复杂网络

高一鸣 王方正 邵逸岚 庞冰瑶 赵帅 张永辉

计算机学院

2018.04.27

复杂网络基础

赵帅

什么是复杂网络？

- 在网络理论的研究中，复杂网络是由数量巨大的节点和节点之间错综复杂的关系共同构成的网络结构。用数学的语言来说，就是一个有着足够复杂的拓扑结构特征的图。
- 具有自组织、自相似、吸引子、小世界、无标度中部分或全部性质的网络称为复杂网络。

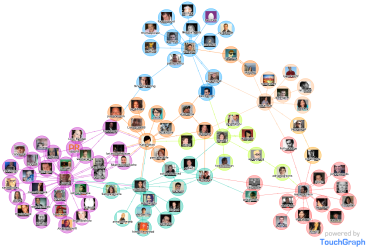


图 1: 社交网络

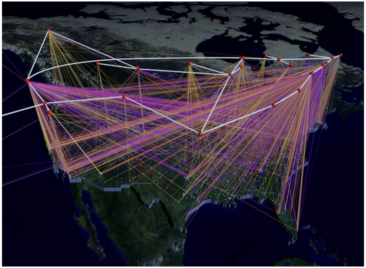


图 2: Internet 一部分的可视化

网络的图表示

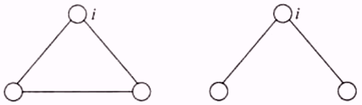
图 3: 七桥问题-Euler 图

图论之父-Euler.

一个具体网络可以抽象为一个由点集 V 和边集 E 组成的图 $G = (V, E)$。节点数记为 $N = |V|$，边数记为 $M = |E|$。一条边对应一对点。

无向网络、有向网络、加权网络、无权网络、重边、自环。

- **平均路径长度**
  网络中两个节点 i 和 j 之间的距离 d_{ij} 定义为连接这两个节点的最短路径上的边数。网络的直径 $D=\max_{i,j} d_{ij}$ 。
  网络的平均长度 $L=\frac{1}{2}\frac{1}{N(N+1)}\sum_{i\geq j} d_{ij}$，也叫特征长度路径。
- **聚类系数**
  k_i 个节点之间最多可能有 $k_i(k_i-1)/2$ 条边，实际存在的边数为 E_i 。聚类系数$C_i=2E_i/k_i(k_i-1)$ 。
  由其几何特点等价为 $C_i=\frac{\text{与点}i\text{相连的三角形的数量}}{\text{与点}i\text{相连的三元组的数量}}$。
  整个网络的聚类系数 C 就是所有节点 i 的聚类系数 C_i 的平均值。



以节点 i 为顶点之一的三元组的两种可能形式

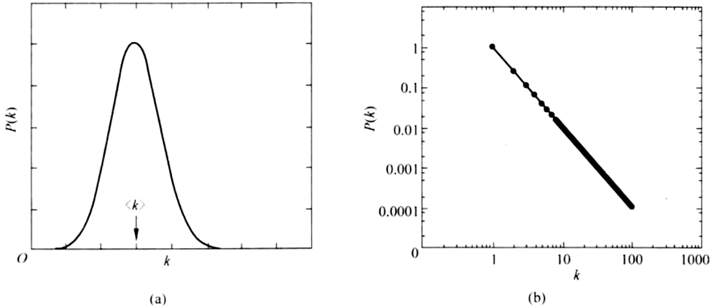
三个指标

- **度与度分布**
  节点 i 的度定义为与该节点连接的其他节点的数目。出度 (out-degree)、入度 (in-degree)。
  度越大从某种意义上意味着该节点越重要。
  所有节点 i 的度 k_i 的期望称为网络的 (节点) 平均度，记为 $<k>$。网络中节点的度分布可用分布函数 $P(k)$ 来描述。

  - **无标度条件**
    对概率分布函数 $f(x)$，对任意常数 a，存在常数 b 使 $f(x)$ 满足“无标度条件”：$f(ax) = bf(x)$，那么一定有：
    $f(x) = f(1)x^{-\gamma}, \gamma = -f(1)/f'(1)$。

    幂律分布函数 $P(k) \propto x^{-\gamma}$ 是唯一满足“无标度条件”的概率分布函数。

三个指标



两种分布的对比

（a）Poisson 分布；（b）幂律分布（对数坐标系）

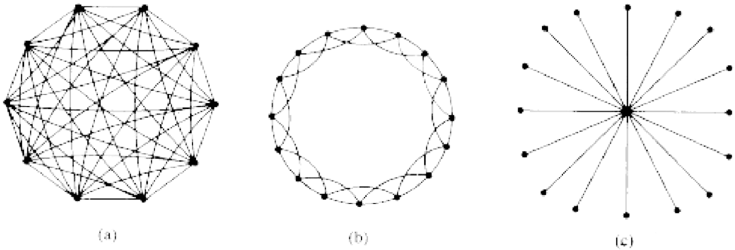
- 均匀网络：公路网等
- 非均匀网络：航空网、Internet 等

规则网络

- **全局耦合 (coupled) 网络：** 任意两点均有边相连。平均路径长度 $L_{gc} = 1$(最小)，聚类系数 $C_{gc} = 1$(最大)。

- **最近邻耦合网络：** 每个节点只和周围的邻居节点相连。
具有周期边界条件的最近邻耦合网络，N 个点围成一个环，每个节点与左右 $K/2$ 个点相连，K 是偶数。对较大的 K 值，聚类系数 $C_{nc} = \frac{3(K-2)}{4(K-1)} \approx \frac{3}{4}$，平均路径长度
$L_{nc} \approx \frac{N}{2K} \to \infty \ (N \to \infty)$。

- **星形耦合网络：** 有一个中心点，其余 $N-1$ 个点只与中心点相连。
聚类系数 $C_{star} = \frac{N-1}{N} \to (N \to \infty)$，平均路径长度
$L_{star} = 2 - \frac{2(N-1)}{N(N-1)} \to 2 (N \to \infty)$。



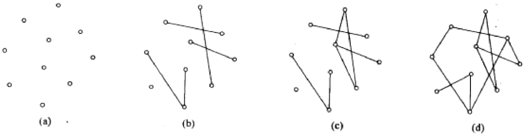
几种规则网络

（a）全局耦合网络；（b）最近邻耦合网络；（c）星形网络

随机网络

- **ER 随机网络** (Erdős-Rényi model,1959)
  N 个点，以概率 p 在两个点之间连线而成的图。

  平均度 $<k>= p(N-1) \approx pN$，二项分布 $\rightarrow$ 泊松分布 $(N \rightarrow \infty)$。平均路径长度为网络规模的对数增长函数，$L_{ER} \propto \ln N / \ln <k>$(小世界特性)。
  聚类系数 $C = p = <k> / N \ll 1$。
  与实际复杂网络模型相比存在明显缺陷。



随机图的演化示意图

(a) $p=0$，给定的 10 个孤立点；(b)～(d) 分别以连接概率 $p=0.1$、$p=0.15$ 和 $p=0.25$ 生成的随机图

小世界网络

- **小世界网络**
  具有较短平均路径长度的同时，又具有较高的聚类系数的一类网络。

WS 小世界模型 (Watts & Strogtz,1998)

(1) 从规则图开始：上文所提到的最近邻耦合网络。环形，每个节点与左右相邻的 $K/2$ 个节点相连。

(2) 随机化重连：以概率 p 随机的重新连接网络中的每条边，即边的一个端点保持不变，另一个节点在网络中随机选取。

$p=0$ 时，没有不同。p 较小时，新的网络与原始网络局部属性差别不大，从而网络的聚类系数变化也不大，平均路径长度却下降很快。上文提到的小世界网络性质。

小世界网络

WS 小世界模型构造算法中的随机化过程有可能破坏网络的连通性。

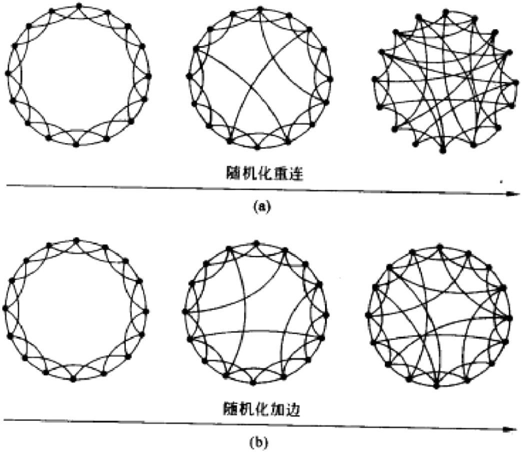
NW 小世界模型 (Newman & Watts,2000)

(1) 从规则图开始：上文所提到的最近邻耦合网络。环形，每个节点与左右相邻的 $K/2$ 个节点相连。

(2) 随机化加边：以概率 p 在在随机选取的一对节点之间加上一条边。任意两各节点之间至多有一条边，无自环。

小世界网络反映了朋友关系网络的一些特性，大部分人的朋友都是其邻居或单位同事，但也有一些人住得很远。

小世界网络



小世界网络模型

(a) WS小世界模型；(b) NW小世界模型

统计性质

- 聚类系数
  WS 小世界网络：$C(p) = \frac{3(K-2)}{4(K-1)}(1-p)^3$
  NW 小世界网络：$C(p) = \frac{3(K-2)}{4(K-1)+4Kp(p+2)}$
- 平均路径长度
  没有精确的显示表达式，但有一些近似表达式。大体上，p K 确定的情况下，$L(p) \propto \ln N$。
- 度分布
  大体上，二项分布，极限情况下逼近泊松分布，是所有节点的度都大致相等的均匀网络。

有一些利用小波分析进行小世界网络分析的研究。

小世界网络

- 六度分离理论
  世界上任何互不相识的两人，只需要很少的中间人就能够建立起联系。
- Kevin Bacon 游戏
- Erdős 数

无标度网络

ER 随机图和 WS 小世界模型的度分布可近似用 Poisson 分布表示 (均匀网络或指数网络)。

许多复杂网络 (Internet、新陈代谢网络) 等的度分布具有幂律形式。这类网络节点的连接度没有明显的特征长度，故称为无标度网络。

为了解释幂律分布的产生机理，提出了 BA 模型。
两个重要特性：

- 增长
  网络规模不断扩大。
- 优先连接
  新的节点趋向于与那些具有较高连接度的节点相连接，“马太效应”。

BA 无标度网络

BA 无标度模型构造算法 (Barabási & Albert,2000)

(1) 增长：从一个具有 m_0 个节点的网络开始，每次引入一个新的节点连接到 m 个已存在的节点上，$m \leq m_0$。

(2) 优先连接：一个新节点与一个已经存在的节点 i 相连接的概率 $\prod_i$ 与节点 i 的度 k_i、节点 j 的度 k_j 之间满足如下关系：

$$\prod_i = \frac{k_i}{\sum_j k_j}$$

BA 无标度网络的演化($m = m_0 = 2$)

BA 无标度网络

- 平均路径长度：$L \propto \frac{\log N}{\log\log N}$。具有小世界特性。
- 聚类系数：没有明显的聚类特征。
- 度分布：大体满足幂律分布。

无标度网络的鲁棒性与脆弱性。

BA 无标度网络中，越老的节点具有越高的度。实际网络中并非如此，有些节点由于自身的特殊性质，更容易被新新加入的节点所连接。如个人的交友能力和科研论文的质量等等。

适应度模型

适应度模型构造算法 (Bianconi & Barabási,2001)

(1) 增长：从一个具有 m_0 个节点的网络开始，每次引入一个新的节点连接到 m 个已存在的节点上，$m \le m_0$。每个节点的适应度按概率分布 $\rho(\eta)$ 选取。

(2) 优先连接：一个新节点与一个已经存在的节点 i 相连接的概率 $\prod_i$ 与节点 i 的度 k_i、节点 j 的度 k_j 和适应度 η_i 之间满足如下关系：

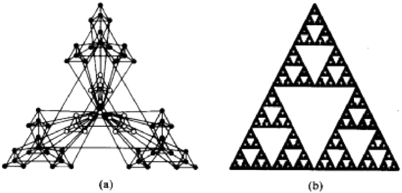
$$\prod_i = \frac{\eta_i k_i}{\sum_j \eta_i k_j}$$

适应度模型中，假如年轻的节点具有较高的适应度，在后续演化过程中会获得更多的边。

适应度分布的性质不同，适应度模型会有不同的行为表现。

复杂网络的自相似性

局部在某种意义上与整体相似，就是自相似性。是分形的基本特征。



等级网络与 Sierpinski 三角形
（a）等级网络；（b）Sierpinski 三角形

盒计数法：用边长 l_B 的盒子来完全覆盖一个几何图形，需要的最少盒子数为 $N_B(l_B)$。图形的维数近似计算公式为：$d_B \approx -\frac{\ln N_B(l_B)}{ln(l_N)}(l_B \to 0)$。分割方式难以寻找。

自组织与吸引子

- **自组织**
  是一系统内部组织化的过程，通常是一开放系统，在没有外部来源引导或管理之下会自行增加其复杂性。
  自组织是从最初的无序系统中各部分之间的局部相互作用，产生某种全局有序或协调的形式的一种过程。这种过程是自发产生的，它不由任何中介或系统内部或外部的子系统所主导或控制。

- **吸引子**一个系统有朝某个稳态发展的趋势，这个稳态就叫做吸引子。吸引子分为平庸吸引子和奇异吸引子。
  如钟摆系统有一个平庸吸引子，其使钟摆系统向停止晃动的稳态发展。
  学术上并没有完善的定义，仅处于概念阶段。

贝叶斯网络

邵逸岚

贝叶斯公式：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

后验概率 =(相似度 * 先验概率)/标准化常量

贝叶斯网络 (Bayesian Network) 是一种经典的概率图模型，它借助有向无环图 (Directed Acyclic Graph, DAG) 来刻画属性之间的依赖关系，并使用条件概率表 (Conditional Probability Table, CPT) 来描述属性的**联合**概率分布。

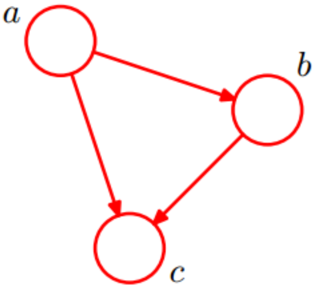


图 4: 贝叶斯网络结构图

贝叶斯网络 B 由结构 G 和参数 Θ 构成，即 $B = \langle G, \Theta \rangle$。给定父结点集，假设每个属性与它的非后裔属性独立，于是将属性 $x_1, x_2, ..., x_d$ 的联合概率分布定义为:

$$P_B(x_1, x_2, ..., x_d) = \prod_{i=1}^{d} P_B(x_i|\pi_i) = \prod_{i=1}^{d} \theta_{x_i|\pi_i}$$

以4中的网络结构为例，联合概率分布可定义为：

$$P(a, b, c) = P(a)P(b|a)P(c|a, b)$$

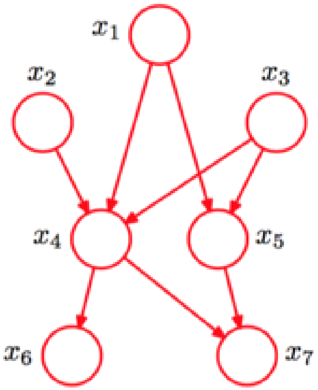


图 5: 较复杂的贝叶斯网络

其联合概率分布可定义为：

$$\begin{aligned} P(x_1,...,x_7) = {} & P(x_1)P(x_2)P(x_3)P(x_4|x_1,x_2,x_3) \\ & p(x_5|x_2,x_3)P(x_6|x_4)P(x_7|x_4,x_5) \end{aligned}$$

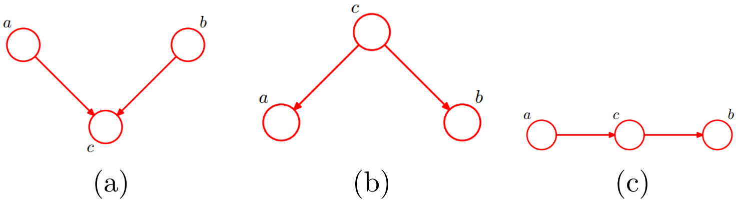


图 6: 贝叶斯网络中属性的典型依赖关系

$$(\mathrm{a})\sum_c P(a,b,c) = \sum_c P(a)P(b)P(c|a,b)$$
$$\Rightarrow P(a,b) = P(a)P(b)$$

$$(\mathrm{b})P(a,b,c) = P(c)P(a|c)P(b|c)$$
$$\Rightarrow P(a,b|c) = P(a|c)P(b|c)$$

$$\begin{aligned}(\mathrm{c})P(a,b|c) &= P(a,b,c)/P(c)\\ &= P(a)P(c|a)P(b|c)/P(c)\\ &= P(a,c)P(b|c)/P(c)\\ &= P(a|c)P(b|c)\end{aligned}$$

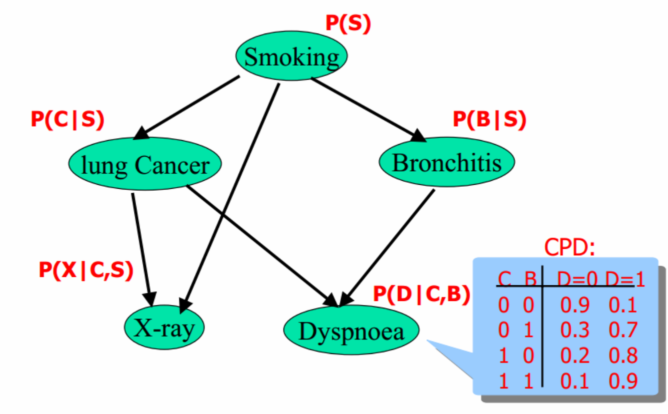

P(S)
Smoking
P(C|S)
lung Cancer
P(B|S)
Bronchitis
P(X|C,S)
X-ray
Dyspnoea
P(D|C,B)
CPD:
C B D=0 D=1
0 0 0.9 0.1
0 1 0.3 0.7
1 0 0.2 0.8
1 1 0.1 0.9

图 7: 贝叶斯网络实例

在给定 x_i 的条件下，x_{i+1} 的分布与 $x_1, x_2, ..., x_{i-1}$ 无关，即 x_{i+1} 的分布只与 x_i 有关。这种顺次演变的随机过程，叫做马尔科夫链。马尔科夫链是贝叶斯网络的一种特例。

图 8: 马尔科夫链结构

马尔科夫链可表示为：

$$P(X_{n+1}=x|X_0,X_1,...,X_n)=P(X_{n+1}=x|X_n)$$

随机变量 $X_1, X_2...$ 取值范围的合集成为“状态空间”，X_i 的值表示在时间 i 的状态。马尔科夫链是时间和状态都是离散的马尔科夫过程。

如果状态空间是有限的，则转移概率分布可以表示为一个具有 (i,j) 元素的矩阵，称之为“转移矩阵”**P**：

$$P_{ij} = P(X_{n+1} = i | X_n = j)$$

对于一个离散状态空间，k 步转移概率的积分即为求和，可以对转移矩阵求 k 次幂来求得。$\mathbf{P}^k$ 即是 k 步转移后的转移矩阵。

平稳分布是一个满足以下方程的向量:

$$\mathbf{P}\pi^* = \pi^*$$

在此情况下，稳态分布 π^* 是一个对应于特征根为 1 的、该转移矩阵的特征向量。

马尔科夫场

王方正

马尔科夫随机场，又称为概率无向图模型，是一个可以由无向图表示的联合概率分布。

图

图是由结点及连接结点的边组成的集合。结点和边分别记作 v 和 e，结点和边的集合分别记作 V 和 E，图记作 $G=(V,E)$。

无向图是指边没有方向的图。

概率图模型

概率图模型（PGM）是由图表示的概率分布。设有联合概率分布 $P(Y), Y\in\mathcal{Y}$ 是一组随机变量。由无向图 $G=(V,E)$ 表示概率分布，即在图 G 中，结点 $v\in V$ 表示一个随机变量 Y_v，$Y=(Y_v)_{v\in V}$；边 $e\in E$ 表示随机变量之间的概率依赖关系。

马尔可夫性：三者等价

- 成对马尔可夫性:$P(Y_u, Y_v|Y_O) = P(Y_u|Y_O)P(Y_v|Y_O)$
- 局部马尔可夫性:$P(Y_v, Y_O|Y_W) = P(Y_v|Y_W)P(Y_O|Y_W)$
- 全局马尔可夫性:$P(Y_A, Y_B|Y_C) = P(Y_A|Y_C)P(Y_B|Y_C)$

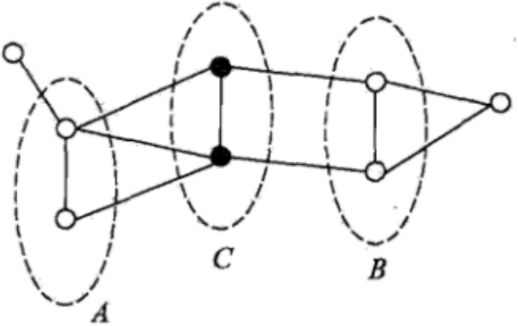


图 9: 局部马尔可夫性

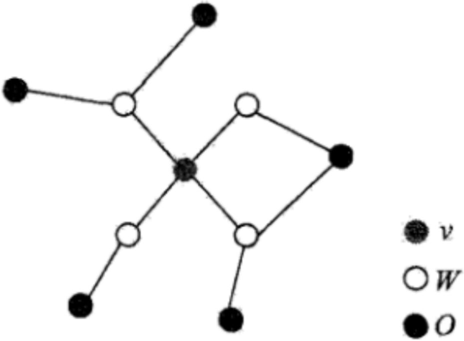


图 10: 全局马尔可夫性

概率无向图模型

设有联合概率分布 $P(Y)$ ，由无向图 $G=(V,E)$ 表示，在图 G 中，结点表示随机变量，边表示随机变量之间的依赖关系。如果联合概率分布 $P(Y)$ 满足成对、局部或全局马尔可夫性，就称此联合概率分布为概率无向图模型或马尔可夫随机场。

团

无向图 G 中任何两个结点均有边连接的结点子集称为团。

最大团

若 C 是无向图 G 的一个团，并且不能再加进任何一个 G 的结点使其成为一个更大的团，则称此 C 为最大团。

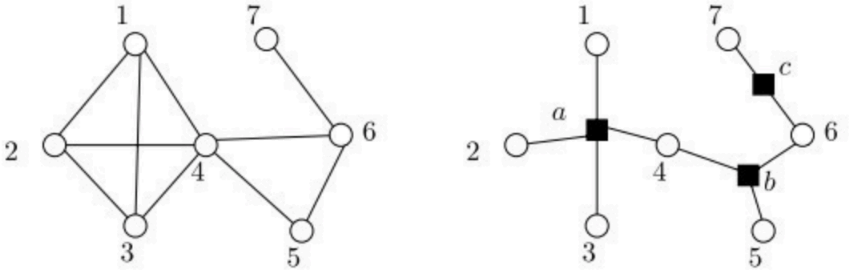


图 11: 无向图与因子图

将概率无向图模型的联合概率分布表示为其最大团上的随机变量的函数的乘积形式的操作，称为概率无向图模型的因子分解。具体公式为：

$$P(Y)=\frac{1}{Z}\prod_{C}\Psi_C(Y_C)$$

其中，Z 是规范化因子，由式

$$Z=\sum_{Y}\prod_{C}\Psi_C(Y_C)$$

给出。规范化因子保证 $P(Y)$ 构成一个概率分布。

函数 $\Psi_C(Y_C)$ 称为势函数，通常定义为指数函数：

$$\Psi_C(Y_C)=exp^{-E(Y_C)}$$

势函数的作用是刻画变量之间的相关关系，它是非负函数，并且在所偏好的变量取值上有较大函数值。

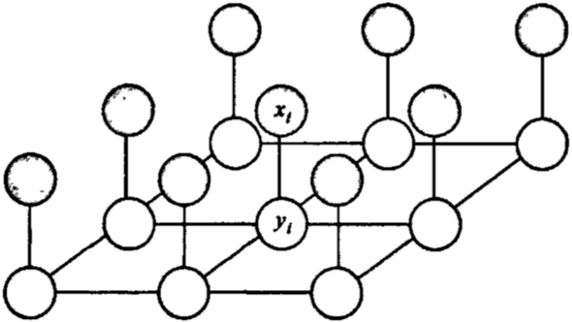


图 12: Pairwise MRF 模型