ordenação dos vizinhos que estão a uma determinada distância de um ponto se preserva quando são consideradas as matrizes de distâncias $\mathbf{D}_E$ e $\mathbf{D}_I$. O valor desta métrica é obtido a partir do coeficiente de Spearman.

Similaridade entre itens vizinhos: O espaço de similaridades gerado deve ser capaz de preservar os itens com maior similaridade (cosseno) próximos, portanto nesta métrica são selecionados 1.000 pontos aleatórios no espaço gerado e é analisada a média entre o cosseno destes itens em relação ao ponto selecionado e seus k vizinhos mais próximos.

Similaridade de gêneros entre itens vizinhos: É esperado que o espaço de similaridades gerado agrupe itens do mesmo gênero o mais próximo possível no espaço métrico (por exemplo, músicas de rock permanecem na mesma vizinhança, bem como séries de TV de suspense). Nesta dissertação esta métrica somente é aplicada aos dados coletados a partir do tvtag[1], e como os itens analisados possuem mais de um gênero, definimos o cosseno entre gêneros, calculado a partir da equação de similaridade entre itens (seção 5.1). Neste cálculo o cosseno entre os gêneros é, portanto, o cosseno das coocorrências entre os gêneros nos filmes e programas de TV. O objetivo deste cálculo é definir gêneros que coocorrem com muita frequência como similares.

Através da seleção aleatória de 1.000 pontos no espaço gerado, esta métrica é a média entre o cosseno dos gêneros do ponto selecionado no espaço de similaridades e os k itens mais próximos desse ponto.

Análise manual de vizinhanças locais: A partir da escolha manual de filmes e programas de TV de referência, foram analisados os 9 pontos mais próximos deles. Através da opinião subjetiva dos autores de [Holanda et al., 2015a], bem como da ajuda de uma especialista externa que mantém um blog sobre séries de televisão, a qualidade destas vizinhanças foram analisadas qualitativamente. Nesta métrica foi considerada o quanto os programas mais próximos são similares em público alcançado, tipo de enredo, etc.

6.3 Estudo de parâmetros

As diversas técnicas de *embedding* apresentadas no Capítulo 5 possuem um conjunto de parâmetros que devem ser definidos para a execução das mesmas. Obviamente, a

[1]Para os dados do Last.fm, esta informação pode ser coletada através de outras fontes de dados, como trabalhos futuros.

escolha destes parâmetros afeta diretamente na qualidade do espaço de similaridades gerado. Desta forma, esta seção mostra um estudo mais detalhado da escolha dos principais parâmetros destas técnicas, utilizando como métrica de qualidade a variância residual entre a matriz de distâncias original e a matriz obtida após o *embedding* do grafo sendo estudado.

6.3.1 Número de dimensões

A Figura 6.7 mostra a variância residual por dimensão gerada através da técnica Iso-Map, considerando os dois tipos de dados analisados. Podemos observar que, considerando poucas dimensões (entre 1 e 5) a variância residual é maior que 0,6. O valor da variância residual é sensível até 30 dimensões. Para valores acima de 30, o aumento no número de dimensões não gera um grande impacto na qualidade do espaço gerado (considerando esta métrica). Isto indica que as amostras analisadas possuem provavelmente uma dimensionalidade intrínseca entre 30 dimensões.

As Figuras 6.8, 6.9, 6.10, 6.11 e 6.12 exibem a similaridade média entre os itens em múltiplas dimensões. Este resultado corrobora o fato de que a dimensionalidade intrínseca dos dados está provavelmente entre 30 dimensões, já que a partir dessas dimensões não há um aumento nas similaridades entre os itens.

Embora seja possível representar os dados com um número pequeno de dimensões sem gerar um grande aumento da variância residual (por exemplo, 40, 50 dimensões), nesta dissertação usaremos o total de 100 dimensões para a comparação entre as técnicas consideradas. Este total de dimensões é capaz de representar bem os dados, sem utilizar muito recurso computacional para armazenamento do espaço de similaridades gerado.

6.3.2 Total de *landmarks*

Um parâmetro importante para a técnica L-IsoMap é o conjunto $\mathcal{L}$ contendo os $|\mathcal{L}|$ *landmarks* que serão utilizados como base para geração do espaço de similaridades. A escolha dos *landmarks* é feita utilizando duas políticas:

- Escolha aleatória entre x% pontos no espaço.
- Função MaxMin: sugerida por de Silva & Tenenbaum [2004], primeiramente é definido um conjunto $\mathcal{S}$ contendo $|\mathcal{S}|$ *landmarks* sementes, escolhidos aleatoriamente tal que $|\mathcal{S}| < |\mathcal{L}|$, e então os $|\mathcal{L}| - |\mathcal{S}|$ *landmarks* restantes são selecionados um de cada vez, onde cada novo *landmark* deve maximizar a distância mínima entre os *landmarks* já selecionados e os demais pontos do espaço.

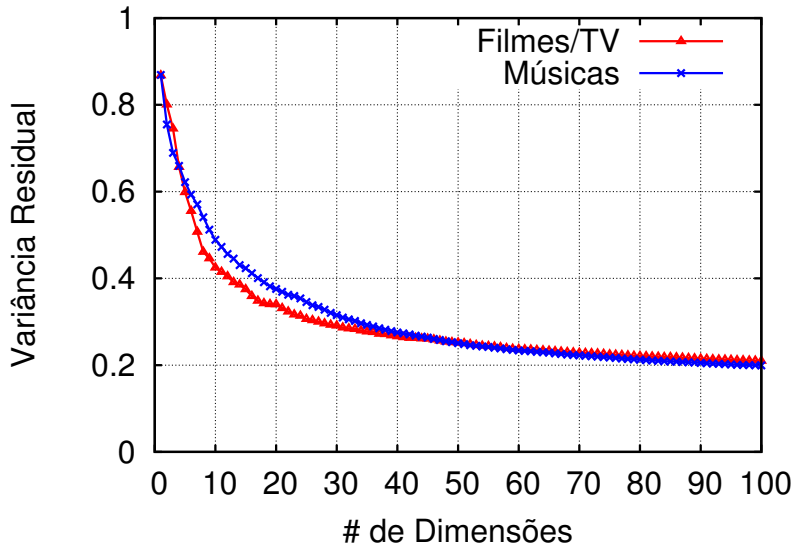

Figura 6.7: Variância residual por dimensões na amostra de filmes/TV com cosseno $\geq 2 \times 10^{-3}$ e na amostra de músicas com coocorrências ≥ 6 com a técnica IsoMap.

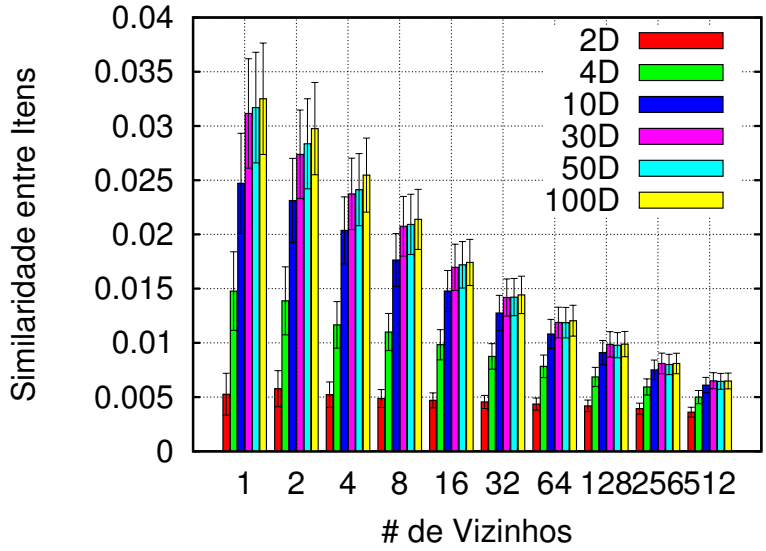

Figura 6.8: Similaridade média entre itens com a técnica IsoMap na amostra de filmes/TV com cosseno $\geq 2 \times 10^{-3}$.

O objetivo desta técnica é que, ao selecionar os pontos mais distantes dos *landmarks* sementes, os novos *landmarks* representem grupos de dados distintos dos que já foram selecionados, porém o custo dessa escolha é de $O(|\mathcal{L}|N)$, onde N é o total de itens da entrada do algoritmo.

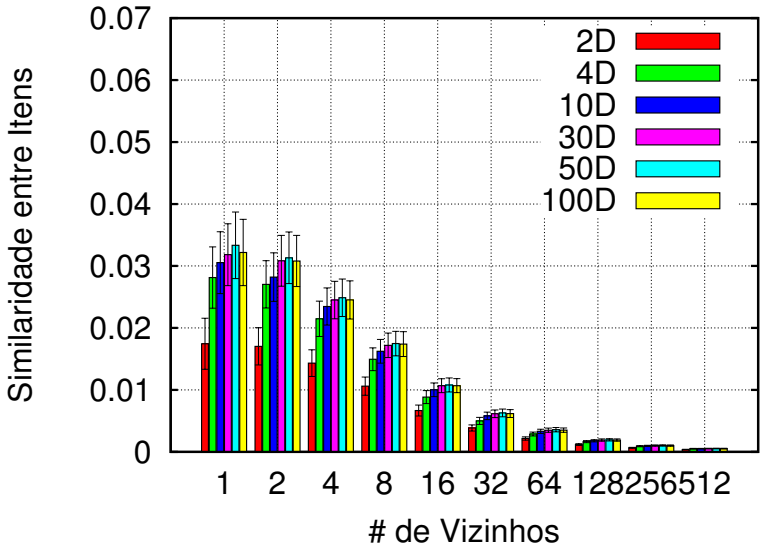

Figura 6.9: Similaridade média entre itens com a técnica IsoMap na amostra de músicas com coocorrências ≥ 6.

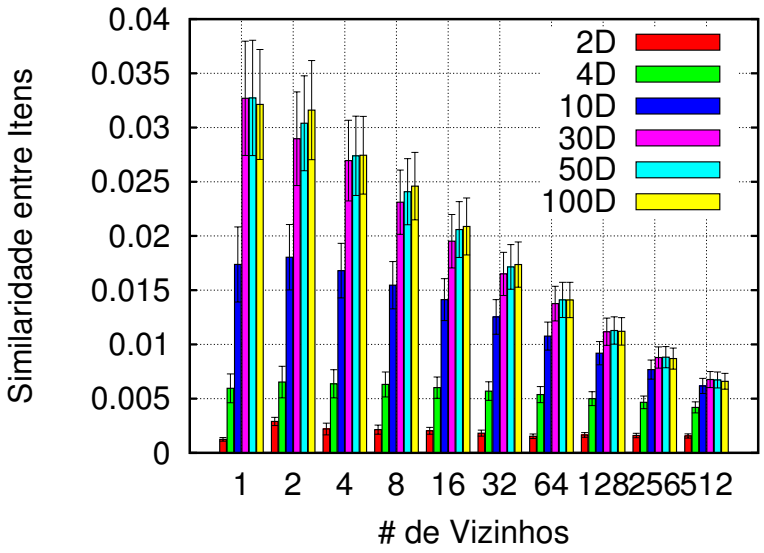

Figura 6.10: Similaridade média entre itens com a técnica LINE na amostra de filmes/TV com cosseno $\geq 2 \times 10^{-3}$.

As Figuras 6.13 e 6.14 mostram a variância residual média por política de escolha de *landmarks*, onde $|\mathcal{S}|$%-Seed e $|\mathcal{L}|$%-IsoMap significa $|\mathcal{S}|$% sementes iniciais aleatórias e $|\mathcal{L}|$% *landmarks*. Como os *landmarks* foram escolhidos aleatoriamente, com exceção da escolha através da função MaxMin que seleciona apenas as sementes aleatoriamente e não todos os *landmarks*, o experimento foi executado dez vezes com cada política

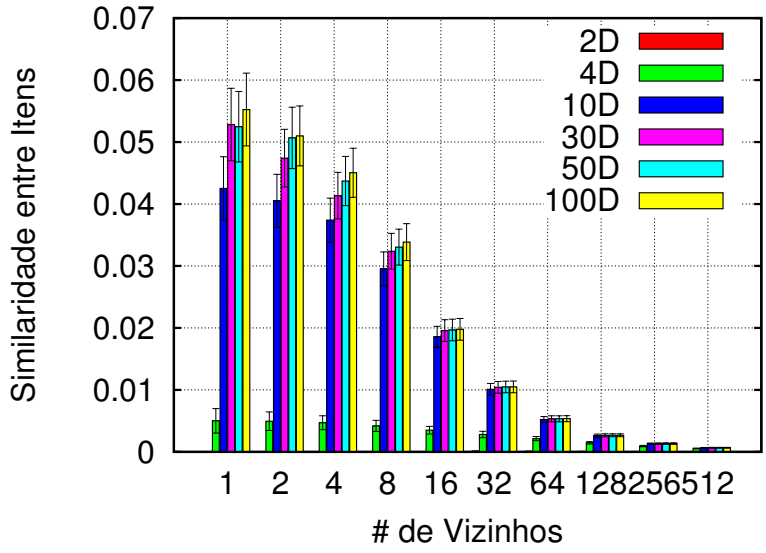

Figura 6.11: Similaridade média entre itens com a técnica LINE na amostra de músicas com coocorrências ≥ 6.

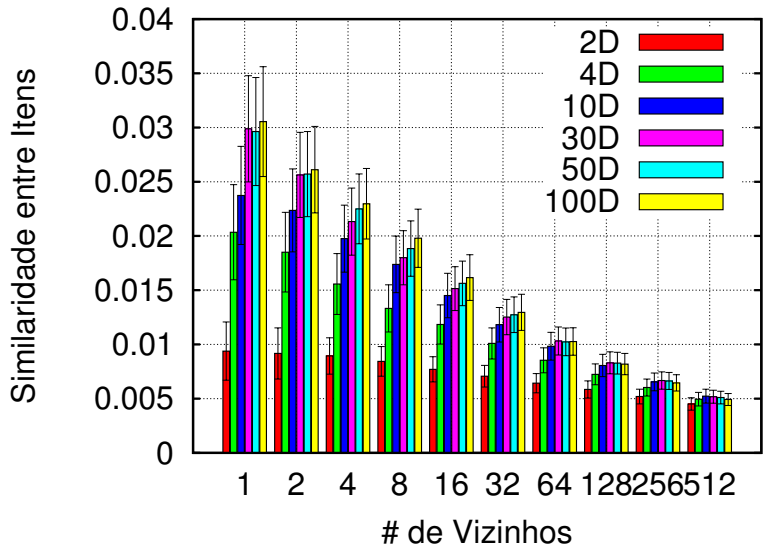

Figura 6.12: Similaridade média entre itens com a técnica LLE na amostra de filmes/TV com cosseno $\geq 2 \times 10^{-3}$.

e exibe um intervalo de confiança de 95% para a amostra de filmes e programas de TV. Considerando as duas amostras, a política de escolha que mais se aproximou do IsoMap completo foi a de 10% de *landmarks*, assim como indicado por de Silva & Tenenbaum [2004]. Na prática a política aleatória funciona muito bem, desde que sejam selecionados uma quantidade de pontos capazes se aproximar do conjunto de

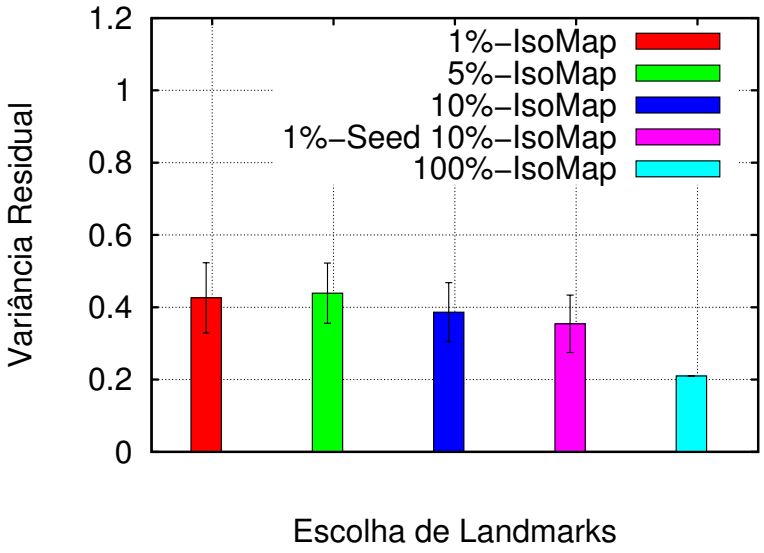

Figura 6.13: Variância residual por escolha de *landmarks* na amostra de filmes/TV com cosseno $\geq 2 \times 10^{-3}$.

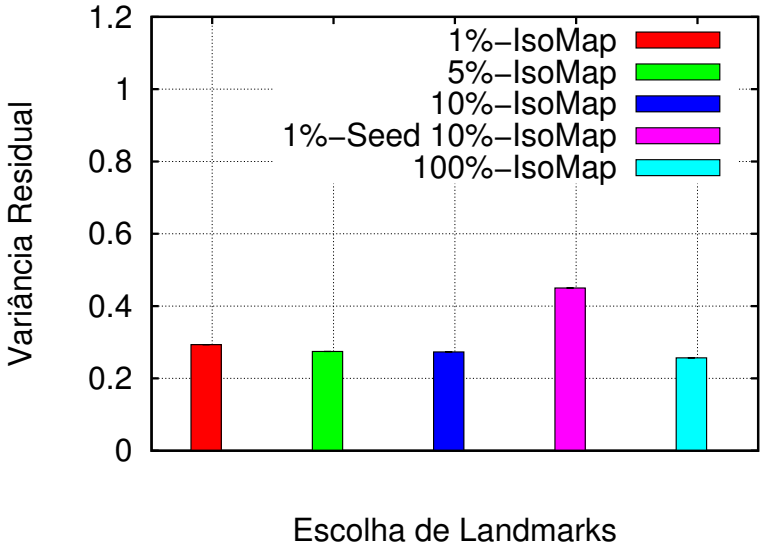

Figura 6.14: Variância residual por escolha de *landmarks* na amostra de músicas com coocorrências ≥ 6.

dados completo. Vale destacar que com apenas 1% da amostra definida como *landmark* a variância residual aumenta em cerca de 20%, o que pode ser um excelente custo-benefício caso a amostra seja muito grande.

6.3.3 Tamanho k da vizinhança

No caso da técnica LLE, o total de vizinhos k a ser considerado é um parâmetro de entrada do algoritmo. Caso este parâmetro seja muito pequeno os pontos podem se separar em sub-grupos no espaço gerado já que ele representa quais vizinhos mais próximos serão utilizados para calcular a posição de cada ponto. Um k pequeno pode separar pontos que sejam próximos no grafo inicial, mas que não estejam em seus respectivos k pontos mais próximos.

Caso o parâmetro k seja muito alto as estruturas de menor escala desse espaço podem ser perdidas, ou seja, os vizinhos mais próximos de cada ponto não terão tanto impacto já que pontos mais distantes serão utilizados no cálculo das coordenadas de cada item [Kayo, 2006].

A Figura 6.15 detalha a variância residual e o coeficiente de Spearman por k avaliado, dentre os os valores observados o valor $k = 84$ foi o que melhor preservou as distâncias observadas (variância residual) e a ordem global de vizinhos (coeficiente de Spearman). Após 84 as duas métricas sofrem pouca alteração nos valores avaliados, porém como o parâmetro k também tem influência direta na complexidade do algoritmo, o valor utilizado nas avaliações é o 84.

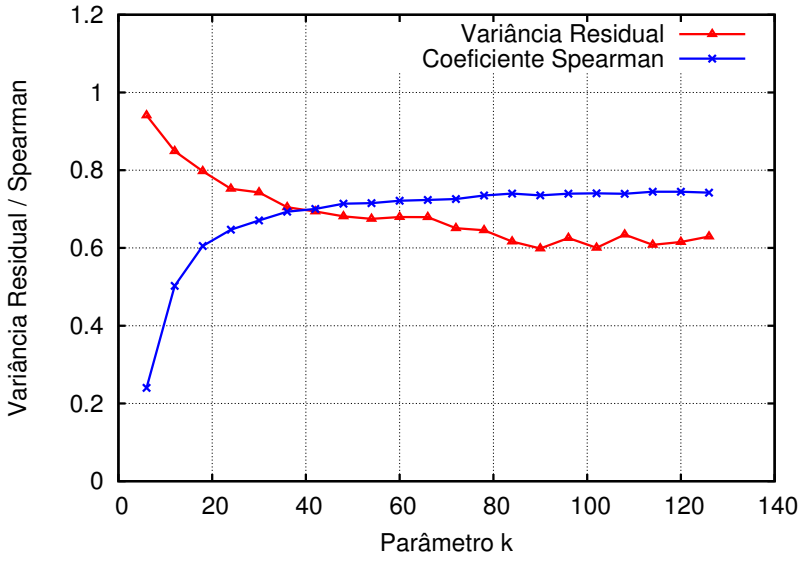

Figura 6.15: Variância residual e Coeficiende de Spearman por parâmetro k na amostra de filmes/TV com cosseno $\geq 2 \times 10^{-3}$ através da técnica LLE.

6.3.4 Parâmetros relacionados ao LINE

Para esta técnica foram utilizados os valores padrão para os parâmetros de taxa de aprendizado $(0,025)$ e para o número de amostras negativas (5) conforme apresentado em Tang et al. [2015]. Já para o grafo de entrada para o treinamento do método foram utilizadas 100% das arestas.

Assim como discutido na seção 5.3, o LINE possui três configurações: (1) considerar apenas os vizinhos de primeira ordem, (2) considerar apenas os vizinhos de segunda ordem e (3) combinar os dois anteriores, considerando as duas ordens de vizinhos. Estas configurações são definidas neste trabalho como LINE-1, LINE-2 e LINE-Ambos, respectivamente.

As Figuras 6.16 e 6.17 contêm a similaridade média entre os itens dos espaços de similaridades gerados a partir das configurações do LINE. Com exceção do LINE-2 na amostra de músicas, os demais resultados apresentam resultados próximos. Como a amostra de música é mais esparsa que a de filmes, os resultados do LINE-2 tem um desempenho inferior nesta amostra, corroborando os resultados em Tang et al. [2015], que indica que o LINE de segunda ordem de proximidade tem problemas quando o grafo é esparso.

O objetivo do LINE de segunda ordem é manter próximos pontos que compartilham os mesmos vizinhos, portanto quanto maior essa informação de vizinhos, desde que eles não representem a adição de itens distintos nas vizinhanças, mais a função objetivo tem dados para o aprendizado do gradiente. Este resultado prejudica o resultado do LINE-Ambos, já que ele é a união entre LINE-1 e LINE-2. No entanto, para as comparações entre os métodos foi utilizado o Line-Ambos, dado que é a técnica mais genérica e apresenta resultados muito próximos, considerando a métrica de similaridade.

6.4 Comparações das técnicas

Grafo de entrada: As Figuras 6.18 e 6.19 exibem a variância residual por método analisado nas diferentes amostras. Os métodos IsoMap e LLE tem uma grande influência em seus resultados dependendo da forma em que foi feita a remoção de arestas. A política que propõe remover arestas dado um limite mínimo de cosseno faz com que o resultado desses métodos seja melhor do que considerando o grafo inteiro ou o limite mínimo por coocorrência.

O príncipio geral utilizado pelo IsoMap e pelo LLE é utilizar uma possível linearidade local entre os dados para obter informações da não-linearidade global de seu

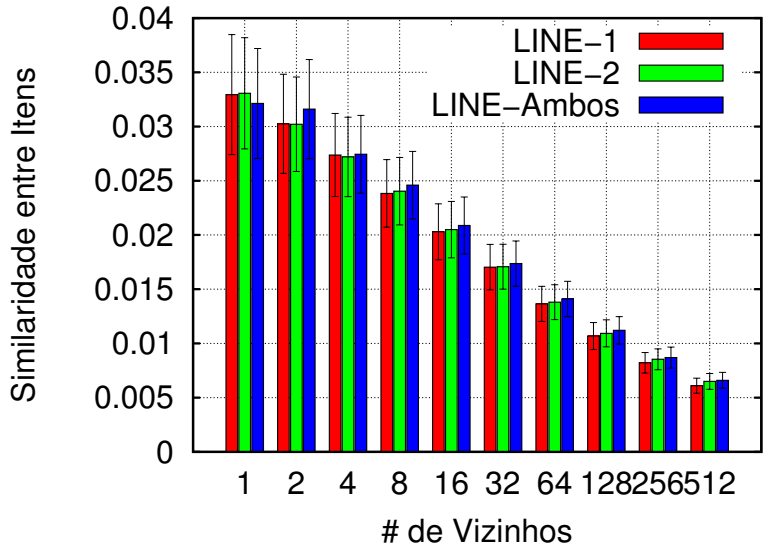

Figura 6.16: Similaridade média entre itens com a técnica LINE considerando primeira ordem, segunda ordem e ambas na amostra de filmes/TV com cosseno $\geq 2 \times 10^{-3}$.

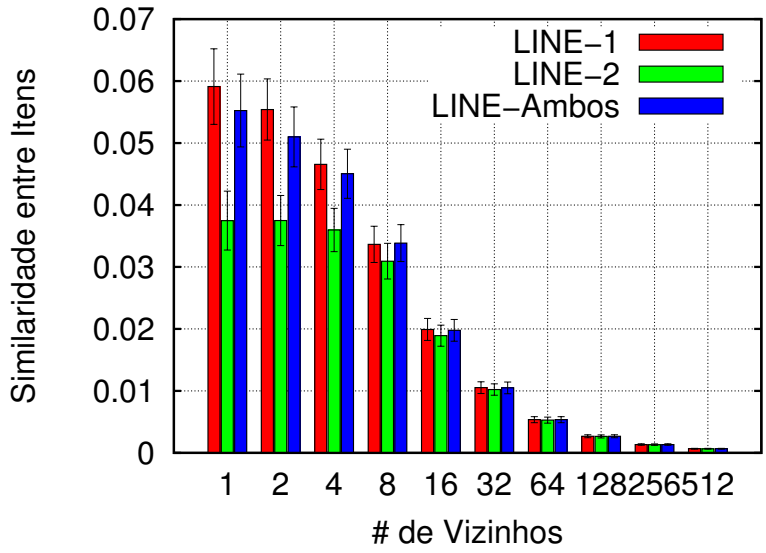

Figura 6.17: Similaridade média entre itens com a técnica LINE considerando primeira ordem, segunda ordem e ambas na amostra de músicas com coocorrências ≥ 6.

manifold. Quando considerados grafos densos como o grafo inicial G_I, principalmente o de filmes e programas de televisão, o desempenho de ambos os métodos são inferiores à de grafos mais esparsos. Este efeito provavelmente ocorre devido ao fato de que grafos com alta conectividade possuem arestas que "encurtam" os caminhos mínimos entre os vértices e, portanto, o príncipio da linearidade local pode ser quebrado ao se conside-

Tabela 6.4: Amostra dos grafos selecionados por técnica.

Técnica	Grafo de filmes e programas de TV	Grafo de músicas[2]
IsoMap	Limitado por cosseno $\geq 2 \times 10^{-3}$	Limitado por cosseno $\geq 2 \times 10^{-3}$
LLE	Limitado por cosseno $\geq 2 \times 10^{-3}$	-
LINE	Grafo inicial	Grafo inicial

rar como próximos, vértices que deveriam estar distantes no *manifold* não-linear como próximos.

O método LINE apresentou o comportamento esperado, assim como discutido em Tang et al. [2015] e na subseção 6.3.4, obtendo melhores resultados considerando grafos mais densos (grafo de filmes e programas de televisão).

As demais comparações são feitas considerando o grafo limitado por cosseno para as técnicas IsoMap e LLE, e o grafo inicial para a técnica LINE. A Tabela 6.4 apresenta as amostras utilizadas em cada uma das técnicas analisadas. Para o grafo de músicas foi utilizada a técnica IsoMap com 10% de *landmarks*, já que não foi possível utilizar a técnica IsoMap completa por limitações computacionais, assim como a técnica LLE para essa mesma base de dados, como discutido na Seção 6.1.

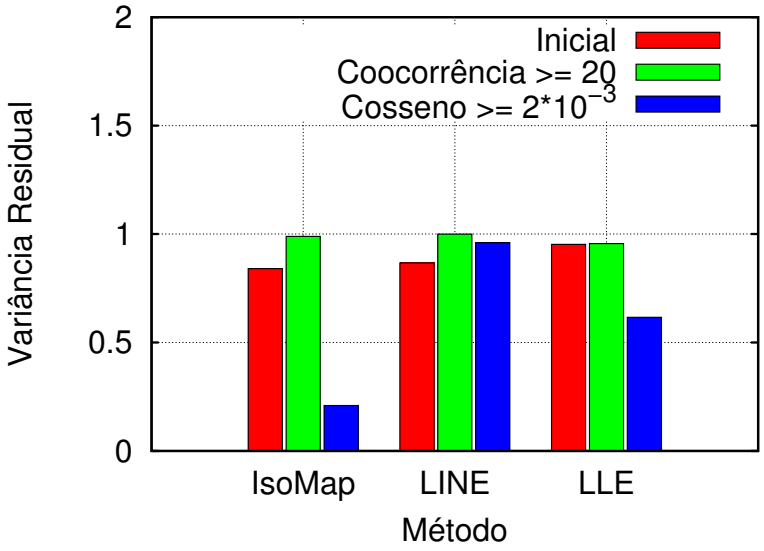

Figura 6.18: Variância residual por método nas amostras de filmes/TV.

Variância residual: A variância residual, exibida na Figura 6.20, mostra que, dentre os métodos e configurações observadas, o método que melhor manteve as dis-

[2]Assim como destacado, foi utilizado o 10%-IsoMap (10% de *landmarks*) na amostra de músicas devido a limitações computacionais.

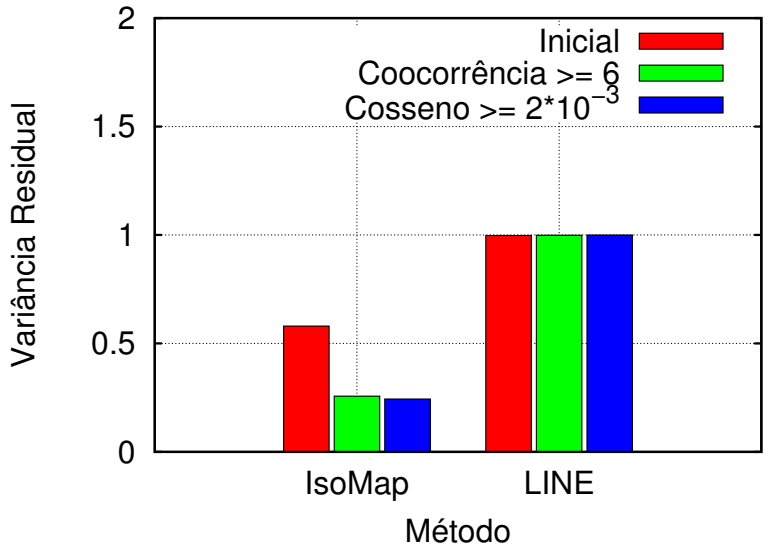

Figura 6.19: Variância residual por método nas amostras de músicas.

tâncias iniciais linearmente foi o IsoMap para as duas bases de dados. O método LINE não foi capaz de preservar as distâncias globalmente (para todos os pontos).

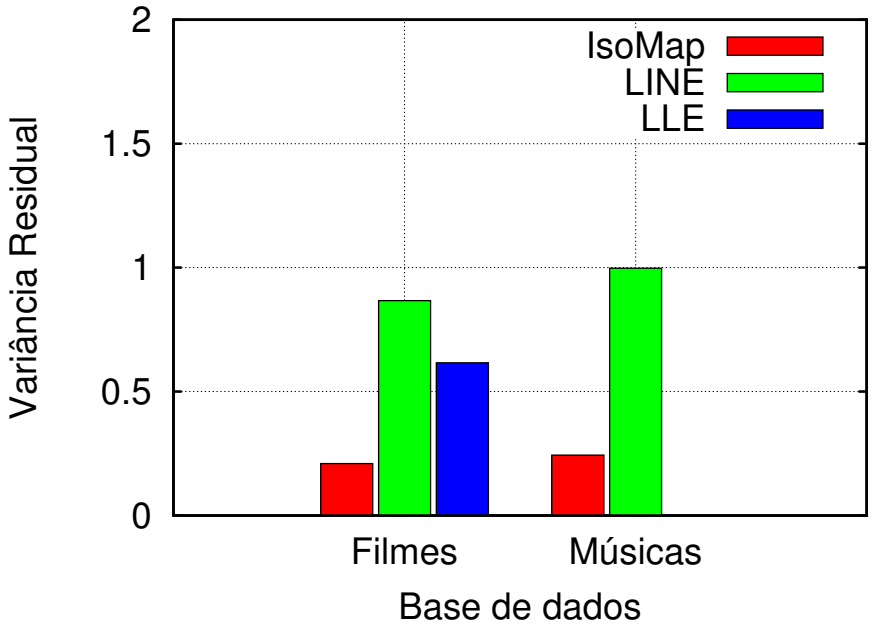

Figura 6.20: Variância residual por método e amostra.

Ordem global dos vizinhos: O coeficiente de Spearman representa se a ordem dos vizinhos é preservada globalmente, sem considerar a magnitude das distâncias em

seu cálculo. A Figura 6.21 contém o coeficiente de Spearman por base de dados utilizada e método. Assim como na variância residual os melhores resultados são apresentados pelo IsoMap, embora o LLE seja capaz de manter a ordem global próxima do IsoMap e o LINE tenha um valor próximo de 0, o que indica uma correlação pequena entre os vizinhos ordenados por distância do grafo inicial e do espaço gerado de maneira global.

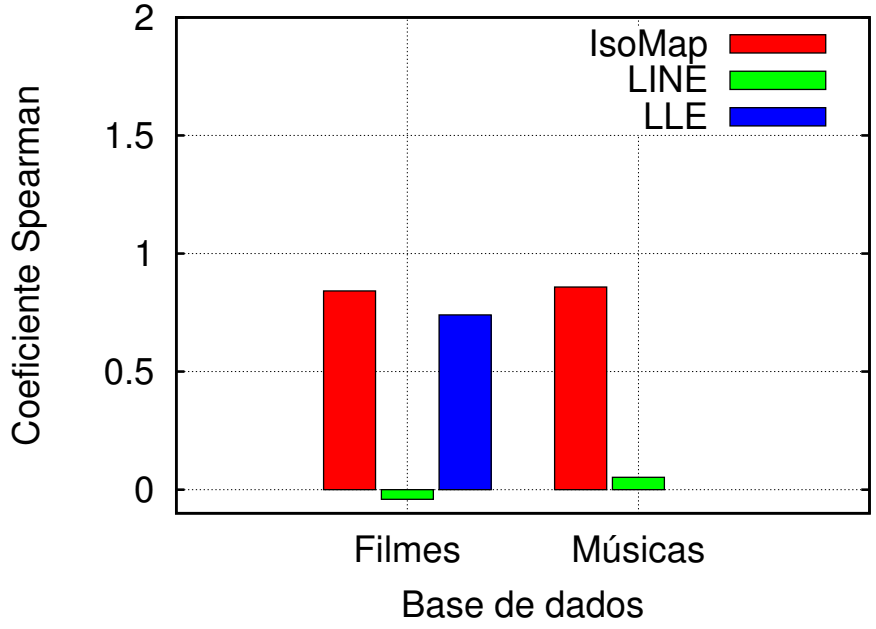

Figura 6.21: Coeficiente de Spearman por método e amostra.

Similaridade entre vizinhos: Dentre as métricas de similaridade entre vizinhos, a Figura 6.22 contém a similaridade entre os gêneros de pontos vizinhos para a amostra de filmes e programas de televisão. Todos os métodos observados são bem similares nesta métrica. A queda de acordo com o aumento da quantidade k de vizinhos considerada é esperada, demonstrando que os gêneros mais similares estão mais próximos e os diferentes mais distantes.

As Figuras 6.23 e 6.24 contêm a similaridade média, ou seja a média dos cossenos, entre os k vizinhos de cada item. Através desta análise é possível observar que o método que melhor preserva a similaridade local entre os itens é o LINE na base de dados de música e é muito similar ao IsoMap na base de dados de filmes e programas de TV. Este resultado indica que embora o método LINE não seja capaz de preservar as distâncias e ordem dos vizinhos globalmente, ele é o que melhor preservou as similaridades locais entre os pontos.

Análise manual de vizinhanças locais: Para a análise manual de vizinhanças foram escolhidos 4 programas de TV e 1 filme: 2 séries de TV, 1 filme de ação e 2

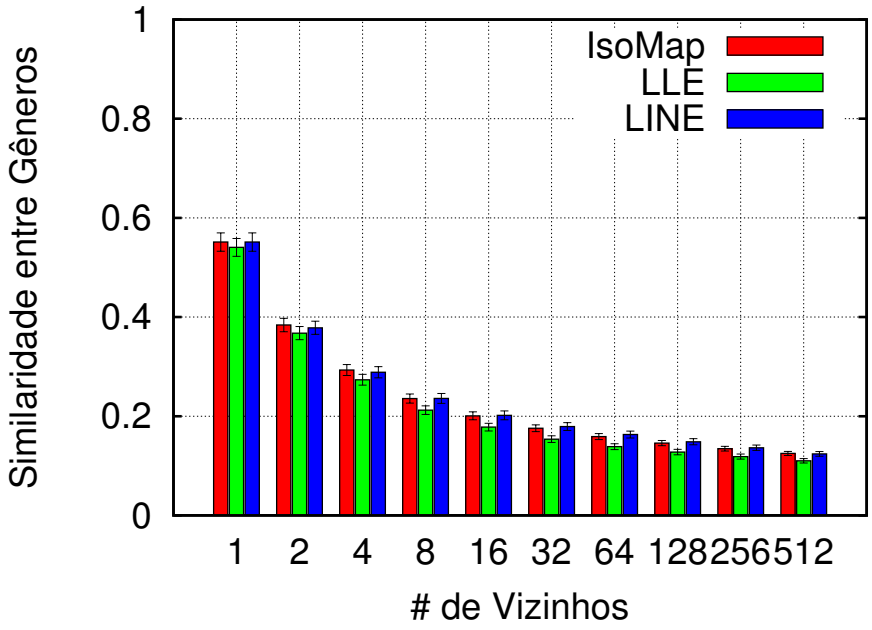

Figura 6.22: Similaridade entre gêneros de vizinhos por método na amostra de filmes/TV.

reality shows utilizando o método IsoMap com a base de dados de filmes e programas de TV:

(a) Série de Comédia - Parks and Recreation: A Tabela 6.5 mostra as vizinhanças da série para 2, 5 e 10 dimensões. Para a dimensão 2, alguns filmes de animação encontram-se na vizinhança, como Ice Age e Shrek (note que ambos pertencem ao mesmo gênero). Entretanto, ao aumentarmos as dimensões, quase toda a vizinhança é formada por séries de comédia, como It's Always Sunny in Philadelphia, 30 Rock, Weeds e Community. Esses títulos são mais representativos do público de Parks and Recreation em comparação aos dos relacionados considerando as outras dimensões, dado que as dimensões menores possuem alguns ruídos como Star Trek, um filme de ficção científica, ou Super 8, um filme de suspense produzido por Spielberg.

(b) Série Teen - Vampire Diaries: Os resultados para a análise da vizinhança desta série é mostrada na Tabela 6.6. Podemos observar que, para a dimensão 2, aparecem várias séries e filmes mais gerais, como Grey's Anatomy e CSI, enquanto, à medida que as dimensões aumentam, itens mais relacionados com a temática da série Vampire Diaries fazem parte da vizinhança (p.e, Twilight é um filme adolescente sobre Vampiros).

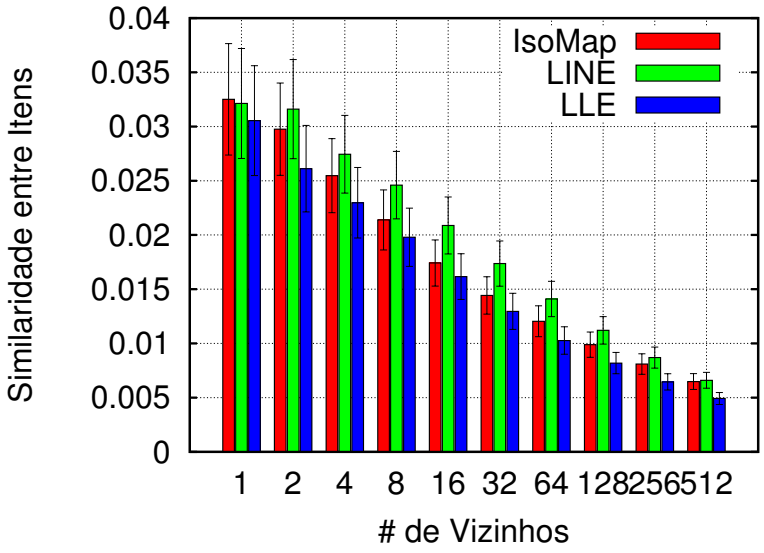

Figura 6.23: Similaridade média entre itens na amostra de filmes/TV.

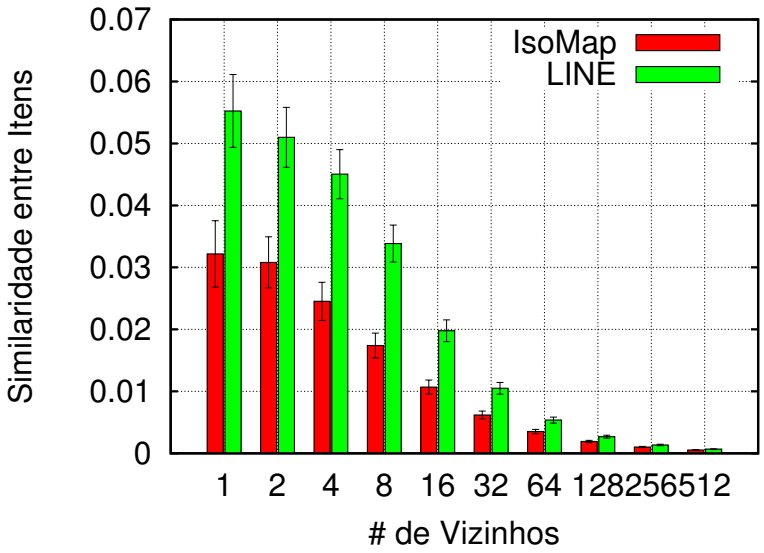

Figura 6.24: Similaridade média entre itens na amostra de músicas.

Vampire Diaries		
2D	5D	10D
Hells Kitchen	Twilight Saga part 1	Twilight Saga part 1
Sex City	Smallville	Twilight
CSI Crime SceneI nvestigation	Twilight	Greys Anatomy
Greys Anatomy	CSI NY	Xfactor
White Collar	notebook	csi_miami
tangled	CSI Miami	Pretty Little Liars
Notebook	Smurfs	Pirates of Caribbean on Stranger Tides
21 Jump Street	Tangled	CSI NY
Mentalist	Greys Anatomy	Smurfs

Tabela 6.6: Vizinhança: Vampire Diaries.

Parks and Recreation		
2D	5D	10D
Ice Age	Cosby Show	Its always sunny in philadelphia
Californication	Conan	Sons of Anarchy
Super 8	Wonder Years	Weeds
Daria	Its Always Sunny in Philadelphia	Community
Knocked up	Golden Girls	Conan
Golden Girls	Boardwalk Empire	Curb your Enthusiasm
Angel	Frasier	Anthony Bourdain no reservations
Its Always Sunny in Philadelphia	I Love Lucy	30 Rock
Shrek Third	Star Trek next Generation	Breaking Bad

Tabela 6.5: Vizinhança: Parks and Recreation.

(c) Filme de Ação - Duro de Matar: A Tabela 6.7 mostra a mudança da vizinhança com o aumento das dimensões do espaço Euclidiano. Podemos observar que, em duas e cinco dimensões, sua vizinhança está em uma região com alguns filmes de ficção científica, como Alien, District 9 e Back to the Future. Em 10 dimensões, os vizinhos mais próximos são Quinto Elemento, que tem inclusive o mesmo ator (Bruce Willis) como protagonista, seguido por Exterminador do Futuro, Batman e Indiana Jones, que são filmes de ação/aventura bem mais relacionados entre si.

Duro de Matar		
2D	5D	10D
American Beauty	Terminator	Fifth Element
Tron	Seven	Terminator
Terminator	Terminator 2 (judgment day)	Terminator 2 (judgment day)
Back to the Future 3	Back to the Future 2	Batman
V or Vendetta	Alien	Indiana Jones Raiders of Lost Ark
Back to the Future 2	American Psycho	Indiana Jones Last Crusade
Terminator 2 (judgment day)	Fifth Element	Back to the Future 2
Fifth Element	District 9	Indiana Jones Temple of Doom
Silence of Lambs	Batman	Alien

Tabela 6.7: Vizinhança: Duro de Matar.

(d) Reality Show de Competição de Moda - Project Runway: Observando a vizinhança deste programa, conforme mostrada na Tabela 6.8, notamos que para 2 dimensões, a vizinhança tem poucos programas relacionados: os únicos *reality shows* são Top Chef, que, de fato, é uma competição e compartilha o público de Project Runway, e Pawn Stars, um *reality show* sobre uma loja de penhores em Las Vegas. No entanto, para 10 dimensões, além do Top Chef ser o ponto mais próximo no mapa, outros *reality shows* relacionados aparecem, como So you think you can dance, Amazing Race e Americas' next Top Model. Este resultado mostra que o aumento das dimensões incluiu o programa num nicho que, segundo nossa especialista, atende a um público

específico e bem definido.

Project Runway		
2D	5D	10D
Hey Arnold	Sweet Home Alabama	Top Chef
Brave	White Collar	Sex City
Crazy Stupid Love	Help	So you think you can dance
Top Chef	Cold case	Americas Next Top Model
Men in Black 3	Cougar Town	Late Night with Jimmy Fallon
Hawaii five0	Top Chef	Raising Hope
Pawn Stars	Private Practice	Beverly Hills 90210
Blind Side	Happy Endings	Amazing Race
Friends with Benefits	So you think you can dance	Masterchef

Tabela 6.8: Vizinhança: Project Runway.

(e) Reality Show de Competição de Artistas - American Idol: Analisando a vizinhança deste programa, mostrada na Tabela 6.9, podemos observar que, em 2 dimensões, apesar de o ponto mais próximo ser o The Voice, que é um programa muito parecido com American Idol, os outros programas não tem muito em comum além, talvez, do público alvo. Aumentando o número de dimensões, aparecem cada vez mais *reality shows*, como Dancing with the Stars, X Factor e America's Got Talent, e a vizinhança passa a fazer mais sentido tanto em público alvo quanto em nicho de mercado.

American Idol		
2D	5D	10D
The voice	The voice	The voice
Snow white Huntsman	Revenge	Dancing with Stars
2 broke girls	CSI NY	Ellen Degeneres Show
Twilight Saga part 1	Xfactor	Twilight Saga part 1
puss_in_boots	Twilight Saga part 1	Pretty Little Liars
Twilight	CSI Miami	Xfactor
Desperate Housewives	New Girl	Greys Anatomy
Wipeout	NCIS los Angeles	Americas got Talent
NCIS los Angeles	CSI Crime Scene Investigation	Vampire Diaries

Tabela 6.9: Vizinhança: American Idol.

Em suma, através da análise de vizinhanças locais dos itens apresentados é possível observar que as vizinhanças em 2 dimensões se mostraram muito genéricas. No entanto, à medida que o número de dimensões aumenta, as vizinhanças passam a fazer mais sentido, entrando cada vez mais nos nichos específicos de público que cada título atrai. Em muitos dos casos analisados em 2 dimensões, foi possível encontrar alguma relação superficial entre títulos que poderiam ser interpretados como ruído de sobreposição (ex. gênero ou época de lançamento semelhantes).

6.5 Resumo do capítulo

Neste capítulo os métodos de *embedding* de grafos são comparados de acordo com amostras e métricas distintas. Dentre os principais resultados analisados pode ser destacada a influência que a estratégia de remoção de arestas exerceu principalmente nos métodos IsoMap e LLE. A remoção de arestas menores que um *threshold* (limite) pré-definido, embora seja feita em outros trabalhos como [Goussevskaia et al., 2008], é utilizada apenas para fins de redução do custo computacional dos métodos.

Além de diminuir o custo computacional, que está diretamente relacionado ao número de arestas do grafo, esta dissertação demonstrou que essa remoção também tem impacto direto na qualidade do espaço gerado através dos métodos IsoMap e LLE. Portanto, o pré-processamento dos grafos iniciais representa um passo importante nesses métodos. Já o método LINE sofreu menos alterações com essa redução de arestas, o que indica que provavelmente seja menos suscetível a ruídos na amostra.

A técnica IsoMap foi a que melhor preservou linearmente as distâncias e as vizinhanças globais, embora localmente a técnica LINE foi capaz de melhor preservar a similaridade entre os itens. Este resultado indica que, caso o espaço de similaridades seja utilizado para navegações ou buscas locais, baseadas na proximidade, a técnica LINE pode ser mais indicada. Um exemplo de navegação local é apresentado em [Cardoso et al., 2016], os autores indicam com maior probabilidade os itens mais próximos no espaço de similaridades.

Caso o espaço de similaridades seja utilizado para análises globais, como verificar qual música está entre outras duas mais distantes, ou qual "caminho" pode ser seguido entre dois filmes, provavelmente o método IsoMap tenha melhores resultados.

O método LLE apresentou preservou as distâncias e as vizinhanças globais melhor que os demais caso sejam consideradas poucas dimensões. Trabalhos futuros devem ser realizados para analisar métodos que melhor representem visualizações (três dimensões ou menos) para esses espaços. Essas visualizações podem permitir outros avanços, não só na navegação em coleções de mídia, mas em outras análises.

Capítulo 7

Conclusões e trabalhos futuros

Nesta dissertação foram analisadas redes sociais online focadas no compartilhamento e uso de mídias, uma rede sobre filmes e programas de televisão (tvtag) e outra sobre músicas (Last.fm). O estudo sobre a rede social tvtag permite um maior entendimento de fenômenos como a TV social e o comportamento *second-screen*. Acreditamos que este trabalho, em conjunto com [Holanda et al., 2015b] são os primeiros passos para compreender essa rede social e esses comportamentos em uma rede social focada especificamente neste tipo de interação.

Trabalhos futuros que considerem outras redes sociais online neste mesmo escopo podem ser considerados, como a análise da rede social telfie.com. Os métodos de predição também devem ser evoluídos, comparando outros dados ou mesmo utilizando dados de audiência final da vida real verificando a real audiência de cada programa e relação disso com as interações nas redes sociais.

O segundo passo deste trabalho foi avaliar formas de derivar similaridades entre músicas ou filmes e programas de televisão através de dados de usuários nas redes sociais analisadas. Este conjunto de dados é então representado por um grafo de similaridades. Entretanto, o objetivo final deste trabalho é transformar este grafo em um espaço Euclidiano com o uso de técnicas de *embedding* de grafos.

O espaço Euclidiano de similaridades apresenta uma série de vantagens em relação a um grafo. A similaridade entre dois itens é calculada em tempo linear, com a distância entre esses pontos. Outras funcionalidades que este espaço possibilita são: trajetórias, volumes e a noção de direção, que podem ser exploradas em aplicações, como feito em [Cardoso et al., 2016].

Trajetórias e direções, por exemplo, permitem "navegar" suavemente entre músicas, filmes e programas de televisão. Este tipo de propriedade pode ser explorada para criar novas formas de navegação e visualização de coleções de mídia.

Para a criação deste espaço de similaridades são detalhadas técnicas de *embedding* de grafos, elas são comparadas de acordo com um conjunto de métricas, algumas propostas nesta dissertação.

Outra importante contribuição deste trabalho é na avaliação de como deve ser construído o grafo para a entrada de cada um dos métodos de *embedding*, o que demonstrou um impacto direto na qualidade do espaço gerado. Além disso, a escolha do método de *embedding* utilizado depende da aplicação do espaço de similaridades, caso ele seja utilizado para comparações globais entre itens provavelmente a técnica IsoMap seja uma melhor escolha. Caso o espaço seja utilizado para navegações locais a técnica LINE obteve melhores resultados.

Este trabalho apresenta uma série de direções nas quais pode evoluir. A construção e remoção de arestas do grafo de similaridades pode ser melhor explorada, com estratégias capazes de extrair o grafo que melhor se comporte em cada técnica de *embedding*. Trabalhos futuros envolvendo outras medidas de similaridade, como o conteúdo do áudio, metadados ou combinações de atributos, também devem ser melhor explorados.

Existe também uma lacuna dentre os métodos de *embedding* de grafos existentes, pois existem poucos métodos que focam em realizar o *embedding* de grafos do mundo real, com grande quantidade de vértices de maneira a preservar as distâncias iniciais desses grafos. Além disso os métodos explorados envolvem apenas em fazer o *embedding* da componente gigante deste grafo, uma forma de inserir vértices fora dessa componente pode ser utilizar outros atributos para realizar uma triangulação desses vértices no espaço gerado.

Finalmente, métodos de visualização e navegação através deste espaço podem ser propostos utilizando sua propriedades. Como este é um espaço de baixa dimensionalidade é possível propor, por exemplo, uma aplicação cliente-servidor em que o cliente precisa apenas das coordenadas de suas músicas, o que pode ser armazenado facilmente em um smartphone, para utilizar estas técnicas de navegação e visualização mesmo que ele esteja *offline*.

Referências Bibliográficas

Aucouturier, J. & Pachet, F. (2002). Scaling up Music Playlist Generation. Em *ICME*.

Backstrom, L.; Boldi, P.; Rosa, M.; Ugander, J. & Vigna, S. (2012). Four degrees of separation. Em *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pp. 33--42, New York, NY, USA. ACM.

Backstrom, L. & Kleinberg, J. (2014). Romantic partnerships and the dispersion of social ties: A network analysis of relationship status on facebook. Em *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pp. 831--841, New York, NY, USA. ACM.

Basapur, S.; Mandalia, H.; Chaysinh, S.; Lee, Y.; Venkitaraman, N. & Metcalf, C. (2012). Fanfeeds: Evaluation of socially generated information feed on second screen as a tv show companion. Em *Proceedings of the 10th European Conference on Interactive Tv and Video*, EuroiTV '12, pp. 87--96, New York, NY, USA. ACM.

Belkin, M. & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373--1396. ISSN 0899-7667.

Bondad-Brown, B. A.; Ricea, R. E. & Pearce, K. E. (2012). Influences on tv viewing and online user-shared video use: Demographics, generations, contextual age, media use, motivations, and audience activity. *Journal of Broadcasting & Electronic Media*, 56:471–493.

Brandes, U. & Pich, C. (2007). *Eigensolver Methods for Progressive Multidimensional Scaling of Large Data*, pp. 42--53. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cardoso, J. P. V.; Pontello, L. F.; Holanda, P. H. F.; Guilherme, B.; Goussevskaia, O. & da Silva, A. P. C. (2016). Mixtape: Using real-time user feedback to navigate large media collections. Em *International Society for Music Information Retrieval Conference, ISMIR*.

Cha, M.; Haddadi, H.; Benevenuto, F. & Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. Em *in ICWSM Õ10: Proceedings of international AAAI Conference on Weblogs and Social*.

Cox, T. F. & Cox, M. (2000). *Multidimensional Scaling*. Chapman and Hall/CRC.

David Gleich, Leonid Zhukov, M. R. & Lang, K. (2005). The World of Music: SDP layout of high dimensional data. Em *InfoVis*.

de Silva, V. & Tenenbaum, B. (2004). Sparse Multidimensional Scaling using Landmark Points. Technical Report, Stanford University.

de Silva, V. & Tenenbaum, J. B. (2002). Global versus local methods in nonlinear dimensionality reduction. Em *NIPS*.

Faloutsos, C. & Lin, K.-I. (1995). Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. *SIGMOD Rec.*, 24(2):163--174. ISSN 0163-5808.

Geerts, D.; Vaishnavi, I.; Mekuria, R.; van Deventer, O. & Cesar, P. (2011). Are we in sync?: Synchronization requirements for watching online video together. Em *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pp. 311--314, New York, NY, USA. ACM.

Gonzalez, R.; Cuevas, R.; Motamedi, R.; Rejaie, R. & Cuevas, A. (2013). Google+ or google-?: Dissecting the evolution of the new osn in its first year. Em *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pp. 483--494, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Goussevskaia, O.; Kuhn, M. & Wattenhofer, R. (2008). Exploring Music Collections on Mobile Devices. Em *MobileHCI*.

Holanda, P. H. F.; Guilherme, B.; Cardoso, J. P. V.; da Silva, A. P. C. & Goussevskaia, O. (2015a). Mapeando o universo da midia usando dados gerados por usuarios em redes sociais online. Em *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, SBRC*.

Holanda, P. H. F.; Guilherme, B.; da Silva, A. P. C. & Goussevskaia, O. (2015b). TV goes social: Characterizing user interaction in an online social network for TV fans. Em *Engineering the Web in the Big Data Era - 15th International Conference, ICWE 2015, Rotterdam, The Netherlands, June 23-26, 2015, Proceedings*, pp. 182--199.

Jimenez, L. & Langrebe, D. A. (1998). Supervised classification in high dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data. *IEEE TRANSACTIONS ON SYSTEM, MAN, AND CYBERNETICS, VOLUME 28, PART C*, 28:39--54.

Kayo, O. (2006). *Locally linear embedding algorithm: extensions and applications*. Tese de doutorado, University of Oulu.

Knees, P.; Schedl, M.; Pohle, T. & Widmer, G. (2006). An Innovative Three-Dimensional User Interface for Exploring Music Collections Enriched with Meta-Information from the Web. Em *ACM Multimedia*.

Kwak, H.; Lee, C.; Park, H. & Moon, S. (2010). What is twitter, a social network or a news media? Em *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 591--600, New York, NY, USA. ACM.

Lee, J. A. & Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*. Springer Publishing Company, Incorporated, 1st edição. ISBN 0387393501, 9780387393506.

Levy, M. & Bosteels, K. (2010). Music recommendation and the long tail. Em *1st Workshop On Music Recommendation And Discovery (WOMRAD)*.

Logan, B. (2002). Content-based playlist generation: Exploratory experiments. Em *ISMIR*.

Moore, J. L.; Chen, S.; Turnbull, D. & Joachims, T. (2013). Taste over time: The temporal dynamics of user preferences. Em *ISMIR*.

Mukherjee, P. & Jansen, B. (2014). Social tv and the social soundtrack: Significance of second screen interaction during television viewing. Em Kennedy, W.; Agarwal, N. & Yang, S., editores, *Social Computing, Behavioral-Cultural Modeling and Prediction*, volume 8393 of *Lecture Notes in Computer Science*, pp. 317–324. Springer International Publishing.

Narasimhan, N. & Vasudevan, V. (2012). Descrambling the social TV echo chamber. *Proceedings of the 1st ACM workshop on Mobile systems for computational social science - MCSS '12*, p. 33.

Neumayer, R.; Dittenbach, M. & Rauber, A. (2005). PlaySOM and PocketSOMPlayer, Alternative Interfaces to Large Music Collections. Em *ISMIR*.

Pálovics, R. & Benczúr, A. A. (2013). Temporal influence over the last.fm social network. Em *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pp. 486--493, New York, NY, USA. ACM.

Pampalk, E.; Dixon, S. & Widmer, G. (2003). On the evaluation of perceptual similarity measures for music. Em *DAFx*.

Pampalk, E.; Pohle, T. & Widmer, G. (2005a). Dynamic playlist generation based on skipping behavior. Em *ISMIR*.

Pampalk, E.; Pohle, T. & Widmer, G. (2005b). Generating similarity-based playlists using traveling salesman algorithms. Em *DAFx*.

Platt, J. (2004). Fast embedding of sparse music similarity graphs. Em *NIPS*, volume 16.

Platt, J.; Burges, C.; Swenson, S.; Weare, C. & Zheng, A. (2002). Learning a Gaussian Process Prior for Automatically Generating Music Playlists. *NIPS*, 14.

Ragno, R.; Burges, C. J. C. & Herley, C. (2005). Inferring similarity between music objects with application to playlist generation. Em *MIR*.

Recht, B.; Re, C.; Wright, S. & Niu, F. (2011). Hogwild: A lock-free approach to parallelizing stochastic gradient descent. Em Shawe-taylor, J.; Zemel, R.; Bartlett, P.; Pereira, F. & Weinberger, K., editores, *Advances in Neural Information Processing Systems 24*, pp. 693--701.

Ribeiro, B. (2014). Modeling and predicting the growth and death of membership-based website. Em *Proceedings of the 23rd International Conference on World Wide Web*.

Roweis, S. T. & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323--2326.

Schölkopf, B.; Smola, A. & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299--1319. ISSN 0899-7667.

Shaw, B. & Jebara, T. (2009). Structure preserving embedding. Em *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 937--944, New York, NY, USA. ACM.

Shlens, J. (2005). A tutorial on principal component analysis. Em *Systems Neurobiology Laboratory, Salk Institute for Biological Studies*.

Tang, J.; Liu, J.; Zhang, M. & Mei, Q. (2016). Visualizing large-scale and high-dimensional data. Em *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pp. 287--297, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J. & Mei, Q. (2015). Line: Large-scale information network embedding. Em *Proceedings of the 24th International Conference on World Wide Web*, pp. 1067--1077. International World Wide Web Conferences Steering Committee.

Tenenbaum, J. B.; Silva, V. & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319--2323.

Torgerson, W. S. (1965). Multidimensional scaling of similarity. *Psychometrika*, 30(4):379--393. ISSN 1860-0980.

Torrez-Riley, J. (2011). The social tv phenomenon: New technologies look to enhance televisionÕs role as an enabler of social interaction.

Ugander, J.; Karrer, B.; Backstrom, L. & Marlow, C. (2011). The anatomy of the facebook social graph. *CoRR*, abs/1111.4503.

Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85.

van Dongen, S. & Enright, A. J. (2012). Metric distances derived from cosine similarity and pearson and spearman correlations. `http://arxiv.org/pdf/1208.3145v1.pdf`.

Vasconcelos, M.; Almeida, J.; Gonçalves, M.; Souza, D. & Gomes, G. (2014). Popularity dynamics of foursquare micro-reviews. Em *Proceedings of the Second ACM Conference on Online Social Networks*, COSN '14, pp. 119--130, New York, NY, USA. ACM.

Zhong, C.; Salehi, M.; Shah, S.; Cobzarenco, M.; Sastry, N. & Cha, M. (2014). Social bootstrapping: How pinterest and last.fm social communities benefit by borrowing links from facebook. Em *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pp. 305--314, New York, NY, USA. ACM.