



计算社会科学导论 —— 复杂网络

金耀辉、许岩岩

2023年4月6日



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

提纲

1

复杂科学简介

2

图的基本概念

3

随机网络

4

无标度网络

5

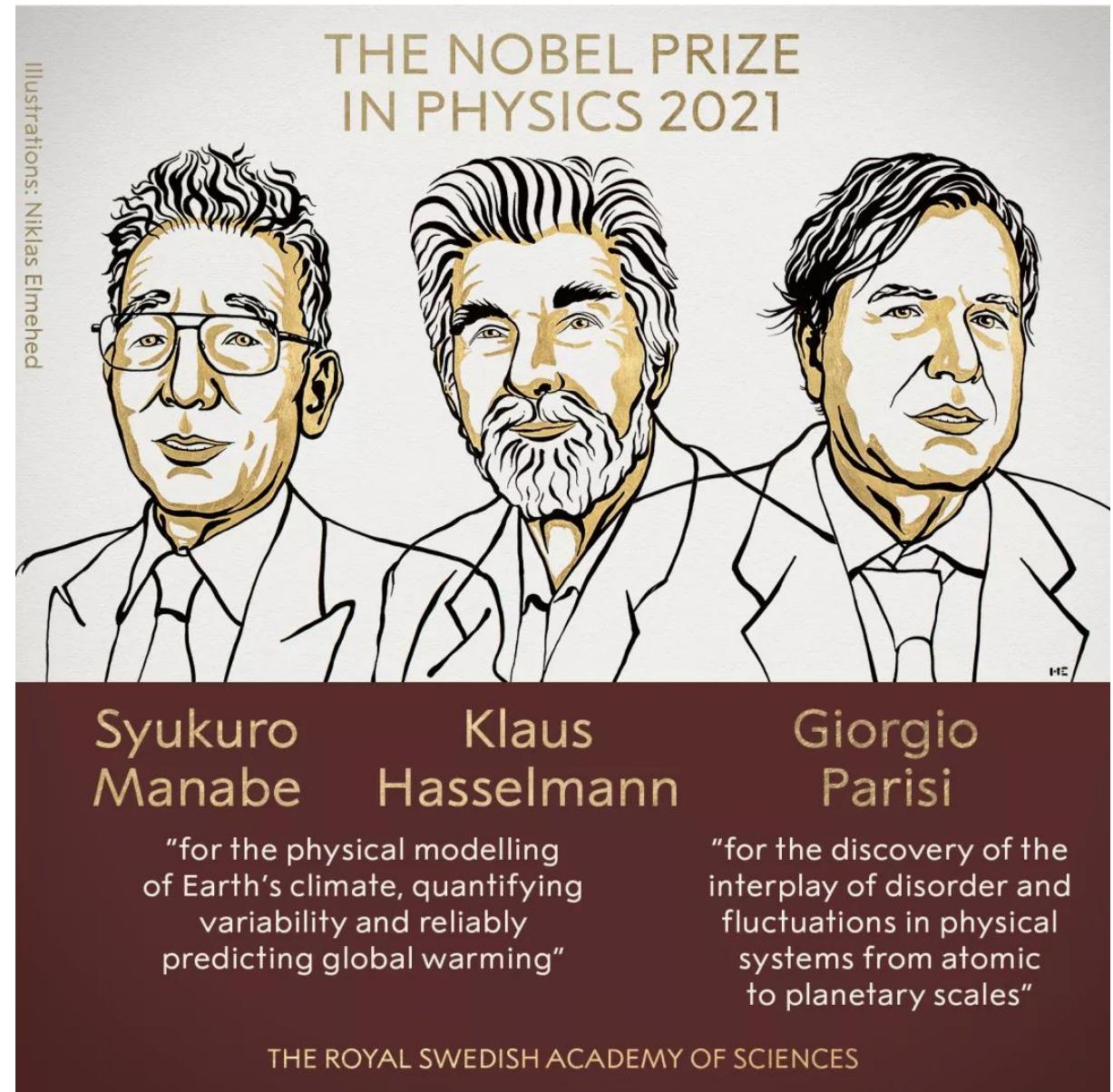
社团检测



北京时间2021年10月5日17时45分许，2021年诺贝尔物理学奖发布，奖项授予了三位物理学家以表彰他们“为我们理解复杂物理系统所做出的开创性贡献”(for groundbreaking contributions to our understanding of complex physical systems)

真锅淑郎(Syukuro Manabe) 和克劳斯·哈塞尔曼(Klaus Hasselmann)，因为对“**地球气候的物理建模，量化可变性并可靠地预测全球变暖**”，共享了诺贝尔物理学奖的一半奖金。

乔治·帕里西(Giorgio Parisi)，因为“**发现了从原子到行星尺度的物理系统中的无序和涨落的相互作用**”，而获得了诺贝尔物理学奖的另一半奖金。

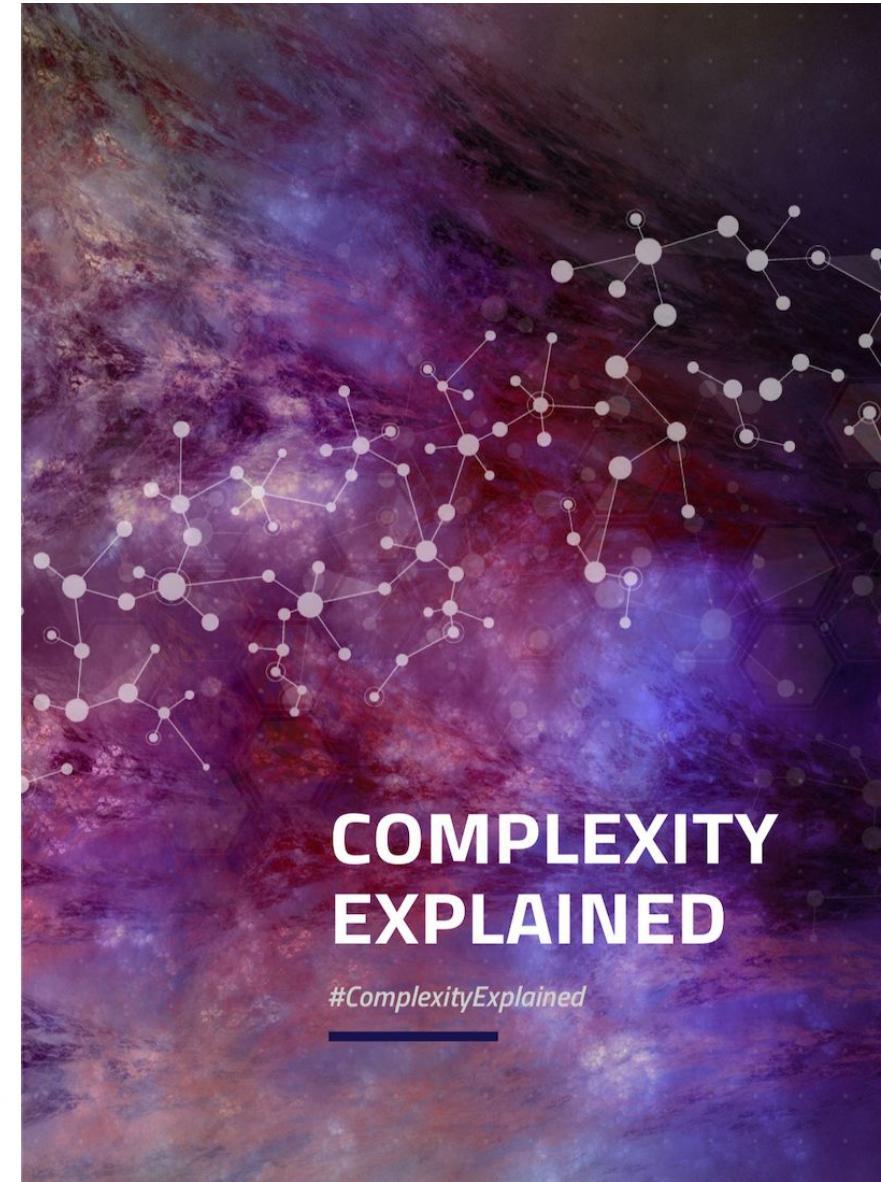




复杂科学(Complexity Science), 亦称复杂系统科学(Complex Systems Science)

研究一个微观上由局部相互作用的单元构成的集体, 如何在缺乏外在干预或中心协调者的情况下, 自发地自我组织, 并展现出某些整体结构与宏观行为。即便我们完全了解每个小单元, 往往无法保证能理解或预测集体的种种性质, 这样的集体成为【复杂系统】，为了研究这种系统, 我们需要新的数学框架和科学方法

钱学森在 2001 年曾说：“23 年来, 系统工程和系统科学已经有了很大发展, 我们已经从工程系统走到了社会系统, 进而提炼出开放的复杂巨系统的理论和处理这种系统的方法论, 即以人为主、人-机结合, 从定性到定量的综合集成法, 并在工程上逐步实现综合集成研讨厅体系。将来我们要从系统工程、系统科学发展到大成智慧工程, 要集信息和知识之大成, 以此来解决现实生活中的复杂问题。”

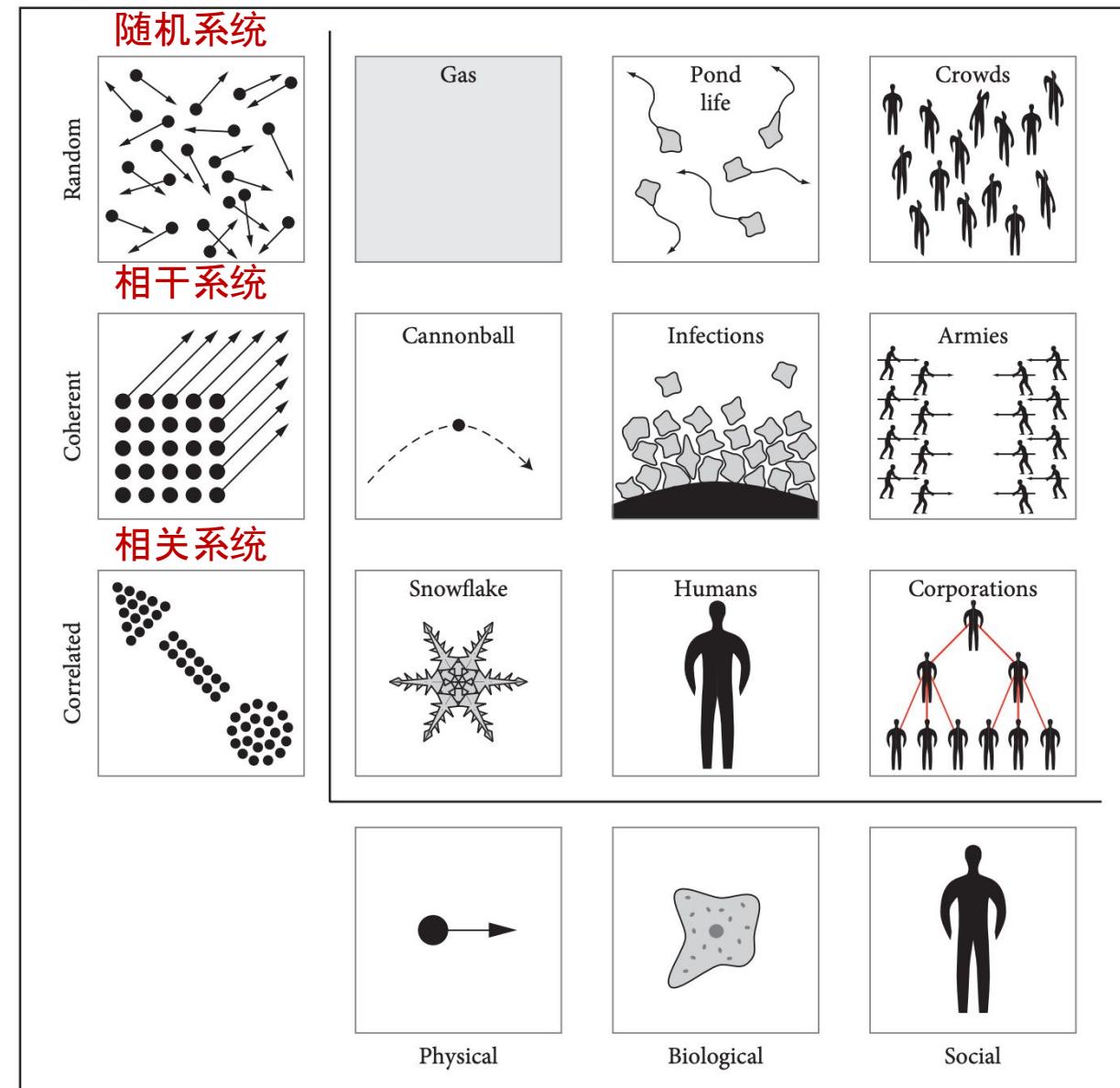


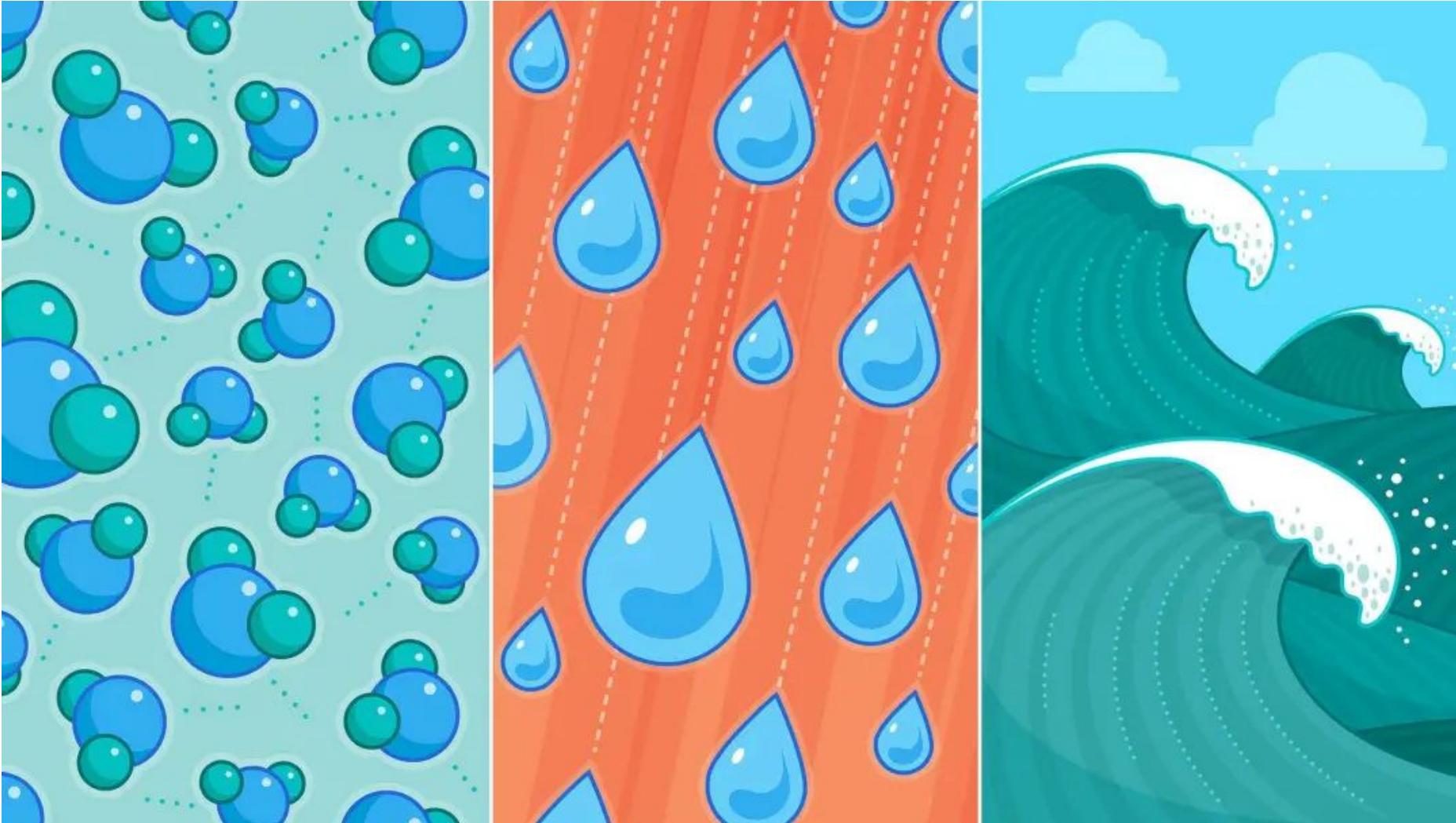
<https://complexityexplained.github.io/>

复杂系统科学研究的对象是包含很多构成组元(components)的系统，范围极广，包括物理系统、生态系统和社会系统等。但不像其他学科那样关注构成系统的组元本身，复杂系统科学关注的是系统中的组元是如何关联起来的。

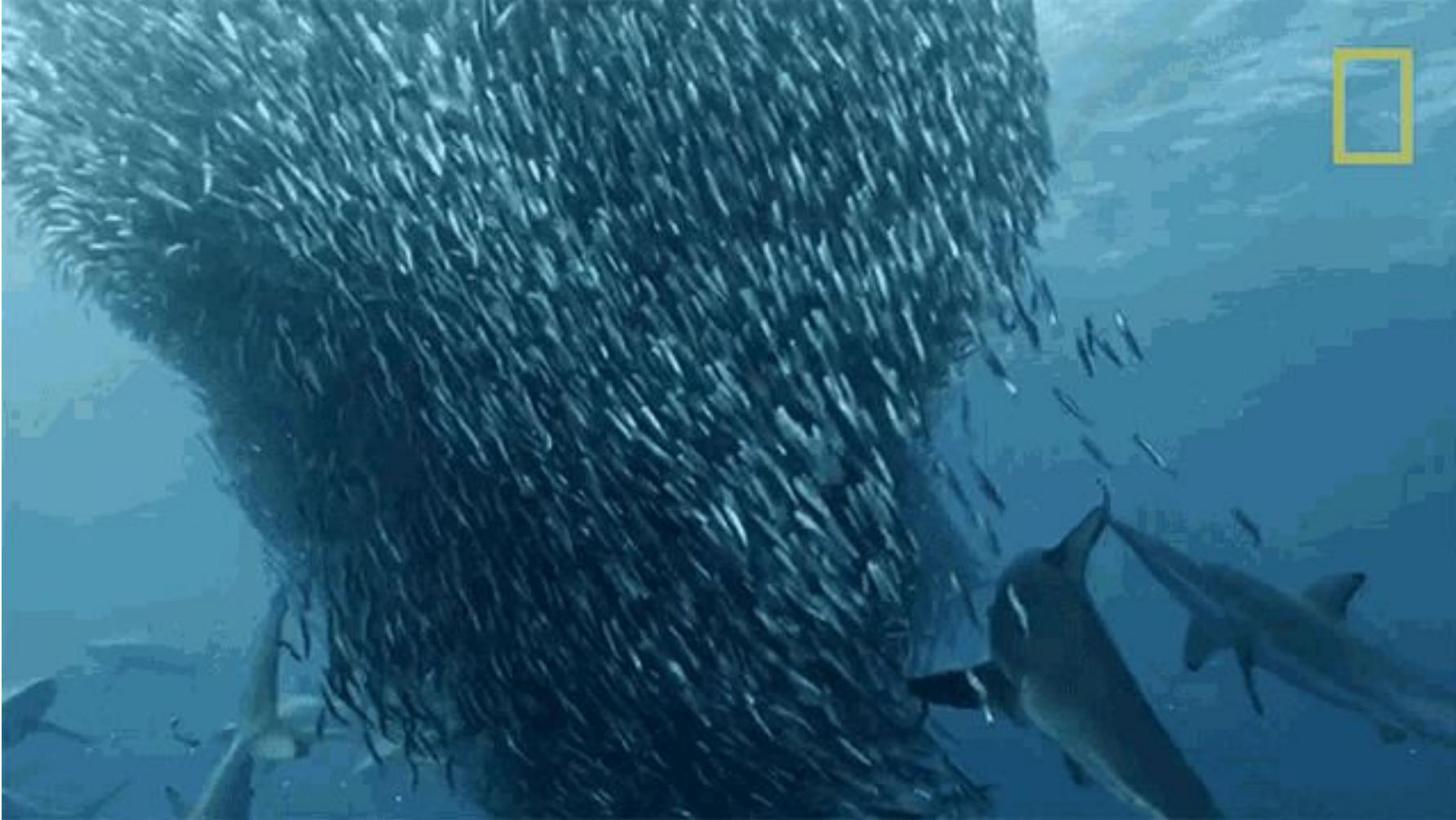
系统的性质主要取决于其组元的关系，而不是组元本身。复杂系统科学的目的是提供统一的科学框架，允许思想的泛化，促使新应用、新连接的发现。

Ref: Siegenfeld, Alexander F., and Yaneer Bar-Yam. "An introduction to complex systems science and its applications." *Complexity* 2020

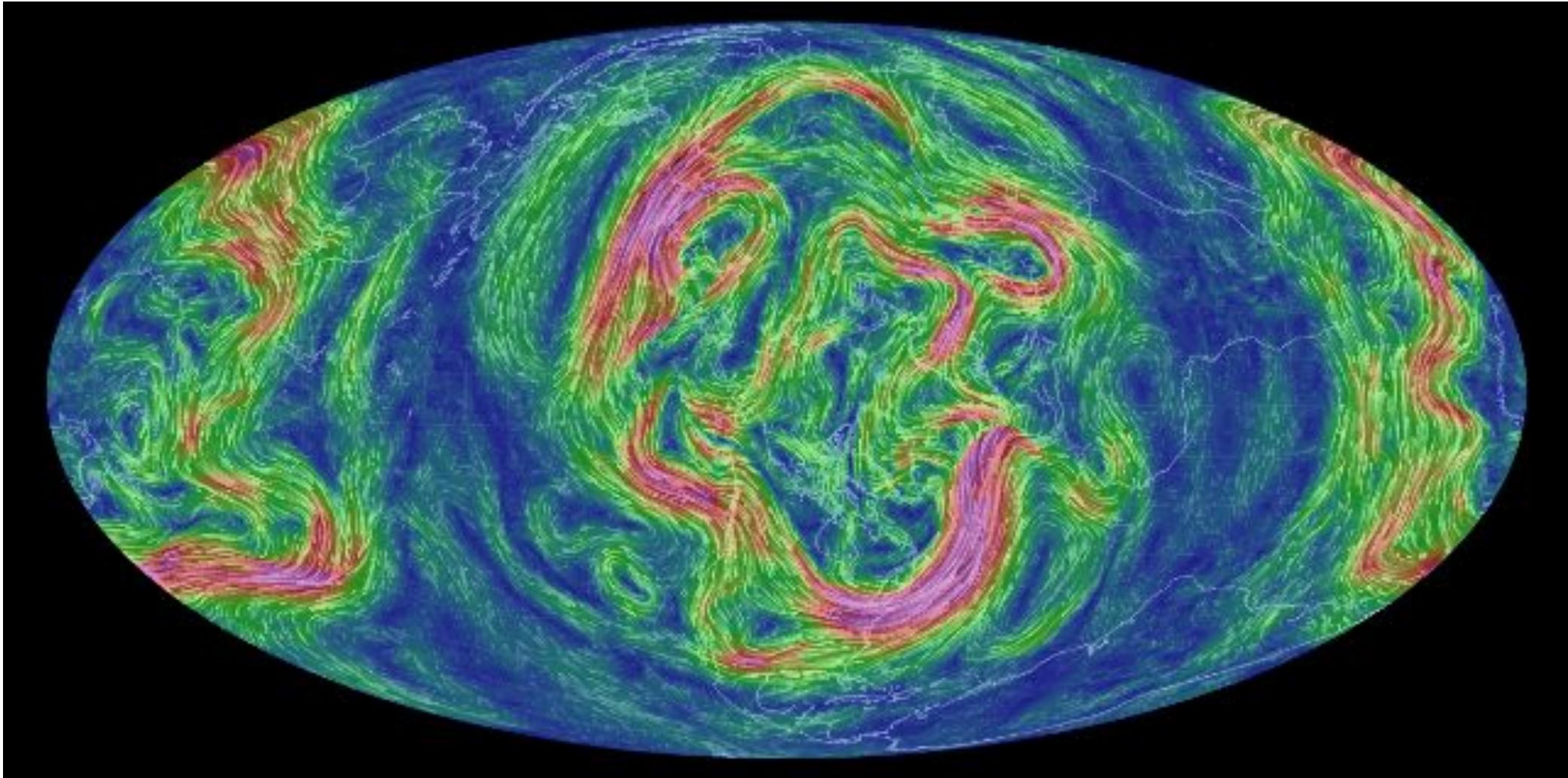




我们不需要分析单个的水分子来理解水滴的行为，也不需要分析水滴来研究水波



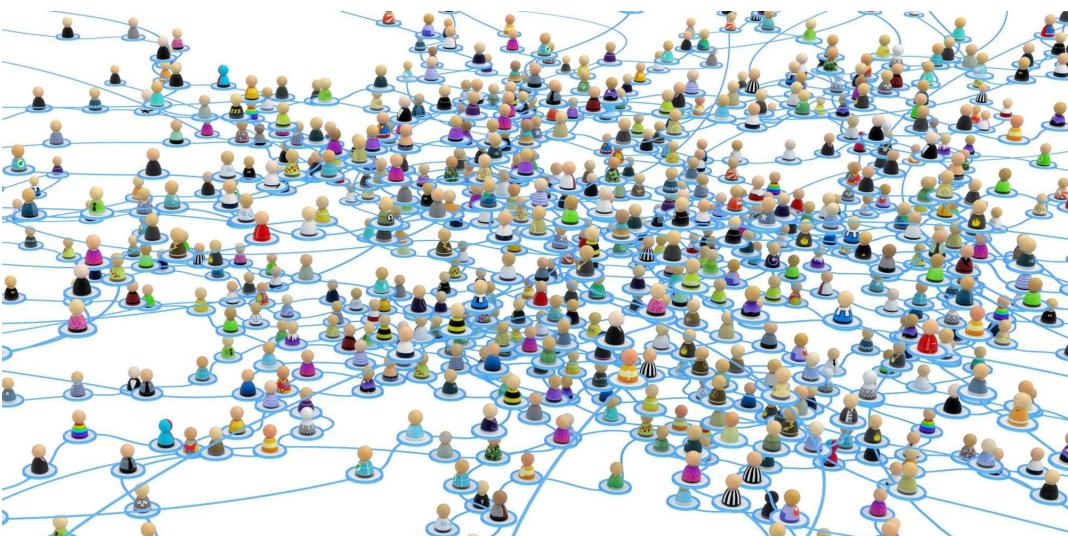
鱼类的群体行为



全球尺度上的风势

什么是**复杂度(Complexity)**

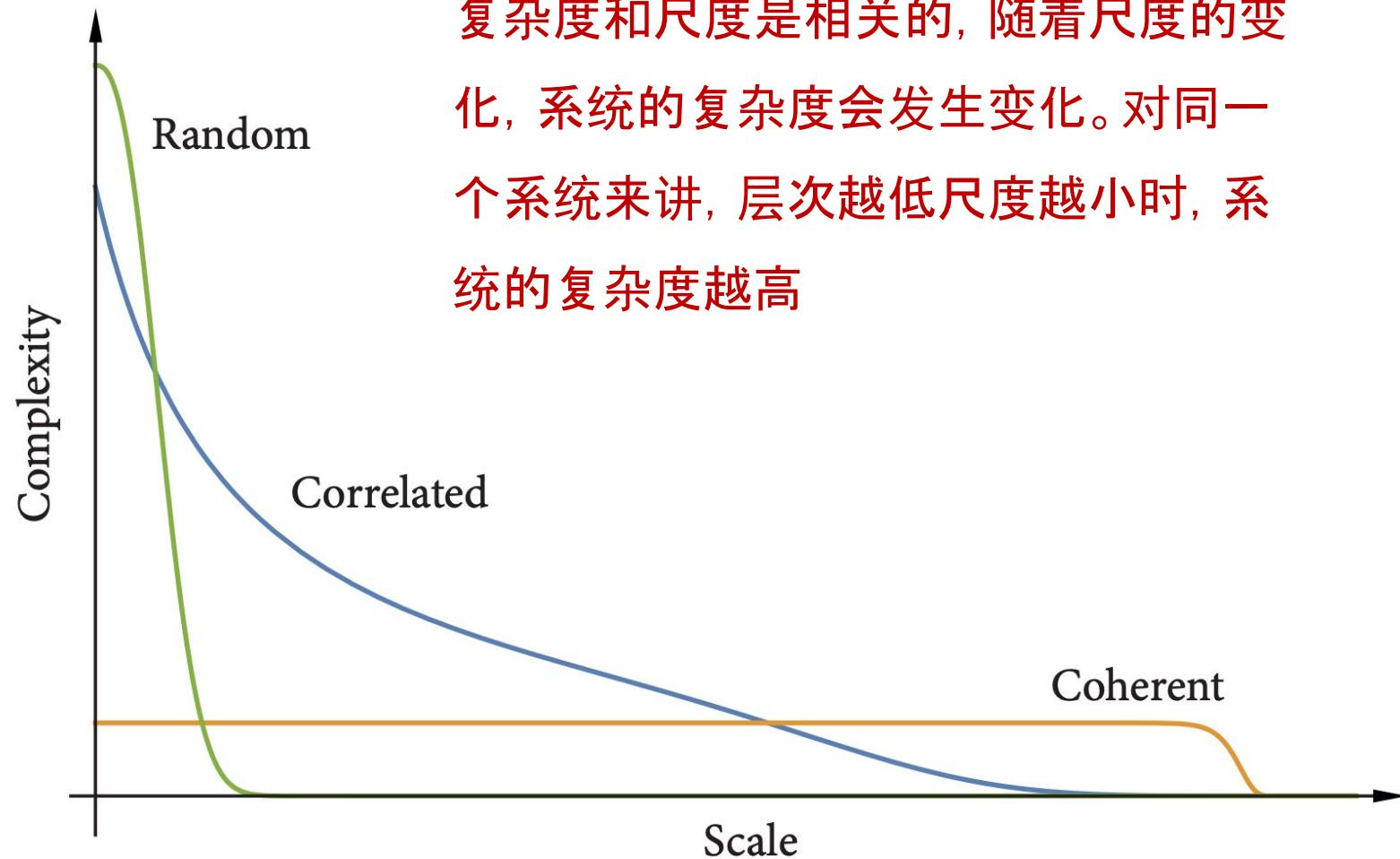
行为的复杂度可以用**描述这个行为的信息长度**(length of its description)来表示, 这个长度取决于**系统可能的状态数量**。例如一个电灯只有两种可能状态——开或关, 这只需要 1 个比特长度的信息就可以表示出来(0 或者 1)。而 2 个比特长度的信息可以用来表示 4 种可能行为, 分别是 00、01、10 和 11。同样的, 可以用 3 个比特的信息来区分 8 种行为。**复杂度可以简单表示为 $C=\log_2 (N)$** , 其中N为可能行为的数量。



**可能的系统行
为数量越多, 系统的复杂度越大**

复杂度依赖于尺度(Scale)

- **随机系统**在微观小尺度上具有大复杂度，在宏观大尺度上复杂度低。
- **相干系统**的复杂度不会随着观察的层次发生变化。
- **相关系统**介于二者之间。
- 所有复杂度随着尺度的减少有增加的趋势。

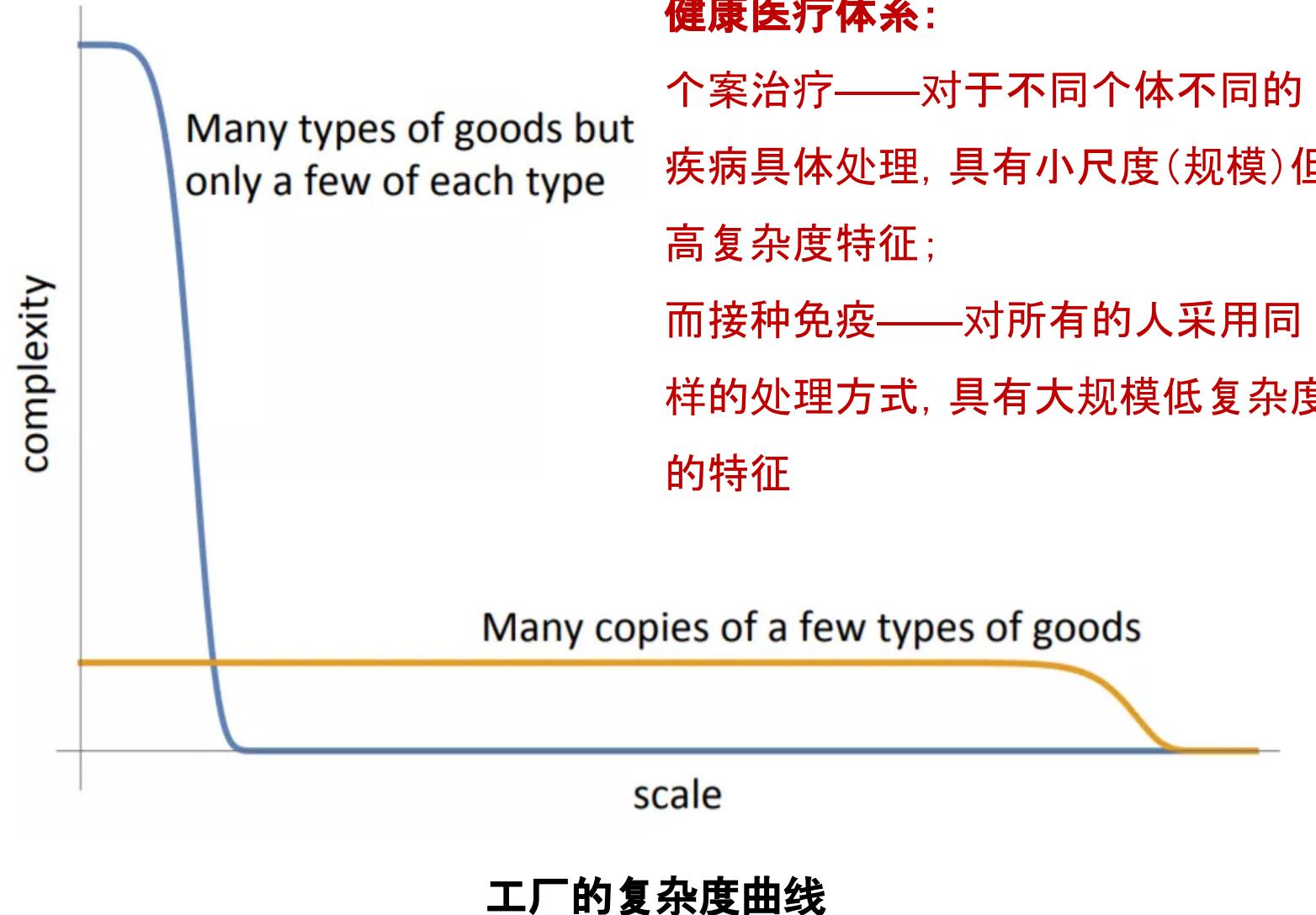


随机、相关和相干系统的复杂度曲线

复杂度、尺度、适应性

复杂度越高，个体行为相对独立，具有更多的行为方式，整个系统会有更大的适应性；

反之，若系统中的很多个体都进行高度协作，可以高效率完成既定任务，满足大规模或者大尺度上的要求，但这种有效系统对于自身或者环境未来不确定变化的适应能力会降低。





相互作用

複雜系統由許多小單元組成，它們用不同的方式或和彼此、或和環境相互作用。



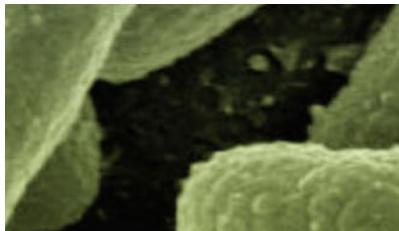
自我組織性

複雜系統可以不需要藍圖的規劃，自我組織並自發地產生不直觀的形態。



突現性質

從整體看複雜系統的性質，往往能意想不到地發現許多和其組成單元不同的特性。



適應

複雜系統會適應環境並演化。



動力學

複雜系統的狀態往往不斷變化，展現出難以預測的長期行為。

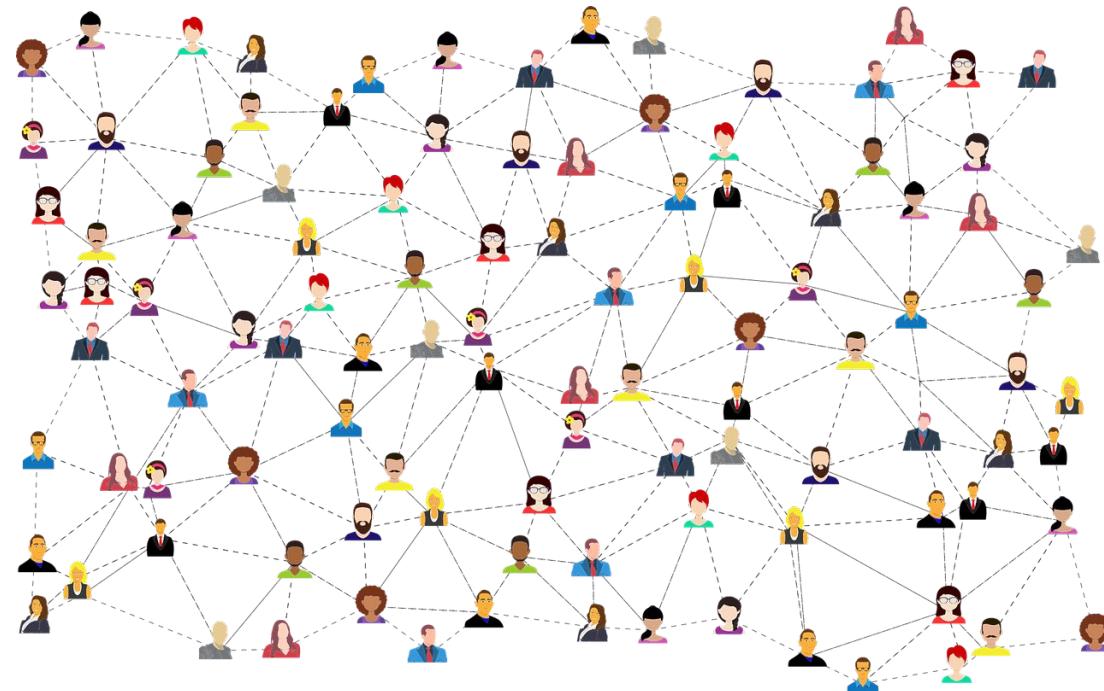
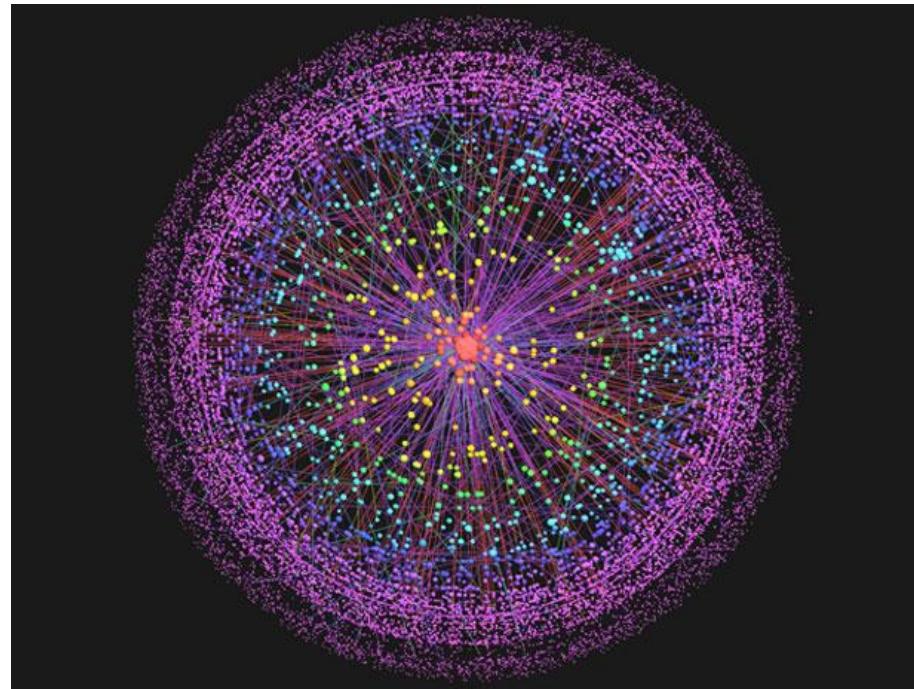


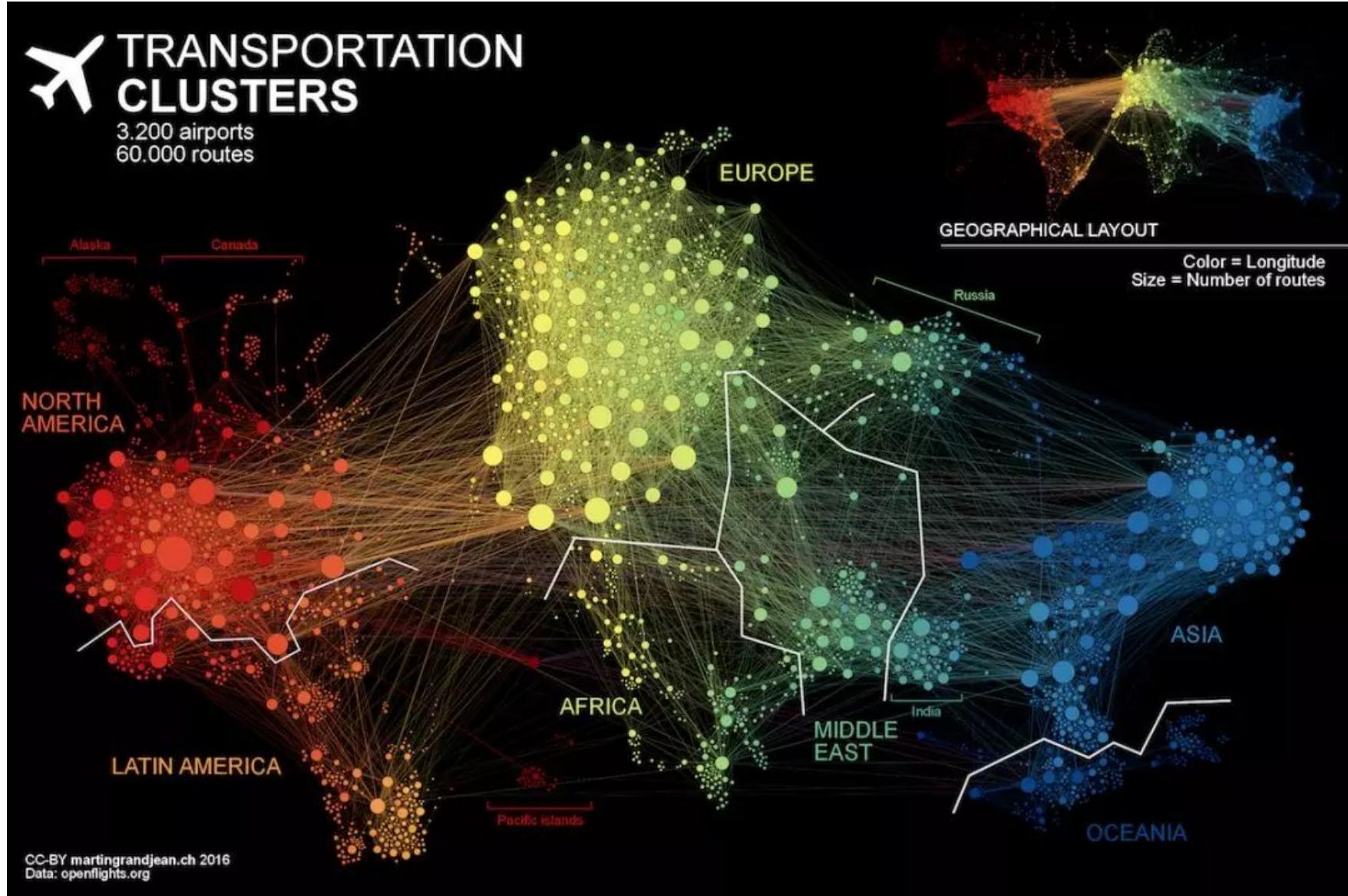
跨領域

複雜系統科學可以用來了解及管理不同領域中形形色色的系統。

“I think the next 21st century will be the century of complexity.” —— Stephen Hawking

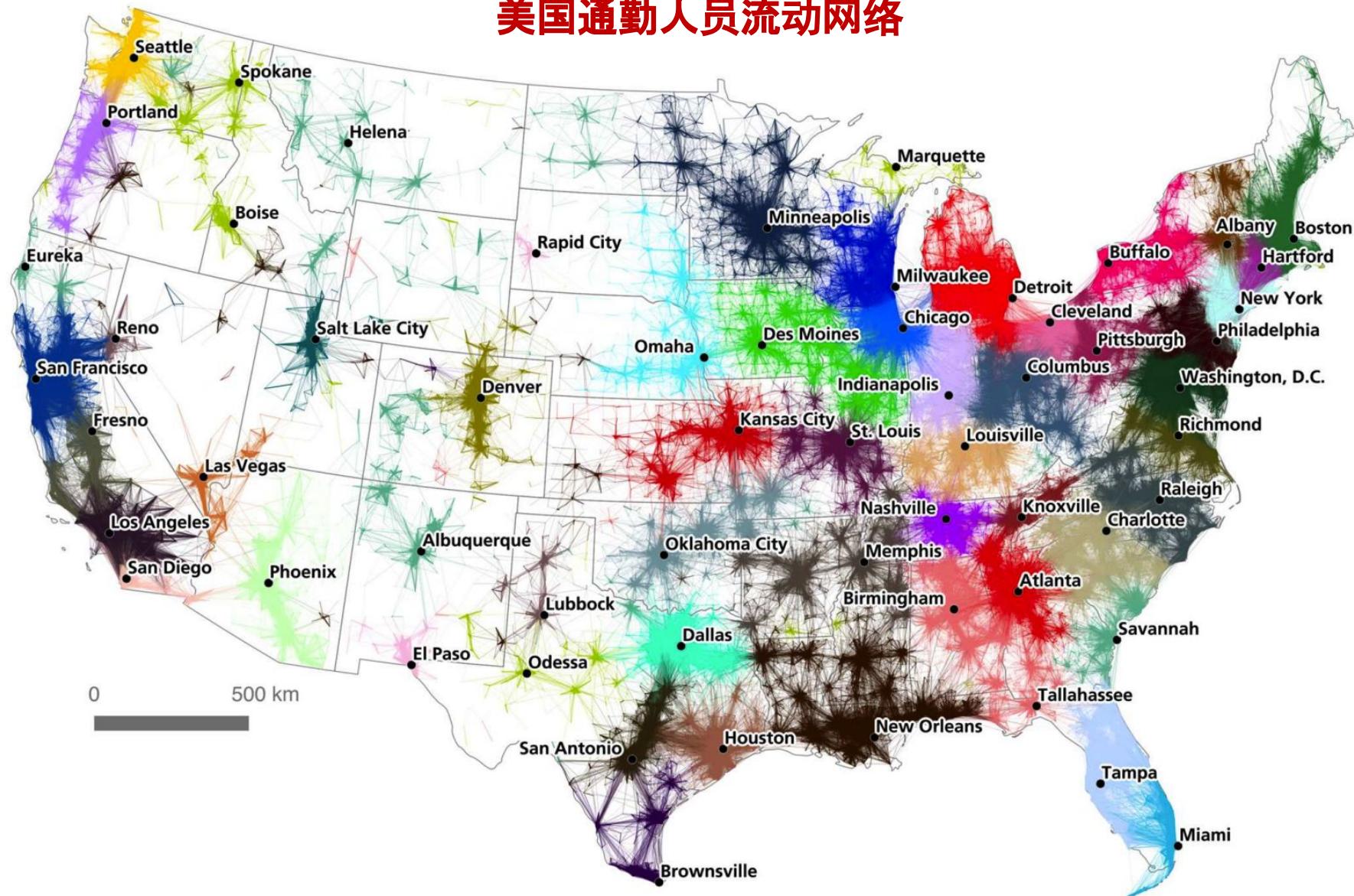
复杂网络(Complex Networks)是将现实世界中各种大型**复杂系统**抽象成网络来进行研究的一种理论工具，在自然界中存在的大量复杂系统都可以通过形形色色的网络加以描述。一个典型的网络是由许多**节点与节点之间的连边**组成，其中节点用来代表真实系统中不同的个体，而边则用来表示节点之间的关系。

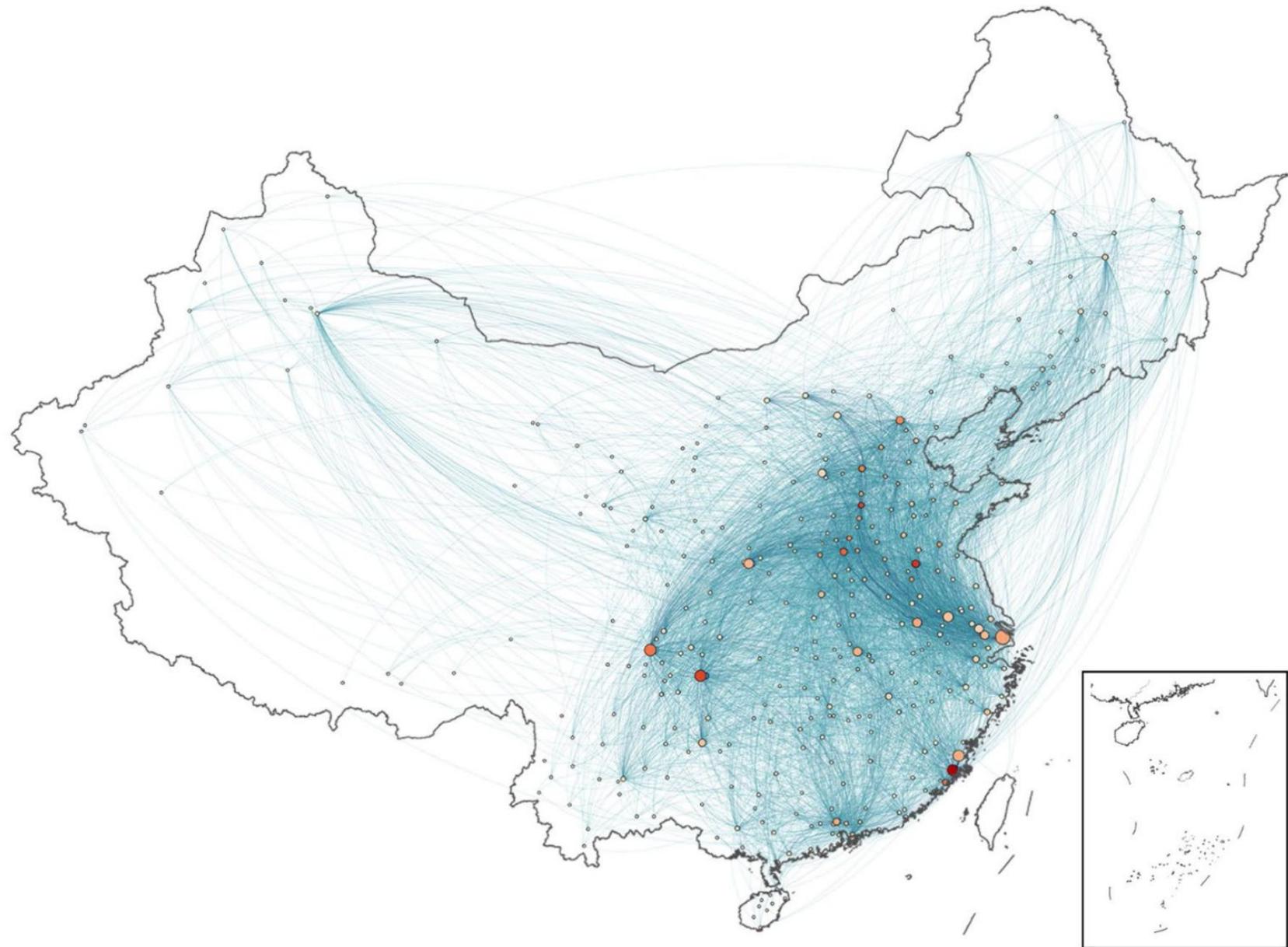




全球航空交通网络

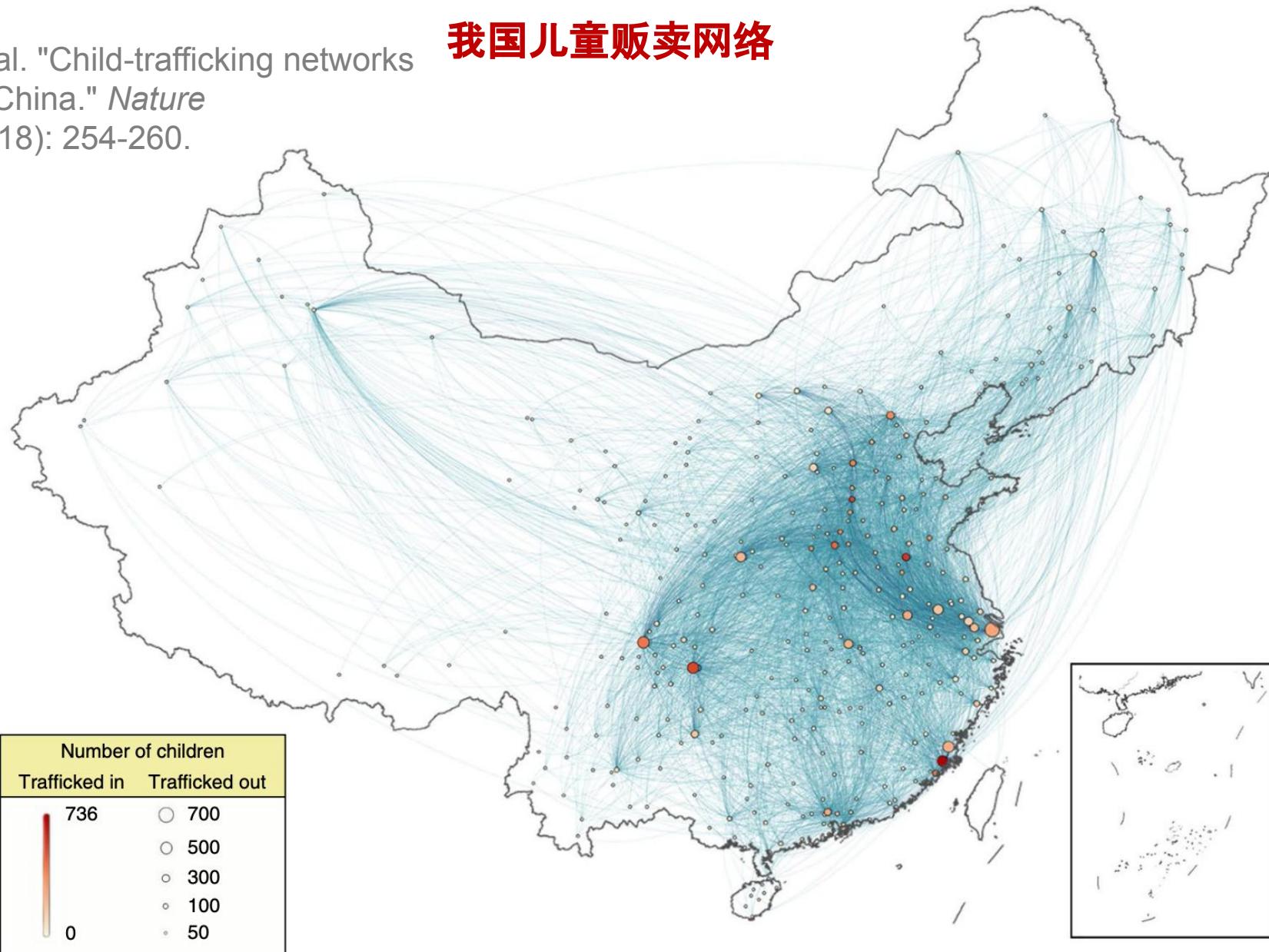
美国通勤人员流动网络

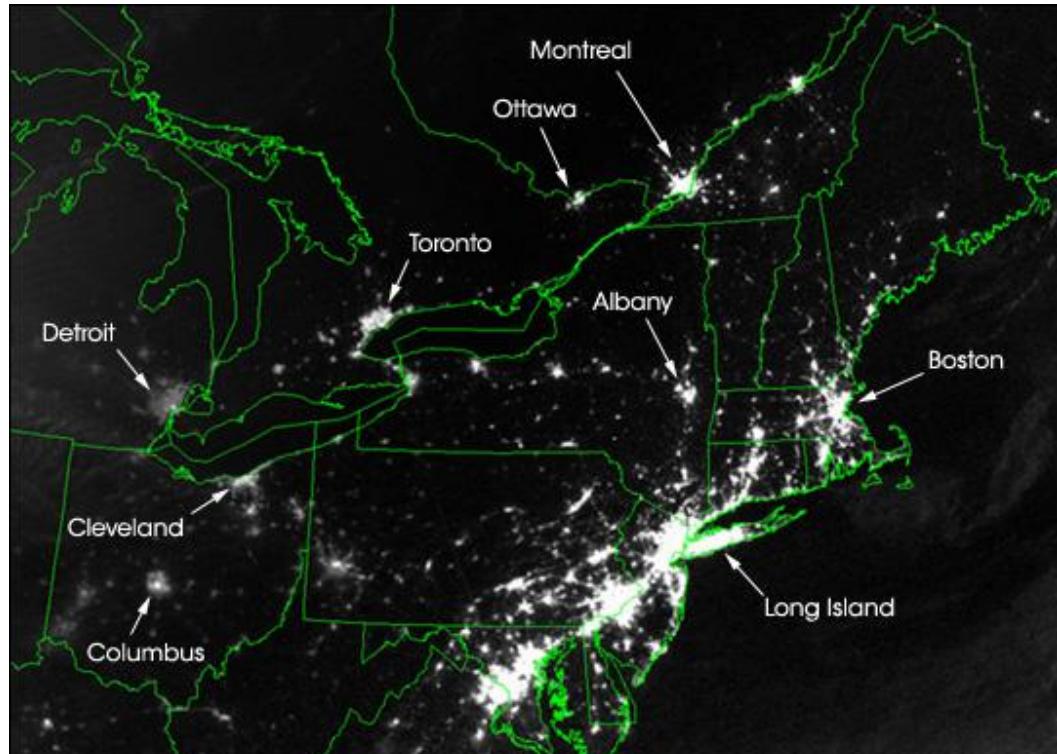




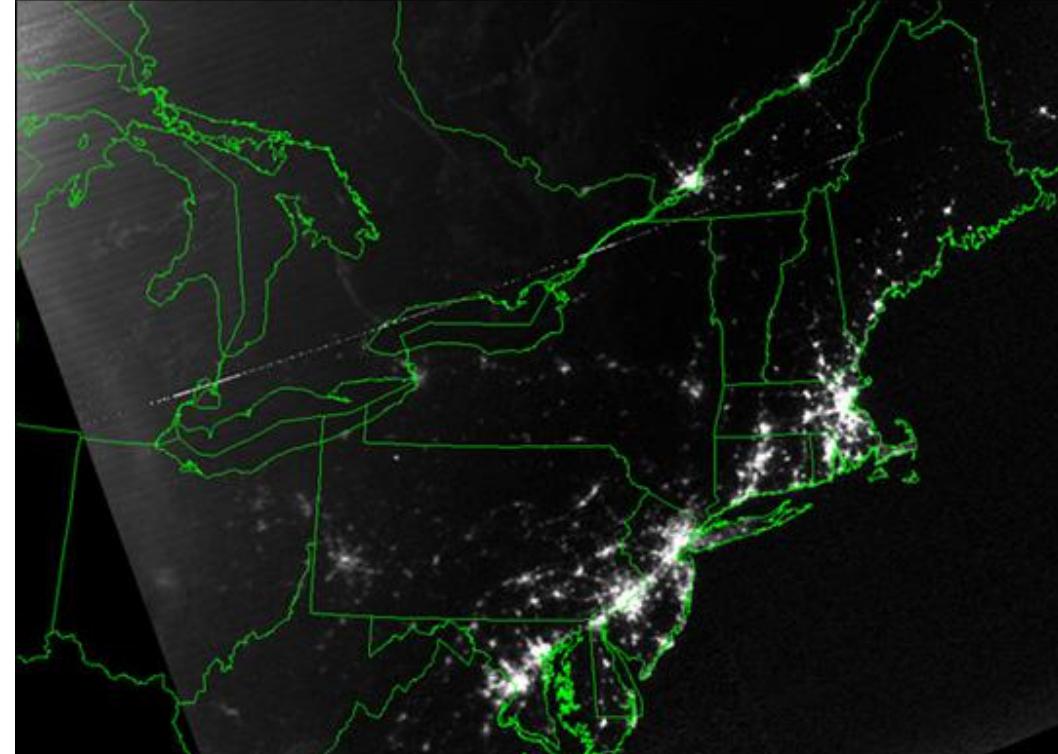
Ref: Wang, Zhen, et al. "Child-trafficking networks of illegal adoption in China." *Nature Sustainability* 1.5 (2018): 254-260.

我国儿童贩卖网络





August 14, 2003 • 9:29 p.m. EDT • About 20 hours before blackout



August 15, 2003 • 9:14 p.m. EDT • About 7 hours after blackout

2003年8月 美国东北部大范围停电

谁在研究复杂网络？

Physicists

Computer Scientists

Applied Mathematicians

Statisticians

Biologists

Ecologists

Sociologists

Political Scientists



it's a big community!

- different *traditions*
- different *tools*
- different *questions*

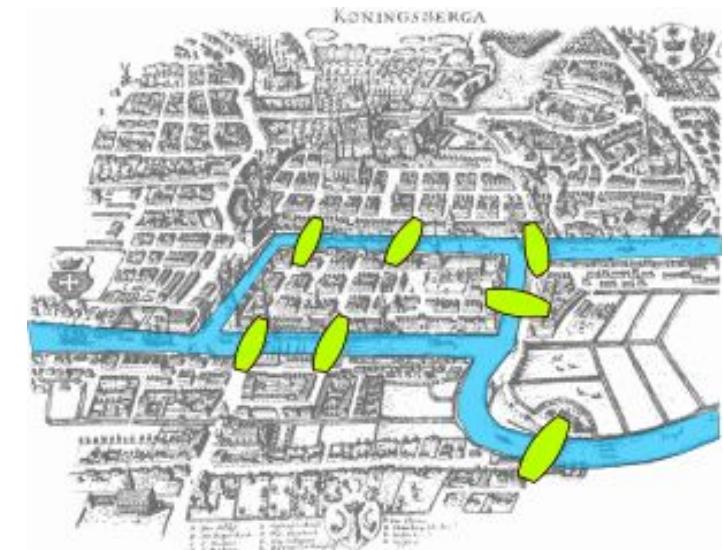
increasingly, not ONE community,
but MANY, only loosely interacting communities

复杂网络科学研究目标:找到不同的复杂网络之间的共性和处理的普适性的方法,并且揭示复杂网络内部的结构和性质,以及不同节点之间的相互作用关系

网络,数学上称为图,最早研究始于1736年欧拉的哥尼斯堡七桥问题,但是之后关于图的研究发展缓慢,直到1936年,才有了第一本关于图论研究的著作。

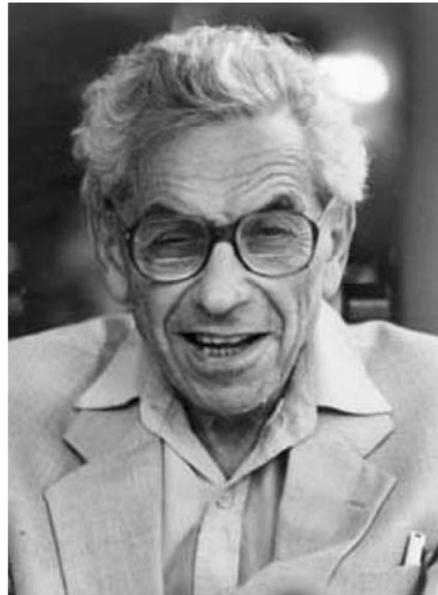


Leonhard Euler



哥尼斯堡七桥问题

20世纪60年代，两位匈牙利数学家Erdos和Renyi建立了随机图理论，被公认为是在数学上开创了复杂网络理论的系统性研究。之后的40年里，人们一直将随机图理论作为复杂网络研究的基本理论。然而，绝大多数的实际网络并不是完全随机的。



Paul Erdős (1913-1996)



Alfréd Rényi (1921-1970)



复杂网络

1998年, Watts及其导师Strogatz在Nature上的文章《Collective Dynamics of Small-world Networks》揭示了复杂网络的小世界性质。

随后, 1999年, Barabasi及其博士生Albert在Science上的文章《Emergence of Scaling in Random Networks》又揭示了复杂网络的无标度性质(度分布为幂律分布), 从此开启了复杂网络研究的新纪元。

nature

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [letters](#) > article

[Published: 04 June 1998](#)

Collective dynamics of ‘small-world’ networks

[Duncan J. Watts](#) & [Steven H. Strogatz](#)

≡ Science

[Current Issue](#) [First release papers](#) [Archive](#)



Emergence of Scaling in Random Networks

[ALBERT-LÁSZLÓ BARABÁSI AND RÉKA ALBERT](#)



2002年Girvan和Newman在PNAS上的一篇文章《Community structure in social and biological networks》，指出复杂网络中普遍存在着聚类特性，每一个类称之为一个社团(community)，并提出了一个发现这些社团的算法。直到现在，有关复杂网络的研究仍然热度不减，并且扩展到了物理、生物、社会等各个领域



RESEARCH ARTICLE

Community structure in social and biological networks

M. Girvan and M. E. J. Newman

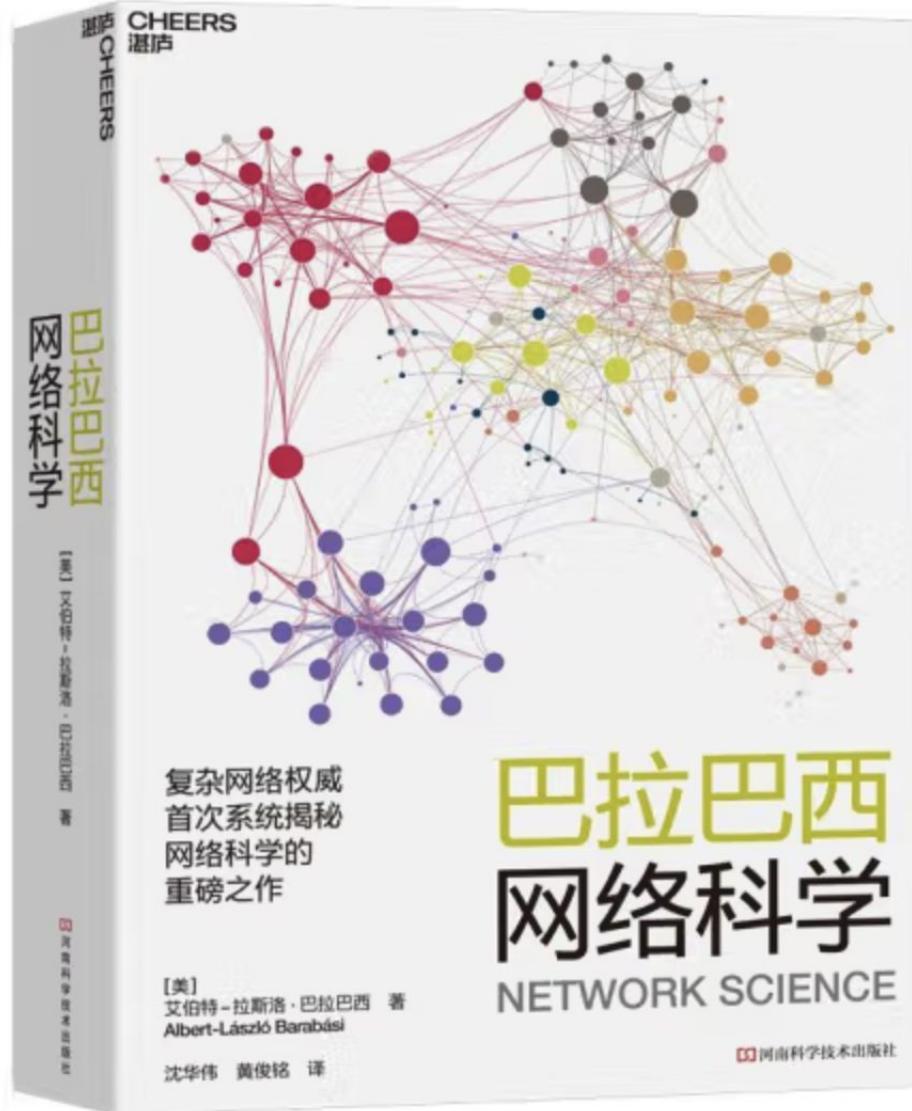
+ See all authors and affiliations



参考教材



人工智能研究院
Artificial Intelligence Institute



Albert-László Barabási
美国东北大学教授

提纲

1

复杂科学简介

2

图的基本概念

3

随机网络

4

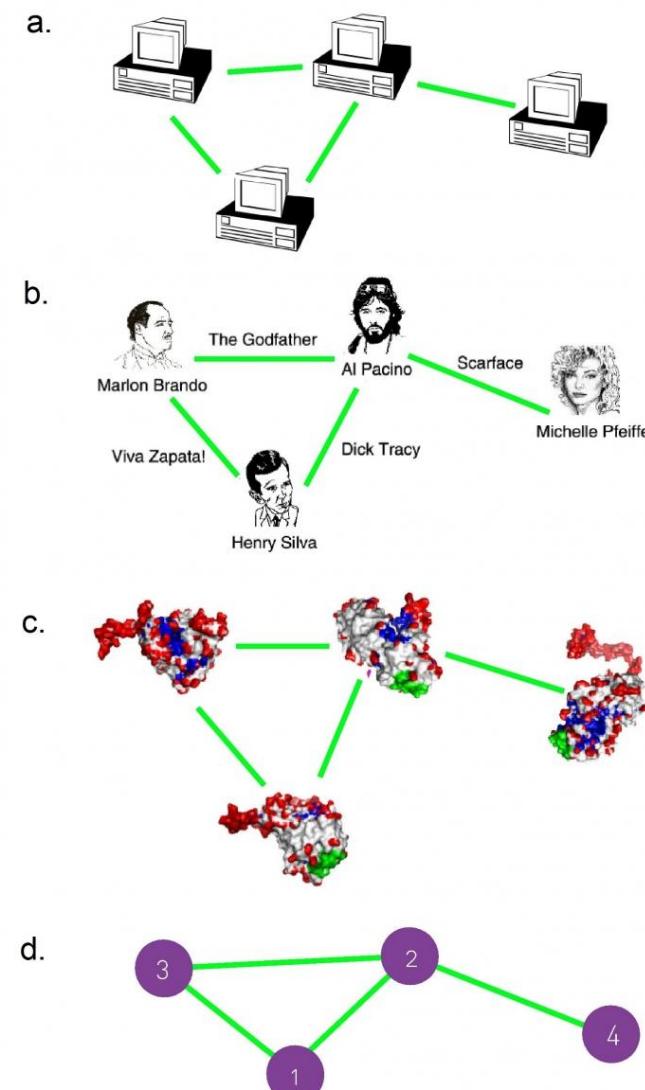
无标度网络

5

社团检测

任何系统都可以抽象为网络或图

Network	Nodes	Links	Directed / Undirected	N	L	$\langle k \rangle$
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.34
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile-Phone Calls	Subscribers	Calls	Directed	36,595	91,826	2.51
Email	Email addresses	Emails	Directed	57,194	103,731	1.81
Science Collaboration	Scientists	Co-authorships	Undirected	23,133	93,437	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Papers	Citations	Directed	449,673	4,689,479	10.43
E.Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930	2.90

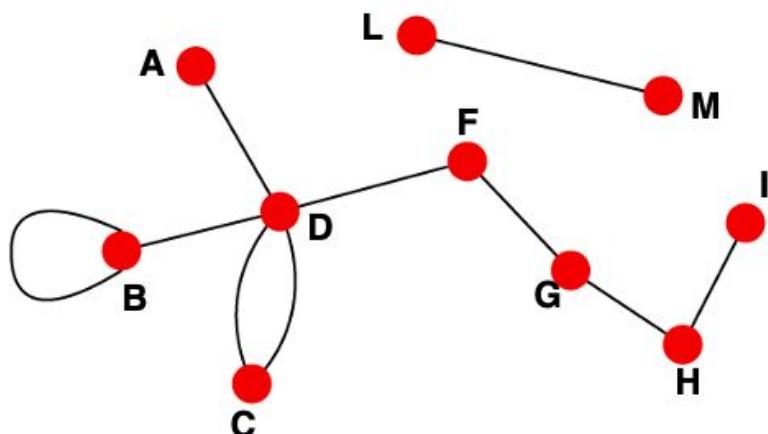


无向图 vs 有向图

Undirected

Links: undirected (*symmetrical*)

Graph:

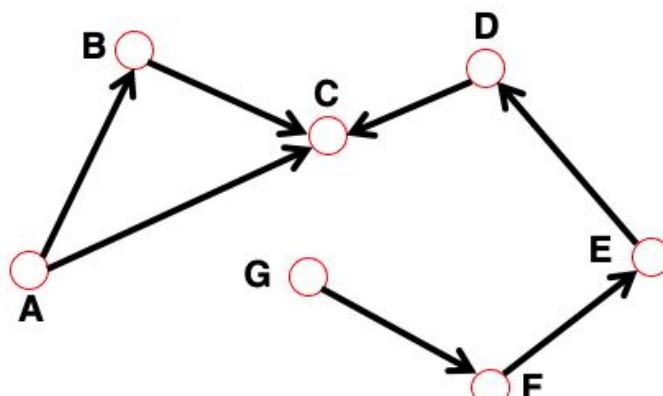


Undirected links :
coauthorship links
Actor network
protein interactions

Directed

Links: directed (*arcs*).

Digraph = directed graph:



An undirected link is the superposition of two opposite directed links.

Directed links :
URLs on the www
phone calls
metabolic reactions

度(Degree)

无向图: 节点*i*相连的边的数量

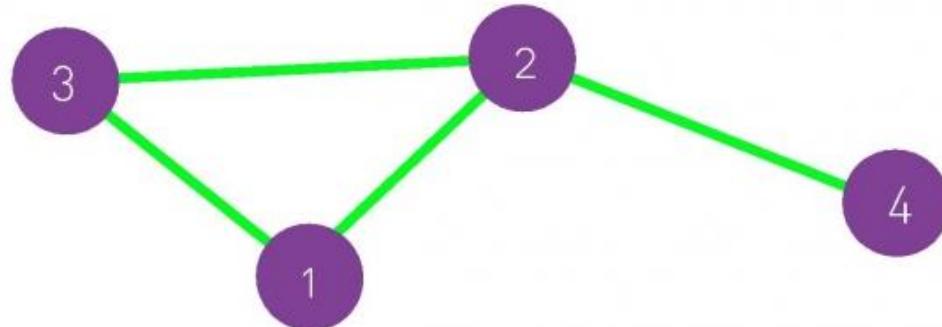
$$k_1 = 2; k_2 = 3; k_3 = 2; k_4 = 1$$

有向图:

入度(Incoming Degree) : 指向节点*i*的边的数量

出度(Outgoing Degree) : 由节点*i*指出的边的数量

$$\text{总度数} : k_i = k_i^{\text{in}} + k_i^{\text{out}}$$



平均度(Average Degree)

无向图:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^N k_i = \frac{2L}{N}$$

有向图:

$$L = \sum_{i=1}^N k_i^{\text{in}} = \sum_{i=1}^N k_i^{\text{out}}$$

$$\langle k^{\text{in}} \rangle = \frac{1}{N} \sum_{i=1}^N k_i^{\text{in}} = \langle k^{\text{out}} \rangle = \frac{1}{N} \sum_{i=1}^N k_i^{\text{out}} = \frac{L}{N}$$

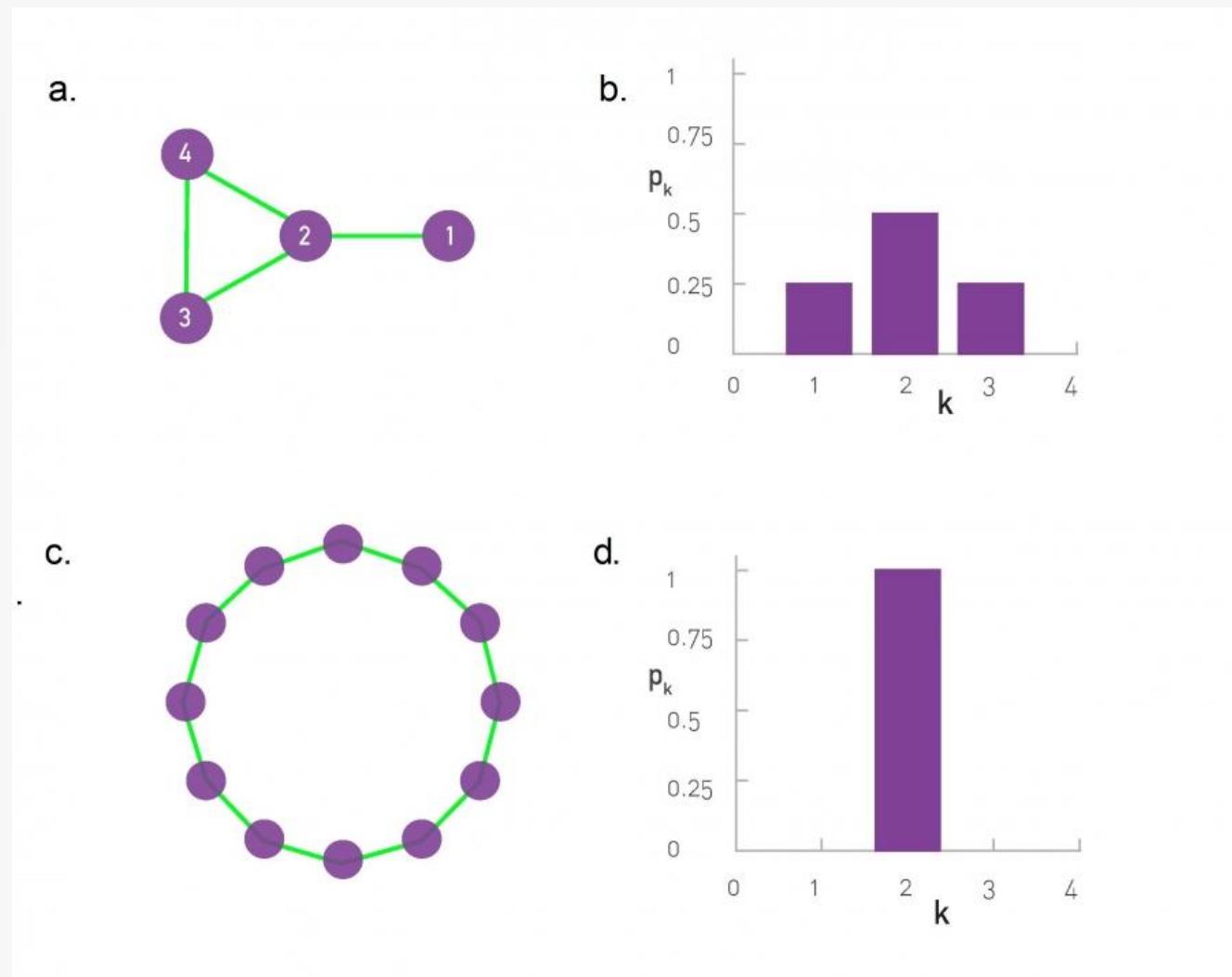
度的分布(Degree Distribution)

$$p_k = \frac{N_k}{N}$$

$$\sum_{k=1}^{\infty} p_k = 1$$

度分布计算平均度：

$$\langle k \rangle = \sum_{k=0}^{\infty} kp_k$$

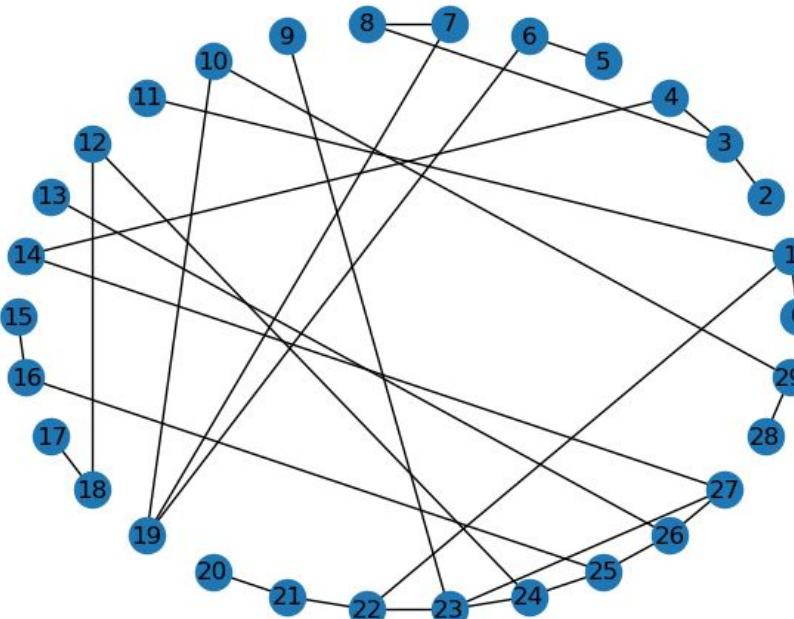


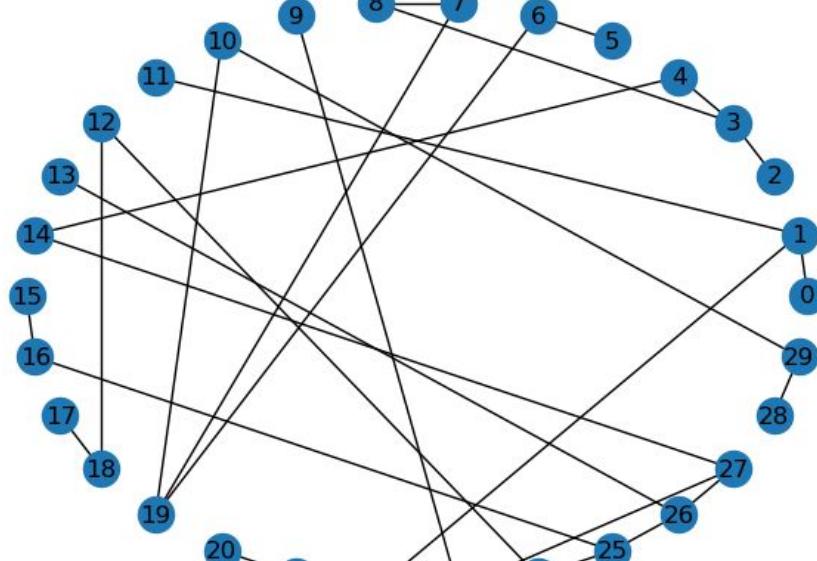
undirected Graph

构建一个小世界模型(无向图)

```
: import networkx as nx
from matplotlib import pyplot as plt

ws = nx.watts_strogatz_graph(30, 3, 0.3) #小世界模型
nx.draw_circular(ws, with_labels=True, )
```





计算平均度

通过统计每个节点的度的平均值计算平均度

```
total_degree = 0
for node, item in ws.degree:
    total_degree+=item
print(f"平均度为: {total_degree/len(ws.nodes)}")
```

平均度为: 2.0

使用列表推导式/元组推导式可以简化:

```
print(f"平均度为: {sum(item for _, item in ws.degree)/len(ws.nodes)}")
```

平均度为: 2.0

度和边的关系: 总度数为边的两倍

通过边数计算平均度

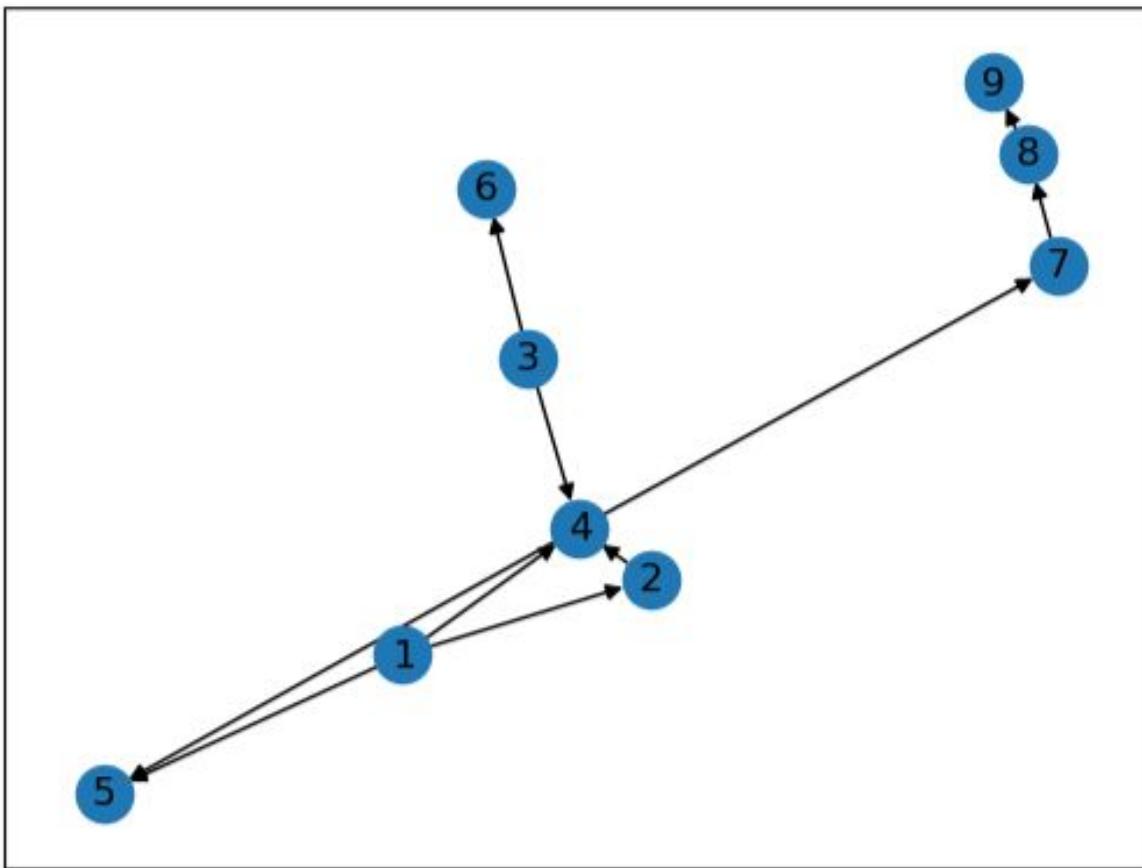
```
print(f"平均度为: {(ws.size())*2/len(ws.nodes)})")
```

平均度为: 2.0

directed Graph ↗

```
: dg = nx.read_edgelist("edgelist.txt", create_using = nx.DiGraph)  
nx.draw_networkx(dg)
```

从文件中读取边列表，构建有向图



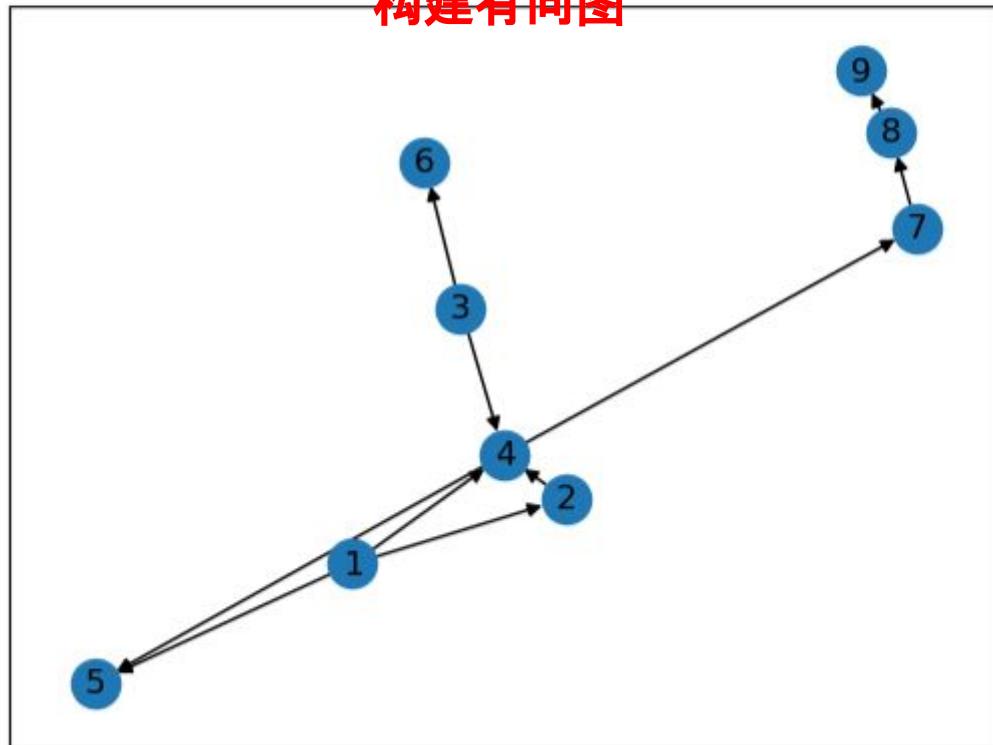
边列表

	1	2
1	1	2
2	2	4
3	3	4
4	1	4
5	1	5
6	4	5
7	7	8
8	4	7
9	8	9
10	3	6

directed Graph 1

```
: dg = nx.read_edgelist("edgelist.txt", create_using = nx.DiGraph)  
nx.draw_networkx(dg)
```

从文件中读取边列表，
构建有向图



```
dg.in_degree() #入度
```

```
InDegreeView({'1': 0, '2': 1, '4': 3, '3': 0, '5': 2, '7': 1, '8': 1, '9': 1, '6': 1})
```

```
dg.out_degree() #出度
```

```
OutDegreeView({'1': 3, '2': 1, '4': 2, '3': 2, '5': 0, '7': 1, '8': 1, '9': 0, '6': 0})
```

平均度：

```
print(f"平均度为: {sum(item for _, item in dg.degree)/len(dg.nodes)}")
```

平均度为: 2.22222222222223

有向图中，出度和等于入度和，等于边的数量

```
print("入度和: ", sum(item for _, item in dg.in_degree()))
```

```
print("出度和: ", sum(item for _, item in dg.out_degree()))
```

```
print("边的数量: ", dg.size())
```

入度和: 10

出度和: 10

边的数量: 10

度的分布直方图

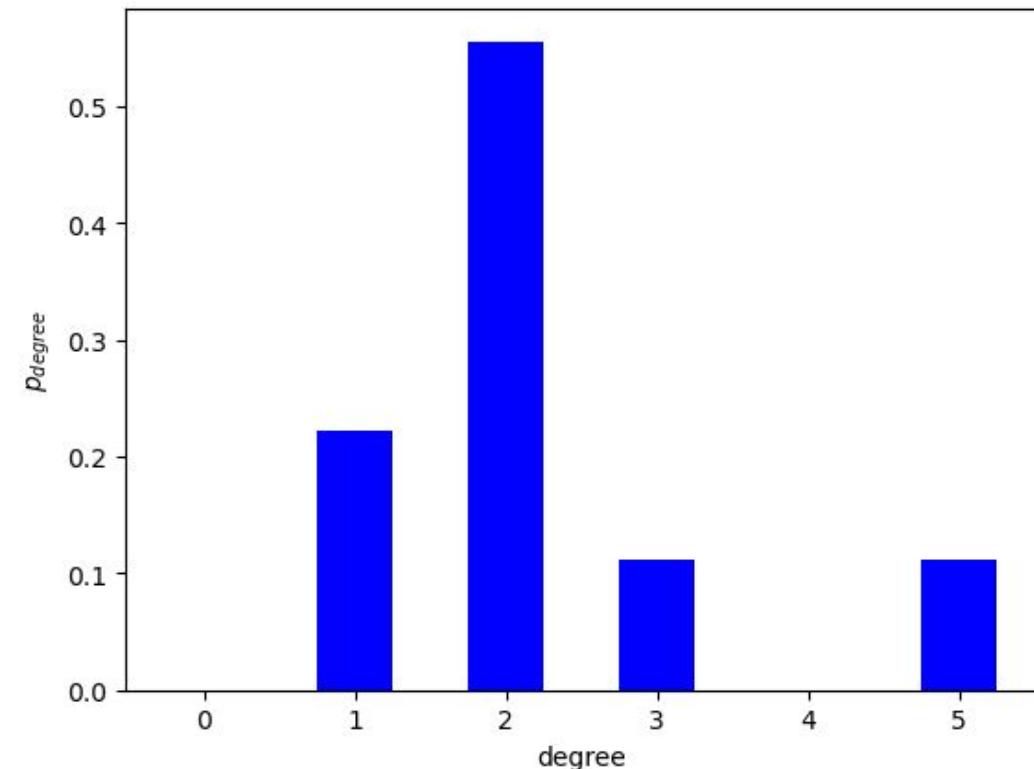
```
nx.degree_histogram(dg) # 返回所有位于区间[0, dmax]的度值的频数列表, 这里就是[0, 5]
```

```
[0, 2, 5, 1, 0, 1]
```

```
def plot_degree_distribution(g):
    maxd = max(item for _, item in g.degree())
    mind = min(item for _, item in g.degree())
    degreehist = nx.degree_histogram(g)#
    plt.bar(range(maxd+1), [item/sum(degreehist) for item in degreehist], width = 0.5 ,color="blue")
    plt.xlabel("degree")
    plt.ylabel("$p_{\{degree\}}$")
    plt.show()
    plt.close()
```

用函数形式方便重复调用

```
plot_degree_distribution(dg)
```



度的分布(Degree Distribution)

$$p_k = \frac{N_k}{N} \quad \sum_{k=1}^{\infty} p_k = 1$$

度分布计算平均度: $\langle k \rangle = \sum_{k=0}^{\infty} kp_k$

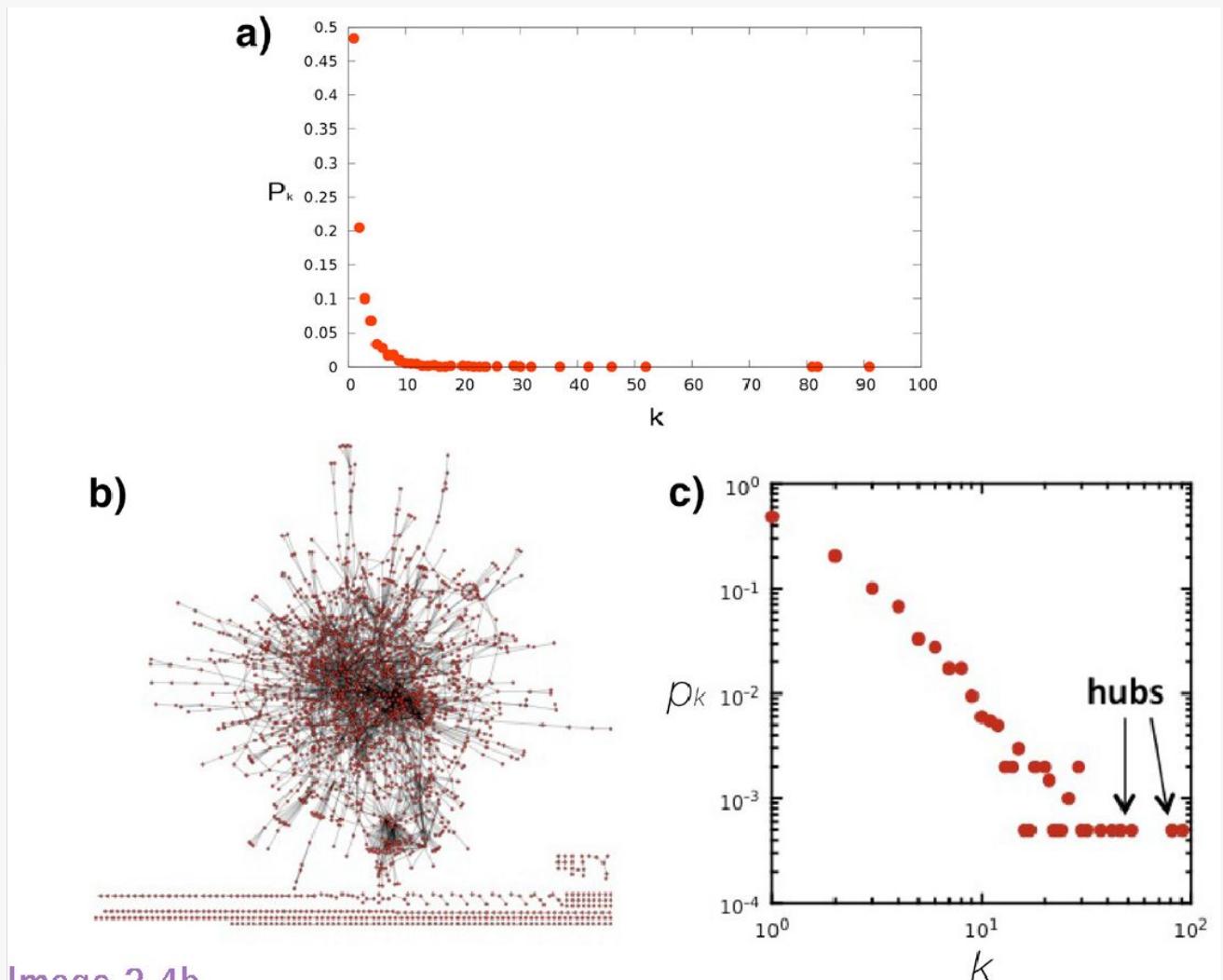


Image 2.4b

真实网络的度分
布



complex Graph

```
: ba = nx.barabasi_albert_graph(1000, 5, seed=None, initial_graph=None):
: print(f"平均度为: {sum(item for _, item in ba.degree)/len(ba.nodes)}")
平均度为: 9.95
```

使用巴拉巴西-阿尔伯特模型生成一个网络，在网络的生长过程中，新加入的点会连接固定数量的节点，但**更倾向于连接网络中度更大的点**，从而使得少量节点逐步成长为枢纽节点。

示例中，生成了一个1000个节点构成的图网
络。

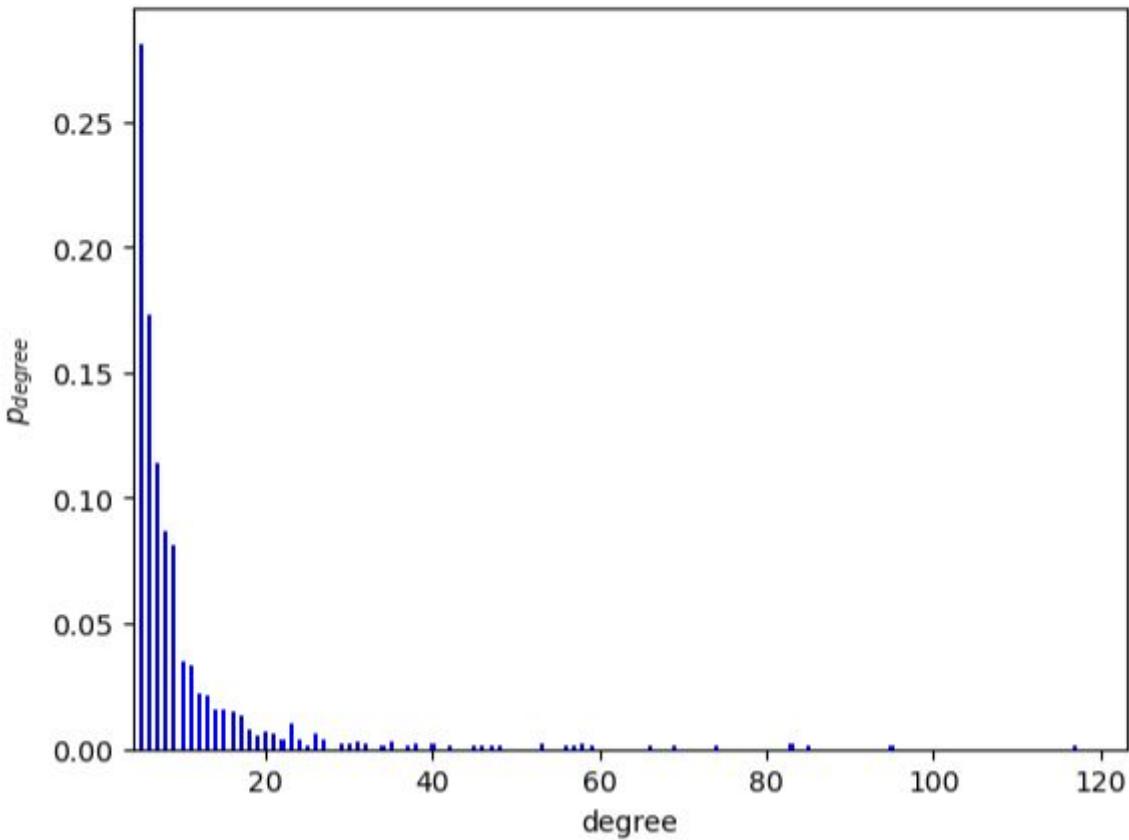
为绘制分布的函数添加新功能：

- 可以绘制scatter plot
- 可以设置为双对数坐标

```
def plot_degree_distribution(g, scatter = False, dual_log=False):
    maxd = max(item for _, item in g.degree())
    mind = min(item for _, item in g.degree())
    degreehist = nx.degree_histogram(g)#
    if scatter:
        plt.scatter(range(maxd+1), [item/sum(degreehist) for item in degreehist], color="blue")
        plt.xlim(mind, )
    else:
        plt.bar(range(maxd+1), [item/sum(degreehist) for item in degreehist], width = 0.5 ,color="blue")
        plt.xlim(mind-1, )
    plt.xlabel("degree")
    plt.ylabel("$p_{\{degree\}}$")
    if dual_log:
        plt.xscale('log')
        plt.yscale('log')
    plt.show()
    plt.close()
```

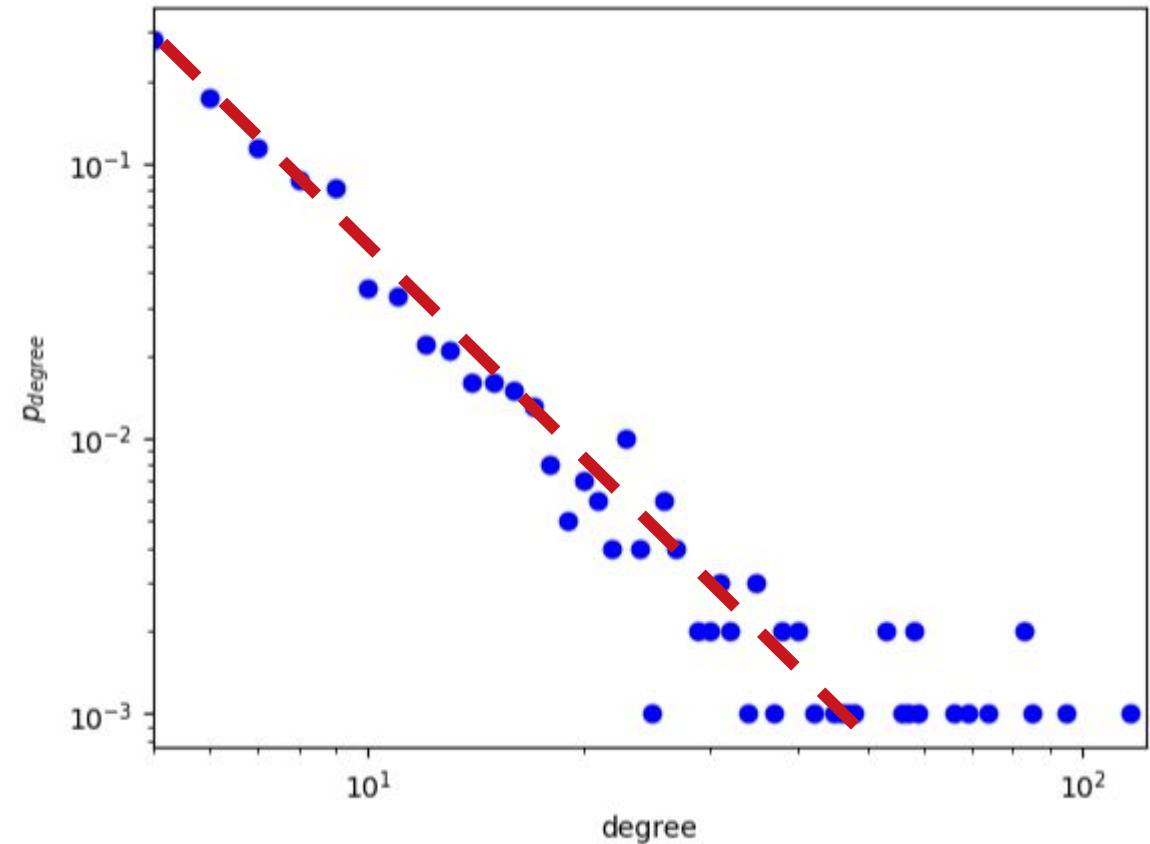
使用bar plot, 采用线性坐标轴

```
plot_degree_distribution(ba, scatter = False, dual_log=False)
```



使用scatter plot, 采用双对数坐标轴

```
plot_degree_distribution(ba, scatter = True, dual_log=True)
```



Hint: 这个网络度的分布有什么规律？

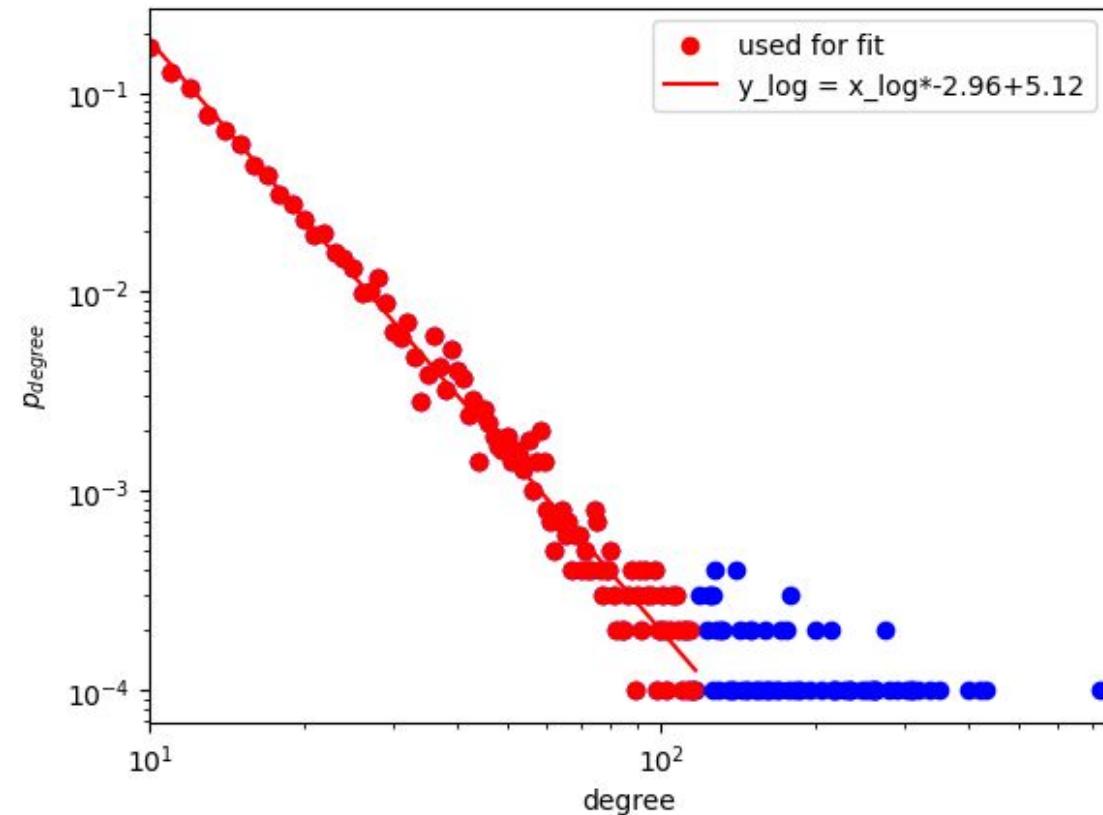
拟合双log下的直线

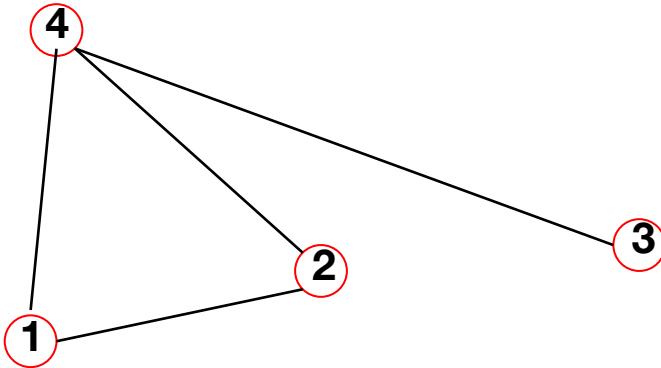
```

def plot_degree_distribution(g, scatter = False, dual_log=False, fit = False):
    maxd = max(item for _, item in g.degree())
    mind = min(item for _, item in g.degree())
    degreehist = nx.degree_histogram(g)#
    if scatter:
        plt.scatter(range(maxd+1), [item/sum(degreehist) for item in degreehist], color="blue")
        plt.xlim(mind, )
    else:
        plt.bar(range(maxd+1), [item/sum(degreehist) for item in degreehist], width = 0.5 ,color="blue")
        plt.xlim(mind-1, )
    plt.xlabel("degree")
    plt.ylabel("$p_{\{degree\}}$")
    if dual_log:
        plt.xscale('log')
        plt.yscale('log')
    if fit:
        cumulate = np.cumsum(degreehist)
        take_first_x = (cumulate/sum(degreehist)<0.99).sum()## 取99%节点的度来拟合
        #筛选可用的数据点:
        x = np.array(range(mind,take_first_x))
        y = np.array(degreehist[mind:take_first_x])/sum(degreehist)
        plt.scatter(x, y, color="red",label='used for fit')
        selected = y>0
        #log
        y = np.log(y[selected])
        x = np.log(x[selected])
        #拟合
        coeff = np.polyfit(x, y, 1)
        #坐标:
        x = np.linspace(np.log(mind),np.log(take_first_x),10000)
        y = coeff[0]*x+coeff[1]
        x = np.exp(x)
        y = np.exp(y)
        plt.plot(x,y,color='red',label=f"y_log = x_log*{coeff[0]:.3}+{coeff[1]:.3}")#绘制log-log下的直线
        print(f"y_log = x_log*{coeff[0]:.3}+{coeff[1]:.3}")
        plt.legend()
    plt.show()
    plt.close()

```

使用99%节点的度分布用于拟合直线



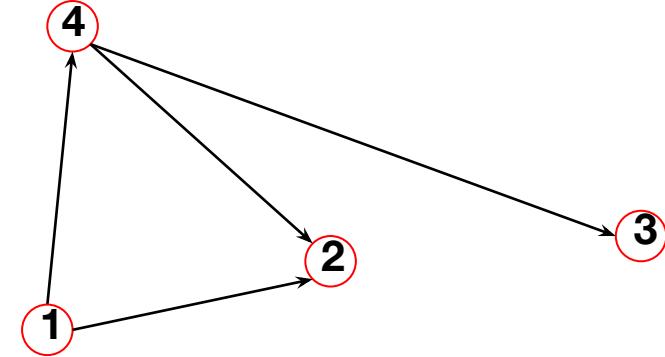


$A_{ij} = 1$ 如果节点 i 和节点 j 之间存在边

$A_{ij} = 0$ 如果节点 i 和节点 j 之间不存在边.

$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

Note: 对于有向图, 邻接矩阵是非对称的

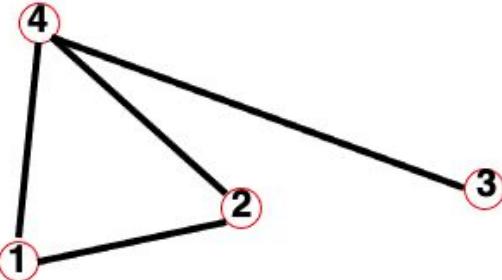


$A_{ij} = 1$ 如果存在边由节点 j 指向节点 i

$A_{ij} = 0$ 如果不存在边由节点 j 指向节点 i.

$$A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Undirected



$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

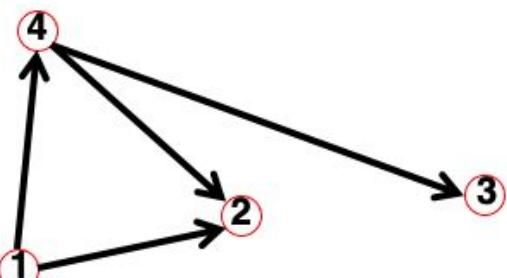
$$\begin{aligned} A_{ij} &= A_{ji} \\ A_{ii} &= 0 \end{aligned}$$

$$k_i = \sum_{j=1}^N A_{ij}$$

$$k_j = \sum_{i=1}^N A_{ij}$$

$$L = \frac{1}{2} \sum_{i=1}^N k_i = \frac{1}{2} \sum_{i,j} A_{ij}$$

Directed



$$A_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

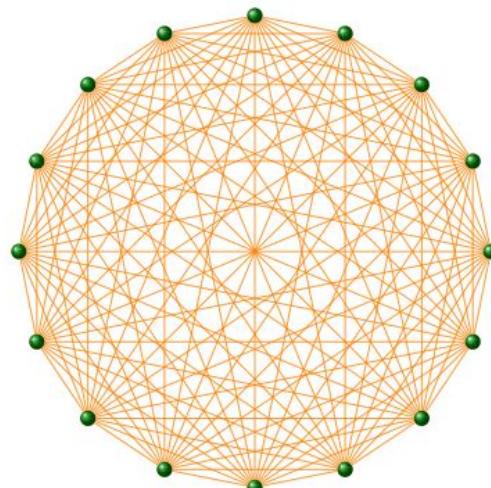
$$\begin{aligned} A_{ij} &\neq A_{ji} \\ A_{ii} &= 0 \end{aligned}$$

$$k_i^{in} = \sum_{j=1}^N A_{ij}$$

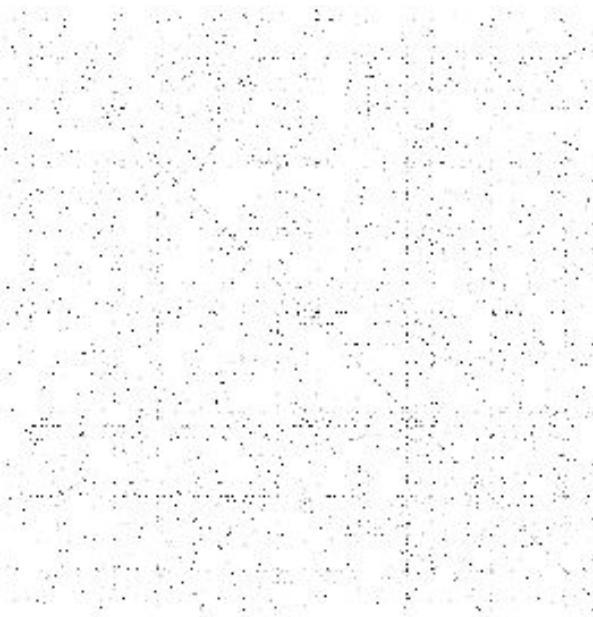
$$k_j^{out} = \sum_{i=1}^N A_{ij}$$

$$L = \sum_{i=1}^N k_i^{in} = \sum_{j=1}^N k_j^{out} = \sum_{i,j} A_{ij}$$

稀疏性(Sparseness)真实网络是非常稀疏的！

全连接网
络

WWW (ND Sample): $N=325,729$; $L=1.4 \times 10^6$
Protein (*S. Cerevisiae*): $N=1,870$; $L=4,470$
Coauthorship (Math): $N=70,975$; $L=2 \times 10^5$
Movie Actors: $N=212,250$; $L=6 \times 10^6$

真实稀疏网络的邻接矩
阵

$L_{\max}=10^{12}$ $\langle k \rangle=4.51$
 $L_{\max}=10^7$ $\langle k \rangle=2.39$
 $L_{\max}=3 \times 10^{10}$ $\langle k \rangle=3.9$
 $L_{\max}=1.8 \times 10^{13}$ $\langle k \rangle=28.78$

(Source: Albert, Barabasi, RMP2002)



不同平均度的邻接矩阵

$G_{n,p}$ 随机图，也叫
Erdős-Rényi图或
二项式图；

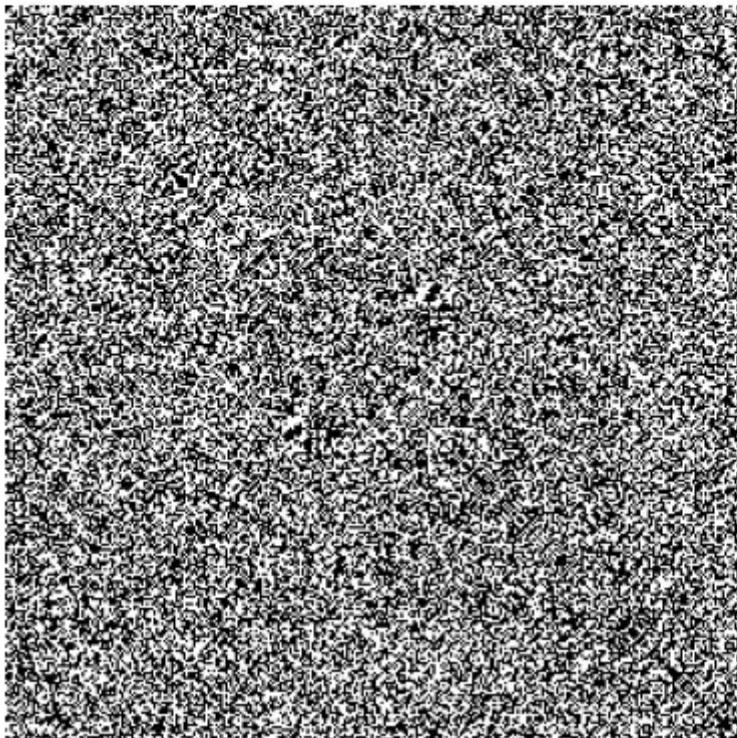
图中白色点表示1，
即有边存在

p=0.5时，生成平均度约等于150

```
er = nx.erdos_renyi_graph(300, 0.5)#Erdős-Rényi graph with p=0.5
print(f"平均度为: {(er.size())*2/len(er.nodes)}")
plt.imshow(nx.adjacency_matrix(er).todense(), cmap = 'gray')
plt.box(False)
plt.axis("off")
```

平均度为: 150.34

(-0.5, 299.5, 299.5, -0.5)

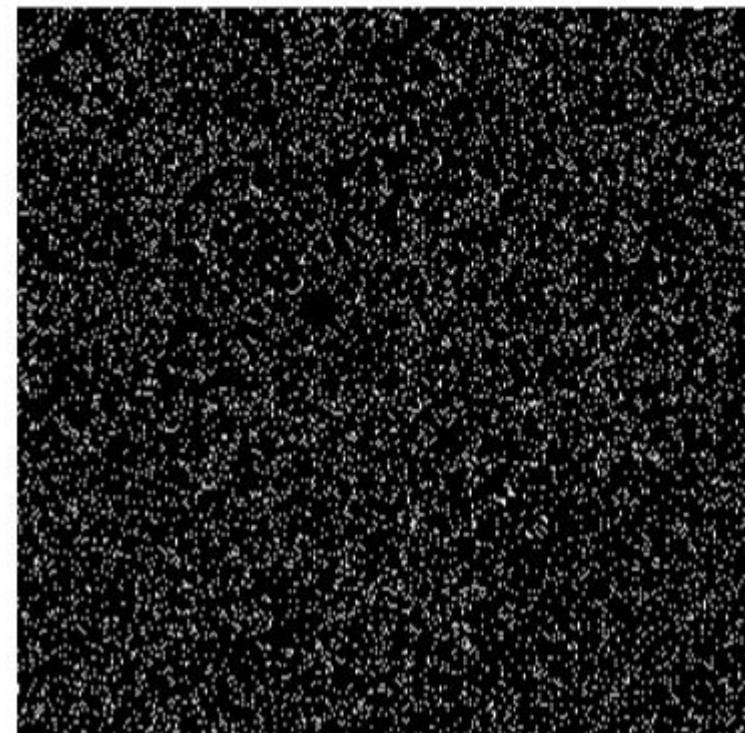


p=0.1时，生成平均度约等于30

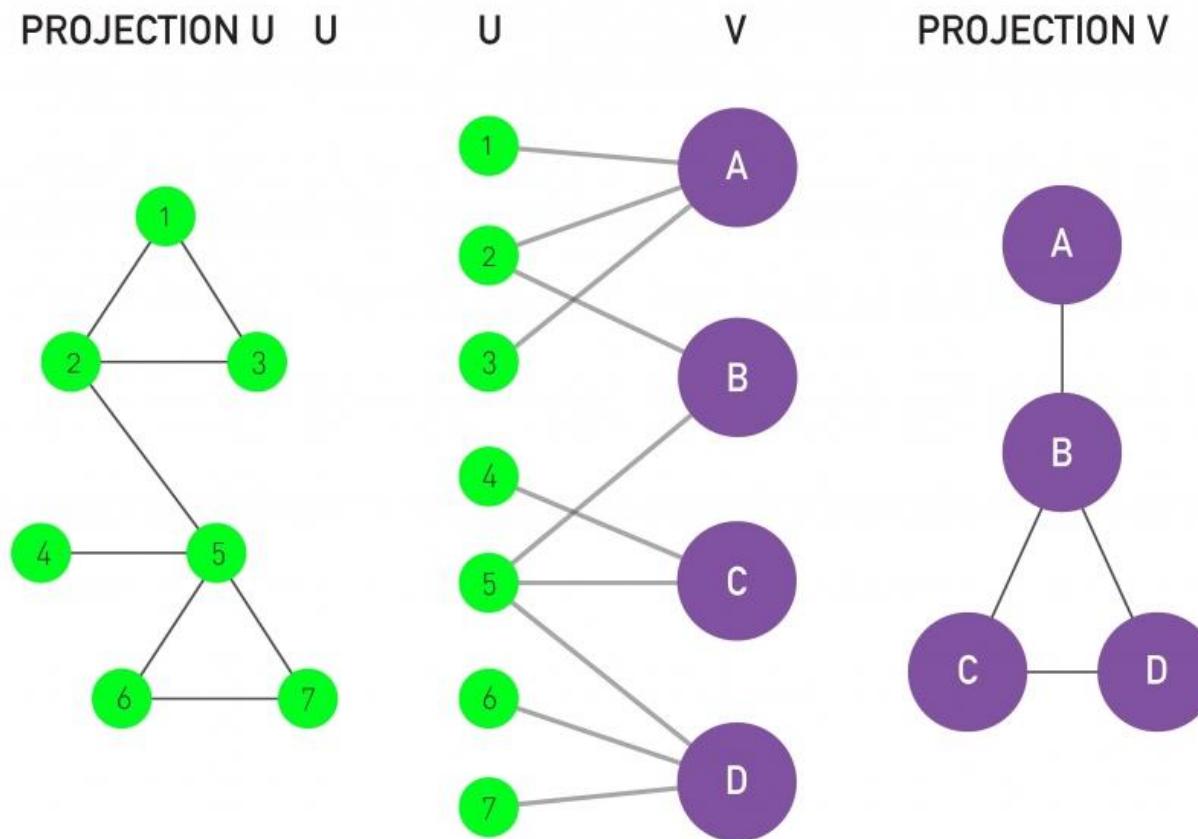
```
er = nx.erdos_renyi_graph(300, 0.1)#Erdős-Rényi graph with p=0.1
print(f"平均度为: {(er.size())*2/len(er.nodes)}")
plt.imshow(nx.adjacency_matrix(er).todense(), cmap = 'gray')
plt.box(False)
plt.axis("off")
```

平均度为: 30.92

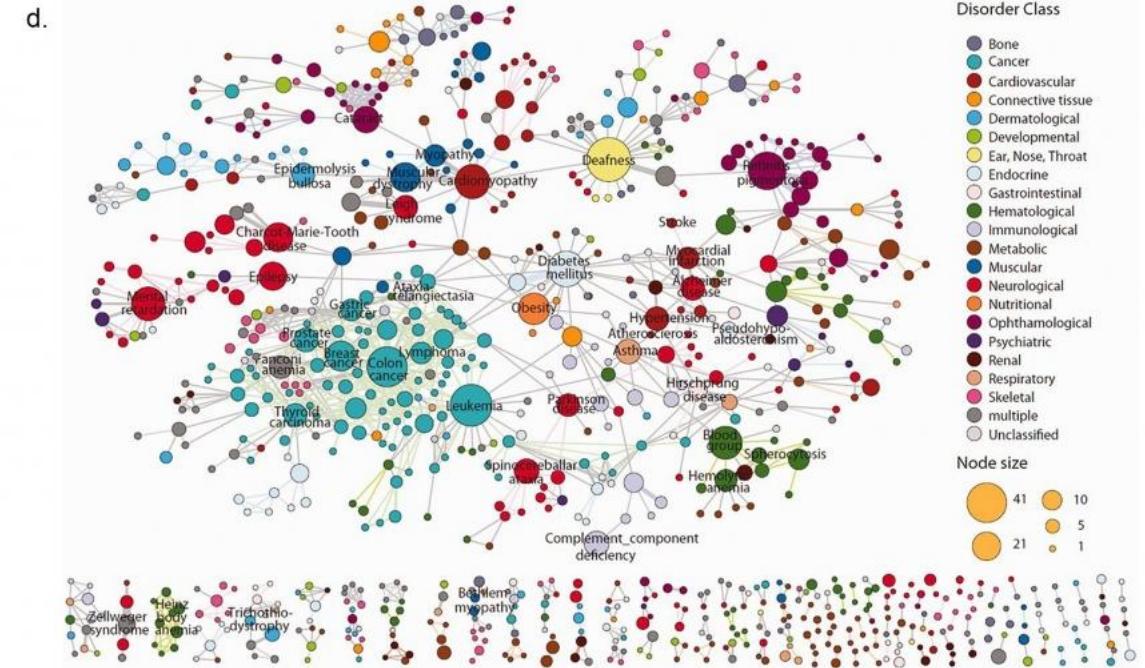
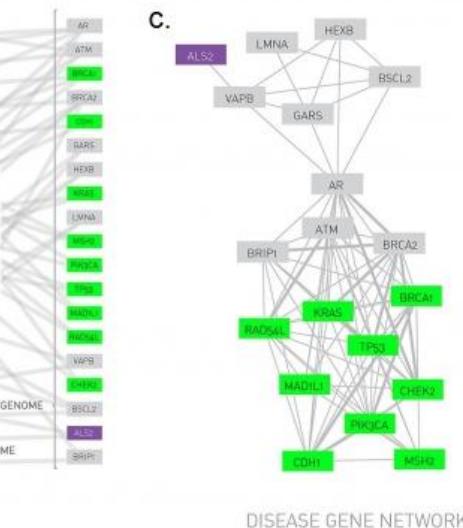
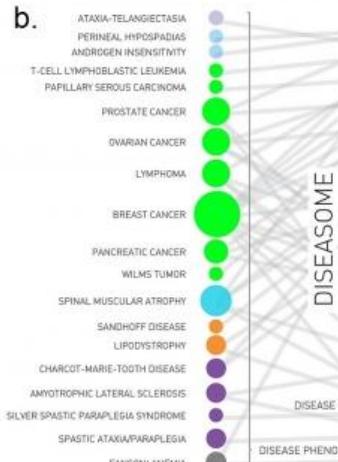
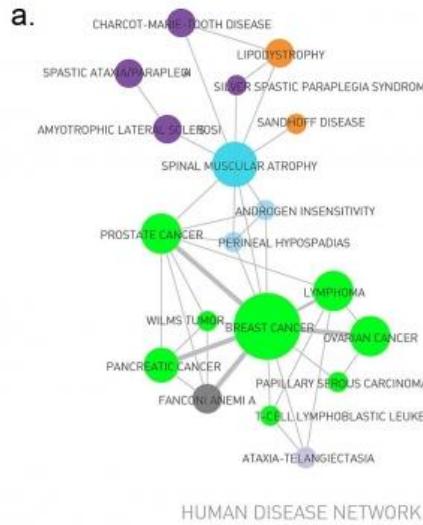
(-0.5, 299.5, 299.5, -0.5)



二部图(Bipartite Network)

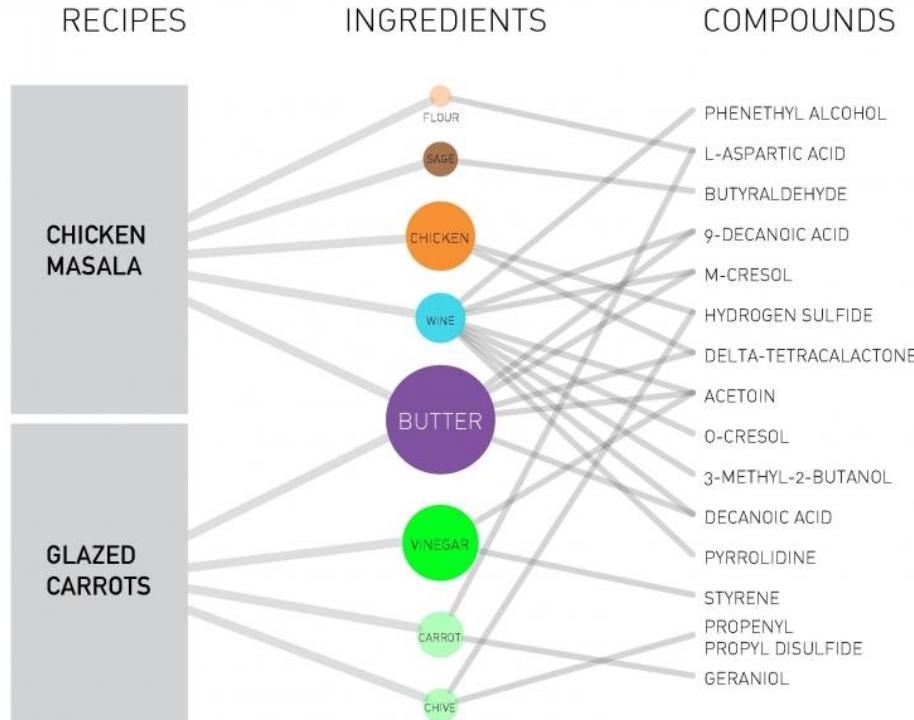


二部图(Bipartite Network)

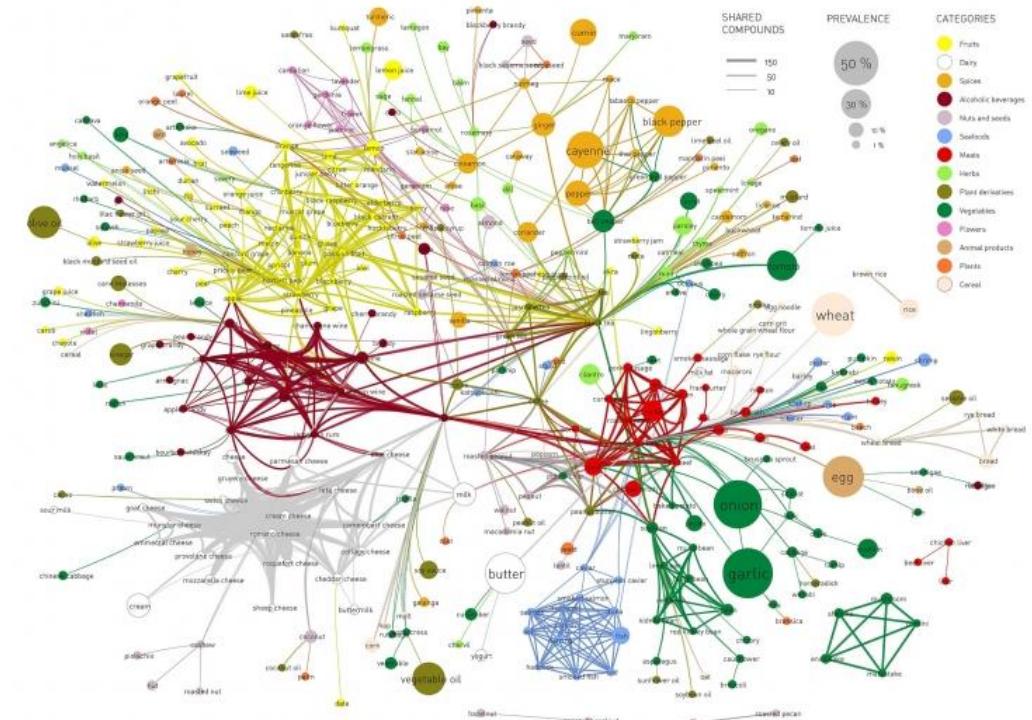


三部图(Tripartite Network)

a. RECIPES

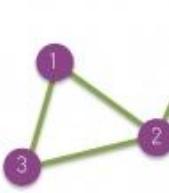


b.



无向图

a. Undirected



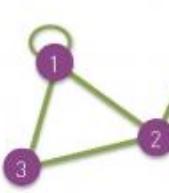
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{ij} = A_{ji}$$

$$L = \frac{1}{2} \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{2L}{N}$$

有环图

b. Self-loops



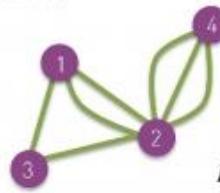
$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

$$\exists i, A_{ii} \neq 0 \quad A_{ij} = A_{ji}$$

$$L = \frac{1}{2} \sum_{i,j=1, i \neq j}^N A_{ij} + \sum_{i=1}^N A_{ii} \quad ?$$

多重无向

c. Multigraph (undirected)



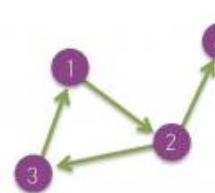
$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{ij} = A_{ji}$$

$$L = \frac{1}{2} \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{2L}{N}$$

有向图

d. Directed

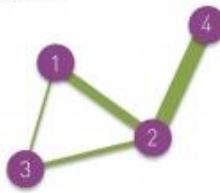


$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_{ij} \neq A_{ji} \quad L = \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{L}{N}$$

带权无向

e. Weighted (undirected)



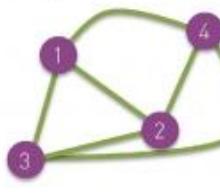
$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{ij} = A_{ji}$$

$$\langle k \rangle = \frac{2L}{N}$$

完全无向图

f. Complete Graph (undirected)



$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{ij} = 1$$

$$L = L_{\max} = \frac{N(N-1)}{2} \quad \langle k \rangle = N-1$$

提纲

1

复杂科学简介

2

图的基本概念

3

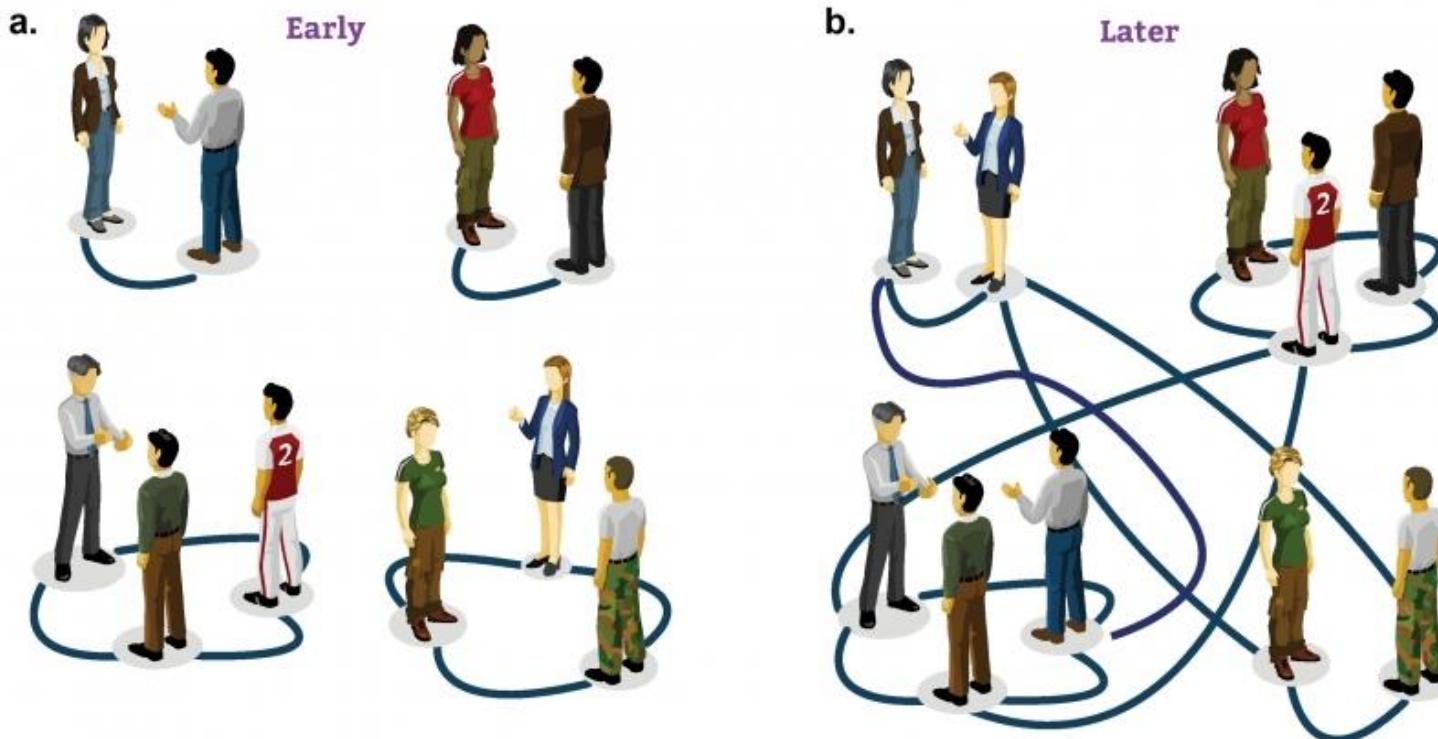
随机网络

4

无标度网络

5

社团检测



鸡尾酒会的熟人网络:一开始宾客之间组成小团体, 随着人与人之间的交往,
一个无形的网络出现了

网络科学旨在建立一个反应真实网络的模型。我们接触到的大多数网络并没有一个确切的结构或者固定模式。它们大多都是随机的。随机网络理论通过随机性来构建模型。

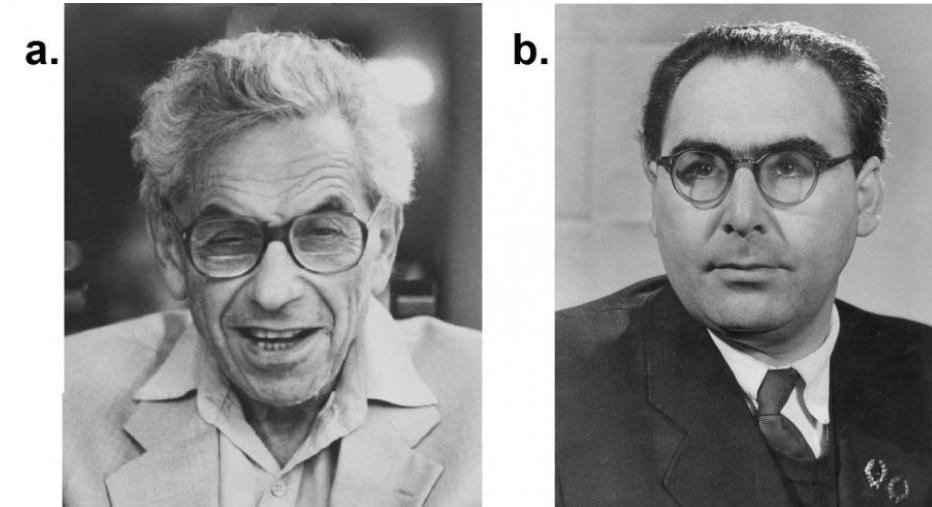
随机网络模型有以下两种定义：

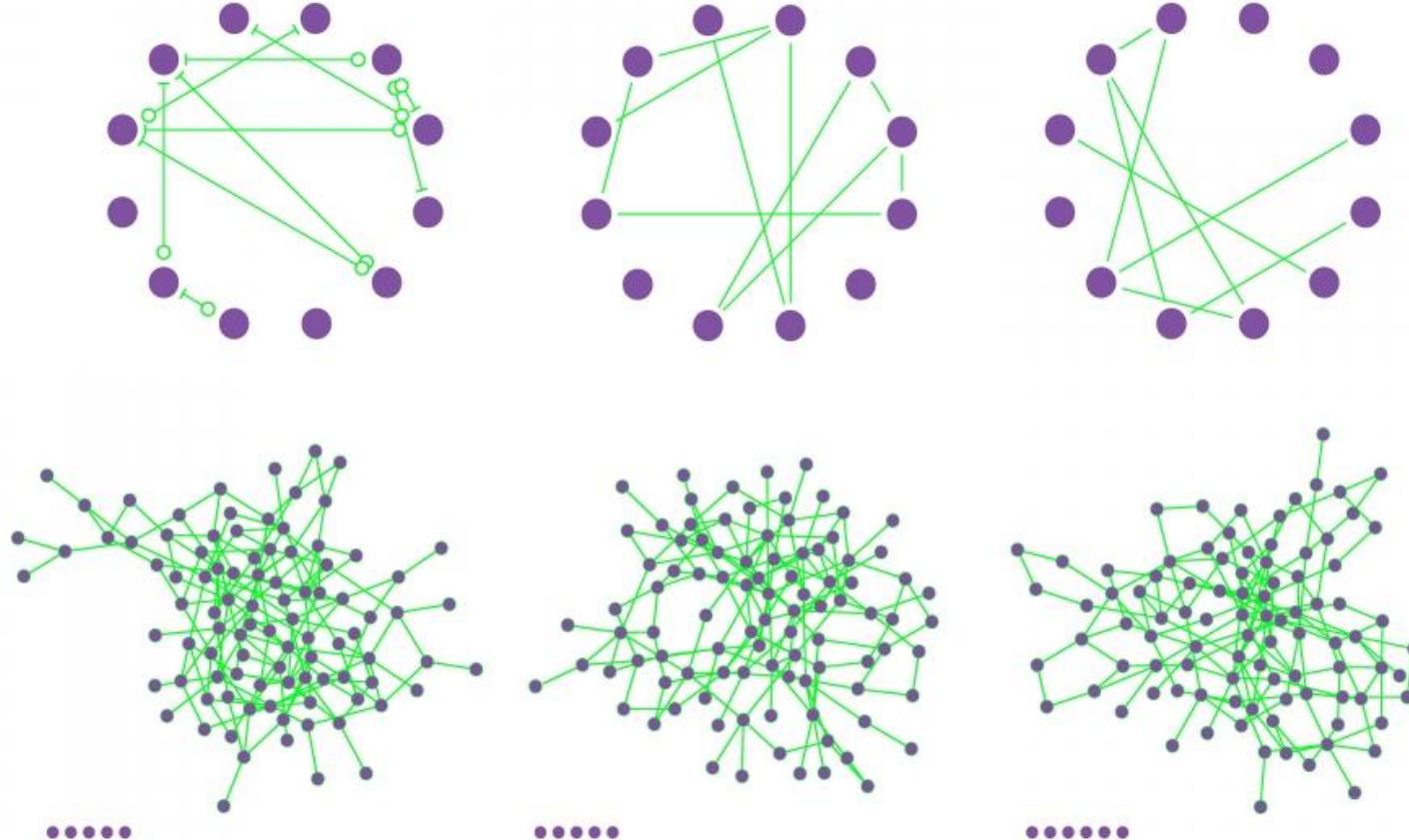
- $G(N, L)$ 模型: N 个节点通过 L 条随机边相连
- $G(N, p)$ 模型: N 个节点两两之间以 p 的概率相连。

生成随机网络可以遵循以下的步骤：

- 从 N 个孤立节点开始
- 选择一个节点对，并随机生成一个0到1之间的随机数。如果这个随机数大于 p ，那么就用一条边将两个节点相连，否则它们保持不连接
- 在所有的 $N(N - 1)/2$ 个节点对上重复步骤二

通过这种方法生成的网络被称作随机图或者随机网络。此外，为了纪念 Pál Erdős 和 Alfréd Rényi 对于随机网络的卓越贡献，随机网络也被称为Erdős-Rényi网络。





随机网络是随机的:同一行为设置了相同 (N, p) 的随机网络, 可以看到他们的结构并不相同。



对于节点数 N 相同的随机网络， p 的大小对网络的形状有着很大的影响。实体的链接情况和边的数量都受到影响。因此，对于固定 N 和 p 的随机网络，估计他们的边的数量的期望是很有意义的。

- L 个链接在 $N(N - 1)/2$ 个节点对之间成功生成的概率为 p^L 。
- 其余 $N(N - 1)/2 - L$ 个节点对之间没有生成边的概率为 $(1 - p)^{N(N-1)/2-L}$ 。
- 组合数 $\binom{\frac{N(N-1)}{2}}{L}$ 表示了从 $N(N - 1)/2$ 个节点对中选择 L 个可能出现的情况数量。

那么我们就可以写出出现 L 条边的概率为：

$$p_L = \binom{\frac{N(N-1)}{2}}{L} p^L (1 - p)^{N(N-1)/2-L}$$



L 服从二项分布，它的期望为：

$$\langle L \rangle = \sum_{L=0}^{\frac{N(N-1)}{2}} L p_L = p \frac{N(N-1)}{2}$$

$\langle L \rangle$ 的上界为网络的最大链接数 $L_{\max} = N(N - 1)/2$ 。利用上式可以得到随机网络的平均度期望：

$$\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N - 1)$$

$\langle k \rangle$ 的上界为节点最多可以有的链接数 $N - 1$ 。

总的来说，网络中的链接数量与 N 和 p 相关。如果我们逐渐增加 p ，链接数的期望会从0增加到 L_{\max} ，节点的平均度期望会从0增加到 $N - 1$



随机网络的度分布

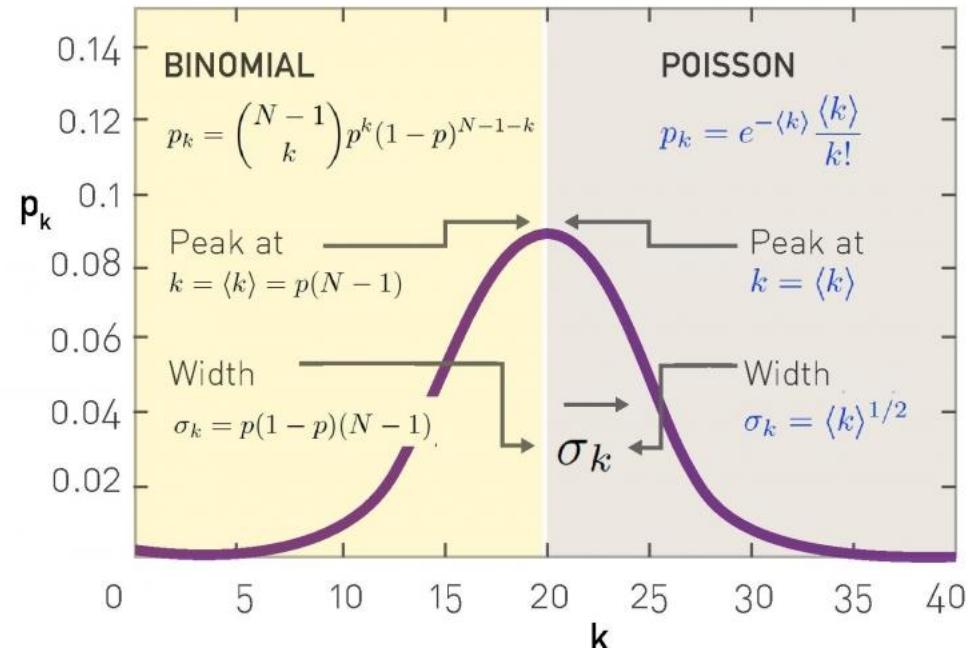
随机网络的节点度的分布服从二项分布：

- 有 k 条边的概率为 p_k
- 不和其他节点连接的概率为 $(1 - p)^{N-1-k}$
- 从 $N - 1$ 个节点选 k 个节点的情况有 $\binom{N-1}{k}$ 种情况

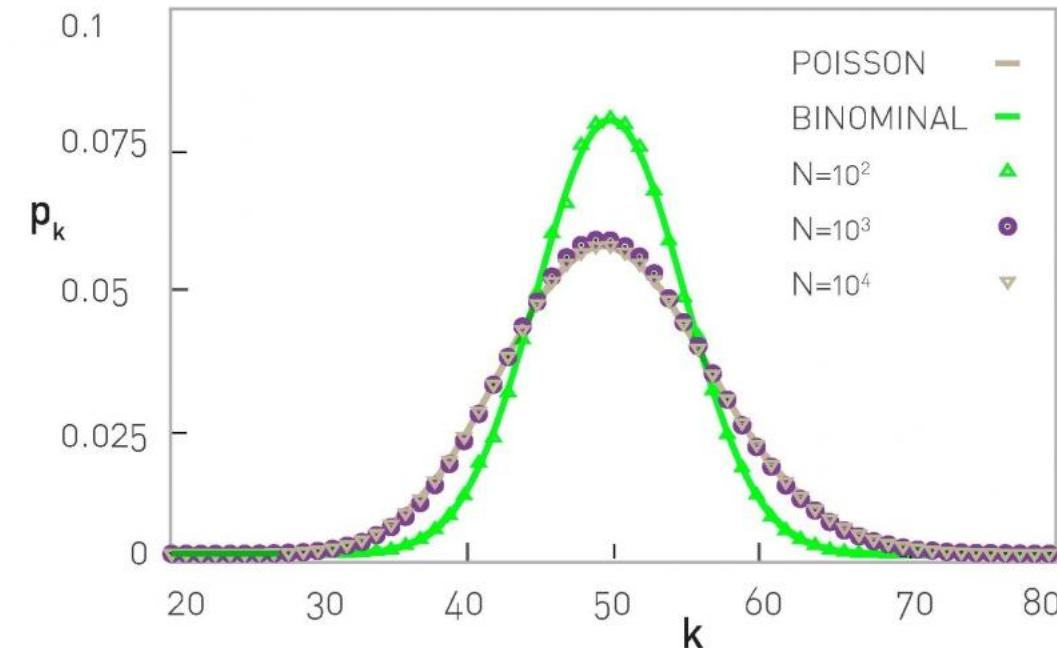
那么度分布为：

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

度分布是由网络规模 N 和概率 p 决定的，通过二项分布的性质，可以计算出平均度 $\langle k \rangle$ ，二阶矩 $\langle k^2 \rangle$ 和方差 σ_k 。



二项分布 vs. 泊松分布: 二项分布的极限分布为泊松分布。 $\langle k \rangle \ll N$ 时, 度分布可以由泊松分布表示。在运算中通常使用泊松分布, 因为它只有一个参数。



度分布是依赖于网络规模的: 不同规模的度分布度分布不同。随着网络规模的增加, 度分布逼近泊松分布。

网络中最大的连接集群的大小 N_G 是如何随 k 变化的

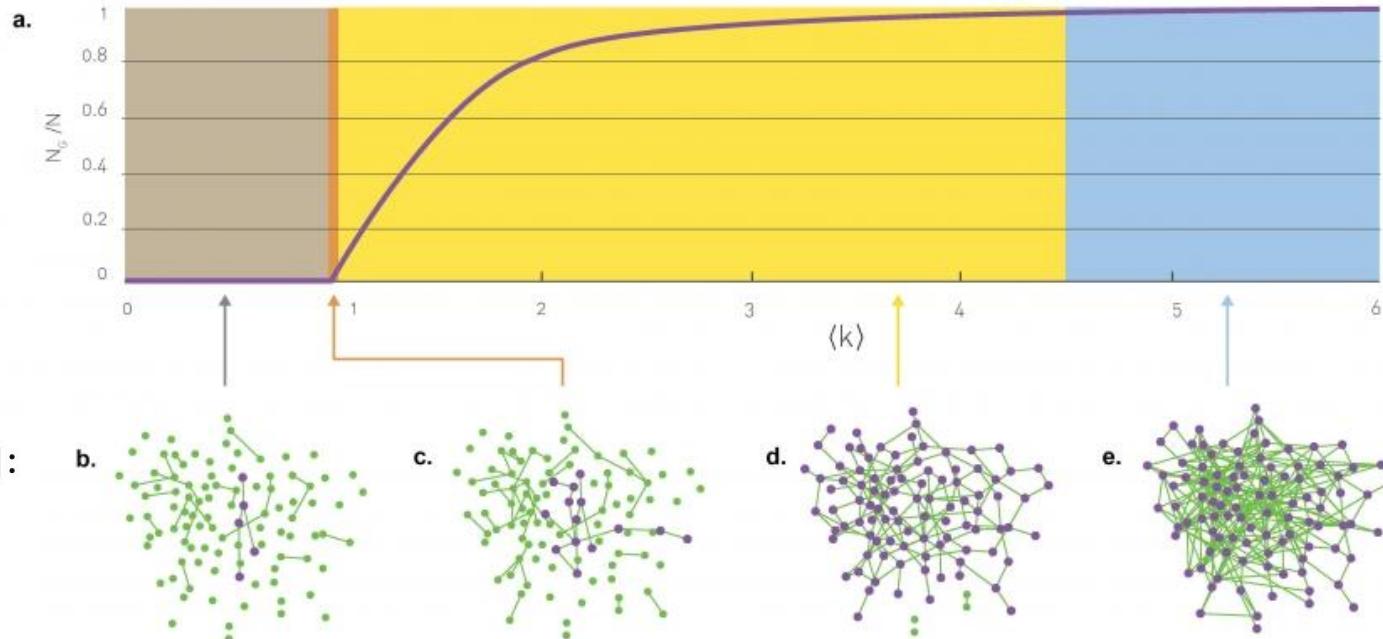
Erdős 和 Rényi 在其 1959 年的经典论文中预测，大型连通子图出现的条件是：

$$\langle k \rangle = 1$$

在 $\langle k \rangle = 1$ 的情况下，通过之前得到的 p 与 k 的关系，可以得到：

$$p = \frac{1}{N-1} \sim \frac{1}{N}$$

因此网络越大，出现大连通分量所需的 p 就越小。



随机网络的演化

亚临界： $\langle k \rangle < 1$

临界： $\langle k \rangle = 1$

超临界： $\langle k \rangle > 1$

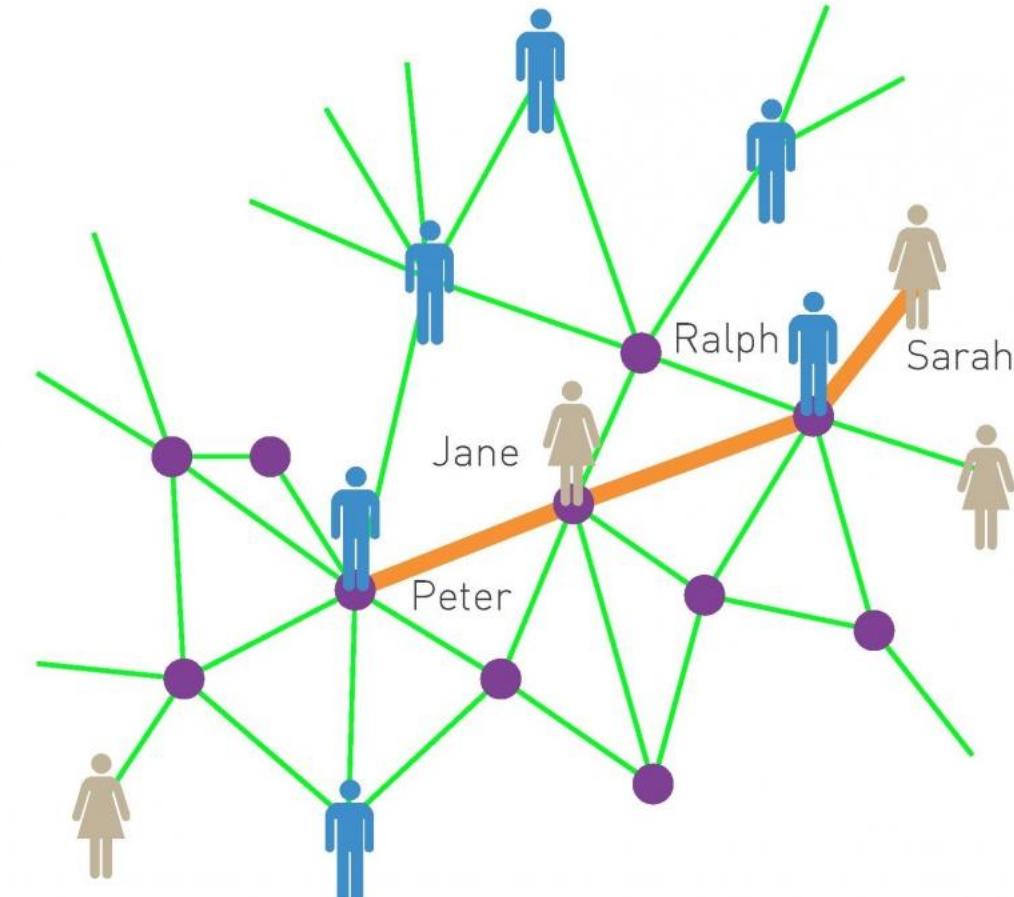
完全链接： $\langle k \rangle > \ln N$

小世界理论(*small world phenomenon*)，也称为六度分隔(*six degrees of separation*)，它指出，如果你选择地球上任何地方的任何两个人，你将在他们之间找到最多六个相识的路径。

考虑一个平均度为 $\langle k \rangle$ 的网络有如下特征：

- 网络中距离为1的节点可以有 $\langle k \rangle$ 个
- 网络中距离为2的节点可以有 $\langle k \rangle^2$ 个
- 网络中距离为3的节点可以有 $\langle k \rangle^3$ 个
- ...
- 网络中距离为 d 的节点可以有 $\langle k \rangle^d$ 个

对于 $\langle k \rangle \approx 1000$ 的社会网络而言，距离为3的熟人就有十亿个



我们可以计算出距离 d 以内的节点数量的上限：

$$N(d) \approx 1 + \langle k \rangle + \langle k \rangle^2 + \cdots + \langle k \rangle^d = \frac{\langle k \rangle^{d+1} - 1}{\langle k \rangle - 1}$$

$N(d)$ 不能超过网络中节点总数 N ，而且 $\langle k \rangle \gg 1$ ，那么有：

$$d_{\max} \approx \frac{\ln N}{\ln \langle k \rangle}$$

这就是代表了小世界理论的公式。

对于社会网络来说，使用 $N \approx 7 \times 10^9$ 和 $\langle k \rangle \approx 1000$ 可以得到：

$$\langle d \rangle \approx \frac{\ln 7 \times 10^9}{\ln \langle 1000 \rangle} = 3.28$$

因此地球上所有个人之间的距离只有三到四个人。这个估计可能比经常引用的6度理论更接近实际值。



提纲

1

复杂科学简介

2

图的基本概念

3

随机网络

4

无标度网络

5

社团检测



回顾: Erdős–Rényi随机网络的度分布

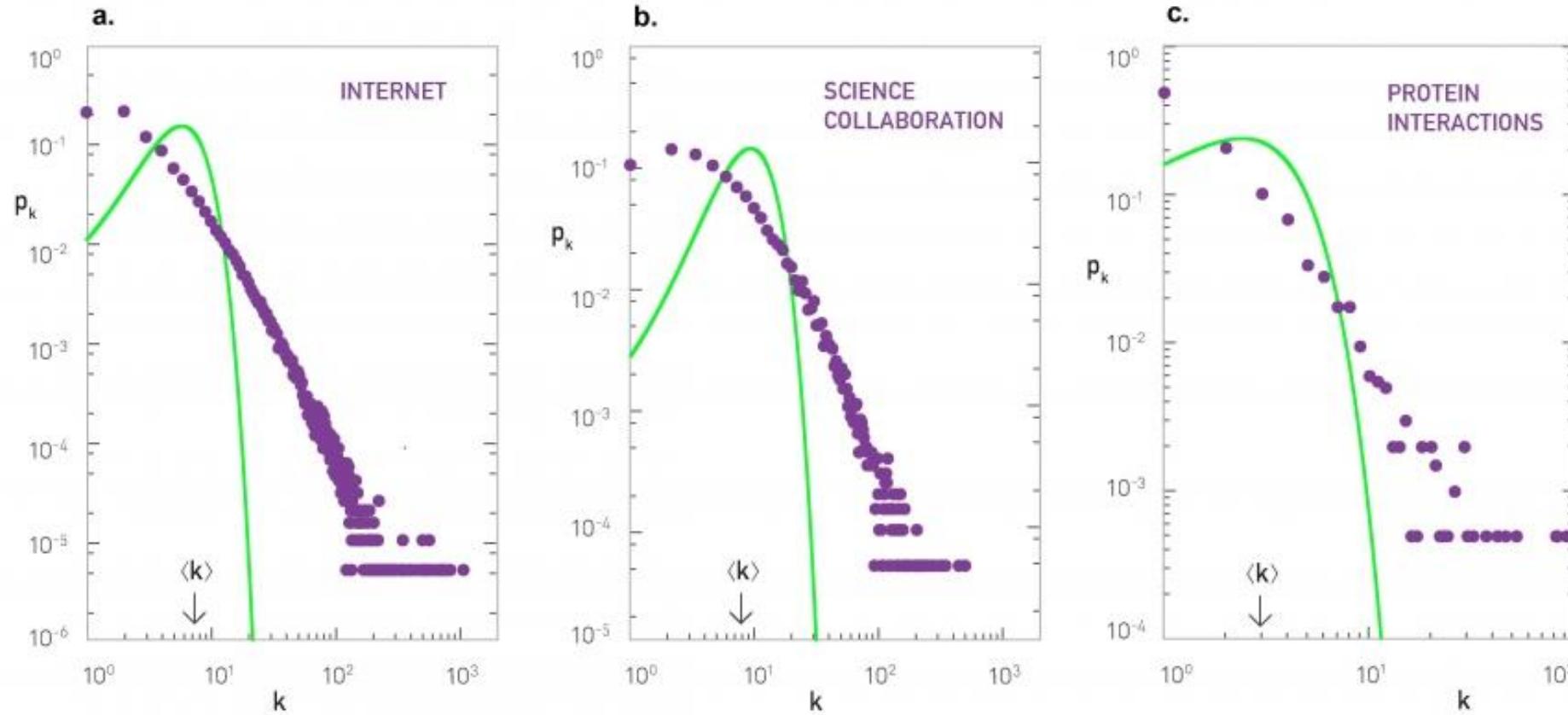
随机网络的节点度的分布服从二项分布:

- 有 k 条边的概率为 p_k
- 不和其他节点连接的概率为 $(1 - p)^{N-1-k}$
- 从 $N - 1$ 个节点选 k 个节点的情况有 $\binom{N-1}{k}$ 种情况

那么度分布为:

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

度分布是由网络规模 N 和概率 p 决定的, 通过二项分布的性质, 可以计算出平均度 $\langle k \rangle$, 二阶矩 $\langle k^2 \rangle$ 和方差 σ_k 。



真实网络的度分布: a, b, c中紫色的点分别是因特网、科学合作网络和蛋白质相互作用网络的度分布。绿色的线为泊松分布的预测值。数据与泊松拟合之间的显著偏差表明，随机网络模型低估了**高度节点的大小和频率**，以及**低度节点的数量**。相反，随机网络模型预测的 $\langle k \rangle$ 附近的节点数量比在真实网络中看到的要多。

美国家庭收入分布

正态分布 X

泊松分布 X

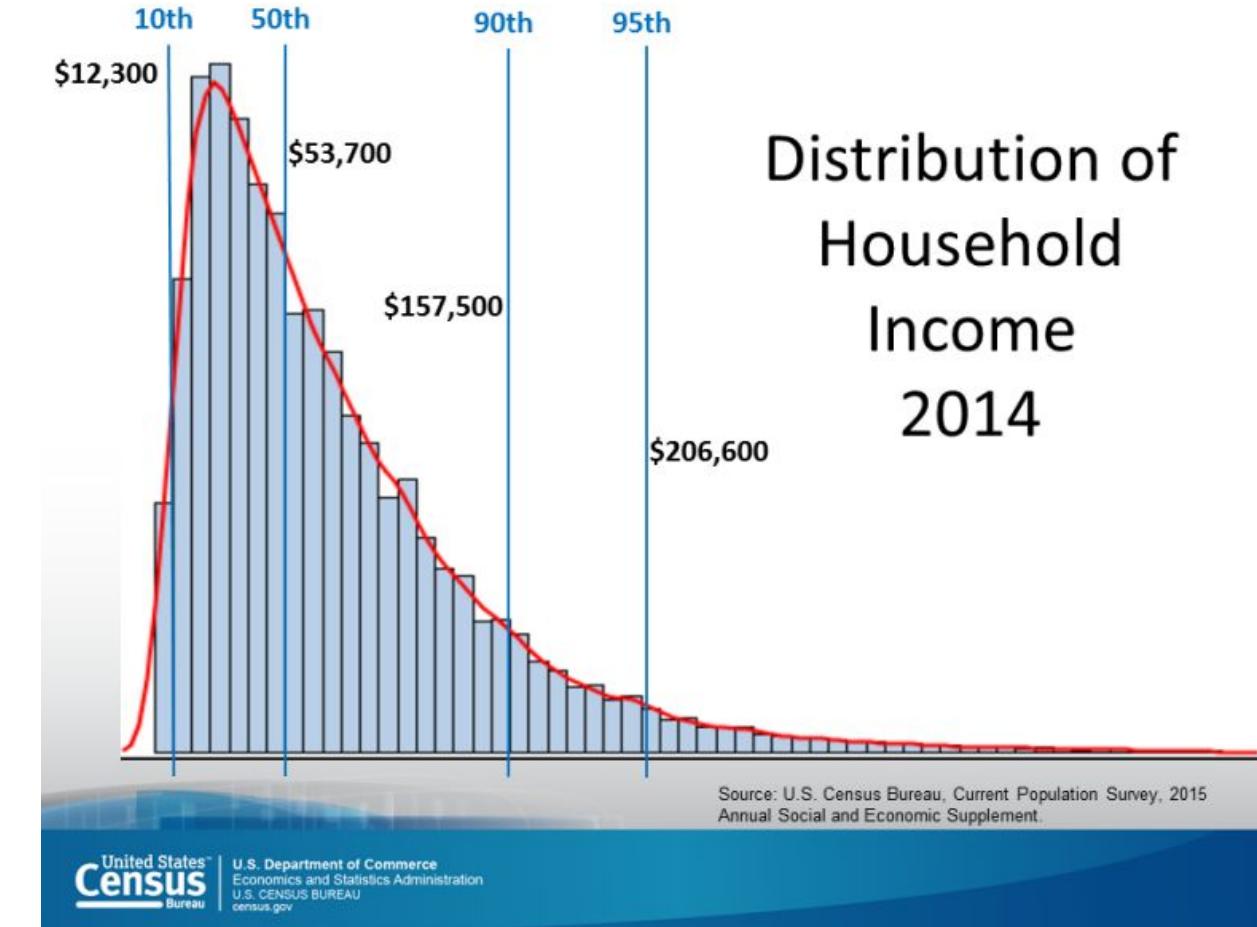
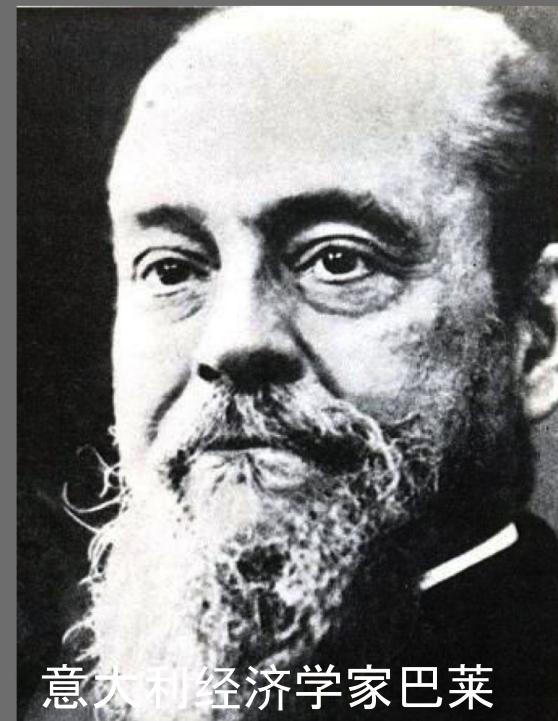
具备长尾分布特征

Box 4.1

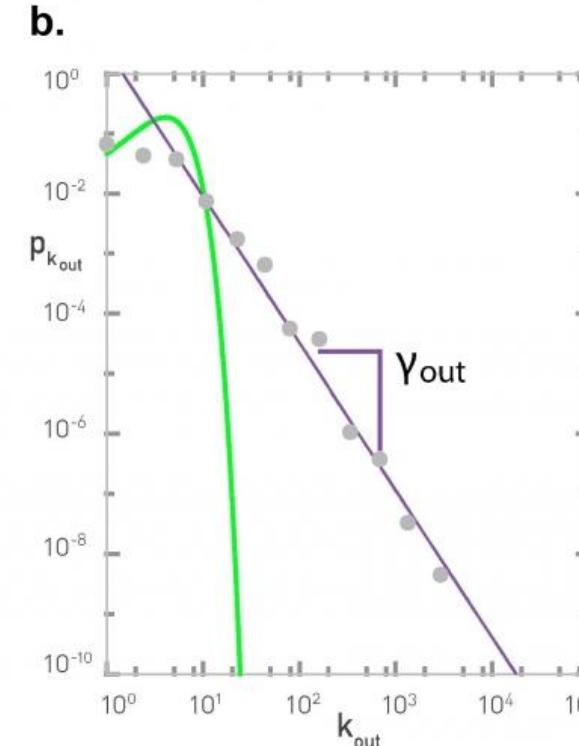
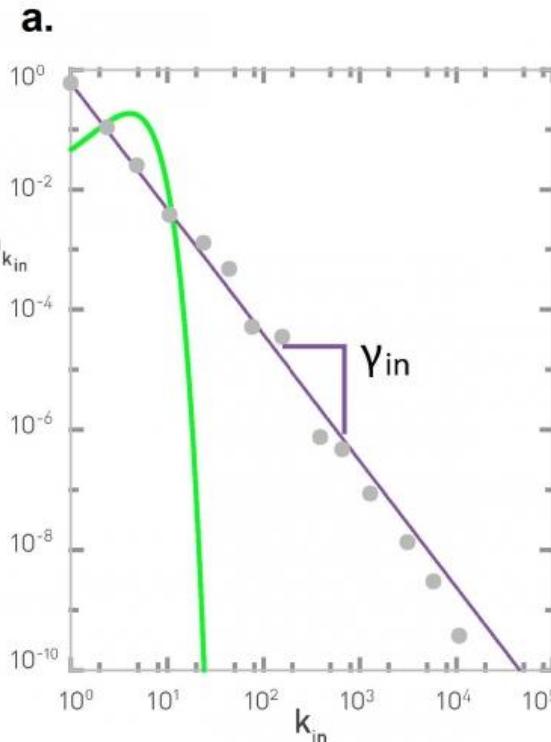
The 80/20 Rule and the Top One Percent

Vilfredo Pareto, a 19th century economist, noticed that in Italy a few wealthy individuals earned most of the money, while the majority of the population earned rather small amounts. He connected this disparity to the observation that incomes follow a power law, representing the first known report of a power-law distribution [3]. His finding entered the popular literature as the *80/20 rule*: Roughly 80 percent of money is earned by only 20 percent of the population.

The 80/20 rule emerges in many areas. For example in management it is often stated that 80 percent of profits are produced by only 20 percent of the employees. Similarly, 80 percent of decisions are made during 20 percent of meeting time.

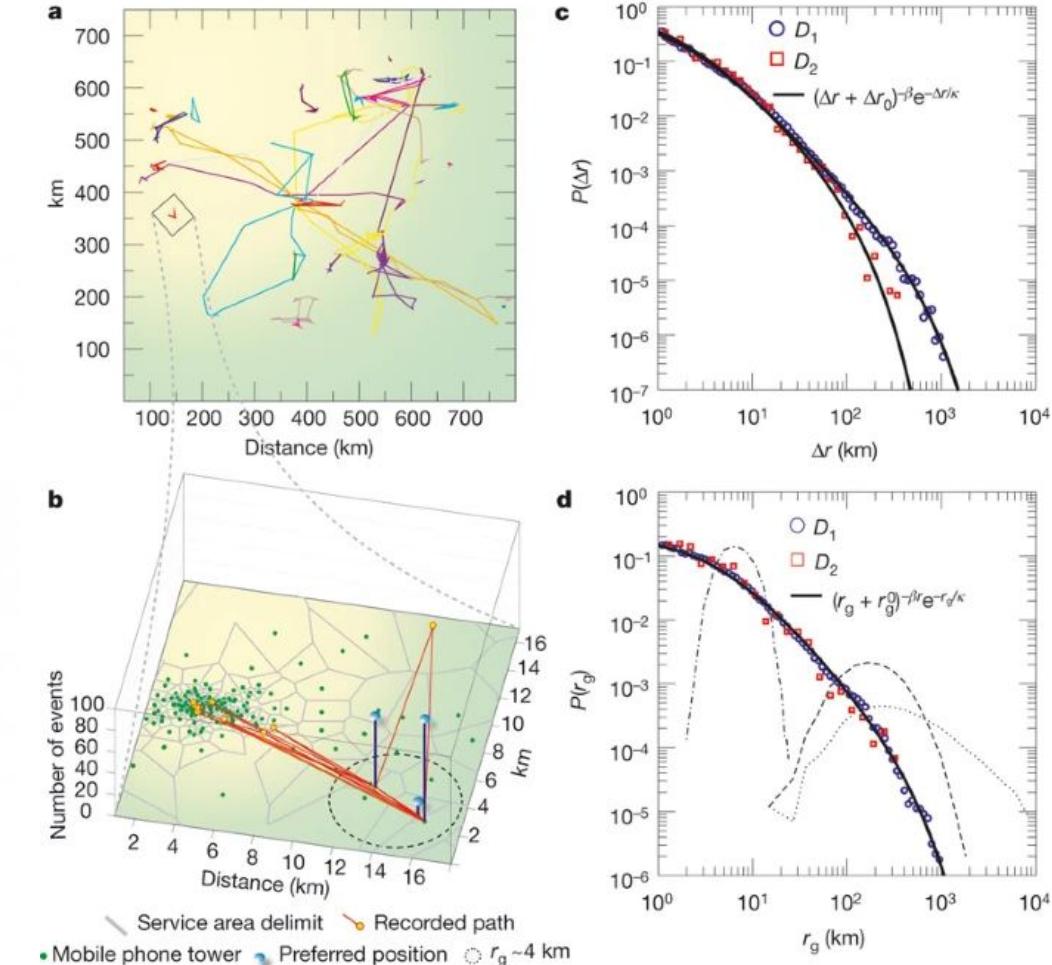


幂律分布(Power-Law)在自然界和社会系统中普遍存在



互联网WWW的入度和出度分布

$$p_k = Ck^{-\gamma}$$



人类移动行为符合幂律分布

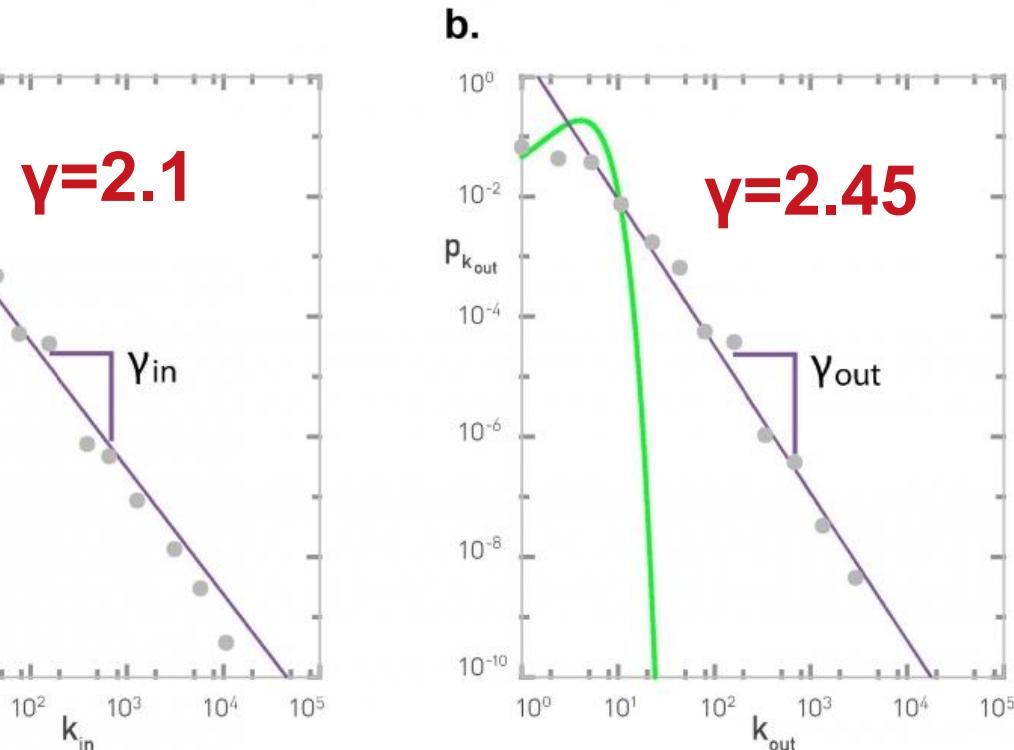
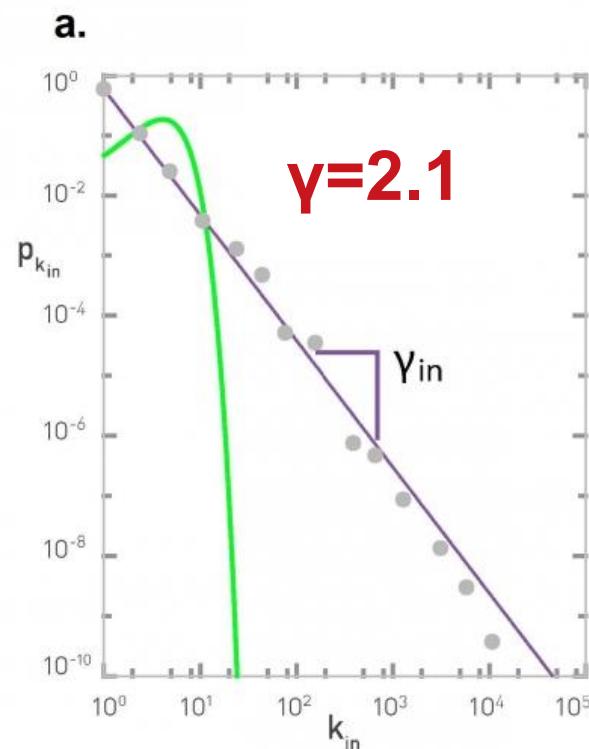
- 对于一个幂律分布 $p_k = Ck^{-\gamma}$

，通常关心指数 γ 的取值；

- 在讨论无标度网络的度分布时，

称其为**度指数**；

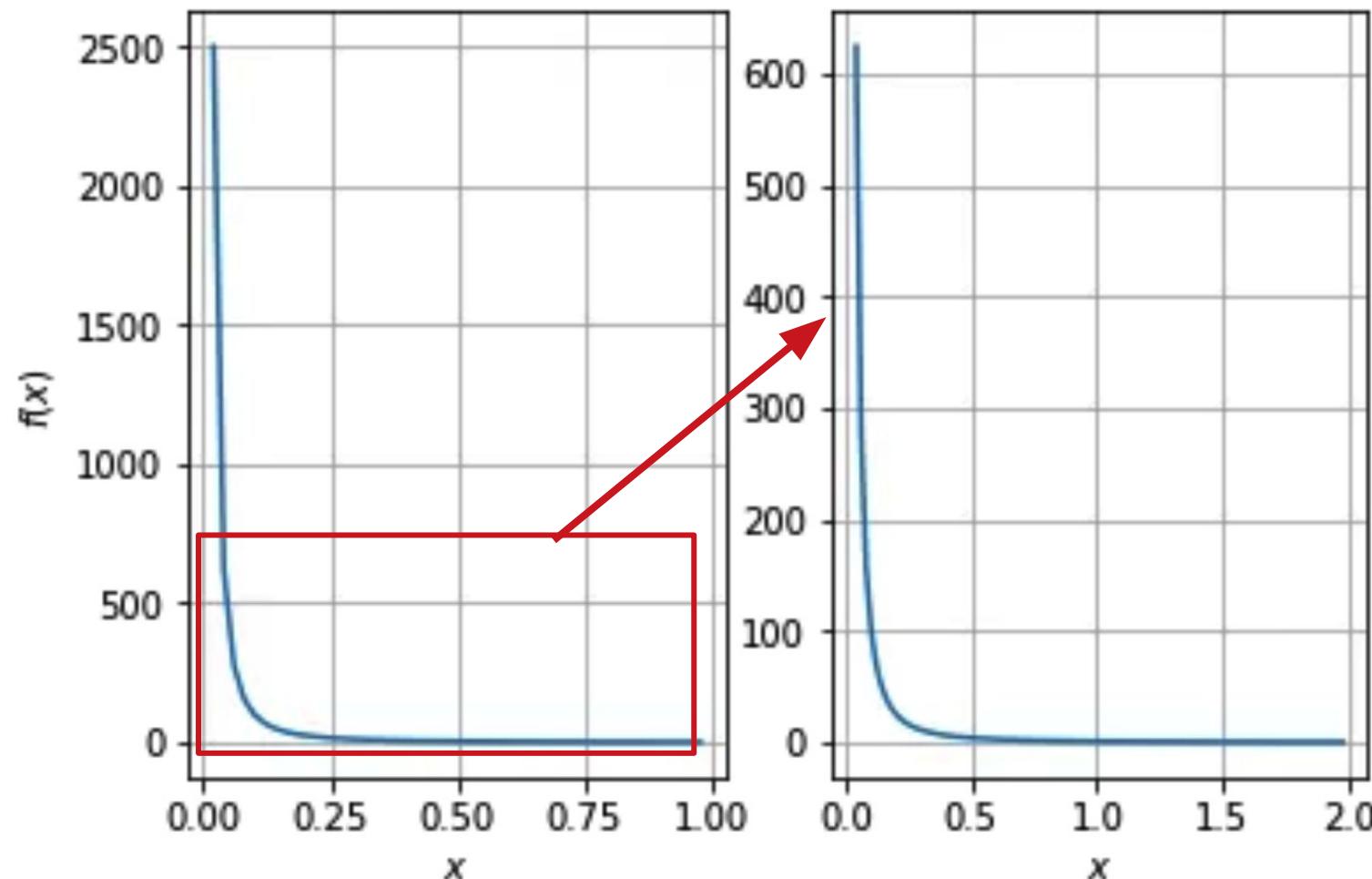
定义：无标度网络是度分布服从幂律分布的网络

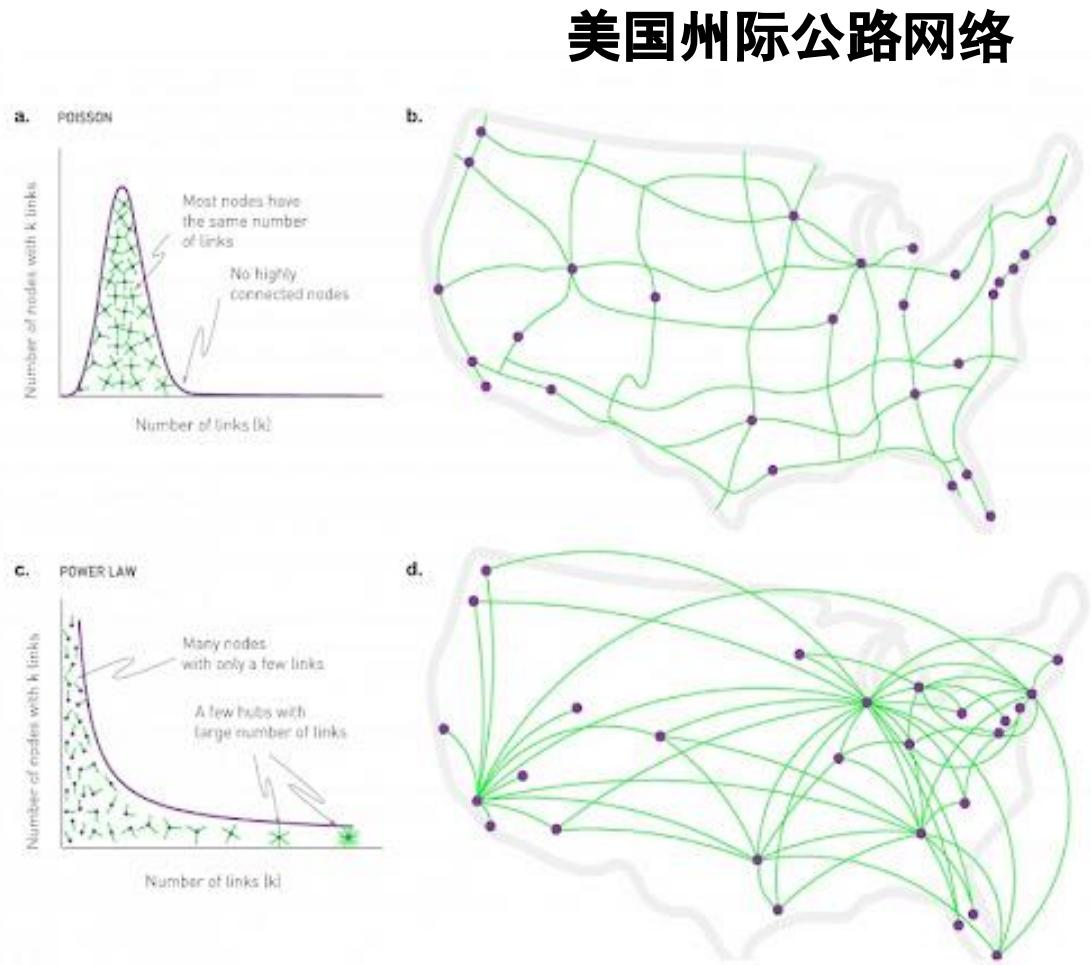


互联网WWW的入度和出度分布

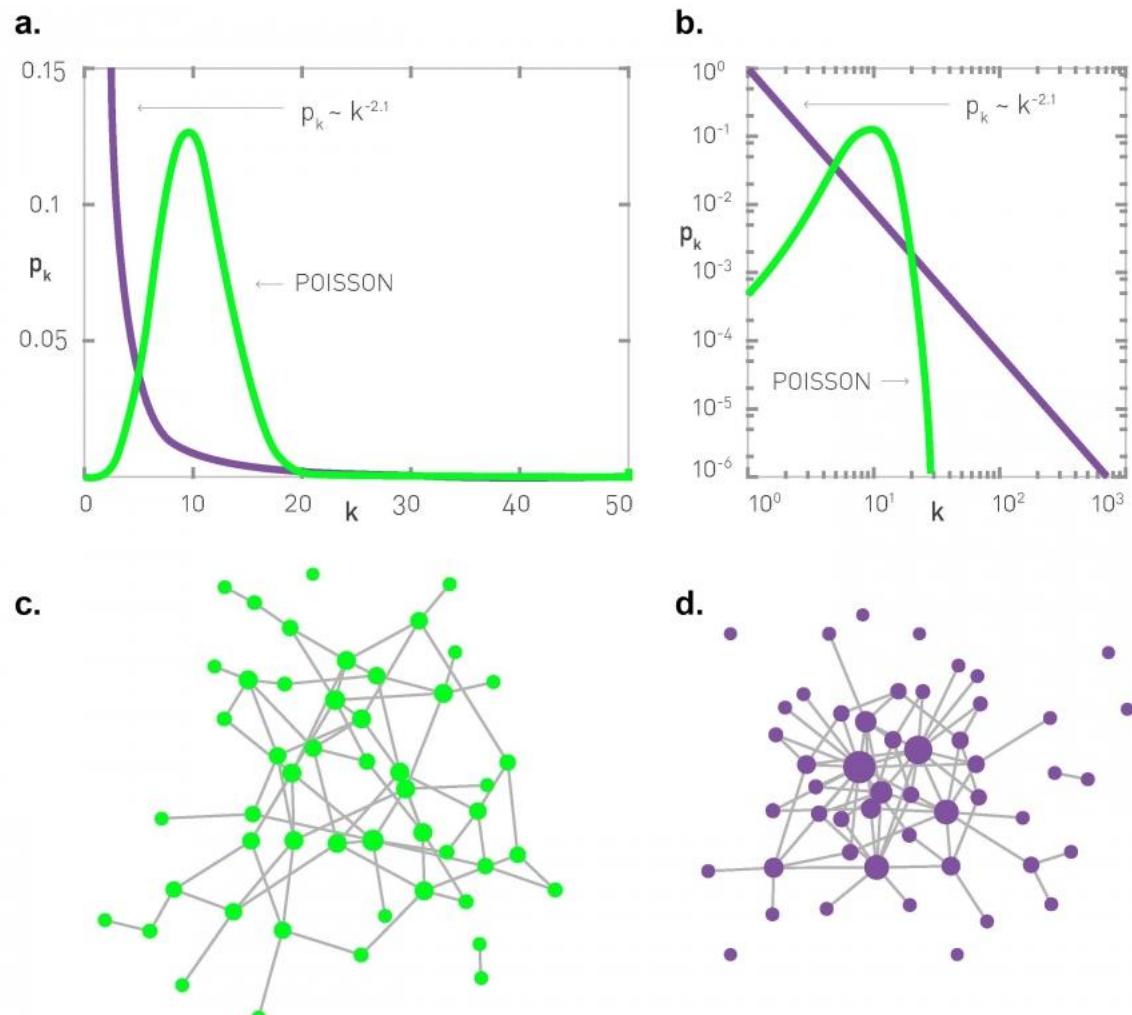
无标度一词来源于幂律分布的标度不变性，指的是幂指数函数 $f(x) = ax^{-k}$ 的自变量 x 经过某个常数放大或缩小 c 倍后，其函数关系仍然为幂指数函数。即： $f(cx) = a(cx)^{-k} = c^{-k} f(x) \propto f(x)$ ，下图可以更直观的看到这一现象，其中左图为 $f(x) = x^{-2}$ ，右图为 $f(x) = (2x)^{-2}$ 。

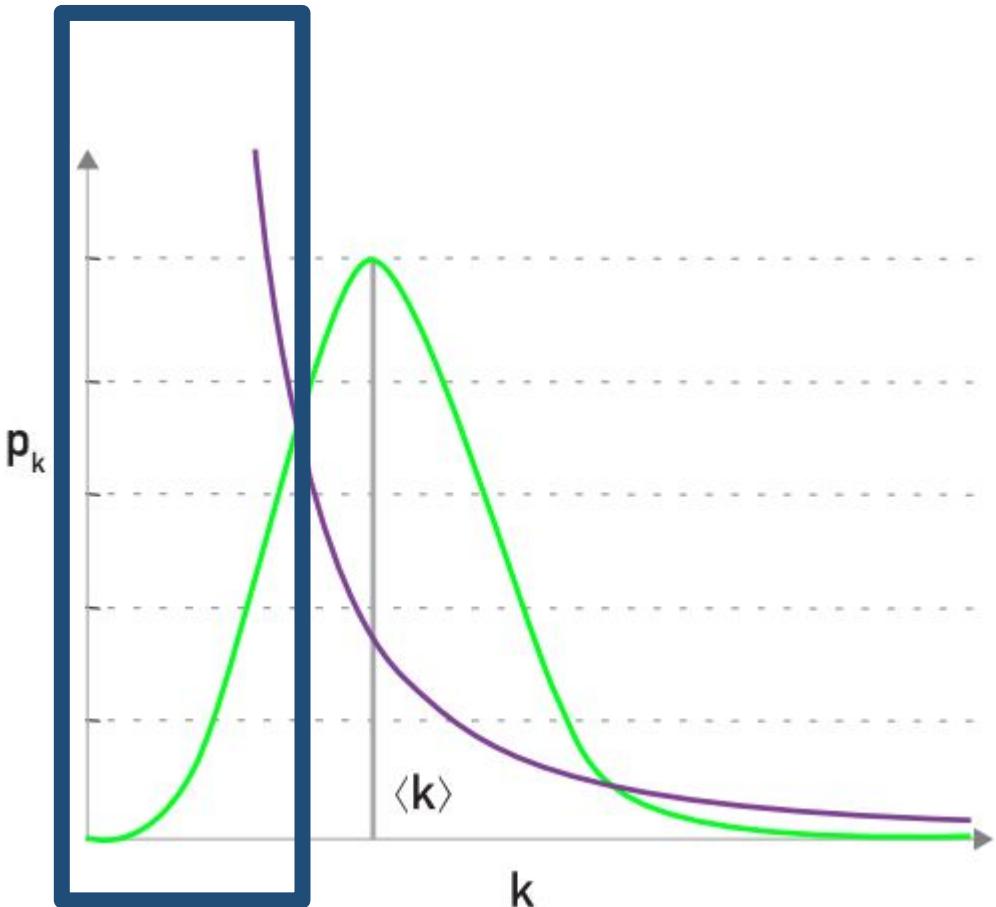
幂律分布标度不变
, $a=1$, $k=2$, $c=0.5$





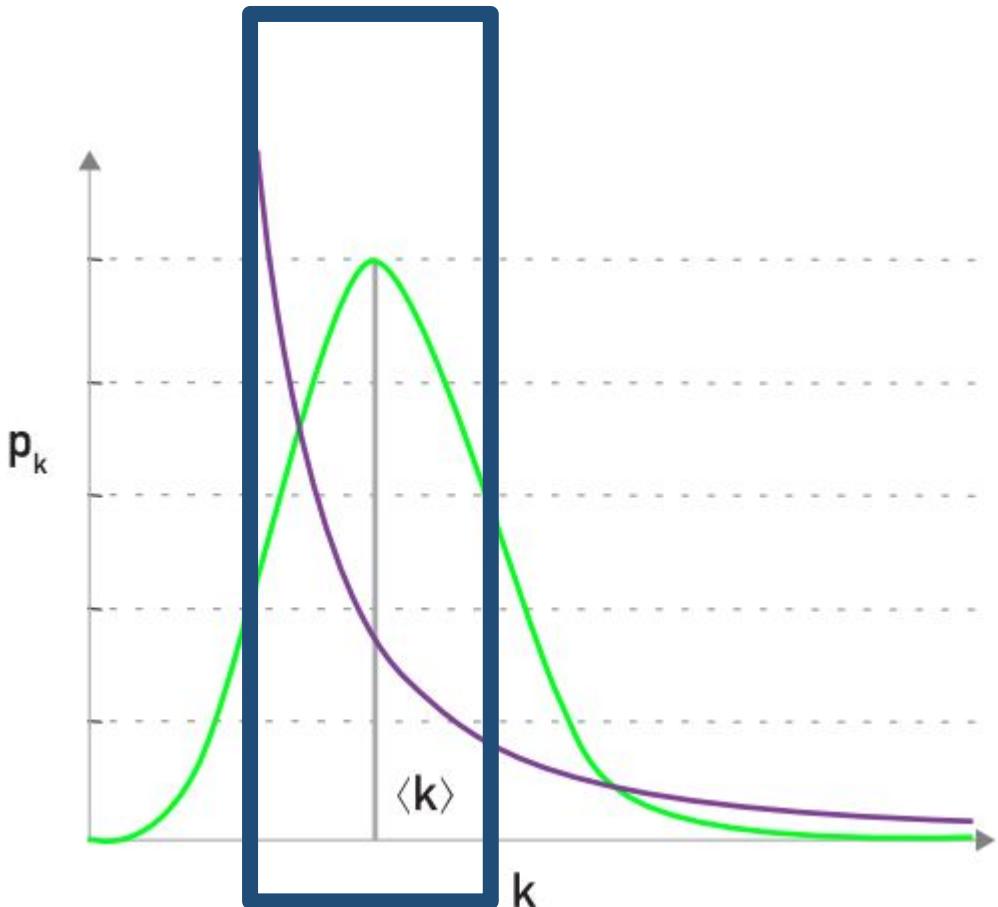
美国航班网络





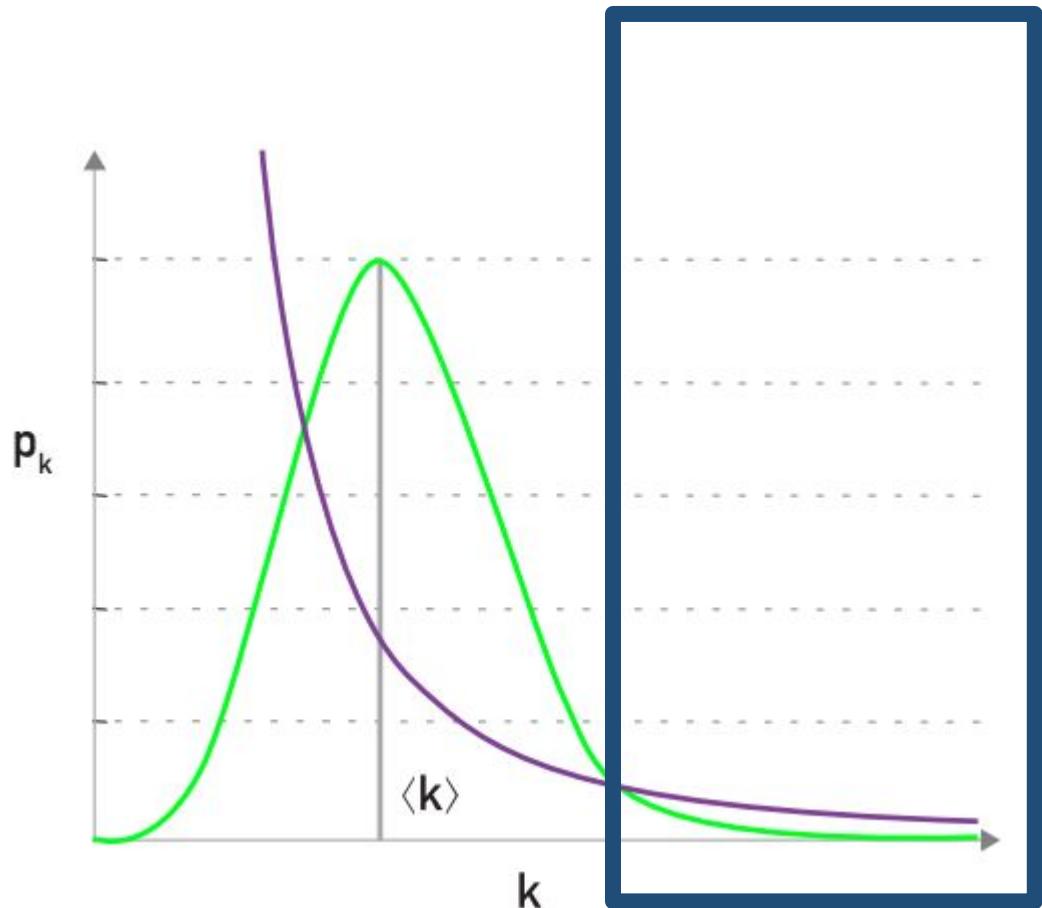
- k 比较小时, 幂律分布在泊松分布之上;
- 说明无标度网络中有大量度比较小的节点, 而随机网络中这样的节点很少

平均度相同的随机网络与无标度网络的度分布

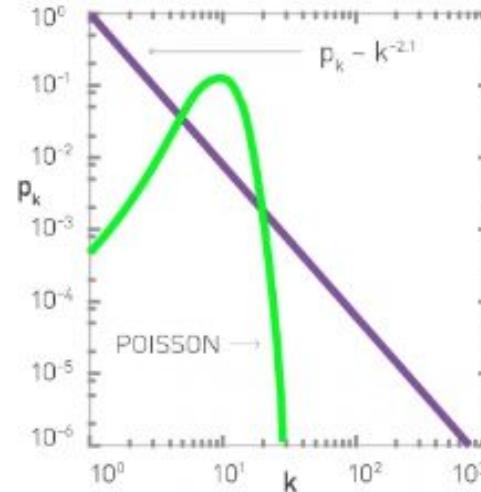


- k 在 $\langle k \rangle$ (平均度)附近时, 泊松分布在幂律分布之上
- 说明随机网络大部分节点的度在平均度附近;

平均度相同的随机网络与无标度网络的度分布



平均度相同的随机网络与
无标度网络的度分布



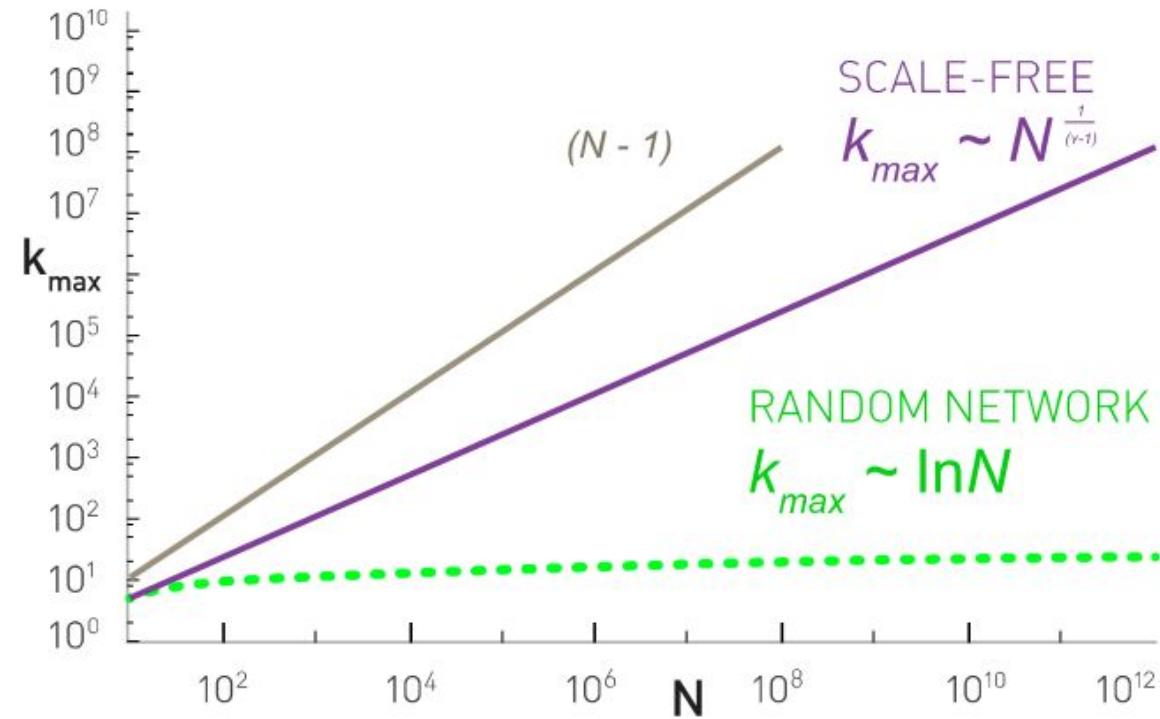
- **k比较大时, 幂律分布再次位于泊松分布之上, 在双指数坐标系下差异格外明显;**
- **这一现象表明, 在无标度网络中观察到一个大度节点——枢纽节点的概率, 要比在随机网络中大好几个数量级**

- 网络大小N如何影响枢纽节点的大小？

- 线性？
 - \log ？

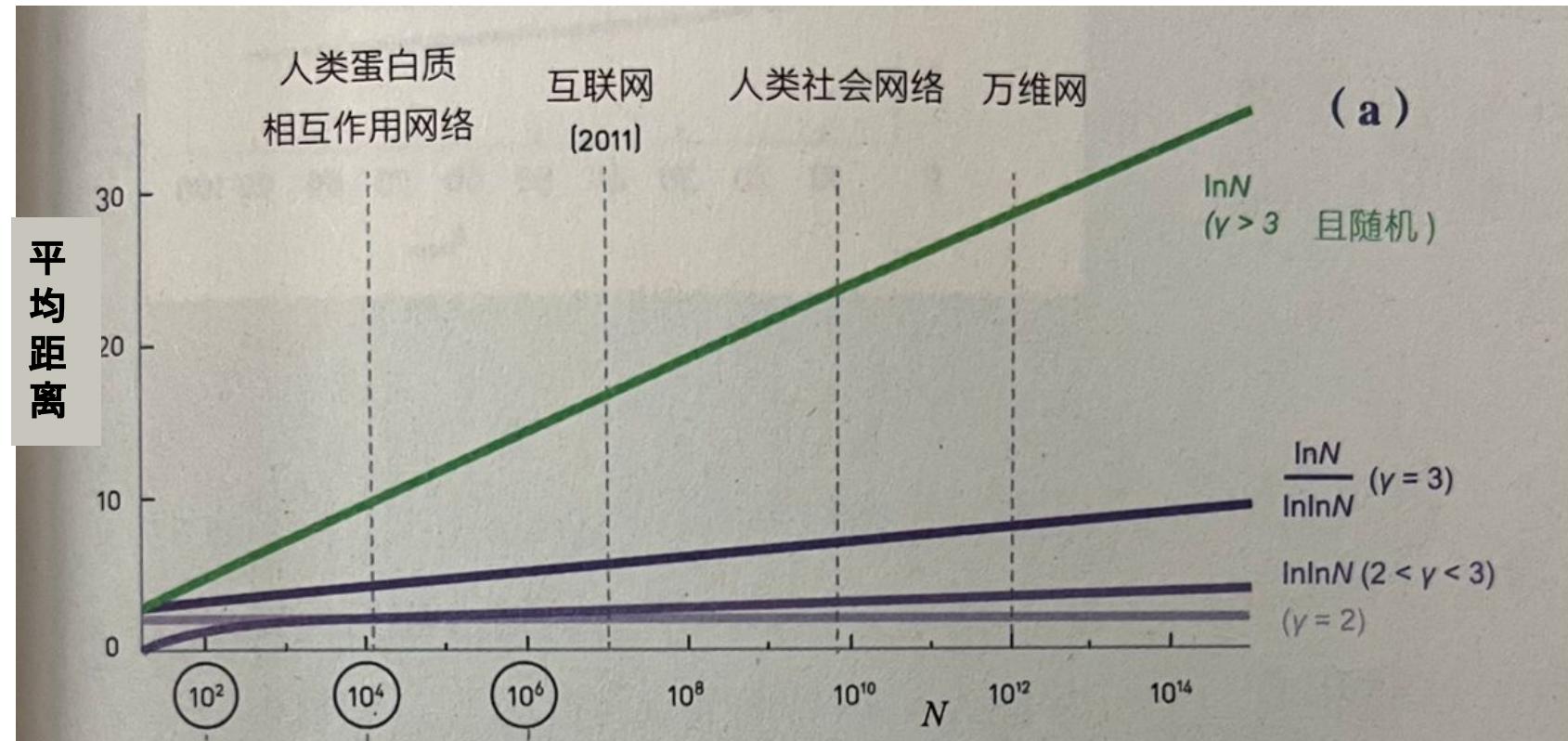
- 定义最大度 k_{\max} , 也被称为度分布的自然截断(natural cutoff), 表示一个网络中最大枢纽节点大小的期望；

$$K_{\max} = K_{\min} N^{\frac{1}{\gamma}-1}$$



无标度网络中枢纽节点 会影响小世界性质吗？

- 计算网络中的平均距离 $\langle l \rangle$
- 航空业中通过修建中枢节点减少转机次数



枢纽节点更明显



**Ultra
Small
World**

$$< l > \sim \begin{cases} const. & \gamma = 2 \\ \frac{\ln \ln N}{\ln(\gamma - 1)} & 2 < \gamma < 3 \\ \frac{\ln N}{\ln \ln N} & \gamma = 3 \\ \ln N & \gamma > 3 \end{cases}$$

对比

- 随机网络中的平均距离和网络大小的关系是对数关系

$$< l > \approx \frac{\ln N}{\ln < k >}$$

**Small
World**

为什么随机网络中没有枢纽节点和幂律分布？

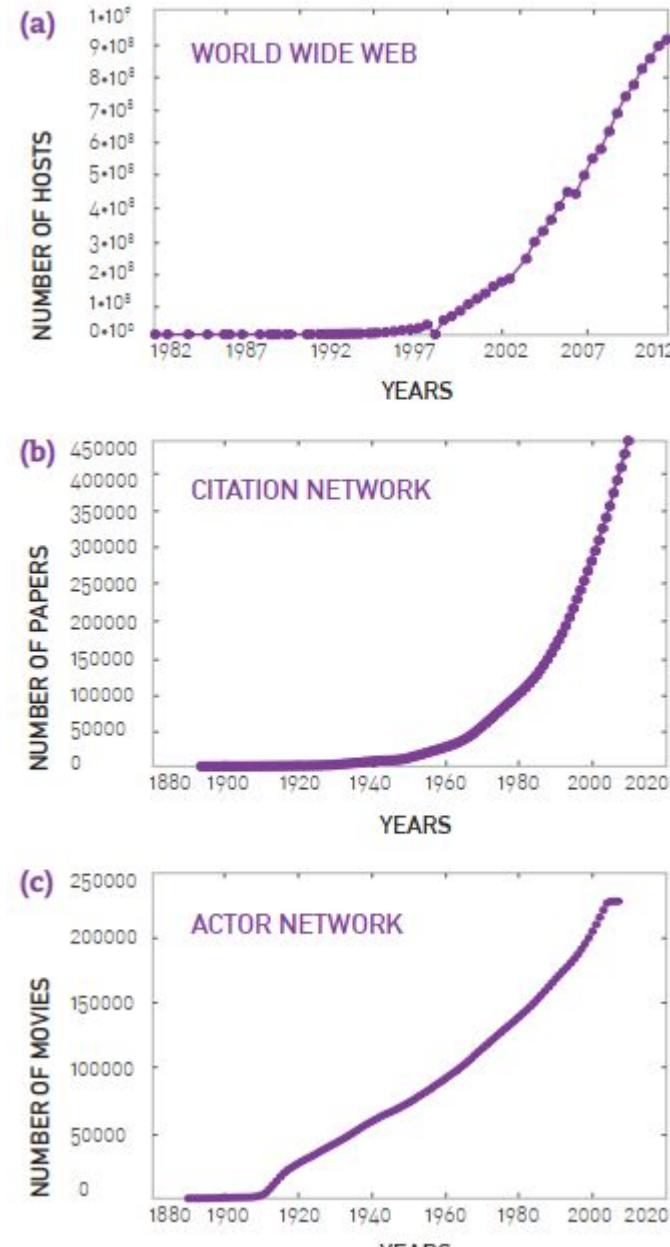
- 随机网络中隐藏了两个假设

1. 网络大小N是静态的

- 然而在真实网络中，由于新节点的加入，节点的数目是不断增长的(**生长**)

2. 节点随机地选择其他节点进行连接

- 在大多数真实网络中，新加入的节点更倾向于与连接数高的节点相连(**偏好连接**)



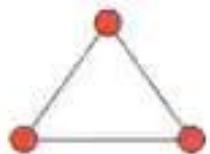
1. 生长

每次向网络中添加一个新节点
, 将其与固定m个节点相连

2. 偏好连接

新节点选择m个节点进行连接
时, 与节点 i 进行连接的概率为

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



BA模型的度分布将服从 $\gamma=3$ 的幂律分布

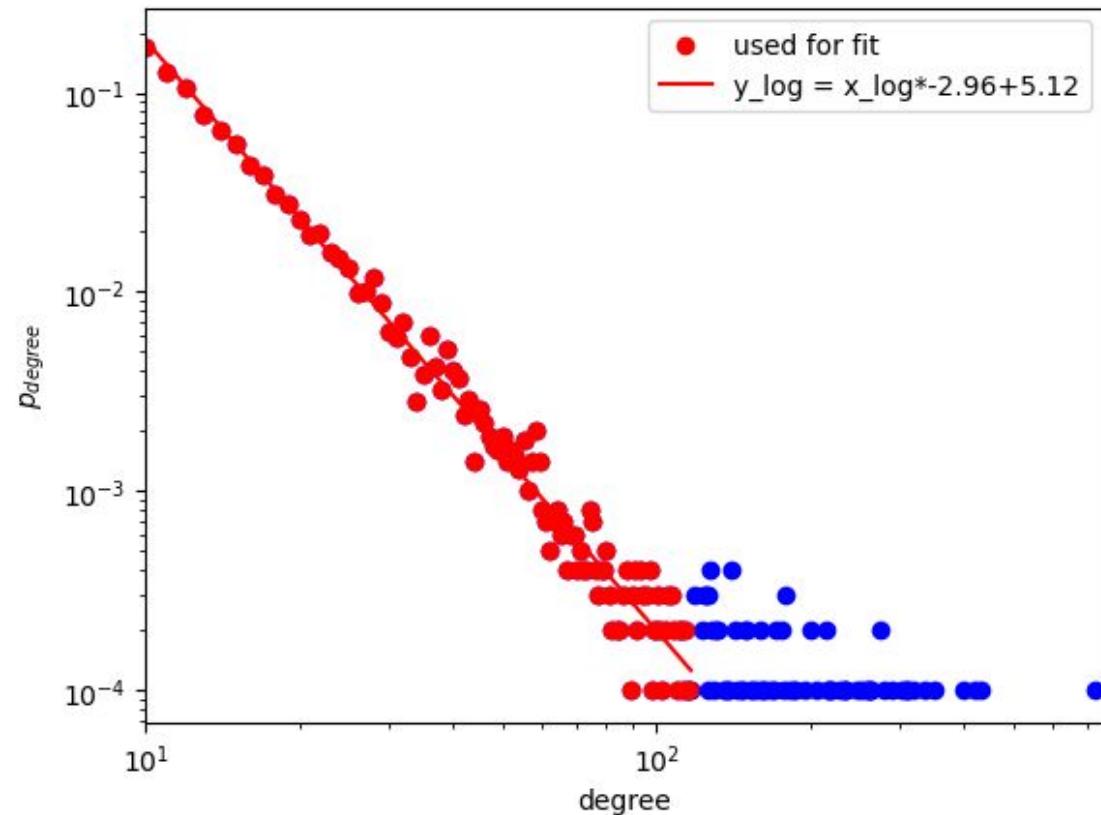
- 准确的度分布

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)}$$

Krapivsky, Redner, Leyvraz, PRL 2000

Dorogovtsev, Mendes, Samukhin, PRL 2000

Bollobas et al, Random Struc. Alg. 2001

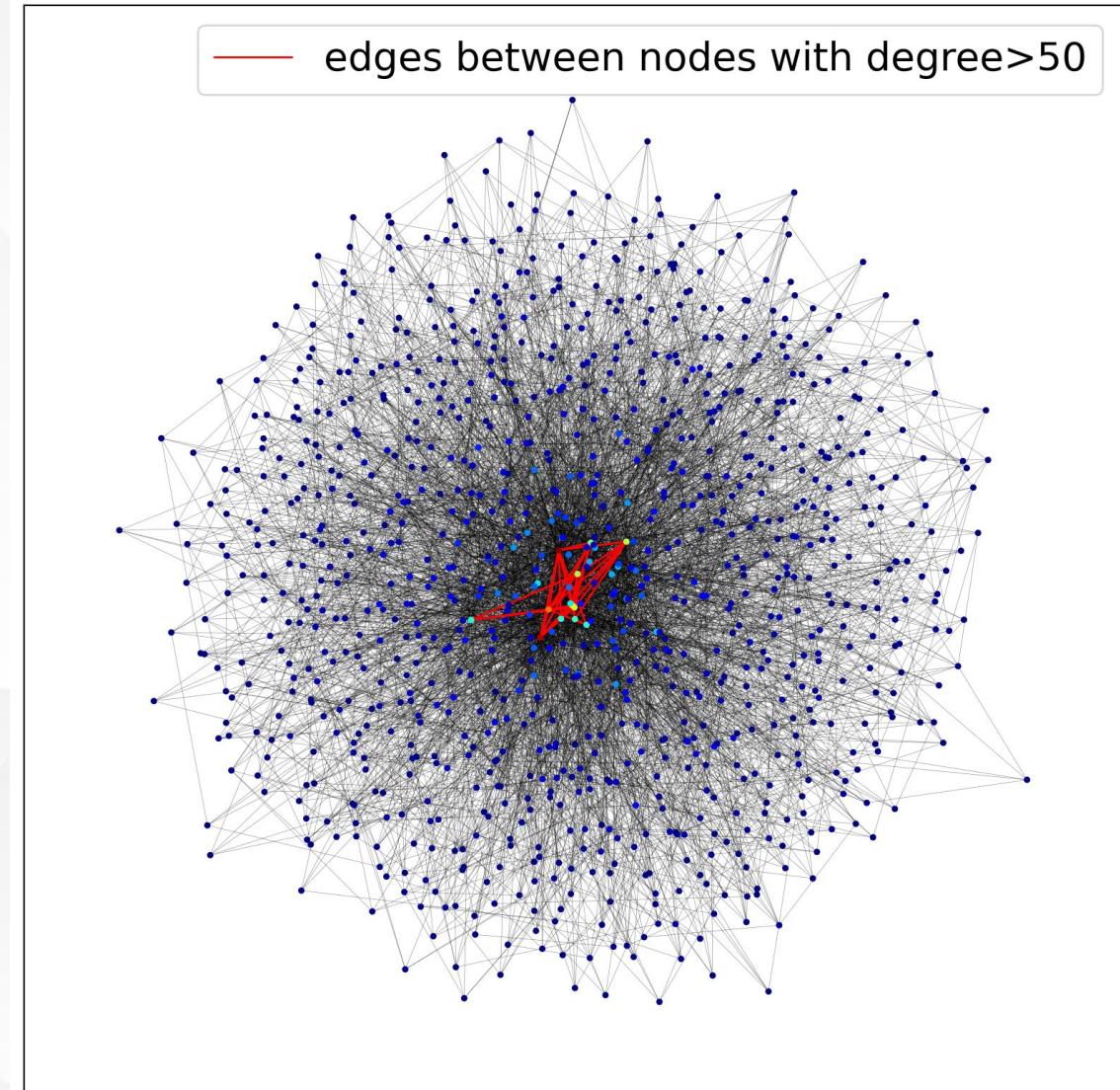


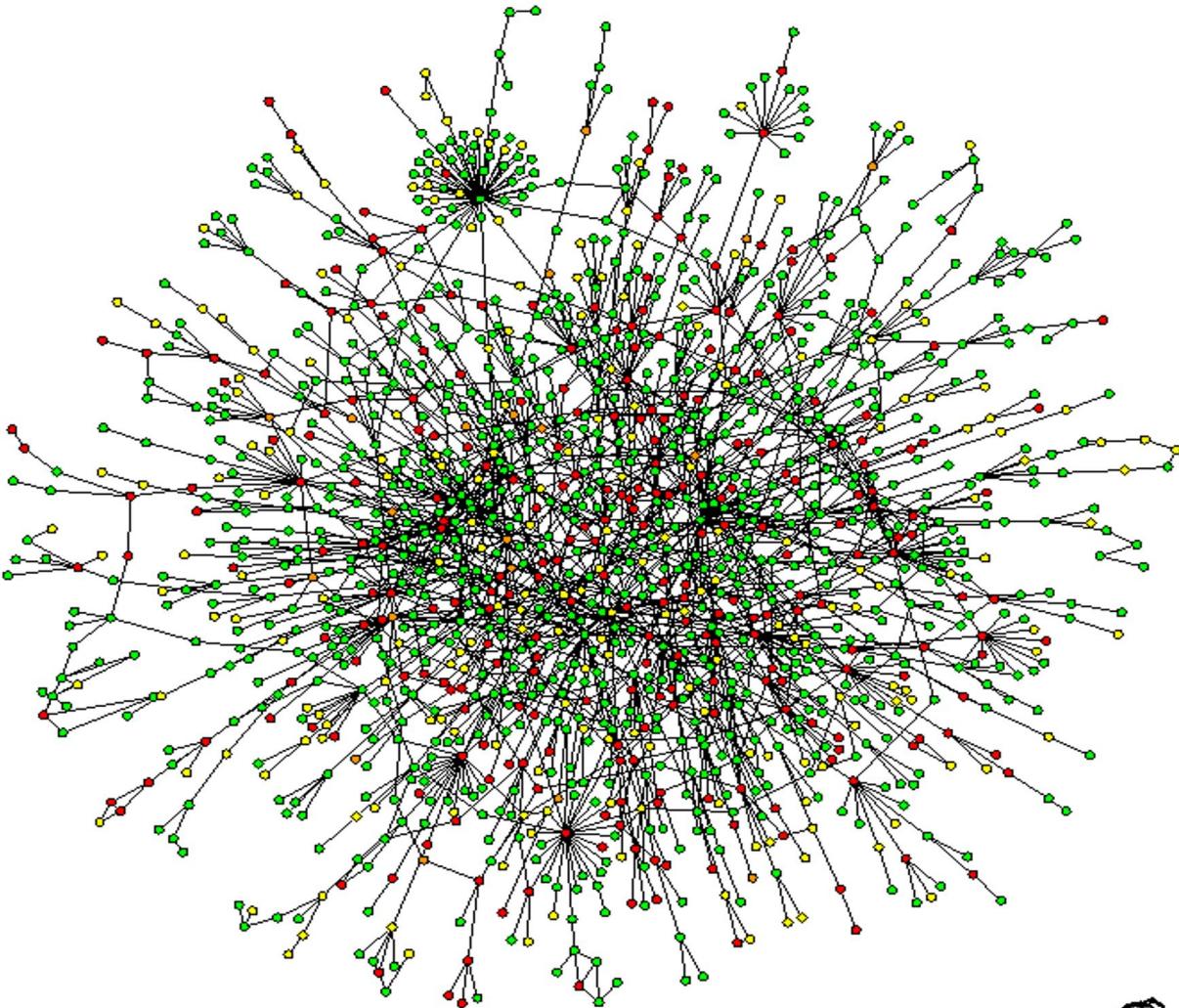
- m 为模型参数, 当 k 很大时, 可以简化为 $P(k) \sim 1/k^3$
- 度分布独立于 N , 度指数 γ 与 m 无关

- 社交网络中明星、政治领袖、大公司的CEO往往认识非常多人，这在社交网络中体现为枢纽节点；
- 此外，明星之间相识、结婚，公司间合作、政商界合作体现出社交网络的另一个特性：枢纽节点倾向于与其他枢纽节点相连；



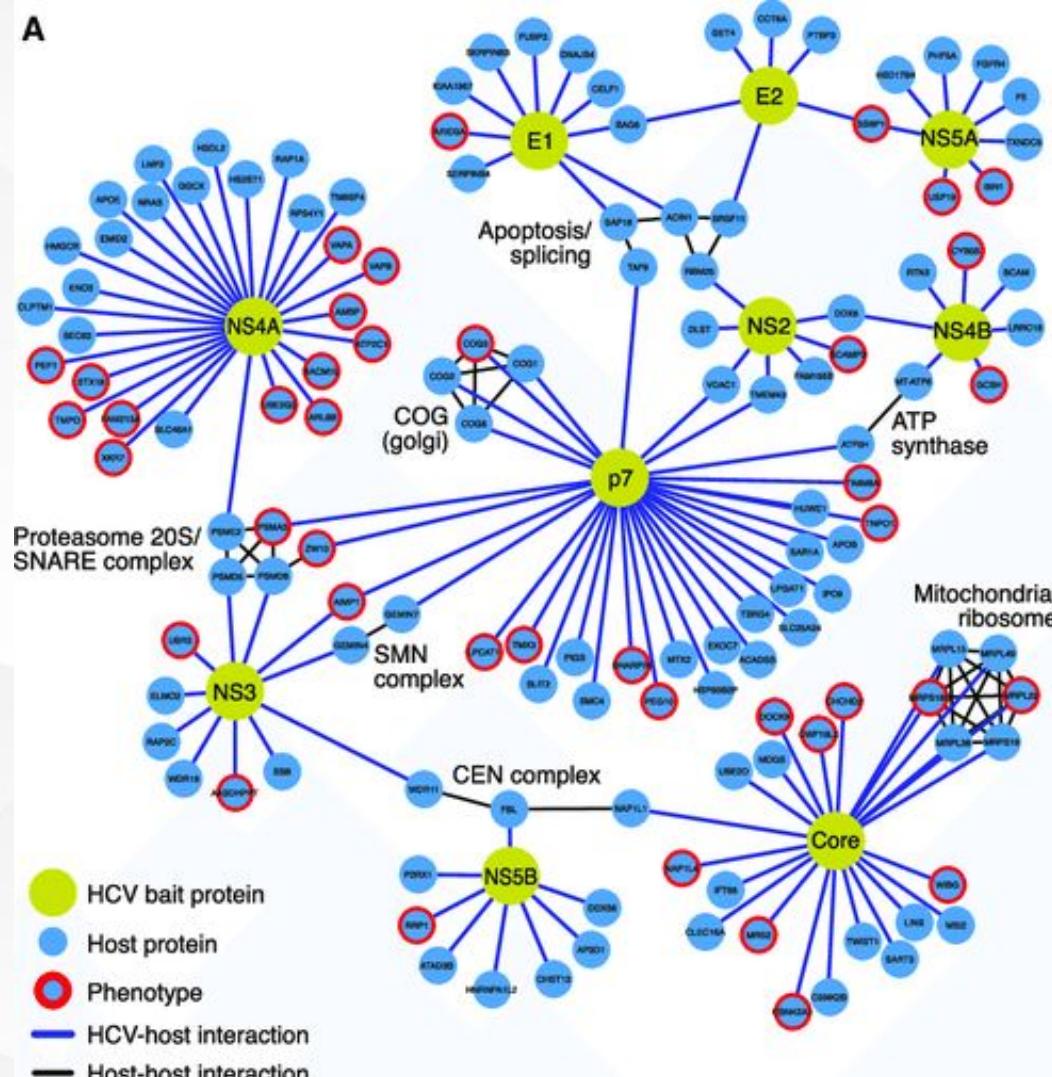
- 这一特性是所有网络都有的性质吗？
- 是所有**无标度**网络都有的性质吗？

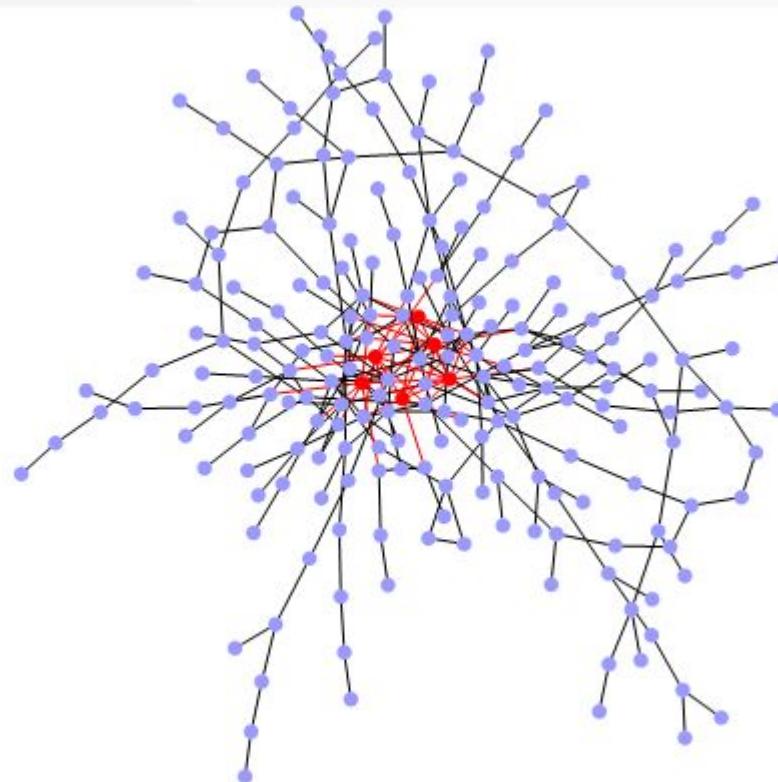




H. Jeong, S.P. Mason, A.-L. Barabasi, Z.N. Oltvai, Nature 411, 41-42 (2001)

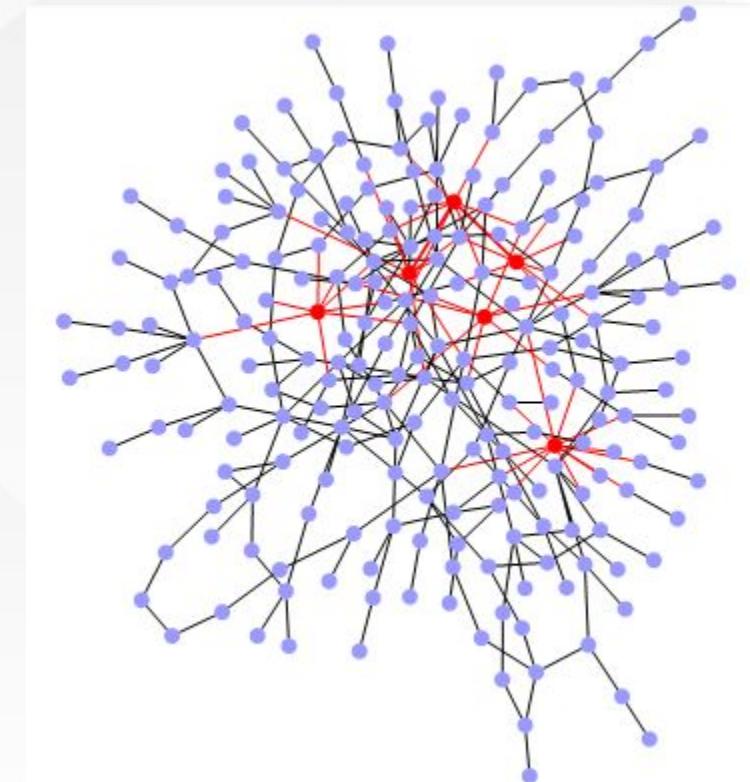
Ramage, Holly R., et al. "A combined proteomics/genomics approach links hepatitis C virus infection with nonsense-mediated mRNA decay." Molecular cell 57.2 (2015): 329-340.





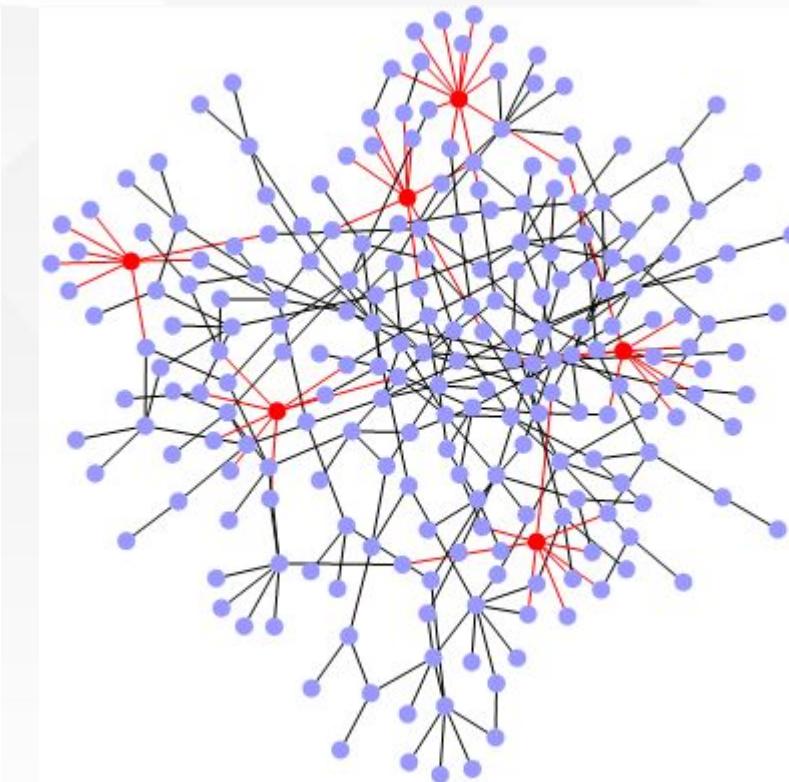
同配网络

枢纽节点倾向于彼此相连，小度
节点倾向于连接小度节点



中性网络

此处展示随机网络

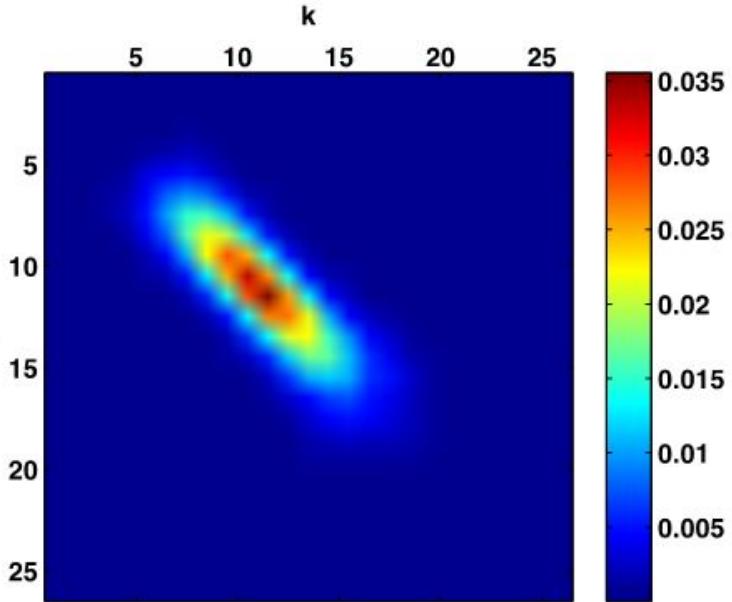


异配网络

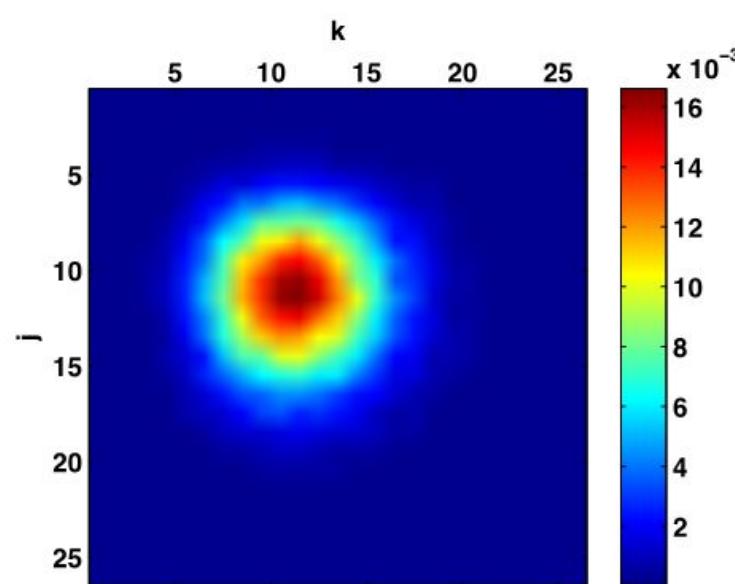
枢纽节点倾向于不彼此相连，
而是连接到小度节点，网络表
现出中心辐射特征



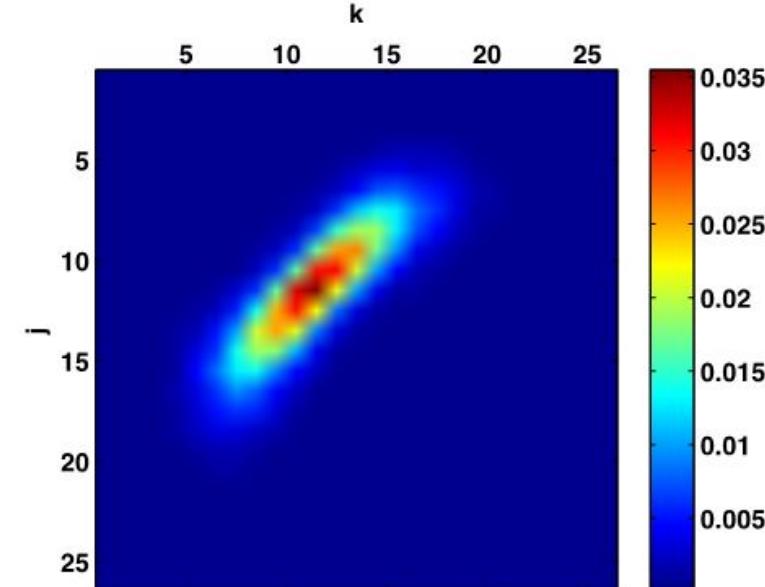
ER assortativ



ER neutral



ER disassortative



同配网络

边分布在对角线上, 也即度相似的节点之间

中性网络

此处展示随机网络

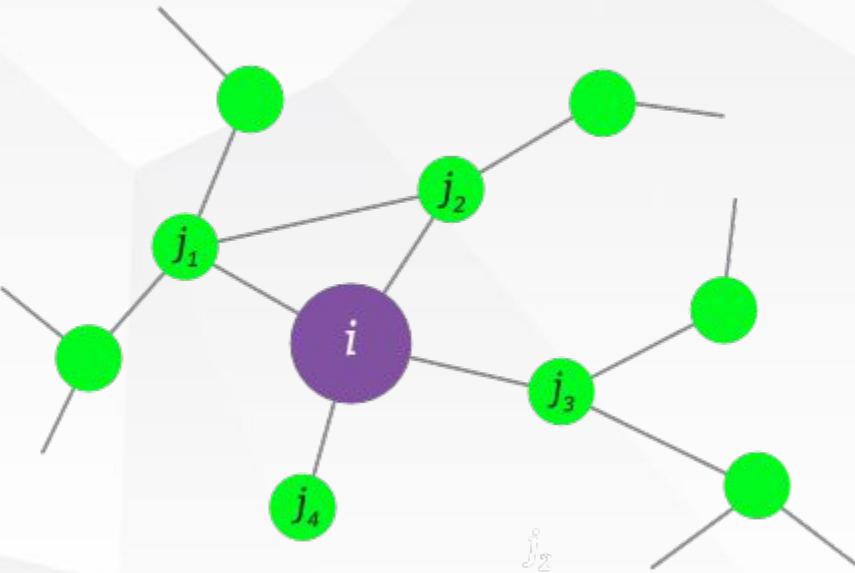
异配网络

边出现在度差异大的节点之间



计算节点*i*的邻居的平均度

$$k_{nn}(k_i) = \frac{1}{k_i} \sum_{j=1}^N A_{ij} k_j$$

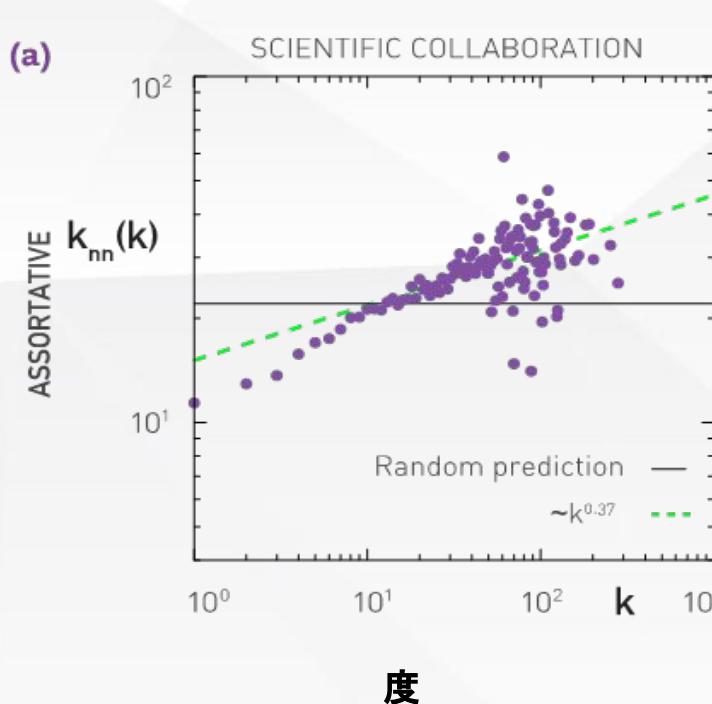


在所有节点上，计算所有度为k的节点上，度相关性函数

$$k_{nn}(k) \equiv \sum_{k'} k' P(k' | k)$$

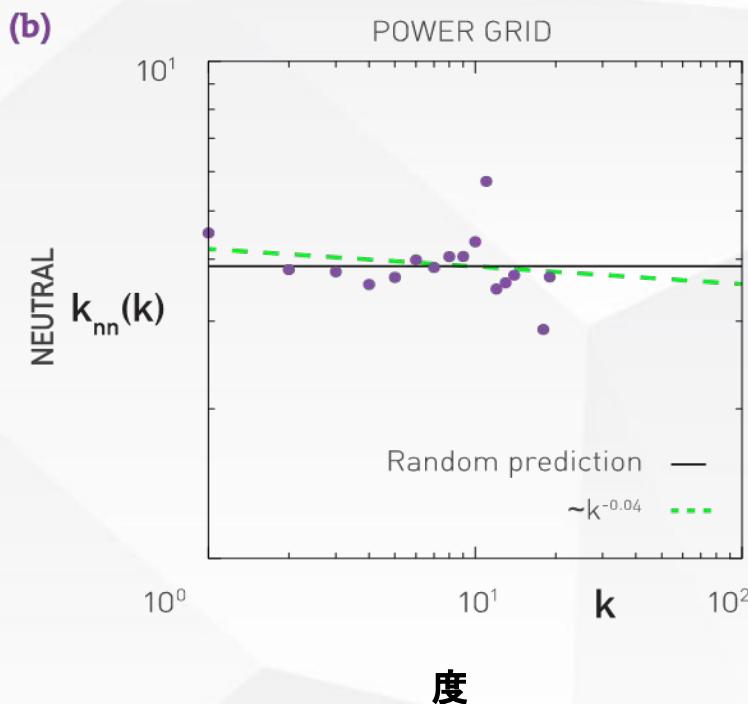
计算图中节点*i*的邻居平均度：





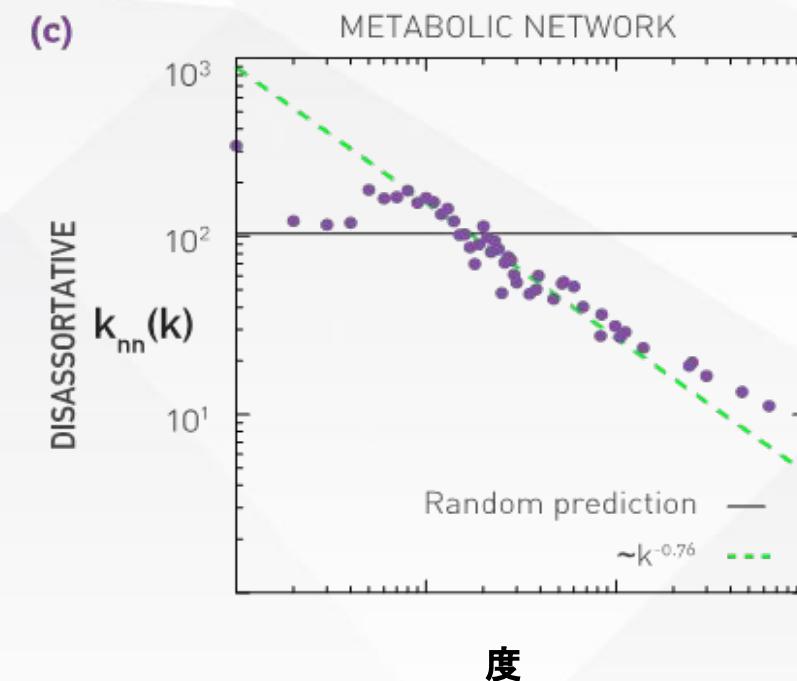
科学合作网络

$k_{nn}(k)$ 随着 k 递增，表明了网络的同配性质



电网

拟合曲线接近水平，表明没有度相关性，符合中性网络的性质



代谢网络

度相关性随 k 递减，表明网络的异配性质(度越大，邻居节点的平均度越小)

*三个子图都可以用幂律分布拟合，此时幂律分布的指数可以用于量化网络的同配性和异配性



提纲

1

复杂科学简介

2

图的基本概念

3

随机网络

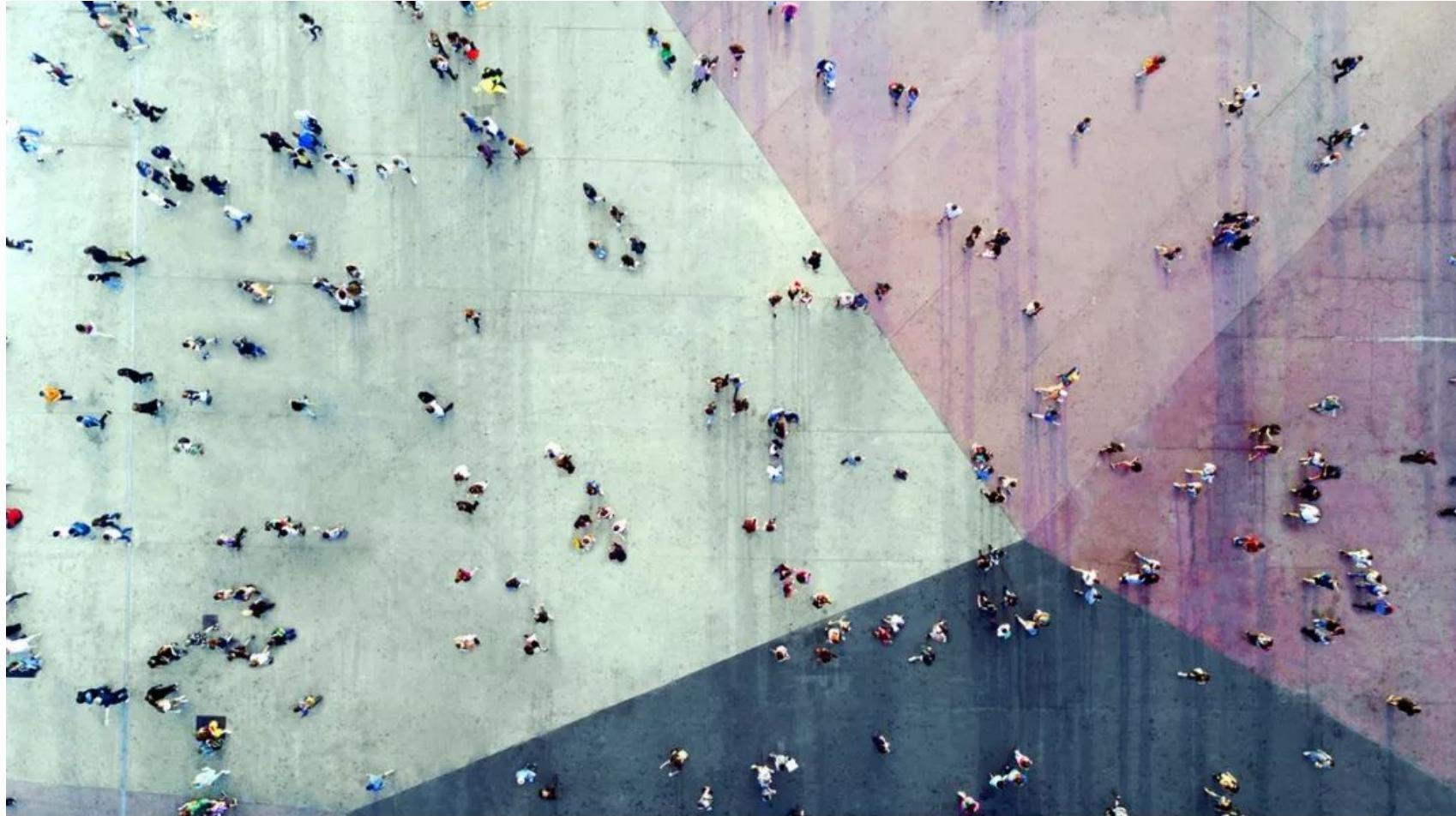
4

无标度网络

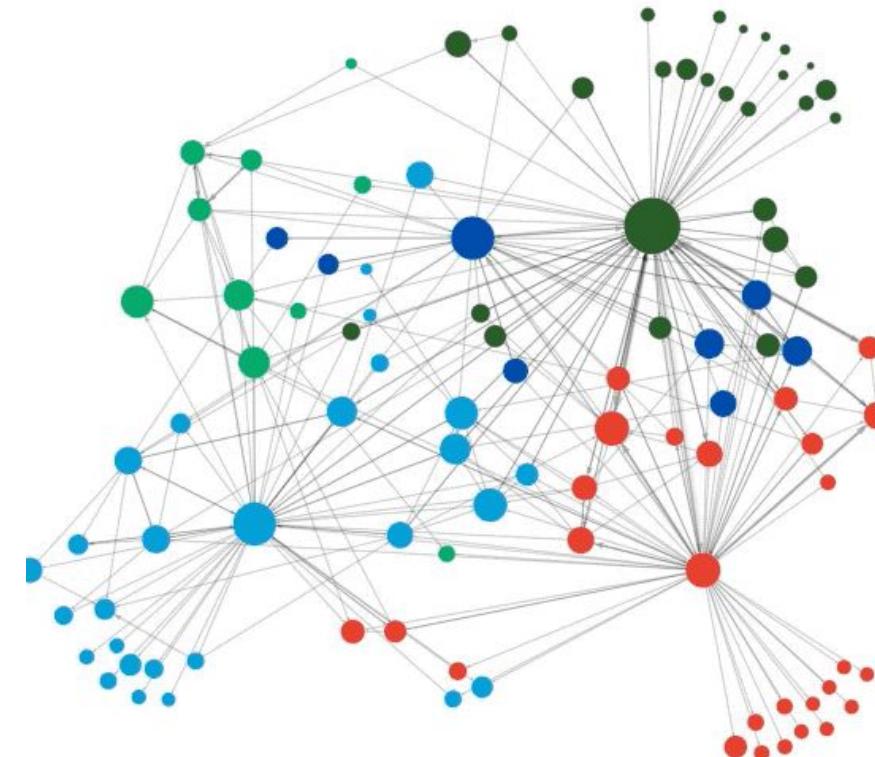
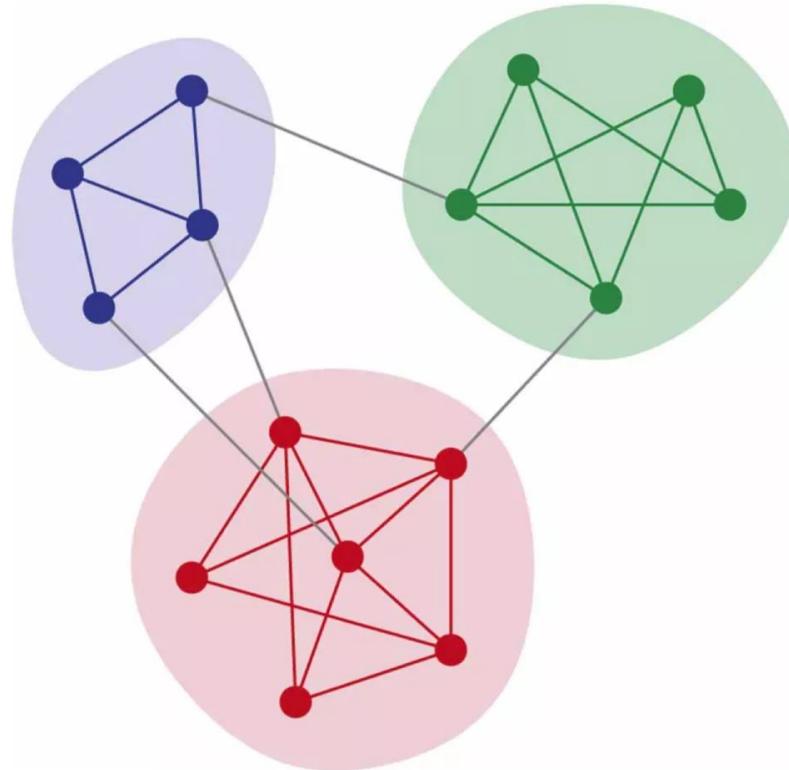
5

社团检测

社团检测(Community Detection):发现网络中存在紧密连接的节点集合或子图(subgraph)

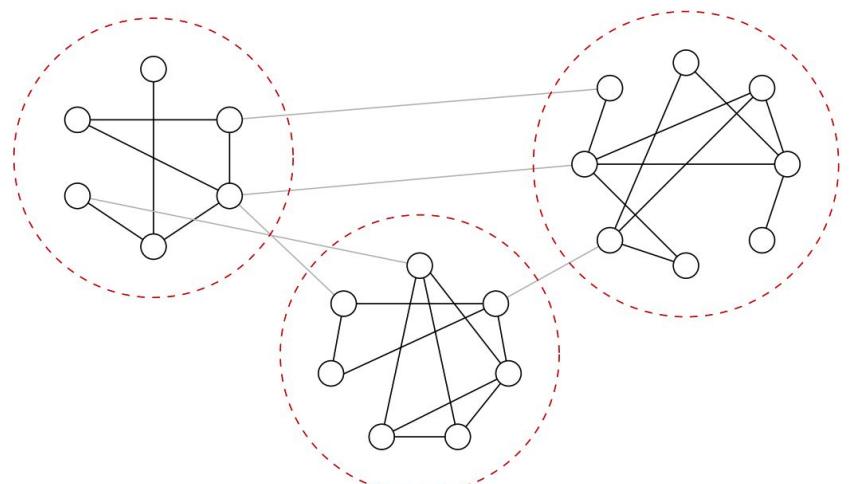


社团内部连接紧密，常对应某种特点或功能，而社团外部的连接则相对稀疏，即“内紧外松”



社团结构可视化

1. 选择初始化的 k 个样本作为初始聚类中心 $a = a_1, a_2, \dots, a_k$;
2. 针对数据集中每个样本 x_i 计算它到 k 个聚类中心的距离并将其分到距离最小的聚类中心所对应的类中;
3. 针对每个类别 a_j , 重新计算它的聚类中心 $a_j = \frac{1}{|c_i|} \sum_{x \in c_i} x$ (即属于该类的所有样本的质心) ;
4. 重复上面 2 3 两步操作, 直到达到某个中止条件 (迭代次数、最小误差变化等) 。





层次聚类 (Hierarchical clustering)

层次聚类是一种传统且有效的方法。它通过一种相似性度量（如欧氏距离），来计算节点间拓扑结构的相似性。然后每次找到距离最短的两个社团，然后进行合并成一个大的社团，直到全部合并为一个社团，形成一个树结构。根据需要可自行决定社团数目，得到最终结果。

算法步骤：

1. 将所有的节点看作一个单独的簇
2. 遍历similar矩阵，得到相似度最高的两个簇
3. 选取最高相似度，检测最高相似的是否高于某个阈值，如果是就结束聚类输出结果 (步骤5)
4. 将相似度最高的两个簇合并，并更新similar矩阵，将合并的两个簇中的最小的similar值更新为新的簇的值(MAX全链)，记录合并的簇
5. 计算模块度量值Q，如果Q值大于当前最大值则输出分类

网络的模块化程度
Modularity

$$Q = \sum_i (e_{ii} - a_i^2)$$

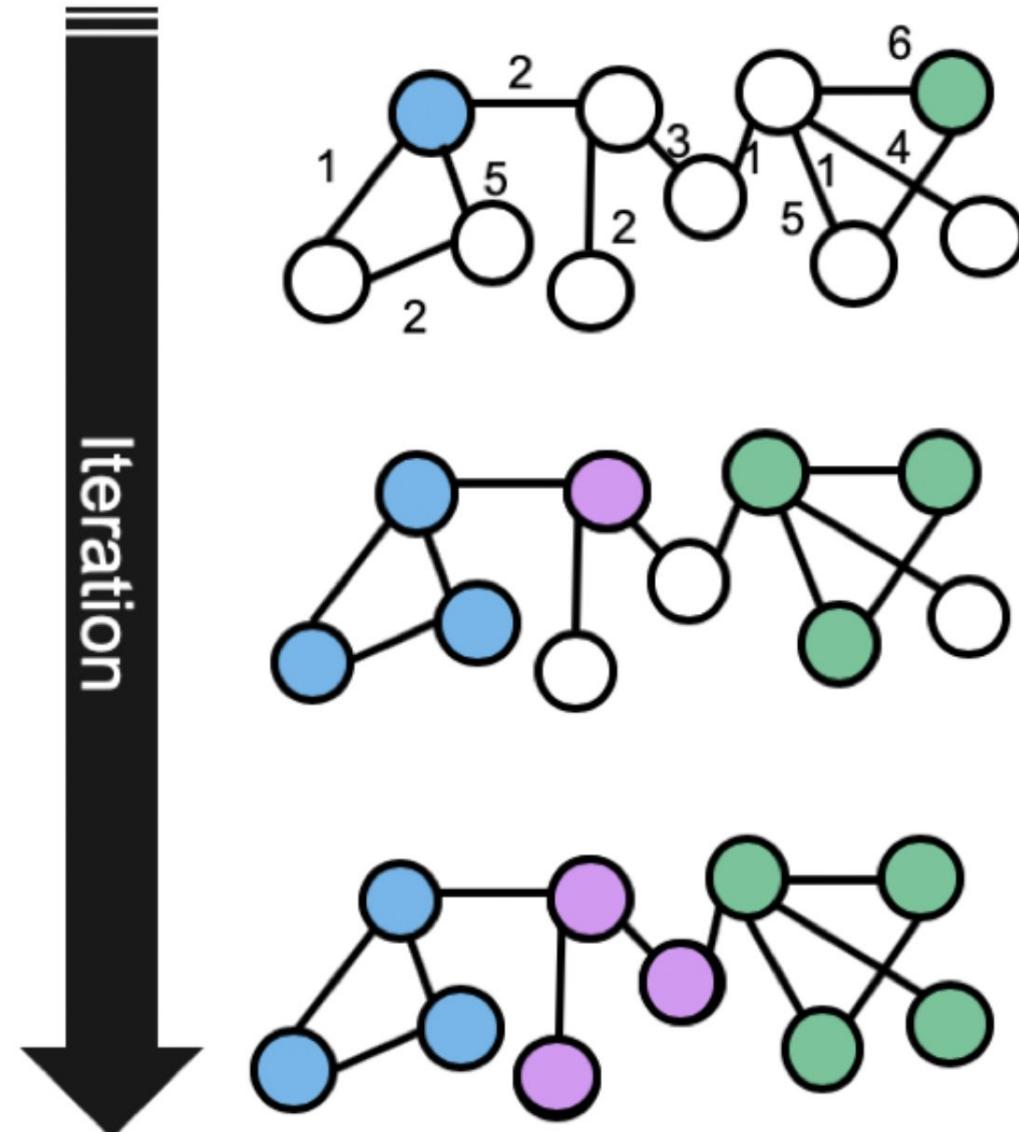
Ref: Newman, Mark EJ, and Michelle Girvan. "Finding and evaluating community structure in networks." *Physical review E* 69.2 (2004): 026113.

标签传播算法(LPA)

标签传播算法基本思想是通过标记节点的
标签信息预测未标记节点的标签情况。

节点之间的标签传播主要依照标签相似度
进行，在传播过程中，未标记的节点根据
邻接点的标签情况来迭代更新自身的标签
信息。

如果其邻接点与其相似度越相近，则表示
对其所标注的影响权值就越大，邻接点的
标签就更容易进行传播。

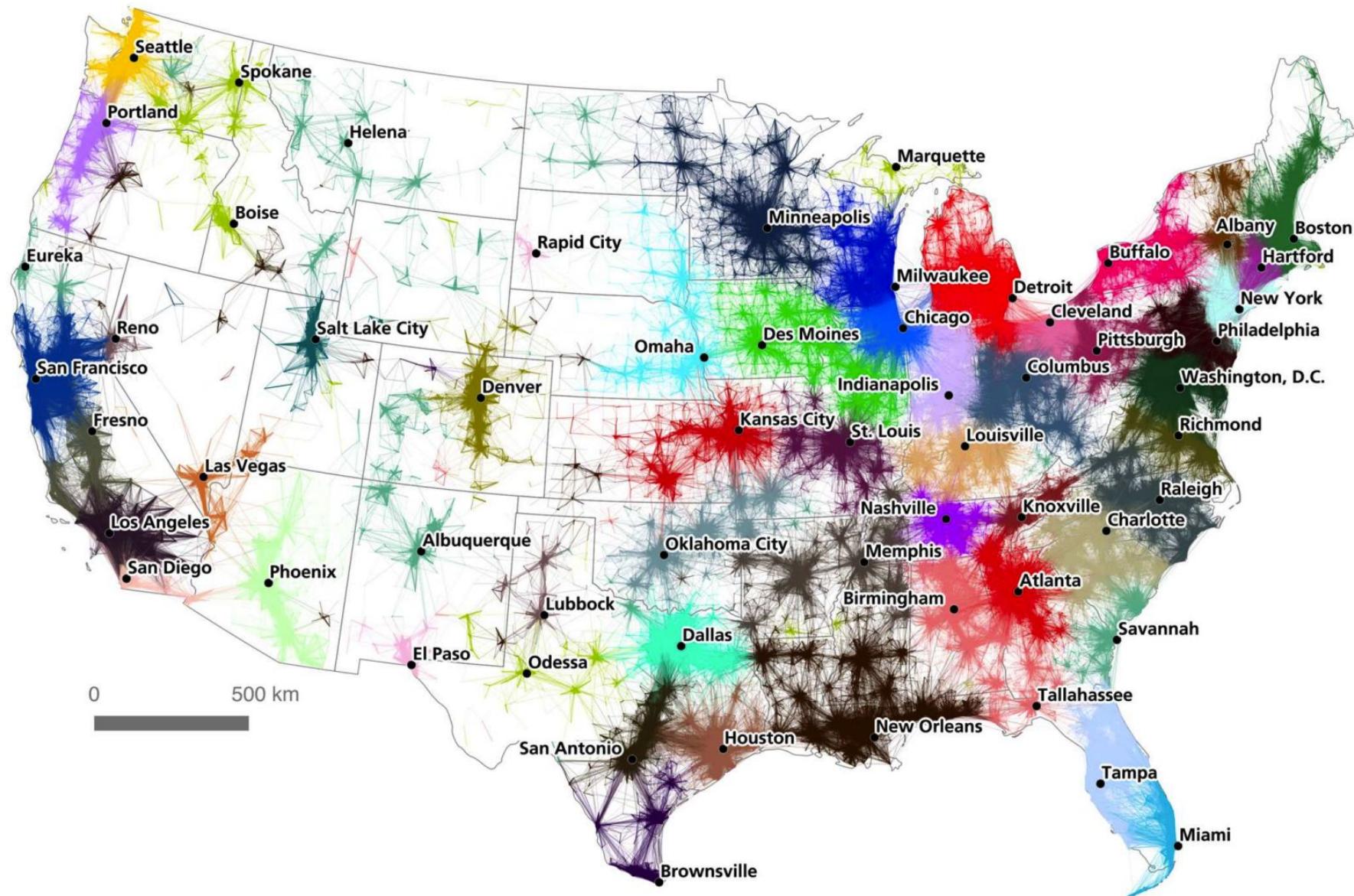




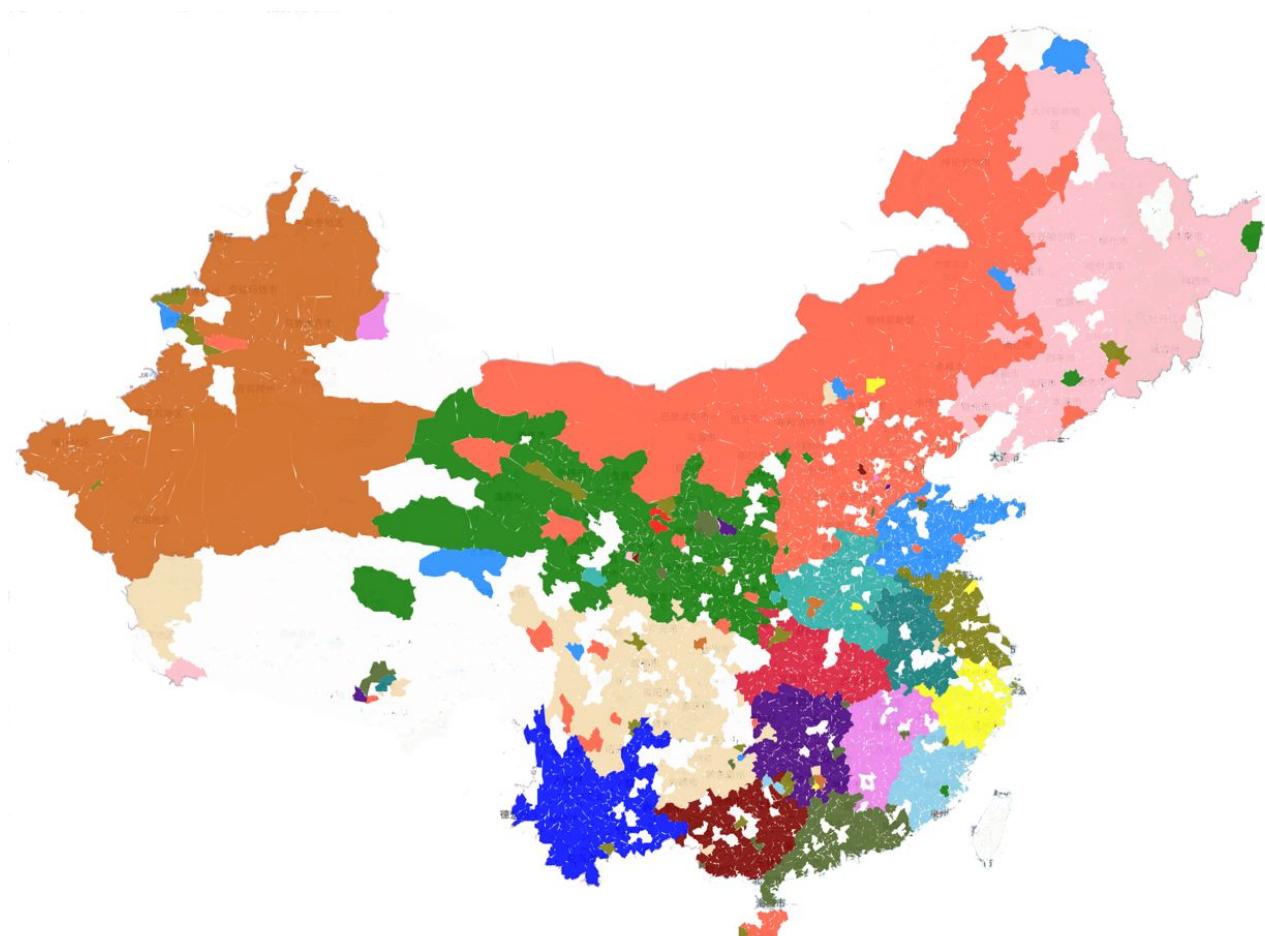
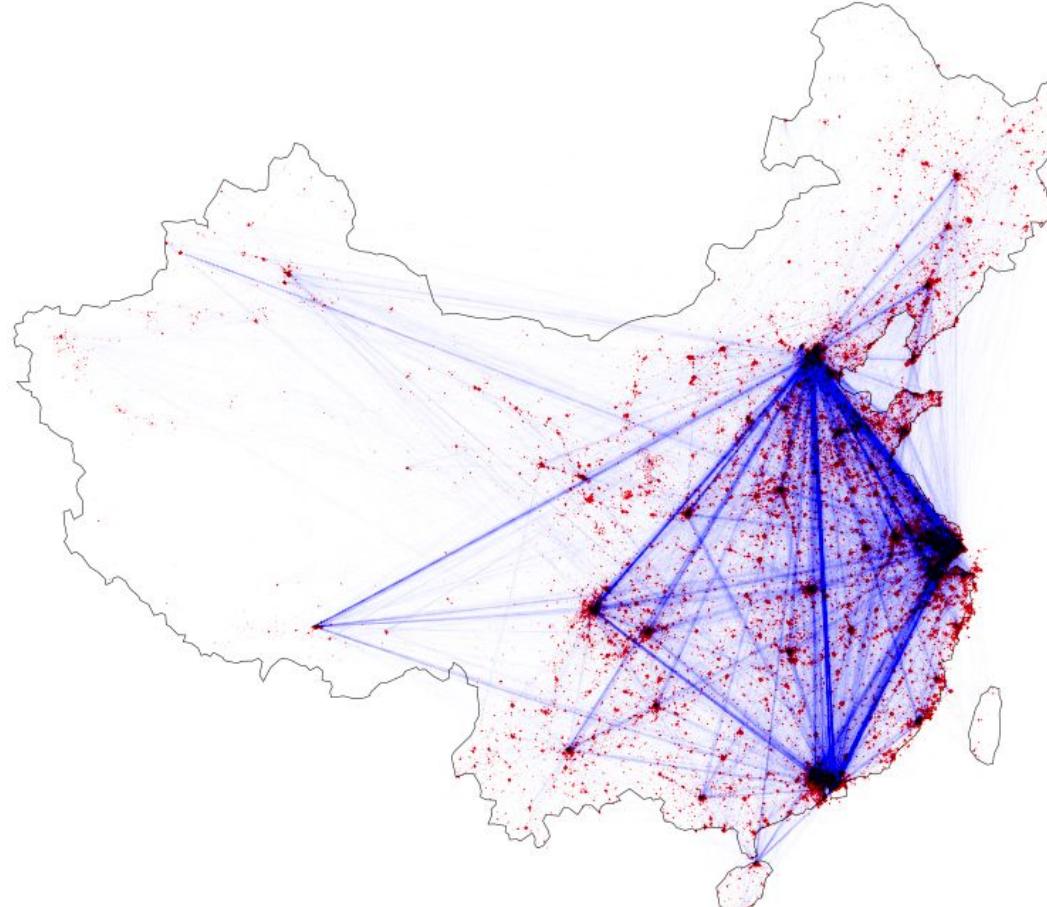
大城市区域发现



人工智能研究院
Artificial Intelligence Institute



区域(县)投资网络社区检测



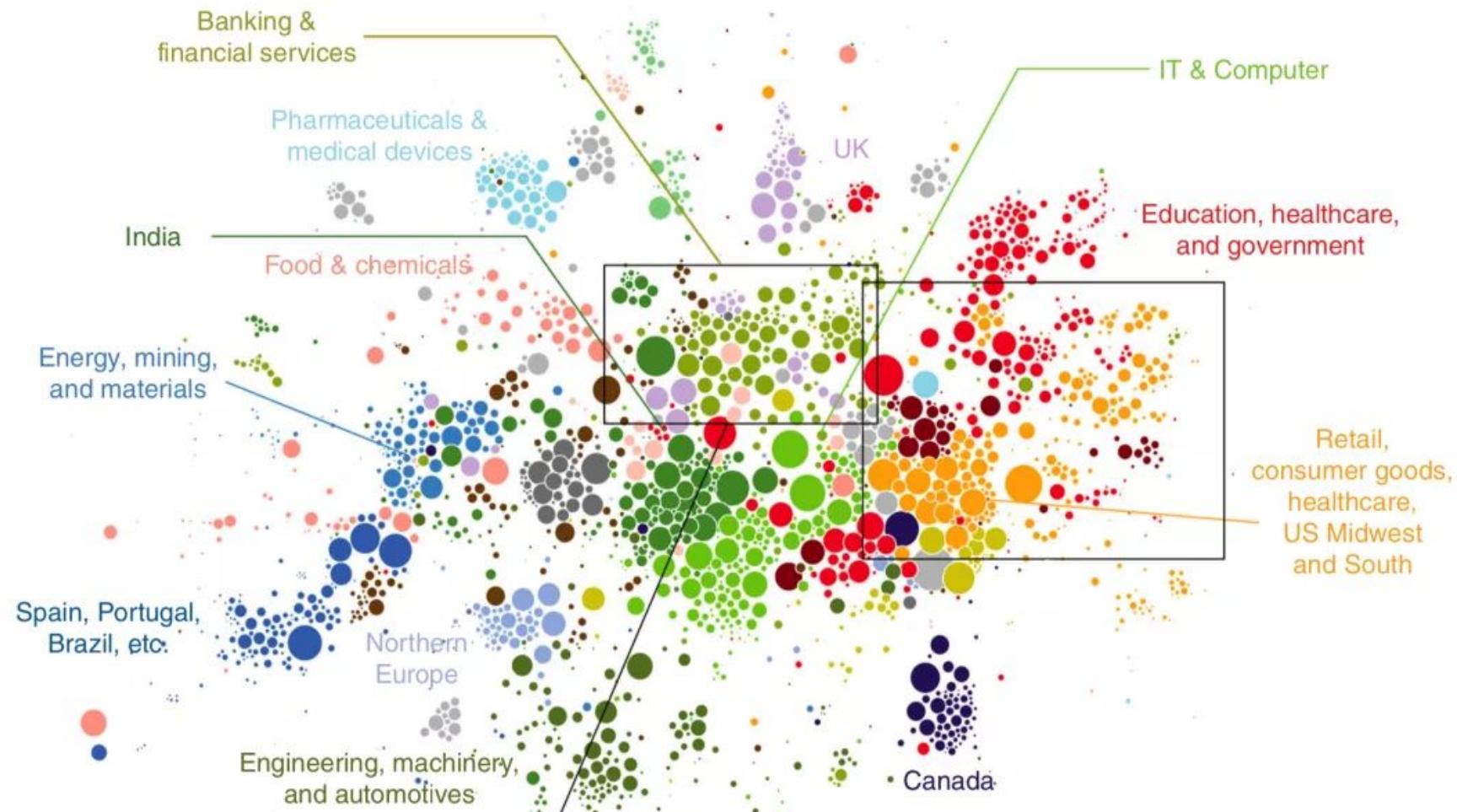
领英职业网络中的社团结构

论文题目：

Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters

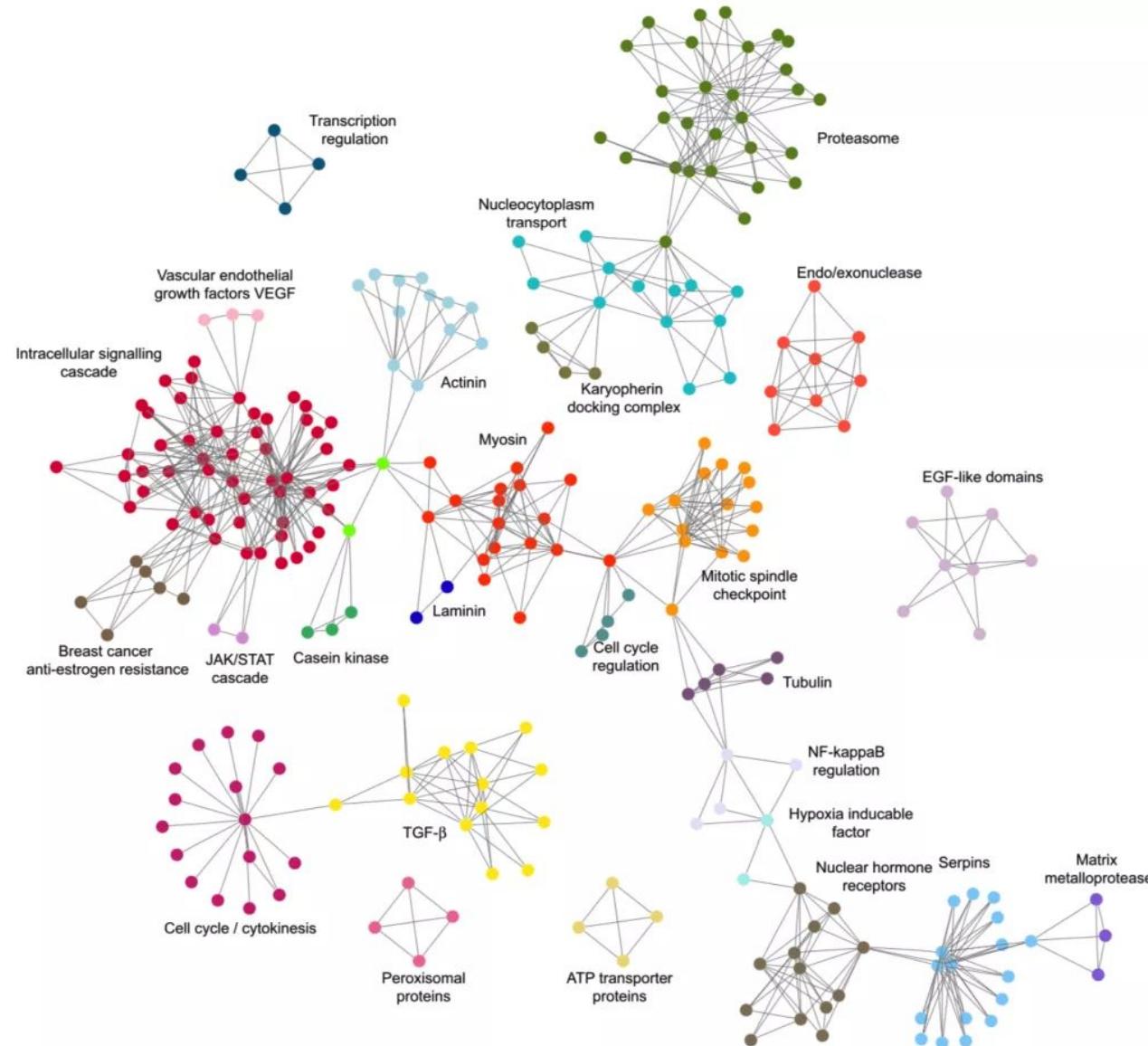
论文地址：

https://www.nature.com/articles/s41467-019-11380-w?fbclid=IwAR35Srs8_E5XN2bE0HgY8deFXC2ZCUW0BAkAowbn6A9gGCXtQryiYLblujc



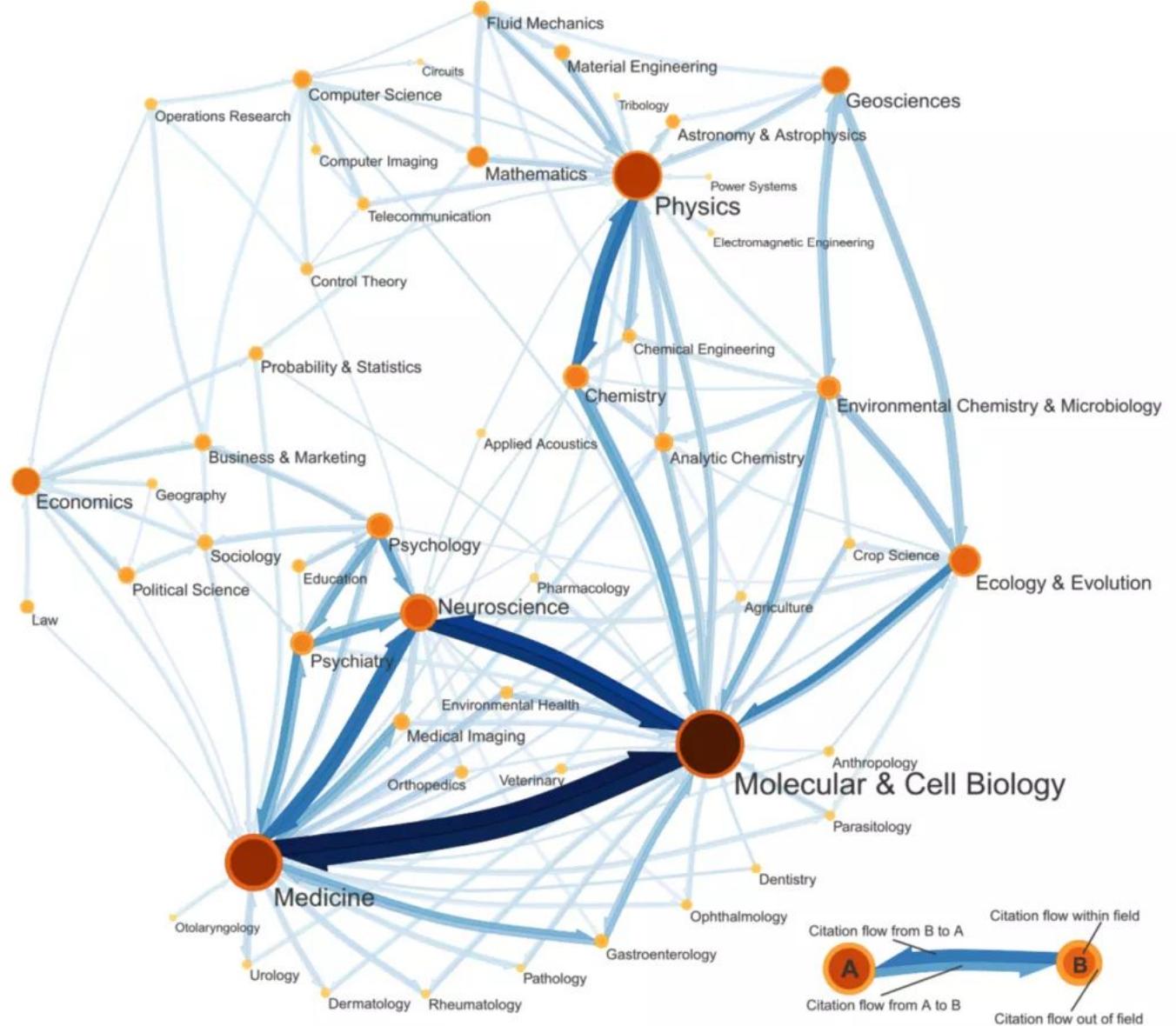
蛋白质相互作用网络 中的社团结构

论文题目 :Community detection in graphs
论文地址 :<https://arxiv.org/abs/0906.0612>





基于论文引用的图谱



论文题目 :Community detection in graphs
论文地址 :<https://arxiv.org/abs/0906.0612>



nature

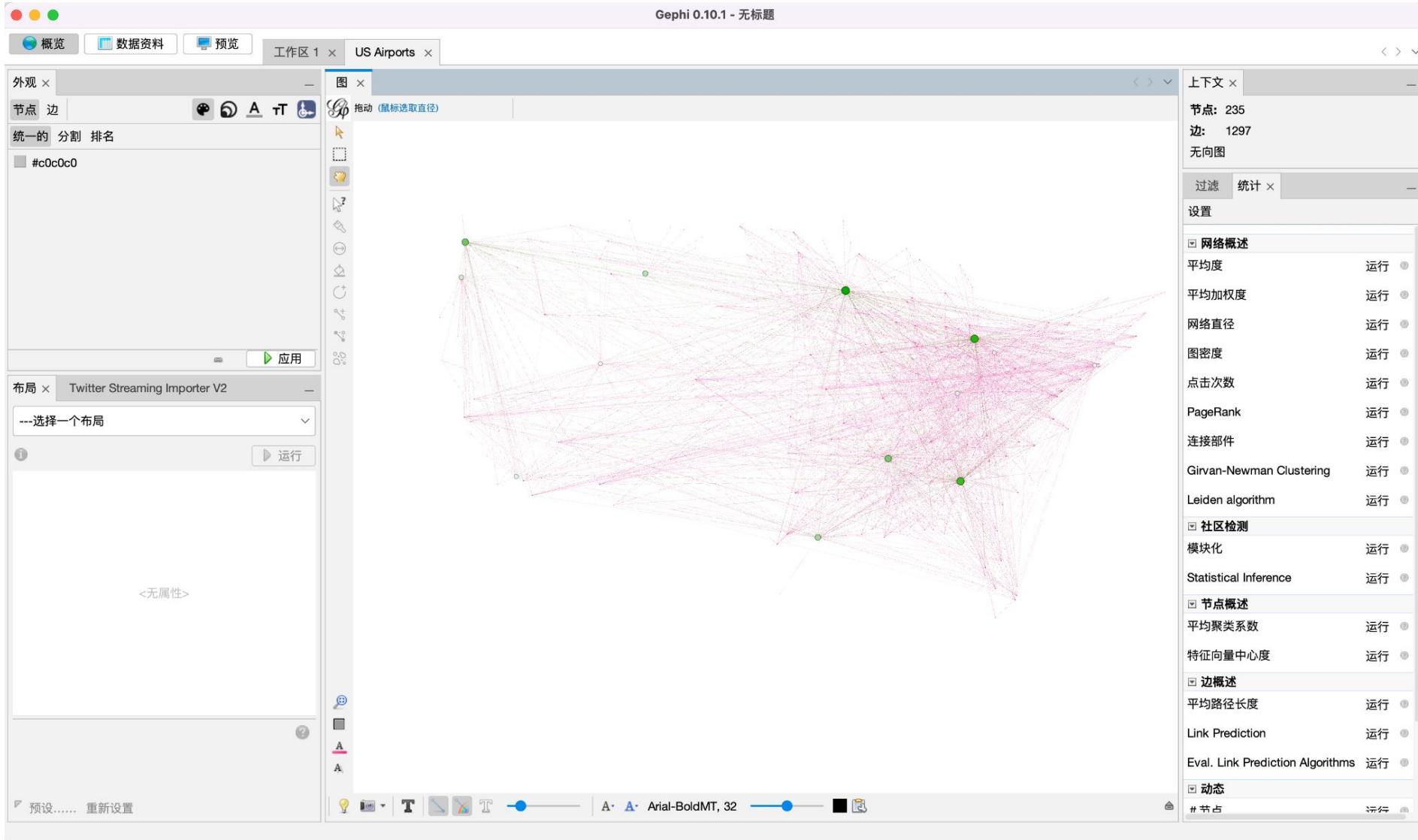
150 years of Nature



网络可视化工具 Gephi



人工智能研究院
Artificial Intelligence Institute



在线：<https://gephi.org/gephi-lite/>



上海交通大学

SHANGHAI JIAO TONG UNIVERSITY