Medidas de Centralidade em Grafos e Aplicações em redes de dados

Elizandro Max Borba, Vilmar Trevisan,
Departamento de Matemática Pura e Aplicada, PPGMAp, UFRGS
Av. Bento Gonçalves, 9500 - Prédio 43111 - Agronomia
91509-900, Porto Alegre, RS
E-mail: elizandro.max@ufrgs.br

***Resumo**: A Análise de Redes trata do estudo da estrutura de uma rede a fim de obter informações importantes sobre seus elementos e suas interações. Um aspecto relevante da análise de uma rede é decidir quais são os elementos mais importantes ou centrais de uma rede, através do uso das medidas de centralidade. Neste trabalho, apresentamos um survey sobre as principais medidas de centralidade, mostrando suas motivações e definições. Em seguida, apresentaremos dois exemplos de aplicações das centralidades às redes de dados: a obtenção da estrutura de comunidades de uma rede de roteadores e a avaliação dos pontos de vulnerabilidade de uma rede.*

1. Introdução

Um grafo é um par ordenado de conjuntos $G = (V, E)$ onde $E = \{ij = \{i, j\}\}$, sendo que i, j devem pertencer a V. Os elementos $v \in V$ são chamados de *vértice* ou *nós* de G, e os elementos $e \in E$ são chamados de *arestas* ou *links*. Grafos são representados graficamente por pontos ligados por linhas. Outras formas de representação incluem listas e matrizes, como a matriz de adjacência $A_{n\times n}$, definida por $a_{ij} = 1$ se $\exists\ ij$, e por zero nas demais posições.

Diversos sistemas no mundo real podem ser representados através de redes. Dessa forma, grafos são uma forma natural de representar matematicamente esses sistemas. A *Análise de Redes* é a área do conhecimento que investiga a estrutura de uma rede a fim de obter informações importantes sobre seus elementos e suas interações.

Nas últimas décadas, o interesse nessa área tem crescido. As redes tecnológicas têm tido um lugar de destaque, especialmente com o advento e a popularização da Internet. As *redes de dados*, cujos elementos mais importantes são os roteadores que direcionam o tráfego de dados na Internet e em redes domésticas e corporativas, são de especial interesse.

Um aspecto relevante dessa análise é decidir quais são os elementos mais importantes ou centrais de uma rede. As *medidas de centralidade* são uma forma de quantificar essa importância. Desde a década de 1950, várias centralidades surgiram na literatura, e várias aplicações têm surgido. Neste trabalho, apresentamos dois exemplos de aplicações das medidas de centralidade no âmbito específico das redes de dados. Inicialmente, utilizamos o algoritmo de Girván-Newman para obter a estrutura de comunidades de uma rede, a fim de sugerir uma distribuição de grupos de roteadores entre equipes que as administrarão. A seguir, apresentamos um algoritmo de simulação de ataque que tem o objetivo de identificar pontos de vulnerabilidade de uma rede.

2. Medidas de Centralidade

A noção de centralidade, em várias aplicações, é associada à importância do elemento na estrutura. Seja um grafo $G = (V, E)$, com $|V| = n$, $|E| = m$ e matriz de adjacência A. Uma *medida de centralidade de nó* é uma função $c_X: V \to \mathbb{R}$ tal que a relação de ordem entre $c_X(i)$ e $c_X(j)$ deve refletir a percepção de que i é *mais central* que j. De maneira análoga, podem-se definir *medidas de centralidade de aresta* como $c_Y: E \to \mathbb{R}$, ou ainda combinar os valores dos vértices ou arestas a fim de se obter uma *medida de centralização* do grafo como um todo. Usaremos a notação $\mathbf{c}_X$ para referenciar o vetor coluna contendo as centralidades de todos os vértices, ou seja, $\mathbf{c}_X = \left(c_X(v_1), \ldots, c_X(v_n)\right)^T$.

DOI: 10.5540/03.2014.002.01.0088

De acordo com Freeman [4], as medidas de centralidade de nó básicas são a *centralidade de grau*, a *centralidade de proximidade* e a *centralidade de intermediação*. Borgatti e Everett [2] usam essas medidas para categorizar outras encontradas na literatura em centralidades similares à de grau, similares à de proximidade e similares à de intermediação.

É comum que medidas de centralidade partam de uma medida *absoluta* e sejam normalizadas por uma cota superior $r \geq c_{\mathrm{X}}$, de modo que $c'_{\mathrm{X}} = c_{\mathrm{X}}/r \leq 1$; c'_{X} então é dita uma medida *relativa*.

2.1 As três medidas básicas: grau, proximidade e intermediação

A medida de centralidade mais básica reflete a ideia de que *"um nó importante está conectado com muitos nós."* A *centralidade de grau* de um vértice v é dada por seu grau, ou seja, $c_{\mathrm{D}}(v) = \deg v$. A partir de A, podemos calcular $c_{\mathrm{D}}(v_i) = \sum_j a_{ij}$ e o vetor de centralidades como $\mathbf{c}_{\mathrm{D}} = A\vec{1}$.

Em vários contextos, entretanto, mais importante que ter muitas conexões é não estar longe demais dos demais nós, ou seja, *"um nó importante está próximo dos outros nós."* A *centralidade de proximidade* de um vértice v é dada pelo recíproco da soma das suas distâncias aos demais nós, ou seja, $c_{\mathrm{C}}(v) = \frac{1}{\sum_{a \in V} d(v,a)}$.

Outra motivação de centralidade é a ideia de mediação; por exemplo, uma cidade que faz parte de várias rotas comerciais tem certamente uma vantagem estratégica. Isto motiva a seguinte ideia de que *"um nó importante faz parte de muitos caminhos."* Dados $v, a, b \in V$, sejam g_{ab} o número de geodésicas entre a e b, e g_{avb} o número dessas que passam por v. A *centralidade de intermediação* do vértice v é dada por $c_{\mathrm{B}}(v) = \sum_{a,b \neq v} \frac{g_{avb}}{g_{ab}}$.

2.2 Outras medidas de centralidade

Das três medidas básicas, diversas outras porem ser derivadas. Faremos aqui um apanhado de medidas encontradas na literatura, seguindo a classificação de centralidades sugerida por Borgatti e Everett [2], separando as medidas em centralidades similares à de grau, similares à de proximidade e similares à de intermediação. Separada dessas há também a classe das centralidades delta, introduzidas por Latora e Marchiori [9].

Medidas similares à de grau: Note-se que centralidade de grau pode ser facilmente calculada a partir de A. Um resultado conhecido é que $(A^p)_{ij}$ dá o número de passeios de tamanho p entre v_i e v_j. Para grafos orientados, A^p e $(A^T)^p$ contêm, respectivamente, as quantidades de passeios orientados "para frente" e "para trás". Se pensarmos $c_{\mathrm{D}}(v)$ como a quantidade de passeios de tamanho 1 saindo de v, podemos expandir essa noção para passeios de um tamanho k qualquer. Isso motiva a definição a seguir:

Definição: a *centralidade de k-passeio* é dada por $C_{\mathrm{k-walk}} = W_k\vec{1}$, onde $W_k = \sum_{j=1}^{k} A^j$, ou seja, $W_k[i,j]$ contém o número de passeios de tamanho k ou menos entre v_i e v_j.

Pode-se também fazer uma combinação linear dos números de passeios de vários tamanhos, ou mesmo todos os tamanhos possíveis, por exemplo, atribuindo um peso α^k aos passeios de tamanho k. Seja $K = \sum_{k=1}^{\infty} \alpha^k A^k$. A *centralidade de Katz* [7] é dada por $\mathbf{c}_{\mathrm{Katz}} = K\vec{1}$. Se $|\alpha| < 1/\lambda_1$, onde λ_1 é o maior autovalor de A, então $\mathbf{c}_{\mathrm{Katz}} = ((I - \alpha A)^{-1} - I)\vec{1}$.

Outra maneira de interpretar a centralidade é pensar que a centralidade de um nó seja uma função da centralidade dos seus vizinhos, ou seja, que *"um nó importante tem vizinhos importantes*." Se considerarmos que a centralidade $c_x(v_i)$ é proporcional (por um fator α) à soma das centralidades de seus vizinhos, temos que $\mathbf{c}_{\mathrm{X}} = \alpha A\mathbf{c}_{\mathrm{X}}$, ou seja, α^{-1} é um autovalor de A e $\mathbf{c}_{\mathrm{X}}$ é o autovetor correspondente. Bonacich [1] definiu a *centralidade de autovetor* como $\mathbf{c}_{\mathrm{eig}} = \mathbf{v}_1$, onde $\mathbf{v}_1$ é o autovetor unitário positivo associado a λ_1.

Uma vantagem desse método é a possibilidade de uso de métodos numéricos (como o método das potências) para a determinação do autovetor. Um problema da abordagem usada na centralidade de autovetor é que, se um nó importante é vizinho de um grande número de nós, todos eles ganham importância. Por exemplo, se eu tenho minha página hospedada no *Yahoo!*, eu ganho importância em função disso - assim como milhões de outros sites. Além disso, nesse

DOI: 10.5540/03.2014.002.01.0088

esquema, uma página poderia inflar artificialmente sua importância gerando uma grande quantidade de links. Seria razoável sugerir que a importância do nó fosse "diluída" entre seus links. Essa é a principal ideia do algoritmo *PageRank*. Similarmente ao que foi feito com $\mathbf{c}_{\text{eig}}$, montamos a equação $\mathbf{c}_{\text{X}} = \alpha AD^{-1}\mathbf{c}_{\text{X}}$, onde D é uma matriz diagonal tal que $D_{ii} = \max\{k_i^{out}, 1\}$. Portanto, a centralidade *PageRank* é um autovetor da matriz αAD^{-1}. Para uma discussão mais aprofundada dessa centralidade, ver [10].

Medidas similares à de proximidade: Outra abordagem possível é primeiro designar o nó (ou conjunto de nós) mais central, e a partir daí usar a distância de cada nó até esse nó (ou conjunto de nós) central. A *centralidade de centroide* é dada por $c_{\text{cent}}(v) = d(v, c(G))$, onde $c(G)$ é o centroide do grafo. A escolha natural é o próprio centro do grafo (o conjunto de vértices de menor excentricidade), porém outro critério poderia ser usado, como até mesmo outra medida de centralidade.

Medidas similares à de intermediação: Muitas vezes a interação entre os nós ocorrem por outros caminhos além das geodésicas. Ford e Fulkerson [5] desenvolveram um modelo de *fluxo* entre dois nós. Freeman, Borgatti e White [3] definiram *centralidade de fluxo* do vértice v como $c_{\text{F}}(v) = \sum_{a,b\neq v} m_{avb}$, onde m_{avb} é o fluxo máximo entre a e b passando por v.

Centralidades delta: Essa nova classe de medidas [9] parte do pressuposto de que *"um nó importante afeta a rede se retirado dela.*" Para tal, define-se uma *medida de coesão* para o grafo como um todo, denotada por $P(G)$. Assim, um nó seria mais central quanto mais decaísse essa coesão em consequência de sua retirada da rede. A *centralidade delta* é dada por $c_{\Delta}(v) = 1 - P(G - v)/P(G)$.

Uma alternativa para P é usar a *eficiência* [8] da rede, definida por $\mathcal{E}(G) = \frac{1}{n(n-1)}\sum_{\substack{u,v\in V \\ u\neq v}} \frac{1}{d(u,v)}$. A eficiência assume que a interação (como um fluxo de informação) entre dois nós é tão eficiente quanto menor é a distância entre eles. A *centralidade de informação* é dada por $c_{\text{inf}}(v) = 1 - \mathcal{E}(G - v)/\mathcal{E}(G)$

3. Aplicações em redes de dados

No contexto de redes, *dados* são informações codificadas em dígitos binários (*bits*) agrupados em *pacotes*, mais comumente codificados pelo protocolo IP (*Internet Protocol*), e por isso essas redes são chamadas de *redes IP*. Os pacotes são direcionados por *roteadores*, e assim a rede formada pelos roteadores é a parte mais importante de uma rede de dados.

3.1 Detecção de comunidades

Em diversas redes, é típica a existência de grupos de nós que apresentam uma maior interação dentro do grupo, em comparação com as interações dentro de cada grupo. Por exemplo, em redes sociais, certos grupos podem ter interesses em comum, como um artista ou time preferido, ou uma área do conhecimento. Denominamos esses grupos *comunidades*. Esse conceito é facilmente estendido para redes biológicas, tecnológicas e outras. Uma *partição* é uma divisão específica de uma rede em comunidades.

Vários métodos são propostos na literatura para a detecção de comunidades, tais como subgrafos maximais, separação por cortes mínimos ou Aglomeração Hierárquica. Girván e Newman [6] definiram a *centralidade de intermediação* do link l como $c_{\text{BL}}(l) = \sum_{a,b} \frac{g_{alb}}{g_{ab}}$, onde g_{ab} é o número de geodésicas entre a e b, e g_{alb} é o número dessas que contêm l. Esperamos que os links de maior c_{BL} sejam justamente aqueles interligam comunidades. O algoritmo de Girván-Newman (ou *algoritmo GN*), é descrito a seguir.

Algoritmo 1 (GN):

1) Calcular c_{BL} para todos os links da rede;
2) Remover o link com maior c_{BL};
3) Recalcular c_{BL} para todos os links afetados por essa remoção;
4) Repetir a partir do passo 2 até não sobrarem mais links.

Comunidades em uma rede de dados: Os roteadores de uma rede de dados podem ser agrupados em comunidades a fim de atender a diferentes propósitos, como definir grupos de roteadores, de modo que cada grupo fique sob a administração de uma determinada equipe.

DOI: 10.5540/03.2014.002.01.0088

Como exemplo inicial, consideremos o caso da Figura 1, uma rede pequena hipotética cuja estrutura de comunidades é bastante aparente.

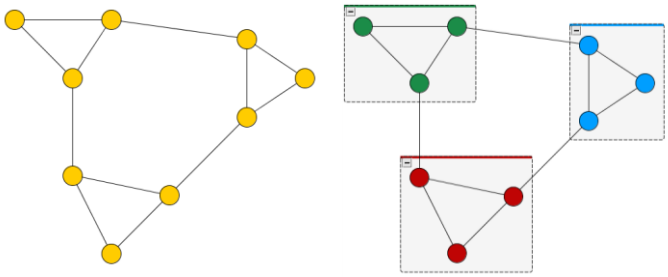

Figura 1: Uma pequena rede antes e depois do algoritmo GN executado pelo software *yEd*.

Apresentamos, por fim, uma rede adaptada de uma porção do backbone de uma empresa de tecnologia. Após aplicado o algoritmo GN, a correspondência com a real divisão de administração dos grupos de roteadores foi quase total: a única diferença foi que os dois pequenos grupos nos cantos superiores da Figura 2 estão, na realidade, juntos com o grupo maior (na parte central, acima).

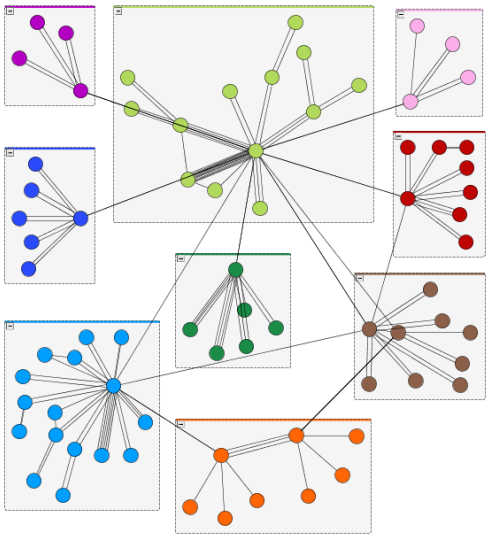

Figura 2: Rede de roteadores após a aplicação do algoritmo GN.

3.2 Análise da vulnerabilidade de uma rede de dados

Ataques em redes são fruto de ação intencional e maliciosa que visam links ou nós. Por exemplo, links podem ser vítimas de atos de vandalismo ou sabotagem; no caso de uma guerra, as instalações onde se localizam os nós podem ser alvo de bombardeios. Dizemos que a rede é *vulnerável* a esses ataques. Hoje em dia tornou-se comum a ocorrência dos chamados *cyberataques*, que são ataques que fazem uso exclusivamente dos protocolos de comunicação de dados para incapacitar nós da rede.

Nossa análise consistirá em assumir a importância de nós a partir de suas centralidades, a fim de escolher os "alvos" mais atraentes para um possível ataque. Usaremos o software *yEd*, que implementa o cálculo das centralidades de grau, proximidade, intermediação e autovetor, possibilitando boa visualização dos resultados, permitindo alterar o aspecto (tamanho ou cor) do nó conforme seu índice de centralidade. O resultado na análise das medidas de centralidade da rede de exemplo está na Figura 3.

DOI: 10.5540/03.2014.002.01.0088

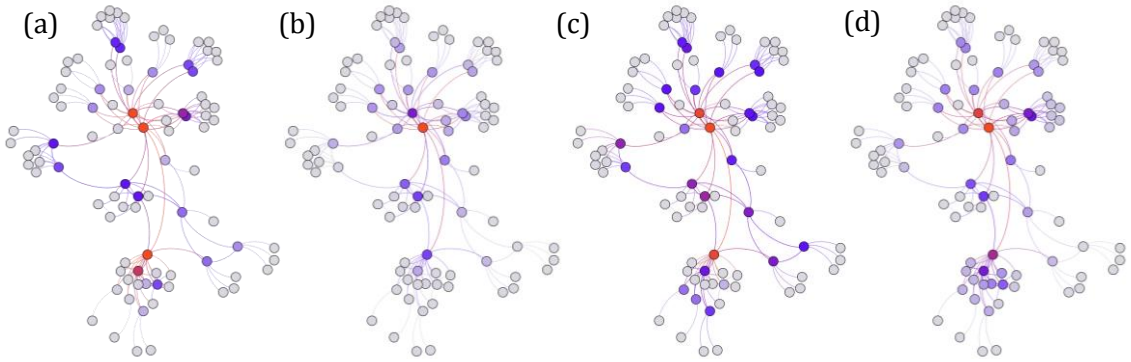

Figura 3: Centralidades calculadas na rede de exemplo: grau (a), proximidade (b), intermediação (c) e autovetor (d). Nós mais avermelhados são mais centrais.

Resta decidir um critério que meça a eficiência de um ataque. Nesse estudo, foi usado para esse fim o tamanho da maior componente conexa, que é um bom estimador do impacto. Descrevemos, a seguir, um algoritmo de simulação de ataque.

Algoritmo 2:

1) Calcular a centralidade dos nós da rede;
2) Remover o nó de maior valor de centralidade da rede no momento;
3) Obter o tamanho da maior componente conexa;
4) Repetir a partir do passo 2.

No Gephi há recursos para executar facilmente as operações dos passos 2 e 3. Na Figura 4 temos um gráfico com os resultados das 10 primeiras iterações dos ataques.

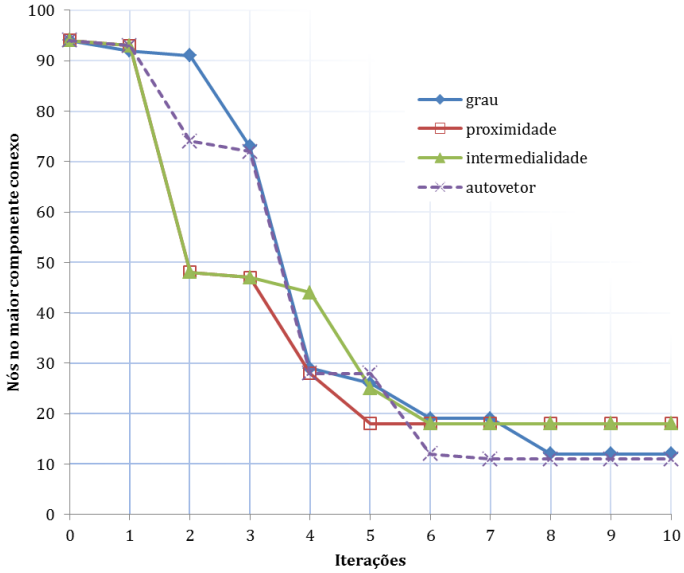

Figura 4: Tamanho da maior componente conexa em função das iterações do Algoritmo 2.

A partir do 6º ataque, a rede já está bem fragmentada em qualquer um dos casos. Isso evidencia, portanto, que esse algoritmo pode servir de guia para direcionar procedimentos de segurança, indicando onde deve ser reforçada a segurança dos nós da rede. Observamos que os ataques iniciais que têm mais impacto são aqueles em que são escolhidos os nós de maior proximidade e maior intermediação, sendo esse último menos custoso computacionalmente.

Conclusão

Dentro da área de Análise de Redes, o conceito de centralidade tem destaque. Fizemos um apanhado das principais medidas de centralidade conhecidas, partindo das três centralidades básicas (grau, proximidade e intermediação), e mostrando como as outras medidas podem surgir

a partir delas. Um exemplo de aplicação das medidas de centralidade é o algoritmo de Girván-Newman, que usa a centralidade de intermediação de link para obter a estrutura de comunidades de uma rede. No caso das redes de dados, mostramos que esse algoritmo pode ser útil para dividir uma rede de roteadores em grupos, de modo que cada grupo fique sob a responsabilidade de uma equipe.

Uma rede de dados está sujeita a eventos que incapacitam seus elementos. No presente trabalho, focamos na questão da identificação dos nós mais vulneráveis de uma rede, apresentando um algoritmo que faz uso das medidas de centralidade para simular um ataque. Emulando o ponto de vista do atacante, partirmos do pressuposto que os nós de maior centralidade são os alvos cuja incapacitação traria o maior impacto. Mostramos, também, que um bom critério para a avaliação do impacto de um ataque é o tamanho da maior componente conexa após efetivação deste ataque.

Verificamos que principalmente as medidas de centralidade de proximidade e intermediação são as que previram melhor a fragmentação da rede, sendo a de intermediação uma melhor escolha devido ao seu menor custo computacional de $O(n^2 \log n)$. Portanto, esse algoritmo pode servir de guia para direcionar procedimentos de segurança.

Referências

[1] Bonacich, "Factoring and weighting approaches to status scores and clique identification.," Journal of Mathematical Sociology, vol. 2, pp. 113–120, 1972.

[2] Borgatti e Everett, "A Graph-theoretic perspective on centrality," Social Networks, vol. 28, no. 4, pp. 466–484, Oct. 2006.

[3] Freeman, Borgatti, e White, "Centrality in valued graphs: A measure of betweenness based on network flow," Social Networks, vol. 13, no. 2, pp. 141–154, Jun. 1991.

[4] Freeman, "Centrality in Social Networks Conceptual Clarification," vol. 1, no. 1968, pp. 215–239, 1978.

[5] Ford e Fulkerson, Flows in networks. 1962.

[6] Girvan e Newman, "Community structure in social and biological networks.," Proceedings of the National Academy of Sciences of the United States of America, vol. 99, no. 12, pp. 7821–6, Jun. 2002.

[7] Katz, "A New Status Index Derived from Sociometric Index," Psychometrika, pp. 39–43, 1953.

[8] Latora e Marchiori, "Efficient Behavior of Small-World Networks," Physical Review Letters, vol. 87, no. 19, p. 198701, Oct. 2001.

[9] Latora e Marchiori, "A measure of centrality based on network efficiency," New Journal of Physics, vol. 9, no. 6, pp. 188–188, Jun. 2007.

[10] Newman, Networks: an introduction. 2009.

