LEC2 Social Network Analysis

- 社交圈呈同心圆结构：
  - 内层：最亲密的朋友（如家人、挚友）
  - 外层：泛泛之交、同事、网友等
- 越靠近中心的人，越容易被频繁沟通和影响
- **关键观点**：真正有影响力的是那些处于“核心圈”的人

📌 这体现了社交网络中的**结构性不平等**——不是所有人都同等重要。

- 每个人都通过朋友、朋友的朋友……连接到全球
- 即使不认识某人，也可能只隔几层就能联系上

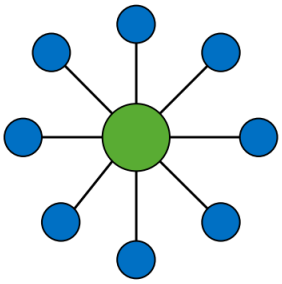
💡 **类比**：“六度分隔理论”（Six Degrees of Separation）的直观体现

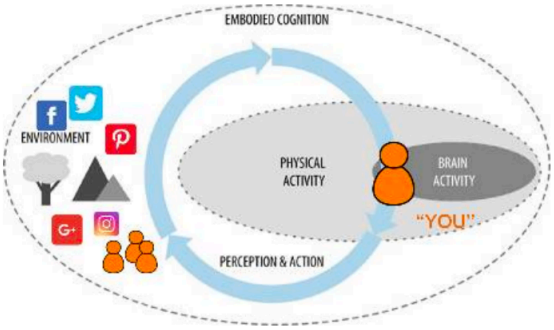
什么是社交网络？

> 在社交网络中，个体（agents 或 actors）相互互动。

🔗 核心元素：

- **节点（Nodes）**：代表实体（如人、组织、品牌）
- **边（Edges）**：代表关系或互动（如友谊、关注、合作）

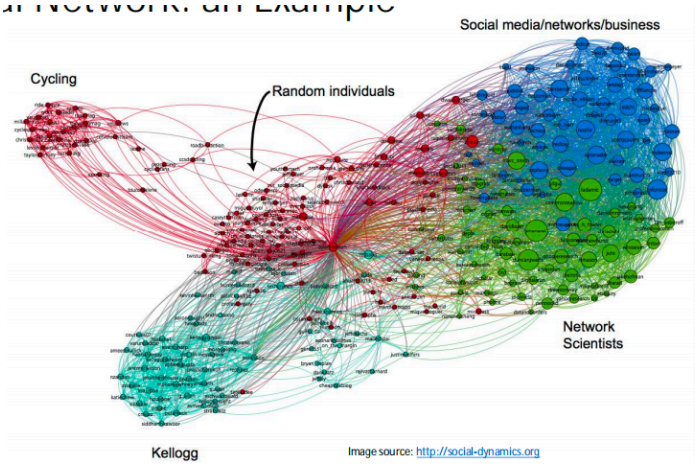




🧩 示例图：

- 中心绿色节点 = 你
- 周围蓝色节点 = 你的朋友
- 边 = 你们之间的连接

📌 补充图示展示了“嵌入式协作”（Embedded Collaboration）模型，强调物理活动与虚拟互动的结合。



- 不同颜色区域表示不同的兴趣群体：
  - 红色：骑行爱好者（Cycling）
  - 蓝色：社交媒体/网络/商业人士（Social media/networks/business）
  - 绿色：网络科学家（Network Scientists）
  - 青色：Kellogg 学院成员
- 中央密集区域为"随机个体"（Random individuals），连接各个子群
- 一条红色箭头指向中央枢纽，显示其作为桥梁的作用

→展现了现实社交网络的**模块化结构**（社区划分）和**跨社区连接**

→某些节点（如中央枢纽）具有高**中介性**（betweenness），是信息流动的关键通道

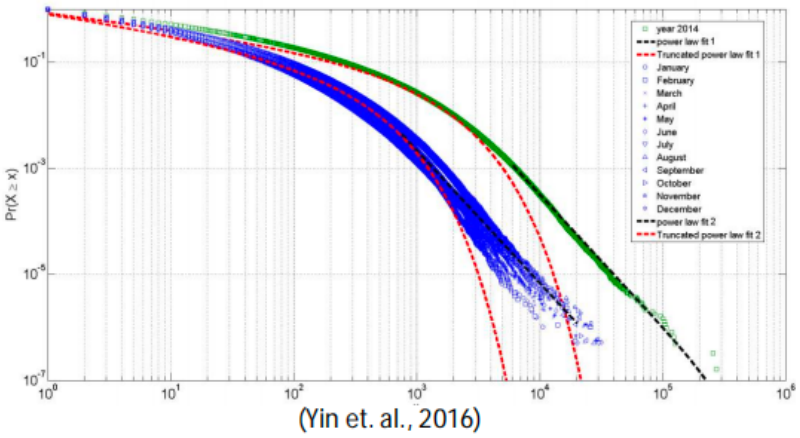
Twitter/X 图的规模与特性

- 截至 2022 年，Twitter/X 图包含超过 **2.29 亿个节点**（即每日活跃用户）
- 其入度（in-degree）和出度（out-degree）均遵循**幂律分布**（Power Law Distribution）

🔍 图表解读：

- X轴：节点度数（log scale）
- Y轴：具有该度数的节点比例（log scale）
- 曲线呈陡峭下降趋势 → 少数节点拥有大量连接，多数节点连接较少

📌 这正是"**无标度网络**"（Scale-Free Network）的典型特征。



(Yin et. al., 2016)

复杂网络的两大共性

✅ 1. 小世界（Small-world）

- 任意两个节点之间仅隔着很短的距离（平均路径长度小）
- 类似于"六度分隔"

✅ 2. 无标度（Scale-free）

- 节点度分布高度异质，服从幂律分布
- 没有典型的尺度（scale），即不存在"典型大小"的连接数
- 几乎总是存在少数"超级连接者"（hubs）

📌 这两种性质共同定义了大多数真实世界的社交网络结构。

多种复杂网络的比较

> 来源：Wang & Chen, 2003

📊 表格对比不同网络的属性：

表格

网络类型	聚类系数	平均路径长度	度指数
Internet (domain level)	0.24	3.56	2.1
WWW	0.1	3.1	2.1
Software	0.06	6.39	2.5
Movie actors	0.79	3.65	2.3

📌 观察发现：

- 所有网络都表现出**低平均路径长度**（小世界）

- 同时具有**高聚类系数**（局部紧密连接）
- 度分布符合幂律（γ ≈ 2–3）

➡️ 证明这些网络都是**小世界 + 无标度**的混合体。

小世界现象（Small Worlds）

全球社交织网（The Global Social Fabric）

> “80亿人口，但你只需6-7跳就能连到任何人！”

🌍 事实：

- 地球上有约 82.3 亿人（UN 2025）
- 每个人只认识地球上极小一部分人
- 但通过短短几步中间人，几乎所有人都能连接起来

📌 **“六度分隔”理论**：任何两个人之间最多相隔六个人。

小世界的社会机制

👫 情境：

- 一个人说：“我知道这个群里每个人都很重要！”
- 另一个人回应：“而且这群里的每个人都通过 Leo 和我认识。”

🔗 结构解释：

- 个体通过家庭、朋友、工作等形成紧密群体（clique）
- 这些群体之间通过某些“枢纽人物”（如 Leo）相连
- 因此，即使看似陌生的两人，也能通过少数中间人建立联系

📌 **关键问题**：两个陌生人之间到底有多远？

小世界网络的图示

> 一个典型的“小世界”网络结构

🌀 特征：

- 大量节点围绕几个中心节点聚集
- 存在许多“跳跃边”（shortcuts），连接原本遥远的节点
- 整体呈现“星型+环形”混合结构

✅ **含义**：

- 尽管每个节点只连接少数邻居，但由于存在少量长距离连接，整个网络仍然非常“扁平”

米尔格拉姆实验（Milgram’s Experiment）

> 1967年，社会心理学家 Stanley Milgram 开展的经典实验

🧪 实验设计：

- 在内布拉斯加州随机选择“起始者”（starter）
- 让他们尝试把一封信寄给波士顿附近的一个指定“目标”（target）人物
- 如果不知道收件人，必须把信交给认为更接近目标的人

📈 结果：

- 大约三分之一的信最终送达目标
- 平均需要 **6 步**（median of six steps）

📌 这就是“六度分隔”理论的实证来源！

第14页：实验地图示意

> 显示从美国中西部到波士顿的信件传递路径

📍 路径示例：

- 起始位置 → 中间城市（如密苏里州） → 新英格兰地区 → 目标城市
- 总距离达 4,305 英里，但仅需 6 步完成传递

💡 **启示**：

- 即使地理距离遥远，社会距离却很近
- 社会网络具有强大的**短路径效应**

小世界网络的特性

> “小世界”网络的特点是：任意两点之间仅需短路径连接

🌐 图解：

- 从起点（Starting Node）到终点（Target Node）只需 6 步
- 尽管网络庞大，但信息可以快速扩散

📌 **应用**：

- 病毒式营销（viral marketing）
- 快速传播谣言或新闻
- 社交媒体上的热点事件爆发

小世界网络模型（Watts-Strogatz 模型）

> 1998 年 Watts 和 Strogatz 提出的经典模型

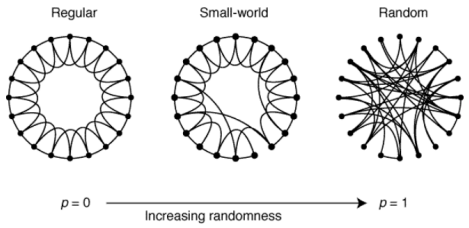
🧩 构建方式：

1. 从一个完全有序的环状晶格开始（Regular Lattice）
    - 每个节点连接 k 个邻居

- 例如 n=20, k=2

2. 以概率 p 随机重连每条边（rewire each edge with probability p）

📈 三种状态：

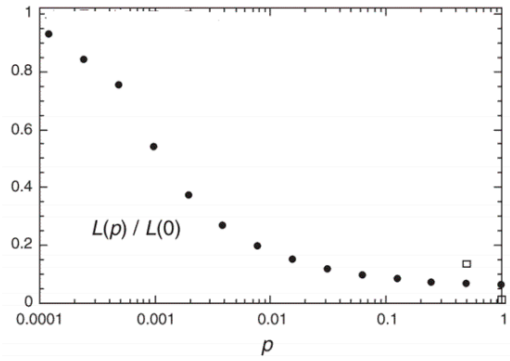


p=0	Regular	完全规则，高聚类，长路径
0<p<1	Small-world	中等聚类，短路径，最佳平衡
p=1	Random	低聚类，短路径，完全随机

📌 **结论**：只需少量随机连接，就能极大缩短平均路径长度，同时保持较高聚类系数。

L(p) / L(0) vs p 曲线

> 描述平均路径长度随随机化程度变化的趋势



📉 图表解读：

- L(p) 是重连概率为 p 时的平均路径长度
- L(0) 是原始规则网络的平均路径长度
- 当 p 很小时，L(p)/L(0) 快速下降 → **非线性效应**
- 当 p 接近 1 时，曲线趋于平稳

💡 **关键发现**：

- 很少的"捷径"（shortcuts）就能显著降低整体距离
- 这解释了为什么现实中人们总能"快速找到"某人

小世界网络的本质

> 网络拓扑结构介于完全规则与完全随机之间

🔁 特点：

- **既规则又随机**
  - 规则：家庭、朋友圈等高度结构化
  - 随机：偶尔结识新朋友（如旅行、会议）
- 这种混合导致“小世界”特性

💡 含义：

- 信息可以在社交网络中迅速传播
- 例如：病毒式营销、社交媒体挑战赛等都能快速扩散

小世界网络的变体

> 一种简化构造方式：不是重连，而是直接添加链接到晶格

🔄 方法：

- 保留原有环状结构
- 添加额外的“捷径”边（shortcuts）

✅ 优点：

- 分析更简单
- 对结果影响不大

📌 图中展示了两种情况：

- (a) k=1：每个节点加一条捷径
- (b) k=3：每个节点加三条捷径

👉 更多捷径 → 更强的“小世界”效应

米尔格拉姆实验中的超级连接者（Hubs）

🧑‍💼 关键人物：

- **Mr. Jacobs**：服装商人，是收件人的朋友
- **Mr. Jones** 和 **Mr. Brown**：其他两位关键中间人

🌐 角色定位：

- Mr. Jacobs 是“枢纽”（hub），连接多个圈子
- 他之所以能成功传递信件，是因为他在社会网络中处于**中心位置**

📌 **结论**：

- 小世界现象依赖于“超级连接者”
- 这些人虽然不多，但在信息传播中起决定性作用

幂律分布（Power Law Distribution）

📌 核心对比：

表格

类型	特征	示例
正态分布（Normal Distribution）	有典型尺度（characteristic scale），呈“钟形曲线”	身高、体重、智商
幂律分布（Power Law Distribution）	无典型尺度，长尾分布（fat-tailed），少数个体占主导	社交网络中的朋友数量

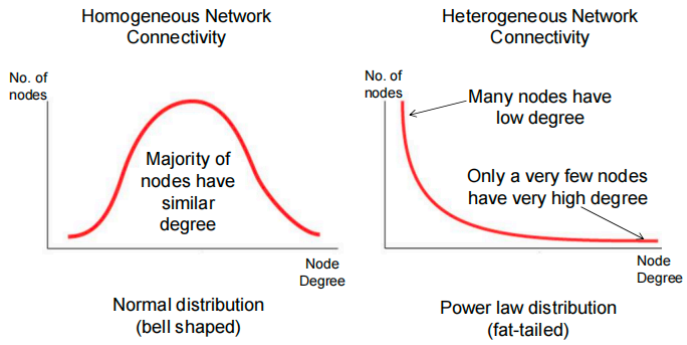
✅ 解释：

- 在现实世界中，许多自然现象遵循正态分布（例如人的身高集中在平均值附近）
- 但在**社交网络**中，节点的连接数（度）往往不遵循这种模式
- 相反，大多数用户只有少量朋友，但极少数人拥有成千上万的朋友（如名人、意见领袖）

📌 **关键点**：

> “没有一个典型的连接数” → 没有“平均值”能代表多数情况

幂律分布 vs 正态分布（图示对比）



📈 左图：同质性网络（Homogeneous Network）

- **正态分布（bell-shaped）**
- 大多数节点具有相似的度（连接数）
- 表现为对称的钟形曲线

📈 右图：异质性网络（Heterogeneous Network）

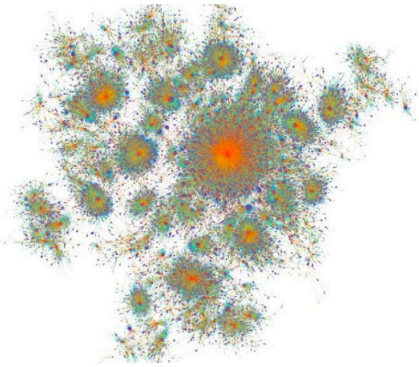
- **幂律分布（fat-tailed）**
- 多数节点度很低（如1～5个连接）
- 极少数节点度非常高（如几千甚至百万级）

- 曲线急剧下降后拖出一条长长的尾巴

✅ **结论**：

- 真实世界的社交网络通常是**异质性的**，即存在“超级连接者”（hubs）
- 这种结构被称为“**无标度网络**”（Scale-Free Network）

无标度网络（Scale-Free Networks）



✅ 定义：

> 如果一个网络的节点度分布遵循幂律，则称为无标度网络

🔍 特性：

- **无特征尺度**：无法用单一数值描述“典型连接数”
  - 因为绝大多数节点连接少，而少数节点连接极多
- **非均匀元素**：存在一些“枢纽节点”（hubs），它们拥有远超平均水平的连接数

🌐 图例说明：

- 图像显示了一个典型的无标度网络结构（来源：https://www.networkpages.nl）
- 中心区域是高度连接的“核心”，周围是大量低连接的小群体
- 呈现出“星状+簇状”的混合结构

📌 **现实意义**：

- 如 Twitter 上的大V、Facebook 上的明星账号
- 这些节点在信息传播中起关键作用

自组织（Self-Organization）

✅ 自组织的本质：

> 网络的异质性和复杂结构不是人为设计的结果，而是由一系列简单规则自动演化而成的过程。

🔍 关键观点：

- 异质性可能来源于某种**规律性、有序的行为**
- 是一种复杂的、自发形成的系统行为（unsupervised process）
- 大型网络的发展受**稳健的自组织现象**驱动，超越个体系统的细节

❓ 为什么重要？

- 理解这些自组织机制是网络科学的核心挑战之一
- 本课程将从跨学科视角探索其背后原理

📌 **类比**：就像蚂蚁群不需要指挥也能建造复杂巢穴，网络也通过局部互动形成全局结构。

自组织与幂律分布的机制（Barabási & Albert, 1999）

✅ Barabási 和 Albert 提出了两个通用机制来解释为何网络会呈现幂律分布：

1. 持续增长（Growth）

- 网络不断扩展，新节点不断加入
- 例如：互联网新增网站、社交媒体新增用户

2. 优先连接（Preferential Attachment）

- 新节点更倾向于连接到那些**已经很受欢迎的节点**
- 即："富者愈富"（the rich get richer）效应

📊 数学表达：

> 节点 v 与其他 k 个节点相连的概率 $P(k)$ 随 k的增加而衰减，遵循幂律：

$$P(k) \sim k^{-\gamma}$$

其中：

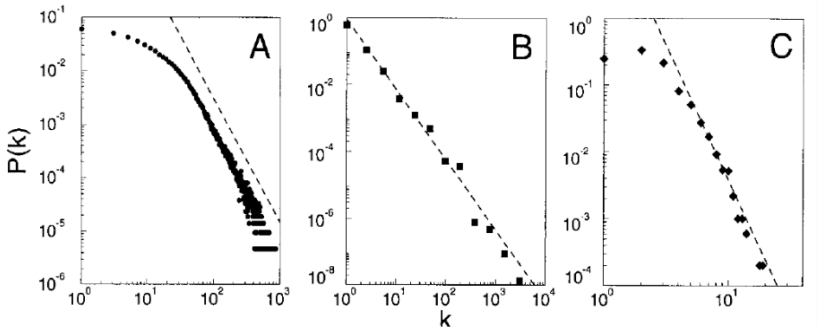
- k ：节点的度（连接数）
- γ：幂律指数（通常在 2～3 之间）

📌 **含义**：

- 度越高的节点，越容易吸引新连接
- 导致少数节点迅速积累大量连接，形成"超级枢纽"

真实世界网络中的幂律分布实例

📈 图表展示三种大型网络的度分布（Barabási & Albert, 1999）



(Barabasi & Albert, 1999) The distribution function of connectivities for various large networks. (**A**) Actor collaboration graph with N = 212,250 vertices and average connectivity $\langle k \rangle$ = 28.78. (**B**) WWW, N = 325,729, $\langle k \rangle$ = 5.46 (*6*). (**C**) Power grid data, N = 4941, $\langle k \rangle$ = 2.67. The dashed lines have slopes

(A) $\gamma_{actor} = 2.3$, (B) $\gamma_{www} = 2.1$ and (C) $\gamma_{power} = 4$.

A. 演员合作网络（Actor Collaboration Graph）

- 节点数：212,250
- 平均连接数： = 28.78
- 幂律指数 γ ≈ 2.3

B. 万维网（WWW）

- 节点数：325,729
- 平均连接数： = 5.46
- 幂律指数 γ ≈ 2.1

C. 电网（Power Grid）

- 节点数：4941
- 平均连接数： = 2.67
- 幂律指数 γ ≈ 4.0

📉 分析：

- 所有三个网络都表现出明显的幂律趋势（数据点近似落在直线附近）
- 使用双对数坐标图（log-log plot），幂律表现为直线，斜率即为 γ
- 越接近直线，说明越符合幂律分布

📌 **结论**：

> 不同领域的复杂网络——从人际关系到技术基础设施——都展现出类似的幂律特性，说明这是一种普遍存在的组织原则

→真实世界的复杂网络之所以如此强大且高效，是因为它们不是随机生成的，而是通过"持续增长 + 优先连接"这两个简单机制自我组织而成的。

→"少数人连接一切，大多数人只连一小部分。" —— 这就是幂律的力量。

问题	回答
什么是幂律分布？	一种长尾分布，少数节点拥有极高连接数，多数节点连接很少；数学形式为
为什么节点度服从幂律？	因为真实网络通过持续增长 + 优先连接这两个简单规则自组织演化而成
核心机制是什么？	优先连接（Preferential Attachment）：新节点偏好连接“已有的热门节点”，导致“强者恒强”
谁提出的理论？	Barabási 和 Albert（1999），提出“无标度网络”（Scale-Free Network）概念
现实意义？	解释了为何社交网络中存在“超级连接者”（hubs），对信息传播、病毒营销、网络安全等至关重要

社交网络分析的核心目标

🎯 核心观点：

> 社交网络分析（SNA）的一个关键目标是识别社交网络中最重要的人物（actors）。

✅ 如何判断一个节点的重要性？

- **直接沟通能力**：能直接连接多少其他节点 → **度中心性（Degree Centrality）**
  - 例如：一个人有多少朋友或关注者
- **接近性**：是否能快速到达网络中的大多数节点 → **接近中心性（Closeness Centrality）**
  - 例如：你是否能在几步内联系到任何人
- **桥梁作用**：是否在不同群体之间起着不可或缺的中介作用 → **中介中心性（Betweenness Centrality）**
  - 例如：你是两个朋友圈之间的“联络人”

→节点的重要性不仅仅取决于它有多少连接，还取决于它的位置和功能。 某些“枢纽”（hubs）可能不是最活跃的，但却是信息流动的关键通道。

从社交媒体数据中提取底层社交网络

🌐 研究背景：

> Machado 等人（2019）利用在线社交网络（OSN）数据，在纽约市开展城市感知（urban sensing）研究。

💡 方法论：

- 使用 OSN 数据创建“**复杂虚拟传感器**”（complex virtual sensors）
- 目标：捕捉、表示和分析城市活动数据
- 优势：成本远低于部署物理传感器（如摄像头、GPS追踪器）

📊 应用价值：

- 实时监测人流分布
- 分析城市热点区域
- 支持交通规划、公共安全等决策

📌 **本质**：将社交媒体用户的行为视为一种“数字足迹”，通过这些足迹重构城市的动态图景。

Twitter/X 的数据来源

📍 数据获取方式：

- 从 Twitter/X 平台收集公开内容样本
- 条件：用户的元数据（metadata）包含发布时的**精确地理位置**（地理坐标）
- 工具：使用 **Twitter Stream API** 获取实时推文流

🧩 示例：

```
[推文内容] Sausage & ricotta or vegetable lasagna...
[地点标签] @LICKlocal ice cream super delicious
[坐标] Latitude: 40.7128, Longitude: -74.0060
```

✅ **关键点**：

- 推文自带地理标签（geotagged），可以直接用于定位
- 这些数据可用于构建用户移动轨迹、热点事件检测等

Instagram 的数据来源

📍 数据获取方式：

- Instagram 用户发布的照片或帖子通常会提及**场所名称**（如餐厅、景点、学校）
- 这些文本可以用来**索引用户的地理位置**

📷 示例：

- 用户上传一张照片并配文："I had the best coffee at Newcastle University!"
- 系统自动识别"Newcastle University"作为地点
- 查找该地点的地理坐标（经度/纬度）

🧩 补充信息：

- 收集的数据增加了上下文信息（contextual information）
- 包括用户注册的位置、常去的地点等
- 可以用于构建个人行为模式、兴趣偏好模型

Facebook 的数据来源

📍 数据获取方式：

- 使用 Facebook Graph API 获取地点信息
- 输入参数包括经纬度范围（center）、查询类型（place）、字段（fields）等

💻 示例代码片段（PHP）：

```
$facebook_data_array=array(
'base_url'=>'https://graph.facebook.com/v2.9/',
```

```
'node'=>'search?',
'query_param'=>'q=',
'root_node'=>'type=place',
'coords'=>'center=-35.7796,-78.6382',
'fields'=>'name,talking_about_count',
'access_token'=>'FB_API_KEY'
);
```

✅ 功能说明：

- 查询某个地理位置附近的场所（如咖啡馆、博物馆）
- 获取场所名称、讨论热度（talking_about_count）等信息
- 用于补充和验证来自 Instagram 或 Twitter 的地点数据

开放数据源 —— NYC OpenData

🏙️ 数据来源：

- 纽约市城市规划局（NYC Department of City Planning）公开发布的数据集
- 具体项目：**Open Data Week 2026**（象征性标志）

📊 数据内容：

- 研究人员调查了 **350 万条移动样本**（mobility samples）
- 来自 **25.6 万名用户**，时间跨度为 **14 个月**
- 数据公开访问地址：https://opendata.cityofnewyork.us/

📌 **意义**：

- 提供官方、权威的城市级人口流动数据
- 与社交媒体数据结合，可进行更准确的城市行为建模

研究人员整合的两大数据源

📚 数据融合策略：

研究人员综合了两类数据源：

1. Mobility Dataset（移动数据集）

- 来源：Twitter 和 Instagram 上带有地理坐标的用户发布内容
- 内容：用户在观察窗口内的地理坐标轨迹
- 用途：反映个体的时空移动行为

2. Venues Dataset（场所数据集）

- 来源：综合 Instagram 的元数据 + Facebook 的地理数据库
- 内容：语义化定义的场所集合（如“中央公园”、“时代广场”）
- 用途：将原始坐标映射为有意义的地点名称，提升数据分析的可解释性

🔗 整合逻辑：

- 移动数据提供“谁在哪里” → 坐标序列
- 场所数据提供“哪里是什么” → 语义标签
- 合并后形成“人在哪些地方活动”的完整画像

步骤	内容	目的
1️⃣	从 Twitter/X、Instagram、Facebook 收集带地理位置的社交媒体数据	获取用户的实时位置信息
2️⃣	利用平台 API 或文本解析提取地理坐标	构建用户移动轨迹
3️⃣	结合 NYC OpenData 等政府开放数据	验证并丰富数据质量
4️⃣	构建“移动数据集”和“场所数据集”	将原始坐标转化为语义化的城市活动图谱
5️⃣	进行社交网络分析	识别关键人物、社区结构、信息传播路径

🎯 最终目标：实现“城市感知”（Urban Sensing）

> 利用社交媒体作为“虚拟传感器”，低成本、高效率地监测城市动态，替代传统物理传感器。

🌍 应用场景：

- 交通拥堵预测
- 公共事件响应（如抗议、火灾）
- 商业选址分析
- 城市规划优化

→**社交媒体不仅是交流工具，更是强大的社会数据源**。

通过对这些数据的挖掘与分析，我们可以揭示人类行为的模式，理解社会结构的运行机制，并为智慧城市提供科学支持。

🔹 第一页：感知社会方面（Sensing Social Aspects）

🎯 核心问题：

> 如何在物理环境中监测与社会背景相关的变量？例如：

- 用户之间的接近程度（proximity）
- 他们是否发生相遇或互动？

✅ 传统方法的局限性：

- 使用蓝牙（Bluetooth）或 Wi-Fi 扫描设备来检测用户附近的人
- 需要安装专用硬件或移动应用 → **对大多数用户来说不友好、侵入性强**

✅ 当前研究的方法：

- 利用**在线社交网络（OSN）数据**（如 Twitter/X、Instagram）

- 结合**时空窗口模型（spatiotemporal window model）**
- 通过用户的地理位置轨迹（trace-based analysis）推断他们的**偶然相遇事件**（opportunistic encounters）

📌 **优势**：

- 无需额外硬件
- 基于用户自愿分享的数据
- 可大规模部署

时空窗口模型（Spatiotemporal Window Model）

📌 定义一个数据样本：

每个来自社交媒体的地理标记样本被表示为一个三元组：

$s = (u, p, t)$

其中：

- u：用户（user），属于用户集合 U
- p：位置，由纬度和经度定义（地理坐标）
- t：时间戳（timestamp），即用户到达该位置的时间

🧩 模型含义：

> 每个样本代表一个“事件”，这个事件受限于：

- **时间窗口**（temporal window）：何时发生的？
- **空间区域**（spatial area）：发生在哪个地点？

✅ 这种方式将离散的社交媒体发布行为转化为连续的空间-时间轨迹。

⏱️ 暂停时间（Pause-time）t_{pi}

- 表示用户 u_i 在某个位置 p 上停留的时间段
- 注意：原始数据中没有直接记录停留时间，需通过相邻样本推断
- 例如：如果用户 A 在 10:00 和 10:30 都出现在同一个地点，则推测其停留了 30 分钟

📏 距离阈值（Distance Threshold）D_{th}

- 设定为 **100 米**
- 若两个用户的地理位置距离 ≤ 100 米，则认为他们在物理上足够接近，可能发生了互动

📌 **关键点**：

> 这些参数是人为设定的“规则”，用于从原始数据中识别潜在的社会接触。

🔹 第四页：偶然相遇事件（Opportunistic Encounter Event）

🤝 定义一次“偶然相遇”：

当两个数据样本 s_i 和 s_j 满足以下三个条件时，就认为发生了相遇：

1. **不同用户**：$u_i \neq u_j$
2. **地理接近**：$d_g \leq D_{th}$
   （d_g 是两点间的地理距离，这里设为 100 米）
3. **时间重叠**：

$$\max(t_i, t_j) < \min(t_i + t_{pi}, t_j + t_{pj})$$

✅ 举个例子：

- Alice 在 10:00～10:30 出现在中央公园
- Bob 在 10:15～10:45 出现在同一地点
- 两者距离 < 100 米→ 判断为一次“偶然相遇”

不依赖用户主动交互（如点赞、评论），而是基于他们在现实世界中的共现（co-location）来推断社会联系。

时间邻近图（Temporal Proximity Graph）

📊 构建图结构：

研究人员构建了一个**时间邻近图** $G = (V, E)$，其中：

- V：在时间 t_s 时刻在线的所有用户集合
- E：表示用户之间因地理接近而产生的“相遇边”

🕰️ 时间序列分析：

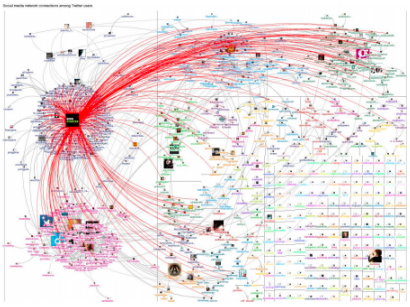
- 图不是静态的，而是随时间变化的
- 定义时间序列：
  $T_g = \{g_1, g_2, ..., g_n\}$
  其中每个 g_i 是图 G 在第 i 个时间点的快照（snapshot）

📌 **用途**：

- 分析社交网络随时间演变的动态特性
- 发现社区形成、信息传播路径等模式

社会网络可视化（Sociogram）

🖼️ Sociogram 示例：



- 展示了一个包含 **1000 个节点** 和 **1908 条边** 的 Twitter/X 用户社交网络
- 使用工具：NodeXL（一款社交网络可视化软件）
- 红色线条表示连接关系
- 中心区域聚集了高连接度的用户（hubs）

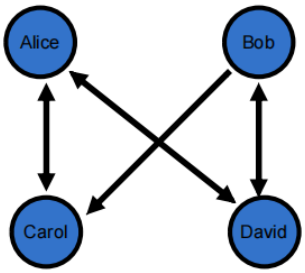
📌 **特点**：

- 显示了网络的结构特征（如簇状结构、核心-边缘结构）
- 可以直观看到哪些用户处于中心地位

简单社交网络（A Simple Social Network）

🧍‍♀️ 场景设定：

考虑一个由四人组成的社交网络：Alice、Bob、Carol、David



🔄 社会关系如下：

- Alice 喜欢 Carol 和 David
- Bob 喜欢 Carol 和 David
- Carol 喜欢 Alice
- David 喜欢 Alice 和 Bob

📌 注意：这是**单向喜欢关系**，不是互惠的！

简单社交网络的图形表示（Sociogram）

🖼️ 图形化展示：

- 四个节点分别代表四个人
- 有向箭头表示“喜欢”的方向
- 例如：Alice → Carol 表示 Alice 喜欢 Carol

✅ 特征：

- 所有边都是**有向的**（directed）
- 存在交叉关系（如 Alice 和 Bob 都喜欢 David）

社会网络的数学表示（Sociomatrix）

🧮 社会矩阵（Sociomatrix）或邻接矩阵（Adjacency Matrix）X

	Alice	Bob	Carol	David
Alice	–	0	1	1
Bob	0	–	1	1
Carol	1	0	–	0
David	1	1	0	–

📌 解释：

- 每行代表一个“出发者”
- 每列代表一个“接收者”
- 如果 $x_{ij} = 1$，表示 i 喜欢 j
- 对角线为空（无自我喜欢）

✅ 数学表达：

$$X = \begin{bmatrix} - & 0 & 1 & 1 \\ 0 & - & 1 & 1 \\ 1 & 0 & - & 0 \\ 1 & 1 & 0 & - \end{bmatrix}$$

📌 **优点**：

- 可以用矩阵运算进行复杂分析（如中心性、聚类）
- 便于计算机处理和建模

练习题（Exercise）

🧩 给出一个 6×6 的矩阵：

	A	B	C	D	E	F
A	–	1	1	0	0	0
B	0	–	1	1	0	0
C	1	1	–	0	0	0

	A	B	C	D	E	F
D	0	0	0	–	1	1
E	0	0	0	0	–	1
F	0	0	0	1	1	–

❓ 问题：

1. 你能画出对应的 sociogram 吗？
2. 这是一个**有向图**还是**无向图**？

✅ 分析：

- 由于矩阵不对称（如 A→B=1，但 B→A=0），说明关系是非对称的
- 因此这是一个 **有向图（Directed Graph）**

👉 你可以尝试自己画出来：A 指向 B 和 C；B 指向 C 和 D；等等。

图的基本概念（Graphs）

📚 图的定义：

一个图 $\mathcal{G}$ 包含两部分信息：

- 节点集合：$N = \{n_1, n_2, ..., n_g\}$
- 边集合：$L = \{l_1, l_2, ..., l_L\}$

🧩 关键术语：

- **节点（node）**：代表个体（actor）
- **边（line/edge）**：代表两者之间的关系（tie）
- **相邻（adjacent）**：若存在边 $l_k = (n_i, n_j)$，则 n_i 和 n_j 相邻

图的类型

🔄 图可以分为两类：

类型	描述	示例
无向图（Undirected Graph）	边没有方向，关系是对称的	Facebook：你加我好友 ⇔ 我加你好友
有向图（Directed Graph / Digraph）	边有方向，关系可能是单向的	X/Twitter：你关注我 ≠ 我关注你

🖼️ 图例对比：

- 左图：无向图（所有边无箭头）
- 右图：有向图（所有边带箭头）

✅ 总结：整套流程的逻辑链条

1. 数据采集

- 从 Twitter/X、Instagram 等平台获取带有地理位置和时间戳的用户数据
- 形成三元组 (u, p, t)

2. 事件识别

- 使用**时空窗口模型**判断用户是否在同一时间、同一地点出现
- 设置距离阈值（100米）、暂停时间等参数
- 识别“偶然相遇事件”

3. 网络构建

- 构建**时间邻近图** $G = (V, E)$
- 每个时间点生成一个快照图
- 分析网络动态演化

4. 网络表示

- 图形化表示（sociogram）→ 直观理解结构
- 数学表示（sociomatrix）→ 支持量化分析

5. 图的性质

- 区分有向 vs 无向图
- 应用于不同类型的社会关系建模

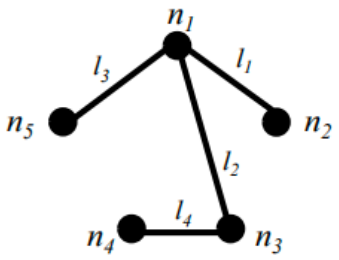
> 我们可以通过分析人们在现实世界中的移动轨迹（即使只是社交媒体上的“打卡”），来重建他们之间的社会联系，从而揭示隐藏的社会结构。

这不仅是一种技术手段，更是一种**新型的社会感知方式**——利用数字足迹反推真实世界的互动。

> “你在地图上的每一次定位，都在默默绘制一张看不见的社会网络。”

这组幻灯片系统地介绍了**社交网络分析（Social Network Analysis, SNA）** 的核心概念，包括图的结构、密度计算以及三大中心性指标：**度中心性（Degree Centrality）、接近中心性（Closeness Centrality）和中介中心性（Betweenness Centrality）**。以下是每页内容的详细解释与逻辑梳理。

无向图示例（Undirected Graph）



📌 图结构：

- 节点集合：$N = \{n_1, n_2, n_3, n_4, n_5\}$
- 边集合：$L = \{l_1, l_2, l_3, l_4\}$

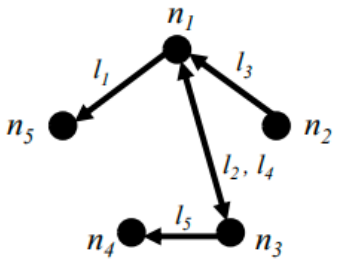
🧩 边的定义：

边	连接
l_1	(n_1, n_2)
l_2	(n_1, n_3)
l_3	(n_1, n_5)
l_4	(n_3, n_4)

✅ 特点：

- 所有边无方向
- 关系是对称的（如 n_1 和 n_2 相连 ⇔ n_2 和 n_1 相连）
- 例如：Facebook 是典型的无向图（好友关系是互惠的）

有向图示例（Directed Graph）



📌 图结构：

- 节点集合：$N = \{n_1, n_2, n_3, n_4, n_5\}$
- 边集合：$L = \{l_1, l_2, l_3, l_4, l_5\}$

🧩 边的定义（带方向）：

边	方向
l_1	(n_1, n_5) → 从 $ n_1 $ 指向 $ n_5 $
l_2	(n_1, n_3)
l_3	(n_2, n_1)
l_4	(n_3, n_1)
l_5	(n_3, n_4)

✅ 特点：

- 边有明确方向
- 关系可能是单向的（如 Twitter 上“你关注我”≠“我关注你”）
- 例如：X/Twitter 或微博是典型的有向图

🔹 第三页：无向图的密度（Density of Undirected Graph）

📐 定义：

> 图的密度（Density）是指图中实际存在的边占所有可能边的比例。

🔢 计算公式：

- 可能的边数（组合问题）：
  $\binom{|N|}{2} = \frac{|N|(|N|-1)}{2}$
  因为任意两个节点之间最多只能有一条边（无重复）
- 实际边数：$|L|$
- 密度公式：
  $D = \frac{2|L|}{|N|(|N|-1)}$

✅ 示例：

若 $|N| = 5$，则最大可能边数为 $\frac{5\times4}{2} = 10$

若有 6 条边，则密度 $D = \frac{2\times6}{5\times4} = 0.6$

📌 **意义**：密度越高，网络越“紧密”，信息传播越快。

🔹 第四页：有向图的密度（Density of Directed Graph）

📐 定义：

> 有向图的密度同样是实际边数占所有可能边的比例。

🔢 计算公式：

- 可能的边数（排列问题）：
  $|N|(|N|-1)$
  因为每个节点对可以有两个方向的边（如 A→B 和 B→A）
- 实际边数：$|L|$
- 密度公式：
  $D = \frac{|L|}{|N|(|N|-1)}$

✅ 示例：

若 $|N| = 5$，则最大可能边数为 $5 \times 4 = 20$

若有 8 条边，则密度 $D = \frac{8}{20} = 0.4$

📌 **注意**：有向图的密度通常低于无向图，因为允许更多连接方式。

社交网络分析简介（Social Network Analysis）

🌐 定义：

> 社交网络是由具有能动性的社会行动者（如人、组织）构成的网络。

📊 分析目标：

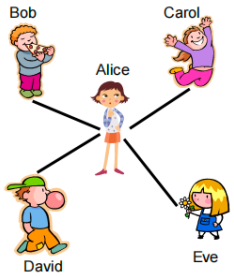
- 使用**量化指标**评估或比较不同社交网络
- 探索谁是关键人物？谁在控制信息流？

🔍 主要指标：

1. **度中心性**（Degree Centrality）
2. **中介中心性**（Betweenness Centrality）
3. **接近中心性**（Closeness Centrality）

📌 配图展示了一个真实的社交活动场景："SOCIAL ACTORS UNITE"，象征社会行动者的互动。

星形网络中的中心性（Centrality in Star Network）



🌟 示例：星形网络（Star-shaped network）

- 中心节点：Alice
- 外围节点：Bob、Carol、David、Eve

✅ 中心节点的独特优势：

1. **最大度（Maximum Degree）**
   - Alice 与其他所有人相连 → 度最高
2. **位于最短路径上（Geodesic Path）**
   - 所有外围节点之间的通信都必须经过 Alice
3. **最小距离（Minimal Distance）**
   - Alice 到其他任何人的距离都是 1 → 最接近所有人

📌 **结论**：

> 在这种结构中，Alice 是绝对的"核心"——她拥有最高的影响力、控制力和可达性。

度中心性（Degree Centrality）

📌 定义：

> 一个节点的度（Degree）是与其直接相连的边的数量。

✅ 无向图示例：

- A 的度 = 2（连接 B 和 C）
- C 的度 = 3（连接 A、B、D）

✅ 有向图中的扩展：

- **入度（In-degree）**：指向该节点的边数 → 被多少人关注/喜欢
- **出度（Out-degree）**：从该节点发出的边数 → 关注了多少人

示例表格：

节点	In-degree	Out-degree
A	1	1
B	2	0
C	0	3

📌 **现实意义**：

- 在 Twitter 上，高 in-degree 表示受欢迎（粉丝多）
- 高 out-degree 表示活跃（关注很多人）

度中心性的数学表达

分析目标	推荐使用的度中心性
谁最受欢迎 / 最有影响力？	入度（In-degree）
谁最活跃 / 最外向？	出度（Out-degree）
谁连接最多人（不区分方向）？	总度（In + Out）（较少用）
网络是无向的？	直接用度（Degree）

🧮 公式：

对于邻接矩阵 X，其中 x_{ij} 表示节点 i 到 j 是否有连接：

$C_D(n_i) = d(n_i) = \sum_j x_{ij}$

即：节点 n_i 的度 = 其所在行的所有元素之和（对无向图也可用列和），算入度

✅ 归一化度中心性（Normalized Degree Centrality）：

为了消除网络大小的影响，进行归一化：

$C'_D(n_i) = \frac{d(n_i)}{g-1}$

其中 g 是总节点数。

📌 **范围**：0 到 1，便于跨网络比较。

接近性（Closeness）

📍 定义：

> 最短路径（Shortest Path）或 测地线（Geodesic）：两个节点之间最少边数的路径。

📏 距离：

- 两点间的距离 = 最短路径的长度
- 例如：A 到 E 的距离 = 4（路径 A-B-C-D-E）

📌 **意义**：

- 距离越短，表示联系越紧密
- 用于衡量一个人能否快速到达其他人

接近中心性（Closeness Centrality）

🎯 定义：

> 接近中心性反映一个节点到网络中所有其他节点的平均距离的倒数。

🧮 公式：

$\text{Closeness}(n) = \frac{1}{\sum_k d(n,n_k)}$

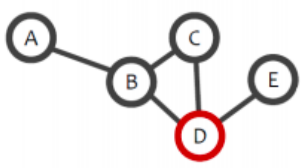
其中 $d(n, n_k)$ 是节点 n 到其他节点 n_k 的最短距离。

✅ 归一化版本：

$C'_c(n_i) = C_c(n_i) \times (g-1)$

👉 使值域在 [0,1] 内，方便比较。

接近中心性计算示例



📌 步骤 1：找出节点 D 到其他节点的距离

- $d(D, A) = 2$（路径 D→B→A）
- $d(D, B) = 1$
- $d(D, C) = 1$
- $d(D, E) = 1$

📌 步骤 2：计算原始接近中心性

$C_c(n_D) = [\sum d(n_D, n_j)]^{-1} = \frac{1}{2+1+1+1} = \frac{1}{5}$

📌 步骤 3：归一化

- 网络有 5 个节点 → 归一化因子为 $g - 1 = 4$
- $C'_c(n_D) = \frac{1}{5} \times 4 = \frac{4}{5} = 0.8$

📌 **解读**：D 是一个非常"接近"整个网络的节点。

中介中心性（Betweenness Centrality）

🎯 定义：

> 中介中心性衡量一个节点在其他节点之间的最短路径中出现频率。

🧮 公式：

$C_B(n_i) = \sum_{i \neq j \neq k} \frac{g_{jk}(n_i)}{g_{jk}}$

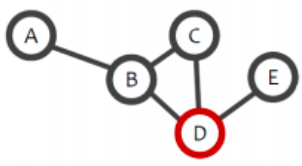
其中：

- g_{jk}：节点 j 到 k 的最短路径总数
- $g_{jk}(n_i)$：这些路径中有多少条经过节点 n_i

📌 **含义**：节点越常出现在他人之间的沟通路径上，其中介性越高。

中介中心性计算示例（Step 1）

📌 目标：计算节点 D 的中介中心性



Node Pair (j,k)	Shortest Path(s)	Total Paths $ g_{jk} $	Paths through $ n_D $	Ratio $ g_{jk}(n_D)/g_{jk} $
(A,B)	A-B	1	0	0/1 = 0
(A,C)	A-B-C	1	0	0/1 = 0
(A,E)	A-B-D-E	1	1	1/1 = 1
(B,C)	B-C	1	0	0/1 = 0
(B,E)	B-D-E	1	1	1/1 = 1
(C,E)	C-D-E	1	1	1/1 = 1

📌 **结果**：D 出现在 3 条最短路径中

中介中心性计算示例（Step 2 & 3）

📌 步骤 2：计算原始中介中心性

$$C_B(n_D) = 0 + 0 + 1 + 0 + 1 + 1 = 3$$

📌 步骤 3：归一化

- 对于 $g = 5$ 个节点，最大可能中介值为：
  $\frac{(g-1)(g-2)}{2} = \frac{4 \times 3}{2} = 6$
- 归一化中介中心性：
  $C'_B(n_D) = \frac{3}{6} = 0.5$

📌 **解读**：D 是一个重要的“桥梁”节点，但不是唯一的关键节点。

✅ 三大中心性对比

中心性类型	含义	测量重点	适用场景
度中心性	连接数量	“我认识多少人？”	发现热门用户、意见领袖
接近中心性	平均距离	“我能多快接触到所有人？”	快速响应、信息扩散
中介中心性	控制路径	“我是否在别人之间的必经之路上？”	控制权、信息瓶颈