Redes Complexas – CPS 765
2020/3

Prof. Daniel R. Figueiredo

Segunda Lista de Exercícios

ATENÇÃO! Para ajudar no processo de aprendizagem, explique suas repostas.

Questão 1: Macacos digitando aleatoriamente. Considere um macaco digitando aleatoriamente diante de um teclado com n letras e uma barra de espaço. O macaco tecla a barra de espaço com probabilidade q, e digita qualquer outra tecla com igual probabilidade, ou seja $(1-q)/n$. O espaço é usado para delimitar palavras. Ao digitar aleatoriamente, nosso macaco vai produzir muitas palavras. Estamos interessados em calcular a frequência de ocorrência das palavras. Em particular, vamos mostrar que nosso macaco produz palavras cujo ranqueamento segue uma lei de potência!

1. Calcule q_k, a probabilidade do macaco digitar uma palavra específica com k letras. Mostre que q_k decresce exponencialmente com k.
2. Calcule quantas palavras distintas podem ser formadas com k letras.
3. Assumindo que todas as palavras com k letras ocorrem no ranqueamento antes das palavras com $k+1$ letras, determine a posição do ranqueamento da primeira palavra com k letras.
4. Considere a palavra na posição j do ranqueamento, para $j = 1, 2, \ldots$. Quantas letras esta palavra possui?
5. Seja p_j a probabilidade do macaco gerar a j-ésima palavra do ranqueamento. Determine p_j.
6. Mostre que p_j segue uma lei de potência, ou seja, $p_j = aj^{-b}$ para duas constantes a, b. Determine o expoente (valor de b) em função dos parâmetros n, q do problema. Dica: use a equivalência $x^{\log_y z} = z^{\log_y x}$ e ignore a função teto.
7. Qual é a relação entre os parâmetros n e q e o peso da cauda? Quando é que a cauda é mais pesada?

Questão 2: MLE e Pareto. Considere um conjunto de dados $x_1, x_2, \ldots, x_n$, com $x_i > 0$ para todo i. Considere a modelagem destes dados por uma distribuição de Pareto. Responda às perguntas abaixo:

1. Determine analiticamente o valor da função de *log-likelihood* em função do parâmetro a da distribuição de Pareto. Que premissa você teve que fazer sobre os dados para chegar a esta função?
2. Determine analiticamente o estimador de máxima *likelihood* (MLE) do parâmetro a, encontrando o máximo da função *log-likelihood*. **Dica**: derivar e encontrar a raiz.
3. Lembrando que o estimador $\hat{a}$ é uma variável aleatória (pois depende das amostras), determine analiticamente seu valor esperado.

Questão 3: Coeficiente de clusterização no $G(n, p)$. Mostre que o valor esperado do coeficiente de clusterização local do modelo $G(n, p)$ é p. **Dica**: Condicionar no grau do vértice, e propriedade da torre da esperança!

Questão 4: *Threshold function* para vértices isolados. Queremos entender quando o grafo $G(n, p)$ passa a não ter mais vértices isolados (vértices com grau zero). Mais especificamente, queremos determinar a *threshold function* para a seguinte propriedade P: "ausência de vértices isolados". Seja E_0 o valor esperado do número de vértices isolados. Determine E_0 em função de n e p. Precisamos agora determinar uma função $p(n)$ tal que $E_0 \to 0$ quando $n \to \infty$. Para facilitar, assuma $p(n) = c\frac{ln(n)}{n-1}$. Para qual valores de $c > 0$ observamos a propriedade P?

Dica: utilize a aproximação $(1 - x) \approx e^{-x}$.

Questão 5: *Threshold function* para diâmetro 2. Queremos analisar mais rigorosamente quando o grafo $G(n, p)$ possui diâmetro igual a 2. Para isto, vamos considerar a seguinte propriedade P: "ausência de pares de vértices a distância maior que 2". Seja $E_{>2}$ o valor esperado do número de pares de vértices a distância maior que 2. Determine $E_{>2}$ em função de n e p. Precisamos agora determinar uma função $p(n)$ tal que $E_{>2} \to 0$ quando $n \to \infty$. Para facilitar, assuma $p(n) = \frac{c\sqrt{ln(n)}}{\sqrt{n-2}}$. Para qual valores de $c > 0$ observamos a propriedade P?

Dica: utilize as aproximações $(1 - x) \approx e^{-x}$ e $np + 1 \approx np$.

Questão 6: Grafos completos com Pareto. Considere o seguinte modelo de grafos aleatório. Seja G_0 o grafo inicial com um único vértice. A cada instante de tempo $t = 1, 2, \ldots$ um novo vértice t é adicionado ao grafo G_{t-1}. O grau do vértice t no instante de sua chegada, d_t, possui distribuição de Pareto. Em particular, $d_t = \lfloor x_t \rfloor$ onde $x_t \sim Par(x_0, a)$ (x_t é uma amostra), e $Par(x_0, a)$ é a distribuição de Pareto com parâmetros $x_0 > 0$ e $a > 0$. Por fim, cada ponta de aresta incidente ao vértice t é conectada a um vértice de G_{t-1} escolhido de forma aleatória e uniforme. Caso $d_t > t$ (grau maior do que o número de vértices em G_t), então tomamos $d_t = t$.

Queremos estudar quando que as instâncias de grafos deste modelo são grafos completos. Em particular, seja $p_{a,x_0,t}$ a probabilidade de G_t ser um grafo completo quando usamos parâmetros x_0 e a. Determine uma expressão para esta probabilidade. Seja $\delta < 1$ uma constante. Determine um limite superior para a tal que $p_{a,x_0,t} > \delta$, em função de δ, t e x_0.

Questão 7: Modelo BA com habilidade. No modelo BA, vimos que a probabilidade de um vértice v_i (que entrou na rede no tempo i) ser incidente a uma nova aresta no tempo $t > i$ é proporcional ao seu grau no tempo t, ou seja $p_{i,t} \propto d_i(t)$. Imagine agora um modelo onde cada vértice possui um atributo $\alpha_i > 0$ aleatório que também afeta sua popularidade mas que é determinado no momento em que o vértice entra na rede (e não muda com o tempo). Neste modelo, temos $p_{i,t} \propto d_i(t) + \alpha_i$. Determine $d_i(t)$ (na verdade, seu valor esperado) para este modelo.

Dica: Siga os passos utilizados para derivar $d_i(t)$ no modelo original.

Questão 8: Modelo WS. Considere o modelo WS e responda as perguntas abaixo:

1. Considerando que $p = 0$, calcule analiticamente o coeficiente de clusterização local e a distância média entre um vértice e todos os outros da rede construída pelo modelo.

2. Considerando que $p = 1$, explique por que o coeficiente de clusterização local e a distância média não são estatisticamente idênticas ao modelo $G(n, p)$. Entretanto, argumente intuitivamente por que o modelo $G(n, p)$ é uma boa aproximação neste caso, principalmente para os valores médios dessas métricas.