Estatística de Redes Sociais

Antonio Galves

Módulo 2

6a. aula

Grafos aleatórios e redes sociais.

Graus com cauda pesada, grafos *rico fica mais rico*.

Recapitulando:

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 pesada.

Recapitulando:

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 $|V|$.*

Recapitulando: coeficiente de aglomeração de um grafo

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

Recapitulando: diâmetro de um grafo

- 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 escolhidos ao acaso.

Grafos com graus de cauda pesada

- Mais precisamente:
- cujos vértices tem graus
- com distribuição de cauda pesada.
- Mas o que é uma distribuição de cauda pesada?

Distribuições com cauda pesada

- Seja $q(k)$, $k = 0, 1, 2, ...$ uma distribuição de probabilidade no conjunto $\mathbb{N} = \{0, 1, 2, ...\}$ dos números naturais.
- Isso significa que $q(k) \in [0, 1]$, para todo $k \in \mathbb{N}$ e a série $\sum_{k=0}^{+\infty} q(k)$ é somável com $\sum_{k=0}^{+\infty} q(k) = 1$.
- Se a série é somável, seu resto

$$R(K) = \sum_{k \geq K} q(k) \to 0,$$

quando $K \to +\infty$.

- Dizemos que a cauda da distribuição q é pesada se $R(K)$ decresce para zero muito mais lentamente do que uma distribuição exponencial.

Exemplo

- Distribuição geométrica de parâmetro $p \in (0,1]$:

$$q(k) = (1-p)^k p, k \in \mathbb{N}.$$

- Observe que a soma $\sum_{k=0}^{+\infty} q(k) = 1$ e

$$R(K) = (1-p)^K = e^{-K[-\log(1-p)]}.$$

- Ou seja, a distribuição geométrica tem um resto que decresce exponencialmente rápido, e portanto, não tem cauda pesada.

DESAFIO

- Considere a seguinte sequência decrescente de números no intervalo $(0, 1)$:

$$\frac{1}{2}, \frac{1}{6}, \frac{1}{12}, \frac{1}{20}, \frac{1}{30}, \frac{1}{42}, \frac{1}{56}, \ldots$$

  como essa sequência foi formada?
- Se $q(k)$ for o k-ésimo termo da sequência, o que pode ser dito da série $\sum_{k=1}^{\infty} q(k)$?
- O que pode ser dito do resto $R(K)$, quando $K \to \infty$?

Resposta do D E S A F I O

- $$q(k) = \frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1}$$
- Portanto,
$$\sum_{k=1}^{K} \frac{1}{k(k+1)} = \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \ldots \left(\frac{1}{K} - \frac{1}{K+1}\right) = 1 - \frac{1}{K+1}.$$
- Portanto, a série
$$\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = \lim_{K \to \infty} \left(1 - \frac{1}{K+1}\right) = 1.$$
- Ou seja, $q(k), k = 1, 2, ...$ define uma distribuição de probabilidade nos números inteiros estritamente positivos.

Uma distribuição com média infinita e cauda pesada!

- Média infinita:

$$\sum_{k=1}^{+\infty} kq(k) = \sum_{k=1}^{+\infty} k\frac{1}{k(k+1)} = \sum_{k=1}^{+\infty} \frac{1}{k+1} = +\infty.$$

- Cauda pesada: o termo $R(K)$ é igual a

$$\sum_{k\geq K} \frac{1}{k(k+1)} = \left(\frac{1}{K} - \frac{1}{K+1}\right) + \left(\frac{1}{K+1} - \frac{1}{K+2}\right) + \ldots = \frac{1}{K}.$$

O peso da cauda dos vértices de $G(N, p)$

- 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.

O que o Teorema-limite central diz sobre $D_N(1)/N$

- O Teorema-Limite Central diz que a distribuição de

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

converge para a normal $N(0, p(1-p))$, 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.

Observação sobre a distribuição normal

- O Teorema-limite Central diz que a distribuição da variável aleatória
$$\frac{D_N(1) - Np}{\sqrt{N}}$$
converge para a distribuição normal de média 0 e variância $p(1-p)$.
- Portanto, a distribuição da variável aleatória renormalizada
$$\frac{D_N(1) - Np}{\sqrt{N}\sqrt{p(1-p)}}$$
converge para a distribuição normal de média 0 e variância 1.

O resto da distribuição de $D_N(1)$

- Vamos usar a notação $q_N(k) = \mathbb{P}(D_N(1) = k), k \in \mathbb{N}$. e calcular o resto $R_N(K)$ dessa distribuição, com $K = Np + t\sqrt{Np(1-p)}$.

$$R(Np + t\sqrt{Np(1-p)}) = \mathbb{P}\Big(D_N(1) \geq Np + t\sqrt{Np(1-p)}\Big).$$

-

$$\mathbb{P}\Big(D_N(1) \geq Np + t\sqrt{Np(1-p)}\Big) = \mathbb{P}\Big(\frac{D_N(1) - Np}{\sqrt{Np(1-p)}} \geq t\Big)$$

- Usando a aproximação normal,

$$\mathbb{P}\Big(\frac{D_N(1) - Np}{\sqrt{Np(1-p)}} \geq t\Big) \approx \frac{1}{\sqrt{2\pi}} \int_t^{+\infty} e^{-\frac{1}{2}s^2} ds.$$

A cauda da distribuição Normal

- Fixado $t > 0$, o resto $R(t)$ da distribuição $N(0,1)$ se escreve

$$R(t) = \frac{1}{\sqrt{2\pi}} \int_t^{+\infty} e^{-\frac{1}{2}s^2} ds = \frac{1}{\sqrt{2\pi}} \int_0^{+\infty} e^{-\frac{1}{2}(t+s)^2} ds.$$

- Observamos que

$$e^{-\frac{1}{2}(t+s)^2} = e^{-\frac{1}{2}(t^2+s^2+2ts)} \leq e^{-\frac{1}{2}(t^2+s^2)}.$$

- Portanto, $R(t)$ é majorado por

$$\frac{1}{\sqrt{2\pi}} \int_0^{+\infty} e^{-\frac{1}{2}(t^2+s^2)} ds = e^{-\frac{1}{2}t^2} \left(\frac{1}{\sqrt{2\pi}} \int_0^{+\infty} e^{-\frac{1}{2}s^2} ds \right) = \frac{1}{2} e^{-\frac{1}{2}t^2}.$$

- Portanto, a cauda da $N(0,1)$ decresce exponencialmente.
- **Conclusão:** A cauda de $D_N(1)$ decresce rapidamente quando se afasta de seu valor médio Np. Logo, desse ponto de vista, $G(N,p)$ não é um bom modelo para redes sociais.

Grafos *rico fica mais rico*

- Tanto o conjunto dos vértices, quanto o conjunto das arestas evolui ao longo do tempo.
- Isso o torna um modelo interessante para descrever como uma rede social se estabelece e evolui ao longo do tempo.
- Vamos trabalhar com tempo t evoluindo de maneira discreta, $t = 0, 1, 2, ...$.
- Para todo $t = 0, 1, 2, ...$, $G_t = (V_t, E_t)$ é o grafo produzido pela evolução do sistema até o instante t inclusive. Para todo $v \in V_t$ denotamos $D_t(v)$ o grau de v no grafo G_t. Também denotamos M_t a matriz de adjacência do grafo G_t.
- Vamos supor por ora que esses grafos são não dirigidos.

Definição da cadeia de grafos $G_t = (V_t, E_t), t \geq 0$.

1. Inicialização: Definimos $V_0 = \{1, \ldots, N_0\}$ com $N_0 \geq 2$ qualquer fixado e um conjunto de arestas E_0 de modo que o grau $D_0(v) \geq 1$ para todo $v \in V_0$.
2. Para todo $t \geq 1$:
   2.1 $V_t = V_{t-1} \cup \{|V_{t-1}| + 1\}$.
   2.2 $E_t = E_{t-1} \cup \{|V_{t-1}| + 1, \xi_t\}$, onde ξ_t é um elemento de V_{t-1} escolhido aleatóriamente com a distribuição

$$\mathbb{P}(\xi_t = v) = \frac{D_{t-1}(v)}{\sum_{v' \in V_{t-1}} D_{t-1}(v')}.$$

Em linguagem de gente

- Os atores que vão entrando sucessivamente nesta rede social vêm do reservatório infinito $\{1, 2, ...\}$.
- Se $V_0 = \{1, ..., N_0\}$, então V_t terá como atores $\{1, ..., N_0, N_0 + 1, ..., N_0 + t\}$.
- Cada novo ator escolhe um ator já presente na rede com o qual se ligar. Essa escolha é influenciada pelo grau dos atores no instante anterior. Quanto maior for o grau de um ator no instante anterior, maior será a sua probabilidade de ser escolhido pelo novo ator que entrou no intante seguinte.
- A probabilidade de escolha de um ator já presente como parceiro do novo ator é proporcional ao seu grau anterior.
- Daí o nome *rico fica mais rico*.

Algumas fórmulas

- Para todo $t \geq 1$, $|V_t| = |V_{t-1}| + 1$.
- Para todo ator $v \in V_t$, $D_t(v) = 1$, se $v = |V_t|$.
- $D_t(v) = D_{t-1}(v) + 1$, se $v = \xi_t$,
- e $D_t(v) = D_{t-1}(v)$, em caso contrário.
- Para todo $v \in V_{t-1}$,

$$\mathbb{P}(D_t(v) = D_{t-1}(v) + 1 | G_{t-1}) = \frac{D_{t-1}(v)}{\sum_{v' \in V_{t-1}} D_{t-1}(v)}.$$

Q U I Z

- Que tipo de situação social é bem descrita pelo grafo do tipo *rico fica mais rico*?

Exercícios

1. Escreva um código para implementar o grafo *rico fica mais rico*. Use esse código para simular a evolução de um sistema começando com $N_0 = 2$ e $M_0(1, 2) = 1$.
2. A classe de grafos *rico fica mais rico* foi feita especificamente para obter vértices com caudas pesadas. Verifique na sua simulação se isso acontece.

Referências

- O modelo *rico fica mais rico*, também chamado de modelo de ligação preferencial (*preferential attachment*) foi introduzido no artigo
- Barabási, Albert-László, and Réka Albert. "Emergence of Scaling in Random Networks." Science 286.5439 (1999): 509–512. Crossref. Web.
- Uma versão aberta do artigo está disponível neste link.