Estatística de Redes Sociais

Antonio Galves

Módulo 2

5a. aula

Grafos aleatórios e redes sociais.

Mundo pequeno, Erdős-Rényi, Watts-Strogatz.

Recapitulando: matriz de adjacência de um grafo

- A matriz de adjacência é uma forma computacional de representar o grafo $G = (V, E)$.
- Se o grafo for dirigido, para todo par ordenado de vértices (v, v'), com $v \neq v'$, definimos

$$M(v, v') = 1 \text{ ou } M(v, v') = 0$$

  se houver (respectivamente, se não houver) uma flecha (uma aresta dirigida), indo de v a v'.
- Se o grafo não for dirigido, a ordem (v, v') ou (v', v) não importa e $M(v, v') = M(v', v)$.

Recapitulando: o grafo de Erdős-Rényi

- Conjunto de vértices: $V = \{1, \ldots, N\}$.
- Os valores das entradas da matriz

$$M = (M(v, v') : v, v' \in V, v \neq v')$$

  são variáveis aleatórias, independentes e identicamente distribuidas (i.i.d.) com

$$\mathbb{P}\{M(v, v') = 1\} = p \text{ e } \mathbb{P}\{M(v, v') = 0\} = 1 - p\,,$$

  e $M(v, v) = 0$, para todo $v \in V$.
- Essa classe de grafos será designada com a notação $G(N, p)$.

Uma noção básica

- Em um grafo não dirigido $G = (V, E)$, tendo M como matriz de adjacência, o *grau* $D(v)$ *de um vértice* v *é definido como*

$$D(v) = \sum_{v' \in V : v' \neq v} M(v, v')$$

- *No caso de um grafo* $G(N, p)$

$$\mathbb{E}[D(v)] = (N-1)p$$

- *e para todo* $k = 0, \ldots, N-1$

$$\mathbb{P}\{D(v) = k\} = \binom{N-1}{k} p^k (1-p)^{(N-1-k)}$$

- **Atenção à mudança de notação:** $D(v)$, *em vez de* $deg(v)$*.*

O que se diz na literatura sobre grafos e redes sociais

- Diz-se que os grafos descrevendo redes sociais deveriam ter:
- as características mundo pequeno
- e para todo vértice v, a distribuição de $D(v)$ deveria ter uma cauda longa.

As duas características de um grafo mundo pequeno

- A primeira característica é que dois *amigos* de um mesmo ator, tem grande probabilidade de também serem amigos.
- Formalmente, dados três vértices v_1, v_2 e v_3, se $M(v_1, v_2) = 1$ e $M(v_1, v_3) = 1$, então com alta probabilidade $M(v_3, v_2) = 1$.
- A segunda característica é que a distância entre dois vértices quaisquer é tipicamente muito menor do que $|V|$.
- A distância entre v e v' é o menor $k \geq 1$ tal que existem vértices $v_0, v_1, ..., v_k$, com $v = v_0$ e $v_k = v'$, tais que $M(v_i, v_{i+1}) = 1$, para $i = 0, ..., k-1$.
- Notação: $|V|$ (lê-se *cardinal de* V*) denota o número de elementos do conjunto* $|V|$*.*

O exemplo de Watts e Strogatz

- Um célebre artigo de Watts e Strogatz, publicado em 1998 na revista Nature, discute essa questão, através do seguinte modelo:
- Começamos com um grafo regular, tendo $V = \{1, \ldots, N\}$ e com matriz de adjacência M satisfazendo: $M(v, v') = 1$, se $1 \leq |v - v'| \leq k$, e $M(v, v') = 0$ caso contrário.
- Aqui, $k \geq 1$ é um inteiro muito menor do que N.
- Usamos a convenção $N + 1 = 1$. Em outras palavras, consideramos os elementos de $V = \{1, \ldots, N\}$ dispostos numa circunferência, de tal forma que os vértices 1 e N são vizinhos.

Q U I Z

O grafo regular apresentado nos dois últimos quadros satisfaz as condições mundo pequeno?

Coeficiente de aglomeração e diâmetro de um grafo

- Duas definições são úteis para tornar mais claras as duas características mundo pequeno.
- Vamos definir o coeficiente de aglomeração e o diâmetro de um grafo.
- O coeficiente de aglomeração mede a proporção de amigos de um ator são amigos entre si.
- O diâmetro de um grafo mede a distância típica entre dois vértices de um grafo.

Coeficiente de aglomeração

- Dado um vértice v, $D(v) = \sum_{v' \in V, v' \neq v} M(v, v')$ é o número de *vizinhos* de v.
- $\binom{D(v)}{2}$ é o número total de triângulos que podemos construir tendo v como vértice e tendo como outros dois vértices vizinhos de v.
- Isto acontece quando todos os vizinhos de v estão conectados por arestas entre si (grafo completo).
- $c(v)$ é a fração que tem como denominador $\binom{D(v)}{2}$ e como numerador o número de triângulos que realmente existem tendo v como vértice e vizinhos de v como outros dois vértices.
- $C(G)$, o coeficiente de aglomeração do grafo, é a média dos valores de $c(v)$ para todos os vértices $v \in V$.

Coeficiente de aglomeração de um grafo

- Dado um grafo $G = (V, E)$ e um vértice $v \in V$,
- $$c(v) = \frac{1}{\binom{D(v)}{2}} \sum_{v_1, v_2 \in V} M(v, v_1) M(v_1, v_2) M(v_2, v)$$
- $$C(G) = \frac{1}{|V|} \sum_{v \in V} c(v)$$

Por que o grafo regular não é um mundo pequeno? Porque o seu diâmetro é muito grande.

- O diâmetro do grafo regular é proporcional a $|V|$.
- O que isso significa?
- Vamos usar a notação $dist(v, v')$ para indicar a distância entre os vértices v e v'.
- Definimos os diâmetro ($L(G)$) do grafo $G = (V, E)$ como sendo a média das distâncias entre dois vértices quaisquer do grafo.
- $$L(G) = \frac{1}{2|E|} \sum_{v,v' \in V, v \neq v'} dist(v, v')$$
- No caso de um grafo aleatório, $L(G)$ é a esperança das distâncias entre dois vértices quaisquer.

Aleatorizando um grafo regular

- Para resolver o problema do grande diâmetro do grafo regular, Watts e Strogatz propõe aleatorizá-lo parcialmente.
- Como isso é feito?
- Escolhemos um valor $p \in [0, 1]$.
- Cada aresta do grafo regular é mantida como está com probabilidade $1 - p$, ou alterada com probabilidade p.
- Se decidirmos alterá-la, uma de suas pontas escolhida com probabilidade $1/2$ é reconectada em um outro vértice escolhido uniformemente dentre todos os vértices, com excessão das extremidades originais da aresta modificada.

Aleatorizando um grafo regular

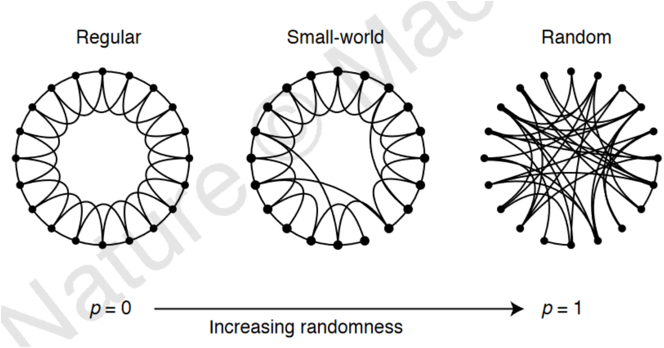

Figura retirada do artigo do Watts e Strogatz na Nature (1998).

Valores de $C(p)$ e $L(p)$ em função de p

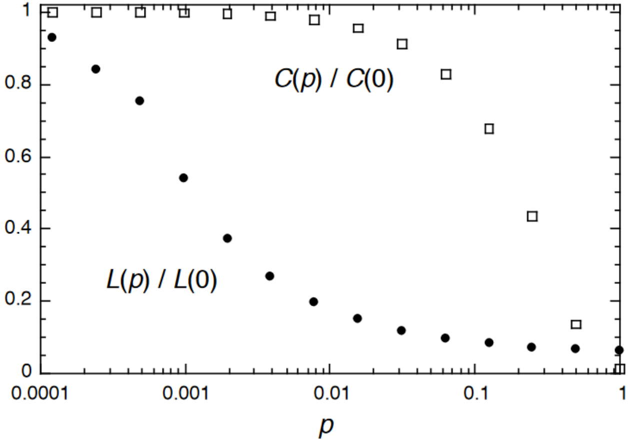

Figura retirada do artigo do Watts e Strogatz na Nature (1998).

QUIZ

- Que tipo de rede social é bem descrita pelo modelo de Watts-Strogatz?

O $G(N,p)$ tem as características de mundo pequeno?

- Na próxima aula faremos contas. Daqui até lá, estime o valor de $C(G(N,p))$.
- Faça isso simulando $G(N,p)$ muitas vezes, independentemente umas das outras, e com $N = 100$.
- Observe que em cada simulação, os valores de $c(v)$ para $v = 1, ..., 100$ são variáveis aleatórias identicamente distribuídas. O mesmo acontece com as variáveis aleatórias $D(v)$.
- Para cada simulação, calcule $c(v)$ para todos os valores de v e em seguida calcule a média para aquela simulação.
- Calcule o histograma dos valores de $C(G(N,p))$ obtidos nas diversas simulações.

QUIZ

- Estamos trabalhando com o grafo $G(100, p)$ onde $p \in (0, 1)$.
- Saber que $D(1) = 0$ nos diz algo a respeito de $D(100)$?

Lembrando

- Na literatura sobre grafos e redes sociais
- diz-se que os grafos descrevendo redes sociais deveriam ter:
- as características mundo pequeno
- **e para todo vértice v, a distribuição de $D(v)$ deveria ter uma cauda longa.**
- O que quer dizer ter cauda longa?
- Na próxima aula discutiremos esta última questão.

Enquanto a próxima aula não chega

- Vamos olhar a distribuição de $D(v)$ em $G(N,p)$, escrevendo $D_N(v)$ em vez de $D(v)$.
- Vamos calcular o limite $D_N(v)/N$ quando $N \to \infty$.
- $$\frac{D_N(1)}{N} = \frac{1}{N} \sum_{v=2}^{N} M(1,v).$$
- As variáveis aleatórias $(M(1,v))_{v=2,\ldots,N}$ são i.i.d. Portanto, pela Lei dos Grandes Números,
$$\lim_{N\to\infty} \frac{D_N(1)}{N} = \mathbb{E}(M(1,2)) = p.$$
- Podemos dizer mais sobre esse limite.

Flutuações de $D_N(1)/N$ em torno de p

- O Teorema-Limite Central diz que a distribuição de

$$\sqrt{N}\,(D_N(1)/N - p) = \frac{D_N(1) - Np}{\sqrt{N}}$$

converge para uma distribuição $N(0, p(1-p))$ quando $N \to \infty$.

- Ou seja, para todo $t \in \mathbb{R}$, vale o limite

$$\lim_{N\to\infty} \mathbb{P}\left(\frac{D_N(1) - Np}{\sqrt{N}} < t\right) = \frac{1}{\sqrt{2\pi p(1-p)}} \int_{-\infty}^{t} e^{-\frac{1}{2}\frac{s^2}{p(1-p)}}\, ds.$$

- Essa é a famosa aproximação normal da distribuição binomial. Ela diz, entre outras coisas, algo sobre a concentração $D_N(1)$ em torno de Np, e isso tem consequências sobre as características que se esperam de um grafo que representa uma rede social.

Exercício

1. Calcule o coeficiente de aglomeração e o diâmetro do grafo regular.
2. Simule várias vezes $G(N, p)$ com N, p fixados. Use essas simulações para estimar o coeficiente de aglomeração e o diâmetro de $G(N, p)$.
3. Leia no artigo de Watts e Strogatz a discussão sobre o diâmetro e o coeficiente de aglomeração do grafo regular com diversas taxas de aleatorização.
4. Reveja a aproximação normal da distribuição binomial e suas aplicações estatísticas.

Referências

- Esta é uma leitura indispensável.
  Watts, D., Strogatz, S. Collective dynamics of 'small-world' networks. Nature 393, 440–442 (1998). https://doi.org/10.1038/30918
- Este artigo pode ser interessante.
  Random graphs with arbitrary degree distributions and their applications M. E. J. Newman, S. H. Strogatz, and D. J. Watts Phys. Rev. E 64, 026118