Modelagem e Caracterização de um Processo de Amostragem de Vértices em Redes*

Vicente M. Pinheiro[1], **Daniel R. Figueiredo**[1], **Antonio A. de A. Rocha**[2]

[1] COPPE/Programa de Engenharia de Sistemas e Computação
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, RJ – Brasil

[2]Instituto de Computação
Universidade Federal Fluminense (UFF)
Niterói, RJ – Brasil

{vicente,daniel}@land.ufrj.br, arocha@ic.uff.br

Abstract. *The increased interest in studying how "things" connect has being leveraged by the growing abundance of a huge amount of data concerning many different networks. In this context, an important aspect is the collection of such data, because in most cases information on network vertices and edges is not publicly available in a centralized or organized repository (eg., Web, P2P, Facebook). Thus, it is necessary to discover these networks through a sampling process in which the process itself fundamentally influences what is discovered about the network. In this work, we study the process of sampling vertices that reveals local information around randomly chosen vertices. Particularly, we develop analytical models to determine the number of vertices and edges discovered by the sampling process according to the number of samples and other network characteristics (eg., average degree). The evaluation of the proposed models against results obtained through simulations for different network models indicates the conditions under which our model is accurate.*

Resumo. *A explosão pelo interesse em estudar como as "coisas" se conectam vem sendo alavancada pela crescente abundância de enormes massas de dados sobre as mais diferentes redes. Neste contexto, um aspecto importante diz respeito à coleta desses dados, pois na maioria dos casos informações sobre vértices e arestas das redes não estão disponíveis publicamente de forma centralizada ou organizada (ex. Web, rede P2P, Facebook). Desta forma, é necessário descobrir estas redes através de algum processo de amostragem, que fundamentalmente irá influenciar o que será descoberto. Neste trabalho estudamos o processo de amostragem de vértices que revela informação local a vértices escolhidos aleatoriamente. Em particular, desenvolvemos modelos analíticos para calcular o número de vértices e arestas descobertos pelo processo em função do número de amostras e de outras características da rede (ex. grau médio). A avaliação dos modelos propostos com resultados obtidos através de simulações em diferentes modelos de rede confirmam nosso modelo exato e mostram quando nosso modelo aproximado é preciso.*

*Esse projeto de pesquisa foi parcialmente financiada pela CAPES, CNPq e FAPERJ.

1. Introdução

A explosão durante a última década pelo interesse em estudar como as "coisas" se conectam e entender as implicações desta conectividade em diversas áreas do conhecimento deu origem a área multidisciplinar conhecida por *Network Science* [Barabási 2009]. Estudos sobre as mais diversas redes vem revelando características estruturais fundamentais e contribuindo na compreensão de fenômenos que operam sobre estas redes.

Uma das principais razões para o grande avanço nesta área é a crescente disponibilidade de grande massa de dados sobre as mais diversas redes, que permitem realizar estudos empíricos e validar modelos matemáticos. Entretanto, um importante aspecto é a obtenção destes dados, pois muitas vezes não estão disponíveis publicamente de forma centralizada ou organizada e precisam ser coletados. Por exemplo, a rede social online do Facebook (usuários e amizades) e a rede do BitTorrent durante um swarm (peers e conexões TCP). Em ambos os casos, a rede precisa ser descoberta através de algum *processo de amostragem* que irá revelar seus vértices e arestas. Um problema fundamental passa a ser a influência do processo de amostragem no que será descoberto, pois em muitos casos é proibitivo coletar a rede inteira. Diversos trabalhos recentes na literatura vem abordando estas questões ao considerar processos de amostragem baseados em passeios aleatórios (*random walks*), busca em largura (BFS), e amostragem de arestas (*edge sampling*) [Ribeiro and Towsley 2010, Kurant et al. 2011a, Pedarsani et al. 2008, Kurant et al. 2011b].

Neste trabalho iremos estudar o processo de amostragem de vértices onde a cada passo um vértice da rede escolhido de forma aleatória é revelado ao observador juntamente com outras informações da rede local ao vértice (ex. seus vizinhos). Iremos considera dois casos: *(i)* vértice escolhido revela sua identidade e todos seus vizinhos; *(ii)* vértice escolhido releva sua identidade, todos seus vizinhos e todos os vizinhos dos vizinhos (detalhes na seção 2).

Estamos interessados em caracterizar como este processo descobre os vértices e arestas de uma rede desconhecida em função do número de amostras. Neste sentido, desenvolvemos modelos analíticos (exatos e aproximados) que caracterizam o valor esperado do número de vértices e arestas descobertos em função do número de amostras para os dois tipos de amostragem. Fazemos ainda uma avaliação numérica utilizando três modelos clássicos de redes, comparando os resultados previstos pelos modelos com resultados obtidos através da simulação detalhada do processo de amostragem. Nossos resultados validam o modelo exato e mostram que o modelo aproximado possui bom desempenho em alguns casos. Além disso, discutimos a influência das diferentes redes no desempenho do processo de amostragem.

Por fim, apesar de considerarmos e avaliarmos o processo de amostragem de vértices de forma abstrata, este processo é uma boa abstração para redes que permitem que um vértice seja inspecionado e que informações locais sejam obtidas. Por exemplo, em redes P2P um par (vértice) controlado pelo observador pode obter informações locais; em redes sociais online podemos amostrar vértices usando identificadores aleatórios e descobrir informações locais. Nosso objetivo neste artigo não é aplicar o método de amostragem de vértices e sim caracterizar como o mesmo descobre uma rede.

O restante deste artigo está organizado da seguinte maneira. Na seção 2 descreve-

mos o processo de amostragem de vértices e as duas variações consideradas. Nas seções 3 e 4 apresentamos os modelos para calcular o número esperado de vértices e arestas descobertos pelo modelo, respectivamente. Na seção 5 apresentamos a avaliação numérica comparando resultados analíticos com simulações. Nas seções 6 e 7 apresentamos os trabalhos relacionados e nossas conclusões, respectivamente.

2. Processo de Amostragem de Vértices

Nosso estudo sobre o processo de amostragem assume que temos uma rede definida por um grafo não-direcionado $G = (V, E)$ onde V é o conjunto de vértices rotulados unicamente de tamanho $|V| = n$ e E é o conjunto de arestas de tamanho $|E| = m$. Podemos interpretar G como sendo uma rede gerada por algum modelo de grafos aleatórios ou uma rede real obtida empiricamente. Esta abstração permite representar qualquer tipo de rede, por exemplo, uma rede P2P em que a identidade dos vértices são seus endereços IPs e as arestas indicam a presença de conexão TCP entre cada par de vértices. Apesar da rede existir e ser estática, iremos assumir que não temos conhecimento da identidade de seus vértices e relacionamentos. O processo de amostragem será responsável por revelar estas informações, conforme descrito abaixo. Por fim, neste trabalho iremos assumir que a rede é estática durante o processo de amostragem, não sofrendo qualquer alteração em sua estrutura (tanto nos vértices quanto em suas arestas).

O processo de amostragem é uma abstração de um processo real de descobrimento dos vértices da rede e consiste em revelar um ou mais vértices que chamaremos de monitores. O vértice monitor será escolhido aleatoriamente entre todos os vértices da rede. Cada monitor irá obter informações locais sobre a rede, tais como a identidade de vértices e as arestas ao seu redor. Este processo poderá se repetir, com monitores sendo amostrados aleatoriamente, de forma iterativa, onde cada novo monitor potencialmente revelará novas informações locais sobre a rede.

Particularmente, iremos considerar dois tipos de monitores, que diferem com relação ao que observam sobre sua localidade, e também podem representar diferentes processos reais de amostragem. São eles:

- Monitor revela sua identidade e a identidade de todos seus vizinhos. Este caso também revela todas as arestas que incidem sobre o monitor. Chamaremos esta caso de *monitor revela vértices até distância 1*.
- Monitor revela sua identidade, a identidade de todos seus vizinhos, e a identidade de todos os vizinhos dos vizinhos. Este caso também revela além das arestas entre o monitor e seus vizinhos, todas as arestas entre os vizinhos do monitor e cada um dos seus respectivos vizinhos. Chamaremos este caso de *monitor revela vértices até distância 2*

Considere a rede ilustrada pela figura 1(a). As figuras 1(b) e (c) ilustram essa rede e a informação revelada pelos dois tipos de monitores, para o caso em que o vértice 9 é escolhido como monitor. Note que as arestas entre os vizinhos não é revelada no primeiro caso (ex., aresta entre vértices 6 e 7 na figura (b)), assim como as arestas entre os vizinhos dos vizinhos também não são reveladas (ex., aresta entre vértices 3 e 4 na figura (c)).

O processo de amostragem considerado permite amostrar mais de um vértice da rede para assumir o papel de monitor e revelar informação local ao mesmo. Desta forma, um parâmetro fundamental do processo é o número de amostras de monitores, denotado

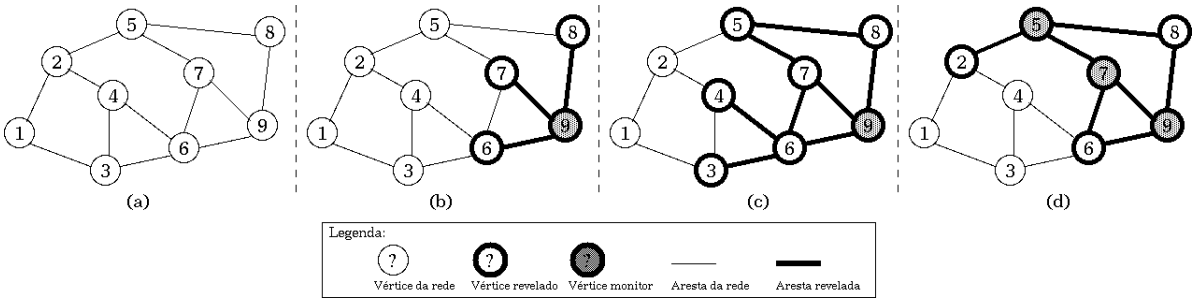


Figura 1. Exemplo de rede e informação revelada nos processos de amostragem.

por k. Note que k não é o número de monitores distintos, mas sim o número de amostras. Isso porque, assumimos que não controlamos o processo de amostragem (que é aleatório) de forma que um mesmo vértice possa ser amostrado mais de uma vez para monitor. Desta forma, k é o número de amostras de monitores.

As informações sobre vértices e arestas, coletadas pelos monitores, serão agregada para que possamos estimar características da rede. Entretanto, é importante destacar que nem toda nova amostra de monitor resultará necessariamente em mais informações. Duas amostras de monitores podem ter vizinhos em comum ou ainda serem o mesmo vértice. Considere, por exemplo, a figura 1(d) que ilustra um exemplo com $k = 3$ amostras de monitores (vértices 5, 7 e 9) para o caso do monitor revelar vértices até distância 1. Note que, se o monitor 7 for amostrado após os monitores 5 e 9, praticamente nenhuma nova informação será obtida, a não ser pela existência da aresta (6,7).

2.1. Medidas de interesse

Estamos interessados em caracterizar como o processo de amostragem de vértices (monitores) revela informação sobre a rede como um todo. Em particular, no número de vértices e arestas que o processo revela em função do número de amostras de monitores, k. Sejam Y_k e W_k, respectivamente, o número de vértices e arestas reveladas pelo processo de amostragem após k amostras de monitor. Como o processo de escolha de monitores é aleatório, assim como a rede também pode ser resultado de um modelo aleatório de grafo, Y_k e W_k são variáveis aleatórias. Desta forma, estamos interessados em caracterizar seus respectivos valores esperados, ou seja, $E[Y_k]$ e $E[W_k]$, o numero médio de vértices e arestas revelados. Outra medidas como clusterização e centralidade de vértices podem ser obtidas a partir dos vértices e arestas descobertos, entretanto neste trabalho não iremos investigar essas medidas.

É também interesse nosso compreender como este processo de amostragem depende da estrutura da rede. Ou seja, obter respostas para questões como: (i) quais são as características estruturais da rede que mais influenciam este processo de descoberta de informação? (ii) qual é a diferença entre o primeiro e segundo tipo de monitor quando aplicadas em diferentes estruturas de rede? Como veremos, o processo de descoberta de vértices e arestas são bem distintos, assim como os tipos de monitores (revelando distância 1 e distância 2). Além disso, o grau médio da rede possui um papel fundamental, enquanto a distribuição de grau pode possuir um papel secundário, mas importante em alguns casos. O cálculo destas medidas de interesse será apresentado nas próximas seções, assim como a avaliação numérica e discussão dos resultados.

3. Análise de descobrimento de vértices

Nesta seção iremos obter analiticamente o valor esperado do número de vértices descobertos na rede seguindo os processos de amostragem definido na Seção 2. A ideia da análise é calcular a probabilidade de um vértice da rede ser descoberto e utilizar esta probabilidade para calcular o número esperado de vértices descobertos.

Considere um vértice u da rede. Estamos interessados em calcular a probabilidade de u ser descoberto após uma amostra de monitor. O vértice u é revelado quando o mesmo é escolhido como monitor ou um de seus vizinhos é escolhido como monitor. E caso estejamos considerando monitores que revelam vértices à distância 2, o vértice u pode ser revelado se um vizinho de seus vizinhos (que não é vizinho direto de u) for escolhido como monitor. Estes três casos estão ilustrados na figura 2. Repare que estes três eventos são mutuamente exclusivos, pois uma amostra de monitor assume a identidade de exatamente um vértice da rede. Para facilitar a exposição, definiremos os seguintes eventos:

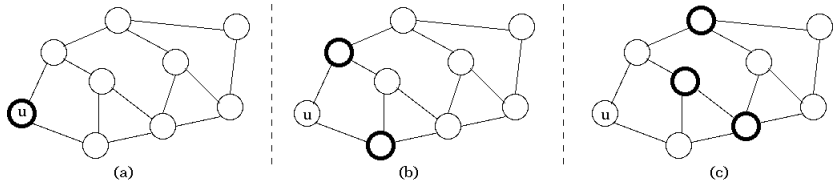


Figura 2. Exemplo das três formas que um vértice u da rede pode ser descoberto por uma amostra de monitor: (a) o próprio vértice u escolhido; (b) um vizinho de u escolhido; (c) um vizinho do vizinho de u (que não é vizinho de u) escolhido.

- $N_u^0 =$ vértice a distância 0 de u (próprio u).
- $N_u^1 =$ vértices a distância 1 de u (vizinhos de u).
- $N_u^2 =$ vértices a distância 2 de u (vizinhos dos vizinhos de u).
- $D_u^k =$ vértice u foi descoberto após k amostras de monitores.

Vamos assumir que o processo de escolha de monitores na rede é uniforme. Ou seja, todos os vértices da rede tem igual probabilidade de ser escolhido como monitor. Desta forma, a probabilidade do vértice u ser escolhido é $|N_u^0| = 1$ em n, pois temos n vértices na rede. Ou seja, $P[N_u^0] = 1/n$.

A probabilidade de um dos vizinhos de u ser escolhido como monitor pode ser calculada condicionando no grau de u. Seja d o grau do vértice u. Desta forma, como qualquer outro vértice da rede, cada vizinho de u pode ser escolhido com probabilidade $1/n$. Como os eventos de escolha dos vizinhos para ser monitor são eventos mutualmente exclusivos, temos $P[N_u^1|Z = d] = \sum_{i=1}^{d} 1/n = d/n$

Desta forma, a probabilidade do vértice u ser descoberto dado que o mesmo possui grau d é dado pela soma das duas probabilidades, de ele ser escolhido ou de um de seus vizinhos ser escolhido. Repare que a probabilidade de u ser escolhido como monitor independe do seu grau. Desta forma, temos:

$$P[D_u^1|Z = d] = 1/n + d/n = (1 + d)/n \tag{1}$$

Vamos considerar que um total de k amostras de monitores serão realizadas na rede com reposição. Vamos assumir que o processo de escolha de amostra é independente

e identicamente distribuído (iid). Desta forma, todo o vértice tem igual probabilidade de ser escolhido como monitor a cada amostra de monitores. Novamente, estamos interessados em calcular a probabilidade do vértice u ser descoberto. Repare que o vértice u pode ser descoberto por qualquer uma das k amostras, o que dificulta o cálculo de maneira direta. Então podemos considerar seu complemento, ou seja a probabilidade do vértice u não ser descoberto por nenhuma das k amostras. Como o processo de escolha de monitores é iid, a esta probabilidade será dada por $P[\overline{D_u^k}|Z = d] = (1 - (1 + d)/n)^k$. Repare que $1 - (1 + d)/n$ é a probabilidade de u não ser descoberto por uma das amostras dos monitores. Por fim, a probabilidade de u ser descoberto é apenas o complemento dele não ser descoberto, ou seja $P[D_u^k|Z = d] = 1 - (1 - (1 + d)/n)^k$.

Podemos agora descondicionar para obter a probabilidade de u ser descoberto, independente de seu grau. Ou seja,

$$P[D_u^k] = \sum_{d=0}^{n-1}(1 - (1 - (1 + d)/n)^k)P[Z = d] \tag{2}$$

Onde $P[Z = d]$ é a probabilidade do vértice u ter grau d.

Vamos definir $X_{u,k}$ como sendo uma variável aleatória indicadora que retorna 1 quando o vértice u é descoberto depois de k amostras de monitores na rede e 0 caso contrário. Repare que $P[X_{u,k} = 1] = P[D_u^k]$. Lembrando que Y_k representa o número de vértices descobertos depois de k amostras de monitores, que pode ser definida como $Y_k = \sum_{\forall u \in V} X_{u,k}$

Repare que cada vértice contribui com 1 (caso foi descoberto) ou 0 (caso não foi descoberto) para a soma que define Y_k. Por fim, estamos interessados no valor esperado de Y_k. E pela linearidade da esperança temos:

$$E[Y_k] = E[\sum_{\forall u \in V} X_{u,k}] = \sum_{\forall u \in V} E[X_{u,k}] = \sum_{\forall u \in V} P[D_u^k] = nP[D_u^k] \tag{3}$$

O penúltimo passo é válido pois o valor esperado de uma variável aleatória indicadora é simplesmente sua probabilidade de assumir valor 1, e o último é válido pois estamos assumindo que todos os vértices da rede são estatisticamente equivalentes e não dependem do seu identificador (rótulo).

É importante notar que a equação 3 depende da distribuição de grau da rede para ser calculada por causa do termo $P[D_u^k]$. Isto indica que a forma como o processo de amostragem revela vértice irá depender da distribuição de grau da rede.

3.1. Monitor revela vértices a distância 2

Estamos agora interessados na probabilidade do vértice u ser descoberto quando uma amostra de monitor revela não somente a identidade do vértice escolhido e de seus vizinhos, mas também dos vizinhos dos vizinhos, ou seja, todos os vértices até distância 2 da amostra escolhida. Desta forma, um vértice u da rede tem mais chance de ser descoberto, pois basta estar a distância 2 ou menor do monitor escolhido.

Lembrando que N_u^2 é o conjunto de vértices a distância 2 de u, temos que a probabilidade de um deles ser escolhido como monitor é dada por:

$$P[N_u^2] = \frac{|N_u^2|}{n} \tag{4}$$

Infelizmente, o valor $|N_u^2|$ (na verdade, sua distribuição) não é trivial e pode depender da estrutura da rede. Entretanto, podemos condicionar no grau do vértice u e utilizar o número de vértices que são incidentes a cada vizinho v de u, medida conhecida como restante de grau de v (já que uma aresta é incidente a u). Seja R a variável aleatória que representa o restante de grau de um vértice v incidente a u. O valor esperado de restante de grau de um vértice qualquer da rede foi obtido por Newman em [Newman 2010], e é dado por:

$$E[R] = \frac{E[Z^2] - E[Z]}{E[Z]} \tag{5}$$

Onde Z é a variável aleatória que representa o grau de um vértice qualquer da rede. Utilizando este resultado, podemos obter uma aproximação para o número esperado de vértices a distância dois, assumindo que cada vizinho v de u terá este número médio de vizinhos. Desta forma, temos:

$$|N_u^2| \approx dE[R] \tag{6}$$

onde d é o grau condicionado do vértice u. Entretanto, dois vizinhos v e w de u também podem ser vizinhos e seus restantes de grau estariam sendo contado erradamente. Este efeito pode ser medido pelo coeficiente de clusterização da rede, que caracteriza a fração de arestas entre os vizinhos de um vértice qualquer. Seja $\overline{c}$ o coeficiente de clusterização médio da rede. Seja A_d o número de arestas entre os vizinhos do vértice u dado que seu grau é igual a d. Podemos estimar o número médio de arestas entre os vizinhos do vértice u da seguinte forma:

$$E[A_d] \approx \overline{c} \binom{d}{2} = \frac{\overline{c}d(d-1)}{2} \tag{7}$$

Utilizando este resultado, podemos melhorar a aproximação para o número de vizinhos a distância 2 de u removendo os vértices que serão contados duas vezes na aproximação dada pela equação 6. Assim temos:

$$|N_u^2| \approx dE[R] - 2E[A_d] \approx \frac{d(E[Z^2] - E[Z])}{E[Z]} - \overline{c}d(d-1) \tag{8}$$

Com isso, podemos aproximar a probabilidade de um vértice a distância 2 de u ser escolhido como monitor, usando a aproximação acima na equação 4.

Por fim, a probabilidade do vértice u ser descoberto é dada pela soma das três possibilidades para a escolha do monitor (distâncias 0, 1 e 2 de u). Ou seja:

$$P[D_u^k | Z = d] = \frac{1+d}{n} + \frac{d(E[Z^2] - E[Z])}{nE[Z]} - \frac{\overline{c}d(d-1)}{n} \tag{9}$$

Utilizando a mesma abordagem para o caso da distância 1, podemos calcular a probabilidade do vértice u ser descoberto depois de k amostras de monitores da rede, de acordo com a equação 2. Em seguida, podemos definir as variáveis aleatórias indicadoras

$X_{u,k}$ para cada vértice u e calcular o valor esperado do número de vértices descobertos, $E[Y_k]$ de acordo com a equação 3. Desta forma, temos:

$$E[Y_k] = \sum_{d=0}^{n-1}(1 + d + \frac{d(E[Z^2] - E[Z])}{E[Z]} - \overline{c}d(d-1))P[Z = d] \tag{10}$$

4. Análise de descobrimento de arestas

O objetivo desta seção é calcular analiticamente o valor esperado do número de arestas descobertas pelo processo de amostragem definido na Seção 2. Assim como no caso de vértices, iremos derivar a probabilidade de uma aresta ser descoberta quando temos k amostras de monitores. Usaremos então esta probabilidade para calcular o número médio de arestas descobertas.

Seja $e = (u, v)$ uma aresta de rede incidente sobre os vértice u e v. Esta aresta será descoberta se o monitor escolhido for o vértice u ou o vértice v. Caso estejamos tratando do tipo de monitor que revela informação à distância 2, então a aresta e também será descoberta se um vizinho do vértice u ou um vizinho do vértice v for escolhido como monitor. A figura 3 ilustra estes dois casos. Repare que todos estes casos são mutuamente exclusivos, pois uma amostra de monitor assume a identidade de apenas um vértice da rede.

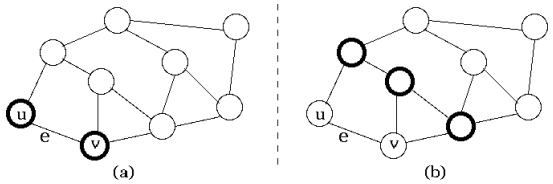


Figura 3. Exemplo das duas formas com a qual uma aresta $e = (u, v)$ pode ser descoberta por uma amostra de monitor: (a) um dos vértices incidentes a aresta é escolhido; (b) um vizinho de um dos vértices incidentes a aresta (que não é incidente a aresta) é escolhido.

Vamos considerar primeiro o tipo de monitor que revela vértices a distância 1. Seja D_e^k o evento que denota que a aresta e foi descoberta depois de k amostra de monitores. Como a escolha de monitores é uniforme, a probabilidade da aresta e ser descoberta após exatamente uma amostra é simplesmente $P[D_{e,1}] = 2/n$. Ou seja, um dos vértices incidentes a aresta e deve ser escolhido como monitor para que a aresta seja revelada.

Considere agora k amostras de monitores. Temos que a aresta e será descoberta se ao menos uma das k amostras for um dos vértices incidentes a aresta e. A probabilidade de nenhuma das k amostras revelar a aresta e é dada por $(1-2/n)^k$. Logo, a probabilidade de e ser revelada é o complemento desta probabilidade, dado por $P[D_e^k] = 1-(1-2/n)^k$.

Seja $Q_{e,k}$ uma variável aleatória indicadora que denota se a aresta e foi revelada (assumindo valor 1) ou não (assumindo valor 0) ao fazermos k amostras de monitores. Lembrando que W_k representa o número de arestas descobertas quando depois de k amostras de monitores, esta pode ser definida como $W_k = \sum_{e \in E} Q_{e,k}$, onde E representa o conjunto de arestas da rede. Temos que $E[W_k] = \sum_{e \in E} E[Q_{e,k}] = \sum_{e \in E} P[D_e^k]$ pois temos que $Q_{e,k}$ é uma variável aleatória indicadora e seu valor esperado é dado pela probabilidade dela assumir valor 1. Repare que a equação acima depende do número de arestas na rede. Seja M a variável aleatória que denota o número de arestas na rede, ou

seja, $M = |E|$. Podemos obter o valor $E[W_k]$ utilizando a regra do valor esperado condicional, ou seja, $E[W_k] = E[E[W_k|M]]$. Repare que temos $E[W_k|M] = MP[D_e^k]$, pois as arestas são estatisticamente equivalentes. E finalmente, temos que:

$$E[W_k] = E[MP[D_e^k]] = E[M]P[D_e^k] \tag{11}$$

pois $P[D_e^k]$ não depende de M e $E[M]$ é o valor esperado do número de arestas na rede. Mais ainda, temos que $E[M] = nE[Z]/2$, pois o valor esperado do número de vértice está relacionado com o valor esperado do número de arestas. Assim sendo, temos finalmente que:

$$E[W_k] = nE[Z]/2\,(1 - (1 - 2/n)^k) \tag{12}$$

É importante notar que diferentemente do número de vértices, o valor esperado do número de arestas descobertas não depende da distribuição de grau da rede, e sim apenas do grau médio, $E[Z]$.

4.1. Monitor revela vértices a distância 2

Vamos considerar agora monitores que revelam vértices a distância 2. Neste caso, precisamos considerar que a aresta $e = (u, v)$ será descoberta também quando um vizinho de um dos seus vértices incidentes for escolhido como monitor, conforme ilustrado na figura 3(c). Precisamos calcular então o número de vértices que são vizinhos aos vértices u e v.

Seja $N_{uv}^{01} = N_u^0 \cup N_u^1 \cup N_v^0 \cup N_v^1$ o conjunto de vértices que ao serem escolhidos como monitor revelam a aresta $e = (u, v)$. Nosso objetivo é calcular a cardinalidade deste conjunto, pois a probabilidade da aresta e ser descoberta quando uma monitor é escolhido, $k = 1$, é simplesmente $P[D_e^1] = |N_{uv}^{01}|/n$.

Considere o vértice u incidente a aresta e. Vamos condicionar em seu grau para obter o número de vizinhos de u. Um destes vizinhos é o vértice v e precisamos considerar o restante de grau de v para contar os vizinhos de v que também podem revelar a aresta e. Entretanto, um vizinho de v também pode ser vizinho de u, em particular se a rede possuir um alto grau de clusterização. Mas podemos aproximar a quantidade destes vértices considerando o coeficiente de clusterização da rede e o grau do vértice u de forma similar ao procedimento para descobrimento de vértices.

Seja $\bar{c}$ o coeficiente de clusterização médio da rede e A_d o número de arestas entre os vizinhos do vértice u dado que seu grau é igual a d, cujo valor esperado é dado pela equação 7. Entretanto, estamos interessados apenas nas arestas entre o vértice v e os outros vizinhos de u e não em todas as arestas entre todos os vizinhos de u. Podemos aproximar o número de arestas entre v e os outros vizinhos de u assumindo que $E[A_d]$ está distribuído igualmente entre os d vizinhos de u, sendo um deles o vértice v. Desta forma, o número de arestas que incidem sobre v e outros vizinhos de u é dado por $2E[A_d]/d$. O fator multiplicativo 2 é necessário pois cada aresta possui duas pontas que serão distribuídas pelos d vizinhos.

Podemos agora estimar o valor para $|N_{uv}^{01}|$ dado que o grau do vértice u é d da seguinte forma:

$$|N_{uv,d}^{01}| \approx 1 + d + E[R] - 2E[A_d]/d \tag{13}$$

onde 1 representa o vértice u, d representa seus vizinhos, que inclui v, $E[R]$ representa o restante de grau de v, ou seja, os vizinhos de v e $2E[A_d]/d$ representa os vizinhos de

v que também são vizinhos de u. Utilizando esta aproximação e substituindo os valores para $E[R]$ e $E[A_d]$ e simplificando, podemos calcular $P[D_e^1|Z=d]$, ou seja:

$$P[D_e^1|Z=d] \approx d/n + E[Z^2]/(nE[Z]) - \overline{c}(d-1)/n \tag{14}$$

Lembrando que como visto na seção 3, $E[R] = E[Z^2]/E[Z] - 1$ e $E[A_d]$ está definido apenas quando $d > 1$.

A probabilidade da aresta e ser revelada ao menos uma vez por uma das k monitores pode ser calculada considerando seu complemento. Como cada amostra de monitor é independente, a probabilidade da aresta não ser revelada é dada por $(1 - P[D_e^1|Z=d])^k$. Logo, temos:

$$P[D_e^k|Z=d] = 1 - (1 - d/n - E[Z^2]/(nE[Z]) + \overline{c}(d-1)/n)^k \tag{15}$$

Podemos agora descondicionar o grau do vértice u e obter a probabilidade de uma aresta e ser descoberta, ou seja:

$$P[D_e^k] = \sum_{d=0}^{n-1}(1 - (1 - d/n - E[Z^2]/(nE[Z]) + \overline{c}(d-1)/n)^k)P[Z=d] \tag{16}$$

onde $P[Z = d]$ é a distribuição de grau dos vértices da rede. Finalmente, podemos calcular o valor esperado do número de arestas descobertas quando temos k amostras de monitores na rede, substituindo a equação acima para $P[D_e^k]$ em 11 e simplificando para $E[M]$, de forma que temos:

$$E[W_k] = nE[Z]/2 \sum_{d=0}^{n-1}(1 - (1 - d/n - E[Z^2]/(nE[Z]) + \overline{c}(d-1)/n)^k)P[Z=d] \tag{17}$$

É importante notar que a equação acima é uma aproximação para o valor esperado do número de arestas descobertos e que ainda assim depende também da distribuição de grau dos vértices da rede. Assim como no caso para descobrimento de vértices, a análise acima indica a dificuldade de se calcular analiticamente a quantidade de arestas descobertas quando amostras de monitores revelam vértices a distância 2.

5. Avaliação numérica

Nesta seção iremos comparar as previsões dos modelos analíticos com o resultado obtido através de simulação detalhada do processo de amostragem. A comparação será baseada em redes geradas por três modelos de redes clássicos: modelo de Erdoős-Rényi, conhecido por $G(n, p)$ [P. Erdos 1960], modelo de Watts e Strogatz, conhecido como modelo *small world* (SW) [Watts and Strogatz 1998]; modelo de Barabási e Albert, conhecido como modelo de *preferential attachment* (BA) [Barabási and Albert 1999].

Todas as avaliações consideram uma rede com 5000 vértices e três diferentes graus médios: 4, 16 e 64. As simulações representam fielmente o processo de amostragem de vértices, conforme descrito na seção 2. Para cada cenário, 100 redes diferentes são geradas pelo modelo de rede, e para cada rede são realizadas 100 rodadas de simulação, onde realizamos 2500 amostras de monitores. Além da média amostral do número de vértices e arestas descobertos em cada cenário, calculamos ainda um intervalo de confiança de 95%.

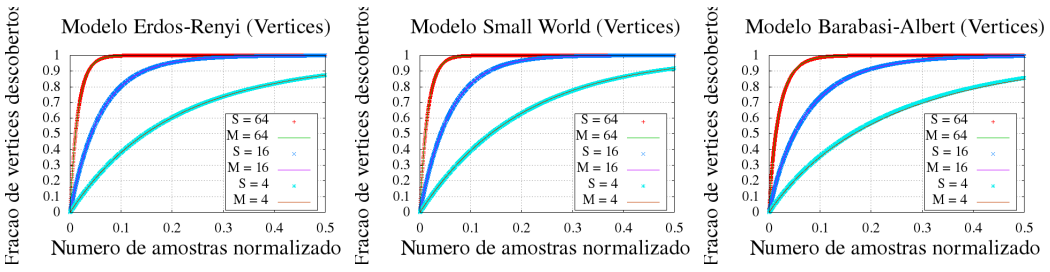
Os modelos analíticos propostos foram parametrizados com as respectivas características das redes geradas pelos modelos sendo utilizados. Em particular, para cada um dos três modelos de rede sendo considerado, utilizamos a distribuição de grau, seu valor esperado e segundo momento, e a clusterização induzida pelo modelo. Com exceção do modelo BA, para o qual não existe um modelo exato para a clusterização média do grafo, e neste caso utilizamos resultados numéricos obtidos a partir das redes geradas (somente para clusterização média).

Para melhor comparação dos resultados apresentaremos o eixo das abscissas como número de amostras normalizado, dividindo o número de amostras pelo número de vértices da rede e o eixo das ordenadas como fração de vértices ou arestas descobertas. As curvas em cada gráfico indicam os resultados obtidos pelo modelo (M) ou simulação (S) para os diferente graus médios.

5.1. Descoberta de vértices

As Figuras 4(a), 4(b) e 4(c) apresentam os resultados de vértices para os três modelos de redes utilizados quando um monitor revela vértices até distância 1. Repare que em todo os casos, o resultado do modelo analítico proposto é idêntico ao resultado obtido por simulação. Podemos ver que com grau médio igual a 4, nem todos os vértices são descobertos pois o grau médio é baixo e mais amostras são necessárias para descobrir a rede toda. Além disso, podemos verificar que quanto maior o grau médio, mais rápido o processo de amostragem descobre os vértices da rede. Em particular, quando o grau médio é 64, verificamos que mais de 90% dos vértices são descobertos com apenas 5% de amostras.

As Figuras 5(a), 5(b) e 5(c) apresentam os resultados quando um monitor revela vértices até distância 2. O modelo analítico aproximado apresenta um resultado idêntico ao de simulação para o modelo $G(n, p)$, pois a rede neste caso é construída conectando os vértices de forma independente, tornando a aproximação para o valor esperado de restante de grau de um vizinho ($E[R]$) adequada para este tipo de rede. Entretanto, este não é o caso para as redes SW e BA e a aproximação para o valor esperado do restante de grau de um vizinho não é boa, influenciando negativamente os resultados do nosso modelo. Um outro aspecto está na clusterização média utilizada para aproximar o número de arestas entre os vizinhos de um vértice, que é necessário para estimar o número de vértices à distância 2 (Equação (8)).



(a) Modelo G(n,p) (b) Modelo SW (c) Modelo BA

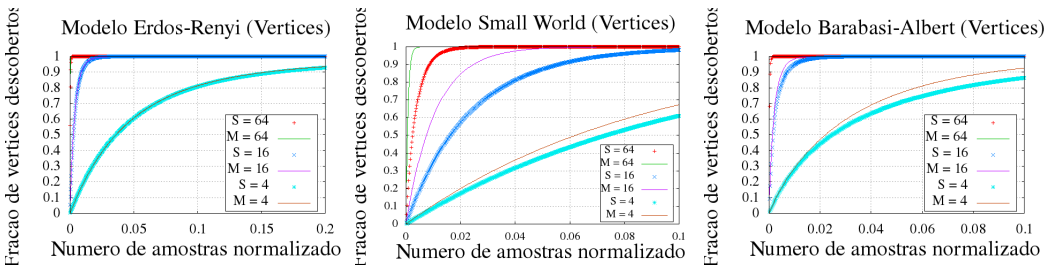
Figura 4. Descoberta de vértices quando monitor revela vértices até distância 1

A Tabela 1 compara os valores da clusterização média e número médio de vértices a distância 2 utilizadas por nosso modelo com resultados de simulação para diferentes

Tabela 1. Clusterização e número de vértices a distância 2 ($n = 5000$).

	Clusterização Média ($\overline{c}$)						Número de vértices a distância 2 (equação (8))					
	Analítico			Simulação			Analítico			Simulação		
E[Z]	4	16	64	4	16	64	4	16	64	4	16	64
BA	$0,001$	$0,001$	$0,001$	$0,004$	$0,017$	$0,054$	$35,2$	$603,7$	10621	$33,6$	$504,6$	3482
G(n,p)	$0,0008$	$0,003$	$0,012$	$0,0008$	$0,003$	$0,012$	$15,9$	$255,1$	4043	$15,8$	$247,4$	2767
SW	$0,48$	$0,68$	$0,72$	$0,47$	$0,65$	$0,69$	$6,1$	$75,5$	1115	$7,3$	41	283

redes e graus médios. Podemos verificar que para as redes BA e SW os valores para número médio de vértices a distância 2 são bastante diferentes, principalmente para grau 64, explicando o baixo desempenho do nosso modelo em calcular o número de vértices descobertos neste caso (Figura 5). Repare que a rede SW possui a maior diferença relativa justificando seu pior desempenho (Figura 5(b)). Por fim, quando o grau médio é pequeno os valores são mais próximos e nosso modelo apresenta melhor desempenho.



(a) Modelo G(n,p) (b) Modelo SW (c) Modelo BA

Figura 5. Descoberta de vértices quando monitor revela vértices até distância 2

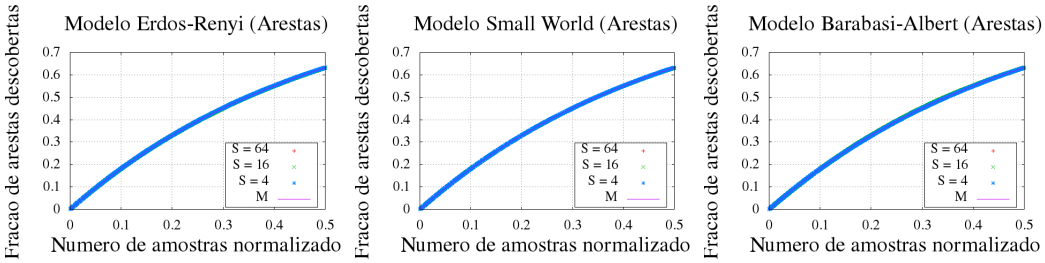
5.2. Descoberta de arestas

As Figuras 6 e 7 ilustram os resultados de descoberta de arestas para os três modelos de redes. As Figuras 6(a), 6(b) e 6(c) apresentam o caso onde o monitor revela vértices até distância 1. O resultado obtido através da equação 12, mostra que a *fração* de arestas descobertas independe da distribuição de grau e do grau médio da rede, o que pode ser confirmado pelo resultado de simulação dos três modelos. Este é um fato muito interessante, pois indica que o processo de amostragem de vértices descobre arestas sempre da mesma maneira, independente de qualquer propriedade estrutural da rede sendo avaliada. Outro aspecto importante é que o descobrimento de arestas é bem mais lento que o de vértices, pois com 10% de amostras descobrimos em média apenas 20% das arestas.

O resultado para o caso onde monitores revelam vértices até distância 2 é apresentado nas Figuras 7(a), 7(b) e 7(c). Notamos que para distância 2, a fração de arestas descobertas depende da distribuição de grau da rede e é proporcional ao grau médio (o que não é o caso para distância 1). Além disso, os resultados para o modelo analítico aproximado é surpreendente para as redes $G(n, p)$ e SW, pois são idênticos aos obtidos por simulação. Para o modelo BA, o resultado difere novamente pois a estimativa da clusterização não é boa, conforme ilustrado na Tabela 1. Além disso, a distribuição de grau utilizada em nosso modelo também é uma aproximação para quando o número de vértices da rede é muito grande.

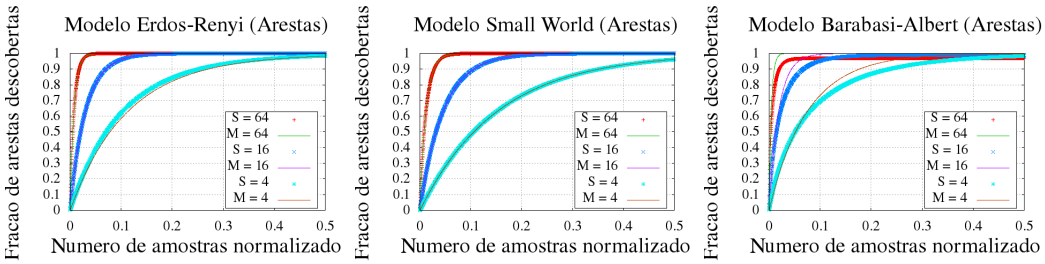
6. Trabalhos relacionados

Informação estrutural das mais variadas redes em geral não está disponível publicamente de maneira centralizada ou organizada e precisa ser coletada através de algum processo de amostragem. Em muitos casos a quantidade de informação disponível é proibitiva para



(a) Modelo G(n,p) (b) Modelo SW (c) Modelo BA

Figura 6. Descoberta de arestas quando monitor revela vértices até distância 1



(a) Modelo G(n,p) (b) Modelo SW (c) Modelo BA

Figura 7. Descoberta de arestas quando monitor revela vértices até distância 2

ser coletada exaustivamente, como é o caso do Facebook, Web, rede de citações, ou uma rede P2P. Desta forma, o projeto e avaliação de processos de amostragem que coletem informações estruturais de forma representativa e sem tendências se torna fundamental no estudo empírico de redes.

Diversos trabalhos recentes propõem e avaliam diferentes processos de amostragem em redes [Ribeiro and Towsley 2010, Kurant et al. 2011a, Gjoka et al. 2011, Kurant et al. 2011b, Avrachenkov et al. 2010, Jin et al. 2011, Pedarsani et al. 2008]. Um dos processos de amostragem mais estudados são os passeios aleatórios (*random walks*) por terem suas propriedades relativamente bem conhecidas [Ribeiro and Towsley 2010, Kurant et al. 2011a]. Uma outra técnica bastante utilizado na prática é a busca em largura (BFS) apesar de sua caraterização teórica ainda ser um desafio [Gjoka et al. 2011, Kurant et al. 2011b]. Além disso, outras técnicas misturando passeios aleatórios com amostragem de vértices vem sendo propostas na literatura [Jin et al. 2011, Avrachenkov et al. 2010].

Nenhum dos trabalhos relacionados acima apresenta uma modelagem do processo de amostragem de vértices conforme proposta neste trabalho. Além disso, o processo de amostragem de vértices é importante, pois serve como abstração para outros processos reais de descobrimento de redes, principalmente quando monitores podem ser utilizados para descobrir informações locais, como no caso de redes P2P. Em [Kryczka et al. 2011], por exemplo, amostras dos peers (e suas conexões) foram coletadas de aproximadamente 250 swarms BitTorrent com o objetivo de analisar as características estruturais e a evolução topológica destas redes. Por fim, vértices que revelam informações localmente, à distância 1 e 2, conforme considerado neste trabalho, também vem sendo propostos e avaliados na literatura no contexto de busca de informação [Mihail et al. 2006].

7. Conclusão

Neste trabalho definimos um processo de amostragem de vértices baseado em monitores escolhidos aleatoriamente em uma rede que revelam informação local até distância 1 e

até distância 2. Apresentamos em seguida um modelo para prever o número esperado do número de vértices e arestas descobertos por estes processos em função do número de amostras. Por fim, fazemos uma avaliação numérica comparando as previsões dos modelos com os resultados de simulação detalhada utilizando três modelos de redes aleatórias.

Nossos resultados confirmam que o modelo para descobrimento de vértices e arestas a distância 1 é exato e que no caso de arestas os resultados não dependem da estrutura da rede. Além disso, os resultados também mostram que o modelo proposto para o caso de distância 2 oferece boas aproximações para diversos casos. Por fim, estes modelos podem ser usados para guiar o projeto de algoritmos que irão descobrir redes reais, tarefa que deixamos como trabalho futuro.

Referências

Avrachenkov, K., Ribeiro, B., and Towsley, D. (2010). Improving random walk estimation accuracy with uniform restarts. In *Workshop on Alg. Models for Web Graph (WAW)*.

Barabási, A.-L. (2009). Scale-free networks: A decade and beyond. *Science*, 325.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286.

Gjoka, M., Butts, C. T., Kurant, M., and Markopoulou, A. (2011). Practical recommendations on crawling online social networks. *IEEE JSAC*, 29(9).

Jin, L., Chen, Y., Hui, P., Ding, C., Wang, T., Vasilakos, A., Deng, B., and Li, X. (2011). Albatross sampling: Robust and effective hybrid vertex sampling for social graphs. In *ACM Workshop on Hot Topics in Planet-scale Measurement (HotPlanet)*.

Kryczka, M., Cuevas, R., Guerrero, C., and Azcorra, A. (2011). Unrevealing the structure of live bittorrent swarms: Methodology and analysis. In *IEEE P2P*.

Kurant, M., Gjoka, M., Butts, C. T., and Markopoulou, A. (2011a). Walking on a graph with a magnifying glass: stratified sampling via weighted random walks. In *ACM SIGMETRICS*.

Kurant, M., Markopoulou, A., and Thiran, P. (2011b). Towards unbiased bfs sampling. *IEEE JSAC - Special Issue on Measurement of Internet Topologies*, 29(9).

Mihail, M., Saberi, A., and Tetali, P. (2006). Random walks with lookahead on power law random graphs. *Internet Mathematics*, 3(2):147–152.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

P. Erdos, A. (1960). On the evolution of random graphs. *Publications of Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61.

Pedarsani, P., Figueiredo, D. R., and Grossglauser, M. (2008). Densification arising from sampling fixed graphs. In *ACM SIGMETRICS*.

Ribeiro, B. and Towsley, D. (2010). Estimating and sampling graphs with multidimensional random walks. In *ACM SIGCOMM Internet Measurement Conference*.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393:440–442.