Instituto Brasileiro de Geografia e Estatística
Escola Nacional de Ciências Estatísticas
Bacharelado em Estatística

Aplicações de modelos para redes sociais em bases textuais obtidas via *Web Scraping*

Pedro Henrique Sodré Puntel

Rio de Janeiro

Pedro Henrique Sodré Puntel

Aplicações de modelos para redes sociais em bases textuais obtidas via *Web Scraping*

Orientador(a): Gustavo da Silva Ferreira

Rio de Janeiro

2019

Resumo

Este projeto de iniciação científica utiliza técnicas para a extração de dados textuais disponíveis na internet, de forma a investigar associações entre os mesmos sob a ótica das redes sociais. Para tal, dois modelos estatísticos foram considerados, sendo o primeiro responsável pela projeção de tal estrutura de relações num sistema de coordenadas de baixa dimensionalidade (Modelo de Escalonamento Multidimensional) e um último cuja finalidade foi modelar a similaridade/dissimilaridade entre os atores da rede (Modelo de Distâncias Latentes). Os resultados obtidos neste projeto proporcionaram um amplo estudo acerca das aplicações das técnicas de *Web Scraping*, bem como permitiram visualizar e interpretar relações implícitas no conjunto de dados considerado, identificando ainda claros agrupamentos dos mesmos a partir dos padrões latentes observados.

Palavras-Chave: *Web Scraping*, Redes Sociais, Billboard, Espaços Latentes, Escalonamento Multdimensional.

1 Relevância e Contextualização do Tema

A popularização das redes sociais traz consigo a necessidade de avanços na modelagem das relações entre os indivíduos que compõem uma rede. Não obstante, os objetivos das análises de redes sociais são hoje bastante diversos, abrangendo tópicos que variam desde a identificação de nós (indivíduos) com papel central, bem como a sumarização e identificação de grupos subjacentes na rede.

Sob este prisma, o avanço dos modelos estatísticos tem permitido análises a cerca de diversos aspectos de uma única rede, como por exemplo: opiniões, hábitos, desejos e graus de relacionamento (intra e entre grupos) dos atores que a compõem. Tais aspectos são muitas vezes utilizados para a personalização da oferta de produtos e serviços sendo esta uma das possíveis aplicações destes modelos.

Como consequência do avanço computacional, grandes bases de dados textuais agora encontram-se disponíveis na Internet, sendo a maioria destas encontradas de forma não-estruturada. Segundo Hoff *et al*., 2008, cerca de 90% de toda a informação contida no universo digital era composta por dados não estruturados como textos, imagens e vídeos. Neste sentido, a habilidade de coletar, estruturar e analisar parte dessa grande quantidade de dados é essencial, sendo o principal objeto de estudo durante o início deste projeto.

Para lidar com a não-estruturação das bases de dados textuais, diversas ferramentas de *Web Scraping* foram desenvolvidas nos últimos anos, permitindo novos níveis de integração dos modelos matemáticos e estatísticos para com as informações latentes disponíveis na Internet.

2 Metodologia

2.1 Web Scraping

Segundo Glez-Peña *et al*., 2013, Web Scraping pode ser definido como um processo sistemático de extração de conteúdos da Internet. Neste processo, um agente de software, um robô, imita a interação de um ser humano com as páginas e servidores da Web. Esse agente é programado para acessar tantos sites da Internet quanto necessário, analisar seus conteúdos para encontrar e extrair dados de interesse e estruturar essa informação conforme desejado.

Para Giannini *et al*., 2017, o Web Scraping pode ser implementado de três maneiras:

- Ferramentas *Ad-Hoc*: programas de uso geral, implementados em uma linguagem de programação, instruídos para navegar e extrair dados de páginas específicas.
- Automação de navegador: ferramentas que imitam interações sequenciais de um usuário através de um navegador para com uma determinada página da Internet, reproduzindo-as de forma automática.
- Ferramentas *point-and-click*: dispositivos capazes de detectar as partes de uma página que contém os dados automaticamente.

As ferramentas *Ad-Hoc* e de automação de navegador constituem uma abordagem mais geral, podendo produzir melhores resultados, uma vez que as propriedades das linguagens de programação muitas vezes auxiliam na coleta e limpeza dos dados. Contudo, é importante mencionar que ambas estão sujeitas a invariável volatilidade de uma página da web, de forma que uma simples alteração na sua estruturação pode comprometer todo o processo.

Para tarefas mais simples e pontuais, as ferramentas *point-and-click* são bastante eficientes pois requerem do usuário um menor conhecimento sobre a estruturação de página. Nas abordagens anteriores, a compreensão de linguagens como *HTML* (*HyperText Markup Language*) e *XML (Extensible Markup Language)* bem como a de ferramentas mais avançadas como expressões regulares e *XPath (XML Path Language)* tornam-se indispensáveis, conforme apontado por Munzert *et al*., 2014.

Entretanto, em alguns casos, o uso de Web Scraping não é necessário para a coleta dos dados. É prática comum de muitas empresas disponibilizarem *APIs (Application Programming Interfaces)*, muitas vezes de forma gratuita, para que usuários ou programas possam acessar seus bancos de dados e informações de maneira muito mais simples e rápida. Porém, frequentemente, existe um limite de chamadas que podem ser feitas com essas *APIs* gratuitas para evitar um congestionamento destes serviços.

2.2 Modelos para Redes Sociais

Os modelos para redes sociais procuram, em geral, identificar nós (ou grupos de nós), de certa forma semelhantes, a partir de uma análise da probabilidade da existência de relações entre eles. Dentre os trabalhos recentes desenvolvidos na literatura, destacam-se àqueles que pressupõem a existência de *espaços latentes sociais* (Hoff *et al*., 2002).

De uma forma geral, os dados para estes modelos costumam ser organizados em uma matriz **Y** simétrica, de dimensão $n \times n$, cujas entradas quantificam a relação entre os indivíduos i e j da rede. Tais entradas podem, em uma modelagem mais simples, sinalizar somente a existência ou inexistência da relação entre os atores i e j (distribuição Bernoulli), ou em cenário mais complexo, quantificar o nível desta relação (distribuição Poisson).

Não obstante, existem ainda modelos que incorporam também a direção da relação, isto é, permitem incorporar ao modelo os casos de relações sociais ”assimétricas”, onde um indivíduo considera outro indivíduo como parte de sua rede social sem que esta consideração seja necessariamente recíproca (Hoff *et al.*, 2002).

Dentre os modelos baseados na existência de espaços latentes, os principais são o Modelo de Efeitos Aleatórios, o Modelo de Classes Latentes (Nowicki e Snijders 2001, Airoldi *et al*. 2008) e o Modelo de Distâncias Latentes (Hoff *et al*., 2002 e Handcock *et al*., 2007). Nesta iniciação, o modelo implementado é o de Distâncias Latentes (Hoff *et al*., 2002 e Handcock *et al*., 2007).

2.2.1 Modelo de Distâncias Latentes

O Modelo de Distâncias Latentes (Hoff *et al*., 2002 e Handcock *et al*., 2007) considera que cada um dos n elementos da rede assume uma posição em um espaço latente definido

em $\Re^d$, onde geralmente $d = 1, 2$ ou 3.

Neste modelo, tradicionalmente, interpreta-se a existência (ou não) de uma relação entre os elementos da rede a partir de uma distribuição Bernoulli. Entretanto, o caráter dicotômico da distribuição Bernoulli, conforme será visto posteriormente, não se aplica ao contexto da nossa base de dados. Destarte, as relações entre os indivíduos i e j, no contexto da nossa rede, seriam melhor captadas assumindo-se uma distribuição Poisson com taxa λ_{ij}, onde $\lambda_{ij} \in [0, 1, 2, ..., \infty)$.

O modelo assume ainda que as probabilidades de uma conexão entre os elementos i e j dependem da distância entre os mesmos no referido espaço latente. Sendo assim, formalmente, o modelo fica então especificado da seguinte forma:

$$[Y_{i,j} \mid \lambda_{i,j}] \sim Pois(\lambda_{i,j})$$

$$ln(\lambda_{i,j}) = \theta - |a_i - a_j|$$

onde $i < j$. Os efeitos latentes $\mathbf{a} = (a_1, ..., a_n)'$ representam as localizações ou coordenadas que os diferentes elementos da rede ocupam no espaço latente $\Re^d$. Assim, elementos que ocupam posições (ou localizações) próximas no espaço latente apresentam maior probabilidade de compartilharem uma relação em comparação com elementos cujas posições (ou localizações) no espaço latente não são próximas.

As localizações no espaço latente associadas a cada um dos elementos da rede social são parâmetros a serem estimados pelo modelo e podem ser definidos para um número arbitrário de dimensões. Por razões de parcimônia e a fim de se produzir resultados mais facilmente interpretáveis, baixas dimensões costumam ser escolhidas (Hoff *et al*., 2002). Tais localizações (coordenadas), são obtidas mediante um segundo modelo denominado Escalonamento Multidimensional (subseção 2.2.2).

Apesar desta flexibilidade aparente, em geral boas representações são obtidas em espaços latentes de até duas dimensões. Além de também permitir a realização de uma análise de agrupamentos, identificados pela proximidade espacial, este modelo fornece uma excelente visão gráfica dos padrões de relacionamento da rede social.

Finalmente, o processo de inferência para este modelo utiliza a abordagem bayesiana, a partir dos métodos de Máxima Verossimilhança e o Método de Monte Carlo via Cadeias de Markov (MCMC) (Metropolis *et al*., 1953 e Hastings, 1970), os quais também serão descritos em mais detalhes nas próximas subseções.

2.2.2 Modelo de Escalonamento Multidimensional

De uma forma geral, o objetivo do Escalonamento Multidimensional é projetar um conjunto de dados num sistema de coordenadas de baixa dimensão de tal forma que toda a distorção causada pela redução de dimensionalidade seja mínima. Essa distorção geralmente se refere às similaridades ou dissimilaridades (distâncias) entre os pontos de dados originais.

Nesse sentido, a credibilidade deste método sustenta-se no fato de que a representação gráfica, simplificada, de dados multivariados, possibilita uma inspeção visual direta e relevante acerca dos dados. Tal inspeção contribui positivamente as interpretações (Johnson e Wichern, 2007).

Formalmente, dado um conjunto de similaridades observadas (ou distâncias) entre cada par de N itens, deve-se encontrar uma disposição dos mesmos em alguma dimensão *q* (onde $q \leq N-1$) tal que as proximidades entre os itens de *q* fiquem o mais próximo possível das proximidades (ou distâncias) originais (Johnson e Wichern, 2007).

Para N itens, existem $M = N(N-1)/2$ similaridades (distâncias) entre pares de itens diferentes. Estas similaridades constituem os dados básicos e podem ser organizadas em uma ordem estritamente ascendente como: $s_{i_1,k_1} < s_{i_2,k_2} < ... < s_{i_M,k_M}$, onde s_{i_1,k_1} é a menor das M similaridades. O subscrito i_1, k_1 indica o par de itens que são menos similares. Os outros índices são interpretados da mesma maneira (Johnson e Wichern, 2007).

Quando se deseja organizar estes N itens em um sistema de baixa dimensão de coordenadas utilizando apenas os postos das $N(N-1)/2$ similaridades originais (distâncias), e não as suas magnitudes, o processo é chamado de *Escalonamento Multidimensional não métrico*. Se as grandezas reais das similaridades originais (distâncias) são utilizadas para obter uma representação geométrica em q dimensões, o processo é chamado de *Escalonamento Multidimensional métrico*. O *Escalonamento Multidimensional métrico* também é conhecido como *Análise de Coordenadas Principais* (Johnson e Wichern, 2007) e também como *Escalonamento Multidimensional Clássico*.

Na forma de um algoritmo, o Escalonamento Multidimensional inicia com uma matriz de distâncias **D** com elementos d_{ij}, onde $i, j = 1, ...N$, e o objetivo é encontrar uma configuração de pontos no espaço p-dimensional, a partir das distâncias entre os pontos, de tal forma que as coordenadas dos n pontos, ao longo da dimensão p, produza uma matriz de distâncias Euclidianas, cujos elementos estejam tão próximos quanto possível dos elementos da matriz de distâncias **D**. No método clássico, a matriz de dissimilaridades é assumida como uma matriz de distâncias (Cardoso-Junior e Scarpel, 2013).

As coordenadas principais, usando as distâncias Euclidianas como dissimilaridades, são obtidas a partir dos seguintes passos segundo Rencher (2002):

- A partir da matriz de dados, calcular as distâncias Euclidianas entre as linhas.
- A partir da matriz de distâncias Euclidianas, construir a matriz A, da seguinte maneira: $a_{ij} = -1/2d_{ij}^2$, onde d_{ij}^2 representa o quadrado da distância euclidiana entre as similaridades dos itens i e j.
- Construir a matriz B, da seguinte maneira: $b_{ij} = a_{ij} - \overline{a}_{i.} - \overline{a}_{.j} + \overline{a}_{..}$, onde $\overline{a}_{i.} = \sum_{j=1}^{N} a_{ij}/N$, $\overline{a}_{.j} = \sum_{i=1}^{N} a_{ij}/N$ e $\overline{a}_{..} = \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}/N^2$.
- Selecionar os p autovalores positivos da matriz B, de modo que $\lambda_1 \geq ... \geq \lambda_p > \lambda_{p+1} = ... = \lambda_N = 0$, e seus autovetores associados $\mathbf{V} = [\mathbf{V}_1, ..., \mathbf{V}_p]$. Obtém-se, então, em p dimensões, a solução das coordenadas principais $\mathbf{X}_i = \mathbf{V}_i\sqrt{\lambda_i}$, onde $i = 1, ..., p$ e cada $\mathbf{X}_i$ é um vetor de dimensão $N \times 1$.
- Caso p seja grande, de pouco interesse prático e se queira obter uma representação em $q < p$ dimensões, basta selecionar os q primeiros autovalores, de modo que

$\lambda_1 \geq ... \geq \lambda_q \geq ... \geq \lambda_p > \lambda_{p+1} = ... = \lambda_N = 0$. Obtém-se, em q dimensões, a solução das coordenadas principais $\mathbf{X}_i = \mathbf{V}_i\sqrt{\lambda_i}$, onde $i = 1, ..., q$ e cada $\mathbf{X}_i$ é um vetor de dimensão $N \times 1$.

Por fim, uma característica interessante do Escalonamento Multidimensional é que as soluções para diferentes dimensões estão agrupadas, isto é, as duas primeiras dimensões de uma solução com 3 dimensões são iguais à solução de duas dimensões. O número de dimensões desejado normalmente é o mais baixo possível de forma a proporcionar interpretações práticas (Cardoso-Junior e Scarpel, 2013).

2.2.3 Máxima Verossimilhança

Para estimação dos parâmetros (localizações) desconhecidos do Modelo de Distâncias Latentes, um dos métodos tradicionais implementados foi o da Estimação por Máxima Verossimilhança (EMV). De forma geral, o objetivo de uma EMV é encontrar valores aos parâmetros desconhecidos que maximizem uma função objetivo especifica, denominada *função de verossimilhança*.

A *função de verossimilhança* contempla toda a informação de uma amostra **x**, de um vetor aleatório **X**, associado a um parâmetro θ. Essa função pode ser expressa por $p(\mathbf{x} \mid \theta)$, que associa cada valor de θ à probabilidade de **x** ser observado.

Supondo uma amostra de tamanho n e independência entre cada variável aleatória X_i, $i = 1, ..., n$ pode-se então estabelecer a seguinte relação entre a função de verossimilhança e a probabilidade de se observar cada x_i: $p(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$.

Assim, partindo da especificação apresentada para o modelo na subseção 3.2.1:

$$[Y_{i,j} \mid \lambda_{i,j}] \sim Pois(\lambda_{i,j})$$

$$ln(\lambda_{i,j}) = \theta - |a_i - a_j|, \quad \forall \quad i < j. \text{ e } \lambda_{ij} = 0, 1, 2, 3, ...$$

Temos então a seguinte função de verossimilhança associada:

$$p(\mathbf{y} \mid \theta, \mathbf{a}) = \prod_{i<j} \left(\frac{\exp(\theta - |a_i - a_j|)^{y_{i,j}} \exp(-\exp(\theta - |a_i - a_j|))}{y_{i,j}!} \right), \quad onde \quad y_{i,j} = 0, 1, 2, 3, ..., \quad 1 < i < j < n.$$

No intuito de simplificar os cálculos, uma alternativa comumente utilizada é considerar como função objetivo a função de log-verossimilhança negativa, uma vez que maximizar a função de verossimilhança é o mesmo que minimizar a função de log-verossimilhança negativa.

Não obstante, conforme mencionado na subseção anterior, utilizou-se como valores iniciais para os parâmetros as posições obtidas via Escalonamento Multidimensional, visando iniciar o algoritmo de maximização em um lugar mais ”próximo do ótimo”. Além disso, as 3 primeiras posições foram fixadas, tornando assim o conjunto de soluções invariante sob as transformações lineares de reflexão, rotação e translação.

2.2.4 Inferência Bayesiana

As conclusões obtidas através da Inferência Bayesiana, a respeito de um determinado parâmetro θ, ou de uma variável não observada **y**, são baseadas em especificações probabilísticas. Tais especificações são feitas condicionalmente a uma amostra de valores observados, sendo esta relacionada de alguma forma com as quantidades de interesse.

Matematicamente, todo o processo de Inferência Bayesiana consiste na atualização da informação obtida *a priori* com base em um teorema denominado Teorema de Bayes. A partir deste é que obtêm-se a distribuição *a posteriori* do(s) parâmetro(s) de interesse mesclando a distribuição *a priori* com a função de verossimilhança associado ao modelo.

A distribuição *a priori* contempla a incerteza a respeito de θ (vetor de parâmetros desconhecidos que pertence ao espaço paramétrico Θ), condicionalmente a H (informação inicial acerca de alguma quantidade de interesse). Denotaremos aqui esta informação por $p(\theta|H)$.

Por sua vez, a distribuição *a posteriori* atualiza a informação inicial sobre uma quantidade de interesse com base na informação da amostra **x**. A informação disponível para a inferência passará a ser $H^* = H \cap \{\mathbf{X} = \mathbf{x}\}$. Denotaremos esta informação por $p(\theta|H^*)$. É importante ressaltar a importância da distribuição *a posteriori* na Inferência Bayesiana: segundo Migon e Gamerman (1999), é através dela a forma mais adequada de avaliar a informação disponível a respeito de uma quantidade desconhecida θ.

$$p(\theta|H^*) = p(\theta|\mathbf{x}, H) = \frac{p(\theta, \mathbf{x}|H)}{p(\mathbf{x}|H)} = \frac{p(\mathbf{x}|\theta, H)p(\theta|H)}{p(\mathbf{x}|H)} \tag{1}$$

onde $p(\theta|H)$ e $p(\theta|H^*)$ representam as distribuições *a priori* e *posteriori*, respectivamente, e

$$p(\mathbf{x}|H) = \int p(\theta, \mathbf{x}|H)d\theta.$$

Como $p(\mathbf{x}|H)$ não depende de θ e sendo H é comum a todos os termos, pode-se reescrever o teorema da seguinte forma:

$$p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)p(\theta). \tag{2}$$

Assim, o resultado obtido em (2) é então conhecido como Teorema de Bayes e se constitui como a base de todos os procedimentos da inferência Bayesiana. Na próxima subseção, será apresentado um último método utilizado, dentro do contexto da inferência bayesiana, para a estimação dos parâmetros do Modelo de Distâncias Latentes: *Método de Monte Carlo via Cadeias de Markov* (MCMC).

2.2.5 Método de Monte Carlo via Cadeias de Markov (MCMC)

Na Inferência Bayesiana, os métodos de simulação estocástica estão relacionados ao processo de obtenção de amostras das distribuições *a posteriori* quando as mesmas não possuem uma forma fechada. Estes métodos visam a estimação das distribuições *a posteriori* e, por conseguinte, das características probabilísticas de interesse.

Em problemas altamente complexos, em que se tem um grande número de parâmetros envolvidos, os métodos de *Monte Carlo via Cadeias de Markov* se tornam a solução para a conclusão do processo inferencial.

Essencialmente, o objetivo do método é construir uma cadeia de Markov cuja distribuição de equilíbrio (também conhecida como distribuição limite) seja igual à distribuição de interesse. Para tal, realiza-se um número finito de simulações desta cadeia, de forma que a mesma eventualmente convirja e, a partir deste ponto, dê origem a uma amostra da distribuição de interesse.

Se os parâmetros $\theta_1, ..., \theta_p$ possuem densidade conjunta $p(\boldsymbol{\theta}) = p(\theta_1, ..., \theta_p)$ e $q(\theta, \theta^*)$ é a distribuição condicional das transições do estado θ, dessa forma, é possível construir uma cadeia de Markov com probabilidades de transição invariantes no tempo, onde qualquer estado pode ser obtido a partir de um outro qualquer em um número finito de iterações.

Assim, independentemente do estágio inicial da cadeia, uma trajetória pode ser gerada e, consequentemente, pode-se alcançar a distribuição de equilíbrio $p(\boldsymbol{\theta})$.

Existem diversos métodos na literatura cujo objetivo é a construção da cadeia de Markov. Dentre os mais famosos, temos o Amostrador de Gibbs, proposto por Geman e Geman (1984) e popularizado por Gelfand e Smith (1990), bem como o método de Metropolis-Hastings, proposto por Metropolis *et al.* (1953) e Hastings (1970).

No contexto desta iniciação, o método para construção da cadeia adotado foi o de Metropolis-Hastings Metropolis *et al.* (1953) e Hastings (1970), pois este já encontra-se implementado no software `R` como sendo uma das rotinas do pacote *latentnet* (Krivitsky, 2018).

Novamente na forma de um algoritmo, o método de Metropolis-Hastings pode ser decomposto nas seguintes etapas:

- Considerar $q(\boldsymbol{\theta}, \cdot)$ como sendo um núcleo arbitrário de transição e assuma que na iteração (j) a cadeia se encontra no estado $\boldsymbol{\theta}^{(j)}$. Denote, a posição da cadeia na iteração $(j+1)$ por $\boldsymbol{\theta}^{(j+1)}$.

- Propor um movimento da cadeia para o estado $\boldsymbol{\theta}^*$ a partir de $q(\boldsymbol{\theta}^{(j)}, \cdot)$;

- Aceitar o movimento proposto com probabilidade

$$\alpha(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^{(*)}) = min\left\{1, \frac{p(\boldsymbol{\theta}^*)/q(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(j)})/q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j)})}\right\}$$

  e fazer $\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^*$ ou rejeitar o movimento com probabilidade $1 - \alpha(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^{(*)})$, neste caso fazendo $\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)}$.

Alguns autores indicam como razoáveis taxas de aceitação entre 20% e 50% dessas propostas durante o MCMC (Gamerman e Lopes, 2006) e, em determinados cenários multidimensionais, Roberts *et al.* (1997) sugere uma taxa ótima de 23.4%.

Após decidido o método a ser utilizado e obtida uma simulação da cadeia, devemos nos certificar acerca da sua convergência. Somente após esta confirmação poderemos formar a amostra da distribuição *a posteriori* das quantidades desconhecidas do modelo.

Dentre as mais variadas formas de se realizar uma análise de convergência, uma delas é a inspeção gráfica, onde se analisa a trajetória de uma ou mais cadeias em intervalos de tempo distintos. Neste estudo, utilizou-se um critério semelhante ao da inspeção gráfica: avaliou-se o comportamento de duas cadeias com valores iniciais distintos, de forma que a convergência foi alcançada se ambas permanecem em torno de um mesmo ponto.

Obtida a amostra, deve-se então analisar a autocorrelação existente entre $\boldsymbol{\theta}^{(j)}$ e $\boldsymbol{\theta}^{(j+1)}$. Como estamos lidando com uma amostra de uma cadeia de Markov, temos uma amostra aleatória, porém não independente. Isto não afeta as estimativas dos parâmetros, mas tem influência sobre as variâncias das estimativas resultantes (Gamerman e Lopes, 2006). Portanto, nos casos em que, após verificada a convergência, for constatada uma forte correlação serial na cadeia, recomenda-se a retirada de uma amostra sistemática de seus valores para compor uma nova amostra. A forma como a amostragem sistemática será realizada pode ser baseada em um gráfico contendo a função de autocorrelação da cadeia.

Finalmente, verificados todos estes itens, o processo de inferência tem prosseguimento a partir do método de Monte Carlo. O objetivo deste método é estimar o valor de uma integral a partir da obtenção de seu valor esperado associado a alguma distribuição de probabilidade. Para se obter, por exemplo, uma estimativa para o valor esperado *a posteriori* de um parâmetro θ do modelo, basta tomar a médias das j-ésimas componentes dos valores amostrados, ou seja:

$$\hat{\theta} = \frac{\sum_j \theta_j}{n}.$$

3 Base de Dados

Os modelos apresentados na seção anterior foram aplicados a dados de artistas que ocuparam o ranqueamento *Artist-100 Chart* produzido pela mídia de entretenimento norte-americana Billboard (www.billboard.com/charts/artist-100).

A motivação por trás da escolha da Billboard como fonte de dados, justifica-se na grande quantidade de informação contida em seu site. No mesmo, o usuário consegue interagir facilmente com a página, acessando diversos ranqueamentos os quais possibilitam não só uma análise acerca da evolução do gosto musical nas últimas décadas, como também acompanhar, por exemplo, a oscilação de popularidade de um determinado artista ou música de interesse. Não obstante, a estruturação consistente do site em termos das linguagens HTML e CSS *(Cascading Style Sheet)* é também um fator que favorece o processo de *Web Scraping*.

Na *Artist-100 Chart*, os artistas são ordenados semanalmente com base nas músicas mais populares dentre todos os gêneros musicais. A respeito dos critérios para a qualificação dos artistas, os divulgados são: impressões de audiência por rádio, dados de vendas de álbuns e *streaming* internacionais.

Dentre os diversos ranqueamentos disponíveis no site da Billboard, a escolha da *Artist-100 Chart* como base de dados, foi motivada pelo fato da mesma ser a única que trata diretamente o artista como objeto de estudo. Outras *charts*, como por exemplo, *Hot 100* e *Billboard 200*, que tratam as músicas e os álbuns, respectivamente, como unidade de

estudo, não se adéquam tão bem ao contexto de uma rede social, uma vez que tratar os indivíduos da rede como sendo álbuns ou músicas, concomita para uma relação menos direta do que tratar os indivíduos como artistas.

No que tange o processo de obtenção dos dados, foi criada uma rotina automatizada de *Web Scraping* no software R (R Core Team, 2019), com o auxílio dos pacotes *RSelenium* (Harrison, 2019) e *rvest* (Wickham, 2019), cujo propósito era interagir repetitivamente com a página, no intuito de determinar a data da primeira *chart* produzida (19/07/2014). A partir desta informação, a rotina coletava sequencialmente todas as *charts* produzidas desde a data mencionada até a data limite em nosso planejamento: 27/07/2019 (263 semanas).

Algumas mudanças foram propostas aos dados durante a etapa de análise exploratória. Em um primeiro momento, observou-se que ranqueamentos produzidos em semanas próximas, constantemente repetiam a ordenação dos artistas, o que acabaria distorcendo a variabilidade nas relações entre os mesmos. Sendo assim, a solução implementada foi extrair somente a última *chart* de cada mês, totalizando 61 *charts*, de forma que cada uma destas atuasse como sendo um ”retrato”das demais outras *charts* daquele mesmo mês.

A segunda e última modificação, diz respeito ao custo computacional associado ao tamanho da rede. Para se ter uma ideia, nos 61 meses considerados, 772 artistas distintos ocuparam a *Artist-100 Chart*. Modelar uma rede social de tal dimensão mostrou ser algo impraticável não só em termos computacionais, como também visuais, haja vista que a visualização de redes é tradicionalmente feita na forma de grafos. Assim, mais uma vez a saída encontrada foi reduzir o conjunto de dados, considerando somente os 100 artistas que mais ocuparam as 20 primeiras posições (Top 20) da *Artist-100 Chart* da Billboard.

Por fim, o banco de dados para o Modelo de Distâncias Latentes ficou então definido na forma de uma socio-matriz **Y**, simétrica, de dimensão 100×100, cujas entradas denotam o número de meses que os artistas i e j compartilharam o Top 20 da *Artist-100 Chart* da Billboard. Convém ressaltar que a diagonal da socio-matriz, ou seja, as entradas Y_{ii}, representam o número de meses em que o próprio artista i ocupou o Top 20 da *Artist-100 Chart*.

4 Resultados

Os resultados que serão apresentados a seguir foram obtidos mediante o uso do já mencionado *software* R, com o devido auxílio de determinados pacotes e das suas funções associadas. Tais pacotes serão referenciados nas suas devidas subseções.

4.1 Estatísticas Descritivas

De antemão a aplicação direta dos modelos aos dados, foram consideradas diversas técnicas de análise exploratória, objetivando melhor resumir as estruturas de popularidade implícitas entre os artistas.

Porém, levando-se também em consideração a alta dimensionalidade da socio-matriz associada aos dados, as estatísticas descritivas apresentadas são também aquelas capazes de lidar com tal magnitude, como é o caso do *heat map* e o *bar chart*.

Um *heat map* nada mais é do que um gráfico que utiliza cores para mapear as discrepâncias numéricas existentes em um conjunto de dados. No nosso contexto, a sua utilidade consiste em identificar, de forma resumida, a magnitude das relações entre os artistas da rede. O *heat map* aqui apresentado, utiliza somente as informações brutas contidas na socio-matriz.

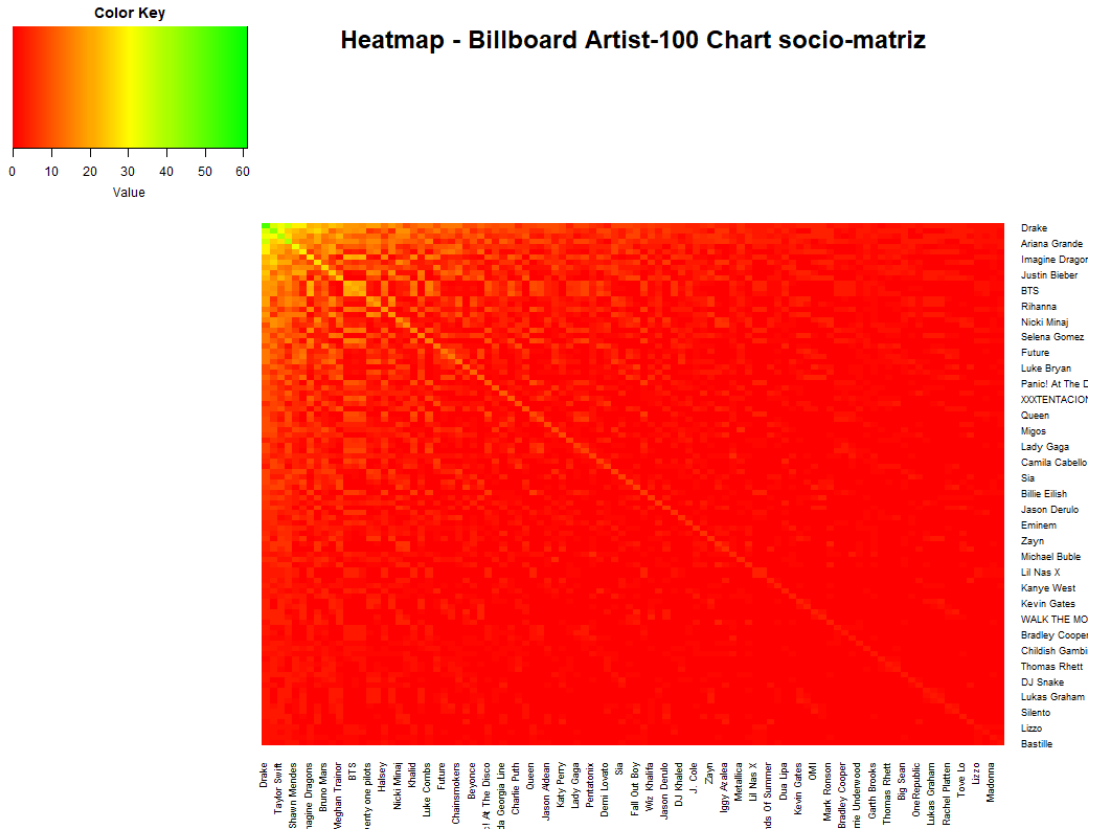

Figura 1: *heat map* da socio-matriz associada a *Artist-100 Chart*. O mapeamento das cores é, em ordem crescente de valor: vermelho, laranja, amarelo e verde.

Analisando a Figura 1, é possível perceber que a vasta maioria das relações entre os artistas, dificilmente ultrapassa os 20 meses compartilhados no Top 20, com a exceção apenas de uma pequena minoria representada nas cores amarela e verde.

No intuito de analisar, em maior profundidade, quais seriam, de fato, os artistas mais populares que já ocuparam a *Artist-100 Chart* desde a sua criação, o *bar chart* que será apresentado a seguir atende especificamente à este propósito, nos dando uma ”ordenação de popularidade”dos artistas.

Em termos de dados utilizados, a diferença entre o *heat map* e o *bar chart* é que este último, considera somente a diagonal da socio-matriz, ou seja, somente a informação a respeito do número de meses em que o n-ésimo artista (sozinho) ocupou o Top 20 da *Artist-100 Chart*.

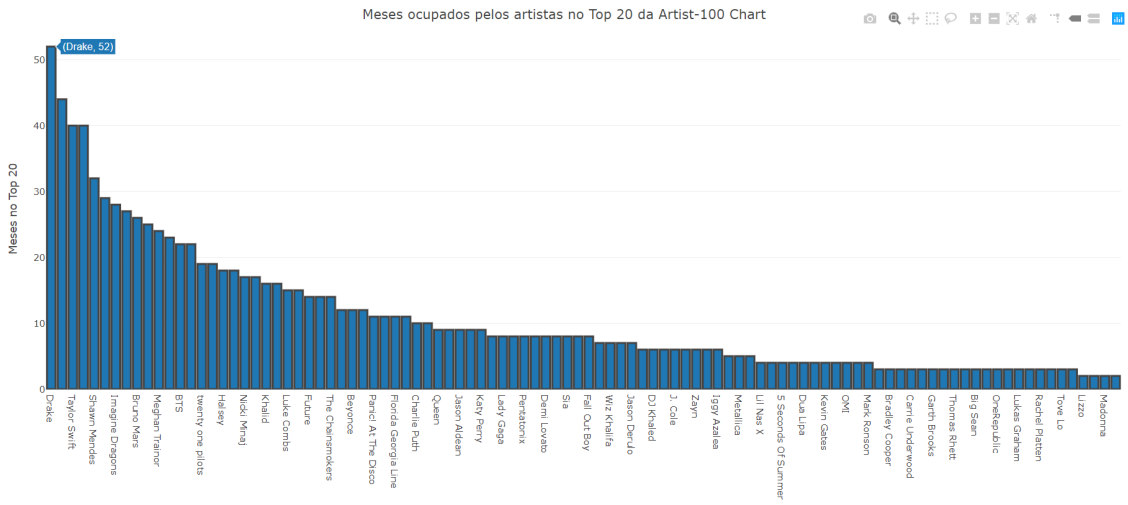

Figura 2: Disposição dos artistas da *Artist-100 Chart* da Billboard em termos de meses ocupados no Top 20. Destaque vai para o artista mais popular: Drake, com 52 meses.

O padrão de ordenação dos artistas observado na Figura 2, complementa aquilo que foi observado no *heat map* da Figura 1, uma vez que as discrepâncias em termos de popularidade entre os artistas, fica agora evidenciada não só pela predominância da coloração vermelha no *heat map*, como também pelo caráter assimétrico do *bar chart*.

Ainda sobre a ordenação dos artistas, destaca-se aqui os 5 primeiros artistas (Top 5), que servirão como referência para os resultados das próximas subseções. São eles, em ordem decrescente: Drake, Ed Sheeran, Ariana Grande, Taylor Swift e Shawn Mendes.

Por último, os pacotes utilizados para a confecção dos gráficos foram o *plotly* (Sievert, 2019), para a construção do *bar chart* e o *gplots* (Warnes, 2019) no caso do *heat map*.

4.2 Analise da rede via Escalonamento Multidimensional

A implementação do Modelo de Escalonamento Multidimensional foi feita com pequena modificação acerca do procedimento apresentado na subseção 3.2.2. Tal modificação foi proposta com o objetivo de minimizar o seguinte efeito identificado nas primeiras replicações do modelo: o número de meses em que os artistas ocuparam no Top 20, isto é, a diagonal da socio-matriz, exercia forte influência na configuração da rede como um todo, menosprezando as relações que os artistas estabeleciam uns com os outros.

Conquanto, uma vez que relegar tal efeito tornaria a interpretação trivial. a mudança introduzida foi a utilização de uma ”socio-matriz de correlações”, em oposição à socio-matriz original. Nesta ”nova”socio-matriz, ainda simétrica e com mesma dimensão 100×100, as suas entradas agora quantificam o coeficiente de correlação de Pearson entre os artistas i e j.

A escolha desta estatística como transformação aos dados é simples de ser justificada, uma vez que, pelo fato de o Escalonamento Multidimensional, por si só, já analisar a rede sob o contexto da semelhança nos padrões de conexões entre os artistas, tomar

o coeficiente de correlação, que dimensiona o quão interligados são os artistas i e j, tornaria ainda mais explícita tal relação. Dessarte, o Escalonamento Multidimensional implementado interpreta a rede acerca do padrão de correlações entre os artistas, onde a proximidade entre os mesmos pode ser interpretada como ”ter compartilhado o Top 20 da *Artist-100 Chart* com artistas iguais”.

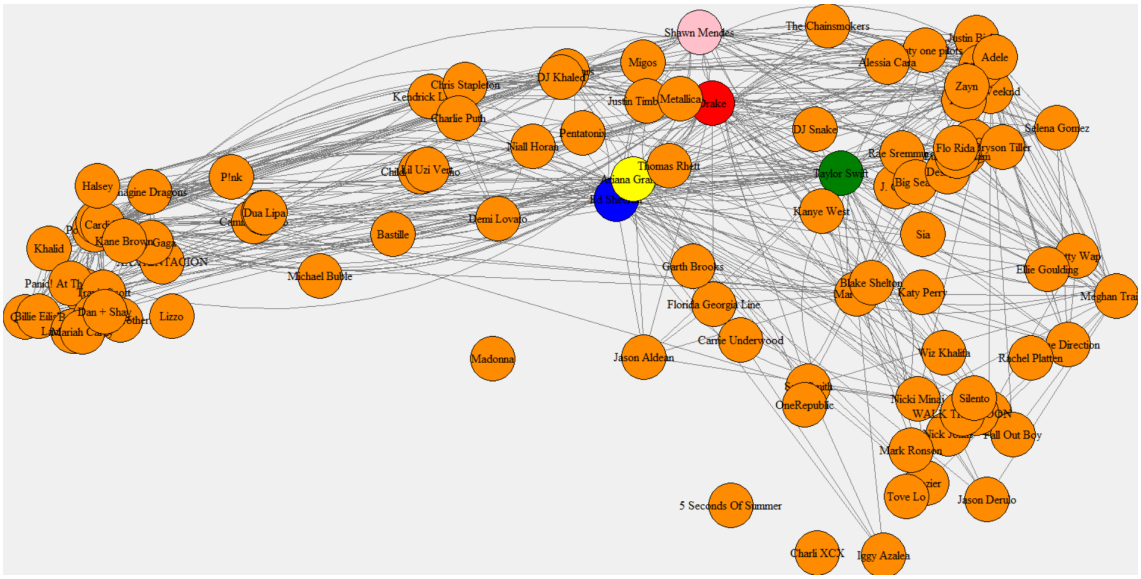

Figura 3: Visualização da rede sob a ótica do Escalonamento Multidimensional realizado com a socio-matriz de correlações. Em cores diferentes, estão destacados os Top 5.

Por uma questão de legibilidade, optou-se por reduzir o número de arestas plotadas, não só no grafo da Figura 1, mas também para com todos os demais grafos deste relatório, de forma que cada aresta simbolize 10 meses compartilhados no Top 20 entre os artistas i e j. É excepcionalmente por este motivo que alguns vértices encontram-se desconexos.

Contudo, o Modelo de Escalonamento Multidimensional nos fornece uma interessante visão acerca das relações entre os artistas. Analisando o grafo, é possível perceber razoáveis agrupamentos dos mesmos, sendo estes influenciados não tão somente pelo fator popularidade, mas também levando-se em consideração a ”afinidade”de um artista com os demais.

Por exemplo, ainda que os Top 5 estejam dispostos relativamente próximos uns aos outros, mais ao centro da rede, artistas apontados como não tão populares pelo *bar chart* da Figura 2, como Metallica, Thomas Rhett e Garth Brooks, encontram-se aqui próximos destes mais populares. Tal fato é compreensível, uma vez que os Top 5 são uns dos poucos artistas que compartilharam o Top 20 com estes 3 últimos citados.

No que tange a implementação do Modelo de Escalonamento Multidimensional no *software* `R`, não utilizou-se nenhum outro pacote estatístico, senão as funcionalidades já presentes por padrão, como é o caso das rotinas *cmdscale* e *cor*. Entretanto, para a elaboração do grafo da Figura 1, bem como a elaboração dos demais grafos deste relatório, utilizou-se o pacote *igraph* (Csárdi, 2019).

4.3 Analise da rede via Modelo de Distâncias Latentes

O estudo da rede sob a ótica do Modelo de Distâncias Latentes, subdivide-se em duas subseções: Resultados obtidos por Máxima Verossimilhança (subseção 4.3.1) e Resultados obtidos por métodos Bayesianos (subseção 4.3.2). As diferenças matemáticas entre as duas abordagens, tão quanto a diferença entre os resultados obtidos, incentivaram tal divisão.

4.3.1 Resultados obtidos por Máxima Verossimilhança

A estimação por máxima verossimilhança dos parâmetros (posições) do Modelo de Distâncias Latentes, foi implementada de duas formas distintas no *software* `R`. Em um primeiro momento, criou-se um subprograma que define a expressão analítica da função de verossimilhança apresentada na subseção 2.2.3. Para maximizá-la, utilizou-se a rotina maximizadora não-linear *nlminb* já disponível no `R`.

No entanto, por entraves computacionais, o nosso subprograma limitou-se em estimar somente as posições referentes aos 50 primeiros artistas da rede. Mesmo com tal restrição, as estimativas obtidas, quando comparadas aquelas de subprogramas mais robustos, como é o caso da rotina *ergmm* do pacote *latentnet* (Krivitsky, 2018), são factíveis e retratam em geral um mesmo padrão na ordenação dos artistas.

As Figuras 4 e 5 a seguir, apresentam os grafos que ilustram a estrutura das relações entre os artistas no espaço latente, utilizando as estimativas de máxima verossimilhança (EMVs) obtidas pelo subprograma de nossa autoria e do pacote *latentnet*.

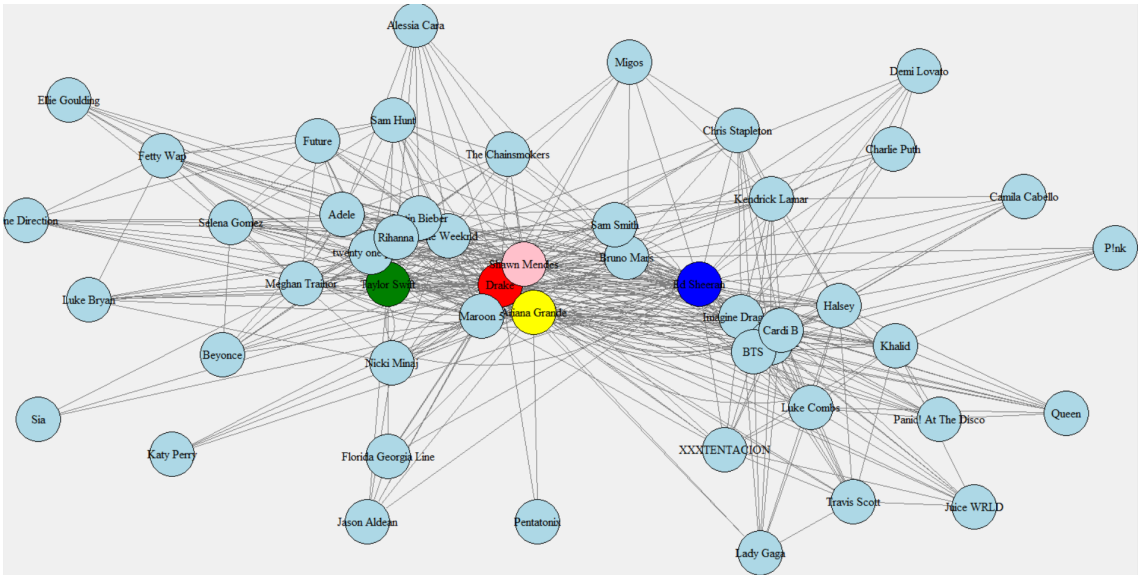

Figura 4: Visualização parcial da rede com base nas EMVs do nosso subprograma.

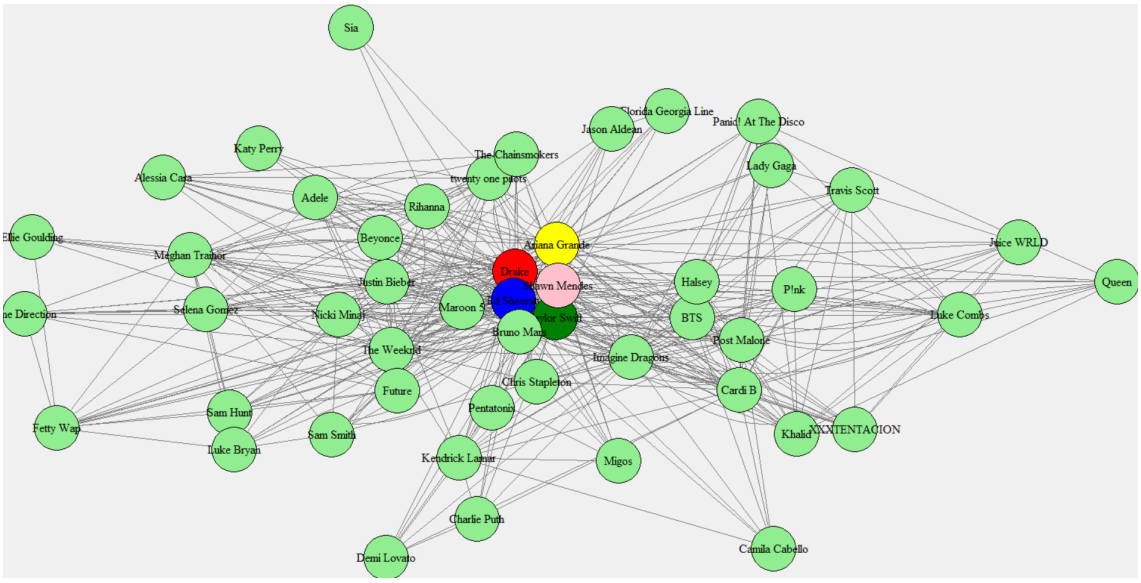

Figura 5: Visualização parcial da rede com base nas EMVs do pacote.

Comparando diretamente as duas estimativas, as suas diferenças na configuração dos artistas devem-se, principalmente, aos já mencionados efeitos de rotação, reflexão e translação inerentes ao modelo, bem como ao condicionamento à clusterização implementado pela rotina *ergmm*, que minimiza o efeito de sobreposição e melhor agrupa os artistas na rede. Maiores detalhes sobre este método podem ser obtidos no artigo [21].

Por fim, a Figura 6 expõem a disposição dos 100 artistas no espaço latente com base nas estimativas de máxima verossimilhança do *latentnet*. Nesta última, é possível perceber que considerar os 50 artistas restantes pouco influenciou no padrão geral da rede, ilustrando o efeito do ”fator popularidade”sobre a mesma.

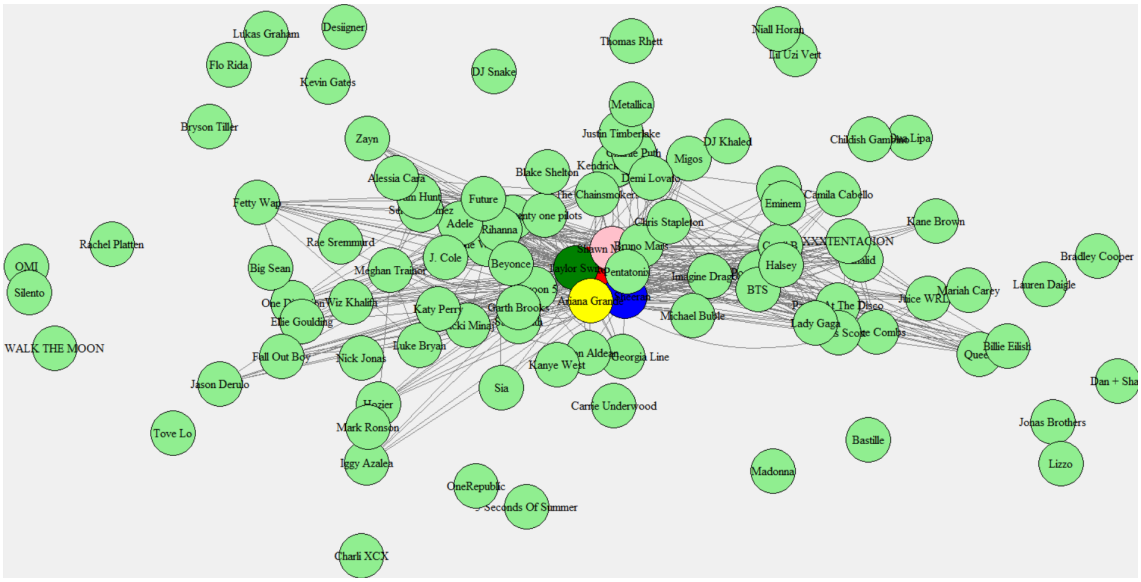

Figura 6: Visualização completa da rede sob a ótica das EMVs obtidas pelo pacote.

4.3.2 Resultados obtidos por métodos Bayesianos

A abordagem bayesiana utiliza o *Método de Monte Carlo via Cadeias de Markov* (MCMC) para a estimação das médias *a posteriori* das posições dos artistas, novamente considerando um espaço latente bidimensional. Visando melhor comparar os resultados entre as seções, as Figuras 7 e 8 a seguir apresentam as configurações dos artistas obtidas pelos métodos Bayesianos considerando, respectivamente, os 50 primeiros artistas e posteriormente todos os 100.

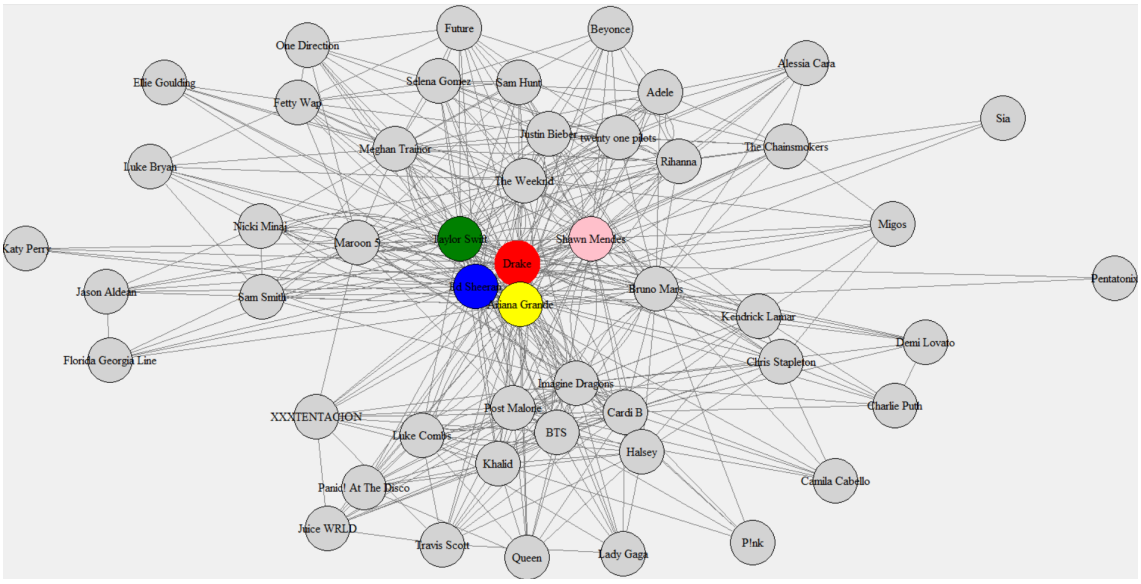

Figura 7: Médias *a posteriori* das posições dos 50 primeiros artistas.

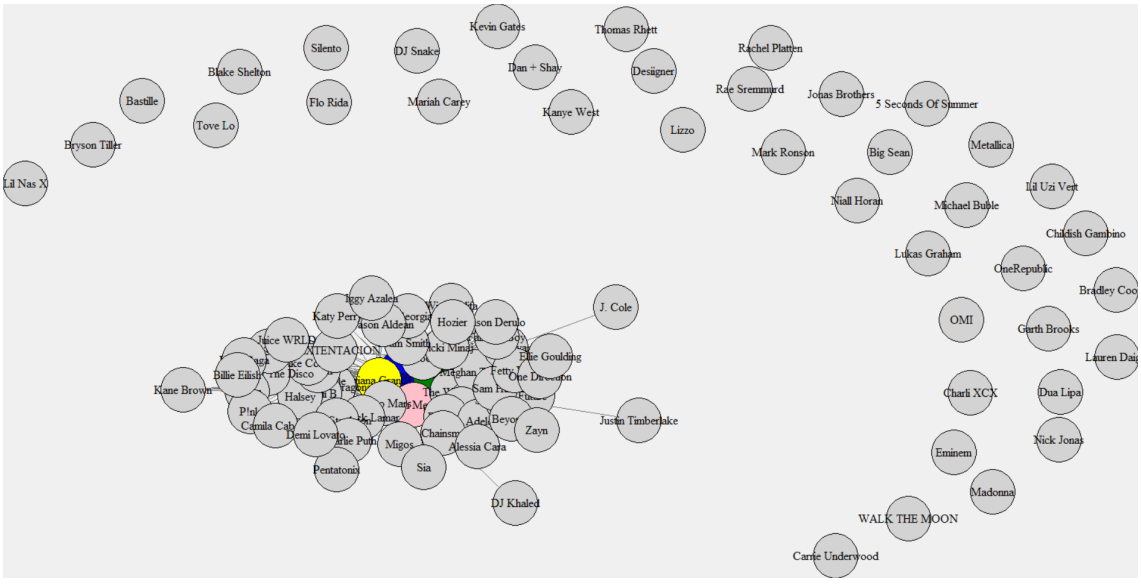

Figura 8: Médias *a posteriori* das posições de todos os 100 artistas.

Comparando as estimativas obtidas pelos métodos Bayesianos e as estimativas de Máxima Verossimilhança, nota-se que estas assemelham-se no sentido de exaltar as relações de popularidade entre os artistas, uma vez que nas Figuras 6 e 8, é possível perceber uma disposição periférica dos artistas menos populares à rede.

Examinando conjuntamente os grafos das Figuras 3, 6 e 8, é notável a diferença de arranjamento obtido pelos 3 métodos considerados. No entanto, contraditoriamente, é interessante notar casos particulares de grupos de artistas, os quais sugerem que determinadas relações pode ser interpretadas de forma igual pelos modelos. São estes o quarteto composto por One Direction, Ellie Goulding, Fetty Wap e Meghan Trainor.

Sobre tal grupo, é curioso notar uma constância na proximidade entre estes artistas ao longo dos modelos considerados. Tal imutabilidade reforça a ideia de que, se um ou mais artistas possuem conexões ”fortes”uns com os outros na rede, esta relação será, sob o escopo da interpretação inerente ao modelo, captada independente da abordagem matemática utilizada. A tabela abaixo mostra os dados da socio-matriz associados à estes artistas.

	One Direction	Ellie Goulding	Fetty Wap	Meghan Trainor
One Direction	10	4	6	10
Ellie Goulding	4	8	6	8
Fetty Wap	6	6	12	11
Meghan Trainor	10	8	11	24

De uma forma geral, observou-se que para o Modelo de Distâncias Latentes, a popularidade de um artista, seja esta expressa na forma de muitos meses ocupados no Top 20 ou na forma de muitas conexões com demais outros artistas, não é necessariamente um fator determinístico para se ocupar uma posição central na rede. O modelo mostra-se capaz de ponderar a ”popularidade geral”de um artista levando-se em consideração a popularidade de ”com quem o mesmo está se conectado”.

5 Conclusões

Os modelos para redes sociais mostraram-se eficazes em identificar relações de popularidade e orgnizar agrupamentos entre os artistas que ocuparam o Top 20 da *Artist-100 Chart* da Billboard. Não obstante, mesmo nos casos em que as justificativas por trás de suas posições não seja tão evidente, é ainda possível interpretar as distâncias entre os mesmos, seja no âmbito de estimar a probabilidade da existência de uma conexão (Modelo de Distâncias Latentes) ou de forma a comparar seus padrões de correlações (Modelo de Escalonamento Multidimensional).

No que tange à parte computacional, a criação da rotina automatizada de *Web Scraping* para extração os dados, tão quanto a implementação dos modelos e a utilização dos seus pacotes associados no *software* R, contribuíram positivamente para a consolidação de uma maior e importante proficiência na linguagem.

As diferentes interpretações e arranjamentos produzidos por cada modelo, concomitaram positivamente para interpretação das estruturas de popularidade entre os artistas como um todo. Entretanto, as similaridades também observadas entre os modelos, contribuíram para ressaltar características marcantes entre os artistas.

Por fim, ideias futuras para este trabalho consistem em ampliar os modelos implementados com a adição de co-variáveis acerca da oscilação nas posições ocupadas pelos artistas, uma vez que tal informação não existia no site da Billboard durante período de coleta.

Referências

Airoldi E. M., Blei D. M., Fienberg S. E., Xing E. P. (2008). *Mixed membership stochastic blockmodels*. *Journal of Machine Learning Research*, 9:1981-2014.

Borg I., Groenen, P. J. (2005). *Modern Mutidimensional Scaling: Theory and Applications*. Springer, New York, 261.

Cardoso-Junior M.M., Scarpel R.A. (2013). Métodos de construção e interpretação do mapa perceptual para estudos de percepção de riscos. *Revista Eletrônica Pesquisa Operacional para o Desenvolvimento*, 5(3):451-452.

Cox T. F., Cox M. A. A. (2001). *Multidimensional Scaling*. 2nd ed. London, Chapman and Hall, 31.

Csardi G, Nepusz T: *The igraph software package for complex network research, InterJournal, Complex Systems* 1695. 2006. Disponível em: <http://igraph.org>.

Gamerman D., Lopes H. F. (2006). *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference*. Chapman and Hall.

Gelfand A.E., Smith A.F.M. (1990). *Sampling-Based Approaches to Calculating Marginal Densities*. *Journal of the American Statistical Association*, 85, 398-409.

Geman S., Geman D. (1984). *Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Gower J. C. (1966). *Some distance properties of latent root and vector methods used in multivariate analysis*. *Biometrika*, 53(3-4):325-338.

Giannini, R. et al. *Web scraping for collecting price data: Are we doing it right? New Techniques and Technologies for Statistics* (NTTS), 2017.

Glez-Peña, D. et al. *Web scraping technologies in an api world. Briefings in bioinformatics, Oxford University Press*, v. 15, n. 5, p. 788–797, 2013.

Handcock M. S., Raftery A. E., Tantrum J. M. (2007). *Model-Based Clustering for Social Networks*. *Journal of the Royal Statistical Society, Series A*, 170(2), 301-354.

Harrison, J. *RSelenium: R Bindings for 'Selenium WebDriver'*. [S.l.], 2019. R package version 1.7.5. Disponível em: <https://CRAN.R-project.org/package=RSelenium>

Hastings W. K. (1970). *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*. *Biometrika*, 57, 97-109.

Hoff P. D., Raftery A. E., Handcock M. S. (2002). *Latent space approaches to social network analysis*. *Journal of the American Statistical Association*, 97(460):1090-1098.

Johnson Richard A., Wichern Dean A. (2007).*Applied Multivariate Statistical Analysis*. 6th ed. New Jersey, Prentice Hall, 706-715.

Krivitsky P, Handcock M (2018). *latentnet: Latent Position and Cluster Models for Statistical Networks. The Statnet Project (http://www.statnet.org)*. R package version 2.9.0. Disponível em: <https://CRAN.R-project.org/package=latentnet>.

Mardia K.V. (1978). *Some properties Classical multidimensional scaling*. *Commun. Stat. Theor. Meth.*, A7(13):1233-1241.

Metropolis N., Rosenbulth A. W., Rosenbulth M. N., Teller A. H., Teller E. (1953). Equation of State Calculations by Fast Computing Machine. *Journal of Chemical Physics*, 21, 1089-1091.

Migon H.S., Gamerman D. (1999). *Statistical Inference - an integrated approach*. Arnold.

Munzert, S. et al. *Automated data collection with R: A practical guide to web scraping and text mining*. [S.l.]: *John Wiley Sons*, 2014.

Nowicki K., Snijders T. A. B. (2001). *Estimation and prediction for stochastic blockstructures*. *Journal of the American Statistical Association*, 96, 1077-1087.

R Core Team (2019). R: *A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. Disponível em: <https://www.R-project.org/>.

Rencher A. C. (2002). *Methods of Multivariate Analysis*, 2nd ed. Wiley-Interscience, New York, 505-506.

Roberts G.O., Gelman A., Gilks W.R. (1997). *Weak convergence and optimal scaling of random walk Metropolis algorithms. Annals of Applied Probability*, 7, 1, 110-120.

Siervert, C. *plotly: Create interactive Web Graphics via 'plotly.js'*. [S.l.], 2019. R package version 4.9.0. Disponível em: <https://CRAN.R-project.org/package=plotly>

Torgerson W.S. (1952). *Multidimensional scaling: I. Theory and method*. *Psychometrika*, 17(4):401-419.

Wang L, Montano M, Rarick M, Sebastiani P. Conditional clustering of temporal expression profiles. BMC Bioinformatics. 2008;9:147. Published 2008 Mar 11. doi:10.1186/1471-2105-9-147.

Warnes, R. G. *gplots: Various R Programming Tools for Plotting Data*, 2019. R package version 3.0.1.1. Disponível em: <https://CRAN.R-project.org/package=gplots>.

Wickham, H. *rvest: Easily Harvest (Scrape) Web Pages*. [S.l.], 2019. R package version 0.3.4. Disponível em: <https://CRAN.R-project.org/package=rvest>.

Young G., Householder A.S. (1938). *Discussion of a set of points in terms of their mutual distances*. *Psychometrika*, 3(1):19-22.