Estudo de um método alternativo de limiarização para gerar grafos a partir de matrizes de dados

Autor: Evandro A. N. Sanches Orientador: Arnaldo Mandel
Universidade de São Paulo

Introdução

Para modelar um problema no contexto da teoria dos grafos é comum gerar a matriz de adjacência a partir de uma matriz de dados experimentais. Para isso, aplica-se aos dados um processo chamado de limiarização. O processo convencional de limiarização consiste em adotar um limiar $p \in]0,1[$ e produzir a matriz de adjacência B de um grafo com base em uma matriz A de dados. A produção ocorre da seguinte forma:

$$b_{i,j} = \begin{cases} 1 & \text{se } a_{i,j} \leq p \\ 0 & \text{c.c} \end{cases} \quad (1)$$

Objetivos

Propor um método de escolha do limiar individualizado por entrada da matriz A, satisfazendo as seguintes duas propriedades:

Propriedade 1 *Para uma matriz A, com entradas $a_{i,j} \sim U[0,1]$ i.i.d., o valor do grau médio do grafo limiarizado pelo processo sugerido deve coincidir com o grau médio de um grafo gerado na limiarização pelo processo controle.*

Propriedade 2 *Cada limiar$_{i,j}$ utilizado no processo sugerido depende apenas do problema e do método de coleta de dados.*

Formulação Teórica

Toda matriz considerada no trabalho é simétrica e pertencente ao espaço $\mathbb{R}^{n\times n}$.

Definição 1 *Uma Matriz de interferência H é tal que na entrada $\eta_{i,j} \in H$ está um valor numérico que reflita alguma caracterítica da modelagem referente à aresta (i,j).*

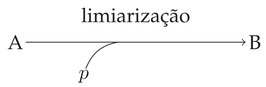
Definição 2 *Uma função de perturbação é em princípio qualquer função crescente.*

Definição 3 *Uma Matriz de interferências perturbadas M é o resultado de aplicar a função de perturbação aos valores de interferência da matriz H normalizados de modo que a média dos valores de M seja* 1.

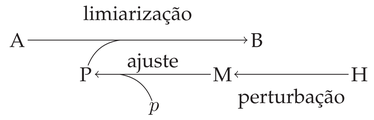
Definição 4 *Uma Matriz de proporções ajustadas P é o resultado da multiplição dos valores de interferência pertubada pelo limiar base ($\pi_{i,j} = \mu_{i,j} \cdot p$). Assim $\pi_{i,j}$ está em $]0,1[$ e contém o valor da proporção de corte para a entrada $a_{i,j}$.*

Experimento

As matrizes de interferência testadas foram matrizes de distância em grafos das famílias listadas no quadro *Grafos utilizados*. Com respeito a função de perturbação a ser utilizada, foram avaliadas funções na forma $f(d) = d^{\alpha}$. Para o experimento foi adotado $\alpha = 0,2$.
Para o tamanho das matrizes foi utilizado $n = 225$. Para a proporção de corte base utilizou-se $p = 0,25$, $p = 0,5$ e $p = 0,75$.
Como dados de entrada utilizou-se matrizes A geradas pseudo-aleatoriamente com a função runif() do pacote básico do R. Cada entrada $a_{i,j} \in A$ sorteada idependentemente e uniformemente no intervalo $]0,1[$.
Gerou-se 100 matrizes A e para cada uma delas aplicou-se o processo de limiarização convencional (aqui denominado controle). Também aplicou-se a toda A gerada, cinco limiarizações pelo método ajustado, um para cada matriz de interferência H mencionada. Foram avaliasdas as estatísticas do quadro *Estatísticas*. E os resultados foram coletados na forma de razões *valor_ajustado* sobre *valor_controle*.



Esquema de *limiarização pelo método controle*



Esquema de *limiarização pelo método ajustado*

Resultados

As tabelas a seguir mostram valores na forma *média* $\pm$ *desvio_padrão*, obtidos de cada uma das estatísticas observadas. A tabela 1 contém os resultados da execução do experimento com $p = 0,25$, a tabela 2 os resultados com $p = 0,5$ e a tabela 3 com $p = 0,75$.

$p = 0,25$	Estrela	Caminho	Ciclo	Tabuleiro	Tabuleiro Toroidal
Grau_médio	1,0000 ±0,0005	0,9999 ±0,0050	0,9997 ±0,0044	0,9995 ±0,0039	0,9996 ±0,0036
Hamming	1.0022 ±0.0005	1.1514 ±0.0046	1.1300 ±0.0050	1.0930 ±0.0037	1.0754 ±0.0037
Fechamento	1,0000 ±0,0008	1,0000 ±0,0007	0,9999 ±0,0006	0,9999 ±0,0005	0,9999 ±0,0005
Centralidade	0,9998 ±0,0079	0,9921 ±0,0157	1,0004 ±0,0146	0,9953 ±0,0147	1,0010 ±0,0135
Transitividade	1,0002 ±0,0007	0,9988 ±0,0062	0,9970 ±0,0053	1,0008 ±0,0050	0,9992 ±0,0044
Clustering	1,0002 ±0,0007	0,9993 ±0,0062	0,9969 ±0,0054	1,0009 ±0,0051	0,9992 ±0,0045
Modularidade	1,0063 ±0,0429	1,0063 ±0,0479	1,0056 ±0,0429	1,0089 ±0,0475	1,0132 ±0,0393
Tamanho_cluster	0,9978 ±0,0369	0,9988 ±0,0443	0,9983 ±0,0450	0,9966 ±0,0455	0,9972 ±0,0431

Tabela 1: Estatísticas para p=0,25

$p = 0,5$	Estrela	Caminho	Ciclo	Tabuleiro	Tabuleiro Toroidal
Grau_médio	1,0000 ±0,0004	0,9997 ±0,0030	1,0004 ±0,0028	1,0002 ±0,0026	0,9999 ±0,0024
Hamming	1.0022 ±0.0004	1.1512 ±0.0035	1.1303 ±0.0032	1.0929 ±0.0025	1.0750 ±0.0022
Fechamento	1,0000 ±0,0001	1,0000 ±0,0010	1,0001 ±0,0009	1,0001 ±0,0008	0,9999 ±0,0008
Centralidade	0,9986 ±0,0052	0,9862 ±0,0110	1,0007 ±0,0097	0,9891 ±0,0104	1,0000 ±0,0091
Transitividade	1,0002 ±0,0004	0,9988 ±0,0033	0,9978 ±0,0029	1,0012 ±0,0027	0,9994 ±0,0025
Clustering	1,0002 ±0,0004	0,9993 ±0,0033	0,9978 ±0,0029	1,0013 ±0,0028	0,9994 ±0,0025
Modularidade	1,0024 ±0,0524	0,9880 ±0,0512	0,9999 ±0,0510	0,9947 ±0,0535	0,9981 ±0,0542
Tamanho_cluster	0,9903 ±0,0586	1,0148 ±0,0615	1,0002 ±0,0599	1,0115 ±0,0557	1,0083 ±0,0663

Tabela 2: Estatísticas para p=0,5

$p = 0,75$	Estrela	Caminho	Ciclo	Tabuleiro	Tabuleiro Toroidal
Grau_médio	0,9999 ±0,0003	1,0000 ±0,0024	0,9999 ±0,0025	1,0001 ±0,0018	1,0000 ±0,0017
Hamming	1.0022 ±0.0003	1.1517 ±0.0027	1.1307 ±0.0024	1.0928 ±0.0021	1.0752 ±0.0020
Fechamento	1,0000 ±0,0001	1,0005 ±0,0014	0,9999 ±0,0015	1,0004 ±0,0011	1,0000 ±0,0010
Centralidade	0,9976 ±0,0036	0,9794 ±0,0098	1,0034 ±0,0090	0,9824 ±0,0103	1,0018 ±0,0086
Transitividade	1,0001 ±0,0003	0,9990 ±0,0025	0,9972 ±0,0025	1,0011 ±0,0018	0,9995 ±0,0017
Clustering	1,0001 ±0,0003	0,9995 ±0,0025	0,9972 ±0,0025	1,0013 ±0,0018	0,9995 ±0,0017
Modularidade	0,9992 ±0,0394	0,9605 ±0,0502	0,9789 ±0,0537	0,9813 ±0,0505	0,9995 ±0,0492
Tamanho_cluster	0,9999 ±0,0444	1,0019 ±0,0501	0,9940 ±0,0530	0,9977 ±0,0471	1,0012 ±0,0513

Tabela 3: Estatísticas para p=0,75

Pode-se observar das tabelas que o grau médio se manteve, assim como requisitado pela propriedade 1. Também é possível ver que os valores da distância de Hamming foram expressivos mas os outros parâmetros não.

Materiais & Métodos

O estudo foi dividido em duas etapas:

- Formulação teórica: Descrição matematica do processo de limiarização pelo método ajustado,
- Testes sobre dados criados: Implementação do processo em uma linguagem de programação para observar seus efeitos sobre um conjunto de dados em particular.

Informações adicionais sobre os testes executados:

- Linguagem e Ambiente
  - Linguagem: R.
  - Ambiente de desenvolvimento: R versão 3.3.1-1~jess que é oferecida no repositório CRAN - *Comprehensive R Archive Network*.
  - Modulos auxiliar: igraph versão 1.0.0 e suas dependências.

Estatísticas

Grau_médio: $\frac{1}{n} \cdot \sum_{v \in V(G)} \deg(v)$

Centralidade: $\frac{1}{n} \cdot \sum_{v \in V} \text{k-core}(v)$

Hamming: $1 + \frac{\sum_{0 \leq i < j \leq n} |b_{ctl_{i,j}} - b_{adj_{i,j}}|}{p \cdot \binom{n}{2}}$

Transitividade: $\frac{3\alpha(G)}{\sum_{v \in V} \binom{\deg(v)}{2}}$

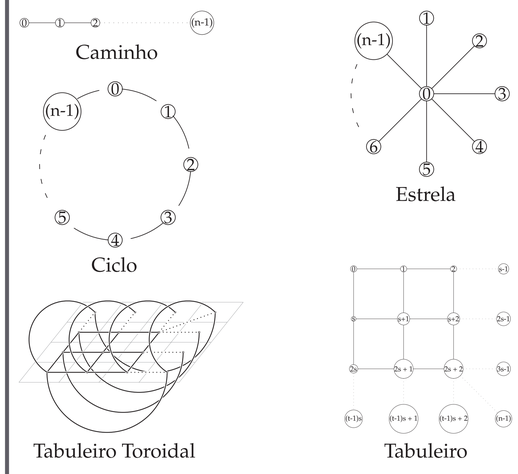
Fechamento: $\frac{1}{n} \cdot \sum_{v \in V} \frac{1}{\sum_{i \neq v} d(v,i)}$

Clustering: $\frac{1}{|V'|} \cdot \sum_{v \in V'} \frac{\alpha(v)}{\binom{\deg(v)}{2}}$

Modularidade: $\frac{1}{2 \cdot |A(G)|} \cdot \sum_{0 \leq i, < j \leq n} \left(a_{i,j} - \frac{\deg(i) \cdot \deg(j)}{2 \cdot |A(G)|} \right) \cdot \delta(c_i, c_j)$

Tamanho_cluster: $\frac{1}{\sum_{c \in C} \text{size}(c)} \cdot \sqrt{\sum_{c \in C} \text{size}(c)^2}$

Grafos utilizados



Conclusão

Os resultados obtidos mostram que as estatísticas selecionadas são robustas e não se alteram diantes de modificações expressivas, 10% da quantidade de arestas esperadas. Por outro lado, talvez seja necessário pensar outras estatísticas mais sensíveis quando esse tipo de alteração no grafo modelado tiver relevância para o problema em questão.

Contato

Web www.linux.ime.usp.br/~evnsan/mac0499
Email vnsanc@gmail.com