MODELAGEM E CARACTERIZAÇÃO DE UM PROCESSO DE AMOSTRAGEM DE VÉRTICES EM REDES

Vicente de Melo Pinheiro

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Daniel Ratton Figueiredo

Rio de Janeiro
Dezembro de 2013

MODELAGEM E CARACTERIZAÇÃO DE UM PROCESSO DE AMOSTRAGEM DE VÉRTICES EM REDES

Vicente de Melo Pinheiro

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Daniel Ratton Figueiredo, Ph.D.

Prof. Mario Roberto Folhadela Benevides, Ph.D.

Prof. Michele Garetto, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
DEZEMBRO DE 2013

Pinheiro, Vicente de Melo

Modelagem e Caracterização de um Processo de Amostragem de Vértices em Redes/Vicente de Melo Pinheiro. – Rio de Janeiro: UFRJ/COPPE, 2013.

X, 47 p.: il.; 29, 7cm.

Orientador: Daniel Ratton Figueiredo

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2013.

Referências Bibliográficas: p. 46 – 47.

1. Redes de Computadores. 2. Redes Complexas. 3. Processo de Amostragem. 4. Amostragem aleatória. I. Figueiredo, Daniel Ratton. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

MODELAGEM E CARACTERIZAÇÃO DE UM PROCESSO DE AMOSTRAGEM DE VÉRTICES EM REDES

Vicente de Melo Pinheiro

Dezembro/2013

Orientador: Daniel Ratton Figueiredo

Programa: Engenharia de Sistemas e Computação

O crescente interesse em estudar como as "coisas" se conectam vem sendo alavancado pela crescente abundância de grandes quantidades de dados relativos as mais diversas redes. Neste contexto, um importante aspecto é a forma como esses dados são coletados, pois na maioria das vezes as informações sobre vértices e arestas não estão disponíveis publicamente de forma centralizada ou mesmo de maneira organizada (por exemplo, Web, P2P, Facebook). Com isso, se torna necessário descobrir essas redes através de um processo de amostragem onde este processo influencia fundamentalmente na forma que a rede é descoberta. Neste trabalho, nós estudamos o processo de amostragem de vértices que revelam informações locais de vértices escolhidos aleatoriamente na rede. Particularmente, desenvolvemos modelos analíticos para determinar o número de vértices e arestas descobertos pelo processo de amostragem de acordo com o número de amostras e outras características da rede (por exemplo, grau médio). A avaliação dos modelos propostos comparada aos resultados obtidos através de simulação para diferentes modelos de redes indicam as condições sobre as quais nosso modelo captura bem a realidade.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

MODELING AND CHARACTERIZATION OF A VERTEX SAMPLING PROCESS IN NETWORKS

Vicente de Melo Pinheiro

December/2013

Advisor: Daniel Ratton Figueiredo

Department: Systems Engineering and Computer Science

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.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

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* [1]. Nos últimos anos, estudos sobre as mais diversas redes vem revelando características estruturais fundamentais e contribuindo na compreensão de fenômenos que operam sobre estas redes. A disseminação de uma epidemia por uma rede social; a distribuição de um arquivo por uma rede peer-to-peer (P2P); o ranqueamento de pesquisadores em redes de colaboração científica são exemplos reais de redes que apresentam ao mesmo tempo similaridade e distinção em suas estruturas.

Uma das principais razões para o grande avanço nesta área é o crescente surgimento de diversas redes de grande porte. Essas redes motivam estudos empíricos e validações de modelos teóricos que capturem as mais diferentes propriedades. Entretanto, um importante aspecto que precisa ser levado em conta é a obtenção destas redes, pois diversas vezes os dados não estão disponíveis publicamente de forma centralizada ou organizada, e na maioria das vezes precisam ser coletados através de algum procedimento.

Podemos utilizar como exemplo a rede social do Facebook, onde representaríamos usuários através de nós e a relação de amizade entre duas pessoas através de uma aresta. Coletar informações de uma rede desse tamanho pode ser um problema maior do que parece. Coletar esse tipo de informação demora muito tempo e em alguns casos ainda são aplicadas restrições a essa busca.

Imagine que queremos amostrar a rede da web. Vamos supor que essa rede tenha 10 bilhões de páginas e que demoramos em média 100 milissegundos para visitar cada uma dessas páginas. Para coletar essa rede seriam necessários pelo menos 32 anos. Logicamente o cálculo puro e simples é uma abordagem ingênua,

mas ilustra bem a grandeza das redes e massas de dados de que estamos falando. Um outro exemplo seria a rede do BitTorrent durante um *swarm*, representaríamos peers através de nós e conexões TCP com arestas. Em ambos os casos citados acima a rede precisa ser descoberta através de algum *processo de amostragem*, que irá revelar seus vértices e arestas para que então seja possível estudar as propriedades das respectivas estruturas.

Um problema fundamental ao descobrir uma rede passa a ser a influência do processo de amostragem que será utilizado, pois em muitos casos é proibitivo coletar a rede inteira. Utilizando novamente o Facebook como exemplo, coletar toda a rede é inviável devido ao seu tamanho (mais de 1 bilhão de nós) e do controle imposto pela empresa que limita a velocidade com a qual os dados podem ser coletados. Uma alternativa é tentar coletar parte desta grande rede para então trabalhar com esses dados, e a forma de amostrar essa rede também influenciará diretamente na estrutura da rede que vamos obter.

Uma técnica muito usada para coletar parte de uma rede é utilizar um processo de amostragem aleatório. Diversas áreas de pesquisa utilizam processos de amostragem baseados por exemplo em passeios aleatórios (*random walks*), onde a ideia é dar sucessivos passos sobre a rede em direções aleatórias para descobrir novos vértices e assim descobrir parte de sua estrutura. A estrutura da rede obtida é apenas uma amostra da rede toda e portanto, não possui necessariamente as mesmas características da rede como um todo.

Uma outra técnica para amostrar uma rede é através de seus vértices. Imagine que temos um grafo qualquer onde um de seus vértices é revelado aleatoriamente. Vamos supor que esse vértice traz consigo informações locais, como suas arestas e seus vizinhos na rede. Agora basta repetirmos esse procedimento diversas vezes para começar a descobrir outros vértices e seus vizinhos. Um ponto importante é entender este processo de amostragem. Por exemplo, quantas amostras são necessárias para descobrir quase toda a rede? Conseguimos estimar o número de vértices e amostras descobertas utilizando este processo? E da rede original como um todo? Estas e outras perguntas fazem parte dos nossos questionamentos neste trabalho.

1.2 Contribuição

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 de redes locais (ex. seus vizinhos). Iremos considerar dois casos de informação local: *(i)* o vértice escolhido revela sua identidade e todos seus vizinhos; *(ii)* vértice escolhido releva sua identidade, todos seus vizinhos, todos os vizinhos dos vizinhos e arestas entre esses vértices. E as principais contribuições

deste trabalho são:

- **Modelagem analítica do processo de amostragem.** Estamos interessados em caracterizar como este processo descobre 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.

- **Simulação e avaliação.** Desenvolvemos um ambiente para simular o processo de descobrimento em redes para gerar resultados empíricos, fizemos uma avaliação numérica utilizando quatro modelos clássicos de redes e comparamos os resultados previstos pelos modelos com resultados obtidos através dessa simulação detalhada do processo de amostragem.

  Nossos resultados validam o modelo exato e mostram que o modelo aproximado possui bom desempenho para todos os casos de descobrimento de arestas, além de um bom desempenho em todos os casos de descobrimento de vértices até distância 1. Para descobrimento de vértices até distância 2, apresentamos um bom desempenho para dois dos modelos. 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, e a partir disso podemos inserir vários vértices controlados e amostrar a rede para estudar características de sua estrutura; Em redes sociais online podemos amostrar vértices usando identificadores aleatórios e descobrir informações locais a estes vértices. Contudo, nosso objetivo neste trabalho não é aplicar o método de amostragem de vértices e sim caracterizar como o mesmo descobre uma rede.

1.3 Organização desta dissertação

Esta dissertação está organizada em 6 capítulos, sendo este primeiro a introdução.

No capítulo 2 serão apresentados alguns trabalhos relacionados e a literatura para compreender este trabalho. Apresentaremos os modelos de grafos aleatórios que serão utilizados, evidenciando as particularidades em sua estrutura, entraremos em detalhes sobre processos de amostragem conhecidos e já utilizados na literatura e finalmente apresentaremos alguns trabalhos relacionados que abordam o problema de descobrimento de redes.

No capítulo 3, definiremos formalmente o processo de amostragem utilizado, justificando, exemplificando e abordando as diferentes variações adotadas neste trabalho. Explicaremos o que consideramos como monitor (vértice revelado) e, em seguida, apresentaremos as medidas de interesse que utilizaremos para caracterizar como o processo de amostragem revela informações sobre a rede.

No capítulo 4, faremos uma avaliação teórica apresentando modelos analíticos. Estudaremos o número de monitores distintos após um certo número de amostras e analisaremos o descobrimento na rede, detalhando a técnica utilizada para deduzir o modelo matemático que captura o número de vértices e arestas descobertos pelo processo.

No capítulo 5, apresentaremos o simulador que foi desenvolvido para validação dos modelos, descreveremos o cenário utilizado (tamanho da rede, grau médio, parâmetros do simulador, etc.) e apresentaremos os resultados de descoberta de vértices e arestas, analisando e criticando seu comportamento ao ser comparado com resultados analíticos.

Por fim, o capítulo 6 apresenta a conclusão desta dissertação, apontando os resultados mais interessantes e trabalhos futuros que podem ser derivados deste.

Capítulo 2

Trabalhos Relacionados

A seguir, será apresentada parte da literatura dos trabalhos científicos que possuem relação com o tema desta dissertação.

2.1 Modelos grafos aleatórios

Grafos aleatórios são grafos gerados a partir de algum processo aleatório definido por um modelo. O número de vértices e arestas e o conjunto de arestas são exemplos de características que podem variar entre realizações do modelo. Grafos aleatórios vem sendo amplamente utilizados na literatura por conta de características estruturais específicas observadas após sua geração. Em muitos casos as características observadas são semelhantes a características existentes em redes reais, o que facilita o estudo e comparação entre resultados obtidos com redes sintéticas e redes reais.

Existem diversos modelos de grafos aleatórios propostos na literatura. Diferentes modelos dão origem a gráficos com diferentes propriedades estruturais, entre elas talvez a mais marcante seja a distribuição de grau.

A clusterização de um vértice representa a probabilidade de dois vizinhos de um mesmo vértice também serem vizinhos. Essa métrica pode ser calculada localmente para um único vértice ou globalmente como uma média para todo o grafo. Outras métricas muito importantes são observadas e estudadas na literatura como por exemplo a distância entre vértices.

Neste trabalho optamos por quatro modelos de grafos aleatórios. Utilizamos esses modelos para gerar redes sintéticas através de um simulador, avaliar e caracterizar o processo de amostragem proposto. Geramos diversas instâncias de cada modelo para avaliar o processo de amostragem aplicado ao conjunto de redes que recebem aquela denominação.

2.1.1 Modelo de Erdős-Rényi - $G(n, p)$

Dentre os diversos modelos de grafos aleatórios existentes na literatura atualmente, o mais antigo e muito utilizado é o modelo Erdős-Rényi.

O modelo Erdős-Rényi, também conhecido como $G(n, p)$, possui dois parâmetros determinísticos, n e p. O parâmetro n define o número de vértices do grafo e p a probabilidade de existência de cada aresta do grafo, independentemente. É importante ressaltar que mesmo com dois parâmetros determinísticos, o grafo gerado é aleatório e pode variar como mostra a figura 2.1. Na verdade, qualquer grafo com cinco vértices pode ser gerado neste exemplo, e cada um deles possui uma probabilidade de ser gerado.

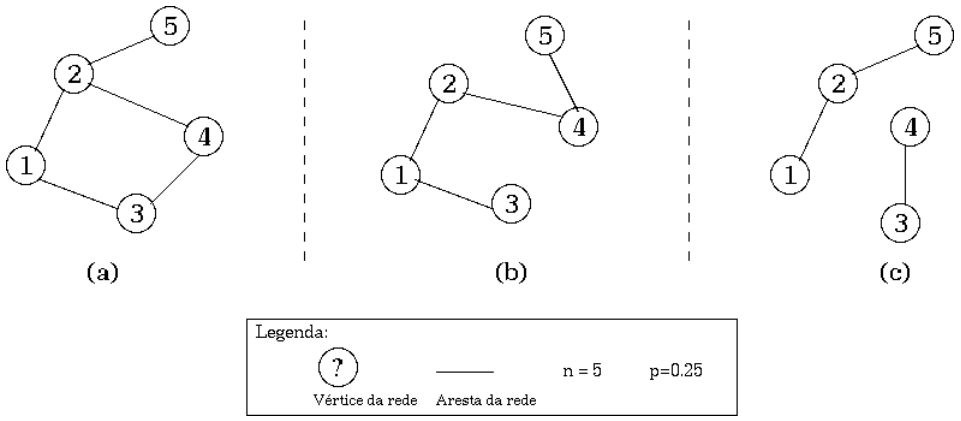


Figura 2.1: Exemplos de grafos diferentes gerados através do modelo $G(n, p)$ com parâmetros iguais.

Podemos ressaltar também outras características marcantes do modelo, como sua distribuição de grau, valor esperado do grau de um vértice, número esperado de arestas e clusterização. A distribuição de grau de um grafo $G(n, p)$ é facilmente deduzida [2] e segue uma distribuição binomial observada na equação 2.1 onde Z representa a variável aleatória que denota o grau de um vértice.

$$P[Z = d] = \binom{n-1}{d} p^d (1-p)^{n-1-d}; 0 \leq d < n \tag{2.1}$$

Consequentemente seu grau médio é:

$$E[Z] = (n-1)p \tag{2.2}$$

Outra medida que também pode ser obtida facilmente é a clusterização média do grafo apresentada na equação 2.3.

$$\overline{c} = p \tag{2.3}$$

O valor da clusterização no modelo $G(n, p)$ é p pois probabilidade de cada aresta é independente.

2.1.2 Modelo de Watts-Strogatz - Small World

O modelo Watts-Strogatz é um modelo de geração de grafo aleatório que pode produzir redes com propriedades *Small World*(SW). O modelo SW representa redes onde temos alta clusterização e curtas distâncias (numero de passos de um vértice a outro).

Um látice inicial é utilizado para iniciar o processo do modelo Watts-Strogatz. Cada vértice é ligado então aos seus k vizinhos mais próximos como mostra a figura 2.2. Em seguida cada aresta é desconectada de seus vértices originais com uma probabilidade p e religada aleatoriamente em outros dois vértices.

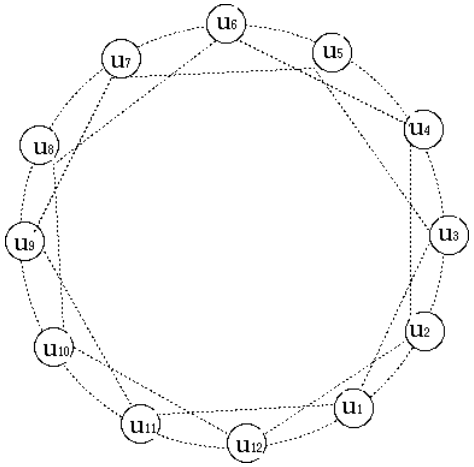


Figura 2.2: Exemplo da primeira etapa de geração de uma rede através do modelo Watts-Strogatz ($k = 4$).

Como o modelo inicia com uma estrutura não aleatória que possui um vértice ligado a k vizinhos, isso origina um grafo com alta clusterização e distâncias médias altas com grau médio k:

$$E[Z] = k \tag{2.4}$$

O segundo passo, que reposiciona as arestas, pode criar "atalhos" que reduzem a distância média do grafo e fazem com que o mesmo seja aleatório e que seja possível gerar um grafo do tipo SW.

A sua distribuição de grau [3] pode ser obtida da seguinte forma:

$$P[Z = d] = \sum_{i=0}^{f(d,k)} \binom{\frac{k}{2}}{i} (1-p)^i p^{\frac{k}{2}-i} \frac{(\frac{k}{2}p)^{d-\frac{k}{2}-i}}{(d-\frac{k}{2}-i)!} \exp(-p\frac{k}{2}) \tag{2.5}$$

Onde $d \geq \frac{k}{2}$ e $f(d,k) = min(d - \frac{k}{2}, \frac{k}{2})$. Sua clusterização [4] é calculada através de:

$$\overline{c} = \frac{3(k-2)}{4(k-1)}(1-p)^3 \tag{2.6}$$

2.1.3 Modelo de Barabási–Albert (BA)

O modelo Barabási–Albert (BA) é utilizado para gerar redes aleatórias livres de escala baseado na ideia de *preferential attachment*. *Preferential attachment* é ideia de que novos vértices que entram na rede tendem a se relacionar com vértices mais populares. Ou seja, vértices populares tendem a ser cada vez mais populares com o passar do tempo. Essa teoria possui diversas aplicações em redes reais, como por exemplo a web, onde novas páginas tendem a criar links para páginas mais populares, mais conhecidas e confiáveis. Outro exemplo é a rede de citações onde novos artigos tendem a citar artigos mais populares.

O modelo Barabási–Albert inicia com um grafo conexo com m_0 vértices. Um novo vértice i é adicionado a cada iteração e ligado à $m \leq m_0$ vértices com probabilidade p_i proporcional ao grau dos vértices existentes, como mostra a equação 2.7 onde d_i é o grau do vértice i e temos a soma do grau de todos os vértice da rede.

$$p_i = \frac{d_i}{\sum_{\forall j \in V} d_j} \tag{2.7}$$

O modelo Barabási–Albert aplica a ideia de *Preferential attachment* em redes de forma que a popularidade passa a ser dada pelo grau dos vértices. Desta forma, depois da adição de um certo número de vértices, a distribuição de grau segue uma lei de potência, tendo sua distribuição de grau calculada da seguinte forma :

$$P[Z = d] = \frac{B(d, \gamma)}{B(m_0, \gamma - 1)} \tag{2.8}$$

onde $\gamma = 3$ no caso BA e B(x,y) é a Função Beta de Euller [5].

Como cada vértice traz para a rede m arestas, quando $n >> |m_0|$ temos aproximadamente mn arestas. Com isso podemos obter o grau médio:

$$E[Z] = \frac{2mn}{n} = 2m \tag{2.9}$$

A clusterização [6] pode ser definida como:

$$\overline{c} = \frac{6m^2((m+1)^2(\ln n)^2 - 8m \ln n + 8m)}{8(m-1)(6m^2 + 8m + 3)n} \tag{2.10}$$

2.1.4 Configuration Model (CM)

Configuration model (CM) é um modelo de redes aleatório que recebe como entrada uma sequência de graus (de soma par) dos vértices e não possui uma distribuição de grau específica. Cada vértice da rede terá um grau fixo e o processo de formação da rede será baseado nesse conjunto de entrada.

O fato do grau dos vértices ser fixo induz que o número arestas m também será fixo. Podemos calcular esse número somando o grau d de cada um dos vértices i e dividindo por 2 para evitar que cada aresta seja contada duas vezes, como mostra a Equação 2.11:

$$m = \frac{1}{2}\sum_{i}^{n} d_i \tag{2.11}$$

Para formar a rede utilizaremos as pontas das arestas. Cada aresta liga dois vértices e, portanto, possui duas pontas. Consideraremos que cada vértice possuirá d pontas de arestas, totalizando $2m$ pontas de arestas. O processo de formação da rede consiste em sortear, iterativamente, duas pontas de arestas aleatoriamente de maneira uniforme dentre todas existentes e então criar uma aresta ligando as mesmas.

Essa forma uniforme de criação das arestas no CM fará com que cada ponta de aresta tenha a mesma chance de ser ligada a qualquer outra ponta de aresta. Essa propriedade é muito importante pois garante independência no processo de formação da rede.

Como falamos anteriormente, esse modelo não possui uma distribuição de grau específica, portanto podemos calcular a distribuição de grau e o grau médio a partir da sequência de graus recebida como entrada como mostram as Equações 2.12 e 2.13 respectivamente :

$$P[Z = d] = \frac{\text{Número de vértices com grau d}}{\text{Número de vértices}} \tag{2.12}$$

$$E[Z] = \sum_{i=0}^{n} iP[Z = i] \tag{2.13}$$

A clusterização [7] do modelo pode ser definida como:

$$\overline{c} = \frac{1}{n}\frac{(E[Z^2] - E[Z])^2}{E[Z]^3} \tag{2.14}$$

2.2 Processos de amostragem em grafos

Informações estruturais das mais variadas redes em geral não estão disponíveis publicamente de maneira centralizada ou organizada e precisa ser coletada através de algum processo de coleta. Em muitos casos a quantidade de informação existente disponível impede que a mesma seja coletada exaustivamente. Muitos casos como a rede de amizade do Facebook, a Web, rede de citações ou até mesmo uma rede P2P são exemplos de redes onde coletar todas essas informações se torna inviável. Desta forma, o projeto e avaliação de processos que coletam informações estruturais dessas redes de forma representativa e sem tendências se torna fundamental no estudo empírico de redes. Processos de coletas de dados, ou seja, amostragem podem ser determinísticos ou aleatórios. No primeiro caso a amostragem é determinada seguindo uma regra determinística que define unicamente a ordem de amostragem. No segundo caso, a regra de amostragem é aleatória, podendo ou não considerar variáveis observadas no passado.

2.2.1 Processos determinísticos

Um processo de amostragem determinístico é um processo onde o método de amostragem possui uma ordem de amostragem pré-definida. Essa ordem pode variar de acordo com o método. As escolhas de amostras serão feitas de acordo com algum critério, evitando assim que amostras sejam escolhidas ao acaso. Os métodos de busca em profundidade ou DFS (*Depth-first search*) e busca em largura ou BFS (*breadth-first search*) são exemplos de dois métodos muito conhecidos para percorrer os vértices de um determinado grafo.

A busca em largura consiste basicamente em a partir de um vértice, amostrar todos os seus vizinhos. Esse procedimento então ocorre iterativamente sobre esses vizinhos amostrados pelo vértice inicial. As figuras 2.3(a), (b) e (c) mostram iterativamente como funciona este processo de amostragem. Primeiramente iniciamos o processo de amostragem em um vértice qualquer (9 no nosso caso), que revela todos os seus vizinhos. Em seguida, visitamos todos os seus vizinhos que nos revelam seus vizinhos, e assim por diante (visitamos 6 e depois 7 na figura).

A busca em profundidade, como o próprio nome diz, tem como objetivo explorar a "fundo" cada vizinho encontrado de um vértice. As figuras 2.4(a), (b) e (c) mostram iterativamente como funciona este processo. Partimos de um vértice (9 no nosso caso) e fazemos a busca em cada primeiro vizinho encontrado (supomos que o primeiro vizinho seria o de menor índice). Sendo assim, amostramos os vizinhos do vértice 6 e em seguida do vértice 3.

Note que os dois mecanismos de busca, BFS e DFS, descobrem redes diferentes nos exemplos acima depois de explorar três vértices.

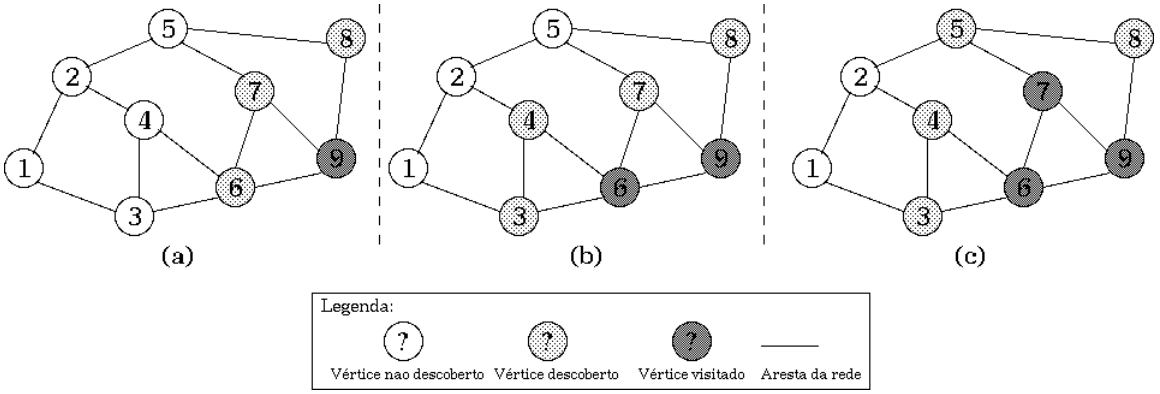


Figura 2.3: Exemplo de amostragem da rede utilizando busca em largura (BFS).

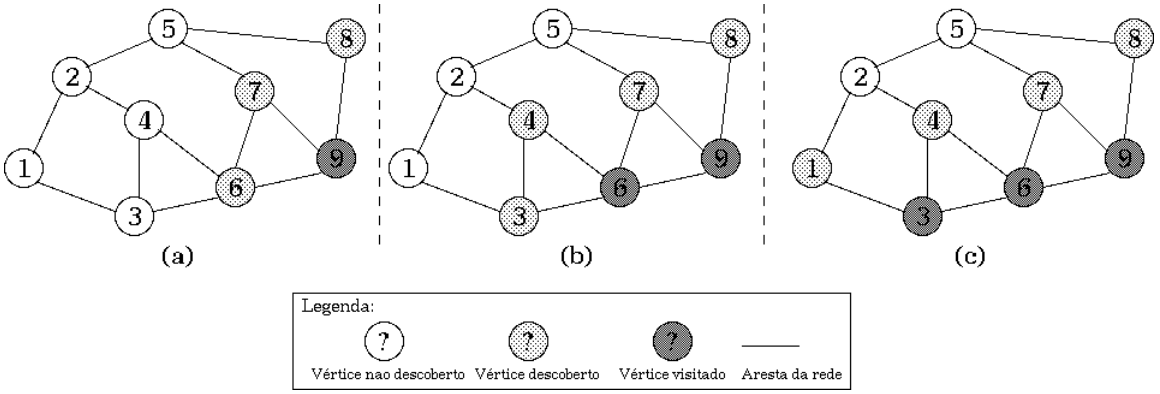


Figura 2.4: Exemplo de amostragem da rede utilizando busca em profundidade.

2.2.2 Processos aleatórios

A amostragem aleatória possui diversas variações. Um método muito conhecido e utilizado é o passeio aleatório (*Random Walk*). O random walk é um processo onde a cada passo, um vértice vizinho do atual é escolhido aleatoriamente para ser explorado. O random walk é um processo de amostragem aleatório sem memória, isto é, cada decisão de próximo passo de amostragem não leva em conta o seu estado anterior ou qualquer outro dado armazenado e decide aleatoriamente qual direção tomar na amostragem. Suponha que temos um passeio aleatório que revela vértices de uma rede representada pelo grafo da figura 2.5. Escolhemos um vértice inicial (9 no caso) e em seguida o passeio aleatório irá escolher um dos seus vizinhos para se posicionar no próximo passo. Note que o vértice 9 possui 3 vizinhos e com isso os vértices 6, 7 e 8 possuem a mesma chance de serem descobertos no próximo passo.

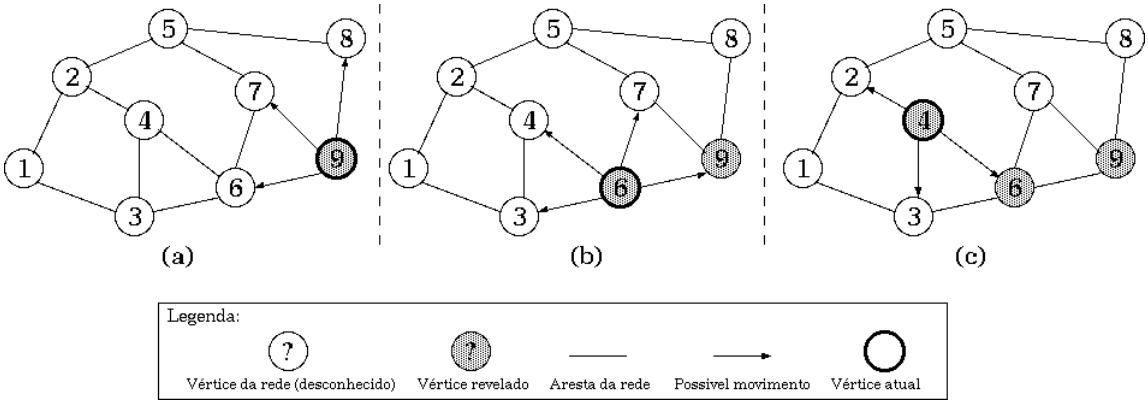


Figura 2.5: Exemplo de amostragem da rede através de um passeio aleatório.

Sendo assim, as figuras 2.5 (b) e (c) ilustram as próximas iterações do random walk.

2.3 Outros trabalhos

Diversos trabalhos recentes propõem e avaliam diferentes processos aleatórios de amostragem em redes [8–10]. Entender e identificar propriedades nas estruturas das redes também é muito importante e diversos trabalhos, principalmente nos últimos anos, vêm surgindo com o propósito de definir metodologias de amostragem que facilitem o estudo dessas redes [11, 12].

Um dos processos de amostragem mais estudados são os passeios aleatórios (*Random Walks*) por terem suas propriedades de amostragem relativamente bem conhecidas e estudadas. Em [8] os autores propõe um novo método de amostragem denominado *Frontier sampling* baseado em passeios aleatórios m-dimensionais. Esse novo método utiliza m passeios aleatórios *dependentes* caminhando sobre a rede e

descobrindo novos nós. Com este trabalho, os autores foram capazes de mitigar os erros de amostragem ao estimar propriedades da rede, tendo assim o *Frontier sampling* resultados mais eficientes na descoberta de vértices que o passeio aleatório tradicional e m passeios aleatórios *independentes*.

Um outro trabalho muito relacionado analisa o processo de amostragem BFS para caracterizar a tendência desse processo [10]. Baseado nisso, os autores propõe uma função de correção de tendência para esse processo e compara com outras funções já conhecidas na literatura, mostrando que essas são menos efetivas que o método proposto.

Com o enorme crescimento das redes sociais, processos de amostragem eficientes e não tendenciosos começaram a ser aplicados nesse universo. Em [9] os autores tem como objetivo obter uma amostra não tendenciosa dos usuários do Facebook através de um processo de amostragem. Neste estudo, várias abordagens são avaliadas afim de determinar qual o método mais eficaz. Após essa análise, uma das abordagens é utilizada para determinar propriedades estruturais da rede de amizades do Facebook.

Uma das aplicações mais populares na internet atualmente é o BitTorrent, um programa de *peer-to-peer* para compartilhamento de arquivos. No entanto, conhecer os detalhes da topologia de um *swarm* BitTorrent ainda é um desafio, tendo em vista que o tamanho e dinamismo da rede e a informação distribuída. Em [11] os autores se dedicam a mostrar através de estudos empíricos que *swarms* BitTorrent não podem ser considerados redes aleatórias como no modelo Erdős-Rényi nem redes *small-world*. Além disso, os autores consideram outros fatores como o caso onde a popularidade afeta a topologia de um *swarm* e o quanto dinâmica é a estrutura de um *swarm* os longo do tempo.

Por fim, diversos trabalham procuram relacionar redes reais com redes sintéticas afim de conseguir entender melhor

Capítulo 3

Processo de amostragem de vértices

O processo de amostragem tem como objetivo coletar iterativamente parte de uma rede definida por um grafo direcionado ou não-direcionado afim de obter dados suficientes para estudar suas características estruturais. Consideraremos um grafo não-direcionado denotado por $G = (V, E)$, onde V representa o conjunto de vértices que são rotulados unicamente de tamanho $|V| = n$ e E é o conjunto de arestas de tamanho $|E| = m$.

Podemos interpretar G como sendo uma rede sintética gerada por algum modelo de grafos aleatórios ou uma rede real obtida empiricamente. Esta abstração nos 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 no nosso caso, iremos assumir que não temos conhecimento da mesma. O processo de amostragem será responsável por revelar estas informações conforme descrito abaixo. Por fim, por mais que a maioria da redes reais sejam dinâmicas, 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).

3.1 Amostragem de vértices

O processo de amostragem de vértices é uma abstração de um processo real de descobrimento dos vértices de uma rede e consiste em revelar iterativamente um ou mais vértices, que chamaremos de *monitores*. O monitor será um vértice escolhido aleatoriamente entre todos os vértices da rede e cada monitor irá nos revelar informações locais a ele sobre a rede, tais como a identidade dos seus vizinhos. Repare que se o

vértice é vizinho do monitor, então existe uma aresta entre eles e a mesma também será revelada. Este processo poderá se repetir, com monitores sendo amostrados aleatoriamente, de forma iterativa, onde cada novo monitor potencialmente revelará novas informações 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. O monitor neste caso, além de revelar a si próprio, também revelará todas as arestas que incidem sobre ele e os vértices que estiverem na outra ponta destas arestas (vizinhos). Chamaremos esta caso de *monitor revela vértices a distância 1.*

- Monitor revela sua identidade, a identidade de todos seus vizinhos, e a identidade de todos os vizinhos dos vizinhos. Neste caso, o monitor além de revelar sua identidade, revelará além de seus vizinhos, a identidade de todos os vizinhos dos vizinhos do monitor e todas as arestas entre eles. Chamaremos este caso de *monitor revela vértices até distância 2*

Para que o processo de amostragem através de monitores seja melhor entendido, apresentaremos exemplos práticos utilizando a seguinte rede.

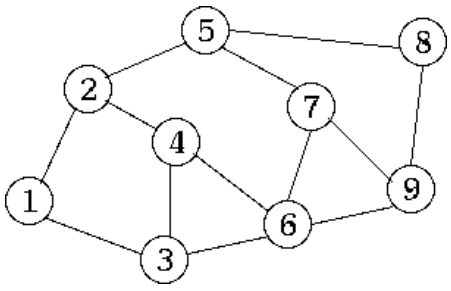


Figura 3.1: Exemplo de uma rede a ser descoberta.

Considere a rede ilustrada na figura 3.1. A descoberta de vértices e arestas tem comportamento diferente para os dois tipos de amostragem citados acima.

A figura 3.2(a) ilustra a informação revelada para o caso onde o *monitor revela vértices a distância 1* quando o vértice 9 é escolhido como monitor. Note que os vértices 6, 7 e 8 são revelados além do monitor e das arestas que ligam esses vértices ao vértice monitor, totalizando 4 vértices e 3 arestas descobertas. É importante ressaltar que para esse metodo de amostragem as arestas entre os vizinhos revelados não são reveladas, pois consideramos que o vértice monitor não possui essa informação. Um exemplo, é a aresta (6,7) que permanece desconhecida.

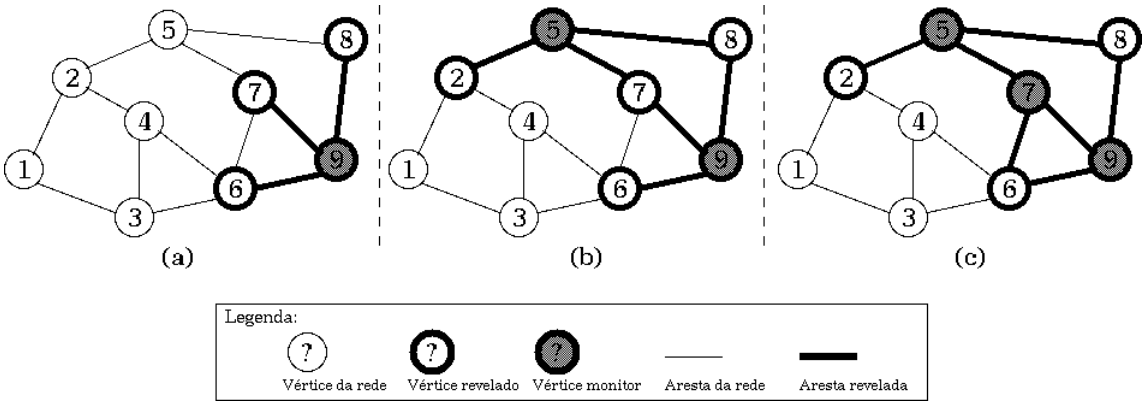


Figura 3.2: Exemplo de rede revelada no caso onde o *monitor revela vértices a distância 1.*

O processo de amostragem considerado permite escolher mais de um vértice da rede para assumir o papel de monitor e assim revelar informações locais aos mesmos. Desta forma, um parâmetro fundamental do processo é o número de amostras de monitores, que denotamos, por k. Note que k não é o número de monitores distintos, mas sim o número de amostras de monitores. Isso porque, assumimos que não controlamos o processo de amostragem de monitores (que é aleatório) de forma que um mesmo vértice pode ser amostrado mais de uma vez para ser monitor. As informações reveladas sobre vértices e arestas, coletadas pelos monitores, serão agregadas para que possamos estimar características estruturais da rede que está sendo descoberta.

A figura 3.2(b) ilustra o caso com $k = 2$ monitores (vértices 9 e 5). Supondo que o vértice 5 tenha sido amostrado após o vértice 9 podemos notar que, por possuir vizinhos em comum com o vértice 9, o mesmo só revela um vértice e três arestas novas. Com isso, podemos facilmente notar que nem toda nova amostra de monitor resultará necessariamente em mais informações sobre a rede. Eventualmente duas amostras de monitores podem ter vizinhos em comum ou ainda serem o mesmo vértice. Considere, por exemplo, a figura 3.2(c) que ilustra um exemplo com $k = 3$ amostras de monitores (vértices 9, 5 e 7 nessa ordem). Note que o monitor 7 praticamente não acrescenta nova informação a não ser pela aresta (6,7). Apesar disso, em redes muito grandes os monitores não devem colidir na informação que revelam quando seu número for baixo.

Agora vamos analisar o caso onde o mesmo vértice 9 é um *monitor que revela vértices até distância 2.* Notamos, obviamente, que mais informação é descoberta com cada monitor. A figura 3.3(a) nos permite observar que além dos vértices 6, 7 e 8 descobertos para o caso anterior, também são revelados seus vizinhos e as arestas que os ligam aos mesmos, totalizando 7 vértices e 8 arestas descobertas.

Nesse caso podemos observar que a aresta (6,7), que não seria conhecida para o

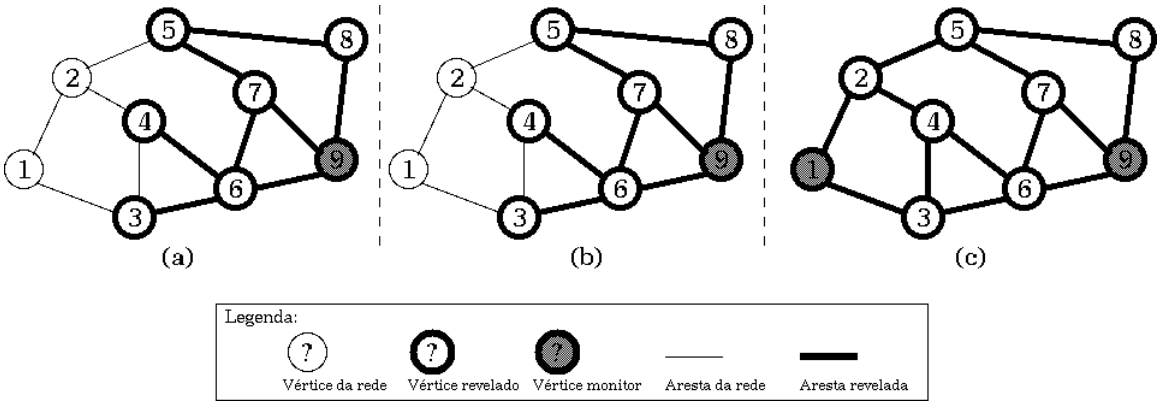


Figura 3.3: Exemplo de rede revelada no caso onde o *monitor revela vértices até distância 2*.

caso anterior (figura 3.2(a)), agora foi descoberta. Isso ocorre pois o vértice 6 além de estar a uma distância 1, também está a uma distância 2 do vértice 9 passando pelo vértice 7. Mas repare que arestas entre vértices que estão a distância 2 não são revelados. Por exemplo, neste caso (3,4) não é revelada.

A figura 3.3(b) apresenta o caso que discutimos anteriormente onde o mesmo vértice é amostrado e, portanto, nenhuma informação nova é descoberta. A figura 3.3(c) apresenta o caso onde o terceiro vértice escolhido como monitor é o vértice 1. Note que a informação revelada por este monitor, em conjunto com o monitor 9, é suficiente para revelar toda a rede.

3.2 Medidas de interesse

O objetivo principal deste trabalho é caracterizar como o processo de amostragem de vértices através de monitores revela informações da rede que está sendo descoberta. Em particular, temos interesse 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 de grafo aleatório, Y_k e W_k são variáveis aleatórias. Desta forma, estamos interessados em caracterizar seus respectivos valores esperados, ou seja, $E[Y_k]$ e $E[W_k]$, que representa o numero médio de vértices e arestas revelados.

Para exemplificar essas váriaveis, vamos utilizar as figuras 3.2 e 3.3 e determinar os valores de Y_k e W_k para $k = 1$, $k = 2$ e $k = 3$ amostras de monitores.

A tabela 3.1 apresenta esses valores.

Como podemos observar, os dois processos de amostragem descobrem a rede de

	Vértices descobertos			Arestas descobertas		
	Y_1	Y_2	Y_3	W_1	W_2	W_3
Revela a distância 1 (figura 3.2)	4	6	6	3	6	7
Revela até distância 2 (figura 3.3)	7	7	9	8	8	13

Tabela 3.1: Número de vértices e arestas descobertas por amostras de monitores.

maneira diferente. É também interesse nosso compreender como este processo de amostragem varia de acordo com a estrutura da rede. Ou seja, entender questionamentos como:

- Quais são as características estruturais da rede que mais influenciam este processo de descoberta de informação?
- Qual é a diferença entre o primeiro e segundo tipo de monitor quando aplicadas a diferentes estruturas de rede?

Como veremos nos próximos capítulos, 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 e a clusterização podem 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.

Capítulo 4

Avaliação teórica

Entender como um processo de amostragem se comporta e suas particularidades é fundamental para que seja possível amostrar uma rede de maneira eficiente. Como o processo de amostragem abordade neste trabalho permite repetição de monitores, iniciaremos este capítulo fazendo um estudo do número de monitores distintos em uma determinada iteração. Em seguida iremos obter analiticamente o valor esperado de vértices e arestas descobertas da rede em todos os casos definidos no capítulo 3.

4.1 Dedução do número de monitores distintos

Seja $G = (V, E)$ um grafo não direcionado onde cada vértice $u \in V$ é identificado por um único inteiro do conjunto $\{1, 2, 3, ..., |V|\}$. Seja $|V| = n$ o número de vértices do grafo.

Um monitor é um vértice $v \in V$ escolhido uniformemente do conjunto de vértices V, com reposição, de forma que um mesmo vértice pode ser escolhido como monitor mais de uma vez. Com isso, as escolhas dos monitores são independentes e identicamente distribuídas.

Apesar da escolha com repetição, sortear o mesmo vértice duas vezes praticamente não irá ocorrer quando o número de vértices do grafo é muito grande. De qualquer forma, quando o número de amostras de monitores aumenta, a chance de termos repetição de monitores também aumenta.

Para entender melhor o processo de repetição de monitores, definiremos C_k como a váriável aleatória que representa o conjunto de monitores distintos após k amostras de monitores e N_k como a cardinalidade deste conjunto. Ou seja, $N_k = |C_k|$.

Podemos obter a distribuição de N_k de forma recursiva. Considere que após $k - 1$ amostras possuímos $N_{k-1} = D$ monitores distintos. Temos então somente duas possibilidades para a etapa k. A nova amostra pode sortear um vértice que já pertence ao conjunto C_k (continuando com D monitores distintos) ou pode sortear um vértice que ainda não pertence ao conjunto C_k (aumentando em um o número

de monitores). Esses dois casos podem ser representados, respectivamente, pelas equações de probabilidade condicional 4.1 e 4.2.

$$P[N_k = D|N_{k-1} = D] = \frac{D}{n} \tag{4.1}$$

$$P[N_k = D + 1|N_{k-1} = D] = \frac{n - D}{n} \tag{4.2}$$

Utilizando esta distribuição condicional, podemos então calcular o valor esperado de N_k dado $N_{k-1} = D$:

$$E[N_k|N_{k-1} = D] = \frac{D}{n}D + \frac{n - D}{n}(D + 1) = \frac{n - 1}{n}D + 1 \tag{4.3}$$

Com esse resultado, podemos utilizar a propriedade da esperança exibida na equação 4.4 e obter uma equação geral para o valor esperado de N_k a partir de 4.3:

$$E[N_k] = E[E[N_k|N_{k-1} = D]] \tag{4.4}$$

$$E[N_k] = \frac{n - 1}{n}E[N_{k-1}] + 1 \tag{4.5}$$

Com a equação 4.5 podemos agora desenvolver a recursão e verificar que o número esperado de monitores N_k se trata de uma soma de prograssão geométrica com razão $q = \frac{n-1}{n}$ onde $a_1 = 1$.

- $E[N_1] = 1$
- $E[N_2] = \frac{n-1}{n}E[N_1] + 1 = \frac{n-1}{n} + 1$
- $E[N_3] = \frac{n-1}{n}E[N_2] + 1 = (\frac{n-1}{n})^2 + \frac{n-1}{n} + 1$
- $E[N_4] = \frac{n-1}{n}E[N_3] + 1 = (\frac{n-1}{n})^3 + (\frac{n-1}{n})^2 + \frac{n-1}{n} + 1$
- $E[N_k] = (\frac{n-1}{n})^{k-1} + (\frac{n-1}{n})^{k-2} + ... + \frac{n-1}{n} + 1$

$$E[N_k] = \sum_{j=1}^{k} (\frac{n - 1}{n})^{j-1} \tag{4.6}$$

Resolvendo este somatório, podemos notar que o número de monitores distintos depende do número de vértices n e do número de amostras de monitor k, e temos o seguinte resultado:

$$E[N_k] = n(1 - (\frac{n - 1}{n})^k) \tag{4.7}$$

Obviamente, para um número muito grande de amostras de monitores, obtemos:

$$\lim_{k\to\infty} E[N_k] = n \tag{4.8}$$

Observar o resultado obtido na equação 4.7 graficamente não é muito simples, pois o número de monitores únicos varia de acordo com o tamanho do grafo e isso torna difícil comparar redes de tamalhos diferentes. Para ficar mais claro, na figura 4.1 apresentamos esse resultado de uma outra forma. Dividimos a equação 4.7 por n para obter a fração de monitores únicos da rede no eixo das abscissas.

Em seguida chamaremos de $\alpha = k/n$ a fração de monitores amostrados em relação a n e o apresentaremos no eixo das coordenadas. Observe que α pode ser maior do que 1 pois k pode ser maior que n.

Feito isso, temos que a fração de monitores únicos e variamos α. Essa equação nós denoratemos por L_α.

$$L_\alpha = \frac{E[N_k]}{n}, k = \alpha \times n \tag{4.9}$$

Simplificando, temos a equação abaixo:

$$L_\alpha = 1 - (\frac{n-1}{n})^{n\alpha} \tag{4.10}$$

e podemos avaliar esse resultado para n grande, e aproximar pela equação 4.11.

$$\lim_{n\to\infty} L_\alpha \approx 1 - e^{-\alpha} \tag{4.11}$$

A figura 4.1 mostra a fração de monitores únicos em relação ao número de amostras normalizado (α). Nele apresentamos a equação 4.10 para diferentes valores de n e comparamos esses resultados com a equação 4.11 e com a equação $y = \alpha$, caso onde só seriam descobertos monitores únicos (sem repetição).

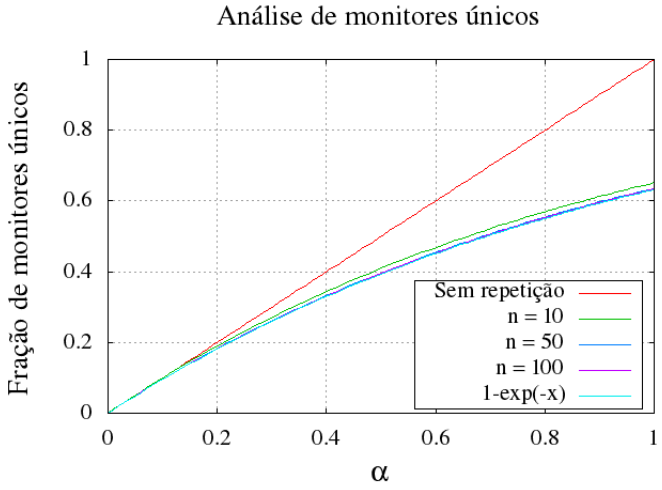


Figura 4.1: Fração de monitores únicos em relação a $\alpha = \frac{k}{n}$.

Observe que para $\alpha <= 0.2$ o caso com repetição tem o mesmo comportamento que o caso sem repetição, independente do valor de n. Ou seja, quando a fração de

monitores em relação a n é baixa, o caso sem repetição praticamente não difere do caso com repetição.

4.2 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 no capítulo 3. 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.

4.2.1 Monitor revela seus vizinhos

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 até 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 4.2. 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:

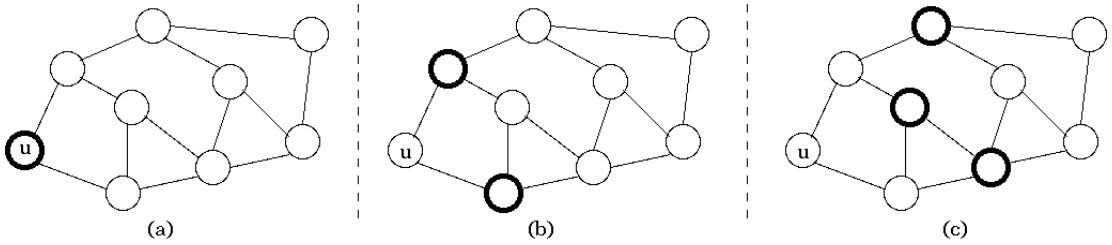


Figura 4.2: Exemplo das três formas que um vértice u da rede pode ser descoberto por uma amostra de monitor: (a) o próprio vértice u é escolhido; (b) um vizinho de u é escolhido; (c) um vizinho do vizinho de u (que não é vizinho de u) é escolhido.

- 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 mutualmente exclusivos, temos

$$P[N_u^1|Z = d] = \sum_{i=1}^{d} \frac{1}{n} = \frac{d}{n} \tag{4.12}$$

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:

$$P[D_u^1|Z = d] = \frac{1}{n} + \frac{d}{n} = \frac{1+d}{n} \tag{4.13}$$

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 monitor. 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, esta probabilidade será dada por:

$$P[\overline{D_u^k}|Z = d] = (1 - \frac{1+d}{n})^k \tag{4.14}$$

Repare que $1 - (1 + d)/n$ é a probabilidade de u não ser descoberto por uma das amostras de monitor. 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,

$$P[D_u^k] = \sum_{d=0}^{n-1} (1 - (1 - \frac{1+d}{n})^k) P[Z = d] \tag{4.15}$$

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, esta pode ser definida como:

$$Y_k = \sum_{\forall u \in V} X_{u,k} \tag{4.16}$$

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:

$$E[Y_k] = E[\sum_{\forall u \in V} X_{u,k}] = \sum_{\forall u \in V} E[X_{u,k}] = \sum_{\forall u \in V} P[D_u^k] = nP[D_u^k] \tag{4.17}$$

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 4.17 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.

4.2.2 Monitor revela seus vizinhos e vértices até 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:

$$P[N_u^2] = \frac{|N_u^2|}{n} \tag{4.18}$$

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, como pode ser visto na figura 4.3. Esta medida é conhecida como restante de grau de v (já que a aresta e já é incidente a u).

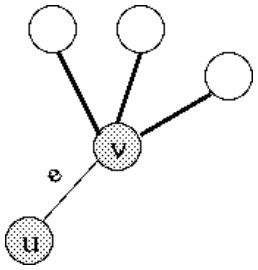


Figura 4.3: Exemplo onde as arestas, exceto a aresta e, são contadas para obter o restante de grau de v (vizinho de u).

Seja R a variável aleatória que representa o restante de grau de um vértice v incidente a u. Repare que R é diferente de Z pois sabemos que v tem grau ao menos um. O valor esperado de restante de grau de um vértice ao final de um aresta qualquer da rede foi obtido por Newman em [7], e é dado por:

$$E[R] = \frac{E[Z^2] - E[Z]}{E[Z]} \tag{4.19}$$

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

$$|N_u^2|Z = d| \approx dE[R] \tag{4.20}$$

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 duas vezes. Este efeito de "triângulos" 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, obtido através da probabilidade de dois vizinhos de um vértice serem vizinhos. Utilizaremos esse valor para calcular o número de arestas entre os vizinhos de u. Essa é uma aproximação pois não temos a clusterização analítica condicionando no grau de u.

Seja A_d o número de arestas entre os vizinhos do vértice u dado que seu grau é igual a d. O valor esperado do número de arestas entre os vizinhos do vértice u da seguinte forma:

$$E[A_d] \approx \overline{c}\binom{d}{2} = \frac{\overline{c}d(d-1)}{2}, d > 1 \tag{4.21}$$

onde utilizamos $\overline{c}\binom{d}{2}$ para calcular a probabilidade de dois vizinhos de u serem vizinhos para cada par de vizinhos. 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 4.20. Assim temos:

$$|N_u^2|Z = d| \approx dE[R] - 2E[A_d] \approx \frac{d(E[Z^2] - E[Z])}{E[Z]} - \overline{c}d(d-1) \tag{4.22}$$

Com isso, podemos aproximar a probabilidade de um vértice a distância 2 de u ser escolhido como monitor, usando a aproximação acima na equação 4.18.

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:

$$P[D_u^1|Z = d] = \frac{1+d}{n} + \frac{d(E[Z^2] - E[Z])}{nE[Z]} - \frac{\overline{c}d(d-1)}{n} \tag{4.23}$$

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 4.15.

$$P[D_u^k] = \sum_{d=0}^{n-1}(\frac{1+d}{n} + \frac{d(E[Z^2] - E[Z])}{nE[Z]} - \frac{\overline{c}d(d-1)}{n})P[Z = d] \tag{4.24}$$

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 4.17. Desta forma, temos:

$$E[Y_k] = nP[D_u^k] = \sum_{d=0}^{n-1}(1 + d + \frac{d(E[Z^2] - E[Z])}{E[Z]} - \overline{c}d(d-1))P[Z = d] \tag{4.25}$$

4.3 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 no capítulo 3. 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.

4.3.1 Monitor revela seus vizinhos

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 até 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 4.4 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.

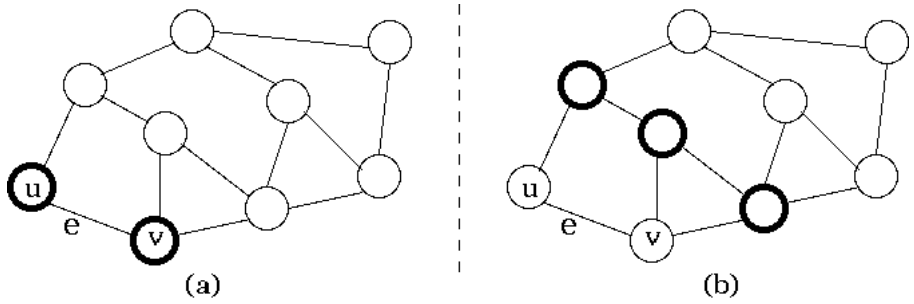


Figura 4.4: Exemplo das duas formas com a qual uma aresta $e = (u, v)$ pode ser descoberta por uma amostra de monitor: (a) um dos vértices incidentes a aresta é escolhido; (b) um vizinho de um dos vértices incidentes a aresta (que não é incidente a aresta) é escolhido.

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:

$$P[D_e^k] = 1 - (1 - 2/n)^k \tag{4.26}$$

Seja $Q_{e,k}$ uma variável aleatória indicadora que denota se a aresta e foi revelada (assumindo valor 1) ou não (assumindo valor 0) depois de k amostras de monitores. Lembrando que W_k representa o número de arestas descobertas depois de k amostras de monitores, esta pode ser definida como:

$$W_k = \sum_{e \in E} Q_{e,k} \tag{4.27}$$

Onde o conjunto de arestas da rede é representado por E. Com isso:

$$E[W_k] = \sum_{e \in E} E[Q_{e,k}] = \sum_{e \in E} P[D_e^k] \tag{4.28}$$

Isso se deve ao fato de $Q_{e,k}$ ser 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 $E[W_k|M = m] = mP[D_e^k]$, pois as arestas são estatisticamente equivalentes. E finalmente, temos que:

$$E[W_k] = E[mP[D_e^k]] = E[m]P[D_e^k] \tag{4.29}$$

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 vértices está relacionado com o valor esperado do número de arestas. Assim sendo, temos finalmente que:

$$E[W_k] = \frac{nE[Z]}{2}(1 - (1 - \frac{2}{n})^k) \tag{4.30}$$

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, $E[Z]$.

4.3.2 Monitor revela seus vizinhos e vértices até 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 4.4(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:

$$P[D_e^1] = \frac{|N_{uv}^{01}|}{n} \tag{4.31}$$

Considere o vértice u incidente a aresta e como pode ser visto na figura 4.5. 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.

Estamos condicionando no grau do vértice u porque além de simplificar a resolução do problema não teríamos a distribuição de grau conjunta de u e v.

Um fato importante que devemos levar em conta é que um vizinho de v também

pode ser vizinho de u (vértice w na figura 4.5) e ser contado duas vezes, 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.

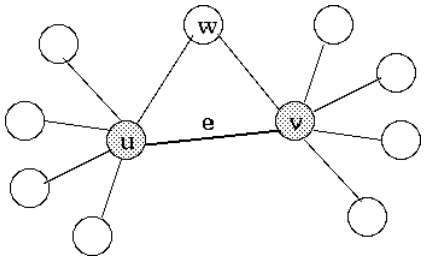


Figura 4.5: Exemplo de vértice vizinho de u e v (vértice w) que pode ser contabilizado mais de uma vez ao se considerar o restante de grau de v.

Seja $\overline{c}$ 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 4.21. 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 (por exemplo, vértice w na figura 4.5) 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:

$$|N_{uv,d}^{01}| \approx 1 + d + E[R] - \frac{2E[A_d]}{d} \tag{4.32}$$

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:

$$P[D_e^1|Z = d] \approx \frac{d}{n} + \frac{E[Z^2]}{nE[Z]} - \frac{\overline{c}(d-1)}{n} \tag{4.33}$$

Lembrando que como visto na seção 4.2.2, $E[R] = E[Z^2]/E[Z] - 1$ 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:

$$P[D_e^k|Z = d] = 1 - (1 - d/n - \frac{E[Z^2]}{nE[Z]} + \frac{\overline{c}(d-1)}{n})^k \tag{4.34}$$

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

$$P[D_e^k] = \sum_{d=0}^{n-1}(1 - (1 - d/n - \frac{E[Z^2]}{nE[Z]} + \frac{\overline{c}(d-1)}{n})^k)P[Z = d] \tag{4.35}$$

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 4.29 e simplificando para $E[M]$, de forma que temos:

$$E[W_k] = \frac{nE[Z]}{2} P[D_e^k] \tag{4.36}$$

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. Em todo caso, no capítulo 5 iremos avaliar estas aproximações comparando-as com resultados obtidos através de simulação.

Capítulo 5

Avaliação Numérica

Neste capítulo apresentaremos o simulador que foi desenvolvido para obter resultados empíricos utilizando os quatro tipos de redes citadas no capítulo 2. Serão apresentados também os cenários específicos que foram avaliados e faremos uma comparação entre os resultados numéricos e os resultados previstos pelo modelo analítico.

5.1 Simulador

Para este trabalho, implementamos um simulador do processo de amostragem de vértices e arestas descrito no capítulo 3, reproduzindo fielmente seu comportamento.

A cada amostragem de monitor, armazenamos o número de vértices e arestas descobertos para comparar com o modelo analítico. As iterações do simulador são bem definidas. A cada iteração um vértice é escolhido como monitor, seus vizinhos identificados, assim como as arestas entre estes. No caso de *vértices até distância* 2, os vizinhos dos vizinhos também são identificados assim como arestas que levam até eles.

Todo esse processo de iterações pode ser melhor entendido através do algoritmo 1. O primeiro passo é sortear aleatoriamente um monitor, como mostra a linha 1. Em seguida, se for o caso, adicionamos esse novo vértice ao conjunto de vértices conhecidos (linhas 2 a 4). Nas linhas 5 à 24 varremos a lista de vizinhos do vértice monitor e adicionamos os vértices e as arestas encontrados no conjuntos de vértices e arestas conhecidas. Repare que nas linhas 13 à 22, fazemos o mesmo procedimento

para os vizinhos do vizinho do monitor se for o interesse do estudo.

Algoritmo 1: Iteração de amostragem de monitor e descoberta de vértices e arestas.

```
1:  M ← sorteiaNovoMonitor() //Uniforme e com repetição
2:  if M ∉ VerticesConhecidos then
3:     VerticesConhecidos ∪ {M}
4:  end if
5:  N ← pegaListaDeVizinhos(M)
6:  for n ∈ N do
7:     if n ∉ VerticesConhecidos then
8:        VerticesConhecidos ∪ {n}
9:     end if
10:    if (M, n) ∉ ArestasConhecidas then
11:       ArestasConhecidas ∪ {(M, n)}
12:    end if
13:    if DescobreAteDistancia2 then
14:       K ← pegaListaDeVizinhos(n)
15:       for k ∈ K do
16:          if k ∉ VerticesConhecidos then
17:             VerticesConhecidos ∪ {k}
18:          end if
19:          if (n, k) ∉ ArestasConhecidas then
20:             ArestasConhecidas ∪ {(n, k)}
21:          end if
22:       end for
23:    end if
24: end for
```

Este é o ciclo fundamental do processo de amostragem, chamado pelo ciclo principal do simulador.

O algoritmo 2 apresenta o ciclo principal do simulador onde redes diferentes são geradas pelo modelo de rede (linhas 2 e 3) . Para cada rede são realizadas várias rodadas de simulação conforme a linha 5, onde realizamos o ciclo de amostras de monitores, conforme linha 9.

Dentro do ciclo de amostras de monitores o algoritmo 1 é chamado na linha 10 e armazenamos o número de vértices conhecidos e arestas conhecidas após aquela iteração nas linhas 11 e 12 através dos vetores "VConhecidos" para vértices e "ACo-

nhecidas" para arestas.

Algoritmo 2: Iteração de chamada do algoritmo 1 e armazenamento dos resultados.

```
1: i ← 0
2: while i < NumeroDeRedes do
3:    G ← GeraNovaRede()
4:    j ← 0
5:    while j < NumeroDeSimulacoes do
6:       k ← 0
7:       VerticesConhecidos = {}
8:       ArestasConhecidas = {}
9:       while k < NumeroMaximoDeAmostras do
10:         Algoritmo1()
11:         VConhecidos[i][j][k] = |VerticesConhecidos|
12:         AConhecidas[i][j][k] = |ArestasConhecidas|
13:         k = k + 1
14:      end while
15:      j = j + 1
16:   end while
17:   i = i + 1
18: end while
```

5.2 Cenários avaliados

Para auxiliar no processo de desenvolvimento, utilizamos a biblioteca igraph [13] para gerar os modelos de redes listados no capítulo 2: modelo de Erdős-Rényi, conhecido por $G(n, p)$ (2.1.1), modelo de Watts e Strogatz, conhecido como modelo *small world* (SW) (2.1.2), modelo de Barabási e Albert, conhecido como modelo de *preferential attachment* (BA) (2.1.3) e configuration model(CM) (2.1.4).

Parametrizamos os modelos analíticos propostos com as respectivas características das redes escolhidos. Em particular, para os modelos SW, BA e G(n,p), utilizamos a distribuição de grau, seu valor esperado, segundo momento, e a clusterização induzida pelo modelo.

O único caso que tratamos de forma diferente é o CM. Para este modelo utilizamos como entrada de cada rede o mesmo conjunto de graus dos vértices originados através do modelo SW. Com isso, nosso objetivo foi preservar para o modelo CM a mesma distribuição de grau do modelo SW e somente variar a estrutura da rede, como a clusterização.

Todas as avaliações consideram uma rede com 30000 vértices e três diferentes

graus médios: 4, 8 e 16. Para cada cenário, 30 redes diferentes são geradas pelo modelo de rede, e para cada rede são realizadas 100 rodadas de simulação, onde realizamos 15000 amostras de monitores.

Ao final, calculamos a média amostral utilizando os vetores de armazenamento de número de vértices conhecidos e de número de arestas conhecidas indicados no algoritmo 2 linhas 11 e 12, respectivamente. Com isso, obtemos o número médio de vértices conhecidos após k amostras ($\overline{Y_k}$) e o número médio de arestas conhecidas após k amostras ($\overline{W_k}$), conforme as equações a seguir:

$$\overline{Y_k} = \frac{\sum_{i=1}^{30} \sum_{j=1}^{100} VConhecidos[i][j][k]}{30 * 100} \tag{5.1}$$

$$\overline{W_k} = \frac{\sum_{i=1}^{30} \sum_{j=1}^{100} AConhecidas[i][j][k]}{30 * 100} \tag{5.2}$$

Repare que esse cálculo é feito para todos as 15000 possíveis valores de k. Além da média amostral do número de vértices e arestas descobertos em cada cenário, calculamos ainda um intervalo de confiança de 95%

Dezesseis casos foram abordados, particularmente:

- **Descobrimento de vértices a distância** 1
  - Análise de descobrimento de **vértices**
    - Modelo de Erdős-Rényi - G(n,p)
    - Modelo de Watts-Strogatz - SW
    - Modelo de Barabási–Albert - BA
    - Configuration Model - CM
  - Análise de descobrimento de **arestas**
    - Modelo de Erdős-Rényi - G(n,p)
    - Modelo de Watts-Strogatz - SW
    - Modelo de Barabási–Albert - BA
    - Configuration Model - CM
- **Descobrimento de vértices a distância** 2
  - Análise de descobrimento de **vértices**
    - Modelo de Erdős-Rényi - G(n,p)
    - Modelo de Watts-Strogatz - SW
    - Modelo de Barabási–Albert - BA
    - Configuration Model - CM

– Análise de descobrimento de **arestas**

  * Modelo de Erdős-Rényi - G(n,p)
  * Modelo de Watts-Strogatz - SW
  * Modelo de Barabási–Albert - BA
  * Configuration Model - CM

Para melhor comparação dos resultados obtidos, apresentaremos o eixo das abscissas como número de amostras de monitores normalizado, ou seja, dividindo o número de amostras pelo número de vértices da rede ($\alpha = \frac{k}{n}$) e o eixo das ordenadas como fração de vértices ou de arestas descobertas com relação ao número de vértices e arestas da rede. Dessa forma nosso objetivo é analisar a relação entre a fração de amostras de monitores e a fração de vértices ou arestas descobertos. As curvas em cada gráfico indicam os resultados obtidos pelo modelo (M) ou simulação (S) para os diferentes graus médios.

5.3 Descoberta de vértices

5.3.1 Monitores descobrem vértices a distância 1

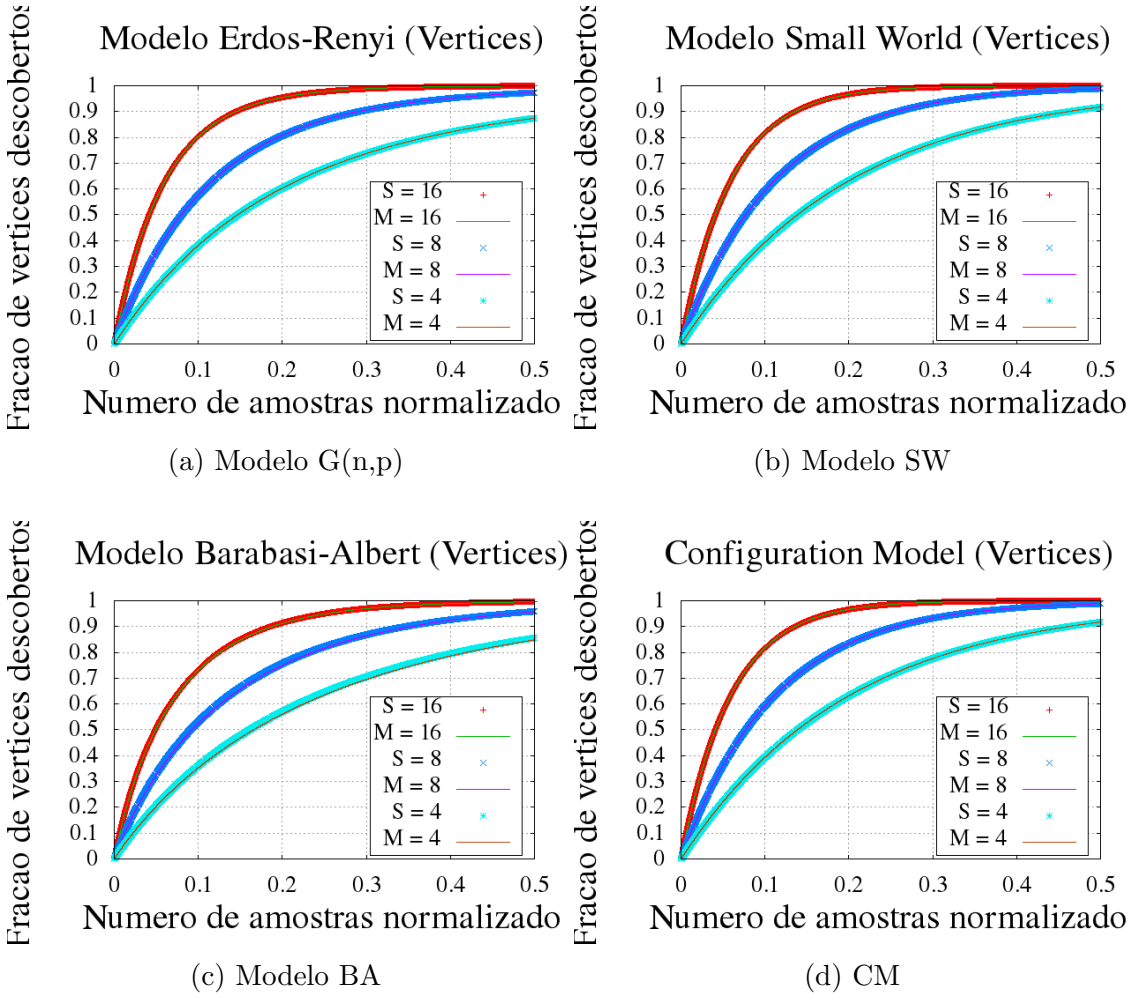
A Figura 5.1 apresenta os resultados de descoberta de vértices para os quatro modelos de redes utilizados quando um monitor revela vértices a distância 1. Em todo os casos o resultado do modelo analítico proposto é idêntico ao resultado obtido por simulação. Isso ocorre pois a equação 4.17 representa exatamente o valor esperado do processo de descoberta de vértices.

Note que quanto maior o grau médio, mais rápido o processo de amostragem descobre os vértices da rede em todas as redes investigadas. Em particular, quando o grau médio é 16, mais de 70% dos vértices são descobertos com apenas 10% da rede em amostras para todos os quatro casos. Isso nos sugere que o processo de amostragem tem resultados muito eficientes para redes muito densas (com grau médio alto).

Além disso, quando o grau médio é igual a 4, mesmo quando temos a metade da rede em amostras, não atingimos 100% de vértices descobertos pois o grau médio é baixo e com isso cada monitor não nos fornece informação suficiente. Repare que quando dizemos metade da rede em amostras, não significa que metade de vértices da rede estão sendo utilizados como monitores, e sim que o número de vezes que amostramos é igual a $\frac{n}{2}$.

Apesar dessa similaridade entre os casos, é importante destacar que as curvas de descobrimento são diferentes para os quatro tipos de redes por conta das diferentes distribuições de grau. Por exemplo, no caso G(n,p) (figura 5.1a), para grau médio

8 com 20% dos vértices em amostras temos aproximadamente 80% dos vértices descobertos. Já para o caso BA (figura 5.1c), esse valor é menor, representando aproximadamente 77% dos vértices descobertos.



(a) Modelo G(n,p)

(b) Modelo SW

(c) Modelo BA

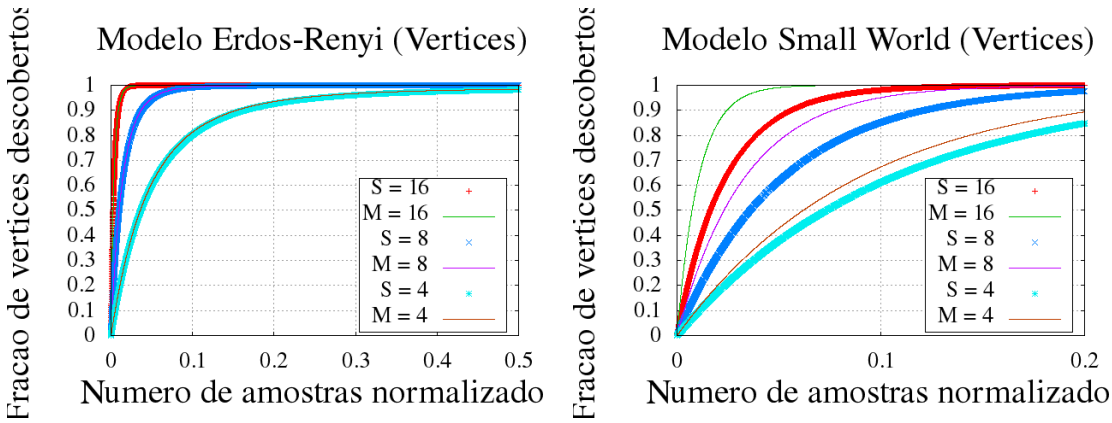
(d) CM

Figura 5.1: Descoberta de vértices quando monitor revela vértices a distância 1

5.3.2 Monitores descobrem vértices até distância 2

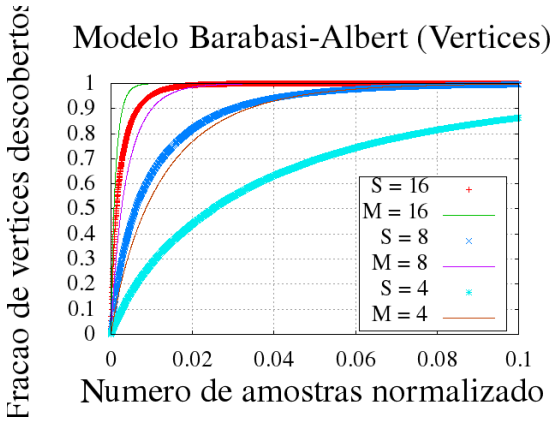
A figura 5.2 apresenta os resultados quando um monitor revela vértices até distância 2. Podemos verificar que o modelo analítico aproximado apresenta um resultado idêntico ao de simulação para o modelo $G(n, p)$ (Figura 5.2a). Isso ocorre pois neste caso a rede é construída conectando os vértices de forma independente, tornando a aproximação que fazemos para o valor esperado de restante de grau de um vizinho ($E[R]$) adequada para este tipo de rede. Entretanto, não vemos a mesma acurácia para as redes SW e BA (figuras 5.2b e 5.2c, respectivamente) e a aproximação para o valor esperado do restante de grau de um vizinho não é muito boa, influenciando negativamente os resultados do nosso modelo.

A Tabela 5.1 compara os valores da clusterização média utilizada por nosso

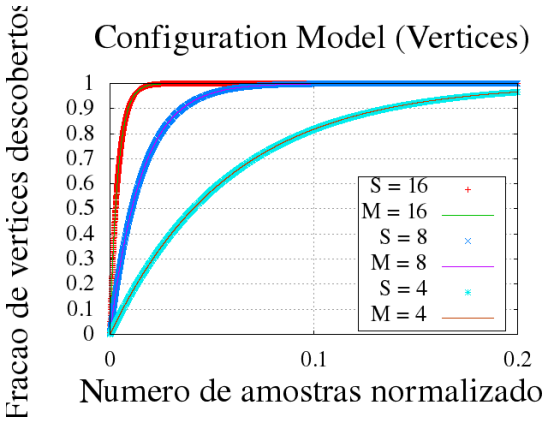


(a) Modelo G(n,p)

(b) Modelo SW



(c) Modelo BA



(d) CM

Figura 5.2: Descoberta de vértices quando monitor revela vértices até distância 2

modelo com resultados de simulação para diferentes redes e graus médios. Esse valor é fundamental pois a clusterização média analítica é parâmetro de entrada para a Equação 4.22, que estima o número de vizinhos a distância 2 de um vértice. Tal função, por sua vez, é utilizada para calcular a Equação 4.25, valor esperado do número de vértices descobertos.

	Clusterização Média ($\overline{c}$)					
	Analítico			Simulação		
E[Z]	4	8	16	4	8	16
BA	$1,8 \times 10^{-3}$	$2,3 \times 10^{-3}$	$3,99 \times 10^{-3}$	$0,82 \times 10^{-3}$	$1,88 \times 10^{-3}$	$3,92 \times 10^{-3}$
G(n,p)	$1,33 \times 10^{-4}$	$2,66 \times 10^{-4}$	$5,33 \times 10^{-4}$	$1,30 \times 10^{-4}$	$2,59 \times 10^{-4}$	$5,31 \times 10^{-4}$
SW	$4,85 \times 10^{-1}$	$6,23 \times 10^{-1}$	$6,79 \times 10^{-1}$	$4,72 \times 10^{-1}$	$6,06 \times 10^{-1}$	$6,59 \times 10^{-1}$
CM	$0,76 \times 10^{-4}$	$2,09 \times 10^{-4}$	$4,70 \times 10^{-4}$	$0,759 \times 10^{-4}$	$2,05 \times 10^{-4}$	$4,7 \times 10^{-4}$

Tabela 5.1: Comparação de clusterização obtida analiticamente (ver equações no capítulo 2) e por simulação ($n = 30000$).

Podemos verificar através desta tabela que a clusterização utilizada para os modelos SW, G(n,p) e CM é bem próxima da realidade, diferente do modelo BA, prejudicando significativamente nossos cálculos para este modelo.

O fato da estimativa analítica da clusterização não ser boa afeta diretamente o resultado para número de vértices a distância 2. A Tabela 5.2 utiliza a equação 4.22 descondicionada para caracterizar o número médio de vértices a distância 2 de um monitor, ou seja:

$$E[N_u^2] \approx \sum_{\forall d} \left(\frac{d(E[Z^2] - E[Z])}{E[Z]} - \overline{c}d(d-1)\right) P[Z = d] \tag{5.3}$$

Note que para os modelos G(n,p), SW e CM os valores analíticos são muito próximos dos valores de simulação. O único caso onde esses valores realmente são muito diferentes é o caso BA por conta da clusterização.

	Número de vértices a distância 2					
	Analítico			Simulação		
E[Z]	4	8	16	4	8	16
BA	$100,45$	$327,35$	$1123,75$	$36,05$	$166,06$	$720,56$
G(n,p)	$15,99$	$63,98$	$255,86$	$15,52$	$63,38$	255.4
SW	$6,19$	$21,09$	$77,04$	6	$21,96$	$81,02$
CM	$12,03$	$56,06$	$240,04$	12	56	240

Tabela 5.2: Comparação de número de vértices a distância 2 obtido analiticamente (equação 4.22) e por simulação ($n = 30000$).

A Figura 5.2c apresenta o gráfico de descoberta de vértices para o modelo BA. Podemos verificar que os resultados para grau médio 4 e 8 são distantes do obtido empiricamente. O modelo analítico só começa a apresentar um melhor desempenho quando o grau médio é 16. Tal fato ocorre pois a Equação 2.10 que usamos para estimar a clusterização neste modelo é mais precisa para redes mais densas, como pode ser visto em [6]. Podemos verificar essa melhora através da Tabela 5.3 que compara a clusterização média obtida analiticamente para o modelo BA segundo a equação 2.10 com o resultado empírico (simulação) conforme o grau médio da rede cresce.

De fato, para graus médios altos, a equação para clusterização é mais precisa afetando diretamente o estimador para para grau médio baixo.

	Clusterização Média ($\overline{c}$) para o modelo BA				
E[Z]	4	8	16	32	64
$\overline{c}$ empírica	$0,82 \times 10^{-3}$	$1,88 \times 10^{-3}$	$3,92 \times 10^{-3}$	$7,76 \times 10^{-3}$	$1,44 \times 10^{-2}$
$\overline{c}$ analítica	$1,80 \times 10^{-3}$	$2,33 \times 10^{-3}$	$3,99 \times 10^{-3}$	$7,49 \times 10^{-3}$	$1,455 \times 10^{-2}$

Tabela 5.3: Comparação de clusterização obtida analiticamente (ver equação 2.10) e por simulação para o modelo BA ($n = 30000$).

Repare que essa análise explica o motivo de não termos uma boa estimativa para o modelo BA, mas mantém em aberto o problema em relação ao modelo SW que abordaremos a seguir.

5.3.3 Descobrindo vértices a distância 2 e o problema dos quadrados

Quando analisamos a descoberta de vértices até distância 2 em algumas redes, podemos nos deparar com o problema do "quadrados". Quadrados são nada mais que subgrafos originados quando dois vizinhos de um vértice v possuem um vizinho w em comum além de v, como mostra a figura 5.3. A existência desse quadrado leva a um erro no estimador apresentado no capítulo 4, pois o mesmo irá contar o vértice w duas vezes, superestimando o número de vértices à distância 2 de v. O estimador irá contar o vértice w uma vez partindo de v e passando por u_1 e a uma segunda vez partindo v e passando por u_2.

No capítulo 4 resolvemos o problema dos "triângulos" (vizinhos de v também serem vizinhos) utilizando o coeficiente de clusterização, mas essa abordagem não é suficiente para resolver o problema dos quadrados apresentado agora o qual não será abordado neste trabalho.

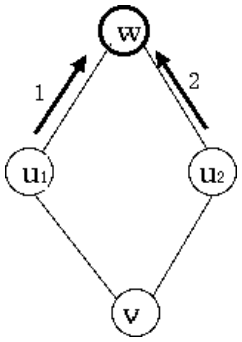
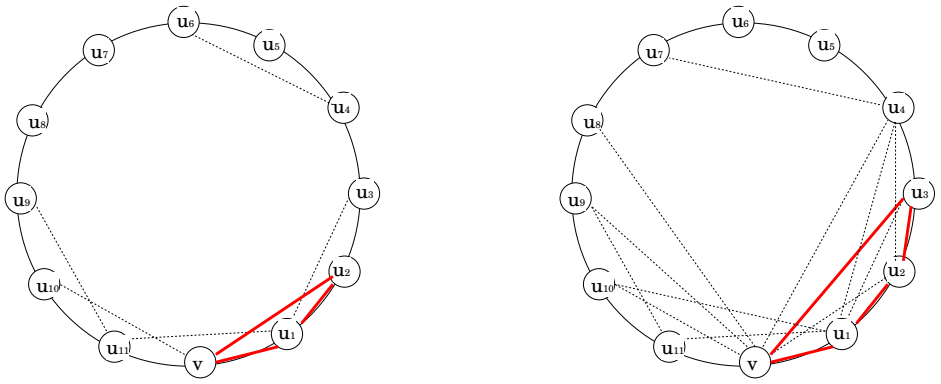


Figura 5.3: Problema do "quadrado" ao contar o número de vértices à distância 2 do vértice v. Nesse caso o vértice w é contado duas vezes.

Em todo caso, esse problema será mais evidente para modelos como o SW, onde há muitos quadrados. Mesmo com o cálculo analítico da clusterização próximo do valor obtido empiricamente, o número de quadrados existentes nessa estrutura será muito grande e com isso a acurácia do nosso estimador será prejudicada.

Visualizar a presença desse problema para o modelo SW é simples se olharmos para o algoritmo de formação dessas redes (apresentado em 2.1.2). Exemplificaremos o caso específico onde o grau médio é 4 e 8, como mostram as figuras 5.4a e 5.4b

A Figura 5.4a apresenta o início do processo de formação de uma rede SW com grau médio igual a 4. Conforme apresentamos no Capítulo 2, inicialmente temos um látice regular onde cada vértice se liga aos seus vizinhos a distância $k = 2$ no anel. Note que o vértice v se liga aos vértices u_1 e u_2 formando apenas um "triângulo", assim como se liga aos vértices u_{10} e u_{11}. Para esse caso, o processo de formação inicial do modelo não origina nenhum quadrado. Os quadrados que serão gerados



(a) Parte do processo de formação de uma rede SW com grau médio 4 ($k = 2$).

(b) Parte do processo de formação de uma rede SW com grau médio 8 ($k = 4$).

Figura 5.4: Parte de dois processos de formação de uma rede SW.

aparecerão somente na etapa aleatória do algoritmo e não serão muitos. Justamente por isso nosso modelo se aproxima melhor quando o grau médio é 4, como podemos ver na Figura 5.2b (comparar S=4 com M=4).

O mesmo não acontece para a Figura 5.4b onde o grau médio será igual a 8. O processo de formação irá conectar cada vértice aos seus vizinhos a distância $k = 4$ no anel. Note que nesse caso o vértice v se ligará aos vértices u_1, u_2, u_3 e u_4 de um lado e u_8, u_9, u_{10} e u_{11} do outro. Com essas arestas teremos originado quatro "quadrados", como por exemplo o composto pelos vértices v, u_1, u_2 e u_3. Como falamos anteriormente, a aproximação do nosso modelo nesse caso não será boa e podemos constatar através da Figura 5.2b (comparar S=8 com M=8).

Conforme apresentado, no modelo CM utilizamos valores de entrada diferentes dos outros modelos. A ideia de utilizar os graus originados das redes SW é garantir que o modelo CM possua a mesma distribuição de grau do modelo SW, porém gerando uma estrutura de redes cujo processo de formação é independente.

A alteração no CM nos gera redes onde a quantidade de "quadrados" é muito menor devido à independência no processo de formação da rede. Como o valor analítico da clusterização para o modelo SW é muito próxima do empírico (ver tabela 5.1), a presença de muitos "quadrados" será o problema fundamental ao estimar o número de vértices a distância 2 nesta rede. A Figura 5.2d apresenta o resultado de descoberta de vértices para o CM.

Como esperado, o resultado do nosso modelo bate exatamente com o resultado de simulação. Esse resultado confirma a hipótese de que o problema do estimador para o modelo SW se deve à estrutura originada pelo seu processo de formação, que induz muitos quadrados, e não tem relação nenhuma com a sua distribuição de grau ou clusterização.

5.4 Descoberta de arestas

Iremos agora apresentar os resultados para descobrimento de arestas.

5.4.1 Monitores descobrem vértices a distância 1

A figura 5.5 apresenta os casos onde o monitor revela vértices a distância 1. Conforme o resultado obtido através da equação 4.30, temos que a *fração* de arestas descobertas independe da distribuição de grau e isso pode ser confirmado quando comparamos o modelo analítico com o resultado de simulação dos quatro modelos. Observe que em todos os casos o resultado obtido analiticamente é idêntico ao resultado de simulação. Um fato muito interessante que podemos perceber, é que o processo de amostragem de vértices descobre a *fração* de arestas sempre da mesma forma, independente de qualquer propriedade estrutural da rede sendo avaliada, seja grau médio, ou distribuição de grau. Como conhecemos o processo de amostragem, isso nos ajudaria por exemplo, a estimar o número de arestas em uma rede onde o grau médio é previamente conhecido, ou possa ser estimado.

Para tornar mais claro, a fração de arestas descobertas pode ser obtida dividindo a equação 4.30 pelo número de arestas do grafo, $m = \frac{nE[z]}{2}$, como mostra a equação a seguir:

$$\frac{E[W_k]}{m} = \frac{nE[Z]}{2m}(1 - (1 - \frac{2}{n})^k) = (1 - (1 - \frac{2}{n})^k) \tag{5.4}$$

Com esse resultado podemos verificar matematicamente a afirmação de que a equação de fração de arestas descobertas não depende do grau médio da rede, apenas de n e k, número de amostras.

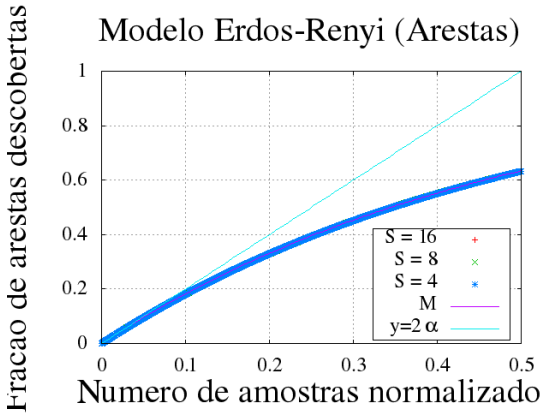
Uma outra característica do descobrimento de arestas através desse método de amostragem é que o descobrimento de arestas é bem mais lento que o de vértices. Como podemos ver em todos os casos na figura 5.5, mesmo com 10% dos vértices da rede em amostras de monitores descobrimos em média pouco menos de 20% das arestas da rede. Mesmo quando consideramos 50% dos vértices da rede em amostras de monitores, menos de 65% das arestas da rede são descobertas. Ou seja, descobrir arestas é mais difícil que vértices. Isso ocorre pois a probabilidade de uma aresta ser revelada é muito menor que a probabilidade de um vértice ser revelado.

Mesmo sendo lento, é importante ressaltar que quando temos um número reduzido de amostras de monitores esse descobrimento é próximo do caso onde cada amostra de monitor descobre exatamente o grau médio da rede (W_k^*). A fração de arestas descobertas nesse caso é representada pela reta $y = 2\alpha$ ($\alpha = \frac{k}{n}$) no gráfico, conforme a equação abaixo:

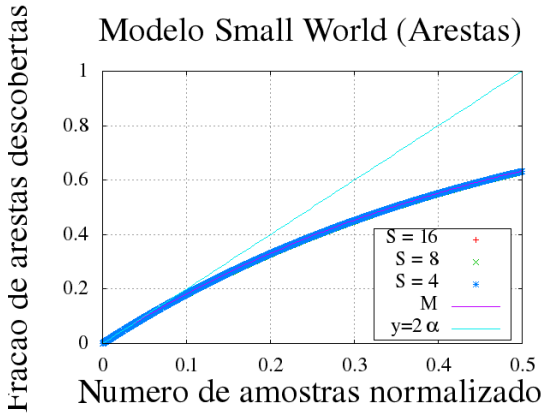
$$\frac{W_k^*}{m} = \frac{kE[Z]}{m} = \frac{k2m}{mn} = \frac{2k}{n} = 2\alpha \tag{5.5}$$

Tal fato ocorre, pois quando temos poucos monitores a probabilidade de um monitor ser amostrado mais de uma vez é muito pequena, e como levamos em conta a média amostral, praticamente não temos repetição. Isso faz com que cada monitor traga de informação aproximadamente o grau médio da rede. Conforme aumentamos o número de monitores passamos a ter mais repetições e nos distanciamos da reta de descobrimento ótimo.

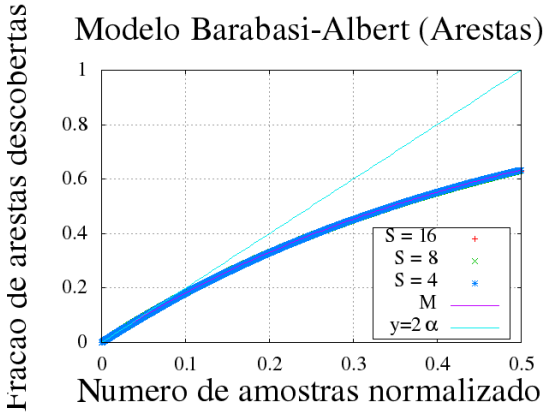
Com um número reduzido de vértices da rede em amostras (10%) o número de monitores únicos (20%) seria muito próximo do resultado com repetição, enquanto com uma grande quantidade dos vértices da rede em amostras (50%) o número de monitores únicos (100%) é muito distante do resultado com repetição (pouco mais de 60%).



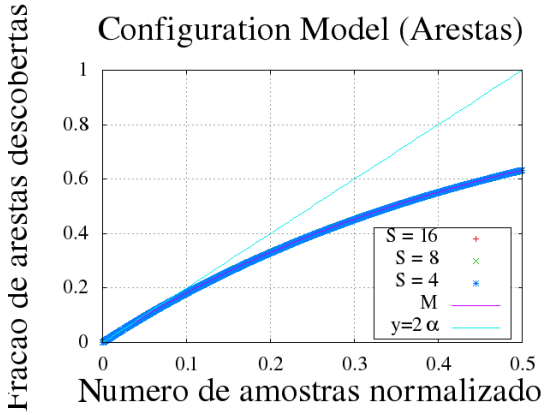
(a) Modelo G(n,p)



(b) Modelo SW



(c) Modelo BA



(d) CM

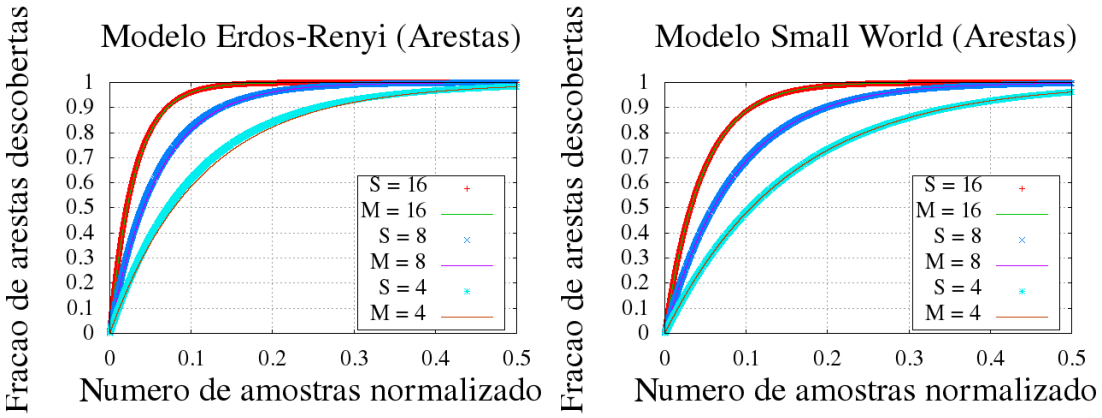
Figura 5.5: Descoberta de arestas quando monitor revela vértices a distância 1

5.4.2 Monitores descobrem vértices até distância 2

O resultado para o caso onde monitores revelam vértices até distância 2 é apresentado na Figura 5.6. Podemos ver que os resultados do modelo analítico proposto são surpreendentes para as redes $G(n, p)$, SW e CM, sendo praticamente idênticos aos obtidos por simulação para todos os graus médios analisados.

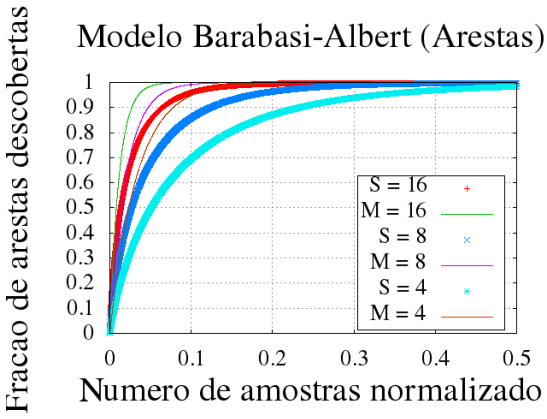
É importante notar que a existência de "quadrados" na rede não afeta os resultados para arestas mesmo quando estamos tratando descobrimento de arestas até distância 2, uma vez que a única possibilidade de contarmos uma aresta duas vezes é quando temos um triângulo, caso esse que estamos tratando no nosso modelo.

Devemos reparar que o bom desempenho não é observado para os modelos BA, pois a clusteriação analítica para este modelo não é muito precisa como apresentado anteriormente e ilustrado na tabela 5.1.

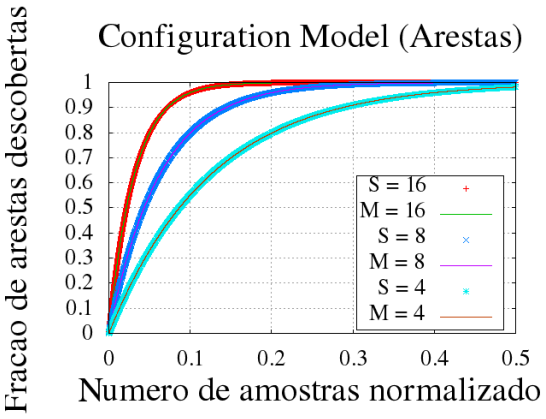


(a) Modelo G(n,p)

(b) Modelo SW



(c) BA



(d) CM

Figura 5.6: Descoberta de arestas quando monitor revela vértices até distância 2

Capítulo 6

Conclusão e trabalhos futuros

Neste trabalho estudamos um processo de amostragem de vértices baseado em monitores escolhidos aleatoriamente em uma rede. A ideia consiste em sortear monitores que tragam informações locais, ajudando assim a reunir conhecimento da rede como um todo.

Consideramos dois tipos de monitores em nosso estudo: i)os que revelam vértices a distância 1 e ii)os que revelam vértices até distância 2. Estudamos as características desses dois modelos, e apresentamos em seguida um modelo analítico para determinar o valor esperado do número de vértices e arestas descobertos por estes processos em função do número de amostras de monitores.

Desenvolvemos um simulador que representa fielmente os dois processos de amostragem e utilizamos o mesmo para avaliar o modelo analítico. Fazemos uma avaliação numérica comparando as previsões do estimador com os resultados de simulação utilizando quatro modelos de redes aleatórias com diferentes graus médios.

Nossos resultados mostram que o modelo analítico para descobrimento de vértices e arestas a distância 1 é exato. Apesar de ser exato, notamos que o mesmo depende da distribuição de grau para o descobrimento de vértices e por isso apresenta diferentes resultados para redes diferentes. Já para o caso de descobrimento de arestas os resultados não dependem de características estruturais da rede como grau médio, distribuição de grau ou clusterização.

Os resultados também mostram que o modelo proposto para o caso de distância 2 oferece boas estimativas para diversos casos tanto para o descobrimento de vértices quanto para arestas. Em particular, nos casos onde a rede é construída conectando vértices de forma independente, obtemos uma boa aproximação para o número de vértices à distância 2, o que é fundamental no nosso modelo para descoberta de vértices. Nos casos onde o resultado não é muito bom, entendemos que o problema do estimador está relacionado com o cálculo da clusterização para o modelo BA e apresentamos o problema dos "quadrados" existente para o modelo SW. Utilizamos o modelo CM para validar que o problema na acurácia para o SW está relacionado

com o problema dos "quadrados" e não com sua distribuição de grau.

O modelo descrito neste trabalho pode ser utilizado para prever o número de vértices e arestas em redes desconhecidas partindo dos valores obtidos ao se aplicar o processo de amostragem de monitores.

Por fim, parte deste trabalho resultou em uma publicação no Workshop em Desempenho de Sistemas Computacionais e de Comunicação (WPerformance) que ocorre no Congresso da Sociedade Brasileira de Computação (CSBC) onde recebeu o título de melhor artigo da conferência em 2012 [14].

6.1 Trabalhos futuros

Os modelos analisados neste trabalho podem ser utilizados para guiar diferentes pesquisas em diferente áreas. Em particular, apontamos os seguintes pontos como trabalhos futuros:

- Descobrimento de grandes redes reais utilizando os modelos apresentados neste trabalho para prever características de redes quando não temos nenhuma informação, tais como distribuição de grau ou clusterização.
- Resolver o problema dos "quadrados" para obter uma melhor estimativa para o número de vértices a distância 2 do monitor.
- Comparar o processo de amostragem proposto com outros processos de amostragem de vértices, como por exemplo o Random Walk.
- Estudar o processo quando monitores revelam vértices até distância $k > 2$. Neste trabalho avaliamos vértices que descobrem até distância $k = 1$ e $k = 2$.

Referências Bibliográficas

[1] BARABÁSI, A. L. "Scale-Free Networks: A Decade and Beyond", *Science*, v. 325, pp. 412, 2009. doi: 10.1126/science.1173299.

[2] NEWMAN, M. E. J., STROGATZ, S. H., WATTS, D. J. "Random graphs with arbitrary degree distributions and their applications", *Phys. Rev. E*, v. 64, pp. 026118, Jul 2001. doi: 10.1103/PhysRevE.64.026118.

[3] BARRAT, A., WEIGT, M. "On the properties of small-world network models", *EUROP.PHYS.J.B*, v. 13, pp. 547, 2000.

[4] ALBERT, R., BARABÁSI, A.-L. "Statistical mechanics of complex networks", *Rev. Mod. Phys.*, v. 74, pp. 47–97, Jan 2002. doi: 10.1103/RevModPhys.74.47.

[5] OLVER, F. W., LOZIER, D. W., BOISVERT, R. F., et al. *NIST Handbook of Mathematical Functions*. 1st ed. New York, NY, USA, Cambridge University Press, 2010. ISBN: 0521140633, 9780521140638.

[6] FRONCZAK, A., FRONCZAK, P., HOŁYST, J. A. "Mean-field theory for clustering coefficients in Barabási-Albert networks", *Phys. Rev. E*, v. 68, pp. 046126, Oct 2003. doi: 10.1103/PhysRevE.68.046126.

[7] NEWMAN, M. E. J. *Networks: An Introduction*. Oxford University Press, 2010.

[8] RIBEIRO, B., TOWSLEY, D. "Estimating and sampling graphs with multidimensional random walks". In: *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, IMC '10, pp. 390–403, New York, NY, USA, 2010. ACM. ISBN: 978-1-4503-0483-2. doi: 10.1145/1879141.1879192.

[9] GJOKA, M., KURANT, M., BUTTS, C., et al. "Walking in Facebook: A Case Study of Unbiased Sampling of OSNs". In: *INFOCOM, 2010 Proceedings IEEE*, pp. 1–9, 2010. doi: 10.1109/INFCOM.2010.5462078.

[10] KURANT, M., MARKOPOULOU, A., THIRAN, P. "Towards Unbiased BFS Sampling", *Selected Areas in Communications, IEEE Journal on*, v. 29, n. 9, pp. 1799–1809, 2011. ISSN: 0733-8716. doi: 10.1109/JSAC.2011.111005.

[11] KRYCZKA, M., CUEVAS, R., GUERRERO, C., et al. "Unrevealing the structure of live BitTorrent swarms: Methodology and analysis". In: *Peer-to-Peer Computing (P2P), 2011 IEEE International Conference on*, pp. 230–239, 2011. doi: 10.1109/P2P.2011.6038741.

[12] WATTS, D. J., STROGATZ, S. H. "Collective dynamics of small-world networks", *Nature*, v. 393, n. 6684, pp. 440–442, jun 1998.

[13] CSARDI, G., NEPUSZ, T. "The igraph software package for complex network research", *InterJournal*, v. Complex Systems, pp. 1695, 2006. Disponível em: <`http://igraph.sf.net`>.

[14] FIGUEIREDO, D. R., PINHEIRO, V. D. M., ROCHA, A. A. A. "Modelagem e Caracterização de um Processo de Amostragem de Vértices em Redes". In: *In: XI Workshop em Desempenho de Sistemas Computacionais e de Comunicação (WPerformance), XXXII Congresso da Sociedade Brasileira de Computação (SBC)*, 2012.