Anais do 13O Encontro de Iniciação Científica e Pós-Graduação do ITA – XIII ENCITA / 2007
Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, Brasil, Outubro, 01 a 04, 2007.

ESTUDO EMPÍRICO DO FENÔMENO *SMALL WORLD* EM REDES SOCIAIS

Hélder Toshiro Suzuki
Instituto Tecnológico de Aeronáutica – Rua H8C, nº311, S. José dos Campos, SP, Brasil, CEP 12228-462
Bolsista PIBIC-CNPq
heldersuzuki@gmail.com

Carlos Henrique Costa Ribeiro
Instituto Tecnológico de Aeronáutica – Pr. Mal. Eduardo Gomes, 50, S. José dos Campos, SP, Brasil, CEP 12228-900
carlos@ita.br

Resumo. *É realizada uma breve introdução sobre o fenômeno Small-World e os principais conceitos são definidos. Os algoritmos clássicos para o cálculo exato de métricas de rede são analisados e verifica-se a inviabilidade do uso desses algoritmos para análise de redes grandes. São então apresentados métodos para estimar essas métricas. É descrita a evolução das ferramentas de análise utilizadas, e – a partir da reprodução de um experimento clássico para verificação do fenômeno – é proposto um novo parâmetro baseado na distribuição da métrica "betweeness centrality" suficiente para identificar a presença do fenômeno Small-World em grafos. Finalmente, uma análise de um grafo real de grande porte retirado de uma rede social é apresentada.*

1. Introdução

Estudos recentes sugerem que diversas redes que ocorrem na natureza apresentam o fenômeno *Small-World*, evidenciado por uma combinação de certos valores estatísticos (*characteristic path length* pequeno e *clusterring coefficiente* grande) medidos sobre a rede de interesse, e caracterizado pela alta eficiência na troca de informações. Este fenômeno pode ser observado em uma grande quantidade de redes biológicas e sociais, tais como redes neurais[9], mapas colaborativos e redes de co-autoria[10]. Pesquisas em redes *Small-World* em geral podem ter aplicações no entendimento de redes *Small-World* naturais ou na reprodução deste fenômeno em redes artificiais de interesse, como redes *peer to peer* (P2P) e comunicação entre agentes autônomos.

A modelagem de redes sociais em grafos pode ser uma tarefa muito subjetiva, pois definir o peso de uma aresta que representa uma relação interpessoal depende muito das propriedades da rede social que se quer analisar[1]. Por exemplo, o critério a ser utilizado na modelagem de grafos sociais para estudar a propagação de doenças sexualmente transmissíveis provavelmente deve considerar se dois indivíduos (representados por vértices) mantêm ou já tiveram contato sexual (representado por uma aresta). Para simplificar a análise de redes visando o fenômeno *Small-World*, neste artigo serão considerados apenas grafos simples não ponderados, não direcionados e conexos.

A análise de redes de relacionamento em geral é feito a partir do cálculo de métricas apropriadas, que são definidas a seguir.

1.1. Definições básicas e convenções

Sejam V e E conjuntos não-vazios tais que $E \subseteq \{\{u,w\} \mid u,w \in V, u \neq w\}$. Um *grafo* G é definido como um par ordenado $G := (V,E)$. Sejam $v \in V$ e $e = (u,w) \in E$, dizemos então que v é um *vértice* de G, e é uma *aresta* de G e que u e w são *vértices adjacentes* ou *vizinhos*. V é o conjunto de vértices de G, E é o conjunto de arestas de G, e podem ser denotados, respectivamente, por $V(G)$ e $E(G)$. $|V|$ e $|E|$ serão usualmente representados por *n* e *m*. O *grau* de $v \in V$ é denotado por $g(v)$ ou g_v e definido por $g(v) := |\{\{v,w\} \in E; w \in V\}|$, i.e. a quantidade de vizinhos de v em G. O grau médio, *k*, de *G* é $k(G) := \underset{v \in V}{media}\{g(v)\} = 2\frac{m}{n}$.

1.2. Definição – characteristic path length

Um *caminho* P em um grafo G é uma seqüência finita $P := (v_1, v_2, ..., v_N), N \geq 1$ em $V(G)$ tal que $\{v_{i-1}, v_i\} \in E(G), i = 2,3,...,N$. P é um caminho de v_1 até v_N e seu comprimento é $N-1$, a quantidade de

arestas que atravessa. Sejam $u, v \in V(G)$, se existir caminho em G de u até v, a distância mínima entre u e v, é o menor comprimento dentre todos os caminhos em G de u até v, e é denotada por $d(u,v)$. O *characteristic path length*, L, é a mediana das médias das menores distâncias entre cada vértice e todos os outros.

$$L = mediana\{\underset{v \in V}{media}\{d(u,v)\}; u \in V\}$$

1.3. Definição – clustering coefficient

A vizinhança Γ_v de $v \in V(G)$ é o subgrafo de G que consiste dos vértices adjacentes a v, não incluindo o próprio v. O *clustering coefficient* de $v \in V(G)$, $\gamma(v)$, mede a proporção de vizinhos de v que são vizinhos entre si.

$$\gamma(v) = \frac{|E(\Gamma_v)|}{\binom{g_v}{2}},\ g_v \geq 2$$

$$\gamma(v) = 1,\ g_v = 1$$

O *clustering coefficient* de G é $C(G) := \underset{v \in V}{media}\{\gamma(v)\}$.

1.4. Definição – betweeness centrality

Seja $\sigma(s,t) = \left|\left\{P \mid P = (s,u_1,...,u_{d(s,t)-1},t) \text{ é caminho em } G\right\}\right|$, $s,t \in V(G)$, o número de caminhos mínimos de s até t em G, e $\sigma(s,t \mid v)$ o número de caminhos mínimos de s até t que contêm o vértice v como intermediário. $\sigma(s,t \mid v) = \left|\left\{P \mid P = (s,u_1,...,u_{d(s,t)-1},t) \text{ é caminho em } G; \exists m \in \{1,...,d(s,t)-1\},\ v = u_m\right\}\right|$, $v,s,t \in V(G)$. A dependência de um par $(s,\ t)$ em um vértice v é definida por $\delta(s,t \mid v) = \frac{\sigma(s,t \mid v)}{\sigma(s,t)}$, e $\delta(s \mid v) = \sum_{t \neq v} \sigma(s,t \mid v)$ é a dependência de s em v. O *betweeness centrality* de v é definido por $c_B(v) = \sum_{s \neq v} \delta(s \mid v)$. O *betweeness centrality* de um vértice v indica a importância do vértice v para as distâncias mínimas entre pares de outros vértices do grafo.

1.5. Algoritmos Clássicos

Esta Seção apresenta os algoritmos clássicos utilizados para o cálculo das métricas apresentadas nas Seções 1.2 e 1.3.

1.5. 1. Algoritmo clássico para o cálculo do *characteristic path length*

Seja G um grafo conexo.
Entrada: (G)

$\forall u \in V(G)$, execute uma busca em largura com raiz em u, determinando $d(u,v)$, $\forall v \in V(G)$.
$\bar{d}(u) \leftarrow media\{d(u,v); v \in V\}$, $\forall u \in V(G)$
Ordene $\{\bar{d}(u)\}$.
Saída: $L \leftarrow mediana\{\bar{d}(u); u \in V\}$

1.5..2. Algoritmo clássico para o cálculo do *clustering coefficient*

Seja G um grafo conexo.

Entrada: (G)
$\forall v \in V(G)$, determine $\gamma(v)$ da seguinte forma:
Se $g_v = 1$, $\gamma(v) \leftarrow 1$

Senão $S_v \leftarrow 0$,
Para todo $\{u,w\} \subset V(\Gamma(v)) : S_v \leftarrow S_v + 1$, se u e w são adjacentes em G.

$$\gamma(v) \leftarrow 2\frac{S_v}{g_v(g_v - 1)}$$

Saída: $C \leftarrow media\{\gamma(v); v \in V\}$

1.6. Análise dos algoritmos clássicos

A determinação da complexidade dos algoritmos clássicos apresentados para os cálculos do *characteristic path length* e do *clustering coefficient* de um grafo é necessária para verificar teoricamente a viabilidade desses algoritmos para grafos grandes ($n > 10^6$).

1.6.1. Teorema para o algoritmo 1.5.1

O algoritmo 1.5.1 retorna o *characteristic path length* e é $O(n(n+m+\log(n)))$.

Demonstração:

Pela definição 1.2 o algoritmo 1.5.1 calcula corretamente o *characteristic path length.*
São executadas n buscas em largura, $O(n^2)$ operações para o cálculo de $\bar{d}(u), \forall u \in V$ e uma ordenação de n elementos para o cálculo de L. Sabemos que uma busca em largura tem custo linear $O(n+m)$ e a ordenação leva $O(n\log(n))$, portanto a complexidade do algoritmo 1.5.1 é $O(n(n+m+\log(n)))$.

1.6.2. Teorema para o algoritmo 3.1.2

O algoritmo 1.5.2 retorna o *clustering coefficient* e é $\Omega(n\,(k-1)^2)$.

Demonstração:

Pela definição 2.3 o algoritmo calcula corretamente o *clustering coefficient.*
Se $g(v) = 1$, é necessário $1 > (g(v)-1)^2$ operação para fazer $\gamma(v) \leftarrow 1$.
Se $g(v) \geq 2$, $|V(\Gamma(v))| = g(v)$, então são necessárias pelo menos $\binom{g(v)}{2}$ operações para determinar $\gamma(v)$, supondo uma representação de grafo que permita determinar se u e v são adjacentes em $O(1)$.
$\binom{g(v)}{2} = \frac{g(v)(g(v)-1)}{2} > \frac{(g(v)-1)^2}{2}$, então para o cálculo de $\gamma(G)$ são necessárias pelo menos $\frac{1}{2}\sum_{v\in V}(g(v)-1)^2$ operações.
Da desigualdade de Jensen [5], temos que $\frac{\sum_{v\in V}(g(v)-1)^2}{n} \geq \left(\frac{\sum_{v\in V}(g(v)-1)}{n}\right)^2$, mas $\sum_{v\in V} g(v) - 1 = 2m - n = n(k-1)$,
então $\frac{1}{2}\sum_{v\in V}(g(v)-1)^2 \geq \frac{1}{2}n\,(k-1)^2$.

1.7. Método Monte Carlo

O custo para determinar os valores exatos do *characteristic path length* e *do clusterring coefficient* utilizando-se os algoritmos 1.5.1 e 1.5.2 é muito grande para grafos de grande porte (n > 10^6). Foram então considerados métodos para determinar valores aproximados dessas métricas de redes utilizando-se um esforço computacional muito menor com relação ao cálculo dos valores exatos.

1.7.1 Definição – (ε, δ)-aproximação

Suponha que $Z_1, Z_2,...$ são independentes e distribuídos identicamente (i.d.i.) no intervalo [0,1] com média μ. Seja $\varepsilon, \delta \in [0,1]$, então $\tilde{\mu}$ é uma (ε, δ)-aproximação de μ se $P(\mu(1-\varepsilon) \leq \tilde{\mu} \leq \mu(1+\varepsilon)) \geq 1-\delta$.

Como C é definido por $C = \underset{v \in V}{media}\{\gamma(v)\}$, uma (ε, δ)-aproximação de C será denotada por (ε, δ)-C

1.7.2. Definição – (ε, δ)-mediana*

Para uma variável aleatória, *X*, com distribuição *F(X)*, M_q é uma p-mediana se $P(X \leq M_p) > p$ e $P(X \geq M_p) > 1-p$. Seja $\varepsilon, \delta \in [0,1]$, então $\hat{\mu}$ é uma (ε, δ)-mediana se $P\left(\hat{\mu} = M_p;\ p \in \left[\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2}\right]\right) \geq 1-\delta$.

Como L é definido por $L = mediana\{\underset{v \in V}{media}\{d(u,v)\};\ u \in V\}$, (ε, δ)-L é definida por $(\varepsilon,\delta) - L := (\varepsilon,\delta) - mediana\{\underset{v \in V}{media}\{d(u,v)\};\ u \in V\}$.

1.7.3. Determinação dos valores aproximados

Para determinar uma (ε, δ)-aproximação do *clustering coefficient* de uma rede, foi utilizado o algoritmo de aproximação *AA* descrito por Dagum[3]. E para determinar uma (ε, δ)-mediana do *characteristic path length* de uma rede foi utilizado o algoritmo *Median estimation* descrito por Huber[2].

1.8. Algoritmos para determinação do *betweeness centrality*

Brandes[6] descreveu um algoritmo exato que determina o *betweeness centrality* de todos os vértices de um grafo com custo O(nm) de tempo e O(n+m) de espaço. Posteriormente, junto com Pich, propôs um método para estimar o *betweeness centrality* por meio da determinação da dependência de alguns vértices-origem com relação a todos os outros[7], baseando-se no trabalho de Eppstein e Wang[8].

Resumidamente, o algoritmo exato de Brandes[6] baseia-se em uma recorrência que permite o cálculo da dependência de um vértice origem para todos os outros em tempo linear, e iterando o vértice de origem obtém-se a dependência de todos os vértices com relação a todos os vértices, e portanto o *betweeness centrality*, em tempo O(nm). O algoritmo de estimação[7], por sua vez, determina a dependência de alguns vértices escolhidos aleatoriamente com relação a todos os outros e, extrapolando o resultado parcial, obtém a aproximação para o *betweeness centrality*.

2. Ferramentas de análise

O desenvolvimento de uma primeira ferramenta de análise foi inicialmente realizado na linguagem Java 1.5 utilizando-se a IDE *NetBeans*[13], o projeto do software foi orientado a objetos, e baseou-se nos algoritmos clássicos para a determinação do *characteristic path length* e do *clustering coefficient* de grafos.

Nesta primeira ferramenta também foi criada uma interface gráfica (Figura 1) de modo a facilitar a utilização do programa na execução de experimentos. O caráter modular permitiu posteriormente a adição de um módulo gerador de grafos para a reprodução do experimento de Watts-Strogatz[1].

Percebeu-se que para grafos grandes, a ferramenta em Java apresentou problemas práticos no tratamento de grafos. O *overhead* causado pelas estruturas da linguagem fez com que a simples leitura do grafo nas estruturas de dados apresentasse um desempenho inaceitável. Para contornar este problema, foi desenvolvida uma ferramenta nas linguagens C e C++, obtendo-se ganhos substanciais no tempo de processamento e leitura do grafo. É importante citar que só o tempo de leitura da entrada foi reduzido por um fator de 500, desde a primeira versão da ferramenta de análise em Java até a última versão desenvolvida em C/C++.

---

* Huber[2] define (p, ε)-mediana com probabilidade 1- δ, mas aqui estamos interessados apenas quando p = 0.5, então adaptações necessárias foram feitas à definição. Nas definições de Dagum[3] para a (ε, δ)-aproximação e de Huber para a (ε, δ)-mediana os papéis de ε e δ são invertidos. Para evitar confusão a definição de Dagum foi adotada como base.

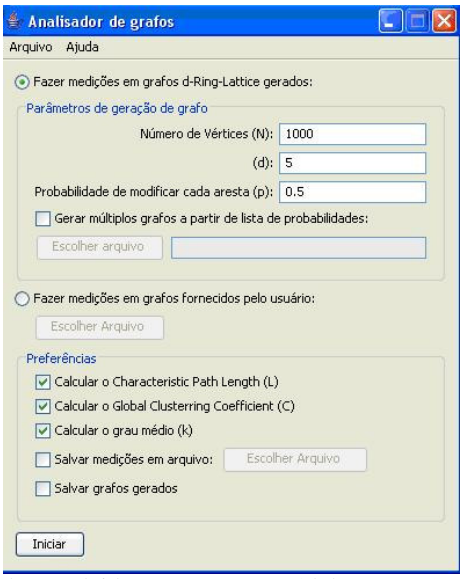

Figura 1: Tela inicial da ferramenta desenvolvida em Java com módulo para reprodução do experimento de Watts-Strogatz.

Os algoritmos clássicos para o cálculo das métricas de redes também foram implementados nesta nova versão, e empiricamente verificou-se a inviabilidade computacional desses algoritmos (tal como previsto na análise teórica descrita na Seção 1.6) . Também foram implementados os métodos Monte Carlo para determinação de (ε, δ)-aproximação e (ε, δ)-mediana de forma genérica e adaptável para os cálculos aproximados das métricas de redes, *characteristic path length* e *clustering coefficient*.

A escolha da estrutura de dados para a representação dos grafos teve que considerar os custos de tempo e memória dos algoritmos em grafos a serem utilizados. Com isto em mente, uma lista de adjacências foi escolhida para representar os grafos, por resultar num custo de memória $O(n+m)$ e permitir uma busca em largura em tempo $O(n+m)$. Entretanto, cabe observar que o custo para verificar a adjacência de dois vértices não é $O(1)$ como suposto na demonstração do Teorema 1.6.2, mas $O(\min\{g_u, g_v\})$ nesta representação. Claramente este fato não invalida a aplicabilidade do Teorema 3.1.3.2 na análise da implementação utilizada, pois o teorema estabelece um limite inferior para o custo de $O(\min\{g_u, g_v\})$ que é maior do que o $O(1)$ suposto na demonstração. Os custos de adicionar e remover arestas também são proporcionais ao grau dos vértices. Essas operações serão importantes na implementação de métodos construtivos de grafos.

É claro que poderíamos utilizar outra representação de grafo que permitisse o custo de $O(1)$ para verificação de adjacências, no entanto o custo para iterar sobre os vizinhos de um vértice (como numa representação por *hash tables*) poderia ser afetado, fazendo com que a busca em largura deixasse de ser linear em *n* e *m*. Apesar disto, a medição do *characteristic path length*, que foi implementada utilizando-se buscas em largura é a operação que mais requer recursos computacionais dentre as métricas utilizadas.

Para facilitar a leitura dos dados na forma de lista de adjacências, o formato dos arquivos de entrada adotado fornecia, para cada vértice, o grau e uma lista dos vértices adjacentes. Inicialmente, o arquivo era armazenado em modo texto, posteriormente, foi observada uma redução no tempo de leitura da entrada por um fator de 10 ao armazenar o grafo em um arquivo binário.

3. Reprodução do experimento de Watts-Strogatz

Até alguns anos atrás, o estudo teórico de grafos considerava primordialmente apenas grafos puramente regulares ou puramente aleatórios[1][12]. Tal modelagem permitia determinar várias características sobre tais grafos, observando comportamentos locais bem definidos, no caso de grafos regulares, ou comportamentos globais estatisticamente bem definidos, no caso de grafos aleatórios. No entanto tais modelos não apresentavam o fenômeno *Small-World* como foi observado em várias redes reais.

A ocorrência do fenômeno *Small-World* em um grafo é caracterizada por um *clustering coefficient* grande e um *characteristic path length* pequeno.

O *clustering coefficient* é considerado grande se for muito maior do que o *clustering coefficient* de um grafo aleatório de mesmas dimensões e próximo do *clustering coefficient* de um grafo regular de mesmas dimensões. O

characteristic path length, por sua vez, é considerado pequeno se for muito menor do que o *characteristic path length* de um grafo regular e próximo do characteristic path length de um grafo aleatório de mesmas dimensões.

$$C_{Regular} \approx C_{Small-World} >> C_{Aleatório}$$

$$L_{Regular} << L_{Small-World} \approx L_{Aleatório}$$

O experimento de Watts-Strogatz[14] procurou explorar fenômenos em grafos intermediários, analisando a variação das métricas em estruturas que topologicamente se encontram entre grafos regulares e grafos aleatórios.

3.1. Definição – *Grafo* d-Ring-Lattice

Um grafo *d-Ring-Lattice* de *n* vértices é tal que existe uma função bijetora $r : V \to \{0,1,...,n-1\}$ e dois vértices $u, v \in V, u \neq v$ são adjacentes se e somente se $r(u) = (r(v) + i) \bmod n, 1 \leq i \leq d$. Visualmente é como se todos os vértices fossem dispostos em torno de um círculo, e cada vértice é conectado por uma aresta aos *d* vértices mais próximos à direita e à esquerda. Da definição, os graus de cada vértice e, consequentemente, o grau médio, de um *d-Ring-Lattice* são $2d$.

3.2. Definição – *Parâmetro B2*

Para analisar a distribuição do valor do *betweeness centrality* de todos os vértices propomos o parâmetro *B2*, que representa a proporção dos vértices de um grafo G cujo *betweeness centrality* supera o dobro da média:

$$\mathrm{B2(G)} = \frac{\left|\left\{ v \in \mathrm{V(G)} \mid \mathrm{B_C}(v) > 2\ \overline{\mathrm{B_C}(G)} \right\}\right|}{\left|\mathrm{V(G)}\right|},$$

onde $\overline{\mathrm{B_C}(G)}$ é o *betweeness centrality* médio de um grafo: $\overline{\mathrm{B_C}(G)} = \frac{\sum_{v \in V(G)} B_C(v)}{|V(G)|}$

3.3. Procedimento

Para gerar grafos intermediários entre grafos regulares e grafos aleatórios, inicia-se com o um grafo *d-Ring-Lattice* regular. Para cada aresta do grafo há uma probabilidade $0 \leq p \leq 1$ para que esta aresta seja removida do grafo e que uma aresta aleatória não existente seja adicionada ao grafo no lugar. Note que este processo não altera a quantidade de arestas do grafo, portanto preservando o grau médio. Para $p = 0$ temos o grafo *d-Ring-Lattice* regular intacto. Quando $p = 1$ temos um grafo aleatório com $k = 2d$. A Figura 2 fornece uma visualização esquemática dos grafos gerados para o experimento.

No experimento em questão as propriedades estruturais dos grafos são medidas por meio do *characteristic path length* e do *clustering coefficient*. Para convenientemente compararmos a variação dessas métricas com relação a p, esses valores estarão normalizados com relação às respectivas métricas quando $p = 0$.

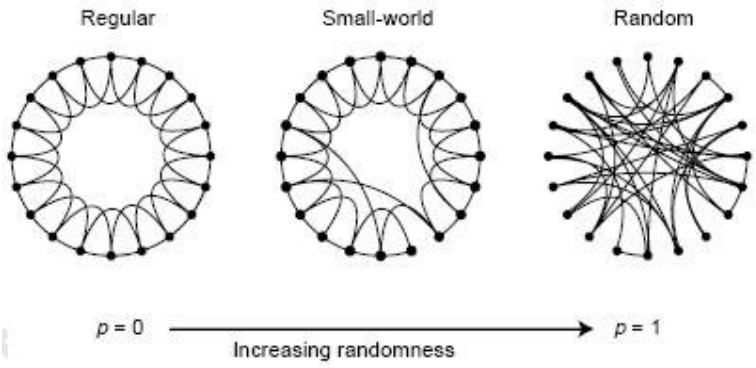

Figura 2: Visualização do experimento de Watts-Strogatz. Ilustração retirada do artigo publicado na revista Nature [1].

Foram adicionadas à ferramenta em Java as classes desenvolvidas especificamente para este experimento. Devido ao caráter modular e orientado a objetos do projeto, foi uma tarefa relativamente fácil a integração das classes de medição e de geração de grafos. Para que fosse possível comparar os resultados com o experimento original de *Watts-Strogatz*, a reprodução foi realizada com os mesmos parâmetros, $n = 1000$ e $d = 5$.

O experimento de Watts também foi reproduzido utilizando-se a ferramenta desenvolvida em C/C++ com as implementações dos algoritmos de cálculo aproximado do *clusterring coefficient*, do *characteristic path length* e do algoritmo de cálculo exato do *betweeness centrality*. O desempenho superior da ferramenta desenvolvida em C/C++ possibilitou a realização de um maior número de medições ao longo do experimento.

3.4. Resultados

Os resultados das medições exatas nos grafos gerados variando-se a probabilidade $0 \leq p \leq 1$ estão na Figura 4. Para comparação, o gráfico gerado originalmente por *Watts-Strogatz* está na Figura 3. É importante observar que devido à quantidade relativamente pequena de vértices e arestas ($n = 1000$ e $m = 10000$), grafos gerados com probabilidades muito pequenas ($p \approx 10^{-3}$) podem apresentar características variáveis, já que nessa faixa têm menor chance de apresentar um comportamento estatístico uniforme, sobretudo quando a métrica é sensível no intervalo em questão, como ocorre com o *characteristic path length* para $10^{-4} < p < 10^{-3}$. Esta dispersão é evidente na Figura 5.

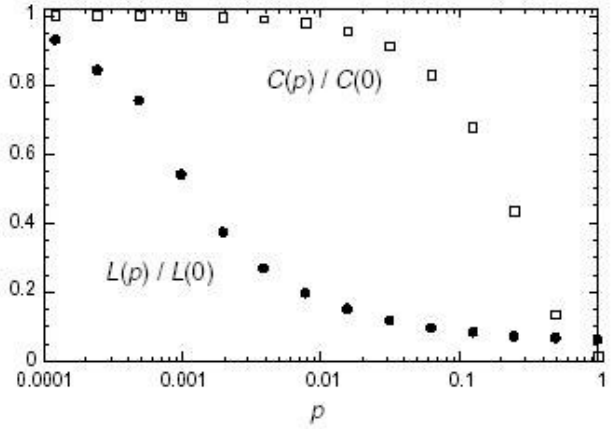

Figura 3: Gráfico do experimento de Watts-Strogatz original. Ilustração retirada de [1].

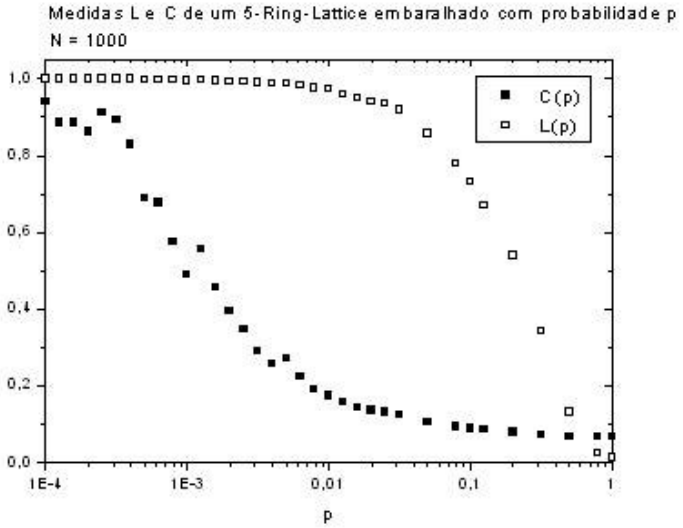

Figura 4: Reprodução do experimento de Watts-Strogatz, usando ferramenta em Java e algoritmos clássicos de medição.

Pode-se comparar a evolução da média do *betweeness centrality* e do *characteristic path length* na Figura 6. Nota-se que não há correlação significativa do fenômeno *Small-World* com relação a média do *betweeness centrality*, porém a noção intuitiva de que quando o *characteristic path length* é pequeno a importância de todos os vértices para os caminhos mínimos é grande, e portanto o *betweeness centrality* médio também é grande, é confirmada para os grafos do experimento de Watts.

Um resultado não-trivial e bastante interessante é que o parâmetro *B2* aparenta ser um indicador suficiente da presença do fenômeno *Small-World* no caso do experimento de Watts. De fato, observa-se na Figura 7 que os maiores valores de *B2* coincidem com a região onde o *characteristic path length* é pequeno e o *clustering coefficiente* é grande.

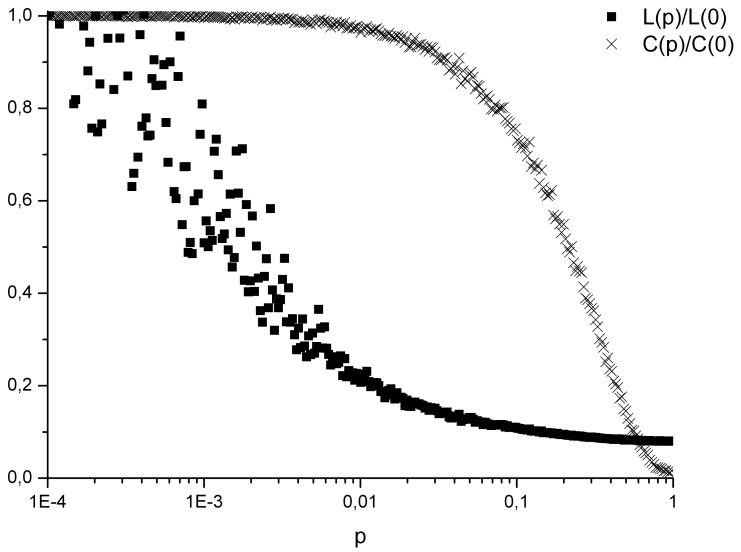

Figura 5: Reprodução do experimento de Watts-Strogatz utilizando a ferramenta desenvolvida em C/C++ com os métodos Monte Carlo

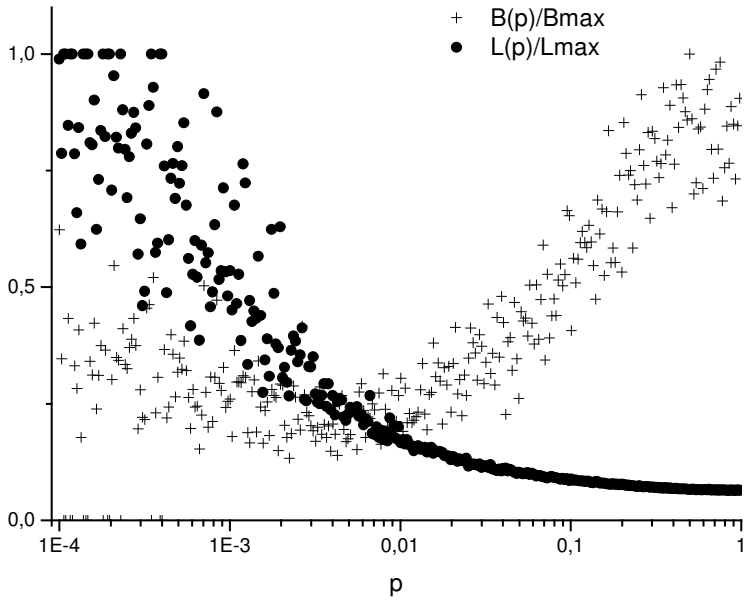

Figura 6: Evolução do characteristic path length e da média do betweeness centrality com relação a probabilidade p.

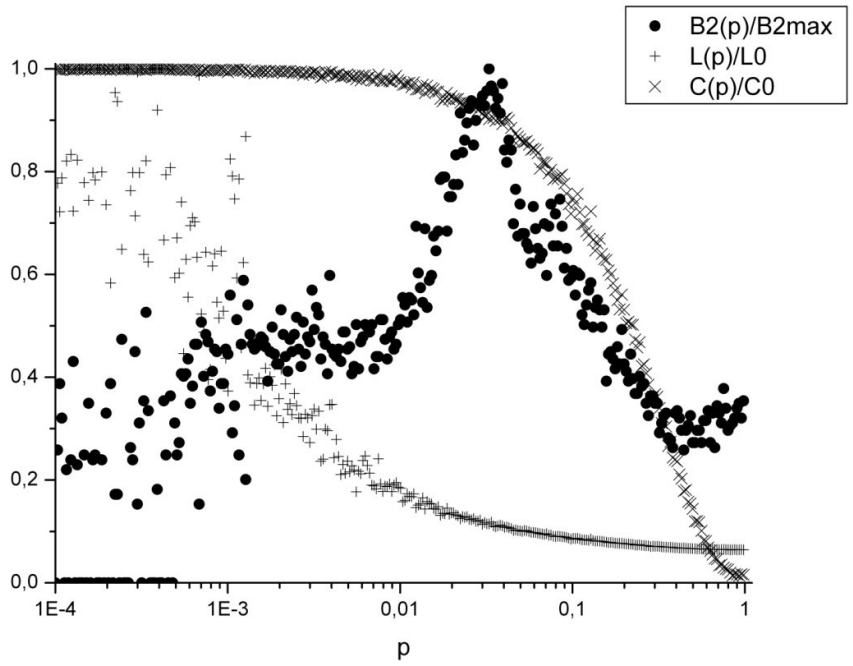

Figura 7: Evolução do parâmetro *B2*, *L* e *C* com relação a probabilidade *p*. O parâmetro *B2* representa um indicador suficiente para o fenômeno Small-World nos grafos do experimento de Watts-Strogatz.

4. Experimentos em uma rede social real

4.1. A rede social Orkut

O Orkut (www.orkut.com) é um famoso sítio de relacionamentos via Internet bastante popular no Brasil, onde cada usuário administra um perfil, podendo adicionar o perfil de outras pessoas na sua lista de amigos utilizando-se critérios pessoais [11]. Como foi comentado anteriormente, uma das principais dificuldades em modelar redes sociais é a subjetividade do critério ao decidir se duas pessoas são ou não conectadas na rede social. No caso da rede social do Orkut, não se tem uniformidade no critério de adição na lista de amigos, pois cada pessoa pode utilizar critérios arbitrários. Além disso, a rede é incompleta, pois nem todas as pessoas possuem conta no Orkut e mesmo que tenham nada garante que as listas de amizade são devidamente atualizadas pelos usuários.

Apesar destas limitações em assumir o Orkut como um modelo de uma rede social real, deve ser observado que o foco deste trabalho não é um estudo específico sobre redes sociais, não interessando fundamentalmente a fonte do grafo a ser estudado, mas sim o fato do grafo ser relativamente grande ($n \approx 10^6$), apresentar características do fenômeno *Small-World* [4] e não ter sido artificialmente gerado como no experimento de *Watts-Strogatz*.

4.2. Tratamento dos Dados

Os experimentos foram realizados utilizando-se uma amostra do Orkut (sub-grafo) fornecida pela empresa Google.

Os dados sobre rede social do Orkut precisaram ser normalizados para que pudessem ser analisados pelos softwares desenvolvidos. Os dados são exatamente os mesmos que foram utilizados em [4], e o procedimento no tratamento dos dados foi análogo.

Os rótulos dos vértices foram renomeados numa seqüência numérica de 1 a *n*. Para esta tarefa, um pequeno software específico foi desenvolvido utilizando *hash tables* para armazenar os rótulos antigos e novos dos vértices. Além disso, os dados foram formatados para facilitar a leitura na estrutura de lista de adjacências, conforme comentado na Seção 2.

O grafo fornecido possui 1.386.332 vértices distribuídos em 1.741 componentes conexas, sendo que a maior componente conexa possui 1.381.565 vértices representando mais de *99,6%* do total de vértices. Como é desejado trabalhar com grafos conexos, o restante do grafo foi descartado, sobrando uma única componente conexa com n = 1.381.565 e m = 40.066.866, esta componente será denotada por G_{Orkut}.

4.3. Resultados

As medições em G_{Orkut} foram realizadas utilizando-se a ferramenta desenvolvida em C e C++ já especificada, rodando em um computador AMD Athlon XP 2800+ com 512 MB de RAM. Devido ao elevado custo computacional, foram calculadas aproximações das métricas, (ε, δ)-L e (ε, δ)-C. Os resultados estão apresentados na *Tabela 6.1*.

(0.1, 0.1)-L	5.10
(0.01, 0.01)-C	0.202

Tabela 0.1: Métricas de G_{Orkut}

Foram necessárias 483.441 amostras para o cálculo de (0.01, 0.01)-C e 1.504 *s* de processamento no total. Para o cálculo de (0.1, 0.1)-L foram necessárias 1.306 amostras e 2.959 *s* (~50 min) de processamento no total. Os resultados são compatíveis com os obtidos por Borges[4].

5. Conclusão

O parâmetro *B2* proposto, baseado na distribuição do *betweeness centrality* dos vértices de um grafo, mostrou ser um indicador suficiente da presença do fenômeno *Small-World* nos grafos do experimento de Watts. Este resultado é importante pois evidencia a importância da distribuição do *betweeness centrality* na caracterização do fenômeno *Small-World*, indicando que estudos mais aprofundado sobre a distribuição do *betweeness centrality* e de outros parâmetros correlatos podem representar um importante passo na caracterização das propriedades de grafos Small-World.

A análise do G_{Orkut} foi realizada com os algoritmos de aproximação, devido à inviabilidade dos algoritmos exatos para grafos desta magnitude, e permitiu verificar resultados anteriores sobre o mesmo grafo. Forneceu ainda uma noção sobre o processamento necessário para análises de grafos dessa magnitude.

Os resultados obtidos são animadores. A ocorrência de redes *Small-World* na natureza sugere que a solução ótima para diversos tipos de problemas, do ponto de vista da sobrevivência e adaptabilidade em um ambiente seletivo, contém agrupamentos e relações com características *Small-World*. Melhorias em ferramentas de análise e a proposta de novos parâmetros de rede podem tornar possível a tipificação topológica dos vértices e a identificação de clusters e comunidades nas redes *Small-World*, possibilitando tanto aplicações em redes existentes como na síntese de redes artificiais. Ficou evidente que a análise de uma rede *Small-World* real demanda grande esforço computacional, portanto o aumento da eficiência das ferramentas de análise é essencial para futuros estudos mais aprofundados do fenômeno *Small-World.*

6. Referências

[1] WATTS, D. **Small Worlds: The Dynamics of Networks Between Order and Randomness**. Princeton: Princeton University Press, 1999. pág. 11-89.

[2] HUBER, M. Nearly Optimal Running Time for Monte Carlo Sampling. School of Operations Research and Industrial Engineering: Cornell University, 1997.

[3] DAGUM, P. *et al*. An optimal algorithm for Monte Carlo estimation. **SIAM Journal on Computing**, [S.l.], vol. 29, n. 5, pág.1484-1496, abril. 2000.

[4] BORGES FILHO, E.. **Análise do Fenômeno *Small-World* em uma Rede de Relacionamentos na Internet**. 2006. 47f. Trabalho de Conclusão de Curso (Graduação). Instituto Tecnológico de Aeronáutica, São José dos Campos.

[5] **Jensen's inequality**. Wikipedia. Disponível em <http://en.wikipedia.org/wiki/Jensen's_inequality>. Acesso em 22 dez. 2006.

[6] BRANDES, U. A Faster Algorithm for Betweeness Centrality. **Journal of Mathematical Sociology**, [S.I.], 25(2): 163-177, 2001.

[7] BRANDES, U.; PICH, C. Centrality Estimation on Large Networks. Department of Computer and Information Science. University of Konstanz. August 18, 2006.

[8] EPPSTEIN, D.; WANG, J. Fast approximation of centrality. **Journal of Graph Algorithms and Applications**, 8(1):39–45. 2004.

[9] STAM, C. J.; JONES, B. F. et al. Small-World Networks and Functional Connectivity in Alzheimer's Disease. **Cerebral Cortex**. Vol. 17, No. 1. pp. 92-99. January 2007.

[10] NEWMAN, M. E. J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. **Physical Review E**. Vol. 64. June 2001.

[11] **Orkut**. Disponível em: <http://www.orkut.com>. Acesso em: 4 de Agosto de 2007.

[12] Bollabas, B. **Random Graphs**, Academic Press, London, 1985.

[13] NetBeans. Disponível em: <http://www.netbeans.org>. Acesso em: 3 de Julho de 2006.

[14] WATTS, D. J.; STROGATZ, S. H**.** Collective Dynamics of Small-World Networks. **Nature**, [S.l.], vol. 393, pág. 440-442, junho. 1998.