Lec1 Introduction to Social Networking

社交网络概述

- 社交媒体与物理技术基础设施结合，形成了人与人之间互动的新方式。
- 这些互动产生了大量用户生成内容（User Generated Content, UGC），是“大数据”（big data）的重要来源。

社交媒体与社交网络

• 社交媒体提供了一个平台，让我们可以理解人们的社交行为、意见、偏好以及个体、社会和文化差异。

社交媒体 vs 传统媒体

类型	特点
Traditional media（如电视、书籍）	- 向大众传播内容- 不支持用户创建或分享内容（无UGC）
Social media	- 支持用户之间的在线互动- 将单向沟通（one-to-many）转变为多向交流（many-to-many）

核心特征：**用户参与性** 和 **双向互动性**。

社交网络示例

实时通知全球数百万粉丝（如Instagram、Threads、X/Twitter）

- Instagram帖子
- 手机查看社交媒体
- 数据可视化地图

→社交网络具有**即时性**和**全球化覆盖**的特点

个性化推荐系统

- 用户可获得基于“相似人群”购买或浏览习惯的个性化推荐。

- 示例：亚马逊图书推荐系统中，“经常一起购买”的商品会被推荐给用户。

→ 展示了社交网络如何利用**用户行为数据**进行精准营销和服务优化。

协作创作

- 以知乎为例，展示“Talk”页面中的协作写作功能。
- 用户可以在平台上共同撰写文章，实现知识共享与合作。

→强调社交网络不仅用于交流，还可促进**集体智慧**和**内容共创**。

社交网络已成为日常生活的一部分

- 截至2025年2月：
  - 全球活跃社交媒体用户达 **5.56亿**（占全球人口67.9%）
  - 平均每日使用时间为 **2小时19分钟**

→社交媒体已深度融入日常行为

▼ 数据分布

社交媒体使用统计概览（2025年2月）

关键数据：

- 活跃用户总数：**5.24亿**
- 年增长率：+4.1%
- 平均每日使用时间：**2小时21分钟**
- 使用者性别分布：女性 45.4%，男性 54.6%
- 年龄分布：18岁以上占86.6%，18岁以下占13.4%

社交媒体用户数量排名（2025年2月）

按广告受众人数排序：

1. YouTube：25.38亿
2. Facebook：23.8亿
3. Instagram：17.1亿

4. TikTok：15.98亿
5. LinkedIn：12.18亿
6. Messenger：9.67亿
7. Snapchat：7.07亿
8. Reddit：6.66亿
9. X（原Twitter）：5.96亿
10. WhatsApp：3.40亿

反映出YouTube和Facebook仍是最大平台，但TikTok增长迅速。

社交媒体平台偏好（按年龄分）

分为两栏：

- **全体互联网用户**中最受欢迎的平台
- **男性互联网用户**中最受欢迎的平台

例如：

- 16–24岁群体最常用：TikTok、Instagram、X
- 35岁以上群体更倾向：WhatsApp、Facebook、WeChat
- 男性用户在某些平台（如Telegram、Discord）上比例更高

社交媒体用户占比（全球地图）

- 显示各国/地区社交媒体用户占总人口的比例（截至2024年4月）
- 高比例区域：
  - 北欧国家（如挪威、瑞典）接近90%
  - 西欧、北美约70%-80%
  - 东南亚、中东部分地区也较高（如泰国、沙特）
- 较低区域：
  - 中东部分国家（如伊朗）、非洲内陆地区较低（如尼日利亚、刚果）

表明社交媒体普及程度存在显著地域差异。

什么是社交媒体？

给出两个权威定义：

1. **Cambridge Dictionary**：“允许人们在互联网上交流和分享信息的网站和计算机程序。”
2. **Investopedia.com**：“通过建立虚拟网络和社区来促进思想、观点和信息共享的基于计算机的技术。”

这两个定义都突出了“**连接性**”、“**信息共享**”和“**虚拟社区**”三大核心要素。

社交媒体是否有法律定义？

指出目前没有统一的法律定义，但不同国家有功能性规定：

- **欧盟 - Digital Services Act (2022)**：
  - 定义“Online Platform”为：向公众存储并传播信息的服务。
- **美国 - Section 230 Communications Decency Act**：
  - 定义“Interactive Computer Service”为：允许多个用户访问计算机服务器的信息服务。

说明社交媒体在法律层面被视为一种**信息服务基础设施**，受特定法规监管。

社交媒体分析——研究社交网络的关键

→ **数据分析**是理解和研究社交网络行为的核心工具，可用于洞察用户行为、情感倾向、影响力传播等。

社交媒体分析的技术定义

> 社交媒体分析（SMA）是对跨社交平台生成的数据进行系统性收集、处理和解释的过程。

- 使用计算、统计和机器学习技术，从用户互动、内容和网络关系中提取**模式、趋势和洞察**。

- 系统性（systematic）
- 数据驱动
- 多种分析方法融合（计算 + 统计 + 机器学习）

SMA 的核心组件

1. 数据源（Data Sources）

- 包括：帖子（posts）、评论（comments）、点赞（likes）、分享（shares）、话题标签（hashtags）、多媒体（图片/视频）、元数据（metadata）
- 这些是原始输入，来自各种社交平台如 Twitter/X、Instagram、Facebook 等。

2. 分析技术（Techniques）

- 自然语言处理（NLP）：理解文本含义
- 情感分析（Sentiment Analysis）：判断情绪倾向（正面/负面/中立）
- 网络分析（Network Analysis）：研究用户之间的连接关系
- 图像/视频分析：识别视觉内容中的信息
- 预测建模（Predictive Modeling）：预测未来趋势或行为

3. 输出结果（Outputs）

- 描述性指标（Descriptive metrics）：如影响力、参与度（reach, engagement）
- 诊断性洞察（Diagnostic insights）：解释“为什么事件发生？”
- 预测性/规范性智能（Predictive/Prescriptive intelligence）：预测“可能发生什么？”或建议“最优行动是什么？”

🔹 **逻辑链条：数据 → 技术处理 → 洞察与决策支持**

SMA 的技术范围 —— 大数据处理框架

阶段	关键任务
1. 数据采集（Data Collection）	准确、完整、实时地获取数据

阶段	关键任务
2. 数据预处理（Data Preprocessing）	清洗、转换、去重、归一化（例如统一时间格式、去除垃圾信息）
3. 数据存储（Data Storage）	支持可扩展性、冗余备份、高效检索（如使用 Hadoop、Spark 或云数据库）
4. 数据分析（Data Analysis）	执行描述性、预测性和规范性分析
5. 数据可视化（Data Visualization）	提供清晰、交互性强的图表，便于理解和沟通

✅ **意义**：SMA 是一个端到端的数据工程过程，不仅限于算法，还需完整的数据基础设施支撑。

SMA 的技术范围 —— AI/ML 算法应用

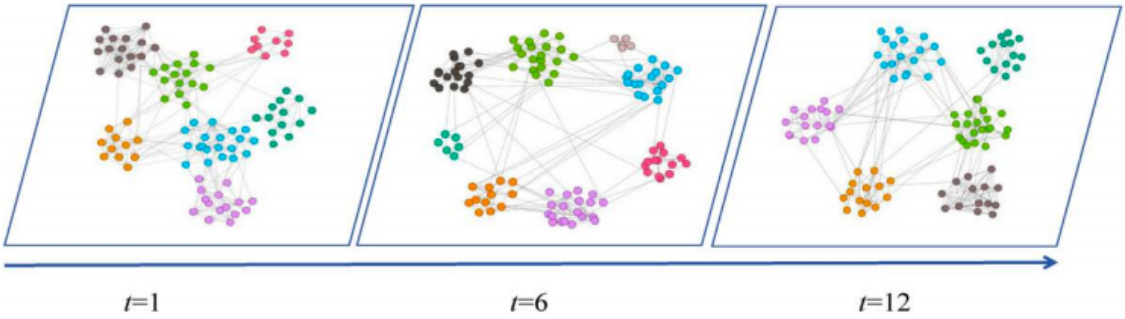
→ 指出 SMA 广泛采用人工智能和机器学习算法进行分类、聚类和推荐

类型	具体算法
无监督学习（Unsupervised Learning）	K-Means 聚类、主成分分析（PCA）、Apriori 算法
监督学习（Supervised Learning）	朴素贝叶斯、支持向量机（SVM）、线性回归、逻辑回归、决策树、随机森林、K-最近邻（KNN）
强化学习（Reinforcement Learning）	AdaBoost、LightGBM、长短期记忆网络（LSTM）
半监督学习（Semi-Supervised Learning）	人工神经网络（ANN）

- 用户分群（聚类）
- 情感分类（监督学习）
- 推荐系统（协同过滤、深度学习）

✅ **重点**：SMA 是高度技术化的领域，依赖现代 ML 框架实现自动化分析。

SMA 的技术范围 —— 图网络模型的应用



- **t=1**：初始状态，节点稀疏，连接较少
- **t=6**：出现小团体聚集（社区形成）
- **t=12**：形成多个紧密相连的子群，部分节点成为中心（枢纽节点）

💡 **含义：**

- 社交网络不是静态的，而是动态演化的
- 可通过图分析识别：
  - 社区结构（communities）
  - 影响力节点（influencers）
  - 信息传播路径（viral spread）

🔧 技术工具包括：

- 社会网络分析（SNA）
- PageRank、Betweenness Centrality 等指标
- 动态网络建模

✅ **应用场景**：品牌营销中寻找意见领袖；危机管理中追踪谣言扩散路径。

SMA 的技术范围 —— 决策支持价值

四个主要应用场景：

1. **B2B 决策支持**

- 通过社交媒体影响采购决策（如企业客户参考同行评价）

2. **销售支持**
   - 利用社交渠道提升销售额（如直播带货、社群运营）
3. **战略性商业资产**
   - 将社交媒体视为一种**可利用的宝贵资源**，用于构建品牌形象、客户关系
4. **数据驱动洞察**
   - 使用数据做出更优决策（如产品改进、市场定位）

🎯 总结：SMA 不仅是技术手段，更是**战略工具**，能为企业、政府、教育机构等提供关键洞见。

社交媒体分析 —— 实践流程

第一步 —— 定义目标与指标（Define your objectives and metrics）

这是整个分析过程的起点。

- **目标设定（Goal Setting）**：
  - 明确希望通过社交媒体实现什么目标。
  - 示例：提升品牌知名度、驱动流量、生成潜在客户（leads）等。
- **关键绩效指标（KPIs）**：
  - 用于衡量进展的具体指标。
  - 常见 KPI 包括：
    - 达人度（Reach）
    - 曝光量（Impressions）
    - 参与率（Engagement Rate）
    - 点击率（CTR）、转化率
    - 情感倾向（Sentiment）

    - 关注增长（Follower Growth）
- **目标受众（Target Audience）**：
    - 明确你的用户是谁？
    - 分析其人口统计特征（年龄、性别、地区）、兴趣爱好、在线行为习惯。
    - 有助于定制内容并选择合适的平台（如年轻人用 TikTok，专业人士用 LinkedIn）。

✅ **要点**：没有清晰的目标，就无法有效评估结果。目标决定数据收集方向和分析重点。

第二步 —— 收集和准备数据

数据来源（Data Sources）：

- 使用平台内置工具：如 Facebook Insights、Instagram Analytics
- 第三方工具：Hootsuite、Sprout Social、Brandwatch 等
- 其他辅助数据源：Google Analytics（用于追踪网站流量来源）

数据采集方法（Data Collection Methods）：

- 选择定量数据（数字型）：点赞、分享、提及次数
- 或定性数据（描述型）：评论、评价、情感分析
- 或两者结合（混合方法）

数据清洗与组织（Data Cleaning and Organization）：

- 清理无效或重复信息
- 过滤垃圾信息（spam）、机器人活动（bot activity）
- 将非结构化数据（如文本、图片）转化为可分析的结构化格式
- 例如：将“#数码产品”标签归类为“科技”类别

📌 **重要性**：“垃圾进，垃圾出”——数据质量直接影响分析结果的准确性。

第三步 —— 分析与解读结果

本阶段是对数据进行深度挖掘，形成洞察。

主要分析维度包括：

表格

类别	内容
Audience Insights（受众洞察）	分析用户画像：年龄、性别、地域、兴趣偏好、活跃时间
Performance Analysis（表现分析）	跟踪 KPI 随时间变化趋势，对比历史活动或竞争对手
Content Analysis（内容分析）	测试不同内容形式（图文、视频、直播）的效果，找出最受欢迎的内容类型
Sentiment Analysis（情感分析）	判断公众对品牌、产品的情感倾向（正面/负面/中立），识别危机信号
Trend Analysis（趋势分析）	发现热门话题、流行标签（hashtags）、新兴趋势，提前布局

💡 **示例应用**：

- 若发现某条视频在周末播放量激增 → 调整发布时间
- 若评论中频繁提到“价格太高” → 可能需要促销或优化定价策略

第四步 —— 行动与策略优化

分析的最终目的是指导行动。

核心步骤：

1. **制定可操作的洞察（Actionable Insights）**
    - 将分析结果转化为具体建议。
    - 如：“年轻用户更喜欢短视频，建议增加 TikTok 内容比例。”
2. **优化营销活动（Optimize Campaigns）**
    - 改进内容创意、发布频率、投放时间、目标人群定位
    - 优化付费广告投放策略（如 Meta Ads、Google Ads）

3. **持续监控与调整（Ongoing Monitoring and Adjustment）**
    - 社交媒体环境快速变化，需定期复盘
    - 动态调整策略以应对新趋势、突发事件或竞争变化

🔁 **循环机制：目标 → 数据 → 分析 → 行动 → 再评估** 是一个持续迭代的过程。

第五步 —— 报告与沟通

分析完成后，必须向利益相关者（stakeholders）清晰传达结果。

关键做法：

- **创建清晰简洁的报告：**
    - 使用可视化图表（柱状图、折线图、仪表盘）
    - 避免技术术语，确保非技术人员也能理解
    - 示例图显示了典型的“社交媒体分析报告”界面，包含总互动数、参与率、增长率等
- **突出关键结论（Key Takeaways）：**
    - 强调最重要的发现（如“本月品牌好感度上升20%”）
    - 提供明确的后续建议（如“建议下周推出新品预告视频”）
- **选择合适的工具：**
    - 推荐使用 Tableau、Power BI、Google Data Studio 等工具
    - 便于制作动态仪表板和自动化报告

🎯 目标：让决策者快速掌握情况，并做出基于数据的判断。

社交媒体分析 vs 社交网络分析

概念	社会媒体分析（SMA）	社交网络分析（SNA）
关注点	内容、用户行为、情感、趋势	用户之间的关系结构、影响力传播
数据类型	文本、图像、评论、点赞等	关系链（follow/friend）、互动模式
目的	了解品牌表现、市场情绪	识别意见领袖、社区结构、信息扩散路径

概念	社会媒体分析（SMA）	社交网络分析（SNA）
典型应用	品牌监测、广告优化	危机传播分析、影响力营销

- **SMA** 更偏向“**说什么**”（content-driven）
- **SNA** 更偏向“**谁跟谁连在一起**”（relationship-driven）

两者常结合使用，例如：先用 SMA 找到热点话题，再用 SNA 查看该话题是如何在特定社群中传播的。

社交媒体分析定义与框架

引用 Alam and Khan (2021) 的研究，提出 SMA 的三个阶段“配方”模型：

三大阶段：

1. **数据识别（Data Identification）**
2. **数据分析（Data Analysis）**
3. **信息解读（Information Interpretation）**

关键技术“原料”：

- 自然语言处理（NLP）
- 机器学习（ML）
- 信息检索（Information Retrieval）
- 数据可视化（Visualization）

✅ **比喻**：就像做菜一样，SMA 是一个“三步法”的流程，依赖多种技术“调料”来完成。

社交媒体数据类型

社交媒体数据是多模态、异构的，主要包括五类：

1. **文本数据（Text Data）**
    - 帖子、推文、评论、标签、评分等
    - 是 NLP 的主要输入，可用于情感分析、主题建模

2. **视觉与多媒体数据（Visual & Multimedia Data）**
    - 图片、表情包、视频、音频、直播流
    - 适用于计算机视觉任务，如品牌识别、场景检测
3. **网络与图数据（Network & Graph Data）**
    - 用户之间的关系：关注、点赞、转发、回复
    - 是社交网络分析的基础，用于构建关系图谱
4. **时间序列数据（Temporal Data）**
    - 发布时间戳、频率趋势、病毒内容生命周期
    - 用于分析事件发展节奏、预测爆发节点
5. **元数据（Metadata）**
    - 用户资料（年龄、地点）、设备信息、地理位置、互动指标（likes/shares）、平台来源
    - 提供上下文背景，增强分析深度

📌 总结：**社交媒体数据 = 结构化 + 非结构化**，需综合处理才能获得全面洞察。

社交媒体分析的三个主要阶段

1. 数据识别（Data Identification）

- **目标**：找到、收集并组织相关的社交媒体数据
- **方法**：

    ▼ API 接口访问（如 X/Twitter API、Facebook Graph API）
        - 使用官方 API 获取结构化数据（如 Twitter API、Facebook Graph API）
        - 可提取：帖子、评论、用户信息等
        - 需要认证（authentication），遵守平台规则和条款

    ▼ 网络爬虫（web scraping）
        - 当 API 不可用或受限时，自动抓取网页内容
        - 必须遵守法律、伦理规范及网站规则（如 robots.txt、请求频率限制）

▼ 使用已整理的数据集（curated datasets）

- 利用研究人员或组织共享的数据资源
- 示例：Kaggle、学术论文附带数据、政府开放数据平台

- **注意事项**：

1. **数据隐私（Data Privacy）**
   - 尊重用户隐私权
   - 遵守服务条款和法律法规（如 GDPR —— 欧盟通用数据保护条例）
   - 不得滥用个人身份信息（PII）
2. **数据质量与相关性（Data Quality & Relevance）**
   - 数据应与业务或研究目标高度相关
   - 聚焦关键词、话题标签（hashtags）、时间范围等维度
   - 避免无关噪音干扰分析结果
3. **抽样策略（Sampling）**
   - 对大规模数据进行代表性采样，以降低计算成本
   - 保证样本能反映整体趋势（例如随机抽样、分层抽样）

2. 数据分析（Data Analysis）

- **目标**：处理和分析数据，发现模式、趋势和洞察

方法	功能
自然语言处理（NLP）	处理文本内容
- 情感分析（Sentiment Analysis）	判断情绪倾向：正面 / 负面 / 中立
- 主题建模（Topic Modeling）	发现讨论热点主题（如 LDA 模型）
- 实体识别（Named Entity Recognition）	提取品牌、人物、地点等关键实体

- **技术手段**：
  - NLP：情感分析、主题提取、实体识别
  - 机器学习：分类（正负情绪）、聚类（用户分群）、预测（未来趋势）

- 信息检索：筛选关键词、提取相关内容
- 可视化：图表、图形、仪表盘辅助探索

3. 信息解读（Information Interpretation）

- **目标**：将分析结果转化为**可执行的洞察和建议**
- **实践方式**：
  1. **情境化锚定（Contextual Grounding）**
     - 将量化结果与现实事件联系起来
     - 示例：某品牌负面情绪突然上升 → 是否有产品质量问题？是否被媒体报道？
  2. **多模态综合（Multimodal Synthesis）**
     - 整合不同类型的分析输出
     - 示例：将“主题分布”叠加到“社交网络中心性指标”上，揭示哪些意见领袖推动了某个话题传播
  3. **推理推导（Inference Derivation）**
     - 把计算结果转化为定性结论
     - 示例：聚类密度高 → 表明存在活跃社区；分类概率高 → 表明公众态度明确

🎯 **终极目标**：不是仅仅展示数据，而是推动决策和行动。

> 社交媒体分析 ≠ 看数据
>
> 它是一个**闭环系统**：**目标** → **数据** → **分析** → **洞察** → **行动** → **反馈** → **优化**

它融合了**技术能力**（NLP、ML、图分析）与**业务理解力**（市场、用户、品牌），是现代数字营销、公共关系、社会治理等领域不可或缺的能力。

实战案例演示 —— “新产品发布公众情绪分析”

Mini Walkthrough Example – 应用 SMA 框架

> 情景设定：你想分析公众对某公司新产品的发布反应，使用 Twitter/X 平台的数据。

🖼️ 展示内容：

- 多条真实推文截图，涉及对微软新AI产品“Generative AI for Beginners”的讨论
- 用户评论包括：
  - 积极反馈：“Perfect to jumpstart our AI journey”
  - 质疑声音：“This card is not the time free.”
  - 建议改进：“They just need your all documents and kyc.”

➡️ 这些是典型的**用户生成内容（UGC）**，可用于情感分析和趋势追踪。

Step-by-Step Application – 数据识别阶段

[数据库] → [Twitter Streaming API]
↓
[Raw Tweets] （原始推文）
↓
[Data Processing] （清洗、过滤）
↓
[Analyzing and aggregating tweets] → [Data Storage]

具体操作：

1. **通过 Twitter/X API 收集**提到该产品的推文（在发布周内）
2. **过滤关键词和话题标签**（如 #NewProduct、#MicrosoftAI）
3. **存储处理后的数据**（用于后续分析）

Step-by-Step Application – 数据分析阶段

操作步骤：

1. **预处理推文**：
   - 清洗文本（去除标点、停用词）
   - 删除垃圾信息（spam）、重复内容（duplicates）
2. **执行情感分析**：

- 使用 NLP 工具将每条推文分类为：**正面、负面、中立**

3. **可视化情绪趋势**：
   - 绘制每日情绪变化曲线图
   - 观察是否有情绪波动高峰（如发布会当天）
4. **（可选）识别关键影响者与热门话题**：
   - 找出转发最多、评论最多的用户（潜在意见领袖）
   - 提取最常出现的话题标签（trending topics）

✅ **输出成果**：一张清晰的情绪走势图，显示公众态度随时间的变化。

Step-by-Step Application – 信息解读阶段

结果总结与行动建议：

🔍 总结发现：

> “候选人 A 的正面情绪占比为 66%；候选人 B 为 56%；事件发生后帖子数量激增。”

- 情绪分析结果可视化（饼图、折线图）
- 显示情绪波动的时间节点

🛠️ 推荐行动：

- 向支持者致谢（thank advocates）
- 回应公众担忧（address concerns）
- 准备常见问题解答（FAQ）应对频繁提问

📌 图中展示了一个完整的分析仪表盘，包含：

- 情绪比例（Pie Chart）
- 时间序列趋势（Line Chart）
- 热门话题列表（Hashtag List）
- 用户画像统计

总结：

步骤	内容	目标
Step 1: 数据识别	通过 API、爬虫或公开数据集获取相关数据	获取高质量、合规的数据
Step 2: 数据分析	使用 NLP、ML、IR 和可视化技术挖掘模式	发现情绪、主题、关系等洞察
Step 3: 信息解读	将技术结果转化为业务意义和行动建议	推动决策与优化策略

社交网络分析（Social Network Analysis）

✅ 定义：

> 社交网络分析是使用图论和网络工具来研究社会结构的过程。

🔍 关键方面：

- 将人、组织或实体建模为**节点（nodes）**
- 将它们之间的关系（如友谊、关注、交易）建模为**边（edges）**
- 揭示连接模式、影响力传播路径以及信息流动机制

📌 **本质**：用数学图形的方式理解人际关系和社会结构。

网络理论 —— 基本组件

1. 节点（Nodes / Vertices）

- 表示网络中的实体（例如：A、B、C、D、E）
- 可以是个人、公司、网页等

2. 边（Edges / Links）

- 表示两个节点之间的关系或互动
- 示例：A-B、C-D 表示 A 和 B 相互连接，C 和 D 也相连

→ 节点（Nodes / Vertices）的属性

节点不仅可以表示“谁”，还可以携带多种属性：

🌐 自身属性（Self-properties）：

- **权重（Weight）**：重要性或价值（如销售额、影响力）
- **大小/位置（Size/Position）**：用于可视化时体现空间或层级意义
- **其他属性**：人口统计信息（年龄、性别）、类别（用户类型）

📊 网络相关属性（Network-based properties）：

- **度（Degree）**：该节点有多少个邻居（连接数）
- **聚类（Cluster）**：属于哪个连通社区或子群

📌 例如：一个高影响力的用户可能有很高的度，并且位于一个活跃的社群中。

→边（Edges / Links）的属性

边不仅仅是"有没有连接"，还可以具有丰富属性：

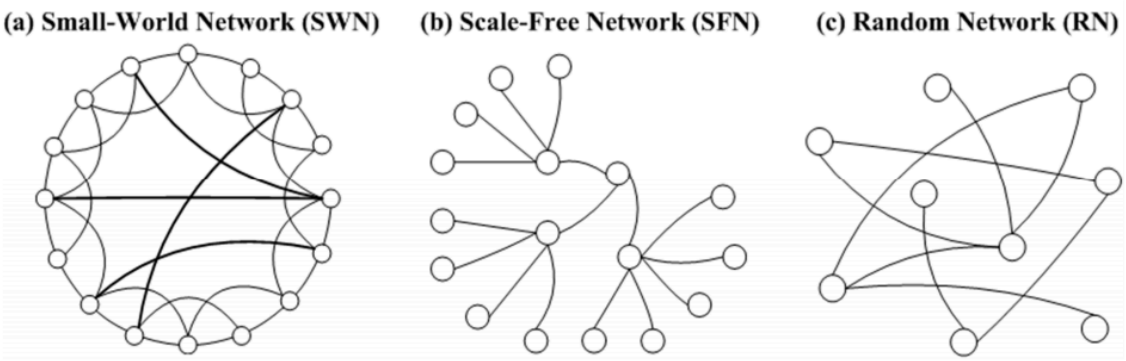
🔗 边的属性包括：

- **权重（Weight）**：连接强度或频率（如点赞次数、消息数量）
- **方向（Direction）**：是否为单向关系（如 A 关注 B 但 B 不关注 A）
  - 有向图（Directed Graph）：箭头表示方向
  - 无向图（Undirected Graph）：没有方向
- **时间（Time）**：交互的时间序列或持续时间

复杂网络（Complex Networks）

现实世界的社交网络通常不是随机生成的，而是具有独特结构。

三种典型模型比较：



类型	特点	示例
(a) 小世界网络（SWN）	大多数节点只与少数邻居相连，但任意两点之间可通过少量中间节点连接	Facebook 好友圈
(b) 无标度网络（SFN）	存在少数“枢纽”节点（hub），拥有大量连接，其余节点连接较少	Twitter 上的意见领袖
(c) 随机网络（RN）	节点间连接概率均等，缺乏结构性	理想化模型，现实中少见

💡 **关键发现：**

- 实际社交网络多为 **小世界 + 无标度** 结构
- 意味着：虽然大多数人只有少量联系人，但少数“超级连接者”可以快速传播信息

中心性度量（Centrality Measures）

衡量一个节点在网络中“有多重要”的指标。

🔎 四种主要中心性指标：

表格

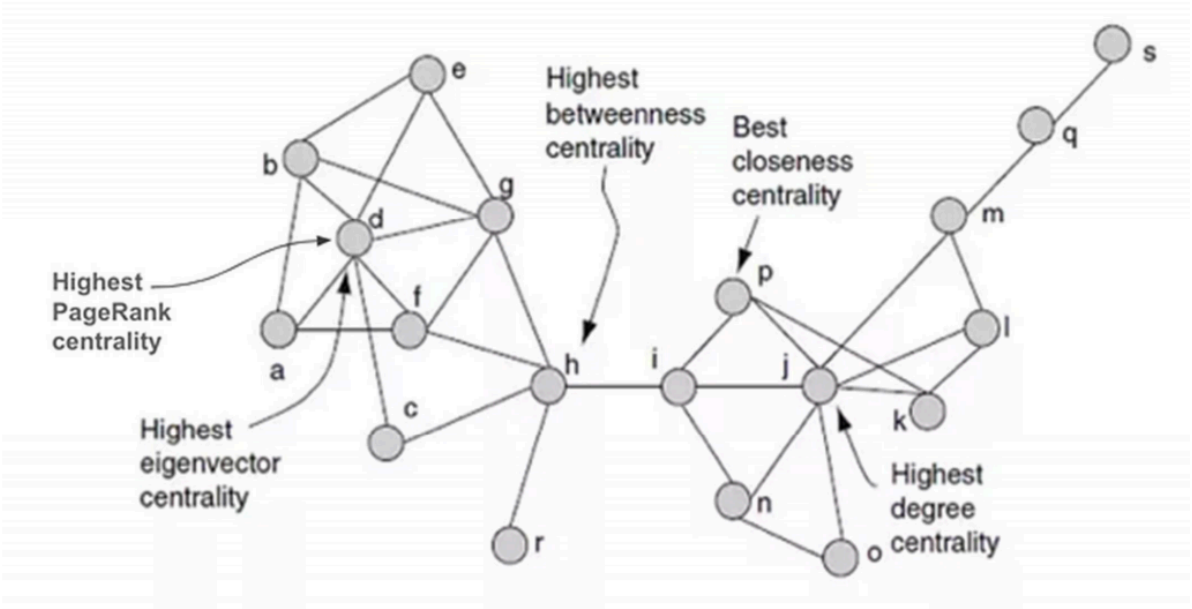
中心性类型	含义
度中心性（Degree）	节点的直接邻居数量（即连接数）
特征向量中心性 / PageRank	迭代计算：不仅看自己有多少连接，还看连接的是不是“重要”节点
接近中心性（Closeness）	到所有其他节点的平均距离越短，说明越“靠近中心”

中心性类型	含义
中介中心性（Betweenness）	多少最短路径经过该节点 → 是信息传递的关键桥梁

📌 **应用场景：**

- 找出关键影响者（KOL）
- 识别传播瓶颈
- 发现社区边界

中心性度量示意图



一张复杂的网络图展示了不同中心性的最高值节点：

- **最高 PageRank 中心性**：被许多高质量节点链接的节点（如核心网站）
- **最高特征向量中心性**：位于多个高连接节点之间的节点
- **最高中介中心性**：处于多个子群之间的“桥梁”
- **最高接近中心性**：到所有其他节点都最近的节点
- **最高度中心性**：连接最多的节点（可能是枢纽）

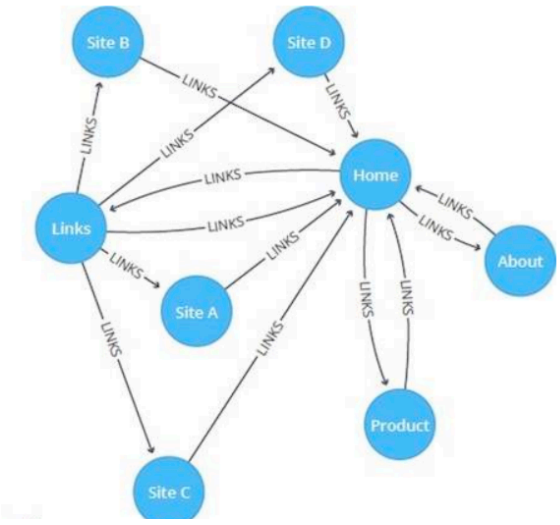
🔍 观察重点：不同的中心性会选出不同的“重要”节点，取决于你关心什么。

中心性度量的应用

中心性	应用场景
PageRank	网页排名（Google 搜索引擎的核心算法）
中介中心性（Betweenness Centrality）	识别关键节点（如关键影响者、信息瓶颈、交通枢纽）
接近中心性（Closeness Centrality）	找出交通中心（如机场、物流中心）——离所有人最近的地方
其他度量	针对特定领域定制（如学术合作网络、疾病传播网络）

📌 **核心思想**：选择合适的中心性指标，才能回答正确的业务问题。

PageRank 算法示意图



> PageRank 是 Eigenvector Centrality 的变体

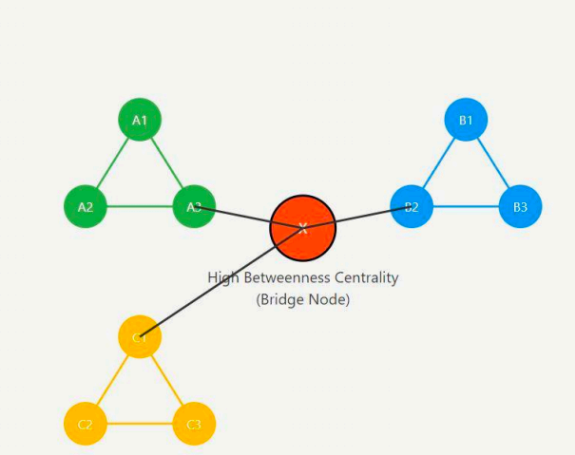
图解说明：

- “Home” 页面被多个站点链接，因此它的重要性高
- 每个页面的 PageRank 值由其入链（in-links）决定

- 如果一个高 PageRank 的页面链接到你，你的 PageRank 也会提升

🧠 **类比**：就像学术界引用论文一样，“被越多权威文献引用的论文越重要”。

中介中心性（Betweenness / Hub Nodes / Bridge Nodes）



🔄 概念：

- **中介中心性高的节点**：是连接不同群体的“桥梁”
- 在图中用红色大圆表示（Bridge Node）

💡 实际意义：

- 是信息传播的关键通道
- 若移除这类节点，可能导致网络分裂
- 在营销中可用于寻找“跨社群影响者”

📌 图中显示：

- 左侧绿色群体和右侧蓝色群体通过一个红色节点连接
- 该节点就是典型的“桥接节点”（Bridge Node）

接近中心性（Closeness）—— 以交通网络为例

🚇 图例：港铁系统地图（MTR system map）

- **接近中心性**：衡量一个站点到所有其他站点的平均距离
- **中心站**（如中环、金钟）往往具有更高的接近中心性
- 它们是“最方便到达所有地方”的站点

✅ **类比**：

- 在社交网络中，接近中心性强的用户能最快接触到整个网络
- 在城市规划中，接近中心性强的地铁站更受欢迎

信息流动（Information Flow）

🌪️ 信息扩散过程类似于病毒传播：

- 从一个人传给他的社交邻居
- 具有传染性动态（contagious dynamics）

🧩 两种经典模型描述此过程：

1. **线性阈值模型（Linear Threshold Model）**
   - 个体是否采纳信息取决于其邻居的影响总和是否超过某个阈值
   - 类似于“说服力累积”
2. **独立级联模型（Independent Cascade Model）**
   - 每次传播都有一定概率成功
   - 成功后，该节点成为新的传播源
   - 更适合模拟社交媒体上的“转发”行为

📌 图中展示了一群人通过手机互相分享信息，形成类似病毒式传播的链条。

线性阈值模型（Linear Threshold Model, LT）

> 定义：

- 描述一个节点如何在邻居的影响下被“激活”（即受感染或采纳某种行为）。

- 适用于如电影推荐等场景：朋友反复推荐后你最终决定去看。

✅ 数学公式：

$\sum_{u \in N(v) \cap A} b_{uv} \geq \theta_v$

其中：

- $N(v)$：节点 v 的邻居集合
- A：当前已激活的节点集合
- b_{uv}：邻居 u 对 v 的影响权重（$\sum b_{uv} \leq 1$）
- θ_v：节点 v 的阈值（通常在 [0,1] 区间）

📌 示例解释：

- 节点 v 有三个邻居：u1、u2、u3
- 它们对 v 的影响分别为 0.4、0.3、0.2
- 如果 v 的阈值 $\theta_v = 0.6$，那么当 u1 和 u2 激活时，总影响为 0.7 ≥ 0.6 → v 被激活

💡 **类比**：就像一个人听到多个朋友说某部电影很好看，积累到一定程度就决定去看。

独立级联模型（Independent Cascade Model, IC）

> 定义：

- 当一个节点被激活后，它有一次机会尝试激活每个未激活的邻居。
- 每次尝试成功的概率是固定的，且与其他尝试无关。

✅ 特点：

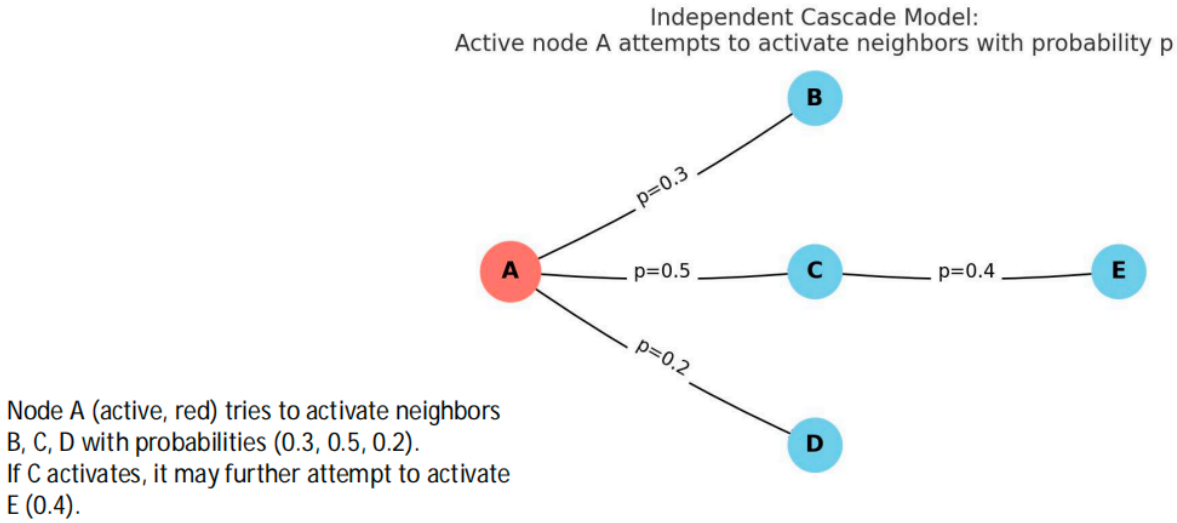
- 每次传播是独立事件（类似病毒传播）
- 成功后该节点不再尝试（只有一次机会）
- 过程持续直到没有新的激活发生

🧩 类比：

- 类似于新冠传播：每次接触都有一定概率传染，但一旦传完就不会再传

📌 **应用场景**：社交媒体上的“转发链”、“谣言扩散”等

独立级联模型（IC）图示说明



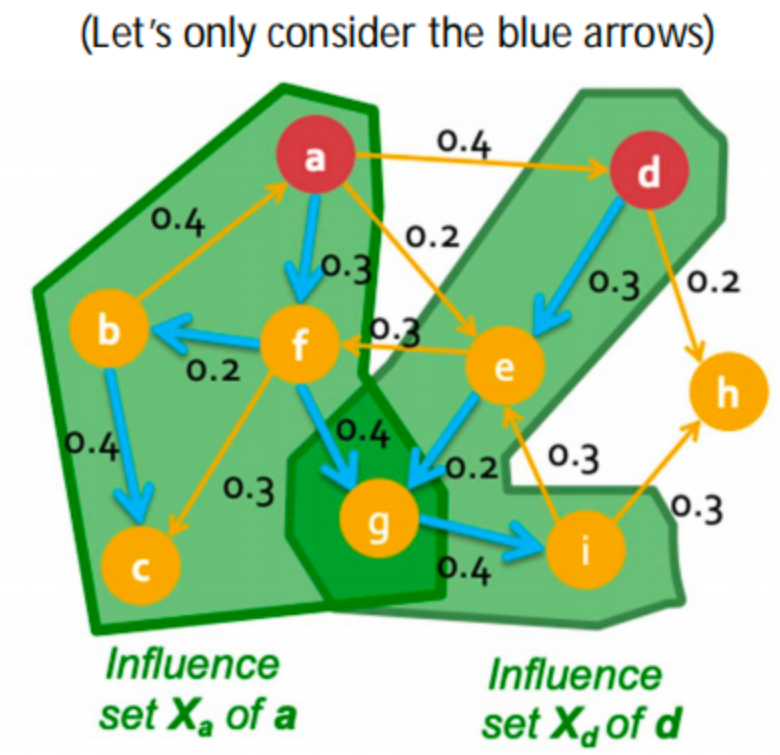
图解流程：

- 红色节点 A 已激活
- A 尝试激活其邻居 B、C、D、E，分别以概率 p=0.3、0.5、0.2、0.4
- 若 C 被成功激活，则 C 又会尝试激活 E（概率 0.4）

✅ **关键点**：

- 传播是**顺序进行**的
- 每一步的成功取决于**随机概率**
- 最终形成的传播路径可能不同（多次运行结果不一致）

影响力最大化（Influence Maximization）



影响力最大化问题

> 目标：

- 在有限预算下选择一组初始节点（称为"种子节点"，seeds），使信息通过自然传播尽可能覆盖更多人。

✅ 应用场景：

- 市场营销：邀请少数意见领袖参加新品发布会，带动整个社群讨论
- 公共卫生：识别关键传播者以控制疫情
- 政治动员：精准投放宣传资源

📊 图例说明：

- 图中绿色区域表示节点 a 的影响力范围 X_a
- 红色区域表示节点 d 的影响力范围 X_d

- 两个节点的影响力有重叠，但整体覆盖更广
- 目标是选择能带来最大影响力的组合（如 a 和 d）

📌 注：仅考虑蓝色箭头（代表传播方向）

影响力最大化 —— 核心目标与应用

✅ 目标：

- 选择有限数量的种子节点（seeds）
- 最大化信息在社交网络中的自然传播范围

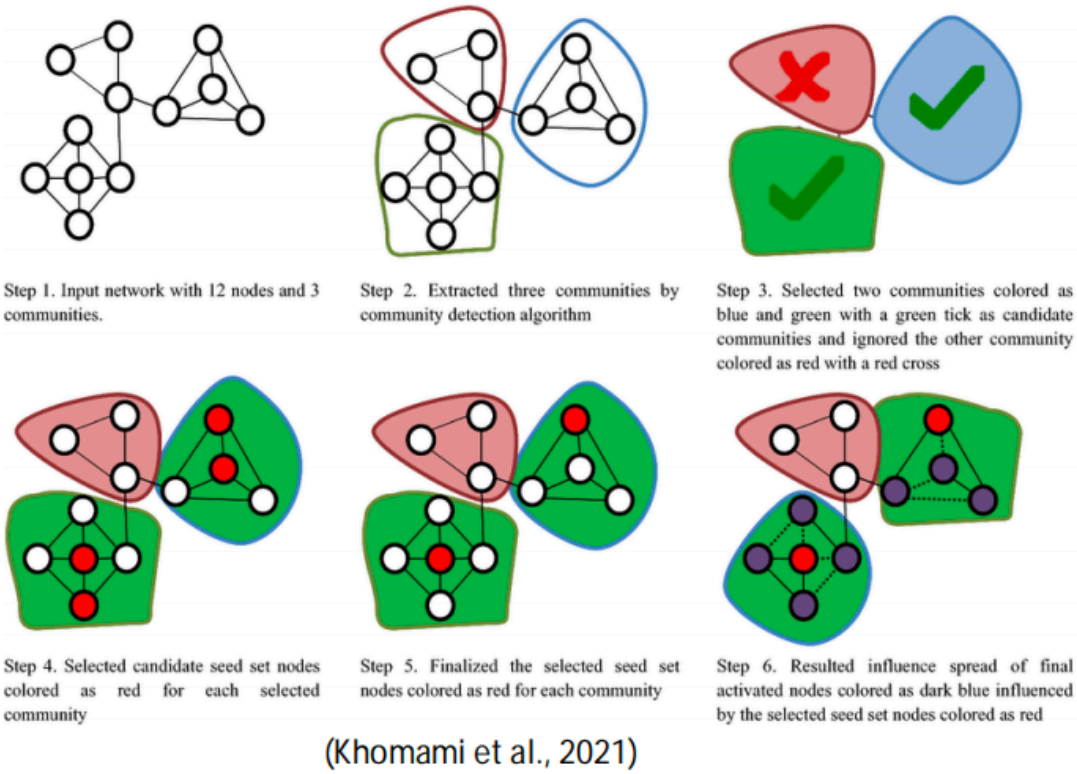
✅ 应用领域：

- 营销活动（如品牌推广）
- 信息战役（如公益宣传）
- 流行病学（如疫苗接种策略）
- 政治动员（如选举宣传）

✅ 如何识别这些种子？

- 使用前面提到的**中心性度量**（如 PageRank、中介性、接近性）
- 或使用专门算法（如 Greedy Algorithm、Degree Discount Heuristic）

影响力最大化 —— 实际操作流程图



(Khomami et al., 2021)

来自 Khomami et al., 2021 的六步流程：

1. **输入网络**：包含 12 个节点和 3 个社区的社交网络
2. **检测社区**：使用社区发现算法划分出不同子群
3. **筛选候选社区**：选出两个最具潜力的社区（绿色勾选），忽略其他（红色叉号）
4. **为每个社区选择种子节点**：在选定社区内挑选高影响力用户（红色圆点）
5. **确定最终种子集**：整合所有候选种子
6. **模拟传播效果**：展示最终被影响的节点（深蓝）和未受影响的节点（浅蓝）

🎯 **结果**：通过合理选择种子，实现了高效的信息扩散。

社交网络的科学与工程

第59页：社交网络的科学 vs 工程

维度	科学（Science）	工程（Engineering）
目的	研究人类社会行为	构建更好的系统支持用户

维度	科学（Science）	工程（Engineering）
问题	- 人们如何互动？- 人们如何被影响？- 信息如何在网络中流动？- 为什么某些话题迅速走红？	- 我们可以给社交平台添加什么新功能？- 如何推荐内容让用户满意？- 用户搜索时如何快速找到相关信息？

💡 **本质区别**：

- **科学**：探索“是什么”和“为什么”
- **工程**：解决“怎么做”和“如何优化”

社交网络分析的科学与工程本质

✅ 定义（IEEE Intelligent Systems）：

> 社交媒体分析致力于开发和评估信息学工具与框架，用于从海量社交媒体数据中提取、分析、总结和可视化信息。通常由特定目标驱动。

✅ 关键观点：

- 社交媒体分析不仅仅是数据分析或统计
- 它是一个**跨学科领域**，融合了：
  - **人文社科**：心理学、社会学、传播学
  - **工程技术**：计算机科学、人工智能、数据挖掘

🎯 总结：

> 社交网络不仅是技术系统，更是人类行为和社会结构的映射。
>
> 因此，研究它需要**科学思维 + 工程能力**的双重结合。
