社交网络分析

基本概念

1. 社交网络分析（Social Network Analysis, SNA）：指基于信息学、数学、社会学、管理学、心理学等多学科的融合理论和方法，为理解人类各种社交关系的形成、行为特点分析以及信息传播的规律提供的一种可计算的分析方法。
   - 特点：SHA主要**关注交互**，而不是个体行为。
   - 应用：SHA可用于**分析网络的配置**如何影响个体和群体、组织或系统功能。
2. 研究技术：（1）节点排序；（2）链路预测；（3）信息传播；（4）社交推荐；（5）舆情分析；（6）隐私保护；（7）用户画像；（8）可视化。

节点排序

3. **重要结点**：相比网络中其他节点而言，能够在更大程度上影响网络结构特征与功能的一些特殊结点。
4. 网络特征结构：（1）**度分布**；（2）**平均距离**；（3）**连通性**；（4）**聚类系数**；（5）度相关性。
5. 节点排序主要方法：（1）基于**结点近邻**的排序方法；（2）基于**路径**的排序方法；（3）基于**特征向量**的排序方法。

基于结点近邻的排序方法

度中心性

6. 度中心性：只考察结点的直接邻居数目。
   - 思路：**认为一个结点的邻居数目越多，其影响力就越大**，这是网络中刻画结点重要性最简单的指标。
   - 在有向图中，度中心性需同时考虑结点的**入度**与**出度**。

k-壳分解法

7. k-壳分解法：用于确定网络中结点的位置。此方法将外围的结点层层剥去，找出处于内存的结点。
    - 思路：**处于网络内层的结点拥有较高的影响力。**
    - 算法步骤：
        1. 假设网络中不存在度为0的孤立结点。从度指标的角度分析，**度数为1的结点是网络中最不重要的结点。**因此,首先将度为1的结点及其连边从网络中删除；
        2. 删除后，网络中将出现新的度为1的结点，接着将这些新出现的度为1的结点及其连边删除；
        3. 重复上述操作，直至不再出现度为1的结点为止；此时，所有被删除结点构成**第一层即1-shell**，结点的Ks值为1；在剩下的网络中，每个结点的度数至少为2；
        4. 重复上述删除操作，得到$Ks=2$的第二层，即2-shell；
        5. 依此类推，直到网络中所有的结点都获得Ks值。

基于路径的排序方法

8. 接近中心性（Closeness Centrality）：通过计算结点与其他所有结点距离的平均值，来消除特殊值的干扰。该平均距离越小，结点的接近中心性越大。
    - 思路：**利用信息在网络中的平均传播时长来确定结点的重要性。**
    - 问题：可能出现高接近中心性的边缘节点。
    - 改进：将结点i到其他所有节点的**距离的倒数和**，作为该结点的接近中心性：

$$C(i)=\sum_{j=1}^{n}\frac{1}{d_{ij}}$$

9. Katz中心性：不仅考虑节点队之间的最短路径，**还考虑它们之间的其他非最短的连通路径。**

> 一个与结点v_i相距p个步长的结点，对v_i的Katz中心线贡献权重为s^p，其中$s\in(0,1)$为固定参数。设p_{ij}为从结点v_i到v_j经过长度为p的路径数目，可得到一个描述网格中任意节点对之间路径关系的矩阵$A=\{p_{ij}\}$，则结点的Katz中心性为：
>
> $$K=sA+s^2A^2+\cdots+s^pA^p+\cdots=(I-sA)^{-1}-I$$

10. 介数中心性（Betweenness Centrality）：一般指最短路径介数中心性（Shortest Path BC）。指的是结点充当某两个结点之间最短路径的中介结点的次数。
    - 思路：**网络中所有结点对的最短路径中，经过一个结点的最短路径越多，该节点就越重要。**

> 形式化定义：
>
> $$C_B(i)=\sum_{s\neq i\neq t,s<t}\frac{\sigma_{s,t}(i)}{\sigma_{s,t}}$$
>
> 其中，$\sigma_{s,t}$表示v_s到v_t的所有最短路径的数目；$\sigma_{s,t}(i)$为从v_s到v_t的最短路径中经过v_i的最短路径数目。

基于特征向量的排序方法

11. 基于特征向量的排序方法：同时考虑了**邻居结点的数量和其质量。**
    - 主要方法：（1）计算**特征向量中心性**；（2）**PageRank算法。**

链路预测

12. 链路预测：根据某一时刻可用的结点及结构信息，来**预测结点和结点之间出现链路的概率。**
    - 任务：（1）**预测新链路将在未来出现的可能性**；（2）**预测当前网络结构中存在的缺失链路的可能性。**
    - 主要方法：（1）基于结点属性的相似性指标；（2）基于局部信息的相似性指标；（3）基于路径的相似性指标。
13. 基于**结点属性**的相似性指标：前提假设为**两个结点之间的相似性越大**，则他们之间**存在链路的可能性越大。**
14. 基于**局部信息**的相似性指标：根据所观察到的**网络结构**来计算结点之间的相似性。
    - **优先连接指标（PA）**：两结点之间存在边的概率**正比于（等于）两结点度的乘积。**
    - **共同邻居指标（CN）**：两结点之间存在边的概率**正比于（等于）它们的共同邻居数量（结点对之间长度为2的路径数目）。**
    - **AA指标**：为共同邻居赋权值，则AA指标等于**两结点的所有共同邻居的权重之和。**
    - **资源分配指标（RA）**：假设每个结点都有一个资源单元，该结点将这些资源平均分配给它的邻居。无链路的结点对v_x, v_y之间，结点v_x可以通过它们的共同邻居将一些资源分配给结点v_y，因此两者的相似度可定义为结点v_y从结点v_x获得的资源数量。
15. 基于**路径**的相似性指标：
    - 路径指标（Local Path, LP）：考虑**结点间的三阶路径数**，即：

$$S = A^2 + \alpha A^3$$

扩散模型

15. 影响力模型：节点排序（中心性分析）-> **影响力建模。**
16. 扩散模型（Diffusion Influence Model）：每个结点都有一个对应的状态，即**active**或**inactive。**
    - 主要模型：（1）线性阈值扩散模型；（2）独立级联扩散模型；（3）影响力最大化模型。

线性阈值模型

17. **线性阈值模型**：每个结点都有一个**信息传导的阈值。**当一个结点从其邻居接收到的**影响大于其阈值**时，该结点就会传播这条消息。

独立级联模型

18. **独立级联模型**：基于概率论，对信息传播过程的一个**动态描述。**结点v在步骤t转发了信息，那么它有**一次机会**去影响它的每一个邻居也转发这条消息，成功的概率由两者连接边上的权重决定，该权重也称为**传导概率。**若v没能让u在步骤$t+1$转发该条消息，它之后**再没有机会使之转发该消息。**

影响力最大化模型

19. **影响力最大化模型**：基于贪心算法求解。

> 在网络中找到一个**种子集合**$\mathcal{S}$，使得：
>
> $$\mathcal{S} = \backslash\text{argmax}_{\mathcal{S}} f_{\mathcal{S}\to\mathcal{V}}$$
>
> 其中$f_{\mathcal{S}\to\mathcal{V}}$表示$\mathcal{S}$影响的结点数。
>
> 利用贪心算法，将问题分K轮求解，即$\mathcal{S} = K$：起始时种子集合为空集，每轮从网络中选取一个能带来最大影响力增量的结点，加入到种子集合中。令第i轮的种子集合为$\mathcal{S}_i$，则：
>
> $$\mathcal{S}_i = \mathcal{S}_{i-1} \cup s_k$$
> $$s_k = \backslash\text{argmax}_{s\in\mathcal{V}\backslash\mathcal{S}_{k-1}} \Delta_s(\mathcal{S}_{k-1})$$

> $\Delta_S(\mathcal{S}_{k-1}) = (f_{\mathcal{S}_{k-1}\cup\{s\}\to\mathcal{V}} - f_{\mathcal{S}_{k-1}\to\mathcal{V}})$是结点$s$加入$\mathcal{S}_{k-1}$时能带来的**影响力增量**。