Sistemas & Gestão 9 (2014), pp 566-574

SISTEMAS & GESTÃO
Revista Eletrônica

UM ESTUDO DO PROBLEMA DE DETECÇÃO DE COMUNIDADES EM REDES

AN ANALYSIS OF THE PROBLEM OF COMMUNITY DETECTION IN NETWORKS

Isabelle Alves[a]; Carla Silva Oliveira[a]; José André de Moura Brito[a]

[a] **Escola Nacional de Ciências Estatísticas** (ENCE/IBGE) - Rio de Janeiro, RJ, Brasil

Resumo

O presente trabalho traz um estudo do problema de detecção de comunidades associadas às redes complexas, ou seja, redes formadas por elementos do mundo real. De forma a realizar esse estudo, as redes foram representadas por uma estrutura matemática denominada grafo. E de forma a detectar as comunidades nessas redes, aplicou-se um algoritmo de detecção de comunidades no grafo associado. Em particular, neste trabalho, optou-se pela utilização de um algoritmo de detecção de comunidades, denominado FastGreedy. Uma vez definida a comunidade, em uma segunda fase do estudo, foi aplicado um algoritmo de agrupamento (K-means) em cada um dos grafos, considerando o número de grupos igual ao número de comunidades. A comunidade definida e os agrupamentos construídos foram avaliados a posteriori mediante o cálculo da modularidade e do coeficiente de silhueta. Ao final desse trabalho, são apresentados alguns resultados computacionais concernentes à aplicação desses algoritmos, disponíveis no software livre R, considerando três bases reais.

Palavras-chave: Grafos, Comunidades, Algoritmos.

Abstract

This paper presents a study of the problem of detecting communities associated with complex networks, this is, networks comprising elements of the real world. In order to perform this study, networks were represented by a mathematical structure called graph. And, in order to detect the communities these networks was applied to an algorithm for detecting communities in associated graph. In particular, in this work, we opted for the use of a community detection algorithm, called FastGreedy. Once defined the community, in a second phase of the study, we applied a clustering algorithm (K-means) in each of the graphs, groups considering the number equal to the number of communities. The community defined and constructed clusters were evaluated retrospectively by calculating the coefficient of modularity and silhouette. At the end of this work are presented some results concerning the application of these computational algorithms available in the free software R, considering three real bases.

Keywords: *Graphs, Communities, Algorithms.*

1. INTRODUÇÃO

A tarefa de agrupar objetos semelhantes, no que diz respeito às suas características, é algo importante e que está presente diariamente em nosso cotidiano. Não só importante, mas também necessária, já que nos possibilita organização. Nesse sentido, podemos desenvolver essa atividade em uma rede, caracterizada por um conjunto n de objetos definidos em função de seus *p* atributos (características ou variáveis). Essa rede, por sua vez, possui estruturas que podem ser modeladas por um grafo, em que os vértices representam os terminais da rede e as arestas o meio físico de comunicação entre eles. Quando objetos de uma rede estão associados com elementos do mundo real, ela é denominada complexa. Portanto, considerando que uma dada rede seja representada por um grafo não orientado, define-se uma comunidade como um conjunto de vértices que possuem propriedades comuns ou que desempenham funções similares dentro da rede. Assim, encontramos na sociedade diversos exemplos de comunidades representadas por redes complexas, quais sejam: as famílias, os grupos virtuais e as nações. De forma a avaliar e quantificar a semelhança entre os vértices da rede, faz-se o uso de medidas de similaridade que estão associadas a algum tipo de distância (métrica) ou coeficiente como, por exemplo, a distância euclidiana

PROPPI / LATEC
DOI: 10.7177/sg.2014.v9.n4.a14

Revista Eletrônica Sistemas & Gestão
Volume 9, Número 4, 2014, pp. 566-574
DOI: 10.7177/sg.2014.v9.n4.a14

ou o índice silhueta. Dessa forma, o valor observado para a distância ou para um coeficiente quantifica o grau de similaridade ou dissimilaridade entre os vértices da rede. No presente trabalho, tanto as medidas de similaridade quanto esses coeficientes foram calculados utilizando um conjunto de funções disponíveis no pacote IGRAPH – pacote computacional disponível no software livre ***R***. A partir desses cálculos, foi possível detectar e analisar algumas comunidades associadas às redes representadas pelos respectivos grafos. Mais especificamente, nessas análises foi possível observar o número de comunidades, os vértices pertencentes a cada uma delas, a qualidade da partição e outras informações concernentes às redes.

Posteriormente, essas redes foram submetidas a um algoritmo não hierárquico de detecção de agrupamentos disponível no pacote CLUSTERSIM (no *software* R). Em geral, esses tipos de algoritmos trabalham, basicamente, com dois parâmetros de entrada, sejam eles: a matriz de atributos dos objetos e número de grupos (k). No caso do presente estudo, o número de grupos foi definido igual ao número de comunidades detectadas anteriormente. Como resultado, o algoritmo produz uma partição (grupos) que são constituídos pelos objetos da base de dados.

Uma vez produzida a partição e analogamente ao que foi feito na detecção das comunidades, são aplicados índices sobre essa partição de forma a estabelecer uma comparação com as medidas obtidas para as comunidades produzidas pelos algoritmos de detecção de comunidades.

O objetivo desse trabalho é estudar algoritmos para a formação de redes complexas. Para isso, na Seção 2 estão reunidos conceitos básicos sobre a Teoria dos Grafos. Na Seção 3, encontram-se as medidas de similaridade para analisar a homogeneidade entre os objetos e uma descrição sucinta dos principais métodos hierárquicos e não hierárquicos de agrupamento. Na Seção 4, são descritos os algoritmos de detecção de comunidades e na Seção 5 o algoritmo utilizado neste trabalho. Na Seção 6, são apresentados os resultados computacionais, considerando a aplicação desses algoritmos em algumas bases de dados reais, quais sejam: PIB dos 100 maiores municípios do Brasil; IDH por município do Rio de Janeiro; informações sobre recursos físicos na área de saúde pública e sistema de informação de atenção básica por municípios do Rio de Janeiro e as considerações finais.

2. CONCEITOS BÁSICOS SOBRE TEORIA DOS GRAFOS

Nesta seção, são apresentados os principais conceitos sobre Teoria dos Grafos utilizados ao longo do trabalho de acordo com Boaventura (1996).

Um grafo simples não orientado G consiste de dois conjuntos V e E, sendo V um conjunto finito e não vazio cujos elementos são denominados *vértices* e E um conjunto de subconjuntos de dois elementos de V cujos elementos são denominados *arestas*. O número de vértices determina a ordem do grafo e é denotado por $n = |V|$ e o número de arestas determina o tamanho do grafo, sendo denotado por $m = |E|$, onde $0 \leq m \leq n(n-1)/2$.

O grau de um vértice v_i, $d(v_i)$ é o número de arestas incidentes a ele e dois vértices são denominados *adjacentes* se existe uma aresta entre eles.

Um grafo $G_s(V_s, E_s)$ é dito um subgrafo de $G(V,E)$ se $V_s \subseteq V$ e $E_s \subseteq E$. Um caminho é uma sequência de vértices tal que, para cada um dos vértices, existe uma aresta para o vértice seguinte. Um caminho que começa e termina no mesmo vértice é denominado ciclo. Um grafo é *conexo* quando, a partir de qualquer um de seus vértices, é possível alcançar aos demais. Caso contrário, ele é denominado desconexo. A conectividade *de arestas de um grafo*, denotada por $\lambda(G)$, é o menor número de arestas que precisam ser retiradas do grafo para torná-lo desconexo. Um grafo é denominado completo e denotado por K_n, quando existir uma aresta para cada par de vértices.

Existem inúmeras maneiras de se representar a estrutura de um grafo. A representação matricial é frequentemente utilizada quando há a necessidade de realização de cálculos envolvendo dados estruturais. Dentre as matrizes mais conhecidas, estão as matrizes de adjacência, incidência, laplaciana e outras. Neste trabalho, é utilizada a matriz de adjacência, denotada por $A(G)$, sendo a entrada a_{ij} dessa matriz igual a 1 se v_i e v_j são adjacentes, ou iguais a 0, caso contrário, $\forall i, j = 1, \ldots, n$.

3. ANÁLISE DE AGRUPAMENTOS

Nesta seção, são apresentadas as medidas de similaridade para analisar a homogeneidade entre os objetos e os principais métodos hierárquicos e não hierárquicos de agrupamento de acordo com Frei (2006) e Hair *et al.* (2009).

Análise de agrupamentos é o nome dado a um conjunto de métodos que têm por objetivo reunir os objetos em grupos homogêneos, sendo essa homogeneidade avaliada através de uma medida de distância. Dada uma amostra de n objetos, cada um deles medindo p variáveis, procura-se um esquema de agrupamento em k grupos, Bussab (1990). Um primeiro passo à aplicação de qualquer algoritmo de agrupamento consiste da definição da matriz de dados. Mais

Revista Eletrônica Sistemas & Gestão
Volume 9, Número 4, 2014, pp. 566-574
DOI: 10.7177/sg.2014.v9.n4.a14

especificamente, selecionada a amostra de *n* objetos, devem-se definir quais atributos (variáveis) serão considerados para os objetos. Esses atributos são considerados no cálculo da medida de homogeneidade dos grupos. Dessa forma, os algoritmos de agrupamento trabalham com uma matriz A constituída por n linhas (objetos) e p colunas (variáveis). A similaridade entre objetos é obtida através de coeficientes específicos para cada tipo de variável.

Neste sentido, primeiramente obtém-se a matriz de dados; segundo, efetua-se a padronização dessa matriz (se necessário); terceiro, o cálculo da matriz distâncias; quarto, a utilização dos métodos de agrupamento; e por último decidir o nº de grupos. Em função dessas considerações, apresentamos nesta seção as medidas de homogeneidade adotadas, bem como os métodos hierárquicos e não hierárquicos, destacando-se o método *k-means*, Hair *et al.* (2009), que foi utilizado no presente trabalho.

3.1 Medidas de homogeneidade

Conforme comentado anteriormente, para avaliar a homogeneidade entre os objetos, utilizamos medidas de similaridade. Consideram-se dois tipos de similaridade: similaridade propriamente dita, que mede quão semelhantes são os objetos, e dissimilaridade, que mede quão diferentes são, sendo uma a complementar da outra. Dessa forma, quanto maior a similaridade, mais semelhante será, e quanto maior a dissimilaridade, menor a semelhança. Assim, definimos a matriz de similaridade, denotada por B, como sendo uma matriz de ordem $n \times p$ em que as n linhas correspondem aos objetos e as p colunas às variáveis.

Nessa matriz, cada entrada $b_{ij}(\forall i, j \;\; 1 \le i \le n, 1 \le j \le p)$ contém o valor do j-ésimo atributo associado ao i-ésimo objeto, conforme matriz abaixo.

$$B = \begin{bmatrix} b_{11} & \cdots & b_{1f} & \cdots & b_{1p} \\ \vdots & & \vdots & & \vdots \\ b_{i1} & & b_{if} & & b_{ip} \\ \vdots & & \vdots & & \vdots \\ b_{n1} & \cdots & b_{nf} & \cdots & b_{np} \end{bmatrix} \quad (1)$$

Supondo atributos quantitativos, as medidas de dissimilaridade mais utilizadas para dois objetos i e j com essa escala são:

- distância euclidiana, que é a distância geométrica dos pontos de coordenadas $(x_{i1},...,x_{ip})$ e $(x_{j1},...,x_{jp})$ representada por $d_{ij} = \sqrt{\sum_{f=1}^{p}(x_{if} - x_{jf})^2}$;
- distância de *Manhattan* $d_{ij} = \sum_{f=1}^{p} | x_{if} - x_{jf} |$;
- outras medidas podem ser construídas com base na distância euclidiana $d_{ij} = \sqrt{\sum_{f=1}^{p} w_f (x_{if} - x_{jf})^2}$, em que cada variável recebe um peso w_f de acordo com sua importância.

Em geral, quando os objetos têm apenas atributos quantitativos e esses atributos têm magnitudes diferentes, efetua-se uma padronização dos mesmos, Kaufman *et* Rousseeuw (1989). Ou seja, calcula-se a média e o desvio padrão associados a cada um dos *p* atributos e são aplicadas as equações definida a seguir:

$$z_{if} = (x_{if} - \bar{x}_f) / s_f \quad (1 \le f \le p, 1 \le i \le n) \quad (2)$$

$$\bar{x}_f = \frac{\sum_{i=1}^{n} x_{if}}{n} \quad (3)$$

$$s_f = \sqrt{\frac{\sum_{i=1}^{n}(x_{if} - \bar{x}_f)^2}{n-1}} \quad (4)$$

3.2 Métodos hierárquicos e não hierárquicos

Os métodos de agrupamento são divididos em hierárquicos e não-hierárquicos. No caso dos métodos hierárquicos, o número de grupos não é um parâmetro de entrada. Esses métodos são divididos em duas categorias, quais sejam: os aglomerativos e os divisivos. Nos aglomerativos, inicialmente há n grupos de 1 objeto cada, sendo efetuada uma série de uniões até obter k grupos; e nos divisivos, inicialmente há um único grupo formado por n objetos, sendo efetuadas sucessivas divisões dos grupos até que sejam atingidos k grupos.

Os agrupamentos gerados por esses métodos são representado sem uma estrutura na forma de árvore, denominada dendrograma que mostra como os objetos foram aglomerados (agrupados).

Dentre os métodos hierárquicos, podemos ressaltar os seguintes: Vizinho mais Próximo (*Single Linkage*), Vizinho mais Distante (*Complete Linkage*), Distância Média (*Average*

Revista Eletrônica Sistemas & Gestão
Volume 9, Número 4, 2014, pp. 566-574
DOI: 10.7177/sg.2014.v9.n4.a14

Linkage), Centróide, método de *Ward*, Otimização da Modularidade, entre outros. Nos métodos não hierárquicos, o número de grupos (k) é especificado previamente. Os grupos formados devem ter: Coesão interna (similaridade interna) e Isolamento (separação). Em geral, os algoritmos computacionais associados a estes métodos são iterativos. Além disso, esses métodos propiciam ter maior capacidade de análise do conjunto de dados de maior porte (com número maior de objetos). Um dos métodos mais conhecido e utilizado em relação aos métodos não hierárquicos é o método *k-means*, apresentado a seguir.

3.2.1 O método k-means

É um dos métodos mais conhecidos e mais utilizados em problemas práticos. As suas principais características são as seguintes: é suscetível a valores atípicos, trabalha com o conceito de centroide e com variáveis quantitativas, sendo de fácil implementação.

Como é um método não-hierárquico, é necessário fornecer *a priori* a quantidade de grupos desejados, sendo esses grupos denominados *clusters*. Assim como uma comunidade, um *cluster* é uma coleção de objetos que são similares entre si e diferentes dos objetos pertencentes a outros *clusters*. Para gerar os *clusters* e alocar os objetos aos *clusters*, avalia-se a distância euclidiana ao quadrado entre cada objeto e o centroide do grupo. Cada centroide é um vetor que contém as p médias em relação aos atributos dos objetos que estão associados a cada um dos grupos. Para a aplicação desse método, são considerados os quatro passos a seguir:

(i) separar os n objetos em k grupos de forma aleatória;

(ii)calcular o centroide de cada grupo $C_g (\forall g, 1 \leq g \leq k)$.

(iii) alocar os objetos ao seu centroide mais próximo, considerando a distância euclidiana ao quadrado, conforme fórmula definida a seguir.

$$\sum_{g=1}^{k} \sum_{x_i \in C_g} \sum_{f=1}^{p} (x_{if} - \bar{x}_{fg})^2 \qquad (5)$$

em que *k* é o número grupos, C_g o centroide do grupo g e *p* é o número de variáveis;

(iv)os passos (ii) e (iii) devem ser repetidos até que não haja uma mudança substancial nos valores dos centroides (considerando duas iterações seguidas).

3.3. Medida de qualidade de Agrupamento

Existem medidas, denominadas medidas de qualidade, que são aplicadas para avaliar *a posteriori* os *clusters* formados. Em particular, a medida utilizada neste trabalho é o coeficiente de silhueta, vide Newman (2004) e Kaufman *et* Rousseeuw(1989). Esse coeficiente permite avaliar, para cada objeto, a sua pertinência em relação ao *cluster*, bem como se os *clusters* formados têm uma estrutura natural de agrupamento. Ou seja, permite identificar se cada um dos objetos está bem posicionado em relação ao *cluster*.

Esta medida é calculada da seguinte maneira: para cada objeto x_i calcula-se o valor s_i dado por $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$, em que a_i é a dissimilaridade média do objeto x_i em relação a todos os outros objetos do *cluster* C_g que contém x_i, e b_i são a dissimilaridade média do objeto x_i em relação a todos os outros objetos do *cluster* mais próximo C_i (vizinho do objeto x_i). Observa-se que, de certa forma, o *cluster* vizinho corresponde ao segundo melhor *cluster* para alocar o objeto x_i. No caso em que o *cluster* C_g contém apenas um objeto x_i, o s_i é zero, isto é, $s_i = 0$.

Após o cálculo de todas as silhuetas $s_i (i = 1,\ldots,n)$, calcula-se a silhueta média associada à solução dada por:

$$\bar{s} = \frac{\sum_{i=1}^{n} s_i}{n} \qquad (6)$$

O coeficiente de silhueta assume valores no intervalo $[-1,1]$ e, de acordo com o valor assumido, há uma interpretação diferente, podendo essa avaliação ser feita por objeto e para a solução. No caso de um objeto, se ele é próximo de 1, significa que o objeto x_i está muito próximo dos objetos do seu grupo em comparação com o seu vizinho; caso seja em torno de zero, a_i e b_i são aproximadamente iguais indicando que o mesmo pode ser um objeto intermediário entre C_g e C_i ; e por fim, se for próximo de −1, o valor de b_i é muito menor do que o valor de a_i , ou seja, o objeto x_i está muito mais próximo do seu vizinho do que do *cluster* ao qual ele foi assinalado, ou seja, x_i foi alocado equivocadamente ao *cluster* C_g.

No caso da silhueta média, temos que um valor de $\bar{s}$ acima de 0.7 indica que há uma boa estrutura de agrupamento, já um valor entre 0.51 e 0.7 indica que há uma estrutura razoável de agrupamento e um valor entre 0.26 e 0.5 indica uma estrutura fraca, ou seja, é aconselhável a aplicação de outro método de agrupamento para os dados. E finalmente, os valores abaixo de 0.26 indicam que não há nenhuma estrutura de agrupamento nos dados, ou seja, não existem grupos naturais, Naldi (2011).

Revista Eletrônica Sistemas & Gestão
Volume 9, Número 4, 2014, pp. 566-574
DOI: 10.7177/sg.2014.v9.n4.a14

4. PROBLEMAS DE REDES E ALGORITMOS DE DETECÇÃO

As redes possuem conjuntos nos quais os nós estão mais fortemente conectados uns com os outros do que com o resto da rede. Esses conjuntos de nós são geralmente chamados de grupos (*clusters*), comunidades ou módulos Palla *et al*. (2005). O objetivo dos algoritmos de detecção de comunidades é encontrar esses grupos de nós em uma rede.

A detecção de comunidades é um problema parecido ao problema de corte em grafos ou de particionamento (Karger (2000), Kernighan *et* Lin (1970), Fiduccia *et* Mattheyses (1982)). Por sua vez, o problema de particionamento em grafos é, em geral, definido como o problema de particionamento em k grupos de tamanho muito parecido ou igual, minimizando o nº de arestas entre os grupos. Esse problema é definido como NP-*hard*, pois não existe nenhum algoritmo na literatura que encontre a solução ótima em tempo polinomial. Sendo assim, para muitos casos reais nos quais o número de nós da rede é da ordem de centenas ou milhares e o grafo tem alta densidade (número de arestas), torna-se necessária a aplicação de algoritmos heurísticos para encontrar uma solução viável em um tempo computacional factível. Considerando esta última observação, encontramos na literatura o algoritmo de *Lin-Kernighan* (Kernighan *et* Lin, (1970)), *SpectralBisection*, Barnes (1982); além dos chamados algoritmos geométricos ou algoritmos multinível, Pothen (1997).

O objetivo do particionamento em grafos é dividir qualquer grafo em grupos de tamanho similar. A detecção de comunidades, por outro lado, tem por objetivo a formação de grupos que tenham similaridade entre seus nós, ou seja, que eles desempenhem funções similares dentro do seu grupo. Além disso, tanto o número de comunidades em uma rede quanto o seu tamanho não são conhecidos *a priori*, sendo esses parâmetros determinados mediante a aplicação de algoritmos de detecção de comunidades. Portanto, dada uma rede associada a um grafo $G(V,E)$ com n nós e m arestas, é possível encontrar subgrupos de nós a partir de qualquer algoritmo de detecção de comunidades. Se considerarmos que $C_1, C_2, \ldots, C_k$ são as comunidades encontradas, elas devem satisfazer as seguintes propriedades: $C_i \cap C_j = 0$, $para\ i \neq j$ e $\bigcup C_i$ =V, vide Fiduccia *et* Mattheyses (1982).

A Figura 1 representa uma rede Simples com três comunidades incluídas nos círculos pontilhados, as quais têm uma alta densidade de conexões internas (várias arestas) e baixa densidade de conexões externas, que se encontra em Fortunato *et* Castellano(2012).

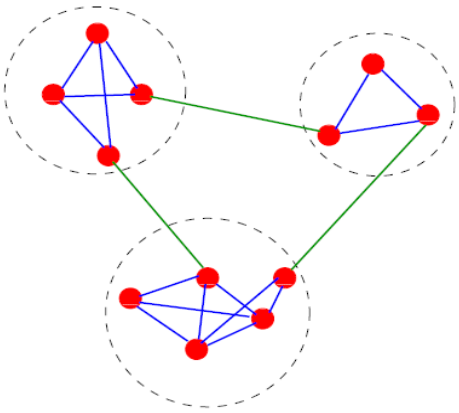

Figura 1 – Exemplo de Três Comunidades

Fonte: Elaborada pelos autores (2013)

Um dos principais problemas que se apresenta à detecção de comunidades é que, em redes reais, normalmente são um conhecimento prévio no que concerne ao número e ao tamanho das comunidades existentes. Consequentemente, não há nenhuma regra que permita definir a melhor divisão da rede em suas comunidades. Para resolver esse problema, Newman *et* Girvan (2004) propuseram uma medida, denominada Modularidade, definida na próxima subseção.

4.1 Modularidade

A modularidade Q é uma medida originalmente criada para definir um critério de parada para o algoritmo de Newman *et* Girvan(2004). Ela representa uma das primeiras tentativas de alcançar uma compreensão do princípio do problema de agrupamento em grafos, baseada no conceito de comunidade. Devido a sua eficácia no que concerne ao tempo de processamento, essa medida foi rapidamente difundida em teoria dos grafos e tornou-se um elemento fundamental de muitos métodos de agrupamento.

Considere uma divisão particular de uma rede em k comunidades e uma matriz simétrica *E* de tamanho $k \times k$, sendo cada elemento e_{ij} correspondente à proporção de todas as arestas na rede que unem os vértices da comunidade C_i com os vértices da comunidade C_j. O traço dessa matriz, denotado por $tr(E) = \sum_i e_{ii}$, dá a fração de arestas na rede que conectam aos vértices da mesma comunidade C_i. Fica claro que uma boa divisão em comunidades deve ter um valor elevado para esse traço. Define-se a soma das linhas (ou colunas) $a_i = \sum_j e_{ij}$, que representa a proporção de arestas que conectam os vértices da comunidade C_i com as outras comunidades C_j. Assim, definimos a medida de modularidade como, $Q = \sum_k (e_{ii} - a_i^2) = Tr\, e - \|e\|$, na qual $\|e\| = (\sum_{ij}^k e_{ij})^2$ é a soma dos elementos da matriz E.

Revista Eletrônica Sistemas & Gestão
Volume 9, Número 4, 2014, pp. 566-574
DOI: 10.7177/sg.2014.v9.n4.a14

Teoricamente, esta quantidade mede a proporção de arestas na rede que conectam vértices da mesma comunidade, menos o valor esperado da mesma quantidade na rede com as mesmas divisões de comunidades, mas com conexões aleatórias entre os vértices. Se o número de arestas em uma comunidade não é maior que o valor aleatório, então o valor da modularidade é $Q = 0$. Valores que se aproximam a $Q = 1$, o qual é máximo, indicam uma estrutura forte de comunidade. Em muitas aplicações reais, os valores de *Q* (para as redes associadas) não ultrapassam 0.7 e valores mais altos não são frequentes.

Há muitos algoritmos de detecção de comunidades que foram formulados e baseados nessa medida. Dentre esses algoritmos, destaca-se o *FastGreedy algorithm*, Clauset *et al.*(2004), que foi o algoritmo utilizado no presente trabalho para efetuar a detecção das comunidades e é apresentado na próxima subseção.

4.2 FastGreedy

O algoritmo *FastGreedy* foi proposto por Clauset *et al.* (2004) e é uma versão aperfeiçoada do algoritmo proposto por Newman (2004), que utiliza uma estratégia gulosa para a detecção de uma comunidade. Considerando a primeira iteração inicial desse algoritmo, cada vértice da rede é associado a um único objeto de uma comunidade. E, em iterações posteriores, os objetos vão sendo agregados (unidos), definindo, dessa forma, novas comunidades com número maior de objetos. Essa agregação (junção) produz um maior incremento no valor da modularidade Q.

Especificamente, para uma rede com n vértices após $n-1$ uniões, o algoritmo terá apenas uma comunidade, e termina. O método pode ser representado como uma árvore, na qual as folhas são os vértices da rede original e os nós internos as correspondentes uniões. A árvore é chamada de dendrograma, e representa uma decomposição hierárquica da rede em comunidades, como pode ser visto na Figura2, em que os círculos na parte inferior da figura representam os vértices individuais da rede. Os vértices se unem para formar comunidades cada vez maiores como indicado pelas linhas, até chegar ao topo, onde todos estão unidos em uma única comunidade. Um corte transversal da árvore em qualquer nível, tal como indicado pela linha tracejada, fornecerá quantas comunidades há nesse nível, bem como os objetos que estão alocados a cada uma dessas comunidades. No caso deste particular exemplo, a aplicação do corte associado à linha vermelha (tracejada) implica na definição de quatro comunidades, e aplicação do corte associado à linha azul (contínua) implica na definição de nove comunidades.

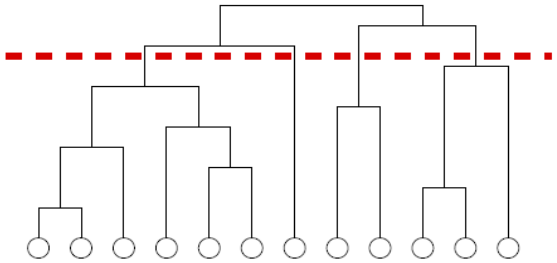

Figura 2 – Exemplo de Dendrograma

Fonte: Elaborada pelos autores (2013)

A mais simples implementação dessa ideia é armazenar a matriz de adjacência do grafo como um vetor de inteiros e, iterativamente, combinar pares de linhas e colunas cada vez que as correspondentes comunidades são unidas. A implementação dessa estratégia tende a consumir uma boa quantidade de tempo e espaço de memória. Neste sentido, o algoritmo proposto por Clauset *et al.* (2004) diminui o tempo de processamento mediante o uso de estruturas de dados mais sofisticadas. O funcionamento do algoritmo envolve encontrar as mudanças Δ na modularidade Q, o que resulta da fusão de cada par de comunidades e escolher a maior delas e realizar a fusão correspondente.

No que diz respeito ao algoritmo *FastGreedy*, ao invés de calcular continuamente a matriz de adjacência e calcular ΔQ_{ij}, é apenas utilizado um procedimento que é responsável pelo armazenamento e atualização do valor ΔQ_{ij} da matriz. A racionalidade desse procedimento consiste na observação de que juntar duas comunidades sem uma aresta comum não produz incremento em Q_{ij}, ou seja, basta apenas armazenar ΔQ_{ij} para os pares de comunidades i, j que estão ligados por uma ou mais arestas. Além disso, o algoritmo usa eficientes estruturas de dados para armazenar valores grandes de ΔQ_{ij}. Como foi mencionado anteriormente, o algoritmo *FastGreedy* começa com cada vértice sendo o único membro de uma comunidade, na qual a proporção de arestas e_{ij} que unem vértices na comunidade C_i com os vértices da comunidade C_j é $\frac{1}{2m}$, se existe uma conexão entre as comunidades C_i e C_j e 0 caso contrário. Assim, $a_i = \frac{d(i)}{2m}$, em que $d(i)$ é o grau do vértice.

Dessa forma, pode-se definir inicialmente:

$$\Delta Q_{ij} = \begin{cases} \frac{1}{2m} - \frac{d(i)d(j)_j}{(2m)^2} \; se\; i, j\; estão\; conectados \\ 0, caso\; contrário \end{cases} \quad (7)$$

Então, para cada comunidade C_i,o algoritmo pode ser definido da seguinte maneira:

(i) Calcular os valores iniciais de ΔQ_{ij} e a_i de acordo

Revista Eletrônica Sistemas & Gestão
Volume 9, Número 4, 2014, pp. 566-574
DOI: 10.7177/sg.2014.v9.n4.a14

com a equação 4.2.1 preenchendo uma *max-heap* H com o maior elemento de cada linda da matriz ΔQ_{ij}.

Obs: O *max-heap* é uma estrutura de dados organizada como uma árvore binária, em que o valor de todos os nós são menores que os de seus respectivos pais. Assim, em um *heap* de máximo, o maior valor do conjunto está na raiz da árvore.

(ii) Escolher o maior ΔQ_{ij} da *max-heap* H, iuntar a correspondente comunidade, atualizar a matriz ΔQ_{ij}, a *max-heap* H, o vetor a_i e incrementar Q mediante ΔQ_{ij}.

(iii) Repetir o passo *(ii)* até obter-se apenas uma comunidade.

5. ALGORITMO UTILIZADO

Para a realização desse estudo, foram utilizados alguns pacotes computacionais disponíveis no *software* livre R, Matloff (2011), mais especificamente, os pacotes IGRAPH e o CLUSTERSIM.

Duas funções foram fundamentais para a execução do trabalho: *fastgreedy.community e k-means*. A primeira pertence ao pacote IGRAPH e, como o seu nome já sugere, é estruturada conforme o algoritmo *FastGreedy*, proposto por Clauset *et al.* (2004). Essa função retorna o número de comunidades detectadas, o valor da modularidade e o vetor de amizades (um vetor cujo valor em cada posição representa a comunidade a qual o objeto está alocado).

Já a segunda pertence ao pacote CLUSTERSIM, a qual também já foi descrita na Seção 3.2.1, e tem a finalidade de determinar *clusters* de acordo com um nº k pré-determinado.

O primeiro passo para a aplicação da metodologia foi selecionar as bases de dados que seriam usadas no trabalho, quais sejam: (1) Dados do PIB dos 100 maiores municípios brasileiros; (2) O IDH dos municípios do Rio de Janeiro; (3) Informações sobre recursos físicos na área de saúde pública e sistema de informação de atenção básica por municípios do Rio de Janeiro.

De forma a manipular e aplicar os algoritmos de detecção e de *clustering*, foi implementada uma função em R que importa cada uma dessas bases e que aplica os seguintes procedimentos:

- padronização dos dados;
- construção da matriz de distâncias, considerando as distâncias entre todos os objetos tomados dois a dois;
- determinação de uma matriz de adjacências dos dados. Neste ponto, cada objeto da base corresponderá a um vértice do grafo e dois objetos serão conectados por uma aresta se a distância entre eles for menor que o 3º quartil calculado em função da matriz de distâncias. Considerando esse critério, foram removidos do grafo os vértices que não tinham nenhuma aresta;
- aplicar o algoritmo de detecção de comunidades;
- determinar *clusters* conforme o nº de comunidades encontradas no passo anterior, mediante a aplicação do algoritmo *k-means*.

6. RESULTADOS COMPUTACIONAIS E CONSIDERAÇÕES FINAIS

A presente seção traz alguns resultados computacionais iniciais concernentes à aplicação da função descrita na seção anterior. Observa-se que todos os experimentos computacionais foram realizados em um computador dotado de um processador intel de 2.2Ghz e com 4 GB de memória RAM e sistema operacional Windows 7. Nestes experimentos foram utilizadas três bases de dados, a saber: PIB (IBGE, 2010), IDH (PNUD, 2000) e SAUDERJ (DATASUS, 2013). A primeira base de dados contém o valor dos PIBs (ano de 2010) correspondentes aos municípios brasileiros com os 100 maiores PIBs; a segunda contém o IDH (índice de desenvolvimento humano) (ano de 2000) dos municípios do Rio de Janeiro e a terceira contém informações sobre recursos físicos na área de saúde pública e sistema de informação de atenção básica por municípios do RJ (ano de 2013).

A Tabela 1 a seguir traz o número de objetos (vértices do grafo) de cada uma das bases de dados, o número de arestas e a densidade do grafo.

Tabela 1. Informações sobre os grafos associados às bases de dados

Base de dados	Vértices de G (n)	Arestas de G(m)	Densidade de G $\left(2m/n(n-1)\right)$
PIB	100	4450	0.8989899
IDH	90	3597	0.8981273
SAUDERJ	68	679	0.2980685

Fonte: Elaborado a partir de IBGE (2010), PNUD (2000) – Programa das Nações Unidas para o Desenvolvimento e DATASUS (2013).

As Tabelas 2, 3, 4, 5, 6 e 7 abaixo sumarizam os resultados obtidos a partir da aplicação do algoritmo de detecção de comunidades (*fastgreedy*) e do algoritmo *k-means* em cada uma das bases de dados descritas na seção anterior.

Revista Eletrônica Sistemas & Gestão
Volume 9, Número 4, 2014, pp. 566-574
DOI: 10.7177/sg.2014.v9.n4.a14

Tabela 2. Resultados do Algoritmo *FastGreedy* - Dados PIB

Número de Comunidades	**Modularidade**	**Número de Objetos por Comunidade**		
		1	**2**	**3**
3	0.04934372	61	34	5

Fonte: Elaborado a partir de IBGE (2010).

Tabela 3. Resultados do Algoritmo *k-means* - Dados PIB

Número de Comunidades	**Silhueta**	**Número de Objetos por Comunidade**		
		1	**2**	**3**
3	0.6589658	74	21	5

Fonte: Elaborado a partir de IBGE (2010).

Tabela 4. Resultados do Algoritmo *FastGreedy* - Dados IDH

Número de Comunidades	Modularidade	Número de Objetos por Comunidade	
		1	**2**
2	0.127555	45	45

Fonte: Elaborado a partir de PNUD (2000).

Tabela 5. Resultados do Algoritmo *k-means* - Dados IDH

Número de Comunidades	**Silhueta**	**Número de Objetos por Comunidade**	
		1	**2**
2	0.3445603	45	45

Fonte: Elaborado a partir de PNUD (2000).

Tabela 6. Resultados do Algoritmo *FastGreedy*– Dados SAUDERJ

Número de Comunidades	**Modularidade**	**Número de Objetos por Comunidade**	
		1	**2**
2	0.08556735	38	30

Fonte: Elaborado a partir de DATASUS (2013).

Tabela 7. Resultados do Algoritmo *k-means* – Dados SAUDERJ

Número de Comunidades	**Silhueta**	**Número de Objetos por Comunidade**	
		1	**2**
2	0.3445603	34	34

Fonte: Elaborado a partir de DATASUS (2013).

Revista Eletrônica Sistemas & Gestão
Volume 9, Número 4, 2014, pp. 566-574
DOI: 10.7177/sg.2014.v9.n4.a14

Uma análise da Tabela 1 mostra que, das três bases de dados consideradas, duas estão associadas a um grafo com razoável densidade. Além disso, analisando as Tabelas de 2 até 7, pode-se verificar que, a partir do número de grupos fornecidos pelo algoritmo de detecção de comunidades, foi possível aplicar o algoritmo de *clustering* e produzir agrupamentos de qualidades razoável e boa no que concerne ao valor da silhueta.

Em experimentos futuros, pretende-se trabalhar com o número substancial de bases de dados associadas com grafos de diferentes densidades. Também é objetivo futuro considerar a aplicação e a comparação entre outros algoritmos de detecção de comunidades e de *clustering*, bem como de outros índices para avaliação da qualidade dos grupos.

7. REFERÊNCIAS

Barnes, E. R. (1982), "An algorithm for partitioning the nodes of a graph", *SIAM Journal on Algebraic and Discrete Methods*, Vol. 3 No.4, pp. 541-550.

Boaventura, Netto (1996), P.O. Grafos: Teoria, Modelos, Algoritmos, 3ªEd., Edgard Blucher, São Paulo.

Bussab, W. O. (1990), Introdução à análise de agrupamentos, USP/Instituto Militar de Engenharia.

Clauset, A., Newman, M. E., Moore, C. (2004), "Finding community structure in very large networks", *Physical Review E*., Vol 70, Nº. 6.

DATASUS (2013), Ministério da Saúde / DATASUS - Departamento de Informática do SUS. Dados acessados em http://www2.datasus.gov.br/DATASUS/index.php?area=02.

Fiduccia, C., Mattheyses, R. M. (1982), "A linear-time heuristic for improving network partitions",In: 19th *IEE Conference on Design Automation*.

Fortunato, S.; Castellano, C. (2012), "Community structure in graphs", *Computation Complexity*, pp. 490-512.

Frei, F. (2006), Introdução à análise de agrupamentos – Teoria e prática, Editora Unesp.

Hair, J.F, Black, W.C, Babin, B.J., Anderson, R.E. e Tatham, R.L. (2009), Análise Multivariada de Dados, 6ª ed., Bookman.

IBGE (2010). Produto Interno Bruto dos Municípios Brasileiros. Dados acessados em http://www.ibge.gov.br/home/estatistica/economia/pibmunicipios/2010/.

Karger, D. (2000), "Minimum cuts in near-linear time", *Journal of the ACM* (JACM), Vol. 47 No. 1, pp. 46-76.

Kaufman L.,Rousseeuw P.J. (1989), Finding Groups in Data – An Introduction to Cluster Analysis, Wiley – Interscience Publication.

Kernighan, B., Lin, S. (1970), "An efficient heuristic procedure for partitioning graphs", *Bell System Technical Journal*, Vol.49 No. 2, pp. 291-307.

Matloff, N. (2011), The Art of R Programming, Publisher: William Pollock, San Francisco.

Naldi, C. N. (2011), Técnicas de Combinação para Agrupamento Centralizado e Distribuído de Dados, Tese de Doutorado, USP, São Carlos.

Newman, M. E. J. (2004), "Fast algorithm for detecting community structure in networks", *Physical Review* E69, 066133.

Newman, M., Girvan, M. (2004), "Finding and evaluating community structure in networks", *Physical Review* E 69, 026113.

Palla, G., Derényi, I., Farkas, I., Vicsek, T. (2005), "Uncovering the Uncovering the overlapping community structure of complex networks in nature and society", *American Journal Experts*, Vol. 435, pp. 814-818.

PNUD (2000), *Programa das Nações Unidas para o Desenvolvimento*. Dados acessados em http://www.pnud.org.br/atlas/ranking/Ranking-IDHM-Municipios-2000.aspx.

Pothen, A. (1997), "Graph Partitioning Algorithms with Applications to Scientific Computing", ICASE LaRC *Interdisciplinary Series in Science and Engineering*, Vol. 4, pp. 323-368.