

# Relatório Técnico 1

Rodrigo R. G. e Souza

April 23, 2009

## 1 Redes de Dependências entre Componentes

Sistemas de *software* precisam ser modificados constantemente. Dependências excessivas entre os componentes de um sistema podem tornar as modificações mais custosas, uma vez que dificultam a compreensão dos componentes isoladamente. Por essa razão o estudo de redes formadas por dependências entre componentes de *software* pode fornecer pistas sobre a atividade de desenvolvimento de *software* em geral.

Neste trabalho consideramos como componentes as entidades de código-fonte, tais como classes, métodos e atributos (no caso de linguagens orientadas objetos), ou procedimentos, funções e tipos abstratos de dados (no caso de linguagens procedimentais). Dizemos que um componente depende de outro quando o funcionamento do primeiro está condicionado à presença do segundo. Dependências podem se originar de diversos tipos de interação, como uma chamada a uma função ou a leitura de um atributo. Dependências são essencialmente assimétricas: o fato de um componente A depender de um componente B não implica que B depende de A.

### 1.1 Extração

A rede de dependências entre componentes de um sistema pode ser extraída automaticamente por um programa construído para este fim, denominado extrator de dependências. As principais abordagens de extração de dependências são a análise dinâmica e a análise estática.

A análise dinâmica envolve a instrumentação do sistema para coletar dados de sua execução, e por isso exige que o programa sob análise esteja corretamente instalado e configurado. Para que a análise seja efetiva, é preciso que o programa seja executado com as mais diversas entradas a fim de capturar todos os comportamentos interessantes, o que normalmente é feito através da execução de casos de teste. Essa imposição restringe o conjunto de sistemas que podem ser efetivamente analisados dinamicamente, e nos demais sistemas o tempo de execução da análise pode ser proibitivo.

A análise estática é realizada sobre o código-fonte ou o código objeto do sistema, dispensando sua execução, mas tende a ser menos precisa do que a análise dinâmica ? e não detecta interações resultantes de introspecção. Apesar

disso a análise estática é mais aplicável no caso geral, pois não depende da existência de testes automáticos representativos e é, em geral, mais rápida que a dinâmica.

Nenhuma técnica automática é capaz de determinar corretamente, em todos os casos, se existe uma dependência entre dois componentes — basta considerar o uso de ponteiros em linguagens como C++, ou a transferências de dados através de arquivos. Assim, redes de dependências extraídas automaticamente devem ser consideradas apenas aproximações das redes de dependências reais.

## 1.2 Simplificação e Detalhamento

A depender do tipo de análise que se deseja realizar sobre uma rede de dependências, pode ser conveniente filtrar ou contrair componentes, ou mesmo tratar dependências como relacionamentos simétricos. A filtragem consiste em remover componentes ou dependências consideradas desnecessárias para a análise. A contração pode ser feita quando há uma relação de composição entre os componentes, como no caso de classes (que são compostas de atributos e métodos). Nesse caso uma classe e todos os atributos e métodos que a compõem são contraídos, isto é, representados como um só componente, a classe. Desse processo surgem dependências implícitas: se na rede original há uma dependência entre dois métodos, na rede contraída surge uma dependência implícita entre as classes que os contêm.

Há análises que, por outro lado, requerem informações mais detalhadas. Pode ser necessário considerar não apenas a existência de dependências, mas também identificar quais são os tipos de interações que ocorrem entre dois componentes e com que frequência essas interações ocorrem.

## 2 Redes Complexas

A teoria das redes complexas estuda propriedades gerais de diversos tipos de redes com o uso de ferramentas estatísticas. Estudos realizados na última década revelaram similaridades entre diversas redes, sejam elas tecnológicas, como a Web e a rede de distribuição de energia elétrica dos Estados Unidos, biológicas, como a cadeias alimentares e ligações entre proteínas, ou sociais, como as relações de amizade entre alunos de uma escola. Essas redes têm sido caracterizadas por observações como a distribuição de graus, o efeito mundo pequeno, o coeficiente de agrupamento e a presença de motivos.

### 2.1 Distribuição de Graus

Barabási e Albert (1999) analisaram uma amostra da *World Wide Web*, modelada como um grafo não-orientado no qual os vértices representam páginas e as arestas representam *links* entre duas páginas. Eles perceberam que a probabilidade de um vértice escolhido ao acaso ter grau  $k$  (isto é, estar ligado a  $k$  arestas) seguia uma lei de potência,  $p(k) \sim k^{-\gamma}$ , como mostra a Figura 2.1.

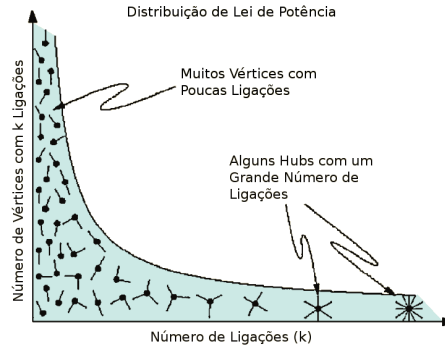


Figure 1: Distribuição de graus como lei de potência. Adaptado de Barabasi (2007).

Dizemos, nesse caso, que a rede possui uma distribuição de graus livre de escala, ou simplesmente que a rede é livre de escala. Desde então esse tipo de distribuição foi encontrado em diversas redes, incluindo redes de dependências entre componentes de programas de computador Valverde and Solé (2003).

Esse resultado foi considerado surpreendente, pois contradiz a hipótese de que redes reais obedecem ao modelo de Erdős-Rényi (1959), também chamado de modelo de rede aleatória. Nesse modelo, a probabilidade de um par de vértices ser ligado por uma aresta é constante e igual a  $p$ . Demonstra-se que a distribuição dos graus de redes geradas por esse modelo é bem aproximada pela distribuição de Poisson.

Uma característica das redes livres de escala é a presença de vértices cujo grau é muito maior do que a média (informalmente chamados de *hubs*). No caso de redes de Erdős-Rényi a probabilidade de existir um vértice com um determinado grau  $k$  cai exponencialmente à medida que  $k$  se afasta do valor médio e, por essa razão, nessas redes os vértices possuem mais ou menos o mesmo grau e a existência de *hubs* é altamente improvável.

Barabási e Albert propuseram um modelo para explicar a formação das redes com distribuição de graus livre de escala, formado por dois mecanismos: crescimento contínuo e ligação preferencial. O modelo propõe que as redes crescem um vértice por vez e cada novo vértice se liga a um número fixo de vértices antigos, dando preferência aos vértices com maior grau (mais formalmente, a probabilidade de um vértice receber uma aresta é proporcional ao seu grau). Hoje sabe-se que redes livres de escala também podem ser geradas por outros modelos Albert and Barabasi (2000); Kumar et al. (2000); Aiello et al. (2000); Dorogovtsev et al. (2002); Bollobás et al. (2003); Deo and Cami (2005).

### 2.1.1 Efeito Mundo Pequeno

A distância entre dois vértices de uma rede é o número de arestas do menor caminho que conecta os vértices. Diz-se que uma rede apresenta o efeito mundo

pequeno quando a distância entre dois vértices é, em média, pequena, mesmo quando a rede é grande. Mais formalmente, a distância média entre vértices é proporcional ao logaritmo do número de vértices Watts and Strogatz (1998). Esse efeito foi detectado em diversas redes reais.

### 2.1.2 Coeficiente de agrupamento

Os vizinhos de um vértice são todos os vértices com os quais ele compartilha uma aresta. O coeficiente de agrupamento de um vértice,  $C_i$ , é a fração de todos os possíveis pares de vizinhos do vértice que estão ligados por uma aresta, e é dado pela seguinte expressão:

$$C_i = \frac{2x}{k_i(k_i - 1)}$$

onde  $x$  é o número de pares de vizinhos do vértice  $i$  que estão ligados por uma aresta e  $k_i$  é o grau do vértice  $i$  Watts and Strogatz (1998). Por definição,  $C_i = 0$  quando  $k_i < 2$ . Define-se o coeficiente de agrupamento de um grafo,  $C$ , como a média aritmética dos coeficientes de agrupamento dos seus vértices. Demonstra-se que o coeficiente de agrupamento de uma rede aleatória de Erdős-Rényi é igual a  $\langle k \rangle / n$  (onde  $\langle k \rangle$  é o grau médio). Muitas redes complexas possuem um coeficiente de agrupamento alto, isto é, muito maior do que o coeficiente das redes aleatórias.

Outra característica observada em algumas redes é que o coeficiente de agrupamento de um vértice é inversamente proporcional ao grau do vértice, ou seja,  $C(k) \sim k^{-1}$ . Segundo Ravasz e Barabási Ravasz and Barabasi (2003), isso indica a existência de uma organização hierárquica na rede.

### 2.1.3 Motivos

Motivos são padrões de vértices e arestas que ocorrem com frequência em uma rede Milo et al. (2002). Estudos recentes encontraram em redes de dependências entre componentes motivos presentes em redes de neurônios e em circuitos eletrônicos Valverde and Solé (2005); Ma et al. (2008). Especula-se que a formação de motivos em sistemas de *software* seja resultado de restrições de custo e otimização impostas a sua evolução.

### 2.1.4 Detecção de Comunidades

Métricas globais como a distribuição de graus escondem o fato de que muitas redes se organizam em grupos de vértices relativamente independentes uns dos outros, chamados de comunidades. Detecção de comunidades é a tarefa de particionar os vértices de uma rede em comunidades de modo que, sob algum critério, os vértices de uma mesma comunidade estão mais relacionados entre si do que com vértices de fora da comunidade.

A tarefa de detectar comunidades é análoga ao que, na mineração de dados, se chama de análise de clustering — sendo cluster, nesse contexto, sinônimo

de comunidade. No domínio de engenharia de software, muitas técnicas de recuperação de arquitetura utilizam algoritmos de clustering para identificar módulos de um sistema de software.

Dada a grande variedade de algoritmos de clustering e de detecção de comunidades, é essencial avaliar os resultados produzidos pelos algoritmos. A avaliação consiste em executar o algoritmo sobre uma rede cuja estrutura de clusters (ou comunidades) é conhecida e então comparar essa estrutura com a estrutura produzida pelo algoritmo, de acordo com alguma função de similaridade entre particionamentos. Infelizmente não há muitas redes cuja estrutura de comunidades é conhecida, e as que existem são pequenas. Por essa razão testes mais extensivos são realizados com redes geradas por computador.

Newman e Girvan 2004 propuseram um modelo que gera redes com 128 vértices, divididos em quatro comunidades de 32 vértices cada. Os vértices possuem aproximadamente o mesmo número de arestas, 16, das quais  $z_{out}$  arestas ligam vértices de comunidades distintas. Newman e Girvan geraram redes com diferentes valores de  $z_{out}$  com a finalidade de comparar dois algoritmos de detecção de comunidade.

Esse modelo possui algumas limitações: os vértices possuem aproximadamente o mesmo número de arestas, as comunidades possuem o mesmo tamanho e a rede é pequena. Com base nessa observação, Lancichinetti et al. 2008 propuseram um modelo no qual a distribuição de graus e a distribuição de tamanho das comunidades são ambas leis de potência e o número de vértices é arbitrário. Cada vértice compartilha uma fração de suas arestas com vértices de outras comunidades; essa fração é o parâmetro de mistura,  $\mu$ .

O modelo de Lancichinetti et al. possui características que, a depender do domínio modelado, tornam-no um modelo pouco realista: as redes são não-orientadas; todos os vértices possuem ligações tanto para vértices da mesma comunidade quanto para vértices de outras comunidades; todas as comunidades estão, potencialmente, ligadas a todas as outras.

### 3 Redes de Dependências entre Componentes como Redes Complexas

Estudos recentes têm aplicado a teoria das redes complexas em redes de dependências entre componentes de *software*. Valverde e Solé Valverde and Solé (2003) detectaram distribuições de graus livres de escala e alto coeficiente de agrupamento em redes não-orientadas formadas por relações de agregação de tipos em diagramas UML, programas em C e programas em C++. Myers Myers (2003) analisou redes de chamadas de função em programas em C e redes de agregação e herança em programas em C++, modeladas como grafos orientados. Em ambos os casos ele identificou organização hierárquica através da distribuição do coeficiente de agrupamento,  $C(k) \sim k^{-1}$ . O estudo revelou uma correlação negativa entre o grau de entrada e o grau de saída de um vértice. Leis de potência foram encontradas em trechos das distribuições de graus de

entrada e das distribuições de graus de saída.

Distribuições de graus livres de escala também foram encontradas em programas escritos em Smalltalk Marchesi et al. (2004); Concas et al. (2007) e em Java Hyland-Wood et al. (2006); Baxter et al. (2006); Ichii et al. (2008), em dependências entre pacotes de *software* LaBelle and Wallingford (2004), em chamadas de sistema, em dependências entre bibliotecas dinâmicas Louridas et al. (2008) e até mesmo em referências entre objetos em tempo de execução Potanin et al. (2005).

É difícil comparar as pesquisas porque nem sempre elas deixam explícito quais tipos de interação foram considerados para a construção das redes. Além disso, alguns trabalhos usam ferramentas estatísticas inadequadas para leis de potência (veja a seção 4).

## 4 Apêndice: Tratamento Estatístico de Leis de Potência

### 4.1 Representação Gráfica

A estratégia mais usada para determinar se um conjunto de dados segue uma lei de potência é plotar o seu histograma em escala log-log. Se os dados seguem a lei de potência, o gráfico resultante deve ser uma reta. No entanto alguns cuidados precisam ser tomados ao traçar o histograma.

Se o histograma divide os dados em classes de mesmo tamanho, o gráfico apresenta muito ruído na cauda, na região onde se encontram os maiores valores. Isso ocorre porque nessa região os dados são esparsos, já que a probabilidade é baixa.

Uma maneira de reduzir o ruído consiste em dividir os dados em classes cujo tamanho cresce exponencialmente. Por exemplo, podemos considerar os intervalos de 1 a 2, de 2 a 4, de 4 a 8, de 8 a 16, e assim sucessivamente. A contagem de amostras de uma classe, nesse caso, deve ser dividida pelo tamanho do intervalo, a fim de normalizar o histograma Newman (2005).

Uma solução um pouco diferente é traçar o gráfico da função de distribuição cumulativa complementar (fdcc). Essa função representa a probabilidade  $\Pr(X \geq x)$ . Não há necessidade de dividir os dados em classes. Se os dados seguem uma lei de potência com expoente  $\gamma$ , sua fdcc será uma lei de potência com coeficiente  $\gamma-1$ .

### 4.2 Estimação de Parâmetros

Poucos dados empíricos seguem uma lei de potência em toda a sua extensão. Normalmente a lei de potência se aplica apenas a valores maiores do que um valor mínimo,  $x_{min}$ , e nesse caso se diz que a *cauda* da distribuição segue a lei de potência.

Ao extrair o logaritmo da lei de potência, obtemos  $\log p(k) \sim -\gamma \log(k)$ , que é uma reta com coeficiente angular  $-\gamma$ . Por essa razão uma técnica comum

para estimar o parâmetro  $\gamma$  de um conjunto de dados consiste em aplicar o método dos mínimos quadrados sobre o logaritmo dos dados a fim de encontrar o coeficiente angular. Esse método, no entanto, introduz erros sistemáticos no caso da lei de potência, e por isso recomenda-se o uso do método da máxima verossimilhança Clauset et al. (2007).

### 4.3 Teste de Ajustamento

Estimados os parâmetros  $x_{min}$  e  $\gamma$ , convém verificar se os dados se ajustam bem a uma lei de potência. Um valor alto para o coeficiente de determinação,  $R^2$ , não é suficiente para afirmar que o ajuste é bom. Uma boa forma de realizar a verificação consiste em gerar leis de potência com os parâmetros estimados e compará-las aos dados através do teste de ajustamento de Kolmogorov-Smirnov. Para um maior rigor, é preciso ajustar outras distribuições teóricas aos dados e verificar qual se ajusta melhor.

A depender da aplicação, pode não ser necessário determinar se a lei de potência é a distribuição que melhor se ajusta aos dados. Em muitos casos é suficiente observar que os dados são distribuídos de forma bastante heterogênea, formando uma cauda pesada, que decai lentamente.

## References

- Aiello, W., Chung, F., and Lu, L. (2000). A random graph model for power law graphs. *Experimental Math*, 10:53–66.
- Albert, R. and Barabasi, A.-L. (2000). Topology of evolving networks: local events and universality. *Physical Review Letters*, 85:5234.
- Barabasi, A.-L. (2007). The architecture of complexity: From network structure to human dynamics. *Control Systems Magazine, IEEE*, 27:33–42.
- Barabasi, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509.
- Baxter, G., Frean, M., Noble, J., Rickerby, M., Smith, H., Visser, M., Melton, H., and Tempero, E. (2006). Understanding the shape of java software. *SIG-PLAN Not.*, 41(10):397–412.
- Bollobás, B., Borgs, C., Chayes, J., and Riordan, O. (2003). Directed scale-free graphs. In *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 132–139, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.
- Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2007). Power-law distributions in empirical data.
- Concas, G., Marchesi, M., Pinna, S., and Serra, N. (2007). Power-laws in a large object-oriented software system. 33(10):687–708.

- Deo, N. and Cami, A. (2005). A birth-death dynamic model of scale-free networks. In *ACM-SE 43: Proceedings of the 43rd annual Southeast regional conference*, pages 26–27, New York, NY, USA. ACM.
- Dorogovtsev, S. N., Goltsev, A. V., and Mendes, J. F. F. (2002). Pseudofractal scale-free web. *Physical Review E*, 65:066122. Citado por Myers, Barabási.
- Erdős, P. and Rényi, A. (1959). On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6:290–297.
- Hyland-Wood, D., Carrington, D., and Kaplan, S. (2006). Scale-free nature of java software package, class and method collaboration graphs. In *Proceedings of the 5th International Symposium on Empirical Software Engineering, Rio de Janeiro, Brasil*.
- Ichii, M., Matsushita, M., and Inoue, K. (2008). An exploration of power-law in use-relation of java software systems. In *Proc. 19th Australian Conference on Software Engineering ASWEC 2008*, pages 422–431.
- Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., and Upfal, E. (2000). Stochastic models for the web graph. In *FOCS '00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, page 57, Washington, DC, USA. IEEE Computer Society. Copying model.
- LaBelle, N. and Wallingford, E. (2004). Inter-package dependency networks in open-source software.
- Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 78(4).
- Louridas, P., Spinellis, D., and Vlachos, V. (2008). Power laws in software. *ACM Trans. Softw. Eng. Methodol.*, 18(1):1–26.
- Ma, Y., He, K., and Liu, J. (2008). Network motifs in object-oriented software systems. *CoRR*, abs/0808.3292.
- Marchesi, M., Pinna, S., Serra, N., and Tuvèri, S. (2004). Power laws in smalltalk. In *ESUG Conference*, Kothen, Germany.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827.
- Myers, C. R. (2003). Software systems as complex networks: structure, function, and evolvability of software collaboration graphs. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68(4 Pt 2):046116.
- Newman, M. E. J. (2005). *Power laws, Pareto distributions and Zipf’s law*.



- Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69:026113.
- Potatin, A., Noble, J., Frean, M., and Biddle, R. (2005). Scale-free geometry in oo programs. *Commun. ACM*, 48(5):99–103.
- Ravasz, E. and Barabasi, A.-L. (2003). Hierarchical organization in complex networks. *Physical Review E*, 67:026112.
- Valverde, S. and Solé, R. V. (2003). Hierarchical small worlds in software architecture. (Directed Scale-Free Graphs).
- Valverde, S. and Solé, R. V. (2005). Network motifs in computational graphs: A case study in software architecture. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 72(2).
- Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.