# 1.复杂网络常用数据集网站

复杂网络的研究很多都离不开数据集，下面这些是个人在做科研的过程中在互联网上搜集到的一些数据集网站，列举出来也方便同行们去使用。（2018年12月14日更）

1、[http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:data:urls:index](https://link.jianshu.com/?t=http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:data:urls:index" \t "_blank)

数据集网站集合。这个网站中列出来很多数据集网站，非常全，其中的很多网站数据描述详细，而且数据可以直接下载，大家可以从中选出一些网站收藏起来。

2、[http://snap.stanford.edu/data/](https://link.jianshu.com/?t=http://snap.stanford.edu/data/" \t "_blank)

这个是斯坦福大学的大型网络数据集网站，大家应该比较熟悉了。

3、[http://konect.uni-koblenz.de/](https://link.jianshu.com/?t=http://konect.uni-koblenz.de/" \t "_blank)

这个网站是我个人最喜欢也是最常用的网站，里面数据集有上百种，数据的分类和描述真的是特别详细，而且还给出了数据集的可视化图以及一些基本统计特性，所有数据均可以直接下载。

4、[http://networkrepository.com/index.php](https://link.jianshu.com/?t=http://networkrepository.com/index.php" \t "_blank)

这是另一个个人特别喜欢也是很常用的网站，和上面一样数据分类也是相当的详细，而且数据都可以在线可视化，大家应该能找到自己想要的数据。

5、[http://gdm.fudan.edu.cn/GDMWiki/Wiki.jsp?page=Network%20DataSet](https://link.jianshu.com/?t=http://gdm.fudan.edu.cn/GDMWiki/Wiki.jsp?page=Network%20DataSet" \t "_blank)

这是复旦大学建立的网络数据集网站，里面有一些常用数据集以及一些相关资源网站。

6、[https://www.aminer.cn/data-sna#Twitter-Dynamic-Net](https://www.aminer.cn/data-sna" \l "Twitter-Dynamic-Net" \t "_blank)

搞数据挖掘应该都或多或少知道Aminer和唐杰，这是Aminer上的一些社交网络数据集。

7、[https://toreopsahl.com/datasets/](https://link.jianshu.com/?t=https://toreopsahl.com/datasets/" \t "_blank)

这个网站里面有十几个数据集，包含社交网络、交通网络、合作网络等常用数据集。

8、[http://netwiki.amath.unc.edu/SharedData/SharedData](https://link.jianshu.com/?t=http://netwiki.amath.unc.edu/SharedData/SharedData" \t "_blank)

这个网站列出了一些数据集以及一些复杂网络大牛的个人数据网站，大家可以去看看。

9、[http://vlado.fmf.uni-lj.si/pub/networks/data/](https://link.jianshu.com/?t=http://vlado.fmf.uni-lj.si/pub/networks/data/" \t "_blank)

这个是Pajek网站提供的数据集，里面的数据都很经典，复杂网络早期研究中很多数据集都是源于这里。

10、[http://socialcomputing.asu.edu/pages/datasets](https://link.jianshu.com/?t=http://socialcomputing.asu.edu/pages/datasets)

社交网络数据集，里面包含国内外一些常用在线社区网站的数据集，中型到大型的数据规模，搞社会计算的同行可能用的比较多。

11、[http://www.sociopatterns.org/datasets/](https://link.jianshu.com/?t=http://www.sociopatterns.org/datasets/)

另一个社交网络数据集，这里面的数据集更多倾向于实体网络，比如联系网、高校朋友网、疾病传播网。

12、[http://www-personal.umich.edu/~mejn/netdata/](https://link.jianshu.com/?t=http://www-personal.umich.edu/~mejn/netdata/)

大牛Newman教授的个人数据网站，里面的数据集特别是科学家合作网络，用到特别多。

以上是一些个人用到比较多的复杂网络数据网站，鉴于自己的了解有限，可能对这些数据集网站给出的描述还不够详细，而且很多单位都没有列出来，敬请谅解。当然了如果想用里面的数据集，一定要记得引用数据集作者们的信息，他们收集数据并公开方便大家使用也是很不容易的。如果转载本文，也请注明出处^\_^.

补充说明:如果具备以下两个技能，构造复杂网络数据集将会变得简单:

(1) 文本处理(尤其是正则表达式)。因为很多原始数据包含很多冗余信息，可以运用文本处理来提取自己想要的关键信息，并在此基础上构造节点和连边关系。

(2) 网络爬虫。有时候从网络上根本找不到想要的现成数据，这时可以运用网络爬虫去爬取数据来构造复杂网络。

# **2. 复杂网络的统计特征**

## **2.1平均路径长度L**

在网络中，两点之间的距离为连接两点的最短路径上所包含的边的数目。网络的平均路径长度L指网络中所有节点对的平均距离，它表明网络中节点间的分离程度，反映了网络的全局特性。不同的网络结构可赋予L不同的含义。如在疾病传播模型中L可定义为疾病传播时间，通网络模型中L可定义为站点之间的距离等。

## **2.2聚集系数C**

在网络中，节点的聚集系数是指与该节点相邻的所有节点之间连边的数目占这些相邻节点之间最大可能连边数目的比例。而网络的聚集系数则是指网络中所有节点聚集系数的平均值，它表明网络中节点的聚集情况即网络的**聚集性**，也就是说同一个节点的两个相邻节点仍然是相邻节点的概率有多大，它反映了网络的局部特性。

## **2.3度及度分布**

在网络中，点的度是指与该节点相邻的节点的数目，即连接该节点的边的数目。而网络的度<k>指网络中所有节点度的平均值。度分布P(k)指网络中一个任意选择的节点，它的度恰好为k的概率。

1） 节点的度的定义：与节点（node）V相连的边（edge）数之和称为节点的度，记为deg（v），简记为：d（v）

（2） 悬挂点：度为1的节点称为悬挂点

悬挂边：连接悬挂点的边称为悬挂边

（3） 任何图中，节点的度之和等于边数的2倍，次数为奇数的节点必为偶数个。

（4） 出度：在有向图中，以节点vi为起始点的边数称为出度

入度：在有向图中，以节点vi为终止点的边数称为入度

## **2.4介数**

包括节点介数和边介数，一个节点的节点介数指网络中所有最短路径中经过该节点的数量比例，边介数则指网络中所有最短路径中经过该边的数量比例。介数反映了相应的节点或边在整个网络中的作用和影响力。

## **2.5小世界效应**

复杂网络的小世界效应是指尽管网络的规模很大(网络节点数目N很大)，但是两个节点之间的距离比我们想象的要小得多。也就是网络的平均路径长度L随网络的规模呈对数增长，即L～In N。大量的实证研究表明，真实网络几乎都具有小世界效应。

## **2.6无标度特性**

对于随机网络和规则网络，度分布区间非常狭窄，大多数节点都集中在节点度均值<k>的附近，说明节点具有同质性，因此<k>可以被看作是节点度的一个特征标度。而在节点度服从幂律分布的网络中，大多数节点的度都很小，而少数节点的度很大，说明节点具有异质性，这时特征标度消失。这种节点度的幂律分布为网络的无标度特性。

## 2.7 图的各种距离定义

1.the distance between two vertices in a graph is the number of edges in a shortest path (also called a graph geodesic) connecting them. This is also known as the geodesic distance

### 2.7.1距离：

两个节点的最短路径包含的边的数目

2.The eccentricity of a vertex v is the greatest geodesic distance between v and any other vertex. It can be thought of as how far a node is from the node most distant from it in the graph.

### 2.7.2 偏心距（可能是这么翻译）：

节点v的最大距离，即所有其他节点中离得最远的节点到他它的距离

3.The radius r of a graph is the minimum eccentricity of any vertex

### 2.7.3半径：

所有节点的偏心距中最小的值

4.The diameter d of a graph is the maximum eccentricity of any vertex in the graph. That is, d is the greatest distance between any pair of vertices or, alternatively. To find the diameter of a graph, first find the shortest path between each pair of vertices. The greatest length of any of these paths is the diameter of the graph.

### 2.7.4直径：

所有节点中偏心距最大的值

5.A central vertex in a graph of radius r is one whose eccentricity is r—that is, a vertex that achieves the radius

### 2.7.5中心节点：

偏心距等于半径的节点

6.A peripheral vertex in a graph of diameter d is one that is distance d from some other vertex—that is, a vertex that achieves the diameter

### 2.7.6边缘节点（次要节点）：

偏心距等于直径的节点

7.A pseudo-peripheral vertex v has the property that for any vertex u, if v is as far away from u as possible, then u is as far away from v as possible.

### 2.7.7伪边缘节点：

距离某节点偏心距最大的节点

### 2.8 用到的一些指标

Average degree **（）**:

Assortativity **（r）**:

Average local clustering coefficient**（）**

Transitivity**（）**

Average shortest path length**（）**

Second moment of degree distribution**（）**:

# 3.据Li Yang等人的[论文](https://xueshu.baidu.com/s?wd=paperuri%3A%2815ed4fc6e6d609d9058ae4300160e127%29&filter=sc_long_sign&sc_ks_para=q%3DIdentifying%20opinion%20leader%20nodes%20in%20online%20social%20networks%20with%20a%20new%20closeness%20evaluation%20algorithm&sc_us=6245285459093995375&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8)总结了四种衡量网络中一个节点的重要程度的方法：

## 1. Degree Centrality

对无向图来说，节点v的degree就是它的直接邻居节点数量。

## 2. Closeness Centrality

节点v的closeness就是v到其他各个节点的最短路径的长度之和的倒数。也就是说如果v到各个节点的路径越短，则closeness越大，说明v越重要。

点的**近性中心度（Closeness Centrality）**：一个点的近性中心度较高，说明该点到网络中其他各点的距离总体来说较近，反之则较远。假如一个物流仓库网络需要选某个仓库作为**核心中转站**，需要它到其他仓库的距离总体来说最近，那么一种方法就是找到近性中心度最高的那个仓库。

## 3. Betweenness Centrality

节点v的Betweenness 就是图中任意两个节点对之间的最短路径当中，其中经过v的最短路径的所占的比例，也就是说经过v的最短路径越多，v越重要。

点的**介性中心度（Betweenness Centrality）**：一个点的介性中心度较高，说明其他点之间的最短路径很多甚至全部都**必须**经过它中转。假如这个点消失了，那么其他点之间的交流会变得困难，甚至可能断开（因为原来的最短路径断开了）。因此假如要**hack一个网络**的话，对哪个结点下手你懂的。从另一个角度想，这些点简直就像是**等在丝绸之路上必经关口的强盗**。不留下买路钱？让你无路可走，生意就别做了。

这两种中心度我目前并未找到很公认的中文翻译，姑且自己翻译了。另外同PageRank和HITS一样，由于指标的计算稍显复杂，这里就不详细叙述了。但是我们都使用的是网络分析库Networkx中的算法实现，对详细算法有兴趣的读者可自行查阅其文档。



# 4. networkx相关函数

 nx.degree(G)// 计算图的密度，其值为边数m除以图中可能边数（即n(n-1)/2）  
       nx.degree\_centrality(G)//节点度中心系数。通过节点的度表示节点在图中的重要性，默认情况下会进行归一化，其值表达为节点度d(u)除以n-1（其中n-1就是归一化使用的常量）。这里由于可能存在循环，所以该值可能大于1.

       nx.closeness\_centrality(G)//节点距离中心系数。通过距离来表示节点在图中的重要性，一般是指节点到其他节点的平均路径的倒数，这里还乘以了n-1。该值越大表示节点到其他节点的距离越近，即中心性越高。

      nx.betweenness\_centrality(G)//节点介数中心系数。在无向图中，该值表示为节点作占最短路径的个数除以((n-1)(n-2)/2)；在有向图中，该值表达为节点作占最短路径个数除以((n-1)(n-2))。

      nx.transitivity(G)//图或网络的传递性。即图或网络中，认识同一个节点的两个节点也可能认识双方，计算公式为3\*图中三角形的个数/三元组个数（该三元组个数是有公共顶点的边对数，这样就好数了）。

     nx.clustering(G)//图或网络中节点的聚类系数。计算公式为：节点u的两个邻居节点间的边数除以((d(u)(d(u)-1)/2)。

# 5.数学知识

## 1.二维数组和三维数组区别

**一维：**举例来说。把一个数当作一个鸡蛋，那么一个int型的变量就是一个鸡蛋。现在建立一个一维数组int[5]：这个就相当于在你面前放了一排5个鸡蛋。

**二维**：接下来到二维数组int[3][4]：首先，由二维数组的第二维[4]我们可以确定一排是4个鸡蛋，这个跟刚才的一维数组一样。那么之前的[3]是什么呢？于是int[3][4]相当于在你面前放了3排的鸡蛋，一排4个。

**三维：**再到三维数组int[3][4][5]：第三维的[5]就是之前一位数组里的一排5个鸡蛋，第二维的[4]就是一共有4排。那么这个[3]呢？那就是如果把这4排每排5个的鸡蛋看成一个小组，即4\*5=20个鸡蛋是一个组，在你面前就放着3组的鸡蛋。

如下图：

