WDS

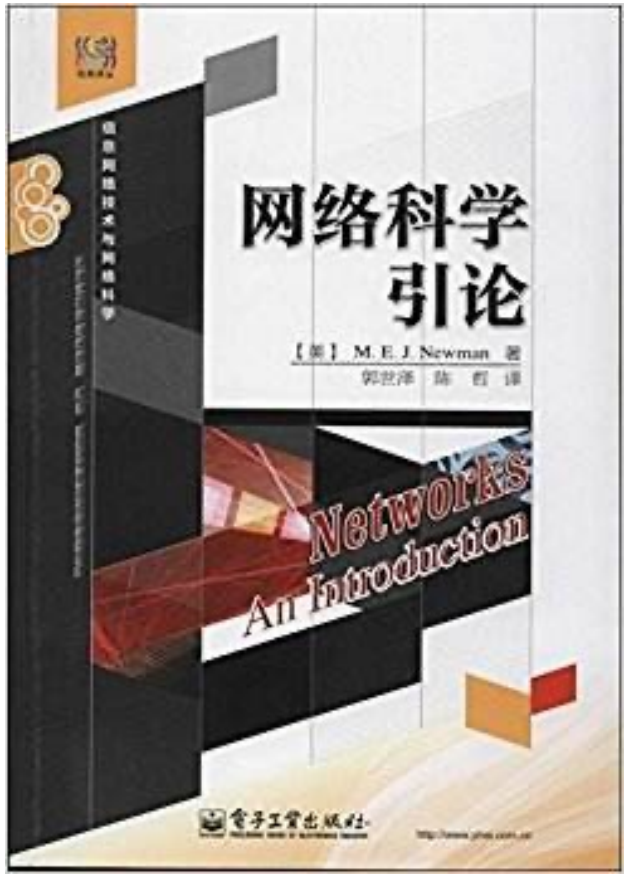
DataScience@Web

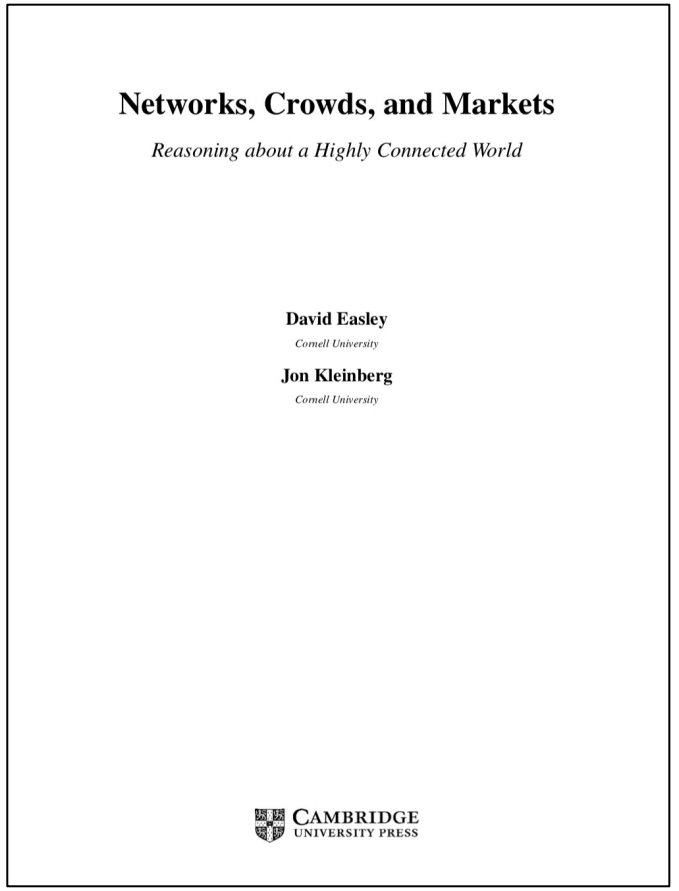
GRAPH | NETWORK ANALYSIS (PART I)

Peng Wang

School of Computer Science and Engineering, Southeast University

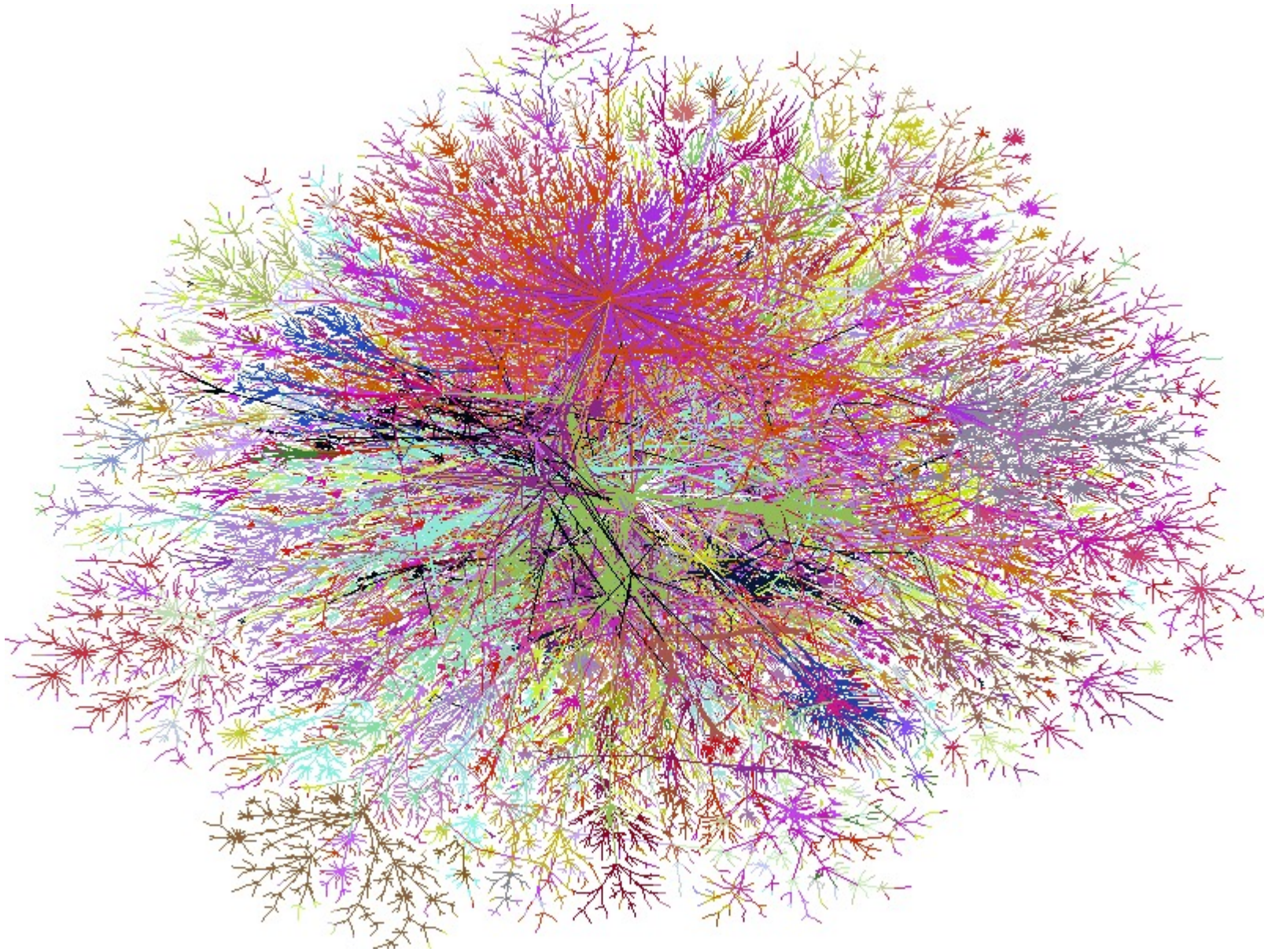
Textbook





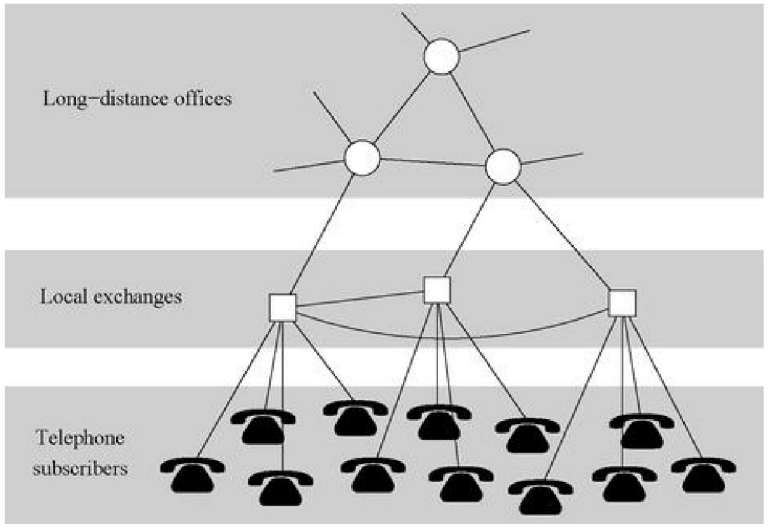
https://www.cs.cornell.edu/home/kleinber/networks-book/

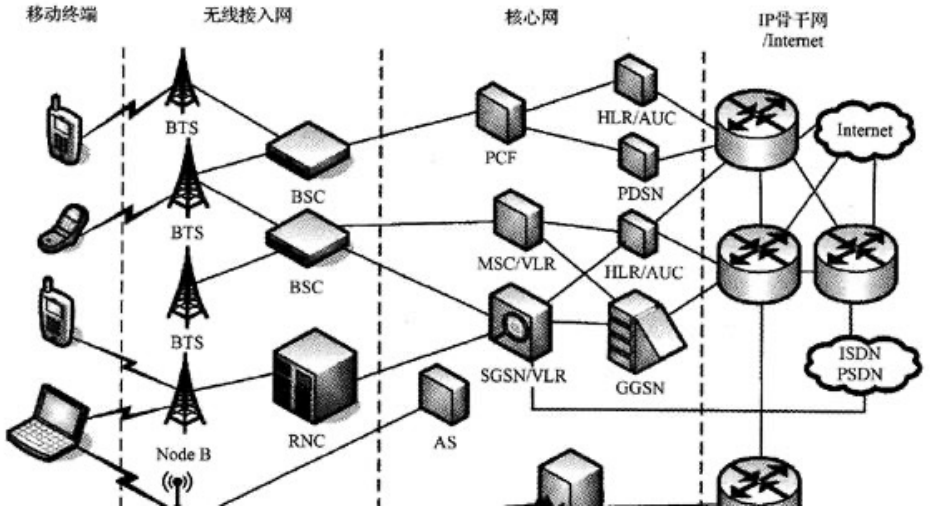
TYPES OF NETWORKS



Technological Networks:
The Internet

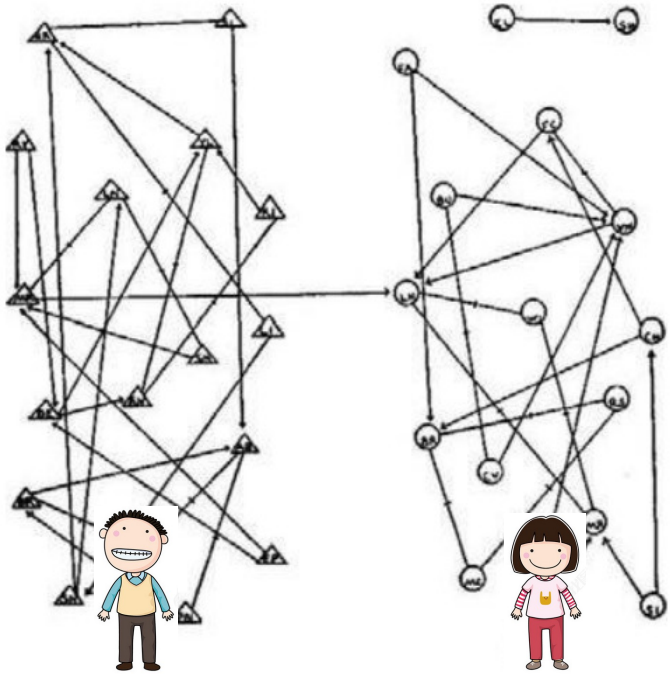
Technological Networks: The Telephone and Mobile Network



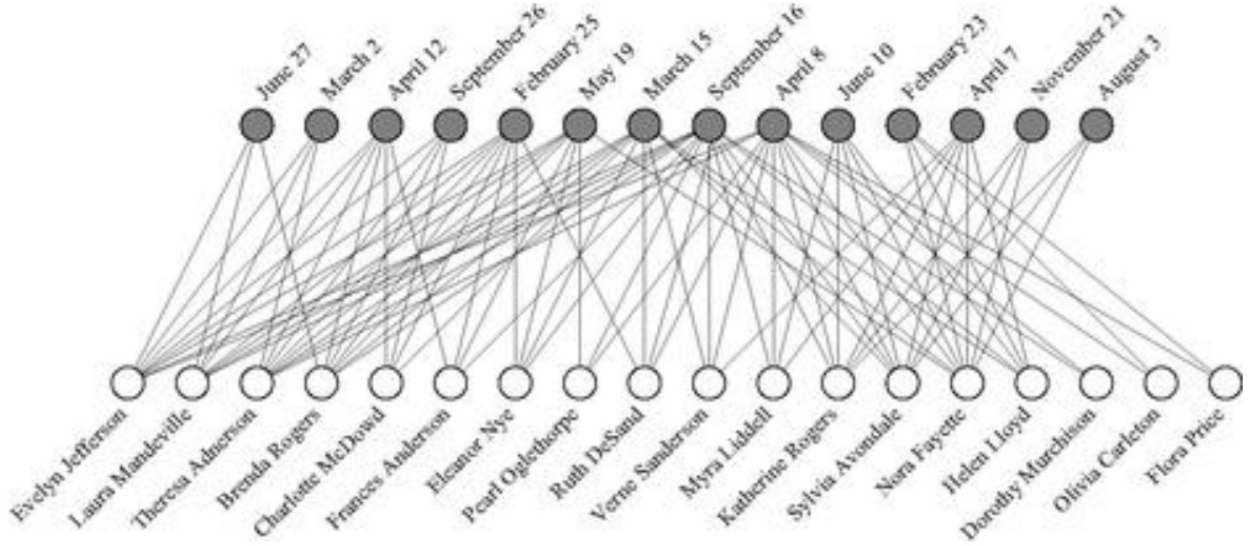


Social Networks:

Moreno: friendships between schoolchildren (1930)

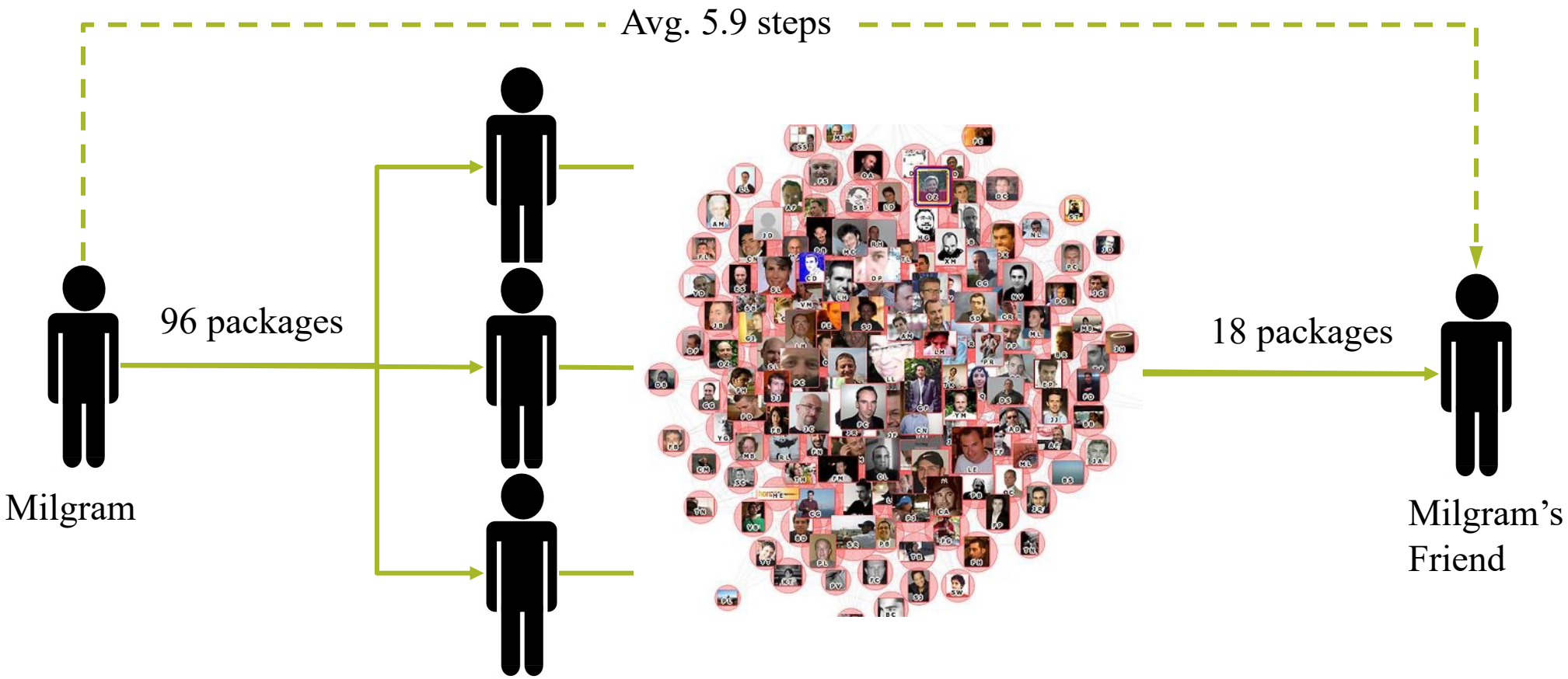


The affiliation network (隶属网络) of the “Southern Women Study” (1940)



bipartite graph

A Small World Experiment



A Small World Experiment

An Experimental Study of the Small World Problem*

JEFFREY TRAVERS

Harvard University

AND

STANLEY MILGRAM

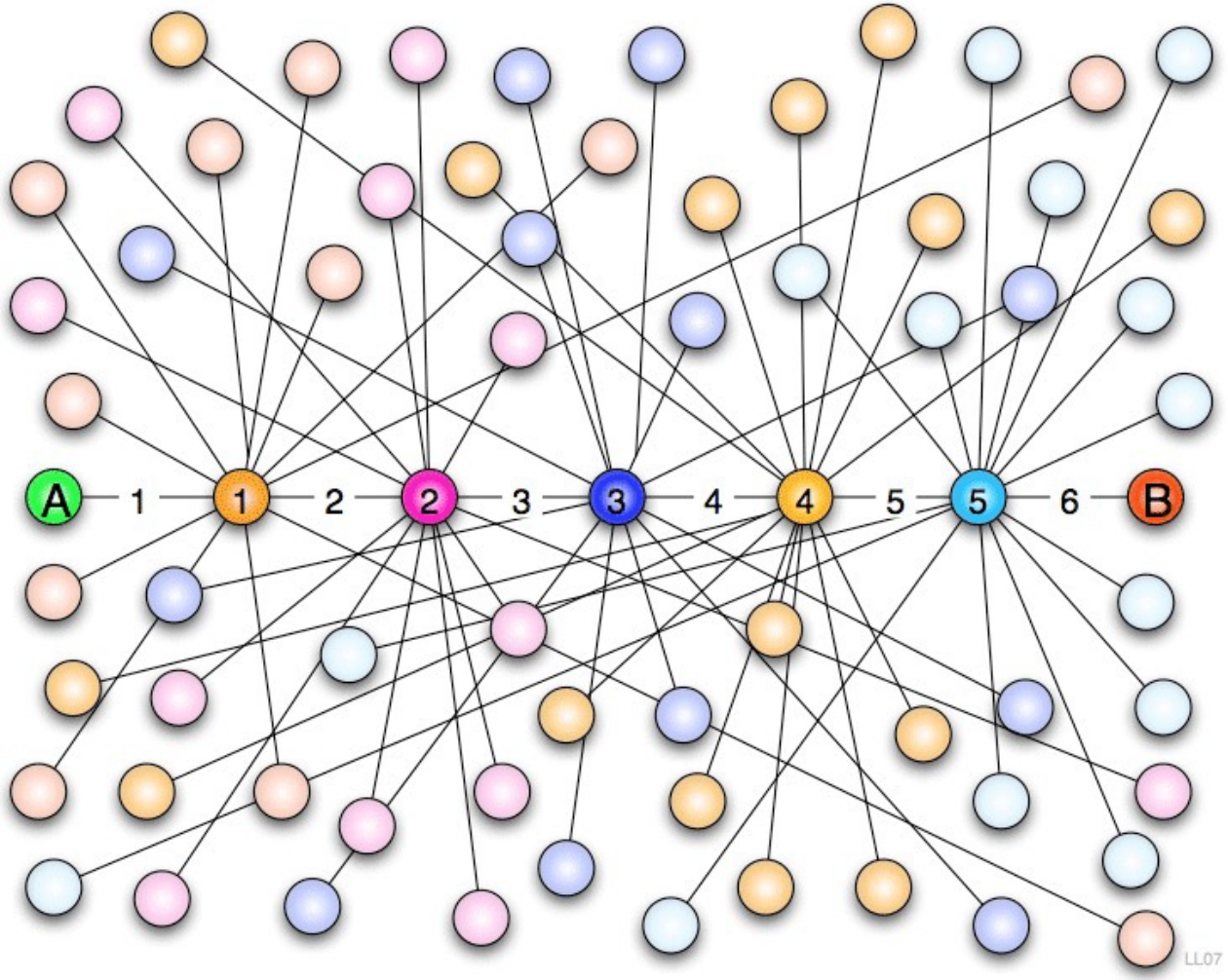
The City University of New York

Arbitrarily selected individuals (N=296) in Nebraska and Boston are asked to generate acquaintance chains to a target person in Massachusetts, employing "the small world method" (Milgram, 1967). Sixty-four chains reach the target person. Within this group the mean number of intermediaries between starters and targets is 5.2. Boston starting chains reach the target person with fewer intermediaries than those starting in Nebraska; subpopulations in the Nebraska group do not differ among themselves. The funneling of chains through sociometric "stars" is noted, with 48 per cent of the chains passing through three persons before reaching the target. Applications of the method to studies of large scale social structure are discussed.

https://snap.stanford.edu/class/cs224w-readings/travers69smallworld.pdf

A Small World Experiment

Six Degrees of Separation
六度分离理论



A Small World Experiment

- Microsoft MSN network

In 2006
Average 6.6 steps in 240,000,000 users

- Facebook network

In 2016
Average 3.74 steps in 721,000,000 users

Dunbar's number

- 150 Law

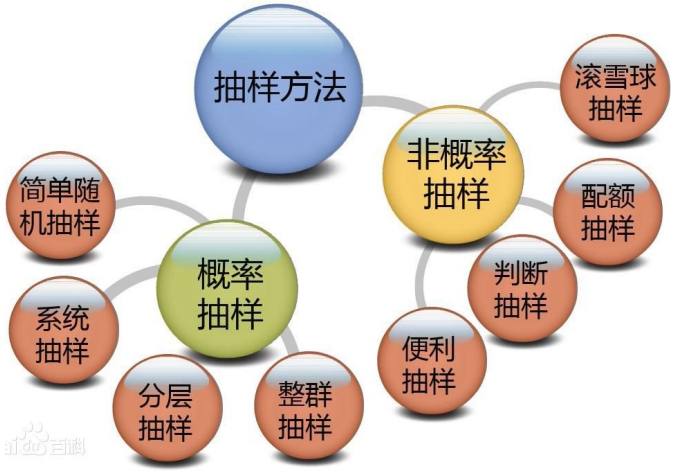
$$150^6 = 11{,}390{,}625{,}000{,}000$$

Bigger than population size on the earth

https://en.wikipedia.org/wiki/Dunbar%27s_number

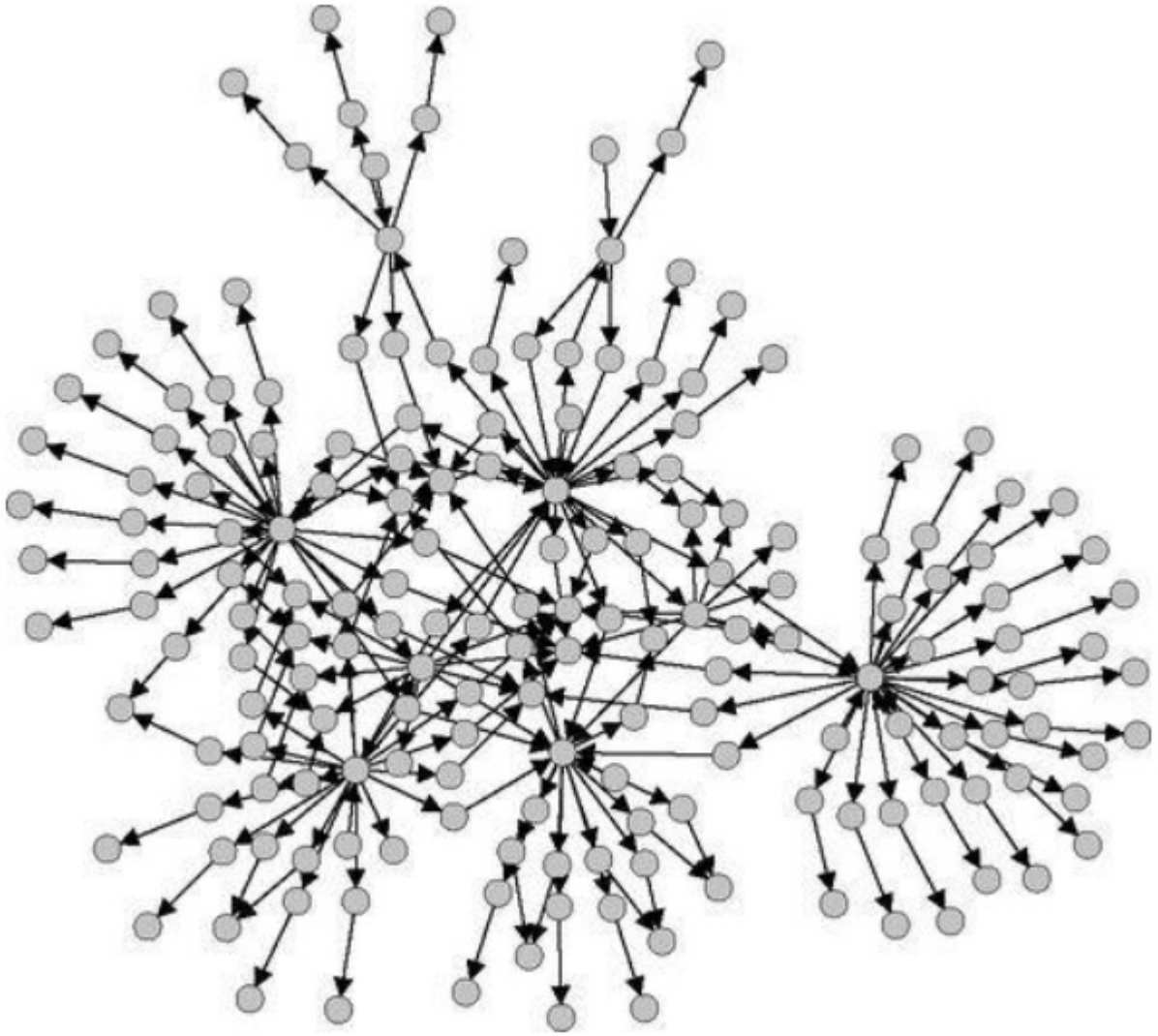
Sampling Hidden Populations

- Scenario: telephone survey - randomly select k men, and ask them if they use drugs.
- Problem:
  - random sampling may fail to find sparsely-distributed positive samples;
  - positive samples have no trust on random visitors;
- Snowball Sampling - non-probability sampling:
  - step1: find a real drug user;
  - step2: let the drug user help to find his neighboring drug users;
  - step3: iteration and propagation;
- Contact Tracing – closely related to Snowball Sampling:
  - scenarios: disease incidence (such as tracing HIV contacts)
  - step1: find a HIV sample
  - step2: find all of his contacts
  - step3: iteration and propagation



Sampling Hidden Populations

- shortage of Snowball Sampling and Contact Tracing
  - focus on local structures
  - sampling with strong bias
- random-walk sampling
  - step1: find a positive sample and its neighbors;
  - step2: randomly select *one and only one* neighbor;
  - step3: follow the selected neighbor and iteration;
  - shortage: slow
  - advantage: less bias. *Why?*
- *What if it is unethical to get a positive sample to name its neighbors?*
  - It is not allowed to publicly list HIV patients by their friends
  - respondent-driven sampling



Information Network:

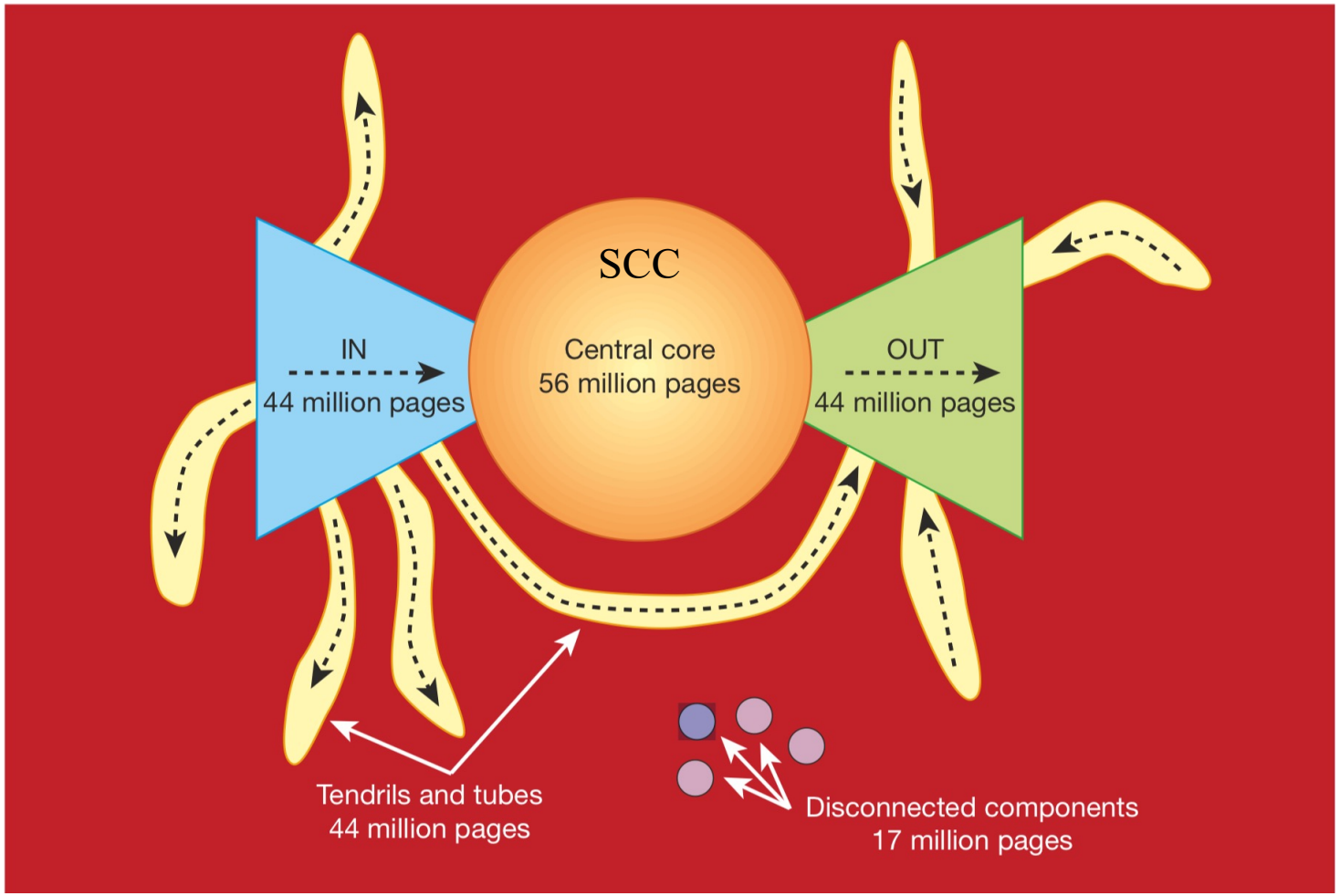
WWW: a network of pages of a corporate website

Web Crawler

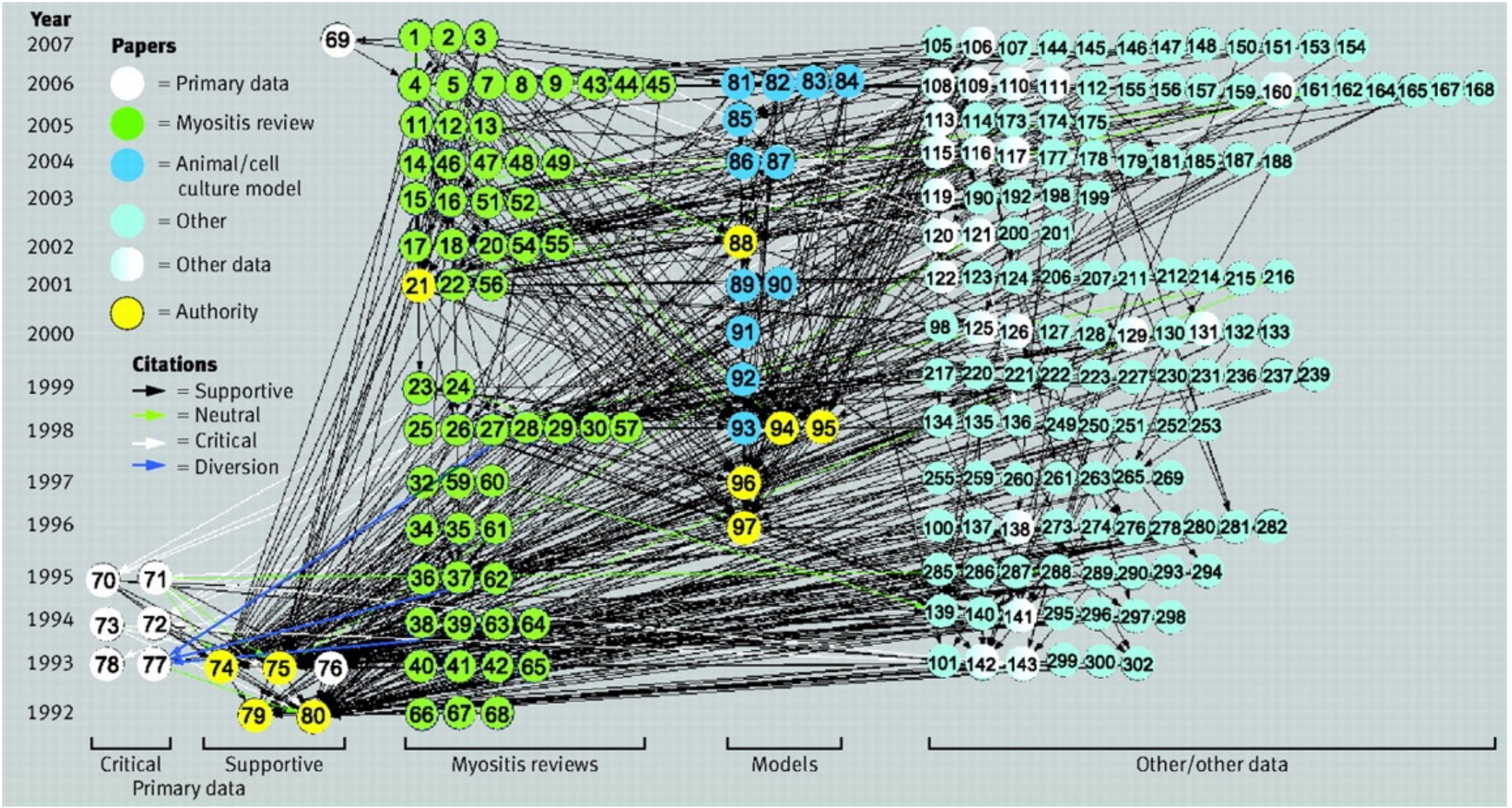
- A robot fetching pages on WWW by following links
- various crawling strategies:
  - breadth first search
  - depth first search
  - focused search
- robots.txt
- Problem: *how to get a proper set of seeds?*

领结

Bow-tie structure of the WWW



https://www.nature.com/articles/35012155



Information Network: Citation Network

What is the major difference between Citation Network and WWW network?

Some fun data in SCI papers

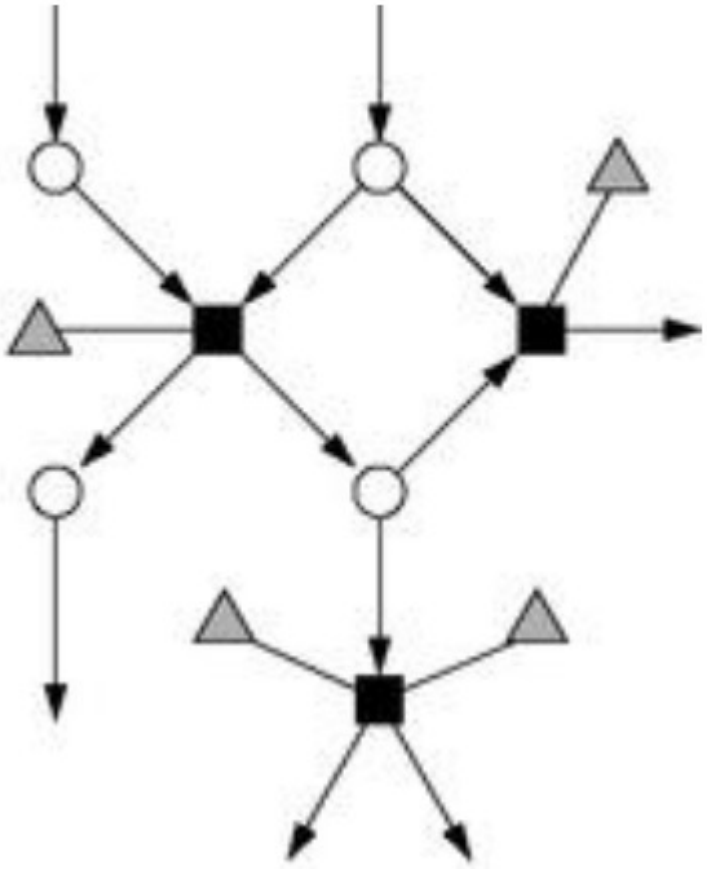
- 47% of SCIs never get cited;
- 9% of remaining only cited once;
- 6% of remaining only cited twice;
- only 1% of remaining have >100 cites

Biological Networks

Metabolite (代谢物)

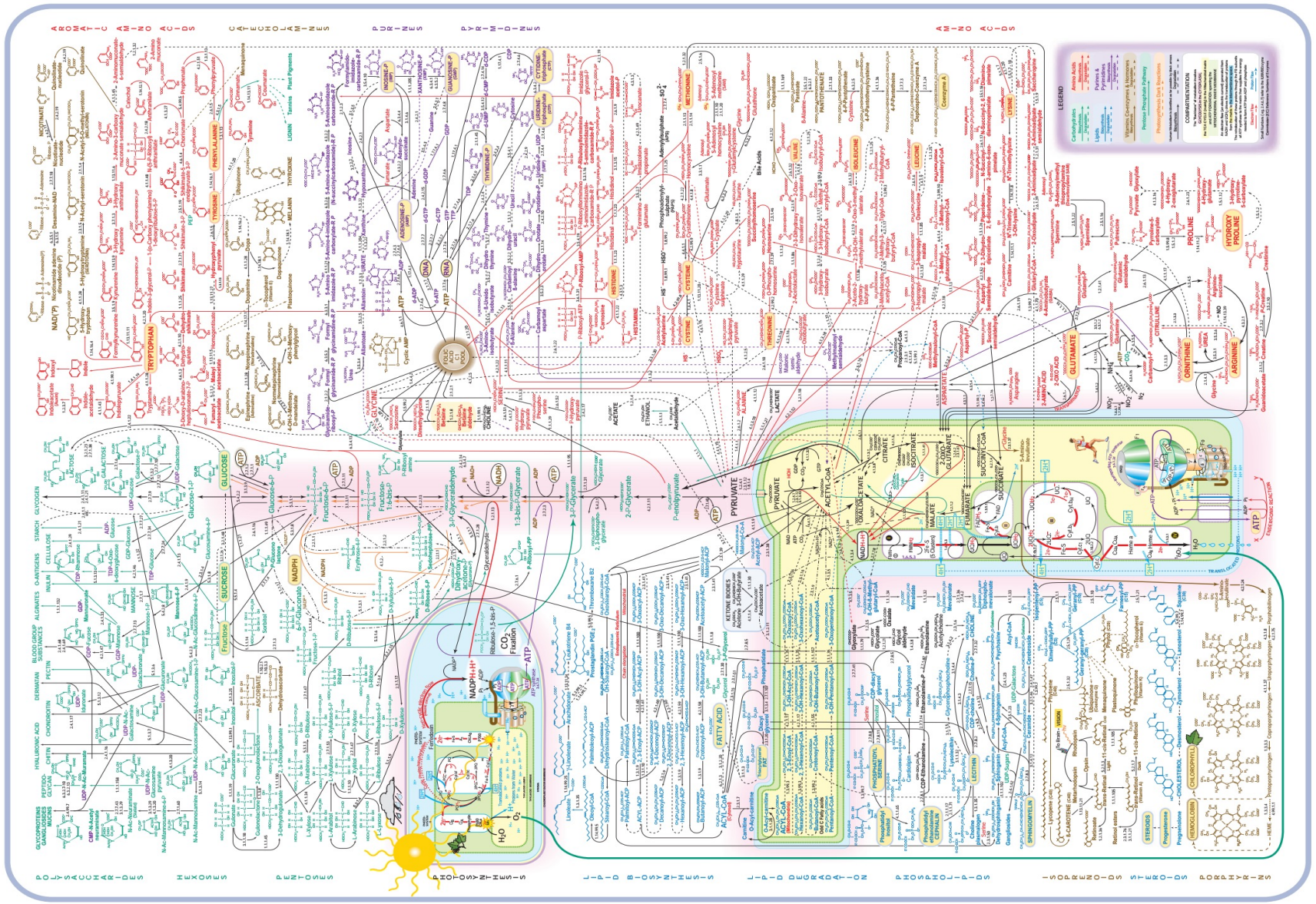
Reactions (化学反应)

Enzyme (酶)

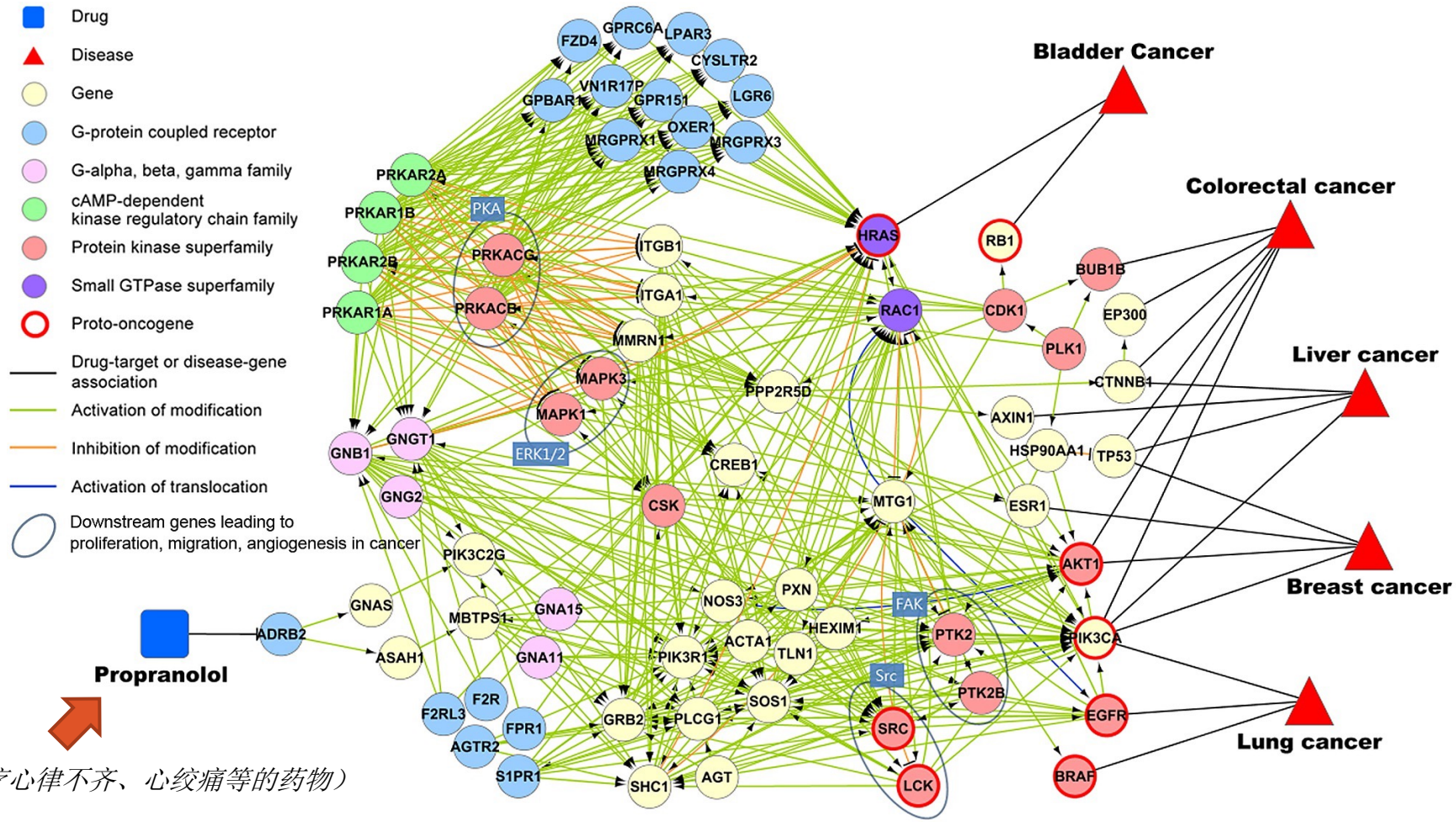


A tripartite representation of a portion of metabolic network

Metabolic Pathways

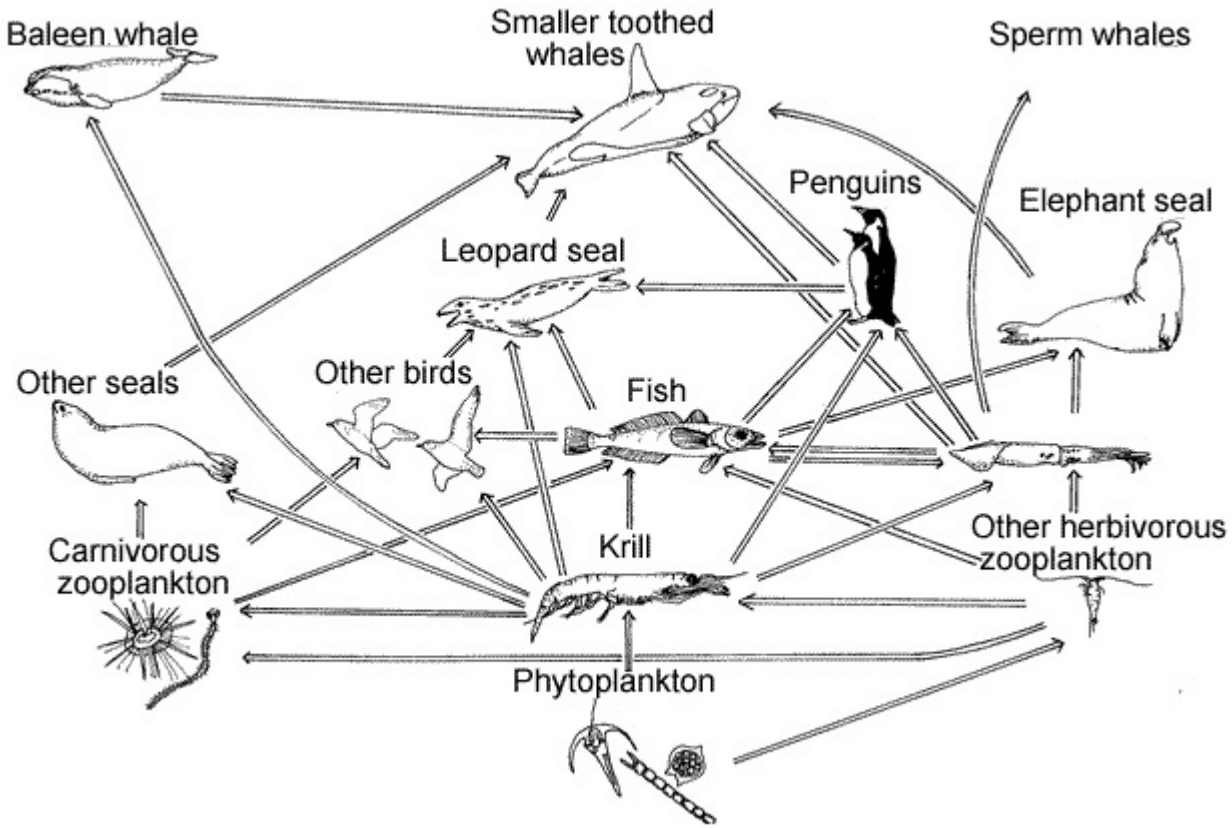


Network of Drugs, Diseases and Genes



心得安（一种用于治疗心律不齐、心绞痛等的药物）

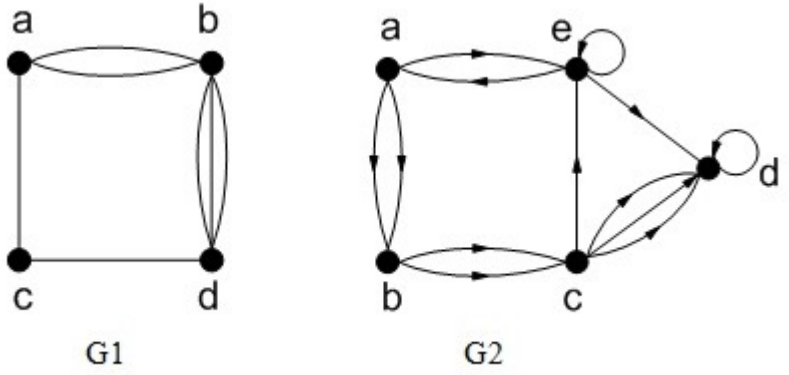
Antarctica Food Network



FUNDAMENTAL OF NETWORKS

Representation of Networks

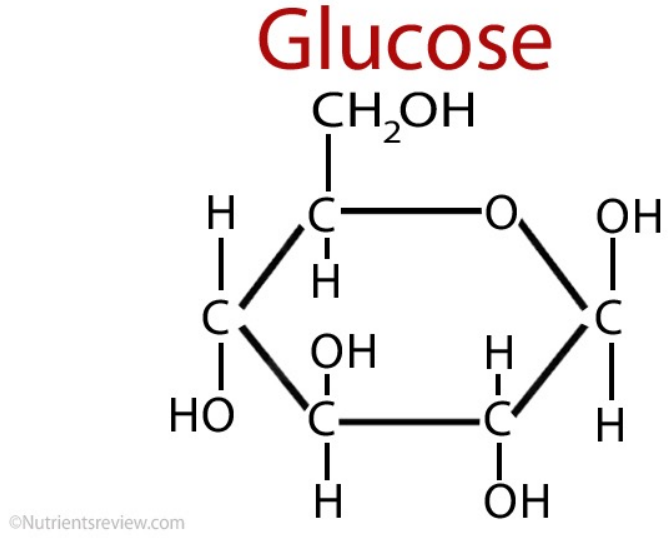
- *Network* – *Graph* (Mathematics)
- *Vertex*, *Edge* – *Node*, *Link* (CS) – *Site*, *Bond* (Physics) – *Actor*, *Tie* (Sociology)
- *n* for #node, *m* for #link
- *multi-edge*: multiple edges between two nodes
- *self-loop*: edges connecting nodes to themselves



Graph Terminology

Type	Direction	Multi-edge	Self-loop
simple graph	undirected	no	no
multigraph	undirected	yes	no
pseudograph	undirected	yes	yes
directed graph	directed	no	yes
directed multigraph	directed	yes	yes

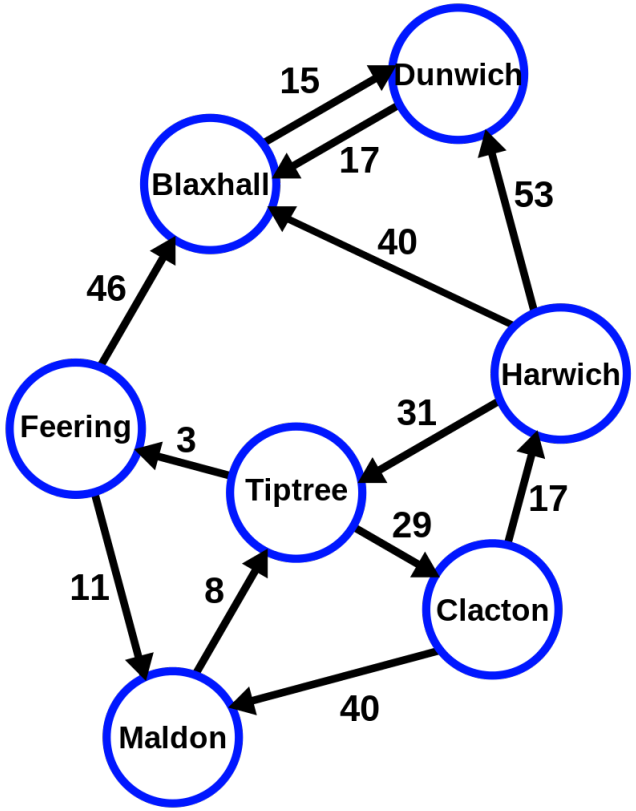
What kind of graph is chemical formulas?



Adjacency Matrix

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Weighted | Directed | Labeled Networks



Adjacency Matrix of directed graph

$$A_{ij} = \begin{cases} 1, & if\ there\ is\ an\ edge\ from\ j\ to\ i \\ 0, & else \end{cases}$$

Notice: $A_{ij} = 1$ if there is an edge from j to i, not i to j. It is for the mathematically convenience.

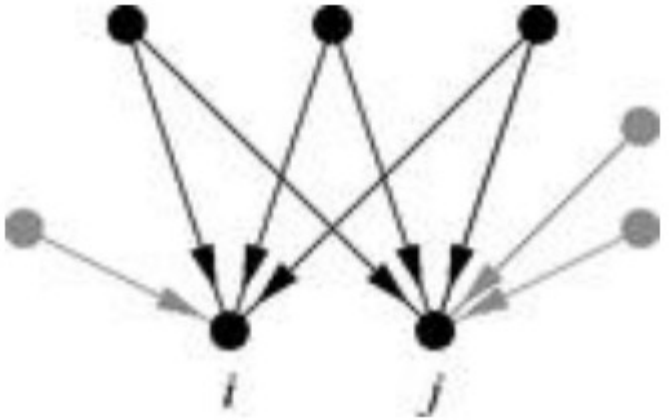
共引 Cocitation Network | 文献耦合 Bibliographic Coupling

- Translating a directed graph into an undirected graph
- *cocitation* C_{ij}: #edge pointing to node i and node j

共引次数

$$C_{ij} = \sum_{k=1}^{n} A_{ik}A_{jk} = \sum_{k=1}^{n} A_{ik}A_{kj}^{T} \ (A_{kj}^{T} \text{ is the element in } A^{T})$$

$$\mathbf{C} = \boldsymbol{A}\boldsymbol{A}^{T}$$

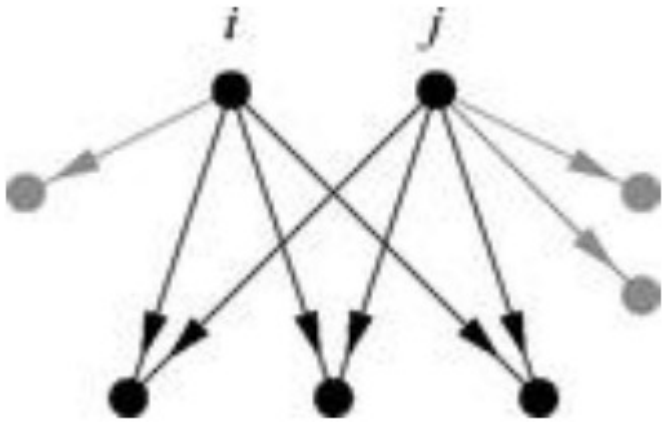


- $\mathbf{C}$ is called a *cocitation matrix*, and $\mathbf{C}$ is symmetric (why?)
- A network constructed by $\mathbf{C}$ is called a *cocitation network*

共引 Cocitation Network | 文献耦合 Bibliographic Coupling

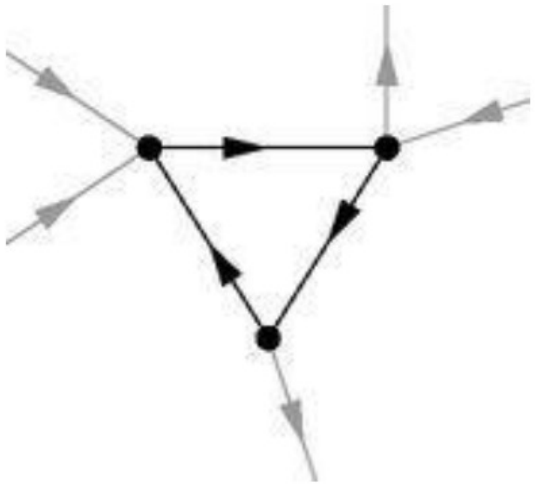
- *Bibliographic coupling* is a reverse translation of *cocitation*
- *Bibliographic coupling* B_{ij}: #edge from node *i* and node *j*

$$\mathbf{B} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$$

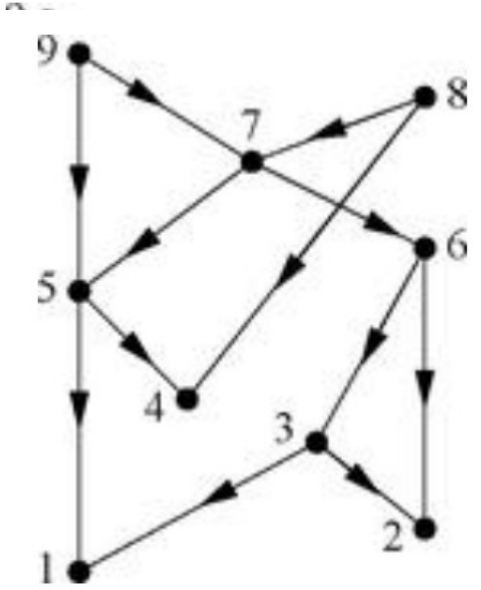


有向无环图

DAG: directed acyclic graph



A directed graph with cycle

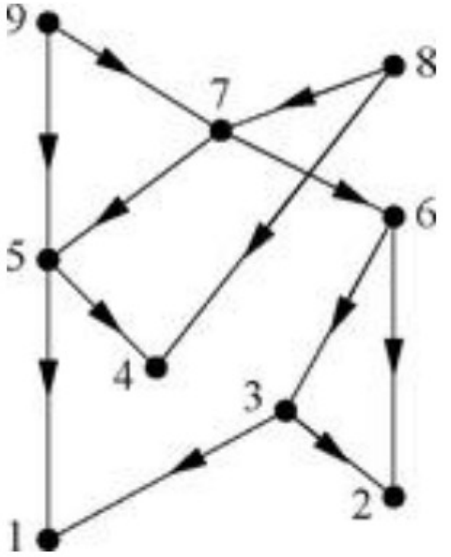


A directed acyclic graph

Judging and representing a DAG

- Each DAG can be drawn in a manner that all edges are pointing downward
- Lemma: At least one node in DAG has no outgoing edges. 出边
- Using this lemma, we can draw the downward representation of each graph , and judge if it is DAG.

? HOW?



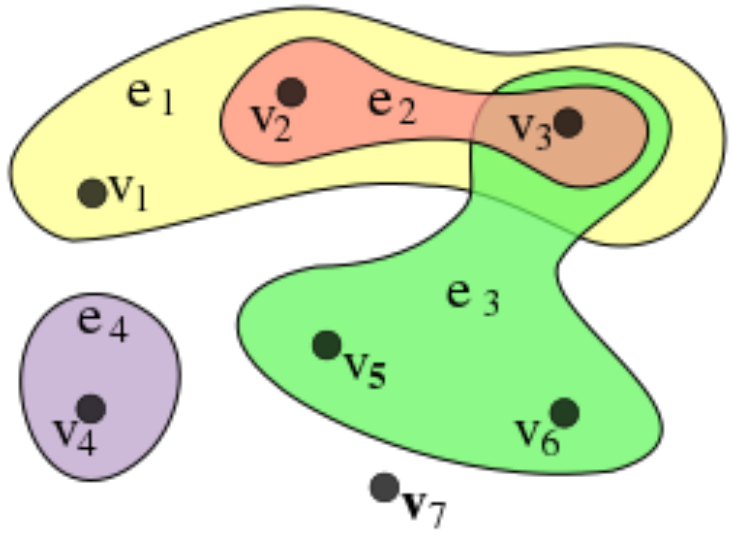
Adjacency Matrix of downward DAG

- adjacency matrix of a downward DAG is upper triangular (上三角矩阵)
- considering that there is no self-loop, the diagonal elements are zero (对角元)
- Triangular matrices with zeros on the diagonal are called *strictly triangular* (严格三角矩阵)
- all of its eigenvalues are zero (N&SC of DAG) (特征值, 充要条件)
- the diagonal elements of a triangular matrix are its eigenvalues
- matrices with all eigenvalues zero are called *nilpotent matrices*. (幂零矩阵)

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

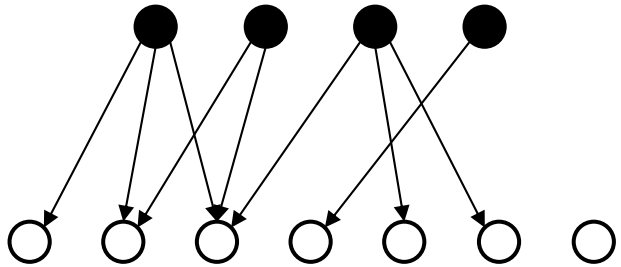
超图 Hypergraph and 二部图 Bipartite Graph

- In some networks, edges connect more than two nodes at a time.
- For example, in a social network, social actors are connected in a complex manner:
  - X = {v1,v2,v3,v4,v5,v6,v7}
  - E = {e1,e2,e3,e4} = {{v1,v2,v3},{v2,v3},{v3,v5,v6},{v4}}
- edges can be generalized into 超边 *hyperedges*
- K均匀子图 *k*-uniform hypergraph: all its hyperedges have size *k*
- *2*-uniform hypergraph: graph



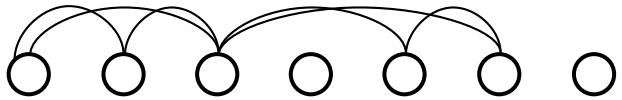
Hypergraph and Bipartite Graph

- hypergraphs can be represented by *bipartite graph*
- a bipartite graph is also called a *two-mode network* 双模网络
- *a bipartite graph can be represented by a incidence matrix* 关联矩阵
-



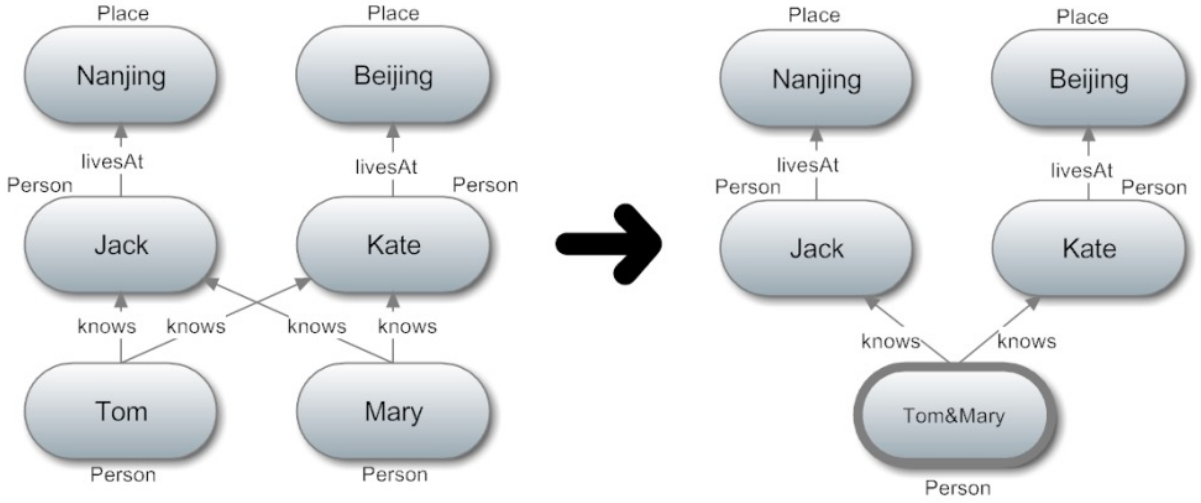
$$B = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

单模投影
- *One mode projection*: deducing a single mode graph

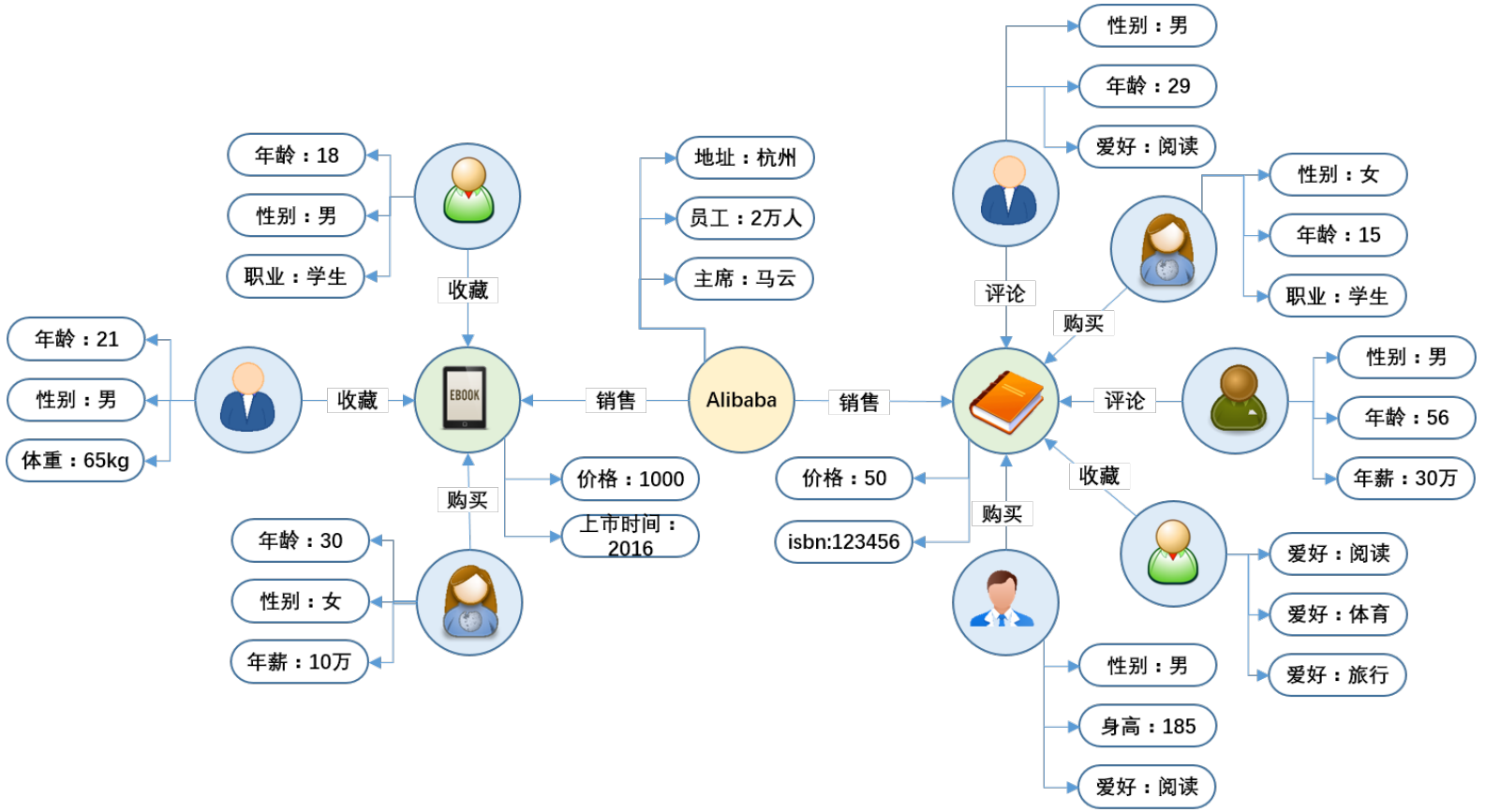


Using hypergraph in Semantic Web

- scenario 1:
  - given a set of statements: $\langle a, isMarriedTo, b \rangle$, $\langle a, hasFather, c \rangle$ ...
  - how to describe WHO, WHEN and WHERE made these statements?
  - Named Graph: https://en.wikipedia.org/wiki/Named_graph
- scenario 2:
  - given a large and redundant RDF graph
  - some subgraphs are compressible
  - how to represent compressed graph?



Multi-Mode and Multi-Dimensional Graph



Tree

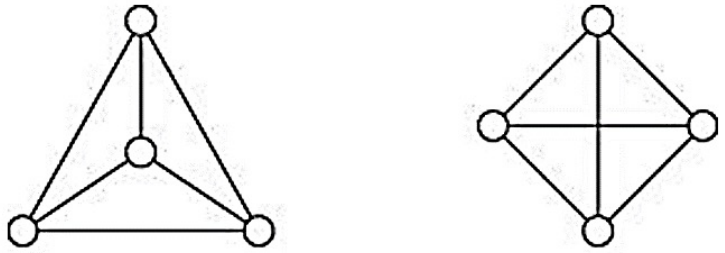
- property 1: there is exactly one path between any pair of vertices
- property 2: a tree of n vertices always has exactly n-1 edges

平面网络

Planar Network

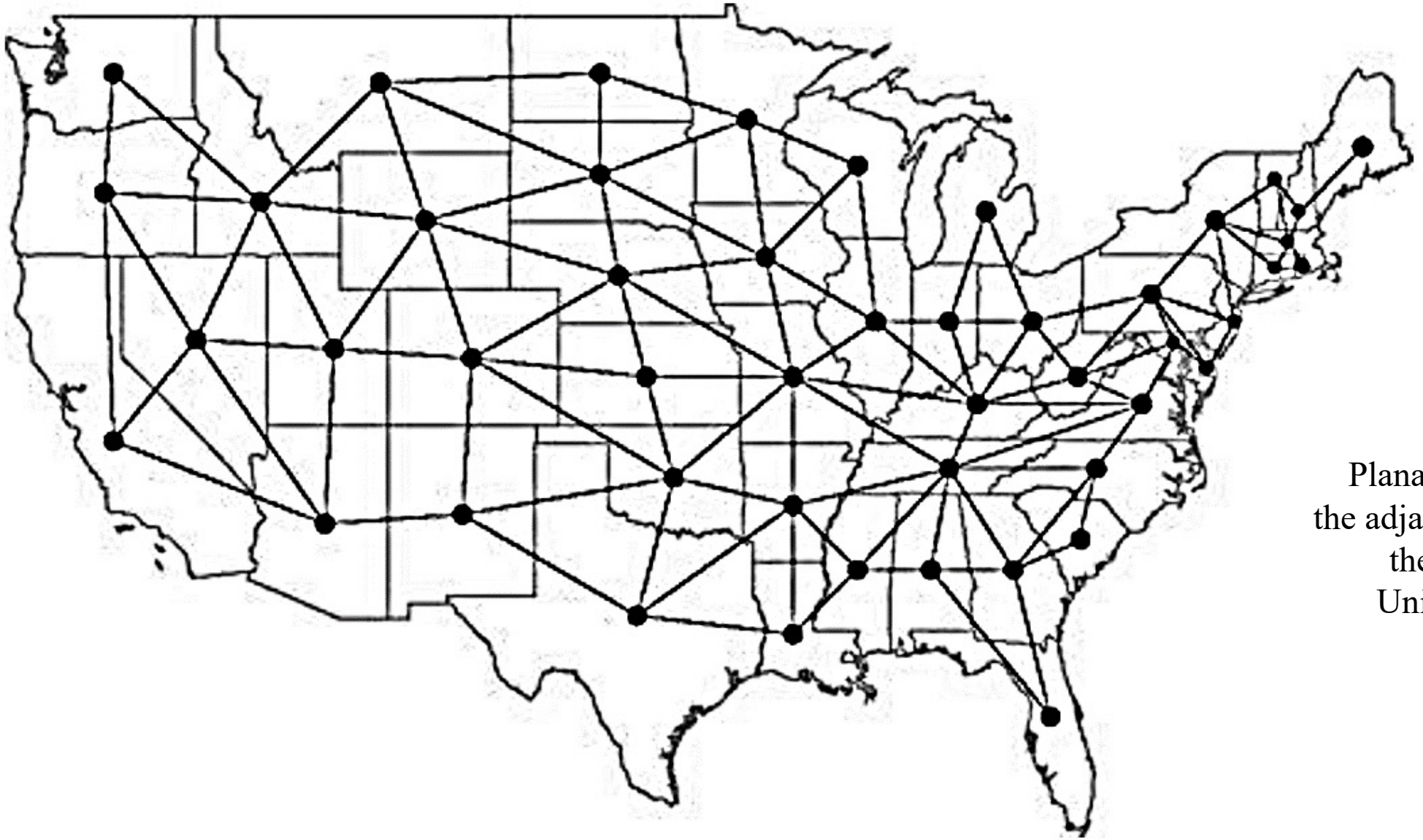
- A *planar network* is a network that can be drawn on a plane without having any edges cross

不存在交叉边



this is also a planar graph!

Two drawings of a planar graph



Planar graph of the adjacencies of the lower 48 United States

Planar Network

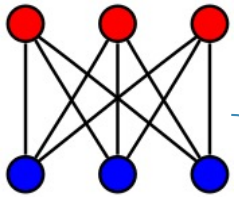
- Most real-world planar network is related to geography
- Planer network is important to *four-color theorem*
- 四色定理

How to determine a network is planar or not

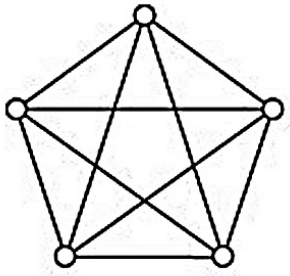
Kuratowski's Theorem: Determine a planar network

- Every non-planar network contains at least one subgraph that is an expansion of k_5 or UG (utility graph, or $k_{3,3}$)

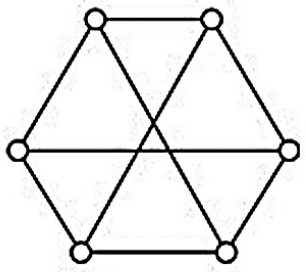
在已有的边上添加新的节点



为什么UG是完全二部图？



(a) K_5

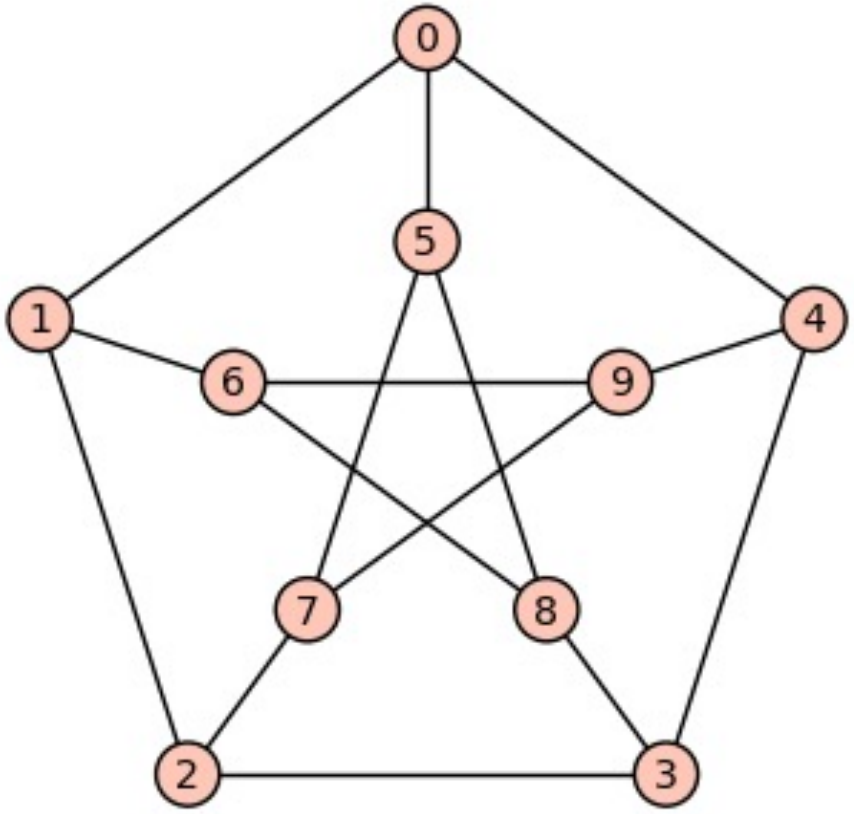


(b) UG

k_5:5个节点的完全图,为非平面图

Utility Graph:应用图，可以表示为有两个三节点群组构成的完全二部图

Kuratowski's Theorem: Determine a planar network



https://www.cs.sfu.ca/~ggbaker/zju/math/planar.html

Degree

Undirected graph		Directed graph
$k_i = \sum_{j=1}^{n} \mathrm{A}_{ij}$	degree and adjacency matrix	$k_i^{\mathrm{in}} = \sum_{j=1}^{n} \mathrm{A}_{ij}\,, k_j^{\mathrm{out}} = \sum_{i=1}^{n} \mathrm{A}_{ij}$
$2m = \sum_{i=1}^{n} k_i$	degree and number of edges	$m = \sum_{i=1}^{n} k_i^{\mathrm{in}} = \sum_{i=1}^{n} k_i^{\mathrm{out}} = \sum_{ij} \mathrm{A}_{ij}$
$c = \frac{1}{n}\sum_{i=1}^{n} k_i = \frac{2m}{n}$	average degree	$c_{\mathrm{in}} = c_{\mathrm{out}} = \frac{m}{n}$

Density

- density ρ (for simple graph)

$$\rho = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)} = \frac{c}{n-1}$$

n个节点的图中最大可能边数

密集图

- if $n \to \infty$, $\rho \to$ constant, the graph is *dense*

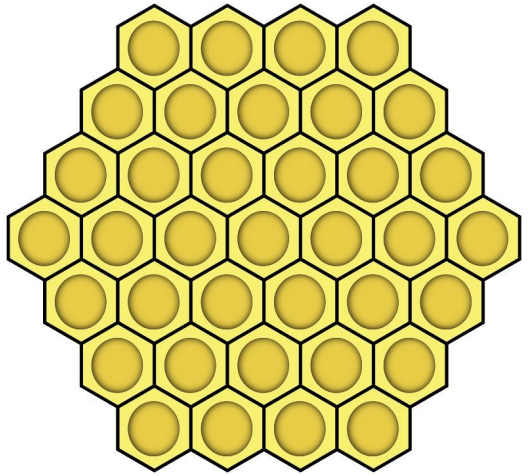
稀疏图

- if $n \to \infty$, $\rho \to 0$, the graph is *sparse*
- Internet, WWW, Social Networks (most networks) are sparse networks
- Food Network is dense network

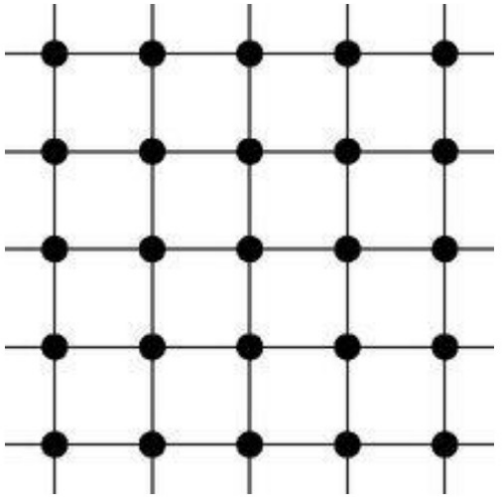
正则图

Regular Graph

- *k*-regular graph: all vertices have degree *k*



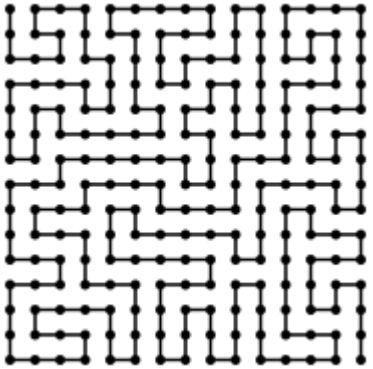
An infinite honeycomb is an example of a 3-regular graph



An infinite square lattice is an example of a 4-regular graph

Path

- self-avoid path: paths that do not intersect themselves
- path length: #edge, NOT #nodes
- calculating number of *r*-path between *i* and *j*:



A self-avoiding walk

既链接*i*又链接*j*的节点*k*

*i*和*j*之间长度为2的路径个数 → $N_{ij}^{(2)} = \sum_{k=1}^{n} A_{ik}A_{kj} = [A^2]_{ij}$ ← A^2矩阵中第*ij*个元素

$$N_{ij}^{(3)} = [A^3]_{ij}$$

$$N_{ij}^{(r)} = [A^r]_{ij}$$

无环图

Acyclic Graph

幂零矩阵

“the graph described by a nilpotent adjacency matrix (i.e., a matrix whose eigenvalues are all zero) must be acyclic ”

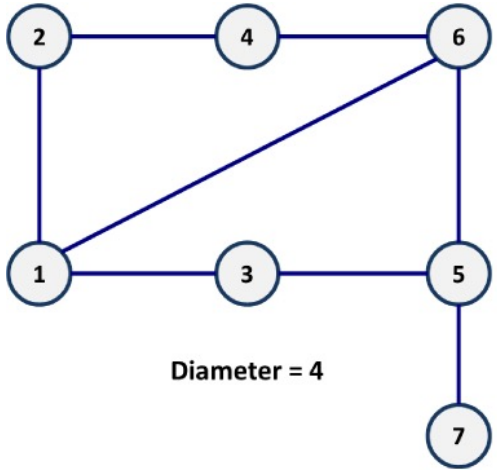
测地路径

Geodesic Path (shortest path)

- simple graph: BFS
- directed and weighted graph
  - single-source shortest path: Dijkstra ($O(n^2)$)
  - all-pair shortest path: Floyd ($O(n^3)$)

图的直径

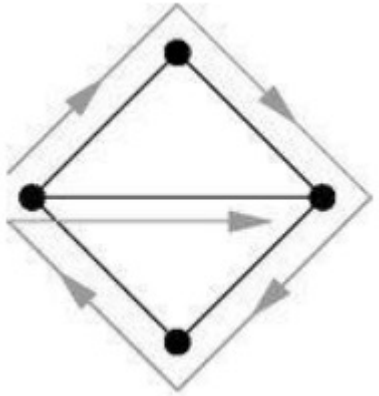
- *Graph Diameter*
  - length of longest shortest path



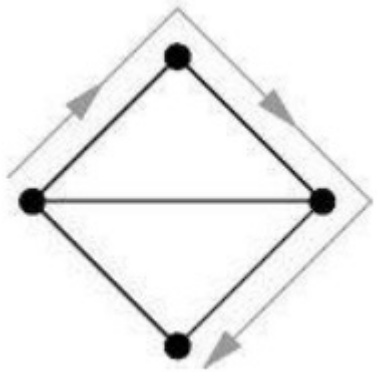
欧拉路径 汉密尔顿路径

Eulerian Path and Hamiltonian Path

- Eulerian path: a path that traverses each edge in a network exactly once. Not necessary to be self-avoiding.
- Hamiltonian path: a path that visits each vertex exactly once. Must be a self-avoiding path.

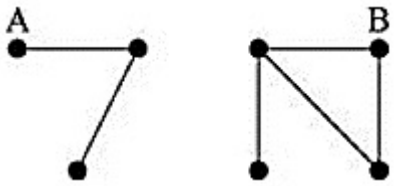


Eulerian path



Hamiltonian path

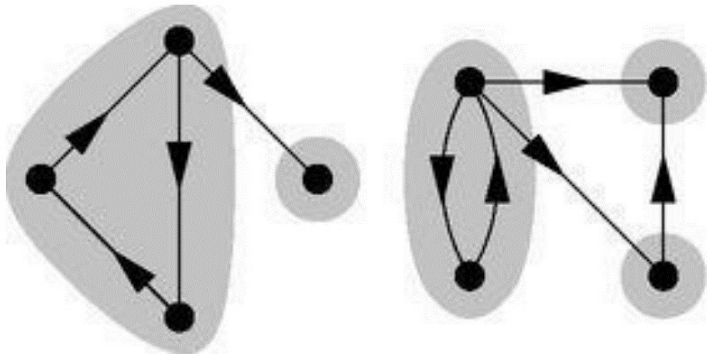
分支
Component



- The adjacency matrix of a network with more than one component can be written in block diagonal form 分块对角形式
  - non-zero elements of the matrix are confined to square blocks along the diagonal of the matrix 非零元素在沿对角线的正方形中
  - all other elements being zero
    其他均为零元素

$$\mathbf{A} = \begin{pmatrix} \square & 0 & \cdots \\ 0 & \square & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

Components in Directed Graph



弱连通分支

A graph with WCC and SCC

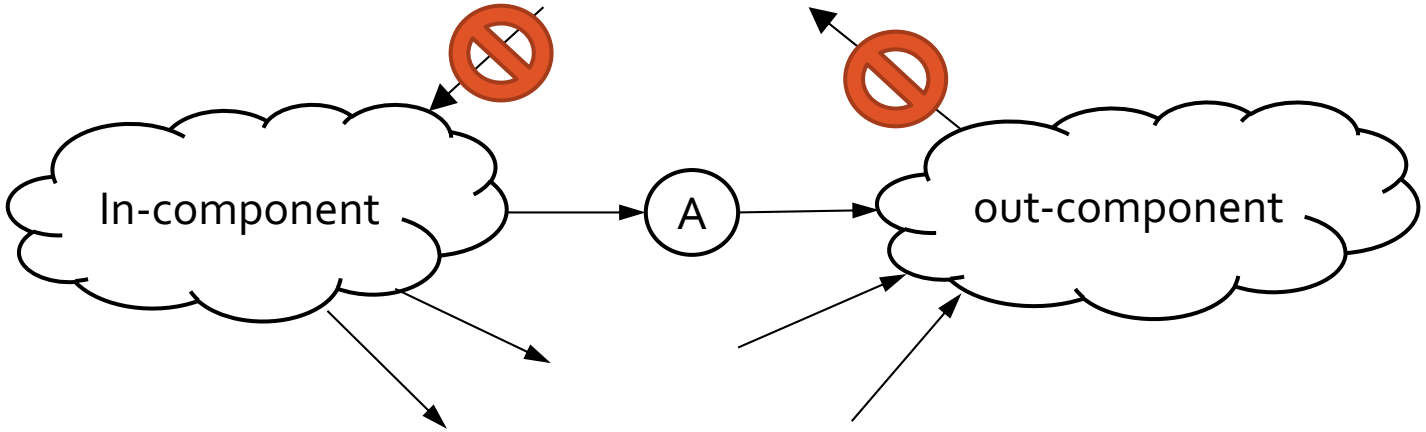
强连通分支(任意两个顶点存在有向路径)

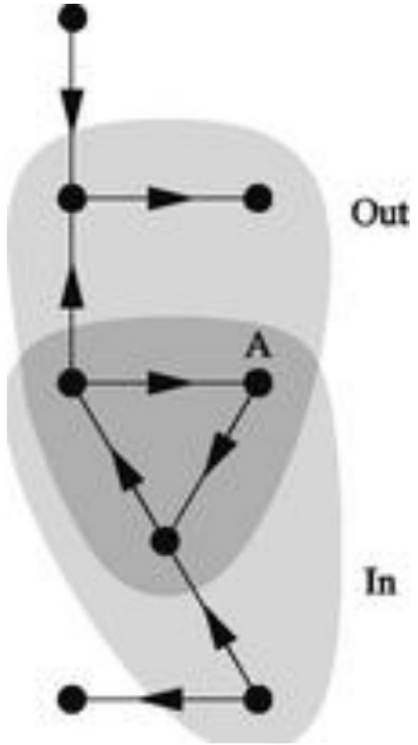
1. there must be circles in a SCC with #node>1;
2. every node in these SCC must belong to a circle;
3. there is no SCC in DAG

外向分支 内向分支

Out-component and In-component

- Out-component: reachable nodes from a specific node
  - edges connecting OC to other vertices only point inward, never outward
- In-component: nodes that can reach to a specific node
  - edges connecting IC to other only point outward, never inward





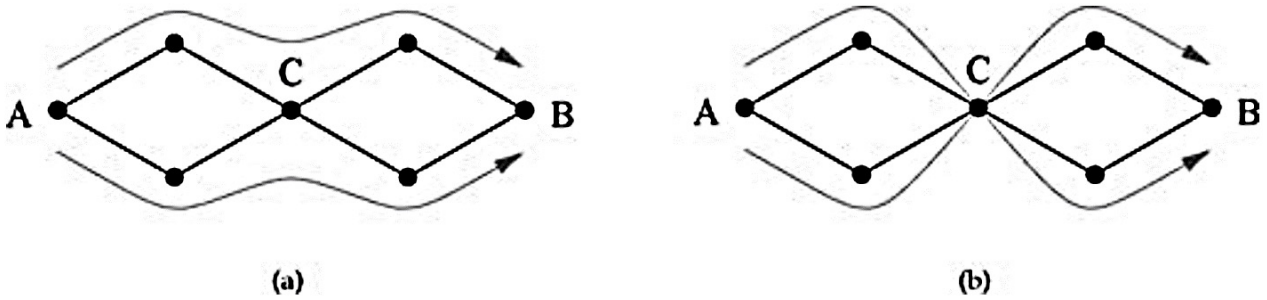
The in- and out-components of a vertex A in a small directed network

独立路径

Independent Path

用来评估两个点之间的连接强度

- there are more than one path between node *A* and *B*
- some paths share no vertex (vertex-independent) or edge (edge-independent)



From *A* to *B*, there are two edge-independent paths, but only one vertex-independent path

连通度

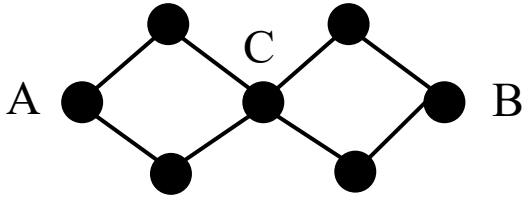
Connectivity

- edge connectivity = # edge-independent path
- vertex connectivity = # vertex-independent path

- e-connectivity(A,B)=2; v-connectivity(A,B)=1

割集
Cut Set

- Given A and B are connected,
  - the cut set is a set of vertex $\{v_1, v_2 \ldots v_k\}$
  - removing the cut set will make A and B disconnected
  - 最小割集 minimum cut set (MCS) is the smallest cut set
  - {C} is the MCS of A and B

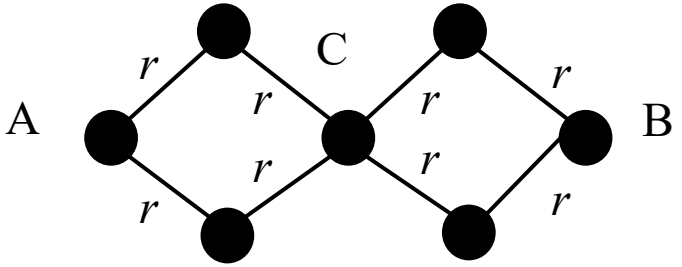


门格尔定理
Menger's Theorem *implies*:
|MCS(A,B)| = v-connectivity(A,B)

最大流

Maximum Flow

自来水管网络



- a water pipeline from A to B
- max rate on each edge: *r*
- max rate from A to B:

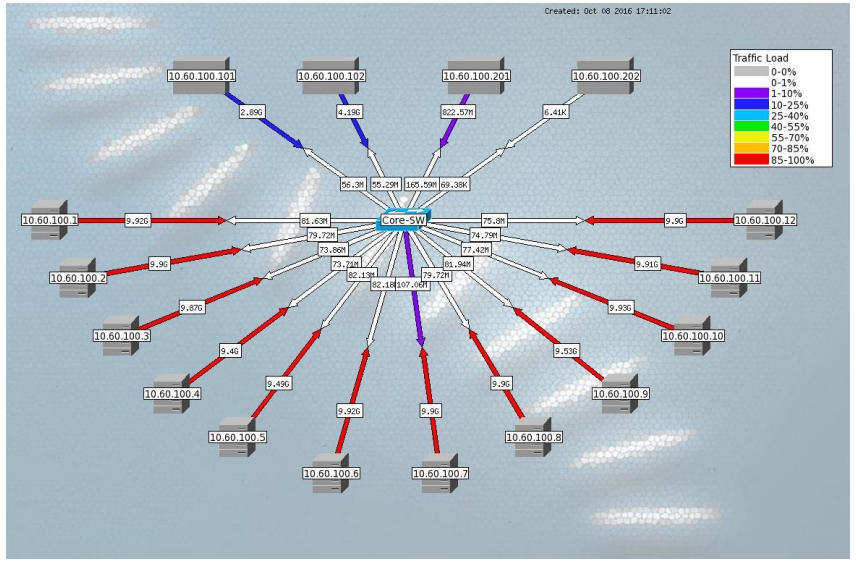
  max-rate(A,B) = e-connectivity(A,B) * *r*

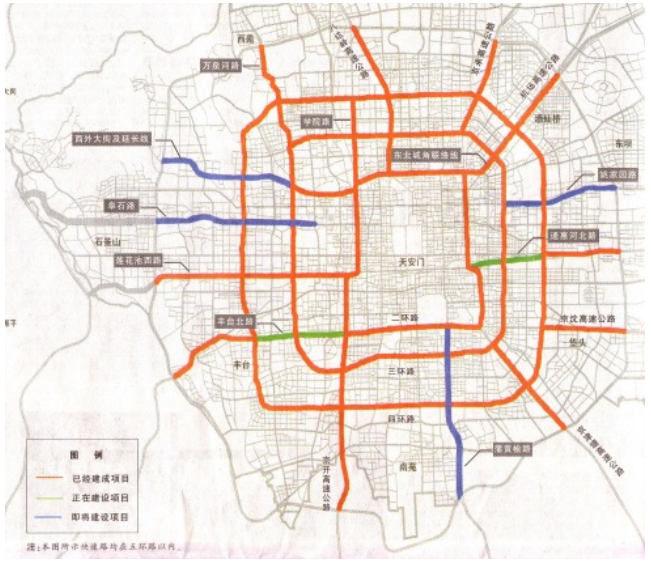
增广路径算法

- Augment Path Algorithm

Application of min Cut | max Flow

- Finding bottleneck in communication network 通信网络中的带宽瓶颈问题
- Predicting maximum traffic in transportation planning 交通规划中的最大流量问题
- border detection in image segmentation 图像分割中的边缘检测问题

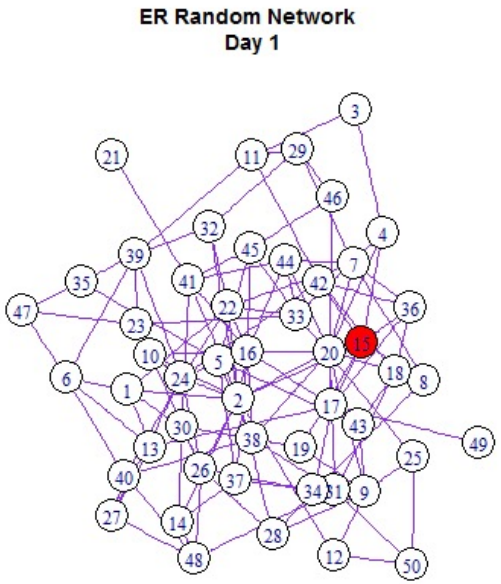





Sobel算子

图拉普拉斯矩阵

Graph Laplacian Matrix



扩散过程

- diffusion process (gas, viewpoint, disease)

类似于气体扩散方程中的拉普拉斯算子∇^2

- Graph Laplacian Matrix: describe the diffusion process in a graph

对角矩阵　邻接矩阵

$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

第i个顶点的度

$$\boldsymbol{D} = \begin{pmatrix} k_1 & 0 & 0 & \dots \\ 0 & k_2 & 0 & \dots \\ 0 & 0 & k_3 & \dots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

t时间点第i个节点的状态　t时间

$$\boldsymbol{a_i(t) = a_i(0)e^{-c\lambda_i t}}$$

第i个拉普拉斯特征值　扩散速率　初始状态

Application of Graph Laplacian Matrix

- calculating the number of components
  - the number of components equals to the number of non-zero eigenvalues of laplacian matrix

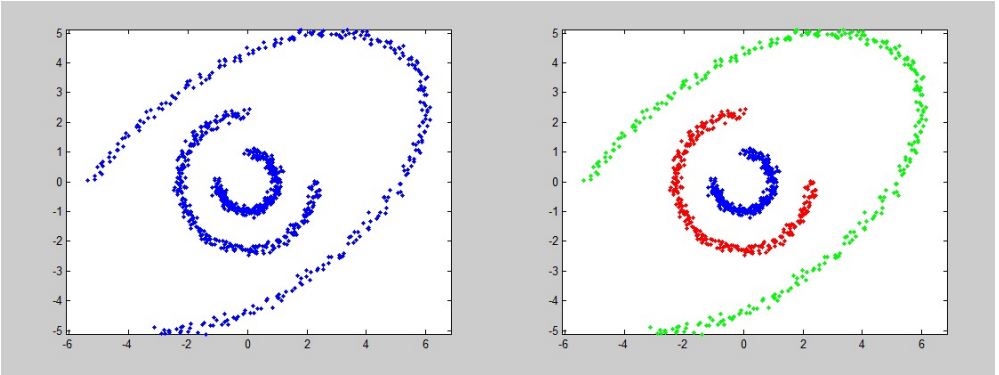
非零特征值的个数

随机游走

- analysis of random walk
  - *given*: a random walk that starts at a specified vertex *i* and takes *t* random steps
  - *goal*: $p_i(t)$, the probability that the walk is at vertex *i* at time *t*
  - link analysis

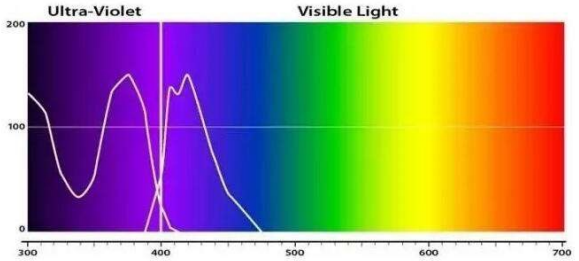
- graph clustering
  - spectrum clustering

谱聚类



图谱理论

Graph Spectrum Theory



代数图论

- Algebraic Graph Theory: using algebra to analyze graph structures

分治法

- divide-and-conquer: divide a complex structure into separated components
  - in optics: optical spectrum 光谱
  - in graph: graph spectrum 图谱

时间维度 空间维度

- understanding graph structure in temporal dimension and spatial dimension

-

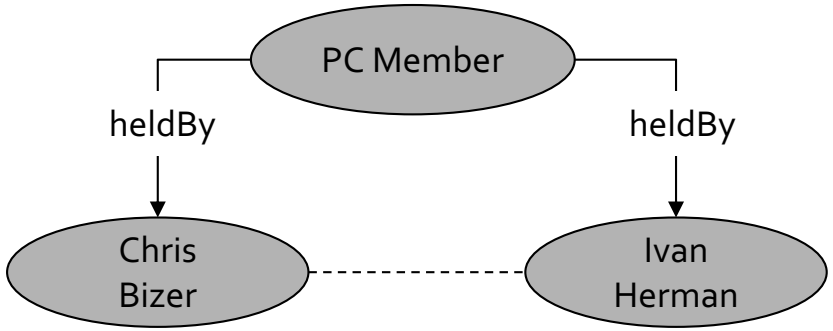
ASSIGNMENT

Text Similarity

- Given two text document (such as news reports) D_1 and D_2
- How to calculate the similarity between D_1 and D_2
- Solutions:
  - Jaccard Similarity
  - TF/IDF + Cosine Similarity
  - Word Embedding + Cosine Similarity
  - Knowledge-based Similarity
- Paper writing: a team paper can be collaborated written including:
  - abstract / problem definition / approach overview
  - details of algorithms / experiments / related works / reference
  - PPT

Community Detection

- Title: Finding Communities in a (Social) Network
- Dataset: Stanford Large Network Dataset Collection (https://snap.stanford.edu/data/)
- Goal: to detect communities in given networks
- Requirements:
  - a paper
  - implements more than two clustering algorithms
  - quantitative evaluation
  - compare the results of different clustering algorithms
  - PPT



ARRANGEMENT OF DISCUSSION AND PRESENTATION

- Time: Next Wednesday afternoon (2 pm ~ 5 pm)
- Style: Online (Tencent online conference)