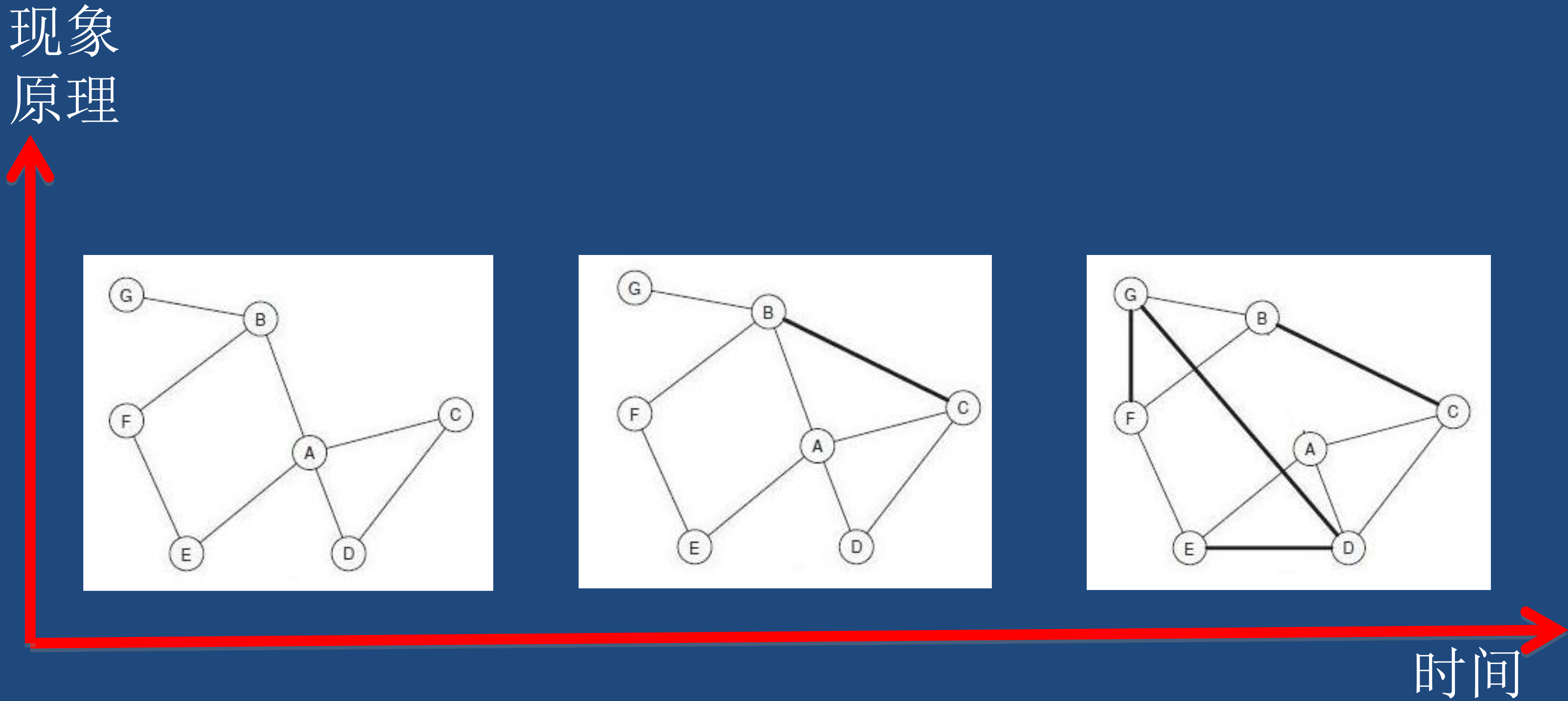
社会网络的基本概念及其在OSN上的体现

三元闭包，关系的强度及其与网络结构的关系，同质性及其影响，正负关系及其平衡，幂率，小世界，节点的地位与关系的均衡，…

社会网络，不仅是人类社会的一个属性

- 但在人类社会体现得最丰富，最多姿多彩
- 人类在社会网络中的行为，能否在基因中找到原因？

讨论社会网络的空间



- 不仅考虑一个时刻（“快照”）上的性质
- 更要研究随时间发生的变化

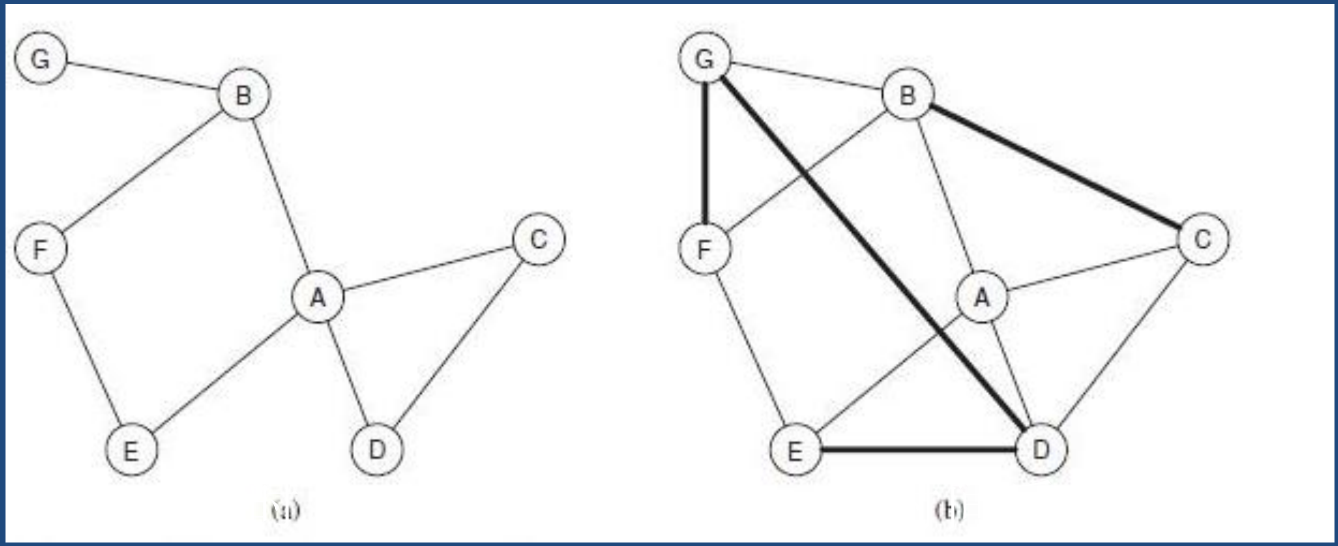
提要

- 三元闭包（triadic closure），社会网络的基本成因
- “关系”的强弱与网络结构的关系
  - 启示，假说，假说的论证（抽象形式化的，数据支持的）
  - OSN中的关系强度
- 在社会网络中跨越“结构洞”的节点的性质分析（“责权利”）

三元闭包（闭合）

- 社会网络的最基本成因
  - Anatole Rapoport（阿纳托尔•拉波波特，1953）

如果两人在社会网络中有一个共同的朋友，则他们俩将来成为朋友的可能性提高。



机会？
opportunity
信任？
trust
动机？
incentive

三元闭包原理的拓展

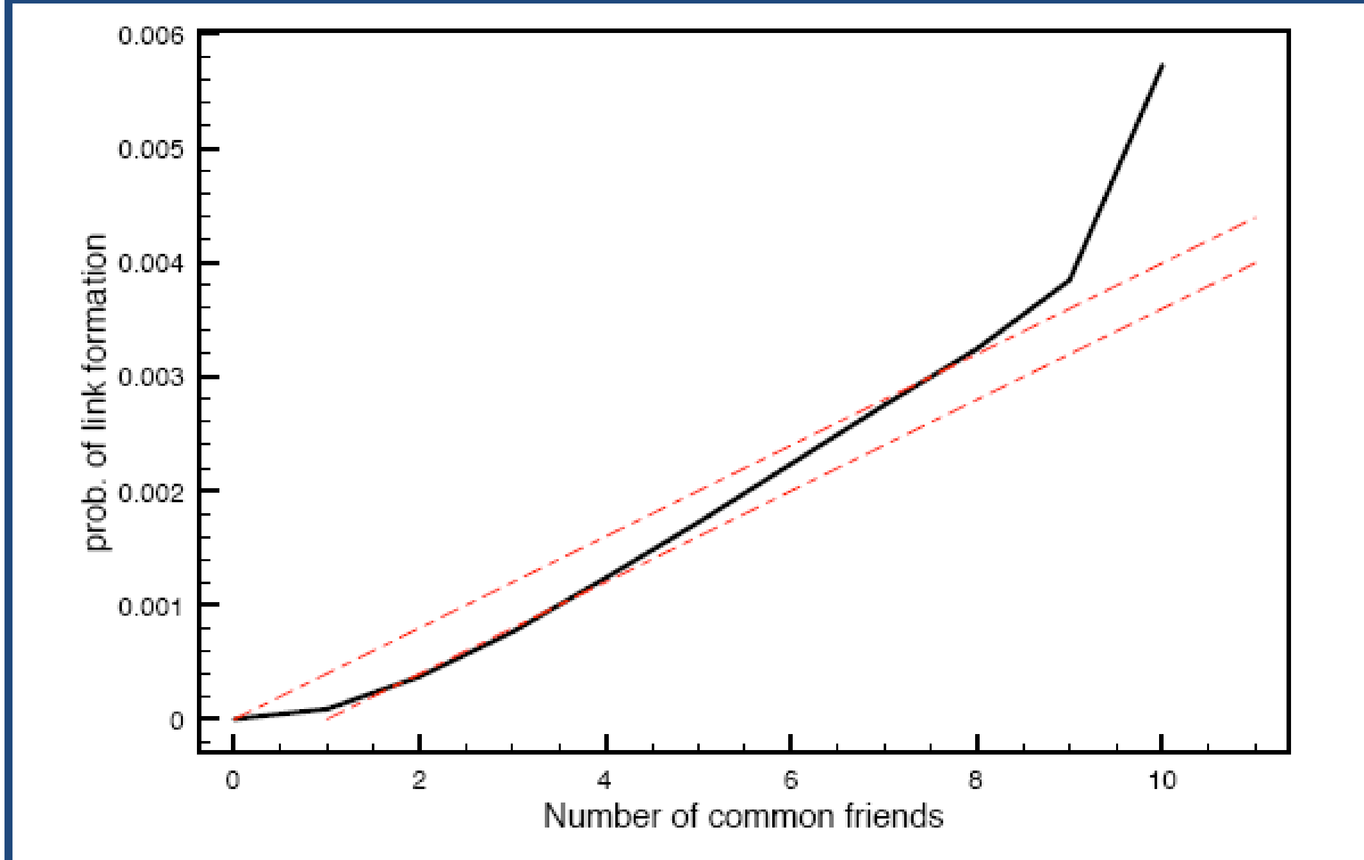
- 两个人的共同朋友越多，则他们成为朋友的可能性越高
  - 这是从“量”方面的拓展
- 两个人与共同朋友的关系越密切，则他们成为朋友的可能性越高
  - 这是从“质”方面的拓展
- 三个原因（机会、信任、动机）的作用在这些拓展的意义上保持一致

如何验证？

一个利用在线数据研究三元闭包的例子

- 电子邮件网络≈社会网络
  - 节点：一定范围的邮件地址（例如一个大学）
  - 边：一段时间（例如一个月）里有邮件通信
    - 单向 vs 双向？多少才算？…
- 网络的演化
  - 什么叫两个相继的网络快照？
  - 两个相继的快照就能说明问题？（回避偶然性事实，大量快照对的平均）
- 如何定义考察三元闭包现象的测度？
  - 简单统计闭包个数 vs 共同朋友数的影响

结果及其含义



＊在电子邮件网络上三元闭包迹象明显——共同朋友有助于关系的建立；

＊突出体现在1—2个共同朋友情形；

＊为什么8—9—10也突出？

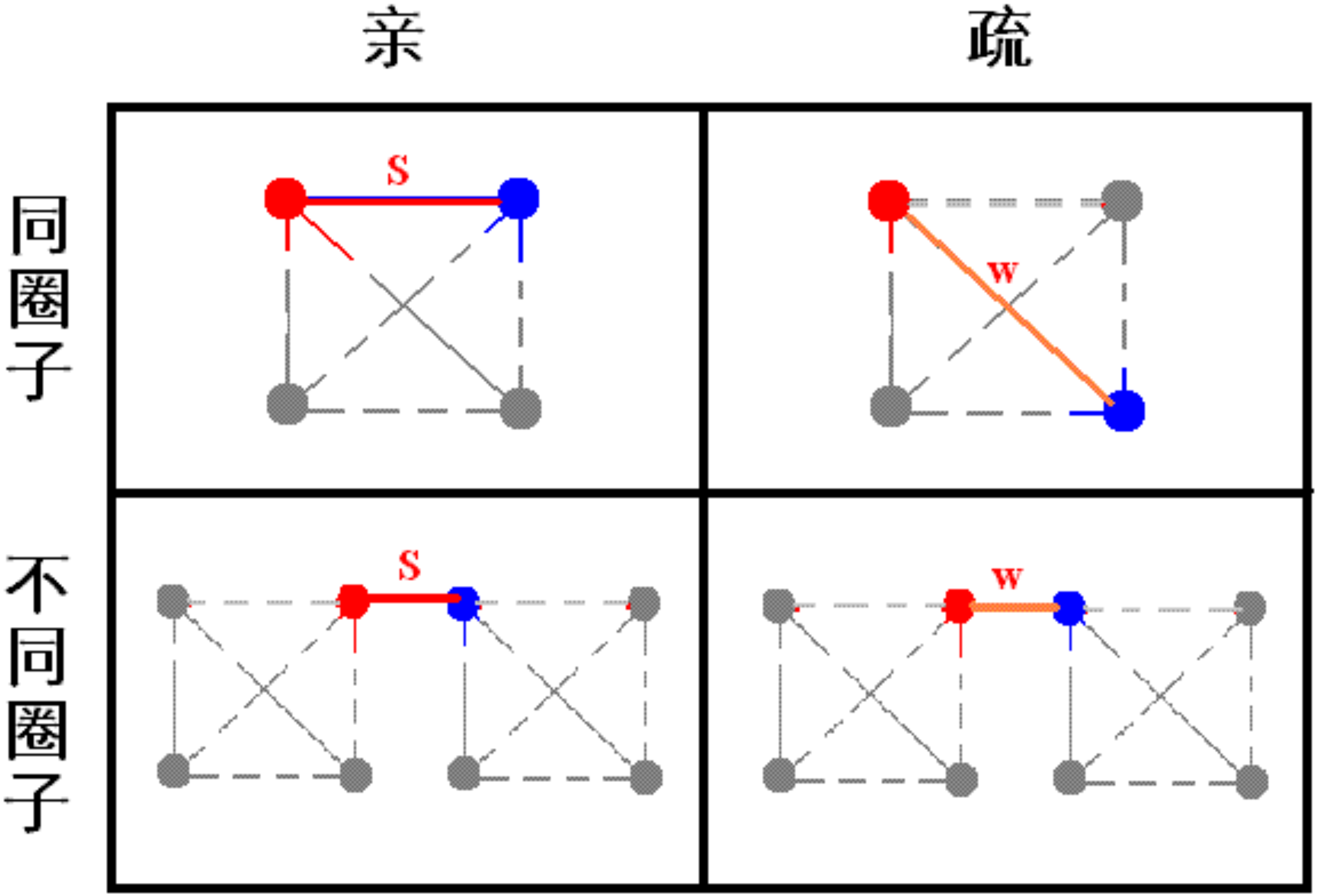
- 特定电子邮件网，其他网络如何？
- 定量分析 vs 定性结论

Science 2006

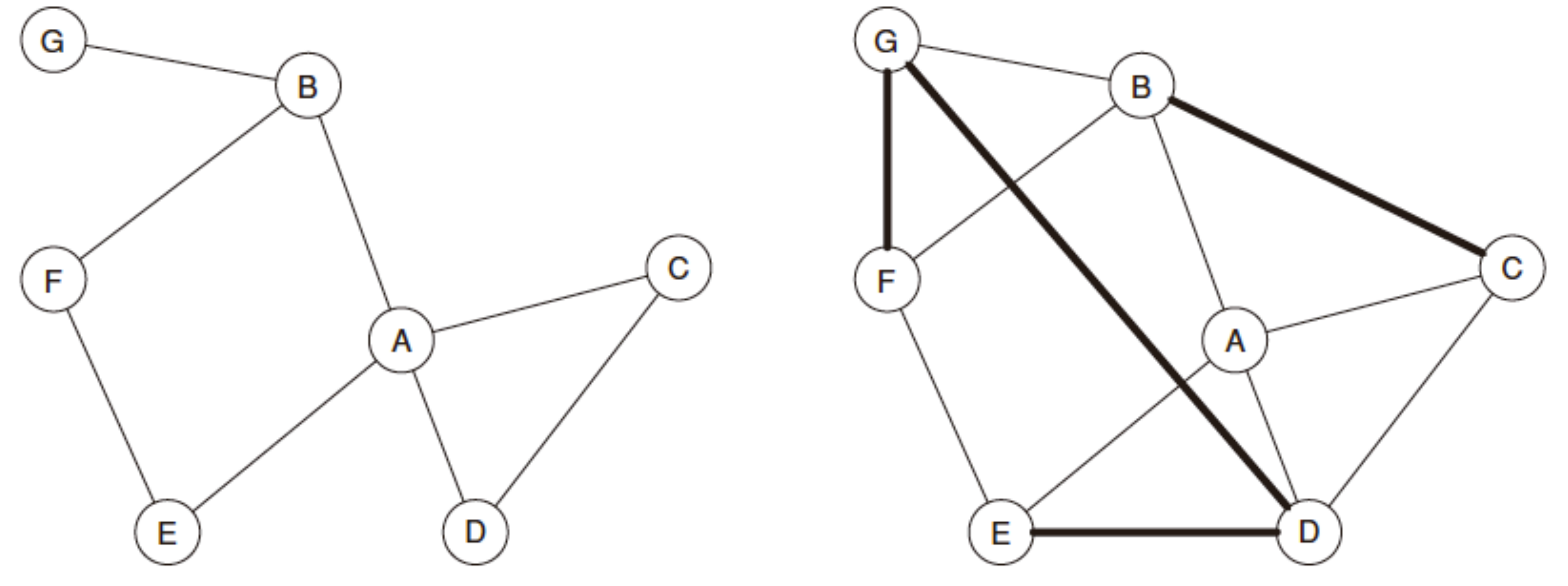
格兰诺维特的诧异

- Mark Granovetter, “The Strength of weak ties” *American Journal of Sociology*, 1973.
- Mark Granovetter, *Getting a Job: A study of Contacts and careers*. University of Chicago Press, 1974.
- 为什么对找工作提供有效帮助的人更多只是一般熟人，而不是亲密朋友？
  – 两个层面的认识，导致对社会关系（网络）两个维度的视角

社会关系的两个视角



节点的聚集系数：邻居间三元闭包体现的强度



(a) (b)

节点A的聚集系数 = A的任意两个朋友之间也是朋友的概率（即邻居间朋友对的个数除以总对数）

聚集系数

- 一个网络的聚集系数 *C满足*:

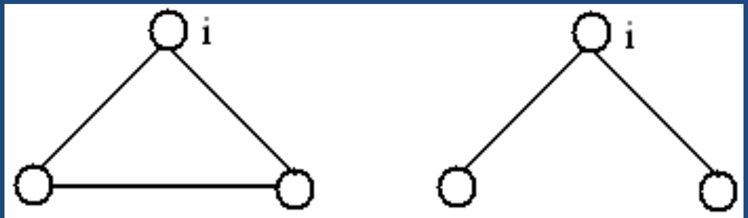
$0 < C < 1$

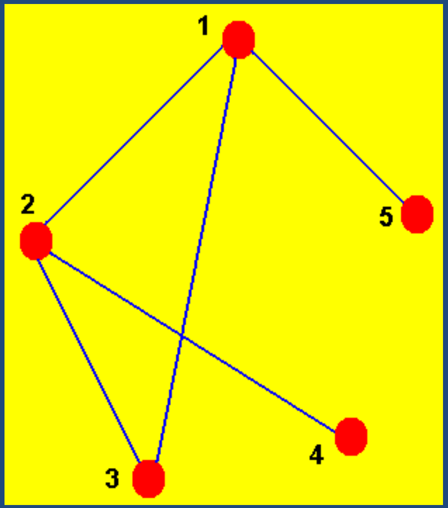
  - $C = 1$ if 任意两个节点有连接
  - $C = 0$ if 无三角形连接

- 大部分复杂网络有较大的 C → 小世界特征
- *富者越富，马太效应*

聚集系数的定义:

$$C(i) = \frac{number\ of\ \ complete\ triangles\ \ with\ \ corner\ \ i}{number\ of\ all\ triangular\ graphs\ with\ corner\ i}$$



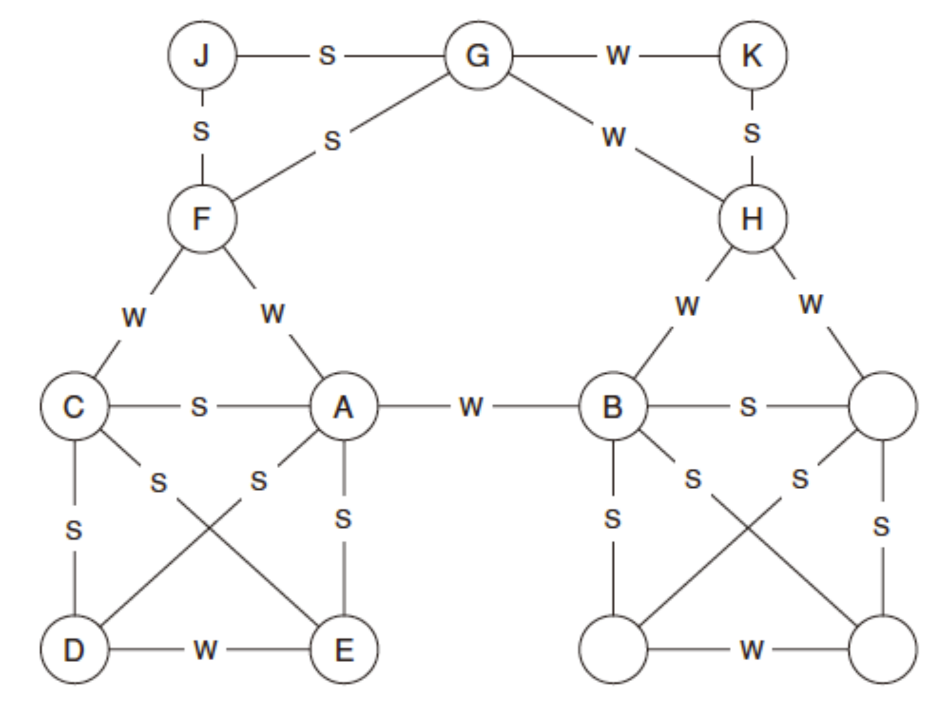


Node-1 has 1 complete triangle and 3 triangular graphs, so *C*(1) = 1/3
Node-2 has 1 complete triangle and 3 triangular graphs, so *C*(2) = 1/3
Node-3 has 1 complete triangle and 1 triangular graph, so *C*(3) = 1
Node-4 has 0 complete triangles, so *C*(4) = 0
Node-5 has 0 complete triangles, so *C*(5) = 0

Average *C* = (1/3+1/3+1+0+0) / 5 = 1/3

关系的强度

- 含义：亲密程度 vs 这一阶段联系频度
  - 尽管“亲密”与“联系的频度”并不是独立的
- 程度：一定范围的数值 vs “强”和“弱”
  - 这里只用强弱，以突出核心思想；后面，会看到也可能用可以测量的某种数值来表达

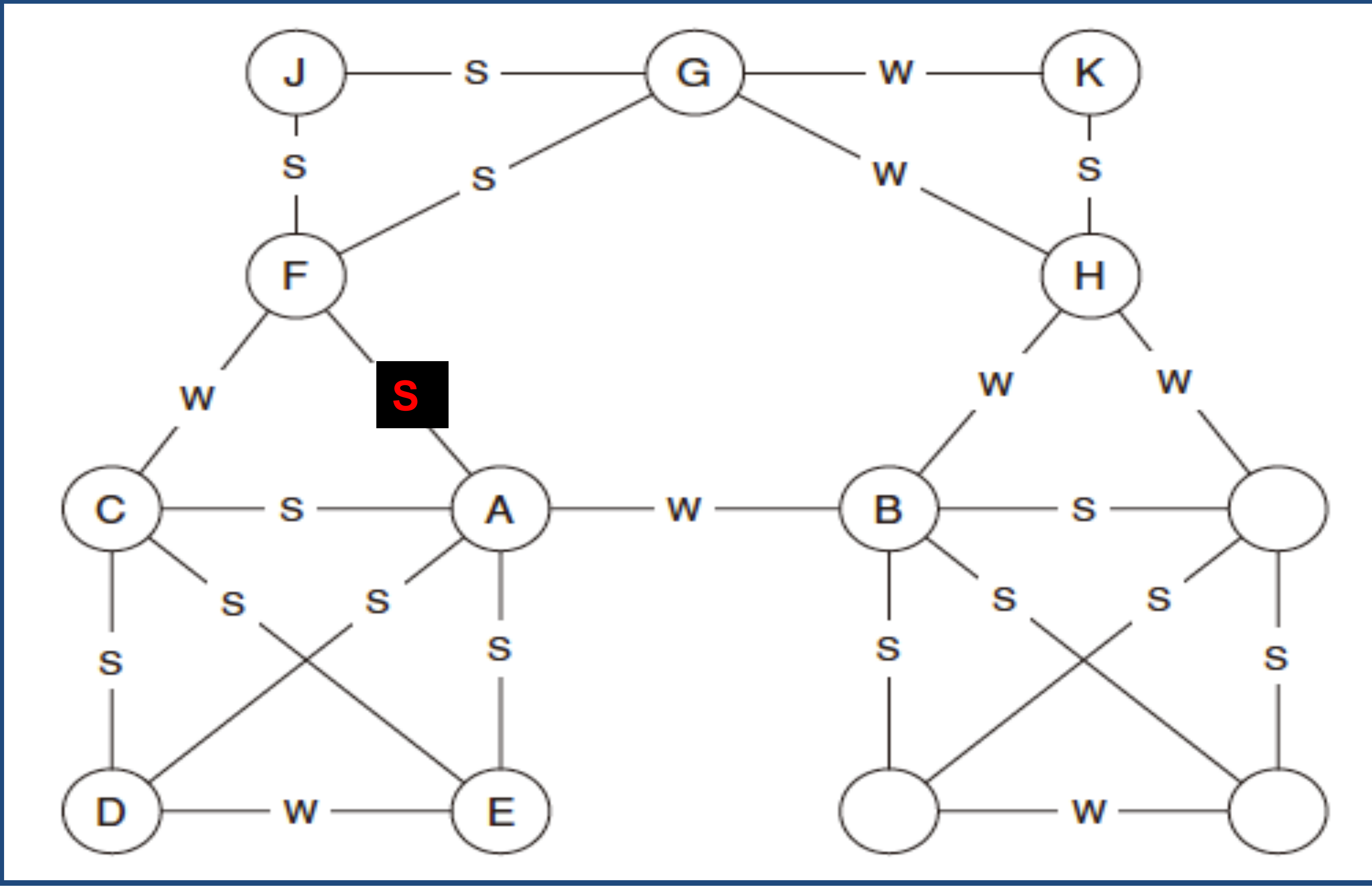


强三元闭包：在标注了关系强弱网络中的节点的一个属性

- 强三元闭包原理（假设）
  - 如果A-B和A-C之间的关系为强关系；则B-C 之间形成边的可能性应该很高；
- 若A有两个强关系邻居B和C，但B-C之间没有任何关系（s或w），则称节点A违背了强三元闭包原理；
- 如果节点A没有违背强三元闭包原理，则称节点A符合强三元闭包原理。

注意：如同聚集系数，一个节点是否符合强三元闭包也是严格定义的，即每个节点要么“符合”，要么“违背”。

哪些节点符合 / 违背强三元闭包？



捷径 = 弱关系？

- 断言：**若节点A符合强三元闭包，且至少有两个强关系邻居，则与A相连的任何捷径必定意味着是弱关系。**
- （证明虽然很简单，但结论的意义重要，以及得到这个结论的思路漂亮）
- 纯数学的证明，得到了一个具有社会学意义的结论
- 这个结论将一个局部概念（关系）和一个全局概念（捷径）连接了起来

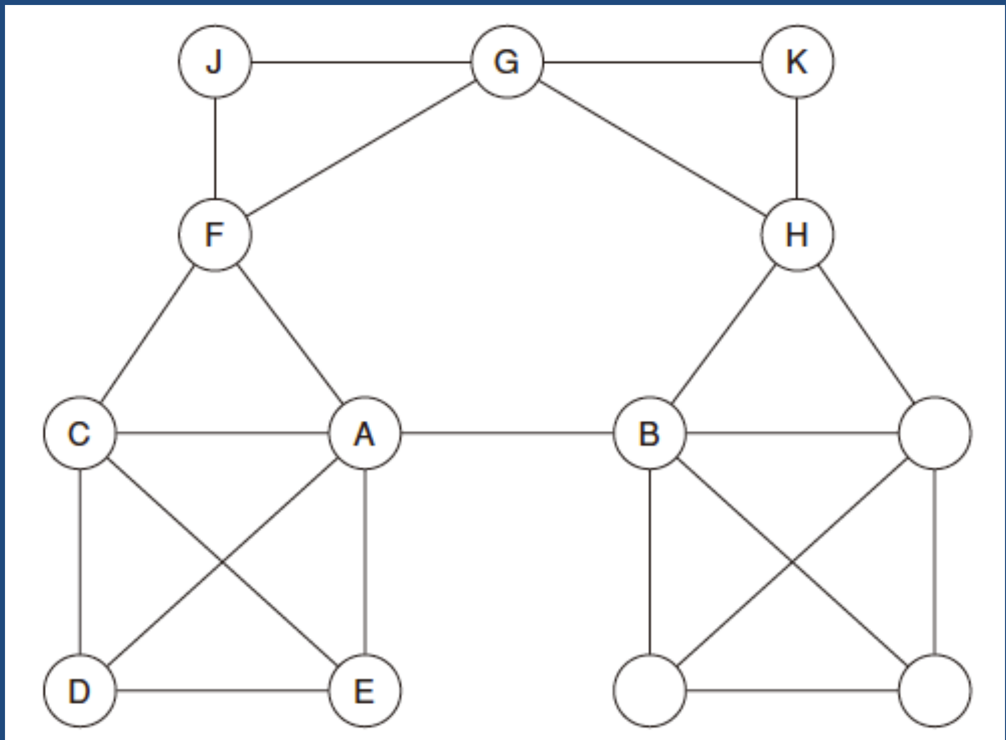
有没有数据来支持这结论？

- 上述结论的精神：两人关系的强度与是否有共同朋友直接相关
  - 捷径意味着没有共同朋友，强度为“弱”。
- 推论：共同朋友数越多，关系的强度越高
  - 精细一些，可以说共同朋友数在总朋友数中的占比（邻里重叠度）
- 我们来找一个能验证这个推论的场景，从而也就间接验证上述结论

用什么社交网络？如何定义关系的强度

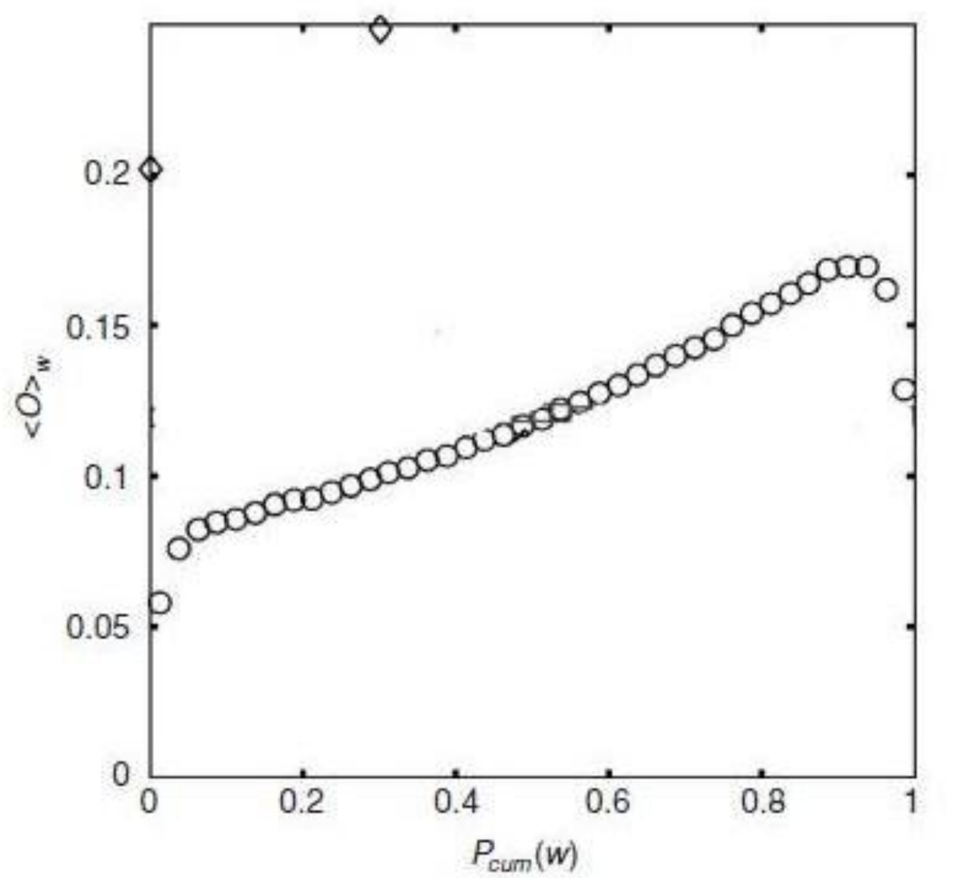
边（A, B）的邻里重叠度

- 与A和B相邻的节点数/与A或B相邻的节点数（不算A和B本身）



A-F边的邻里重叠度：
C既与F相邻，也与A相邻
与A或F相邻的则有B, C, D, E, G, J
则邻里重叠度为1/6

捷径 = 邻里重叠度为0的边



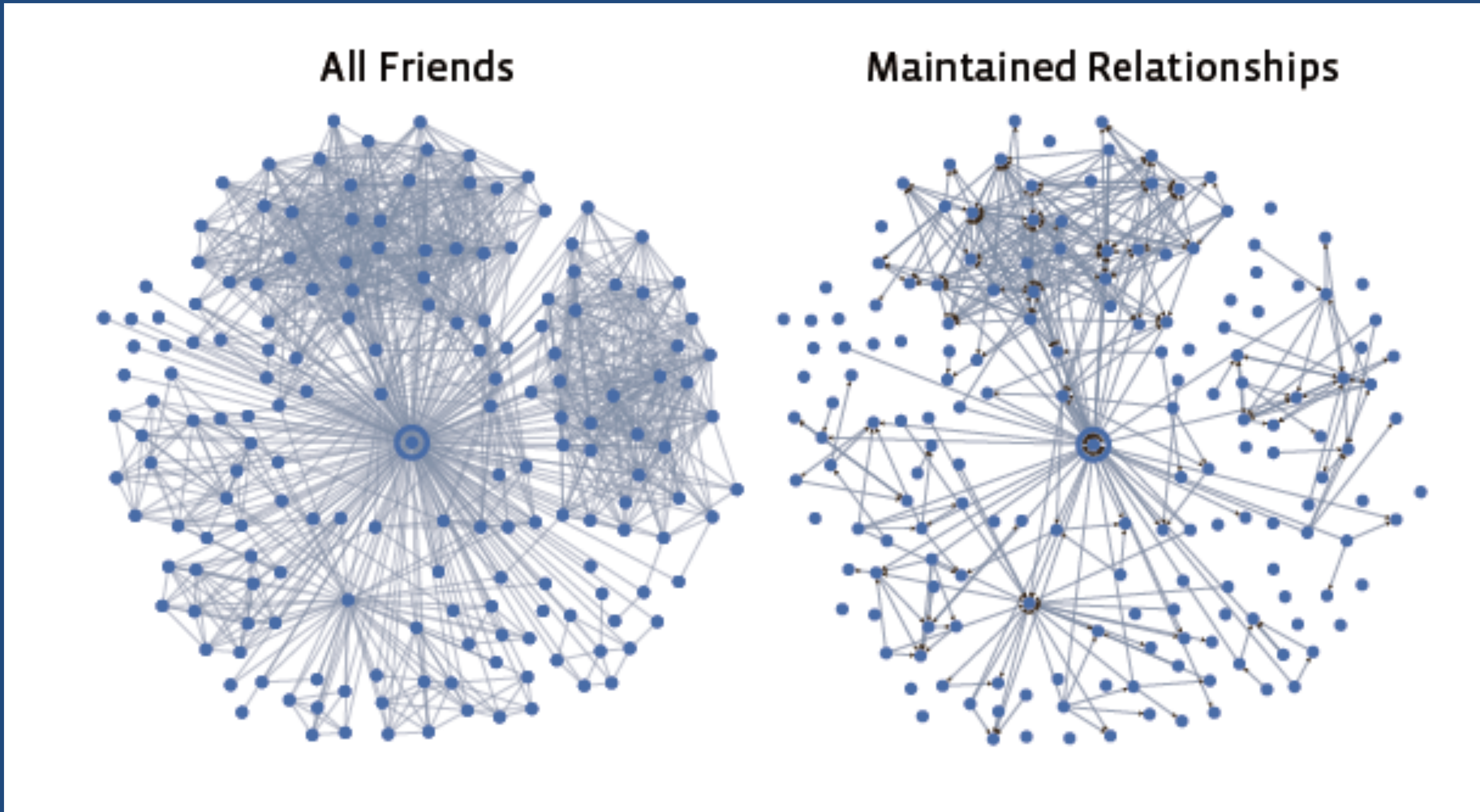
在手机通信网上的数据结果

- 美国全国人口的20%，18周的通信数据
- 节点：手机号
- 边：通话关系
- 关系强度：通话时长

- 横轴表示边的关系强度（由低到高，%）
- 纵轴表示邻里重叠度
- 曲线表明这两个量正相关

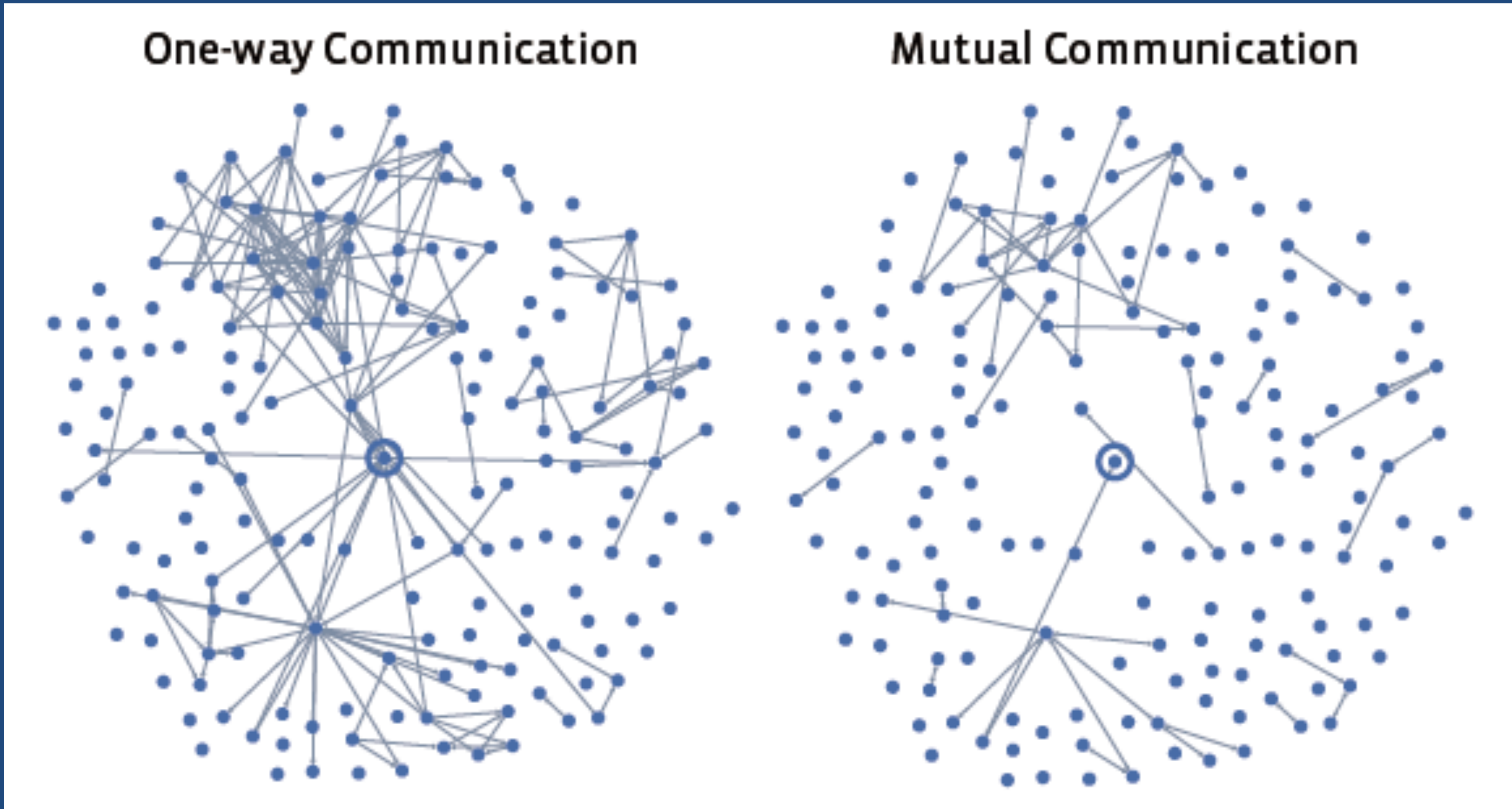
Onnela, Structure and tie strength in mobile communication networks, PNAS 2007

OSN上关系强度的不同体现形式



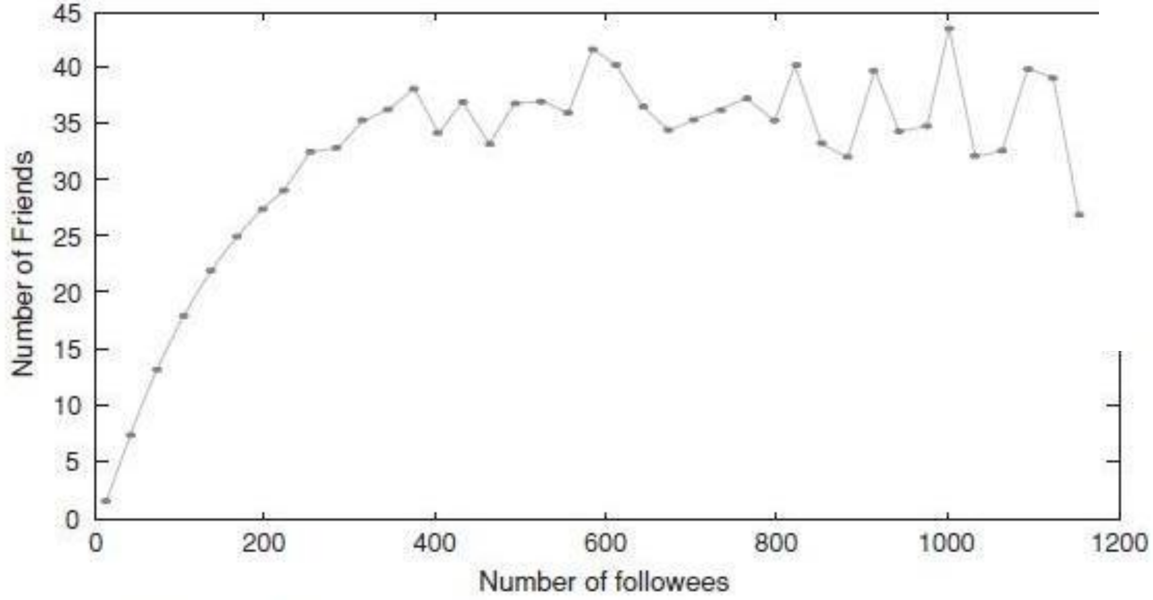
以Facebook为例，图中给出的是一个用户及其“朋友”之间的关系情况。
左图表示有关用户自己给出的“好友”情况，其中许多实际没发生任何通信联系

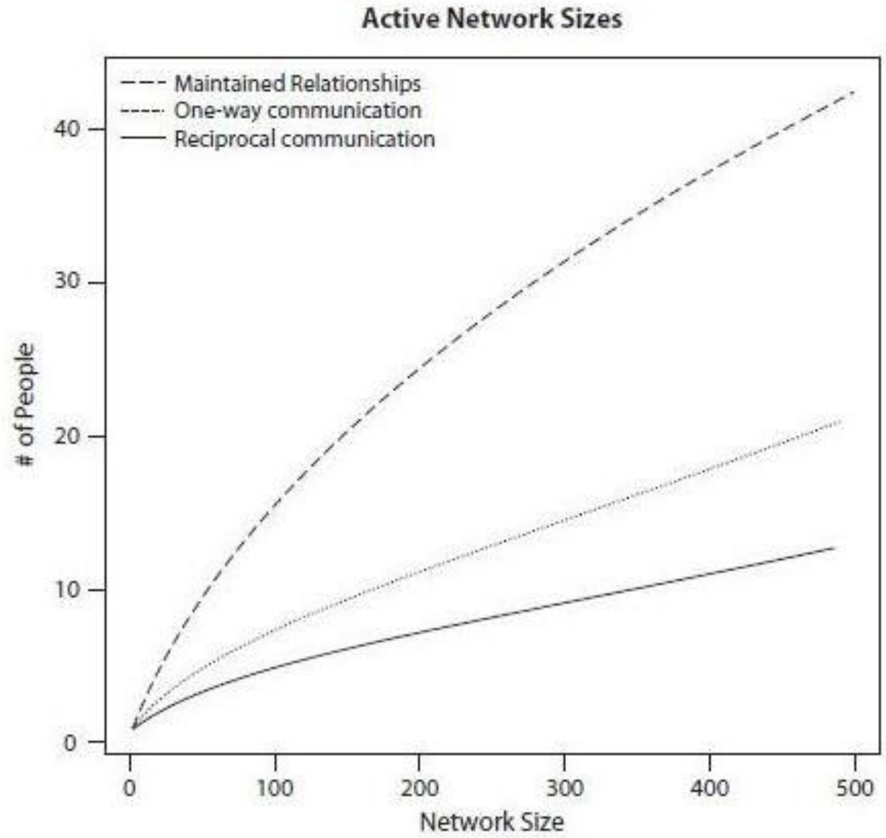
Facebook的关系强度体现形式



按照该项工作研究人员的定义，one－way包含mutual，因此我们看到左图包含右图所有的边。

Twitter上一个用户追随对象的个数与他实际联系的人数之间的关系



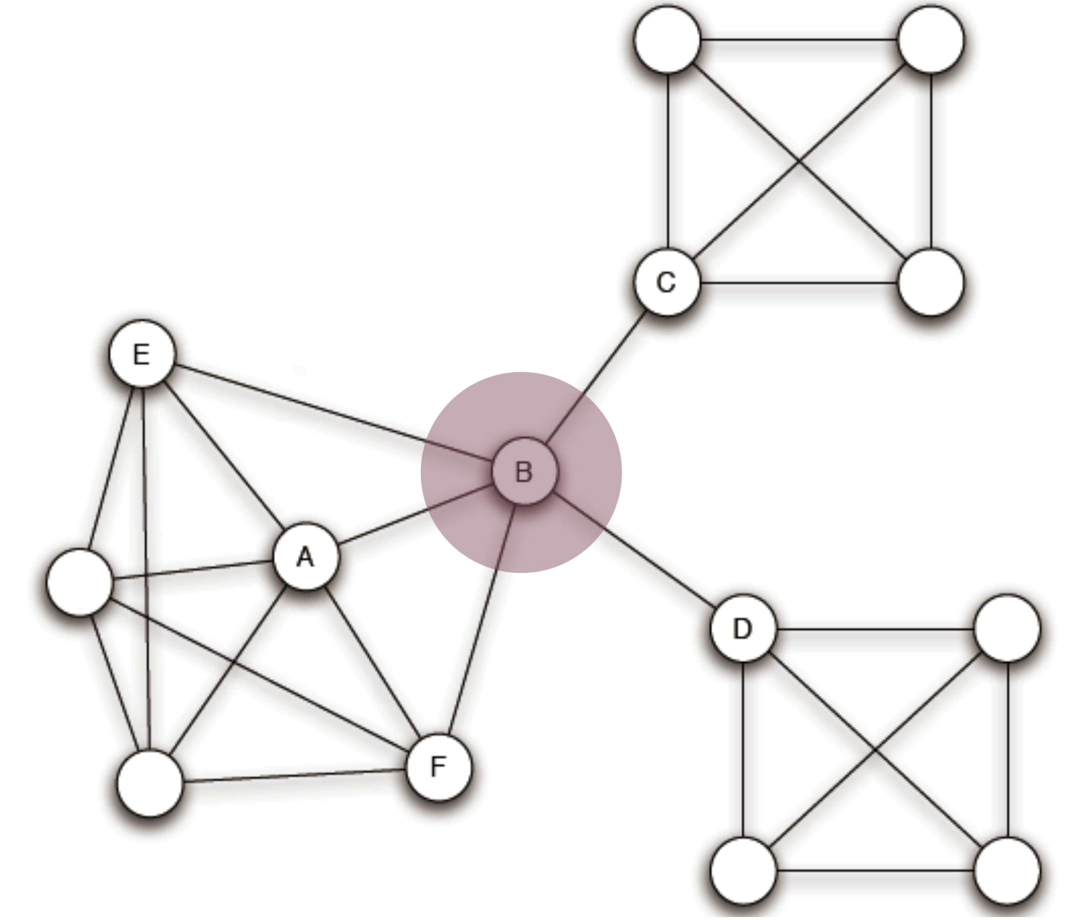


Facebook上一个典型用户好友之间联系的强度情况

从上述可以得到的一个定性结论是：在**OSN**上，尽管一个用户可以声明他关注大量（几百）其他用户，但实际关注的大约在**50**以下，而真正有联系的则更少，在**20**以下。

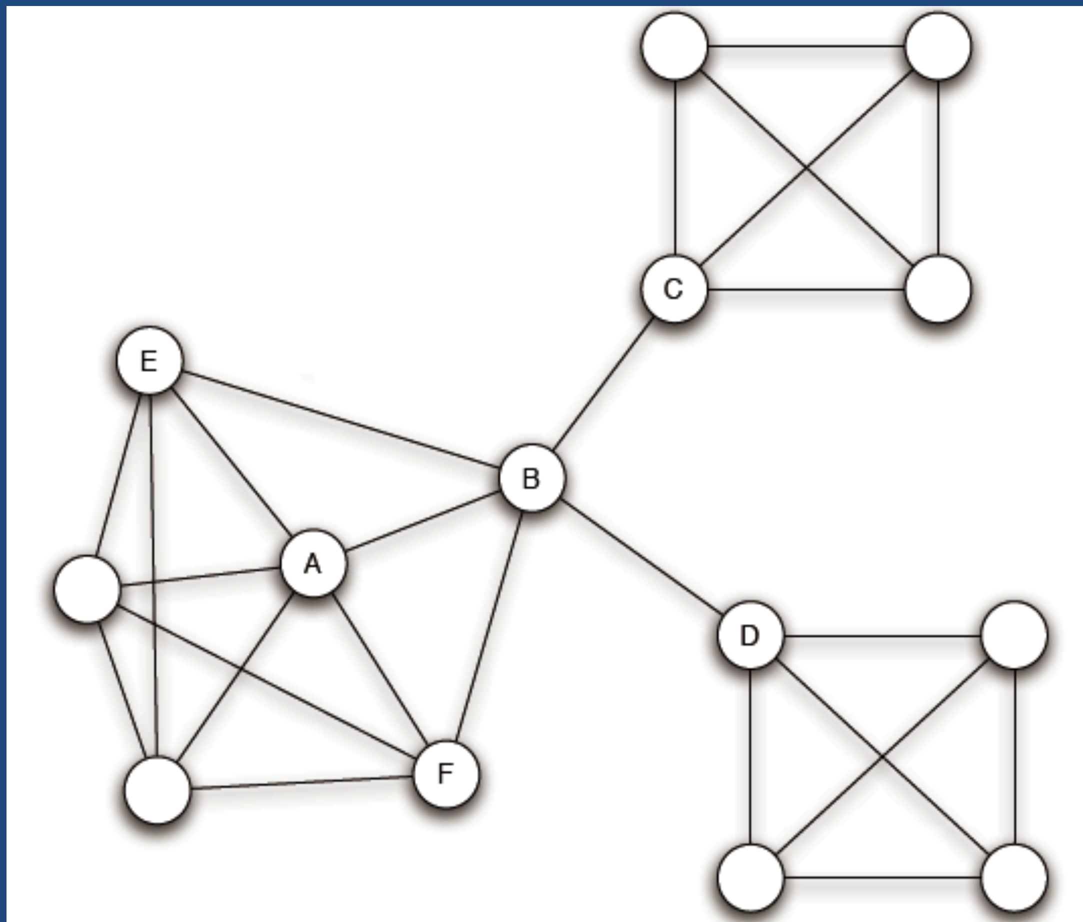
由上述，可得社会网络结构的一个基本意象

- 用桥（或者捷径，或者**邻里重叠度**很低的边，弱关系）连接起来的相对比较密集互连的节点群
  - **边的嵌入性**：两个端点共同邻居的数量
- 其中，那些是多个桥的端点的节点（B）值得特别讨论
  - **聚集系数**较低
  - 她与群组内部的节点（A）相比，有什么利弊？
  - 怎样与她打交道？

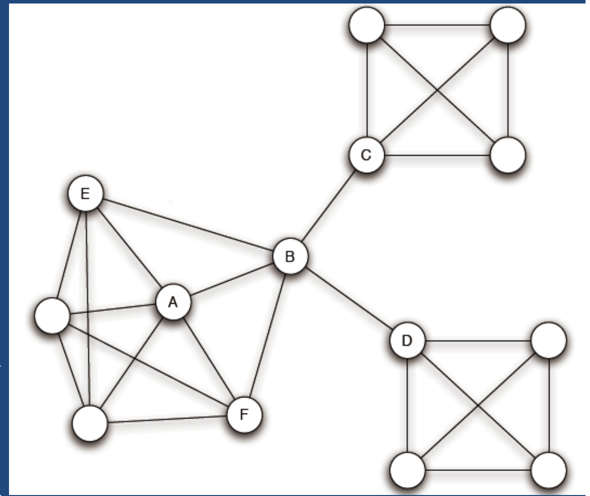


结构洞

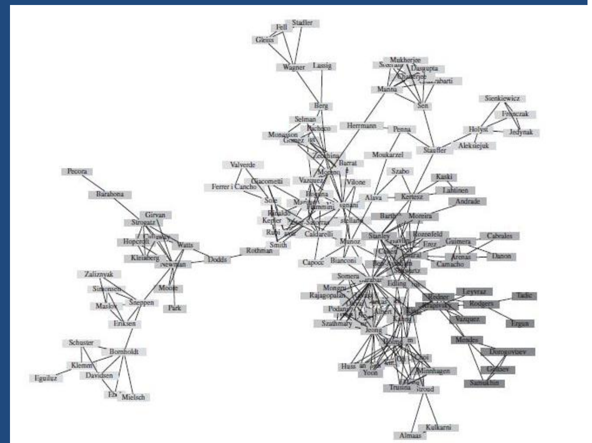
- 结构洞：存在网络中两个或多个没有紧密联系的节点集合之间的“空地”

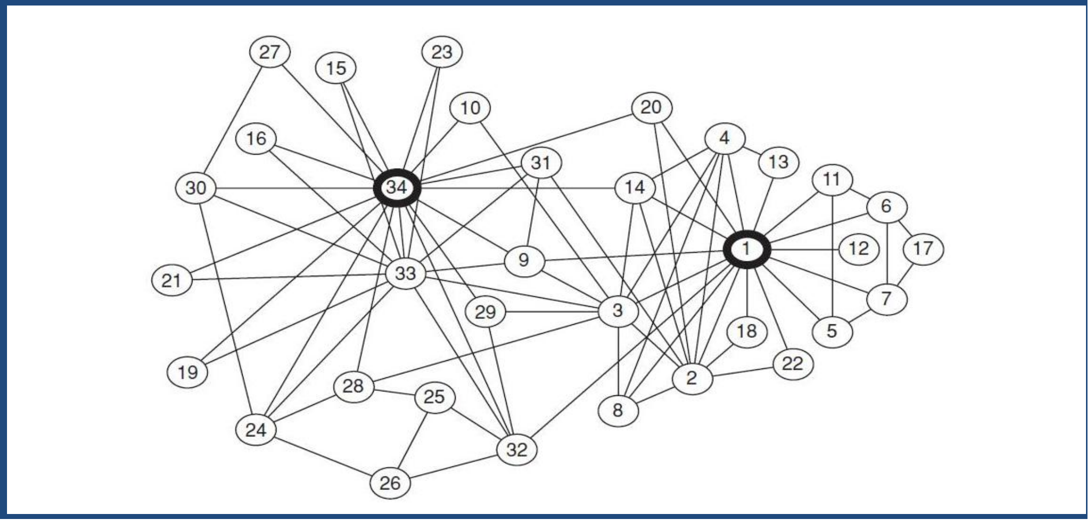


图划分算法

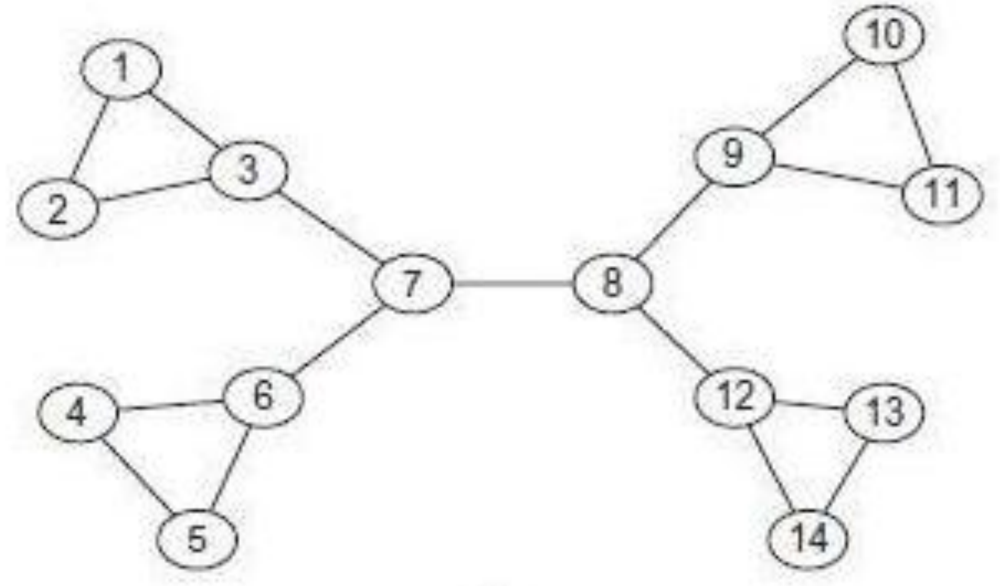


- 如何刻画社会网络中“相互紧密连接的节点群”？能否有一种精确的方法将它们找出来？
  – 分割法
    - 逐步去掉“跨接边”
  – 聚集法
    - “滚雪球”
  – 近似
    - 准确与效率的平衡



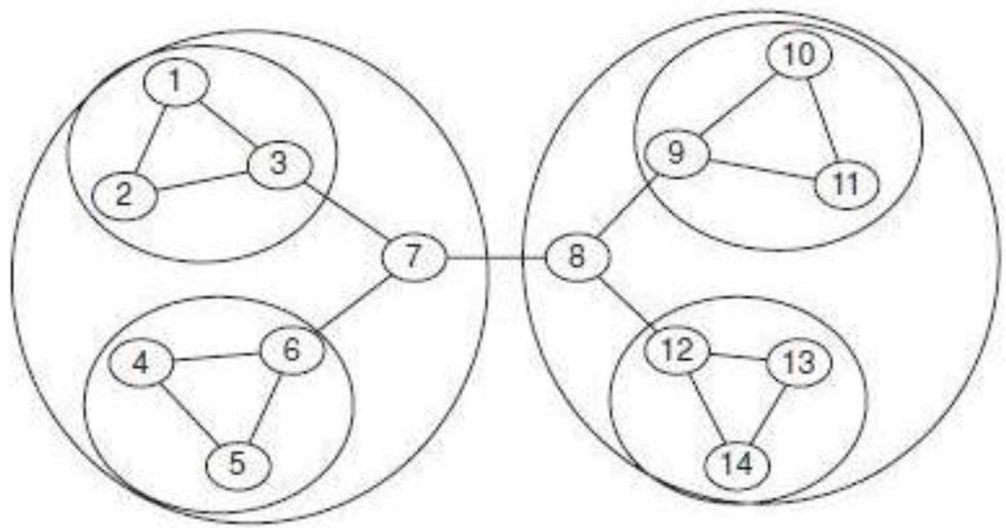


Girvan-Newman方法



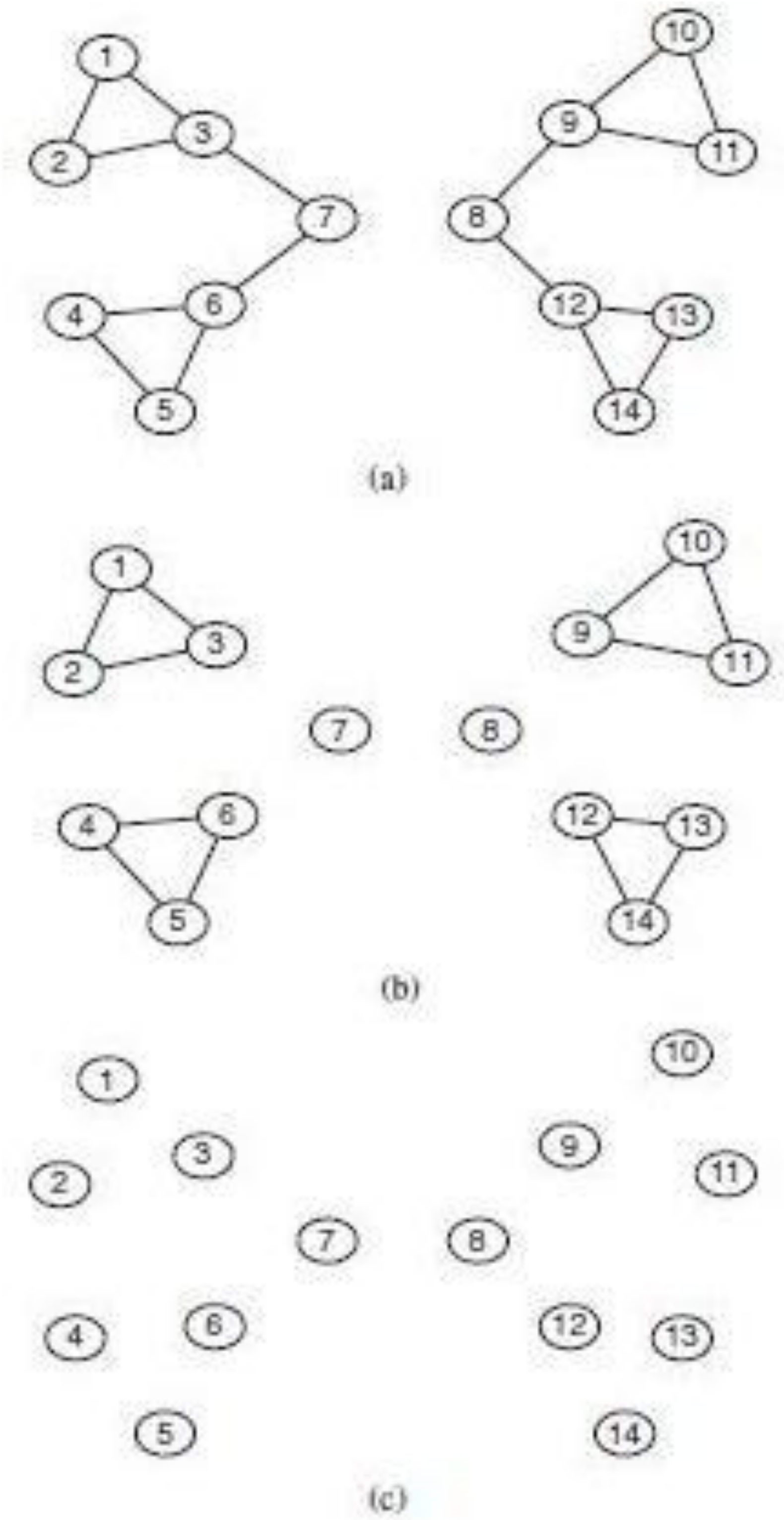
- 最先应该删除哪条边？

- 可以“一层层”进行



如何发现那些最“弱”的边？

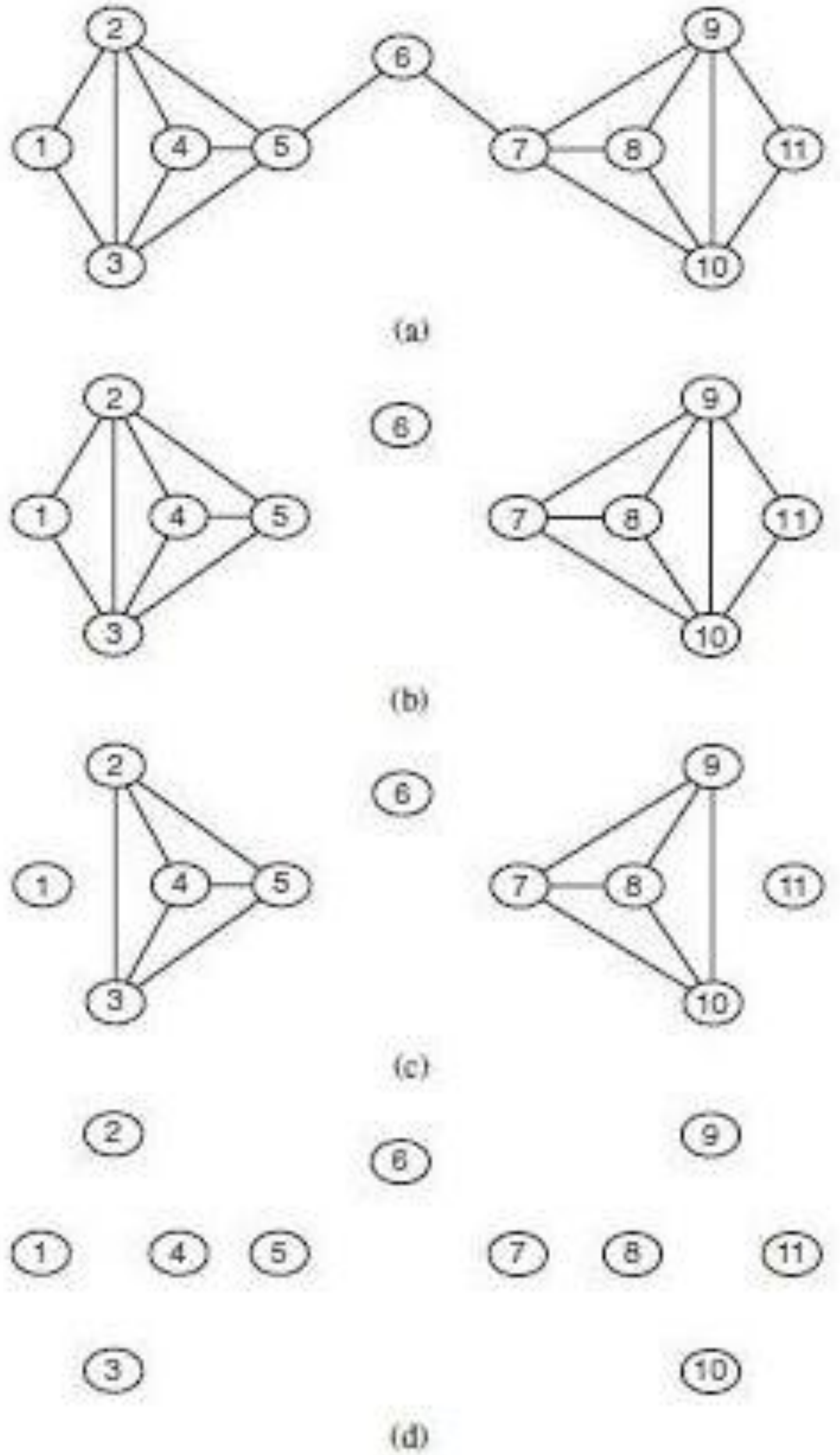
- 或者“最关键”的边：许多节点之间的最短路径都要经过它
- **介数**———条边承载的一种“流量”
  - 两个节点A和B，设想1个单位的流量从A到B，均分到它们之间所有的最短路径上
    - K条路径，则每条路径上分得1/k，
    - 若一条边被m条路径共用，则在它上面流过m/k
  - 所有节点对都考虑后，一条边上的累记流量就是它的介数（betweenness）



逐步删除高介数边：例

- b(7,8) = 49
  - 两边各7个节点，都要经过它，7＊7；7个节点内部则不经过
- b(3,7)=b(6,7)=b(8,9)=b(8,12) = 33
  - 3＊7＋3＊4
- b(1,3)=… = 12
  - 涉及1和3-14等12个节点
- b(1,2)=… b(13,14) = 1
  - 仅涉及1和2两个节点

去掉最高介数边后，重新计算剩下的…



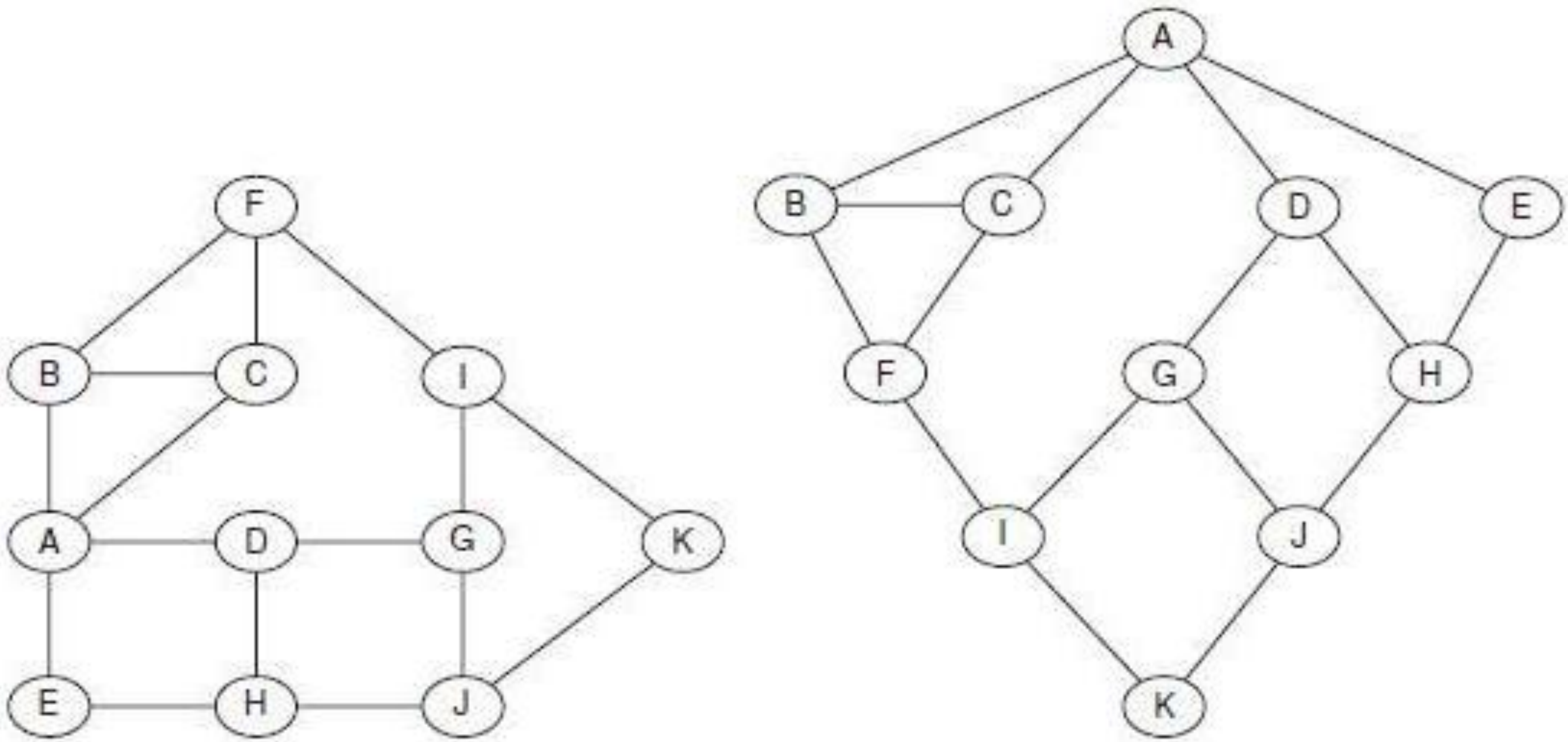
课堂练习

介数计算的一种算法

- 从一个节点（A）开始，做宽度优先搜索，将节点分层（以便于下面的步骤）
- 确定从A到其他每个节点的最短路径的条数
- 确定当从节点A沿最短路径向其他所有节点发送1个单位流量时，经过每条边的流量。

对每一个节点，重复上述过程，累计，除以2，即得每条边的介数。

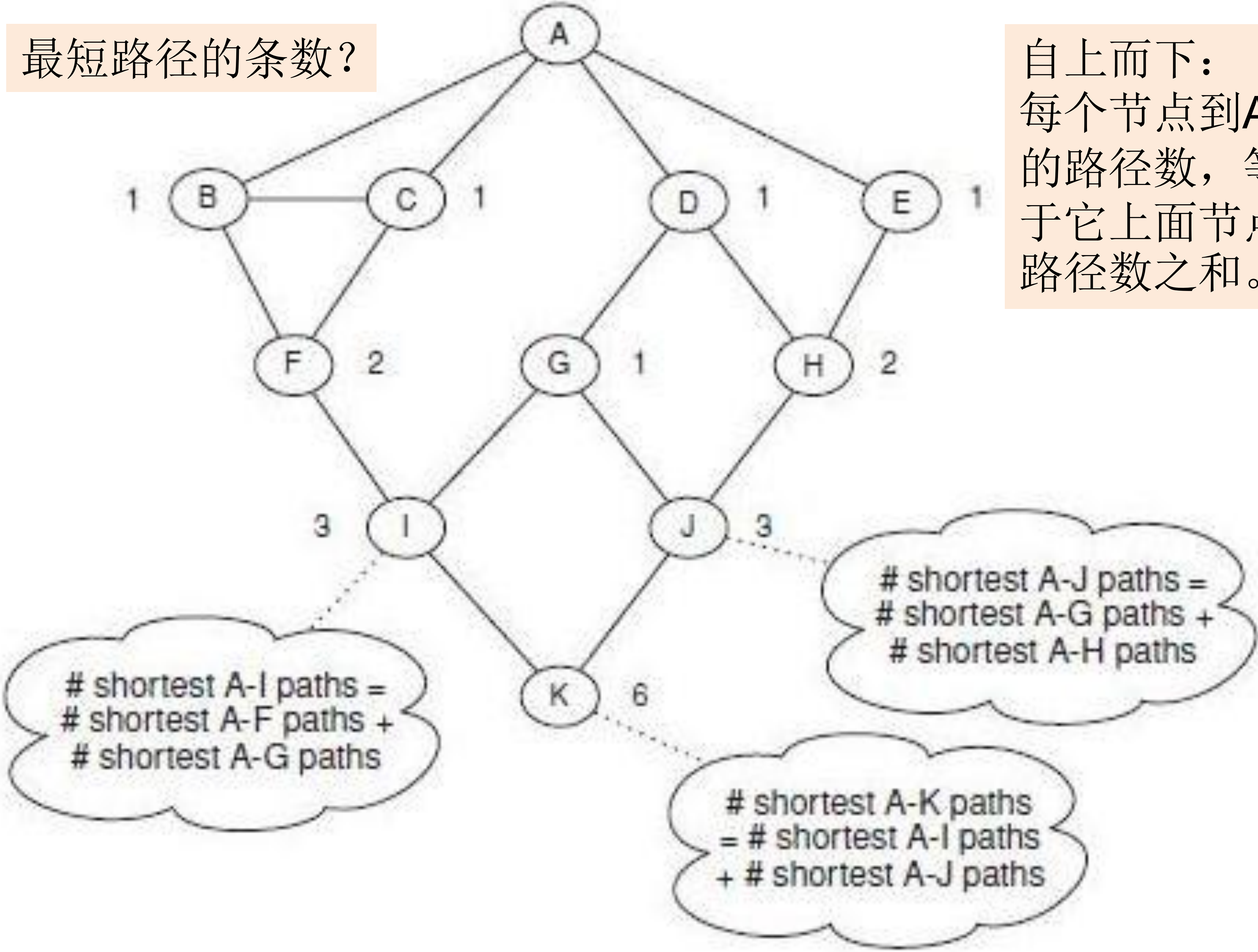
例子：从A开始先宽搜索结果



(a) (b)

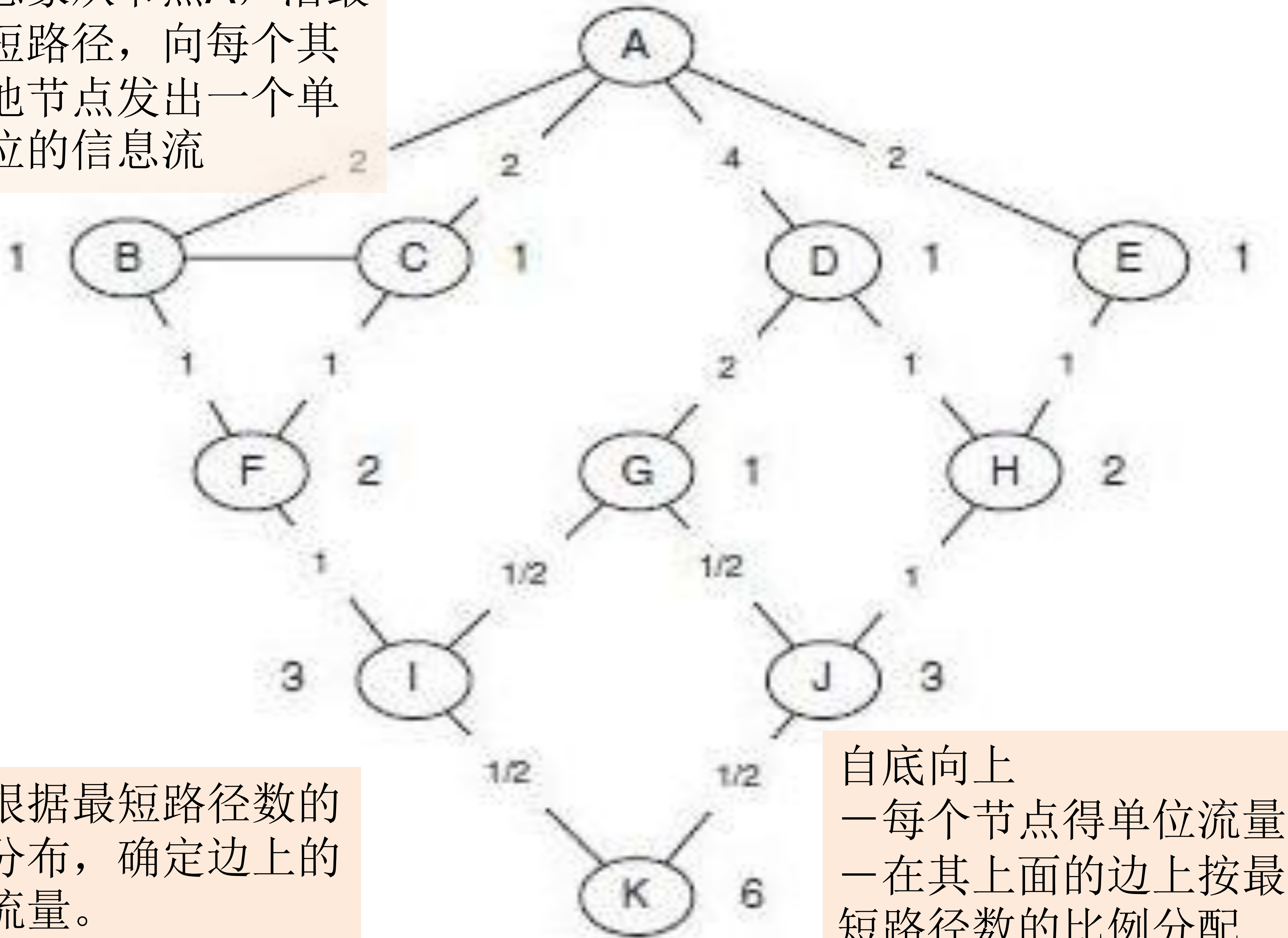
- 从A到K有多少条最短路径？（系统化方法）
- 层次就是最短路径的长度（距离）

最短路径的条数？



自上而下：
每个节点到A的路径数，等于它上面节点路径数之和。

想象从节点A，沿最短路径，向每个其他节点发出一个单位的信息流



根据最短路径数的分布，确定边上的流量。

自底向上
—每个节点得单位流量
—在其上面的边上按最短路径数的比例分配