Difusão Dirigida por Centralidade em Redes Complexas

Abraão Guimarães Flores[1]
Ana Paula Couto Silva[1], **Alex Borges Vieira**[1]

[1]Departamento de Ciência da Computação (DCC)
Instituto de Ciências Exatas (ICE) – Universidade Federal de Juiz de Fora (UFJF)
Rua José Loureço Kelmer, s/n, São Pedro – 36.036-900 – Juiz de Fora – MG – Brasil

abraaoguimaraes@ice.ufjf.br
{ana.silva, alex.borges}@ufjf.edu.br

Abstract. *Diffusion processes on complex networks can arise, for instance, on data search, data routing and information and/or disease spreading. Understanding how optimize the diffusion process is an important topic on the study of complex networks. In this letter, we shed some light on how centrality measures and nodes' dynamic can help on accelerating the network cover time.*

Resumo. *Processos de difusão em redes complexas podem aparecer, por exemplo, em pesquisa de dados, roteamento de dados e informações e/ou propagação de doenças. A compreensão de como otimizar o processo de difusão é um tema importante no estudo de Redes Complexas. Neste artigo são apresentados alguns indícios sobre como métricas de centralidade e a dinâmica dos nós podem ajudar na aceleração do tempo de cobertura da rede.*

1. Introdução

A presença do conceito de redes como conjunto de entidades que possuem ligações entre si está presente em diversas áreas da vida cotidiana. Entre os diversos exemplos destes tipos de redes, estão presentes as redes de comunicação, sociais, biológicas e tecnológicas.

Uma característica em comum entre estes sistemas é o dinamismo. A estrutura física muda com o tempo, entidades são adicionadas ou retiradas e as ligações entre estas surgem ou desaparecem ao longo do período de observação. Modelar e estudar tais sistemas com uma única fotografia que agrega uma grande quantidade de informação para um tempo longo de observação pode resultar em conclusões equivocadas sobre, por exemplo, o comportamento global e a interação entre as entidades que compoẽm estes sistemas. Como consequência, ignorar o dinamismo pode comprometer propostas mais eficazes de gerência, construção e caracterização das redes encontradas na vida real.

Poucos trabalhos na literatura consideram a dinâmica do sistema na caracterização do mesmo. Isso ocorre principalmente devido à alta complexidade introduzida tanto na modelagem quanto na análise das características comportamentais de maior relevância. O próprio conceito de dinâmica em um sistema real não é trivial, e proposições devem ser estabelecidas *a priori* da análise a ser realizada.

A difusão de informação entre os elementos da rede é um dos diversos problemas importantes que devem ser abordados no estudo de redes reais, quando se considera a dinâmica. Difundir informação em uma rede pode ter diversos significados dependendo do sistema real em estudo. Por exemplo, em redes par-a-par (P2P), difundir informação significa distribuir arquivos no caso de sistemas *file sharing*; ou vídeo no caso de alguns sistemas de tempo real. No caso de redes biológicas, pode-se considerar a difusão de informação como sendo o processo de disseminação de uma doença ou de uma vacina entre pessoas que estabelecem algum contato direto ou indireto.

Claramente, o processo de difusão da informação em uma rede qualquer, seja uma rede social ou biológica, está intimamente atrelado ao comportamento e qualidades dos integrantes das mesmas.

Diversas redes possuem um grau de dinamicidade elevado, e assim, suas características se modificam em um curto intervalo de tempo. Ao se analisar estas redes após um longo intervalo de observações, pode-se perder informações valiosas para uma análise detalhada do comportamento destes sistemas. Este fenômeno ocorre por conta de se agregar, em uma única fotografia, um conjunto grande de modificações estruturais. Tais modificações, em muitos casos, deveriam ser verificadas em intervalos de tempo menores.

Assim, ao se avaliar um determinado sistema real, incluindo o aspecto dinâmico do mesmo, devem-se considerar duas questões importantes. Primeiro, como definir a *dinâmica* propriamente dita. Em segundo, como capturar mudanças de suas características evitando agregar uma grande quantidade de informação.

2. Fundamentação Teórica

2.1. Modelagem Matemática

A notação e a definição matemática para modelar as redes dinâmicas utilizadas nesse trabalho são baseadas em [Basu et al. 2010].

Uma rede é representada matematicamente por um grafo. Seja $G(V, E)$ a representação desta rede, sendo V o conjunto de nós e E o conjunto de arestas. Seja o período total de observação do sistema real denotado por T_N, iniciado no tempo $T_1 = 0$. Sem perda de generalidade, considera-se que a visão agregada da rede, em outras palavras a modelagem do sistema considerando o tempo total de observação T_N, é feita através da construção do grafo estático G.

Seja $\mathcal{G}_t(\mathcal{V}_t, \mathcal{E}_t)$ a representação dinâmica do grafo agregado G, sendo $\mathcal{V}_t$ o conjunto de nós e $\mathcal{E}_t$ o conjunto de arestas. O grafo $\mathcal{G}_t$ e os conjuntos $\mathcal{V}$ e $\mathcal{E}$ estão indexados no tempo t, com $T_1 \leq t \leq T_N$. A análise do comportamento dinâmico da rede é feita através de uma sequência de grafos organizados em *snapshots* incrementais no tempo.

A duração de cada *snapshot* é definida por $\Delta = T_i - T_{i-1}$, com $0 \leq i \leq N$. Esta representação permite capturar a evolução da rede no espaço e no tempo. Ou seja, mudanças estruturais são capturadas a cada Δ unidades de tempo. Claramente, se $\Delta = T_N$, é representada uma única fotografia do sistema, desconsiderando toda a dinâmica, e assim, reduzindo a análise ao grafo estático G. Em contrapartida, a sequência de grafos $\mathcal{G}_t$ pode ser interpretada como um conjunto de fotografias do sistema, cada uma representando as mudanças estruturais ocorridas em um *snapshot* em particular.

A escolha do parâmetro Δ influencia na dinamicidade capturada na modelagem do sistema. Quanto menor o seu valor, maior é a aproximação de uma análise dinâmica formada por várias fotografias. Para obter uma análise com menor granularidade, Δ pode ser definido como o menor intervalo de acontecimento de um evento no sistema (e.g. adição/remoção de um nó ou aresta). No entanto, diminuir a granularidade de observação implica em aumentar a complexidade na amostragem dos dados e na caracterização do sistema analisado. De forma análoga, quanto maior o valor de Δ, maior é o nível de agregação de informação incorporada ao grafo e menor a percepção da dinâmica associada.

2.2. Métricas Topológicas

Métricas topológicas são definidas como medidas baseadas em atributos estruturais de um grafo. Estas métricas podem considerar cada participante (nó) em específico ou a visão global do grafo. Neste artigo são utilizadas as métricas relacionadas à caracterização dos nós, conhecidas como métricas de centralidade.

Sem perda de generalidade, será considerado o grafo estático G para formalização das métricas. Para o conjunto de grafos indexados em t, com $T_1 \leq t \leq T_N$, as métricas são definidas em cada *snapshot*. As principais métricas consideradas neste artigo são:

(1) Grau: O grau de um nó v, g_v, é definido como o total de arestas incidentes a este nó. NEste artigo foi definida a métrica centralidade do grau de um nó v, i.e. g_v como a fração entre o grau de um nó e maior valor possível de grau de um nó:

$$d_v = \frac{d(v)}{\max_{\forall v \in \mathcal{V}_t} d(v)}$$

(2) Betweenness: O *Betweenness* de um nó v é a fração dos caminhos mínimos, calculados usando *breadth-first search*, que ligam qualquer par de nós e que passam pelo nó v.

Em outras palavras, seja $\sigma_{u,j}$ o total de caminhos mínimos entre u e j, e $\sigma_{u,j}(v)$ o número total de caminhos que passam por v. A métrica de *Betweenness* é definida como:

$$Betweenness(v) = \sum_{\forall v \neq u \neq j \in V} \frac{\sigma_{u,j}(v)}{\sigma_{u,j}}.$$

(3) Closeness: A métrica de *closeness* de um nó v captura o quão perto este nó está de todos os nós que podem ser alcançados a partir deste na rede. Dado o tamanho de um caminho mínimo entre v e j, definido por $l(v, j)$, a métrica de *closeness* é dada por:

$$Closeness(v) = \sum_{\forall v \neq j, j \in V} l(v, j)^{-1}.$$

Para os algoritmos de difusão baseados em métricas de centralidade, são calculadas as métricas através da definição clássica das mesmas, apresentadas nos itens 1, 2 e 3. No entanto, diversos trabalhos na literatura buscam aproximações para os seus valores, visando a diminuição do custo computacional [Wehmuth and Ziviani 2011]. Como trabalhos futuros, para avaliação de redes com maior quantidade de nós, poderão ser aplicados os resultados aproximados propostos na literatura.

2.3. Modelos de Difusão de Informação

Nesta seção são descritos os principais modelos básicos de difusão de informação, sendo que estes podem ser aplicados tanto no caso onde o sistema é modelado com múltiplas fotografias, ou seja, com o conjunto de grafos $\mathcal{G}_t$, $T_1 \leq t \leq T_N$, quanto no caso onde o sistema é modelado através da visão agregada representada pelo grafo G. Será apresentado o modelo de difusão *Epidêmico* e, baseando-se em [Lovasz 1993], o modelo de difusão *Random Walk* (RW).

2.3.1. *Epidêmico*

A denotação *Epidêmico*, adotada nesse artigo, refere-se à implementação onde todos os vizinhos habilitados recebem a informação a cada iteração. É fácil notar que este modelo dará o menor tempo de difusão partindo de uma origem, tendo em vista que todos os vizinhos dos nós que possuem a informação, receberão o dado qualquer que seja o *snapshot* atual. No entanto, o custo computacional deste método, baseado no número total de mensagens trocadas entre os nós, é elevado. Como consequência, a implementação deste modelo em sistemas reais é complexa e pode-se tornar inviável.

2.3.2. *Random Walk*

Conforme descrito em [Lovasz 1993], o funcionamento do modelo de difusão *Random Walk* (RW) é simples: dado um grafo qualquer e um nó aleatório v como ponto de partida, um vizinho u de v é escolhido aleatoriamente e a informação é repassada a este. De forma análoga, um vizinho i escolhido aleatoriamente entre os vizinhos de u recebe a informação na próxima iteração do algoritmo. A sequência aleatória dos nós

selecionados neste caminho é definida como um passeio aleatório no grafo, ou seja, um *Random Walk*.

Dada a dinâmica do modelo RW, espera-se que o tempo de difusão da informação seja maior que nos modelos epidêmicos. Adicionalmente, este algoritmo não considera características do sistema real que podem acelerar a difusão da informação entre os nós do grafo. No caso de sistemas onde o tempo de entrega da informação é crucial, torna-se importante identificar nós que potencialmente possam acelerar o processo de difusão.

Nos resultados apresentados na Seção 5, são consideradas duas implementações diferentes do modelo RW: a tradicional, como descrita anteriormente, onde a cada instante de tempo somente um vizinho é escolhido para receber a informação que será denotada somente por *RW*.

3. Modelos de Difusão utilizando Métricas Topológicas

Com o objetivo de acelerar o processo de difusão de informação em redes que representam sistemas reais considera-se o conhecimento de características das entidades que formam o sistema real sendo analisado para decidir qual será a próxima entidade a receber a informação a ser difundida no sistema.

Os modelos de difusão, denominados *Betweenness Walk* e *Closeness Walk* se assemelham ao modelo RW. Seja o grafo G ou $\mathcal{G}_t$ que modela o sistema real. A cada escolha do próximo nó a receber a informação, seleciona-se o nó com o maior valor da métrica de *Betweenness* (ou *Closeness*) entre todos os possíveis.

Para o cálculo da métrica de *Betweenness* (ou *Closeness*), supõe-se que o sistema possua uma entidade com visão global da topologia do grafo que modela o mesmo. Obviamente, para grafos com centenas de milhares de nós, o cálculo desta métrica pode ser custoso. Como trabalho futuro, pretende-se verificar a possibilidade do cálculo distribuído desta métrica, bem como a aplicação de outras métricas topológicas.

4. Descrição dos *Logs*

As experiências foram conduzidas em diferentes tipos de redes sintéticas e reais, com dinâmicas distintas. A descrição de conjuntos de dados é mostrada na sequência.

4.1. Redes Sintéticas

Foram criadas redes com os modelos Erdös-Rényi (ER), Watts-Strogatz (também conhecida como modelo *Small-World* (SW)) e Barabási-Albert (BA) usando a ferramenta *NetworkX* [1]. Todas as redes sintéticas possuem $n = 1000$ nós.

Foram considerados 100 *snapshots* de uma unidade de tempo, i.e, $T_1 = 1$ e $T_N = 100$, criando sucessivos grafos independentes. A tabela tab:parameters1 mostra os parâmetros do modelo.

Para o modelo ER, o parâmetro p é a probabilidade de existir uma aresta entre um par qualquer de nós. Para o modelo SW, p é a probabilidade de rearranjo nas arestas. No modelo BA, um novo nó é adicionado com m arestas.

[1] http://networkx.lanl.gov/

Parameters	Erdös-Rényi	Small-World	Barabási-Albert
p	0.01	0.01	-
m	-	-	10

Tabela 1. Parâmetros dos Grafos Sintéticos

4.2. Rede Real

SopCast [Sopcast 2012] é uma das aplicações mais populares para difusão de vídeo em tempo real baseada em redes P2P. Clientes SopCast que utilizam a aplicação estão conectados a um canal em particular, pertencendo a rede sobreposta na qual o vídeo está sendo transmitido.

O *log* modelado neste artigo representa uma coleta de $1h$ realizada no dia 14 de outubro de 2011. Para a construção do grafo $\mathcal{G}_t$, considera-se o valor de $\Delta = 1s$, capturando uma quantidade significativa de troca de informação entre vizinhos. São considerados 3.601 *snapshots* de $1s$, com o total de 334 nós no grafo que modela o sistema. Uma aresta é estabelecida entre dois nós (clientes) da rede sobreposta, se existe, pelo menos, uma troca de pacote maior que 200 *bytes*, considerado pacote de vídeo (descartando pacotes de controle) [Tang et al. 2009].

5. Resultados

O principal foco dos resultados apresentados neste artigo é avaliar o impacto da dinâmica inerente aos sistemas descritos na Seção 4, bem como o desempenho dos diferentes modelos de difusão apresentados nas Seções 2.3 e 3.

Seja o total de nós alcançados a partir do nó j denotado por TN_j e $|\mathcal{V}_{T_i}|$ o total de nós presentes no grafo no instante de tempo T_i. Em todos os resultados relativos aos modelos de difusão de informação considera-se a medida de interesse *percentual de nós alcançados em* T_i (π_{T_i}), definida por:

$$\pi_{T_i} = \frac{\sum_{\forall j \in \mathcal{V}_{T_i}} TN_j}{|\mathcal{V}_{T_i}| * (|\mathcal{V}_{T_i}| - 1)}. \quad (1)$$

Esta métrica calcula a média dos nós alcançados a cada *snapshot*, considerando cada um dos nós em $\mathcal{V}_{T_i}$ como nó inicial de difusão.

5.1. Resultados Numéricos

Os primeiros resultados apresentados são calculados na rede do *SopCast*. A figura 1 mostra uma melhoria no tempo de difusão ao ser consideradas as métricas topológicas de centralidade.

Serão mostrados agora os resultados calculados nas redes sintéticas geradas pelos modelos citados anteriormente.

Como esperado, para o modelo ER (Fig. 2(a)) a utilização de métricas não aceleram o processo de difusão. Isto porque o modelo se baseia na homogeneidade do grau dos nós.

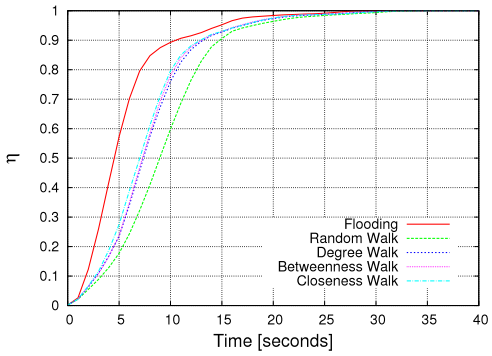
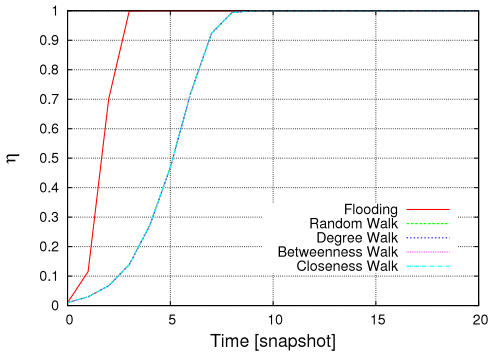


Figura 1. Impacto da dinâmica do sistema no processo de difusão.

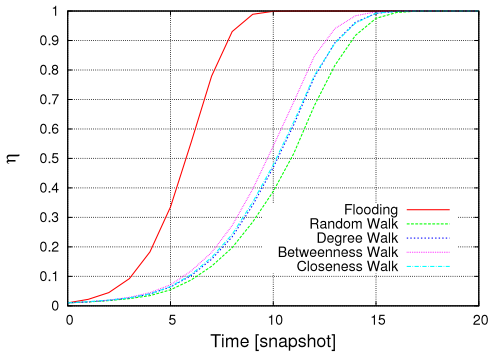
Para o modelo BA, é possível ver na figura. 2(b), a distribuição nó grau segue uma distribuição de lei de potência. Portanto é mais provável, mesmo ao acaso, e escolha de nós com poucos vizinhos, correspondendo aos nós com menores valores das métricas de centralidade. Consequentemente, o tempo de difusão do *random walk* é maior, mesmo perdendo para os algoritmos baseados nas métricas de centralidade.

Para o modelo de SW, os algoritmos de difusão se comportam de maneira diferente. O processo de difusão é mais sensível às métricas de centralidade. Como é possível ver na figura. 2(c), as métricas de centralidade apresentam o mesmo padrão de difusão.



(a) Modelo Erdös-Rényi

(b) Modelo Barabási-Albert



(c) Modelo Small-World

Figura 2. Difusão nas redes sintéticas

6. Trabalhos Futuros

6.1. Gerador Markoviano de Grafos

Em sistemas reais, ao observar *snapshots* em instantes subsequentes, é possível notar certa dependência do instante atual com os instantes anteriores. Essa dependência pode ser maior ou menor, dependendo do sistema observado, mas ela sempre está presente. Uma forma de simular este comportamento é utilizar modelos que geram o instante atual a partir de características do sistema encontradas no instante anterior.

A partir de um grafo inicial, é possível definir grafos subsequentes seguindo o modelo *Markovian Edge* presente na literatura em [Coolen et al. 2009]. O grafo atual é gerado a partir do modelo markoviano de definição de arestas, onde a existência de cada aresta é definida pela Cadeia de Markov, tendo como base as arestas do grafo anterior.

A Cadeia possui dois parâmetros, p e q, que são as probabilidades de mudança de estado. Desta forma, a probabilidade de permanencia no estado será dada por $1 - p$ (ou $1 - q$).Se a aresta existe no grafo anterior, ela não existirá no atual com a probabilidade p. Da mesma forma, se a aresta não existe no grafo anterior, ela existirá no grafo atual com probabilidade q.

Com isto, é possível simular o comportamento de sistemas reais, mantendo uma dependência do instante atual com os anteriores.

6.2. Previsão

Buscando acelerar o tempo de difusão de conteúdo em redes dinâmicas, está sendo calculada a previsão do próximo instante de tempo. Se for possível definir, com certa probabilidade, qual a formação topológica da rede no próximo instante de tempo, será possível utilizar essa informação para que se alcance mais rapidamente o destino da informação.

A previsão está sendo feita baseando-se no histórico da rede. A medida com que os instantes de tempo vão passando, a probabilidade de existência de cada aresta vai sendo calculada levando em conta uma maior quantidade de informação.

A probabilidade de cada uma das arestas do próximo instante é baseada no modelo *Weighted Moving Average*, apresentado em [Roberts 1959]. Esta probabilidade é dada pela equação (3), que leva em conta a probabilidade calculada no instante anterior e uma média ponderada do histórico da rede, realizada da forma a seguir.

- Seja a_{ij} a aresta $i \rightarrow j$.
- Seja $E_{t,a_{ij}}$ a função a seguir:

$$E_{t,a_{ij}} = \begin{cases} 1, & a_{ij} \in \mathcal{E}_t \\ 0, & c.c \end{cases} \tag{2}$$

- Seja $\varphi_{n_{ij}}$ a "probabilidade de existência" da aresta a_{ij} no instante n, dada por:

$$\varphi_{n_{ij}} = \begin{cases} \varphi_{(n-1)_{ij}} * f + \frac{\sum_{\tau=1}^{n} E_{\tau,a_{ij}} * \tau}{\sum_{\tau=1}^{n} \tau} * (1-f), & n > 1 \\ \mathcal{D}_{ij} * f + E_{\tau,a_{ij}} * (1-f), & c.c \end{cases} \tag{3}$$

onde:

- f: fator de "importância" do histórico;
- $\mathcal{D}_{ij}$: distribuição inicial de probabilidade da aresta $i \rightarrow j$

É possível perceber que este cálculo leva em consideração todo o histórico da rede. Isto pode ser um problema ao ser aplicado em sistemas reais, pois a manutenção de todo o histórico em *cache* implicaria na necessidade de se ter um *buffer* infinito para isso. Diante desta limitação, é razoável considerar um *buffer* de tamanho finito b, utilizado para armazenar o histórico. Assim, é possível realizar uma alteração no cálculo do φ, considerando apenas um número finito b de *snapshots*.

$$\varphi_{n_{ij}} = \begin{cases} \varphi_{(n-1)_{ij}} * f + \frac{\sum_{\tau=n-b}^{n} E_{\tau,a_{ij}} * \tau}{\sum_{\tau=n-b}^{n} \tau} * (1-f), & n > 1 \\ \mathcal{D}_{ij} * f + E_{\tau,a_{ij}} * (1-f), & c.c \end{cases} \tag{4}$$

Onde b é o tamanho do *buffer* de histórico da rede.

Dessa forma, tem-se a distribuição da probabilidade de existência de cada uma das arestas do grafo. Os valores destas probabilidades são comparadas com um "limiar" e todas arestas cuja probabilidade alcance ou ultrapasse este limite são consideradas existentes no grafo previsto. Analogamente, todas arestas que não alcançam este limiar são consideradas inexistentes na previsão.

Com este grafo criado através da previsão das arestas, é feito o cálculo dos valores das métricas. Todos os nós que possuem a informação no instante avaliado encaminhará a mensagem para seu vizinho que possui o maior valor da métrica calculada no grafo previsto. Com isto, espera-se uma aceleração no tempo de difusão, pois existe a possibilidade do grafo previsto ter uma correlação com o estado da rede no próximo instante de tempo.

Referências

Basu, P., Bar-Noy, A., Ramanathan, R., and Johnson, M. P. (2010). Modeling and Analysis of Time-Varying Graphs. Published on arXiv.org;cs;arXiv:1012.0260.

Coolen, A., Martino, A., and Annibale, A. (2009). Constrained markovian dynamics of random graphs. *Journal of Statistical Physics*, 136:1035–1067.

Lovasz, L. (1993). Combinatorics, Paul Erdos is Eighty. *Bolyai Society Mathematical Studies*, 2:1–46.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250.

Sopcast (2012). http://www.sopcast.org/.

Tang, S., Lu, Y., Hernández, J. M., Kuipers, F., and Mieghem, P. (2009). Topology dynamics in a p2ptv network. In *Proceedings of the 8th International IFIP-TC 6 Networking Conference*, NETWORKING '09, pages 326–337, Berlin, Heidelberg. Springer-Verlag.

Wehmuth, K. and Ziviani, A. (2011). Um Novo Algoritmo Distribuido para Avaliação e Localização de Centralidade de Rede. In *Proceedings of X Workshop em Desempenho de Sistemas Computacionais e de Comunicação (WPerformance)*.

Submissão dos Artigos

Os orientadores Ana Paula Couto da Silva e Alex Borges Vieira estão cientes e de acordo com a submissão dos artigos dos seguintes alunos:

1) Bianca Portes;
2) Thiago Guarniere;
3) Rodrigo Duarte;
4) Abraão Guimarães;
5) Rafael Barra;
6) Thiago Boubee;
7) Francisco Henrique.

Atenciosamente,

Ana Paula Couto da Silva

Alex Borges Vieira