北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

社交网络分析与应用

尹清宇 叶小伟 王浩 刘秉杰 吴优 胡仲则

Contents

结构大纲

社交网络的概述

节点的运作方式

社交网络与安全

内容相关社交网络

社交网络的应用

社交网络前沿方向

demo展示

1、社交网络的概述

社交网络

社交网络是由一组节点（如个人或组织）、节点之间的关系组成的社会结构。

当我们想到社交网络时，我们会想到Facebook、Twitter、微博或其他被称为“社交网络”的网站，而事实上，这些网络是狭义上的社交网络。

广义的社交网络，即所有能够连接人与人的媒介技术所构成的关系都可称为社交网络（例如邮件网络、道路网络、web网络、论文引用网络、twitter、微博等网络社区）。

社交网络的基本特征

基本特征

- 有一个构成网络的实体的集合。通常情况下。这些实体是人，但也可以是完全不同的东西。
- 网络中的实体之间至少有一种关系。
- 有一个非随机性（nonrandomness）或地域性（locality）的假设。这个条件是最难正式化的，但直觉是，关系倾向于聚集。也就是说，如果实体A与B和C都有关系，那么B和C有关系的概率就会比平均高。

基于结构的分析和基于内容的分析

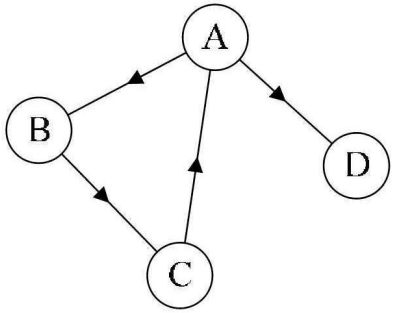
现存两种类型的数据：
1、表示网络的连接、交互和拓扑结构的结构数据
2、内容数据，重点关注用户在社交网络中包含和共享的信息。

（对这两种数据的孤立分析将提供存储在网络中的信息的不完整视图，因此可能会丢失潜在的模式和知识）

考虑到这两种类型的数据，可以区分两种不同的分析方法：
1、基于结构的分析方法（Structural-based Analysis）
2、基于内容的分析方法（Content-based Analysis）

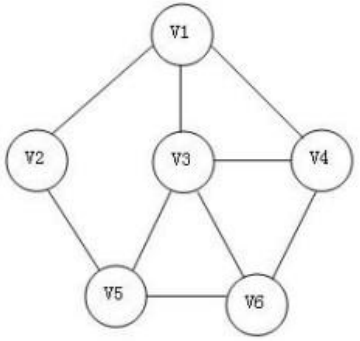
基于结构的分析方法

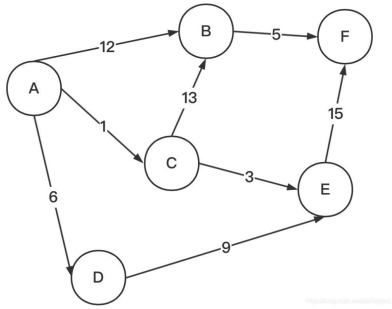
使用网络理论（通常称为图论）进行分析。图论将社交网络表示为一个图，表示为边和节点，其中节点是个体，边是连接它们的关系。通过图论，可以概括和分析用户之间现有的社区互动，以及他们在他们的联系网络中的行为。



图：G=(V,E)，V为节点集合，E为边集合。

关于边（表示为个体之间的关系）
（1）可以是有向的（比如用户之间的关注关系）
（2）可以是无向的（如在Facebook或其同类产品上，个体之间存在一种称为朋友的关系，这种关系是全有或全无的且是相互的）
（3）也可以拥有权重（关系有一个程度。这个程度可以是不连续的，例如，朋友、家人、熟人；或者也可以是一个真实的数字；一个例子是两个人平均每天花在互相交谈上的时间的分数。）





各种不同类型的图：
树、森林、生成树、斯坦纳树、完整图、平面图、二叉图、规则图或桥等等。

图模型特别适用于基于网络的领域和问题。

任何网络（或图）模型的主要目标是重现或模仿真实网络的主要特征。为了确定这些特征，需要确定和评估一些措施以确保 这些测量方法与真实世界的网络相一致。三个网络属性通常被用来评估这些行为：

1、度分布（degree distribution）

度分布表示节点度（node degrees）在网络中的分布情况。

2、聚类系数（clustering coefficient）

聚类系数衡量一个网络的跨度（transitivity）。

3、平均路径长度（average path length）。

平均路径长度表示任何一个节点之间的平均距离 (最短路径长度)。

1、真实世界的网络（Real-World networks）：
节点之间的连接（或联系）取决于特定的概率。在现实世界的网络中，如社交网络，这意味着如果两个节点之间存在任何类型的关系（友谊、职业等），那么它们之间建立联系的概率会更高。
现实世界网络中的度分布值，如社交网络，遵循幂律分布（简单来说就是：节点具有的连线数和这样的节点数目乘积是一个定值）。
社会网络与其他类型的网络不同，主要是因为社会网络比非社会网络更容易被划分为社区。因此，这一特性影响了度分布，通常度是正相关的，而且聚类程度高。

2、随机网络（Random networks）：
从理论的角度来看，随机网络，最常见的是随机图，属于图论和概率论的交叉领域。任何随机图都可以被描述为一个概率分布，或者说是一个随机过程，它将被用来生成图中任何一对节点之间的连接。
因此，要生成一个随机图，从一组n个孤立的节点开始，只需要随机地添加（按照预定的概率分布）连续的 预先确定的概率分布）在任何一对节点之间连续添加边。
随机图通常被用来比较任何图的结构和属性。

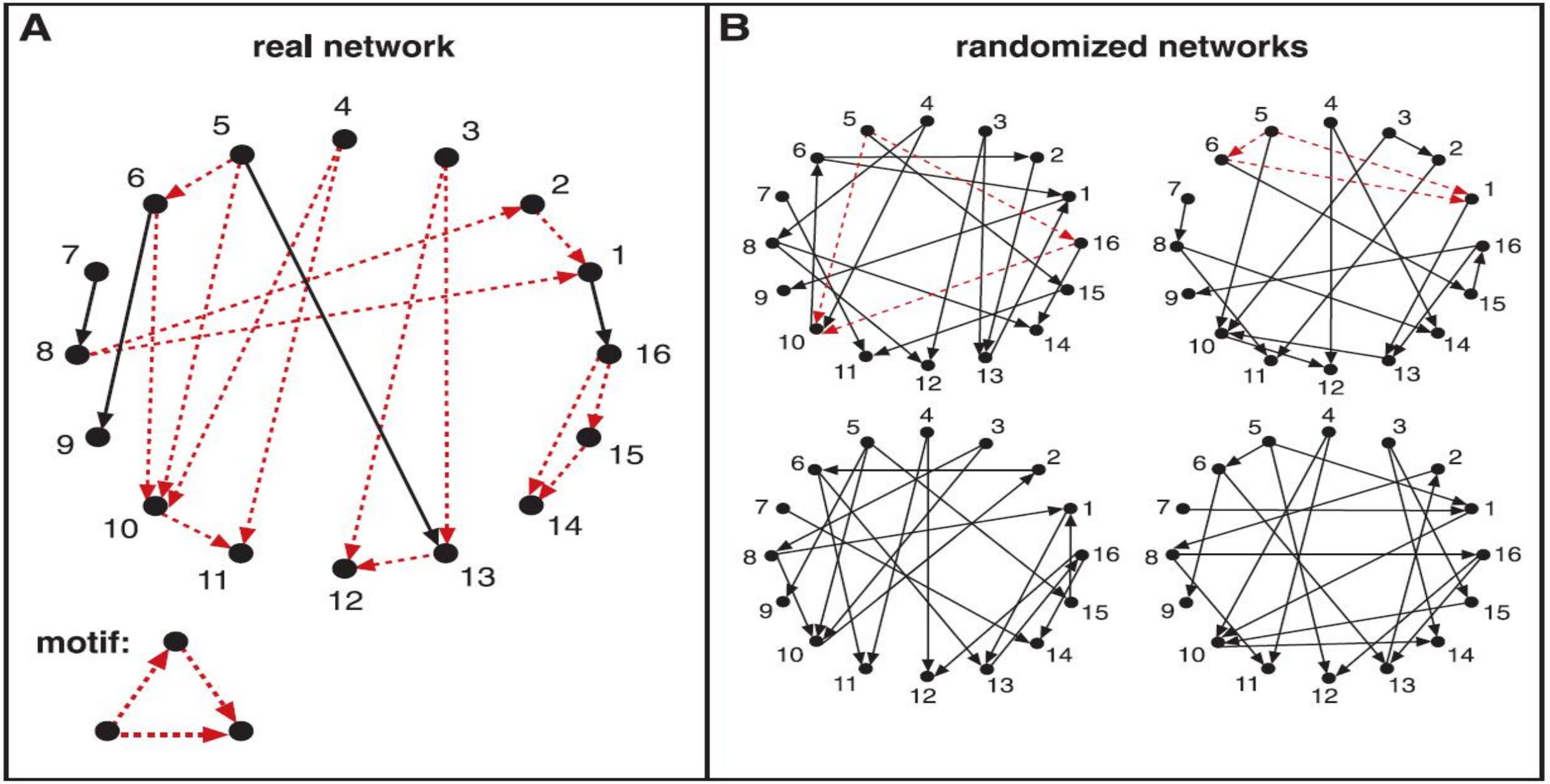
3、小世界网络（Small-World networks）：
许多现实世界的网络中，有两个基本属性，第一个是两个节点之间的距离通常很小，第二个是网络的传递性（或者说是聚类系数）比较高。
这种被大量现实世界网络所共享的属性，通常被称为小世界属性。这个概念由两个人之间的 "六度分隔 "这样的术语推广开来，意思是任何两个人之间的 距离最多只有六个友谊链接。

4、无尺度的网络（Scale-free networks）：
这类网络通常是根据其度分布来定义的，节点的度分布遵循幂律（至少是渐进式的），也就是说，网络中与其他节点有k个连接的节点的比例$P(k)$在k的大值下会变成$P(k) \sim k^{-\gamma}$，其中γ是一个参数，其值通常在2和3的范围之间。

在引入之前的概念和模型之后，就可以开始定义一些基本的度量标准了。
它们通常被图算法所使用。
1、中心性（Centrality）
中心度定义了一个节点在网络中的重要程度。在在线社交网络中，这个度量可以用来检测或识别网络中最有影响力的人。
根据评估方法不同，可以分为：度数（Degree）中心性、特征向量（Eigenvector）中心性、PageRank、间性（Betweenness）中心度、紧密性（Closeness）中心度、组（Group）中心性等。
2、传递性（Transitivity）和相互性（Reciprocity）
被用来表示网络中的链接行为。传递性分析链接行为，以确定它是否在三个节点之间表现出传递性行为。拥有更高传递性的图更接近于一个完整的图。
3、平衡（Balance）和地位（Status）
在一个有符号图（signed graph）中，其中每条边都有一个正负符号（可以代表人际关系，如朋友或敌人，老板或下属，社会地位）。
如果每个环的每条边的符号的乘积是正的，这种图就是平衡的。
在现实世界的社会网络中，我们期望关于这些互动关系有一定程度的一致性。例如，一个人的朋友成为朋友比成为敌人更有可能。在有符号图中，这种一致性转化为观察有三条正边（即所有的朋友）的循环（三边，三角形）比有两条正边和一条负边（即朋友的朋友是敌人）的循环更频繁。

2、节点的运作方式

节点运作方式——子图模式



子图模式计数
(子图模式枚举)
motif 计数

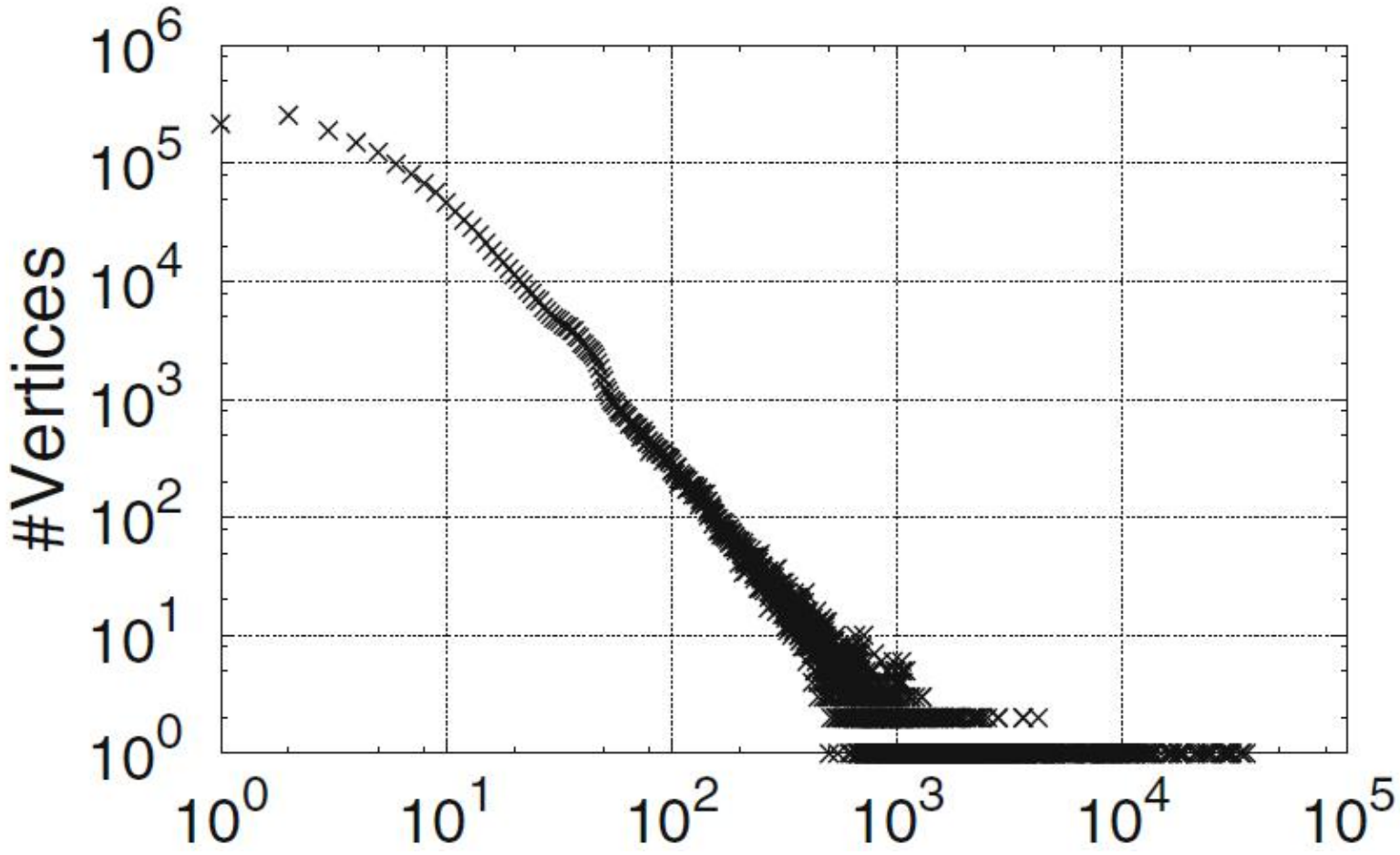
- 三角形枚举、计数
- 团枚举、计数
- 环枚举
- k-plex 枚举计数
- 所有 k-motif 计数（k<5）
- k-truss
- k-star
- ……

社交网络分析永远的前沿研究！

power law

局部稠密性

社交网络生成



社交网络具有连通性

网络名称	Nodes in largest WCC	Edges in largest WCC	Nodes in largest SCC	Edges in largest SCC
Facebook	1.000	1.000	1.000	1.000
Twitter	1.000	1.000	0.841	0.953
LiveJournal	0.999	1.000	0.790	0.954
Pokec	1.000	1.000	0.799	0.953
Wikipedia vote network	0.993	1.000	0.183	0.381

1967年哈[illegible]拉姆根据这概念做过一次连锁信件实验，[illegible]以联系任何两个互不相识的美国人。

小世界现[illegible]证，根据2011年Facebook 数据[illegible]亿用户中任意两个用户间的平均路径长度[illegible]67。可以说，在五步之内，任何两个网络上[illegible]

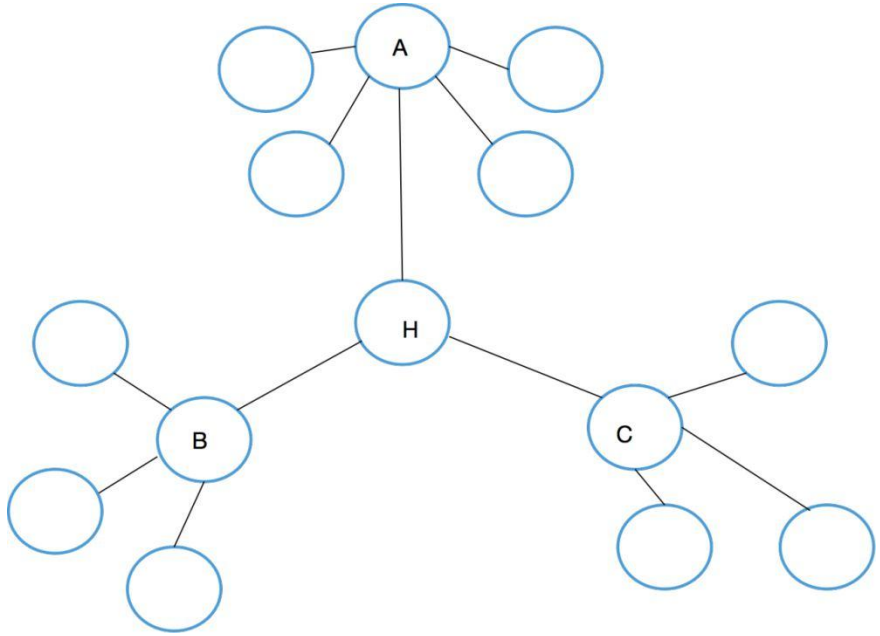
网络名称	Diameter
Facebook	8
Twitter	7
LiveJournal	16
Pokec	11
Wikipedia	7

$$\frac{3 \times G_{\wedge}}{3 \times G_{\triangle} +}$$

网络名称	Average clustering coefficient
Facebook	0.2647
Twitter	0.5653
LiveJournal	0.2742
Pokec	0.1094
Wikipedia	0.1409

为图中某节点承载整个图所有最短路径的数量

$$\sum_{v_s \neq v_i \neq v_t, s<t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$



What：基础理论定义：图论，motif，稠密子图，谱图理论……

How to compute：如何计算?

Analysis：社交网络分析相关研究

Application：应用

谱图理论：pageRank、GCN

稠密子图：k-core，最稠密子图

高维数据结构：超图，单纯复型

3、社交网络与安全

信息泄露

Facebook曾被报道泄露大规模用户隐私数据，超过5 000万用户的个人信息数据被第三方公司获得

Linked In服务器被黑客入侵，超过1亿名用户的个人信息资料被窃取

黑客窃取信息

维基解密的创始人阿桑奇

在美国大选时，疯狂曝光希拉里黑料

曝光美军在伊拉克屠杀百姓的视频

曝光教廷神父的性侵案

斯诺登通过维基解密曝光棱镜门事件

美国前总统杜鲁门曾说：“美国有95%的秘密情报，都在报纸和其他刊物上发表过。”

美国中央情报局80%的情报源于公开资料。

以色列情报“摩萨德”公开承认其获取的情报65%来自报刊、广播、电视和学术研究论文等公开渠道。

人肉搜索本身是一个有别于机器搜索的中性词，是指以互联网为媒介，通过海量人工互助提供知悉信息的方式，不断汇总和清晰信息，以查找人物或者事件真相的群众运动。

我们反对的侵犯人权的人肉搜索和作为其孪生兄弟的网络暴力，因为它们的杀伤力之强，社会危害性之大，通过众多事件已让人们有了清晰认识。

人肉搜索和网络暴力的违法性，主要体现在泄露个人隐私，侵犯公民个人信息，以及侮辱、诽谤他人人格，损害他人名誉等当面

(1) 身份认证方案

(2) 社交网络身份加密方案

(3) 数据安全管理

(4) 分组数据访问控制

4、内容相关社交网络

人类社会中的社交

- 强社交 -- 天生，无法躲避，没有隐私，身心疲惫，耗竭感，有基本的安全感。
- 弱社交 -- 利益衍生，合同，责任，不看出身，重能力，个人品质。
- 兴趣社交 -- 快乐，心理健康，自然，短连接，经历认同，体验决定质量。

社交网络中的社交

- 熟人社交（链接关系的社交网络） -- 微信、QQ
- 陌生人社交（与内容、兴趣相关的社交网络） -- 微博、贴吧、知乎

社交图谱 -- “我认识你”

- 2010年，Facebook提出。
- 它反映了用户通过各种途径认识的人，是人们线下关系在线上的简单映射。
- 社交图谱主要由一些主流的社交网络产生，例如Facebook或者微信QQ。
- 用户们互相向自己认识的人们发送邀请来构建和维持他们的社会关系。

兴趣图谱 --“我喜欢这个”

- 2010 年 Twitter 开发者大会上，Dick Costolo （当时twitter CEO）指出的概念。
- 是以人和人的共同兴趣 为线索的图谱，以分享共同的兴趣为基础，但是不一定互相认识。
- 通过人们相同的兴趣将众人聚集在一起。

两种图谱实际上是用户在互联网上另种不同角度的诉求。
关系图谱承载了用户与好友进行沟通互动的情感需求
兴趣图谱则体现了用户追求品位、获得知识的自我实现需求。

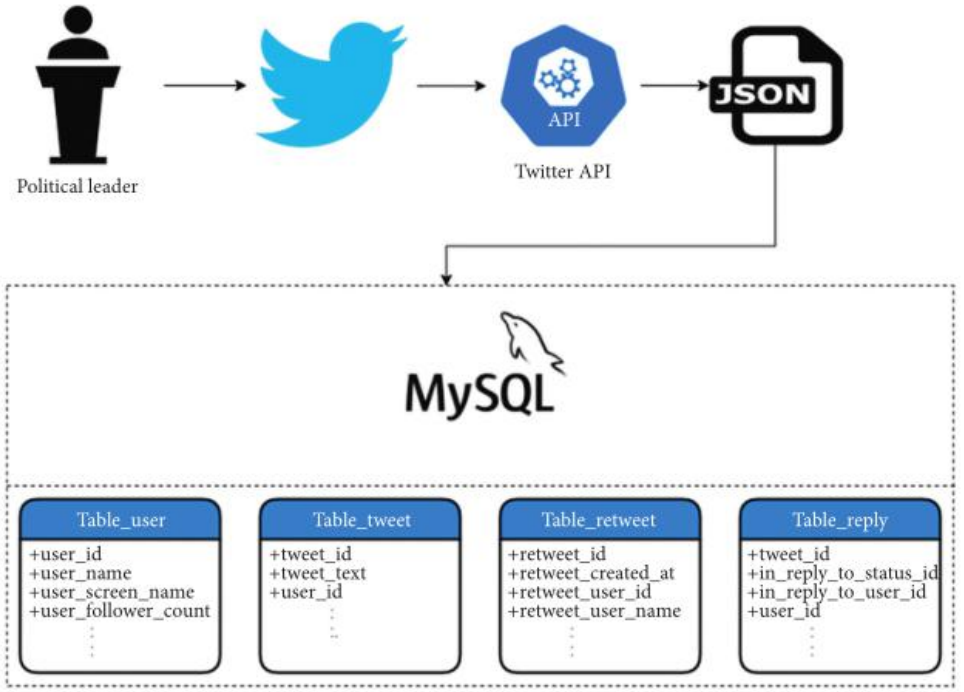
兴趣图谱	社交图谱
兴趣图谱以兴趣点为最小单位	社交图谱以人为最小单位
多为单向关系	多为双向关系
弱关系	强关系
关系拓展	关系维系
默认公开	默认私人

使用社会网络分析（SNA）技术 举例

- 使用开源python库“Tweepy”，可以方便地访问Twitter API，收集tweet上的推文。它提供了对官方Twitter API的所有对象和方法的访问。
- 数据(Tweets)以JSON的形式存储，json中包含大量数据，如用户信息、文本、转发、回复、提及、链接、标签、位置等。然后数据被解析到MySQL数据库，同时忽略不相关的信息。

用户兴趣挖掘

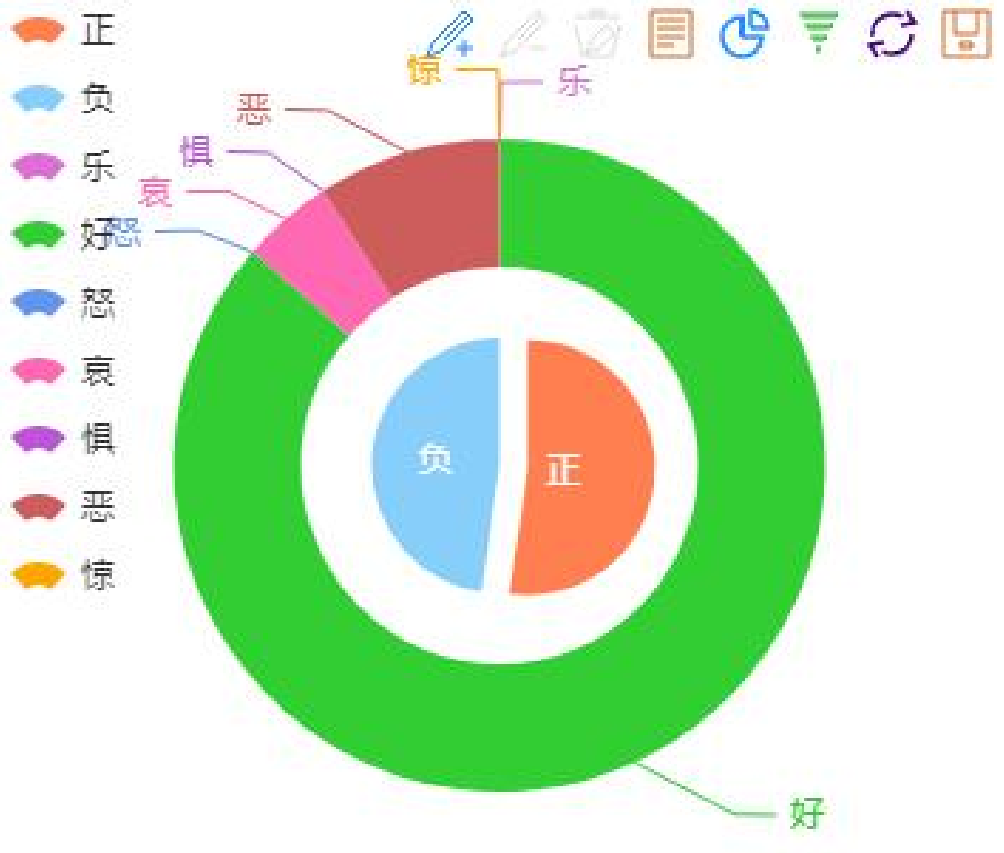
举例：政治支持者分析



Name	Party name	Followers (millions)	Following (K)	User name	Tweets (since joining) (K)	Joined since
Imran Khan	Pakistan Tehreek-i-Insaf (PTI)	10.7	18	@ImranKhanPTI	6.2	March 2010
Maryam Nawaz Sharif	Pakistan Muslim League Noon (PMLN)	5.4	9.7	@MaryamNSharif	62	January 2012
Bilawal Bhutto Zardari	Pakistan People's Party (PPP)	3.6	1.9	@BBhuttoZardari	11.4	July 2011

情感分析sentiment analysis

- 情感分析又称观点挖掘或情感人工智能，是指计算语言学、自然语言处理和文本分析，通过分析提取、识别，并分析情感状态和主体信息。简而言之，就是对带有情感色彩的主观性文本进行分析、处理、归纳和推理，最后得出文本的二元化或者多元化极性分类的方法
- 情感分析主要采用三种方法:基于词汇的方法、基于机器学习的方法和混合方法。



语义词云semantic word cloud

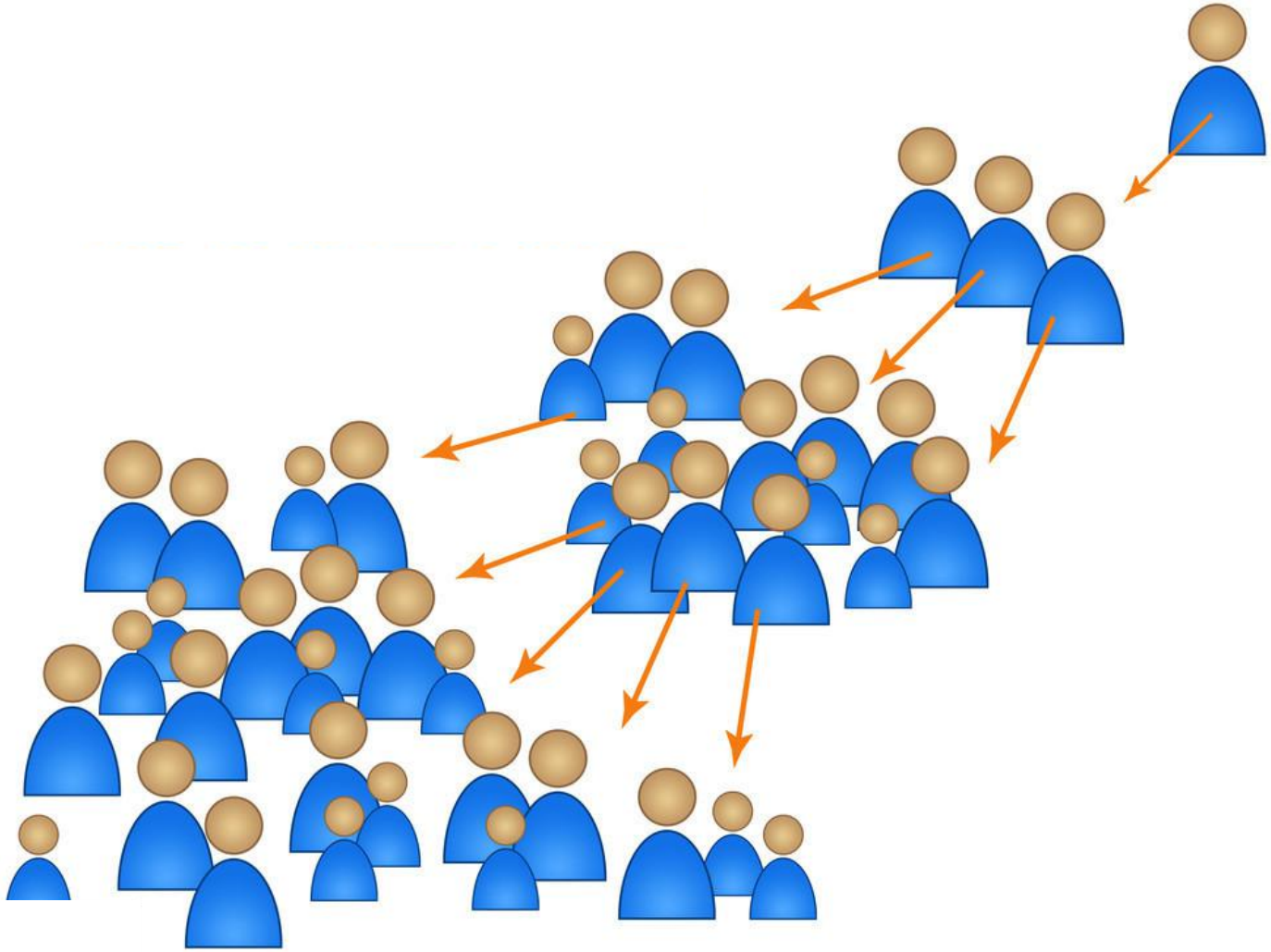
- 它描绘了有不同兴趣的用户推文上出现的不同词语的同时性和频率。
- 词云被用来直观地表示文本。它增加了文本分析的清晰度，单词频率很容易被看出。
- 所有的推文都首先经过预处理。它可以被平台研究人员用来理解用户的情绪、话题和讨论。

5、社交网络的应用

社交网络的应用

- 社交推荐
- 选举
- 恐怖分子挖掘
- ISIS邪教传播

- 社交推荐概念
- 社交推荐经典场景
- 社交推荐算法
- 社交推荐任务
- 社交推荐前沿
- 社交推荐挑战



- 推荐系统的出现早于社交网络，从亚马逊将其用于推荐商品，推荐系统一直在蓬勃发展。社交网络的推荐，我们常见的就是推荐好友，这是一种显性推荐。根据社交关系和社交行为进行的推荐属于隐性推荐，例如根据你微博的内容或者你好友的行为来给你推荐广告和商品。

- Location-Based Social Networks (LBSN)

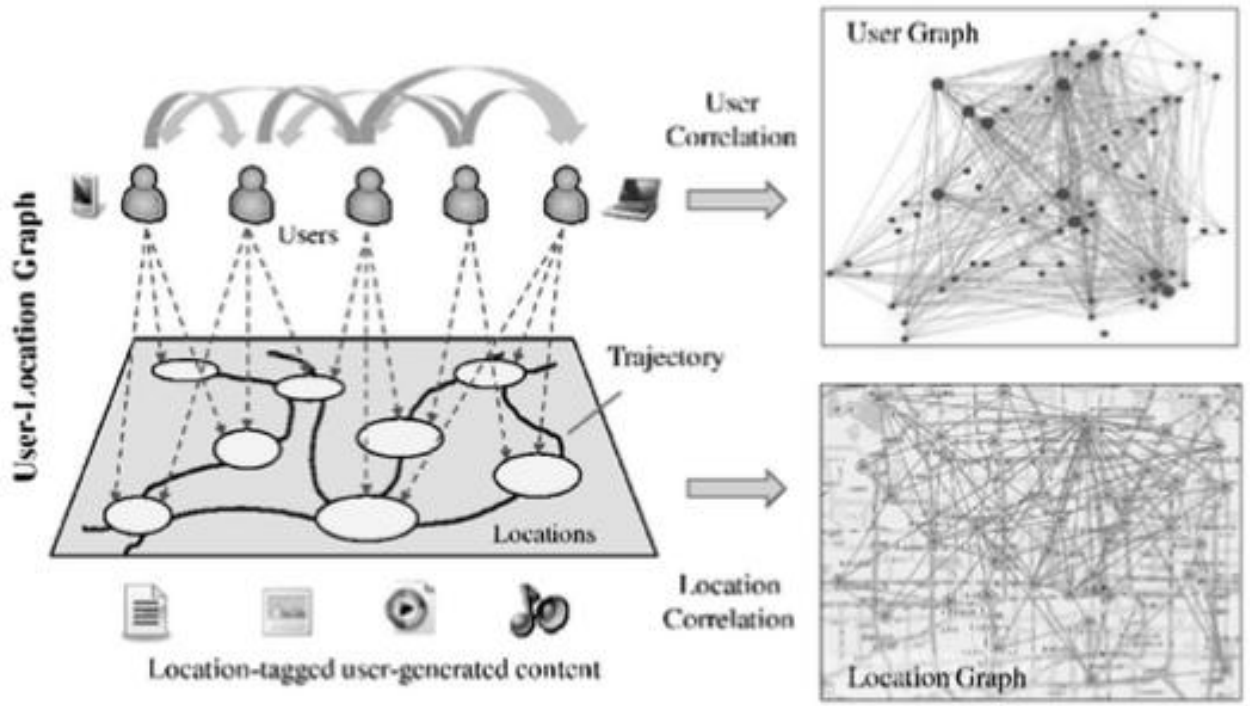


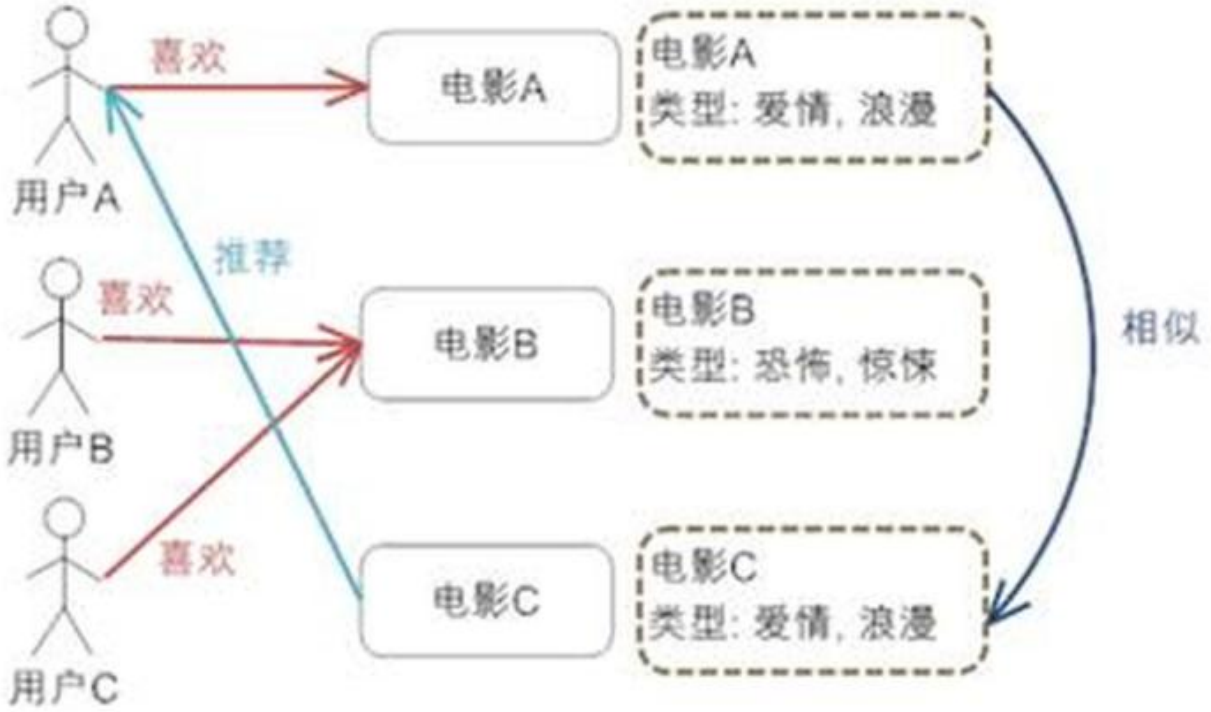
图1 LBSNs应用场景

图 1 是一个典型的 LBSNs 应用场景:用户之间有好友关系，用户可以在现实世界中的位置场点签到并产生带有位置标记的媒体内容。这样，我们便可以得到 LBSNs 中三种类型的图:

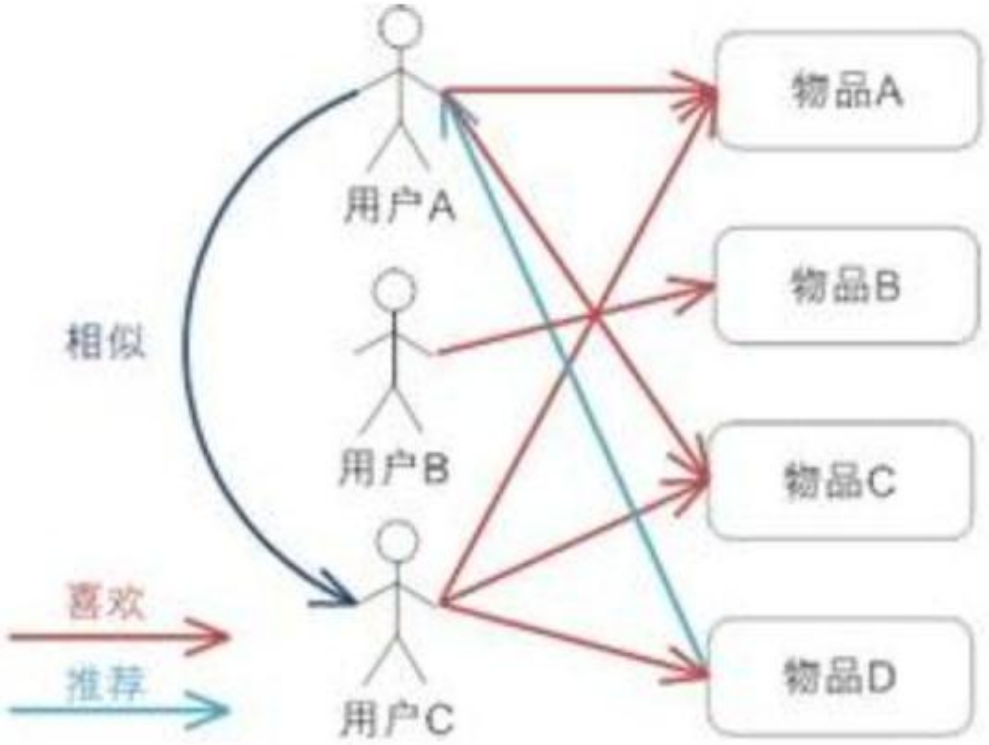
- 位置-位置图
- 用户-位置图
- 用户-用户图。

推荐系统算法	优势	劣势
基于内容的推荐算法	推荐结果单一简单	难以高效简单的向新用户推荐，分类器过多
协同过滤的推荐算法	高度个体化和系统化，处理非结构化对像，多样的推荐对象	不能快速扩展，数据集质量要求高，稀疏性
基于规则的推荐算法	快速发现	很难找到规则，同类产品相同处理，个性化低
基于效用的推荐算法	无稀疏问题，对偏好变化明显，无产品特性	依赖用户手动输入，静态选择
基于知识的推荐算法	要求产品形成映射，无产品属性	难以获取知识和知识结构，静态推荐

邻居模型、矩阵分解模型、基于蒙特卡洛的随机游走模型 ……


喜欢
电影A
电影A
类型：爱情，浪漫
用户A
推荐
喜欢
电影B
电影B
类型：恐怖，惊悚
相似
用户B
喜欢
电影C
电影C
类型：爱情，浪漫
用户C

基于内容


用户A
物品A
物品B
相似
用户B
物品C
喜欢
推荐
物品D
用户C

协同过滤

推荐内容和方法	用户推荐	好友推荐	直接匹配历史轨迹衡量用户相似度
			使用层次树图模型描述用户历史轨迹在空间和时间上的分布特征
			兴趣标签、历史轨迹信息、社交网络关系的三层好友关系模型
		旅行专家发现	(Hypertext-Induced Topic Search，HITS)算法计算经验值和流行度
	位置推荐	单个位置推荐	贝叶斯网络模型计算用户配置文件
			协同过滤(Collaborative Filtering，CF)方法为用户推荐场点
		行程推荐	协同过滤
			离线学习、在线询问
	活动推荐	提取位置-活动矩阵，集体矩阵分解	
		分解用户- 位置-活动三维张量	

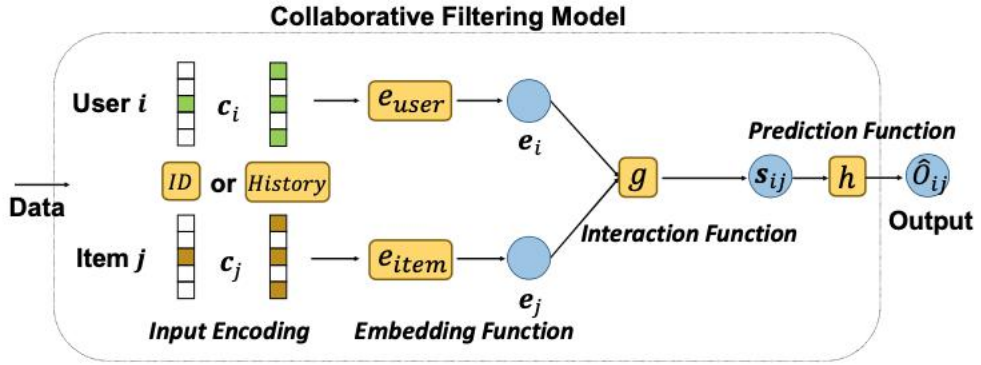


Figure 1: A unified framework of CF models, which contain four stages: input encoding, embedding function, interaction function and prediction function.

KEY WORDS

- Recommender System
- AUTO Machine Learning
- Collaborative Filtering

Algorithm 1: AutoCF: Automated Model Search for CF

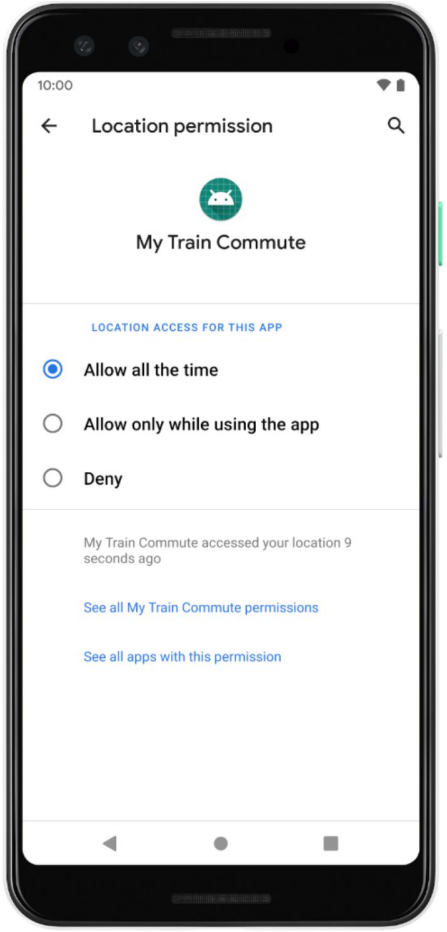
input : Search space $\mathcal{F}$, a learnable predictor $\mathcal{P}$, performance measurement $\mathcal{M}$, a empty set $\mathcal{H}$, size of training batch for predictor K_1 and K_2, training data $\mathcal{S}$.

Initialize the predictor $\mathcal{P}$ with random parameters;
do
  Randomly select a $(K_1 + K_2)$-size model set $\mathcal{F}^b$ from $\mathcal{F}$;
  Generate one-hot encodings $\mathbf{x}_o$ to represent models in $\mathcal{F}^b$;
  Estimate the performance of models in $\mathcal{F}^b$ with $\mathcal{P}$;
  Choose top-K_1 model sets $\mathcal{F}^t$ to train with $\mathcal{S}$;
  Evaluate the trained models in $\mathcal{F}^t$ with $\mathcal{M}$;
  Update the set of evaluated-model $\mathcal{H} \leftarrow \mathcal{H} \cup \{(f, \mathcal{M}(f)) | f \in \mathcal{F}^t\}$;
  Update $\mathcal{P}$ with records in $\mathcal{H}$ via loss function in (8).
while *not meet stop criteria*;
return desired CF models in $\mathcal{H}$.

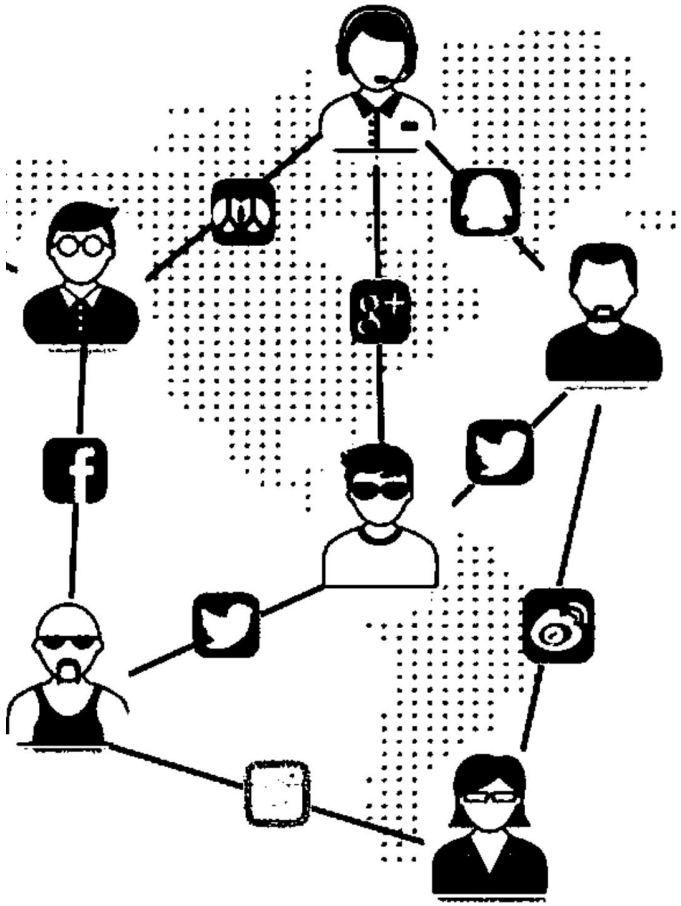
Chen Gao, Quanming Yao, Depeng Jin, and Yong Li. 2021. Efficient Data-specific Model Search for Collaborative Filtering. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21). Association for Computing Machinery, New York, NY, USA, 415–425. DOI:https://doi.org/10.1145/3447548.3467399

社交推荐挑战

- 准确的位置信息



- 多层异构网络结构



- 数据稀疏性

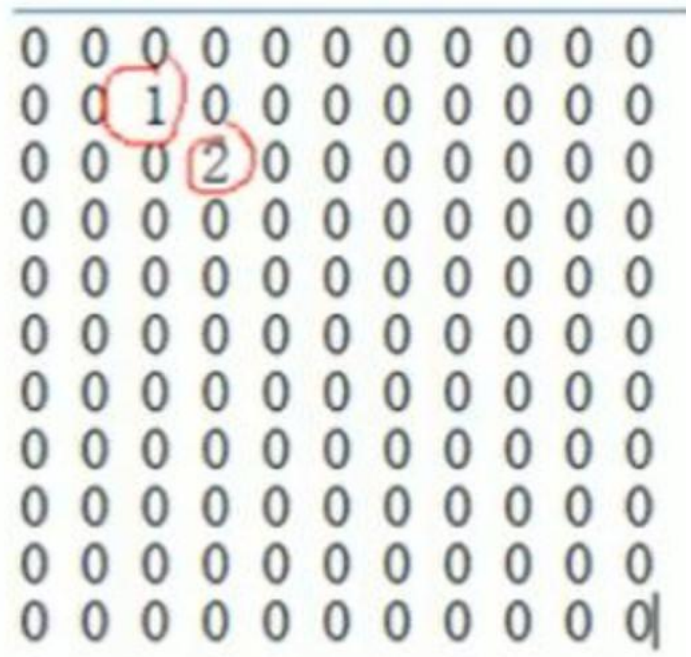
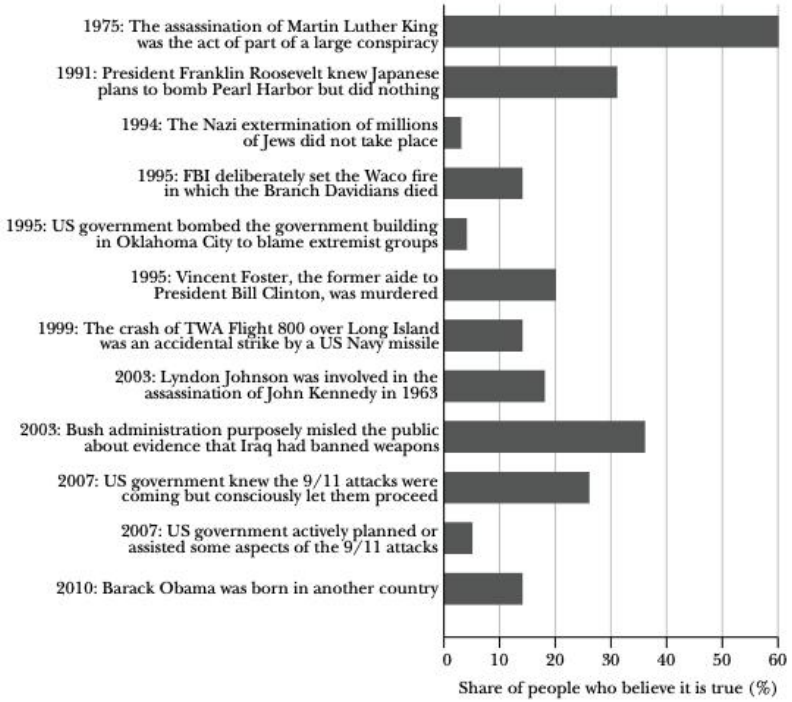


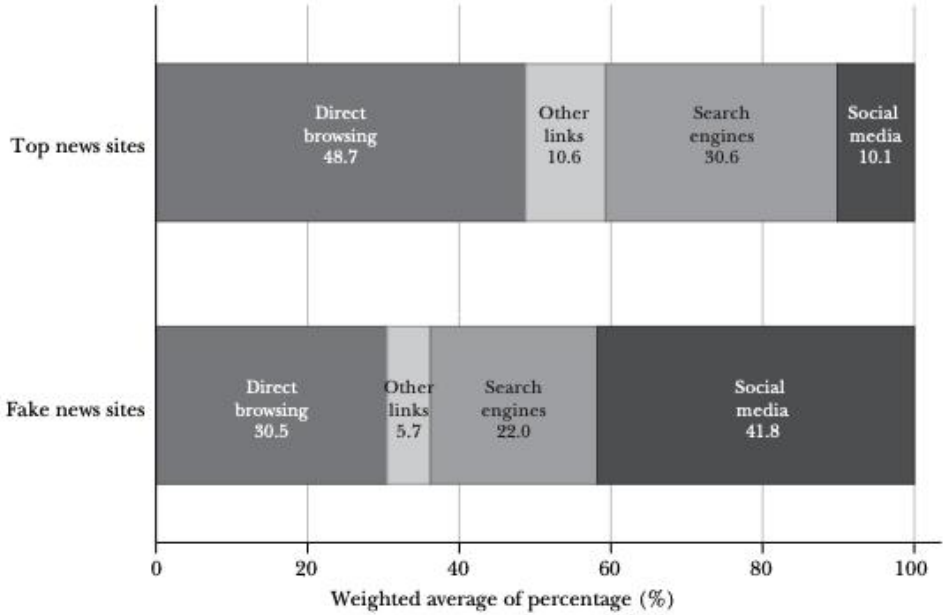
Figure 1

Share of Americans Believing Historical Partisan Conspiracy Theories



Note: From polling data compiled by the American Enterprise Institute (2013), we selected all conspiracy theories with political implications. This figure plots the share of people who report believing the statement listed, using opinion polls from the date listed.

Share of Visits to US News Websites by Source



Note: This figure presents the share of traffic from different sources for the top 690 US news websites and for 65 fake news websites. "Other links" means impressions that were referred from sources other than search engines and social media. "Direct browsing" means impressions that did not have a referral source. Sites are weighted by number of monthly visits. Data are from Alexa.

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives*, 31 (2): 211-36.DOI: 10.1257/jep.31.2.211

- 监控信息和情报?
- 监控资金来往?

ISIS邪教传播

- 平台：脸书、推特、油管、电报
- 策略：Fajr al-Bashaer应用程序、标签、
  邮件、博客
- 思路： 购买僵尸粉、 宣传和招募信息、ISIS行动的大量细节、美好世界描述、爱好共情（Islamic State of Cat，经常发布可爱的猫咪照片，深得用户欢心，进而传播ISIS的极端思想)

Global Terrorism and Social Media: A Study of ISIS

BY

DEEPENDRA CHHETRI

M.Phil

ROLL NO. 16MPPLO4

SUPERVISOR- AMIT KUMAR GUPTA

DEPARTMENT OF POLITICAL SCIENCE

SCHOOL OF SOCIAL SCIENCES

SIKKIM UNIVERSITY

YEAR 2018

6、社交网络前沿方向

Tasks:

Graph construction, reconstruction, network inference, graph identification
Sparsification, sketching, and compression of network data
Subgraph and motif discovery
Influence propagation, information diffusion, spreading and epidemics
Link prediction
Graph summarization and visual analytics
Succinct data structures for network-related data
Network representation learning and graph embeddings
Reinforcement learning and advanced machine learning for graphs
Graph neural networks
Self-supervised learning on graphs
Pre-trained models and zero-shot learning for networked data
Causal inference in relational data
Querying and indexing algorithms for massive graphs

Ethical impacts of algorithms:

Privacy-preserving graph algorithms
Explainable graph algorithms
Fairness, bias, and transparency of graph mining and learning algorithms
Adversarial attacks on network algorithms and graph neural networks

Data types:

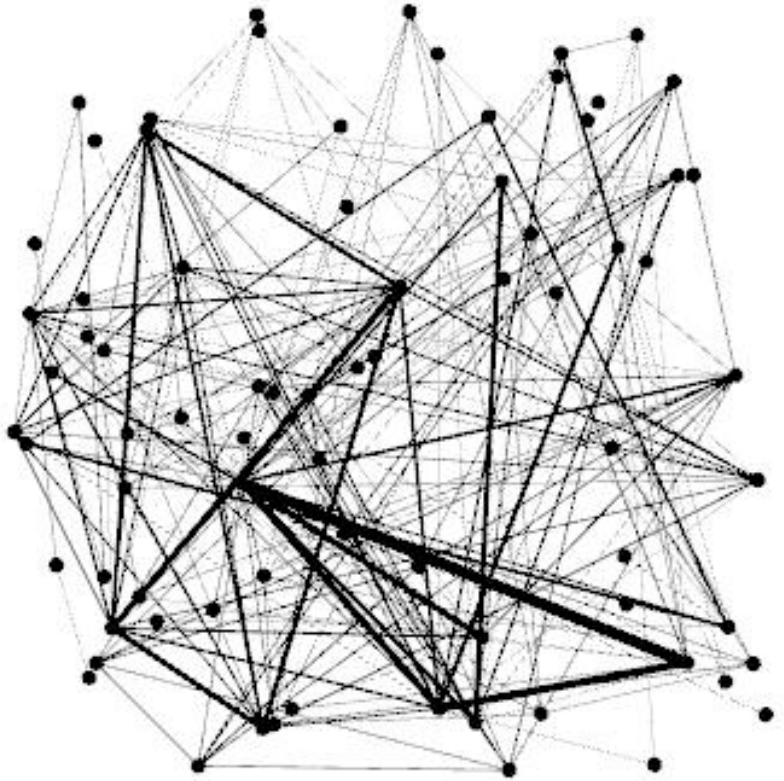
Analysis of heterogeneous, signed, attributed, and labeled networks
Dynamic network analysis and algorithms for graph streams
Multi-relational graph analysis
Higher-order graph and network algorithms
Knowledge graph mining and learning-based reasoning
Mining and learning in graphs with missing information and noise

Applications:

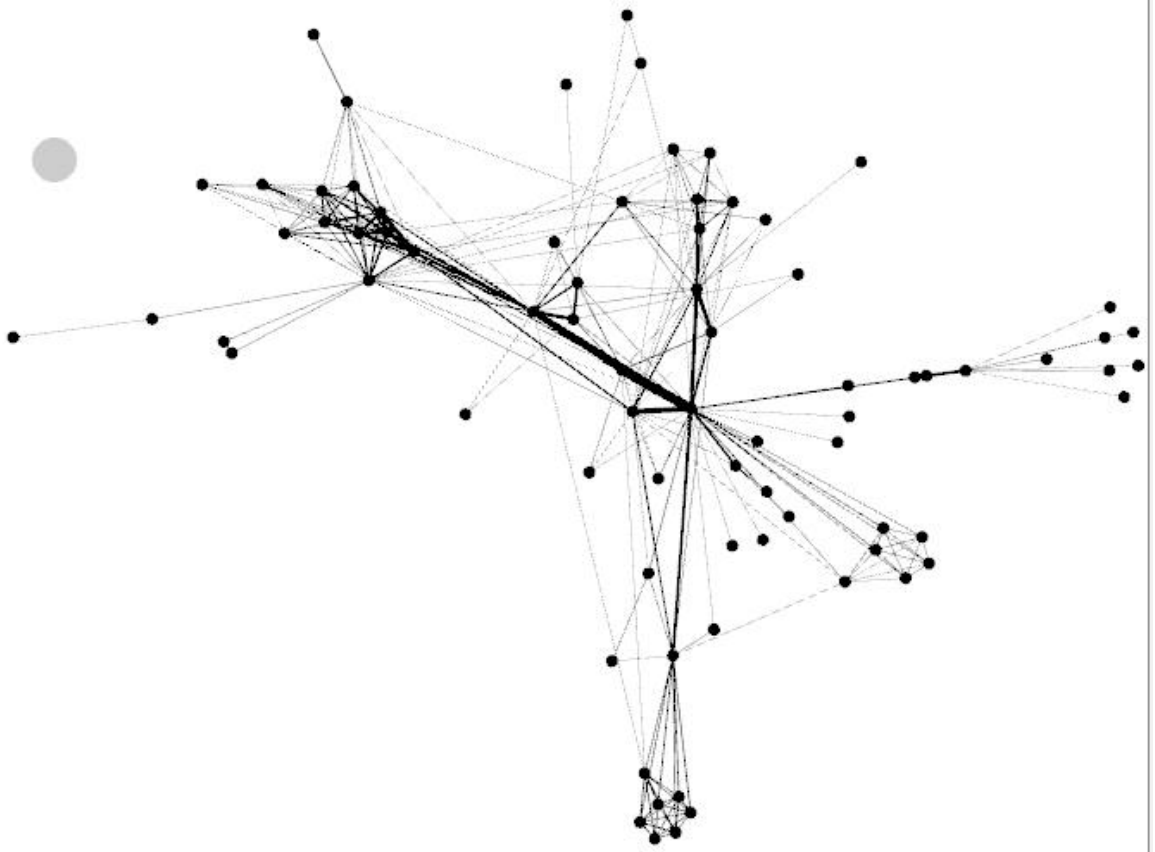
Social media analysis through the lenses of networks
Social mining, social search, and social recommendation systems
Social reputation and trust management
Game theoretic and economic aspects on graphs and networks
Detecting, understanding, and combating misinformation and fake news
Fraud, spam, and malice detection in relational domains
Web-based applications of graph mining (e.g., in economics, sociology)

7、demo展示

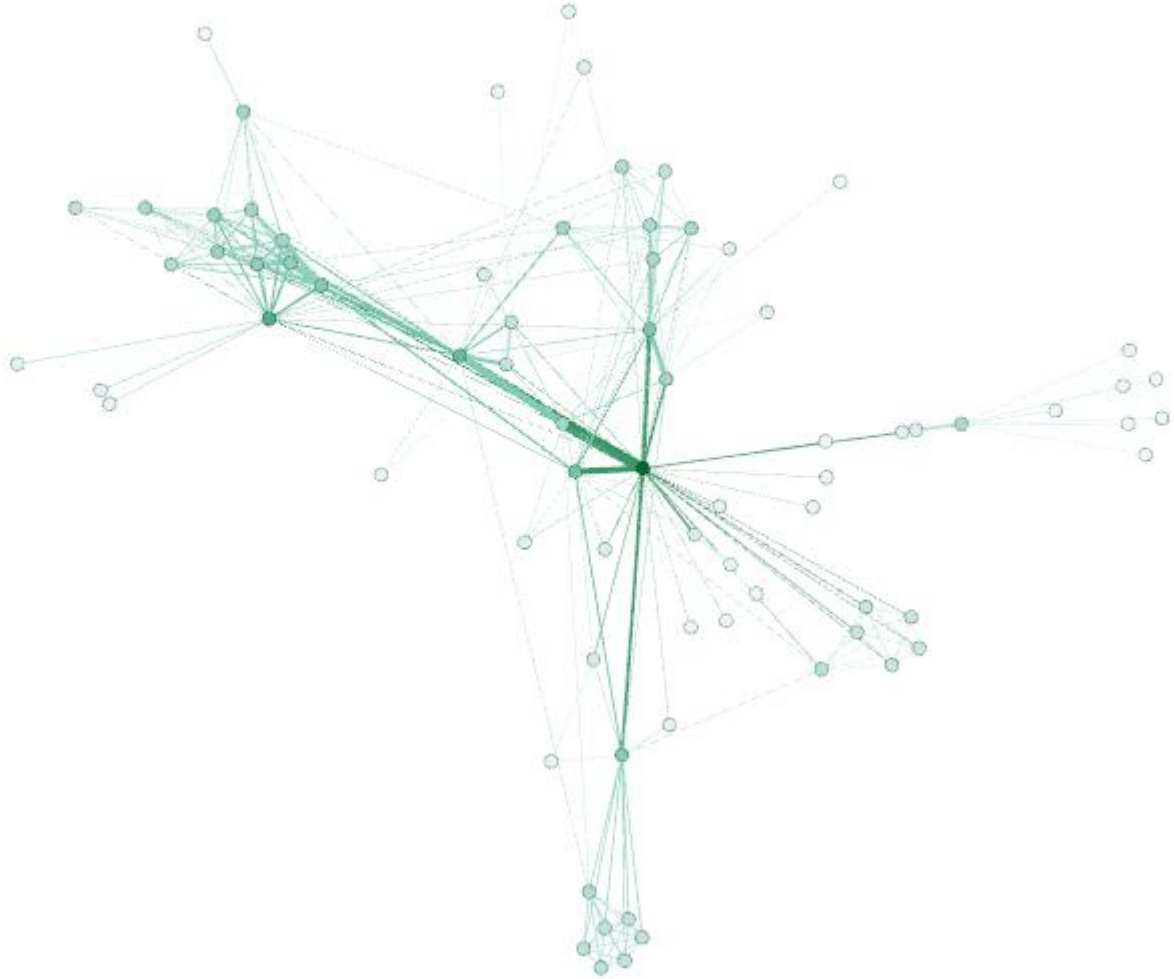
通过可视化的方式分析右图的社交网络性质



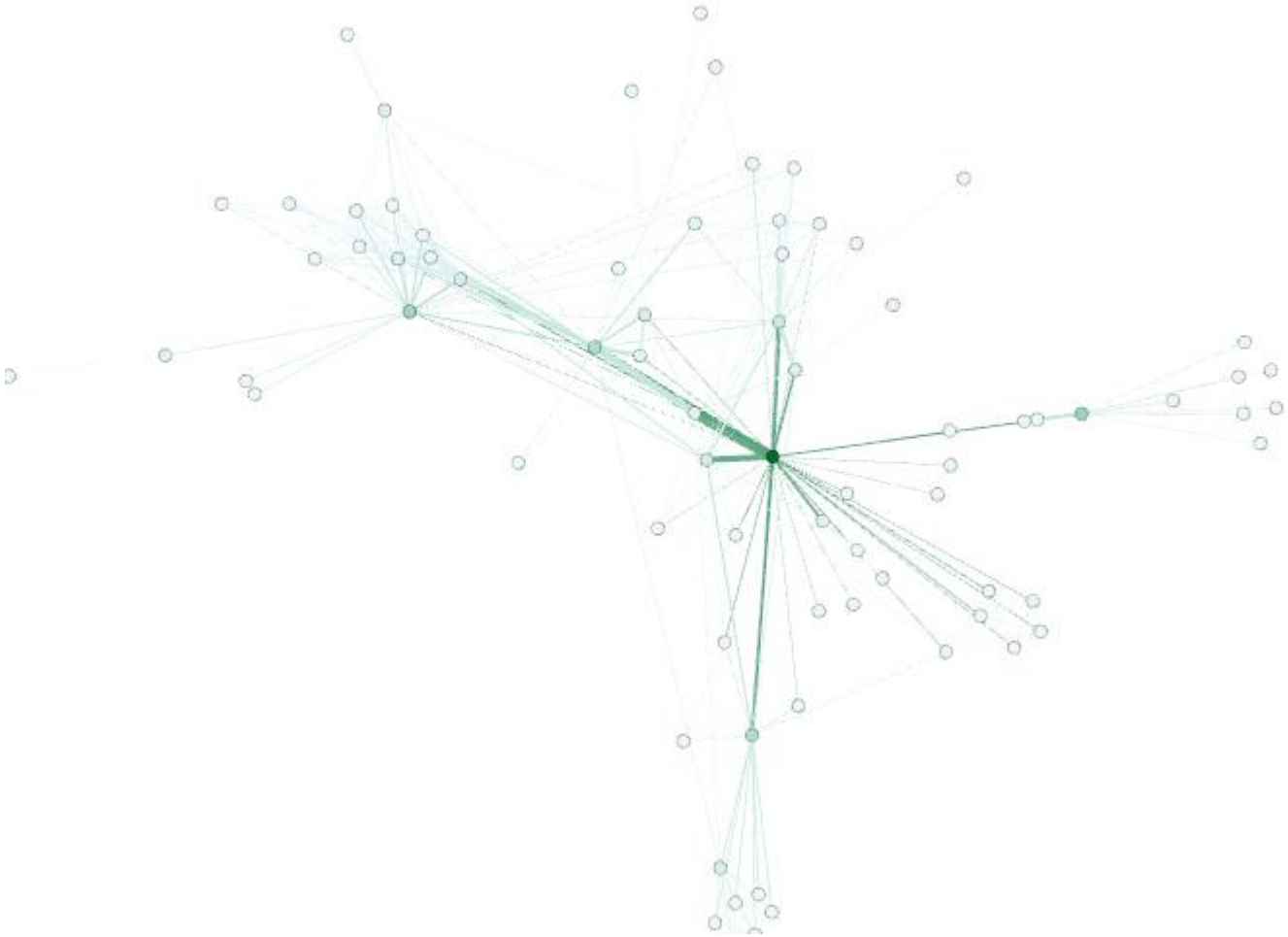
通过引导力模型布局结果



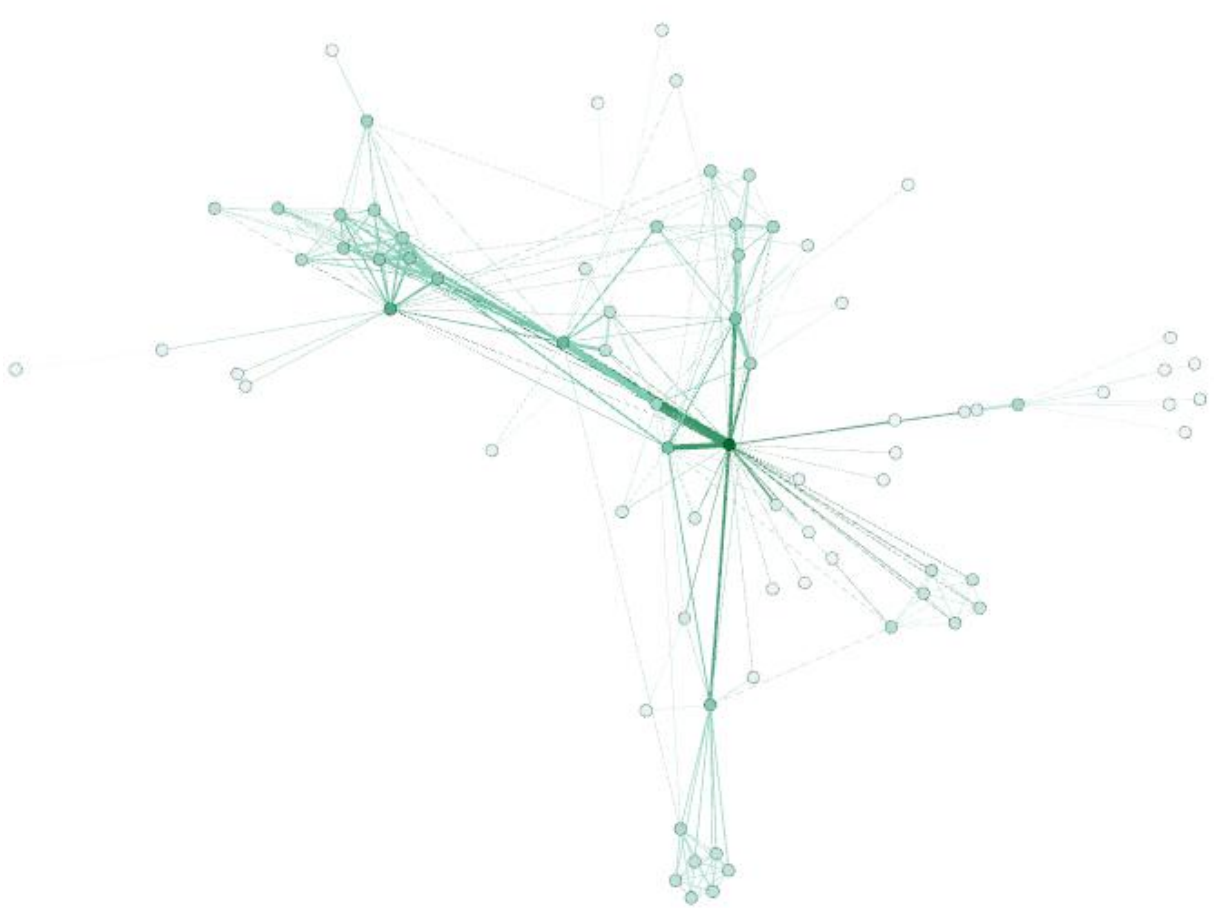
按照度数给节点染色，颜色越深度数越大

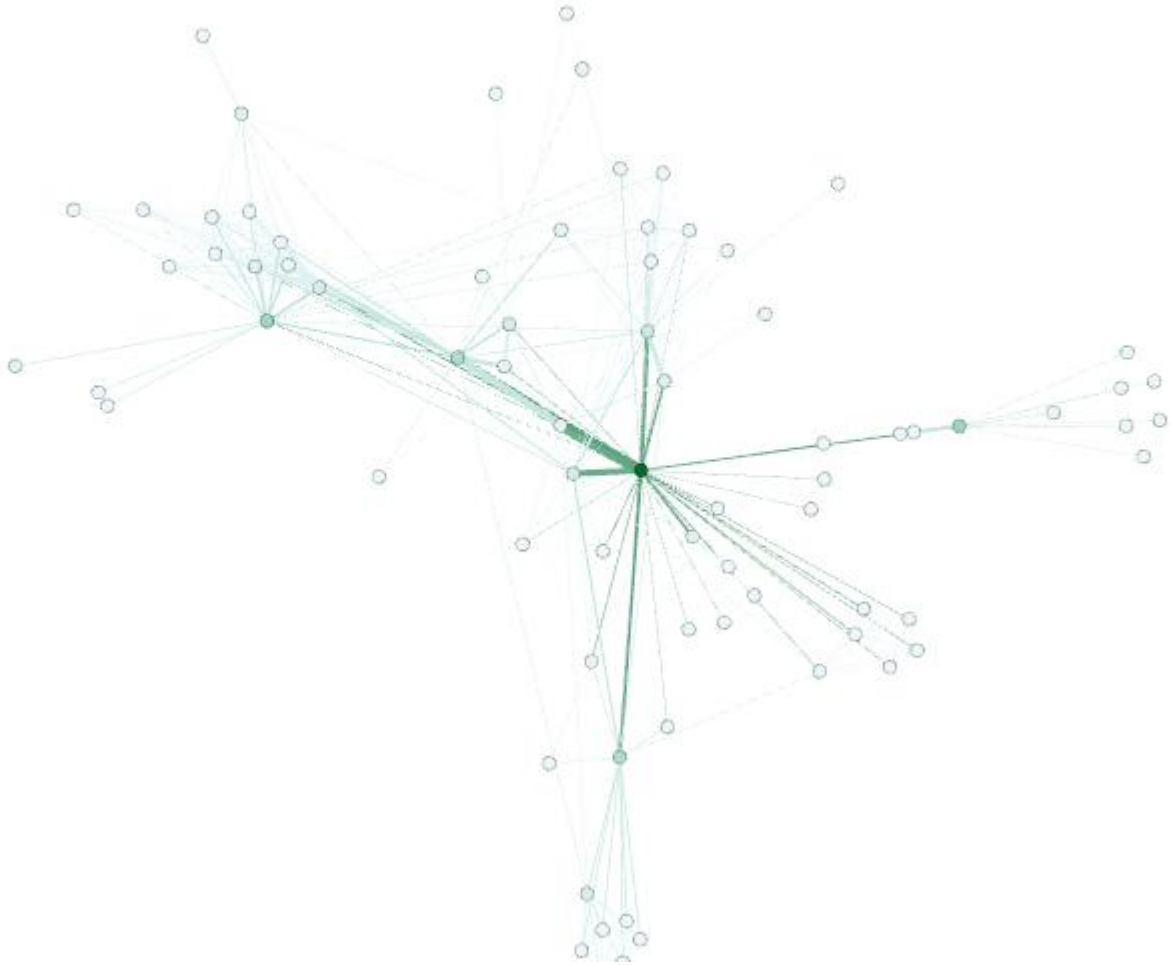


按照介数中心性给节点染色，颜色越深度数越大

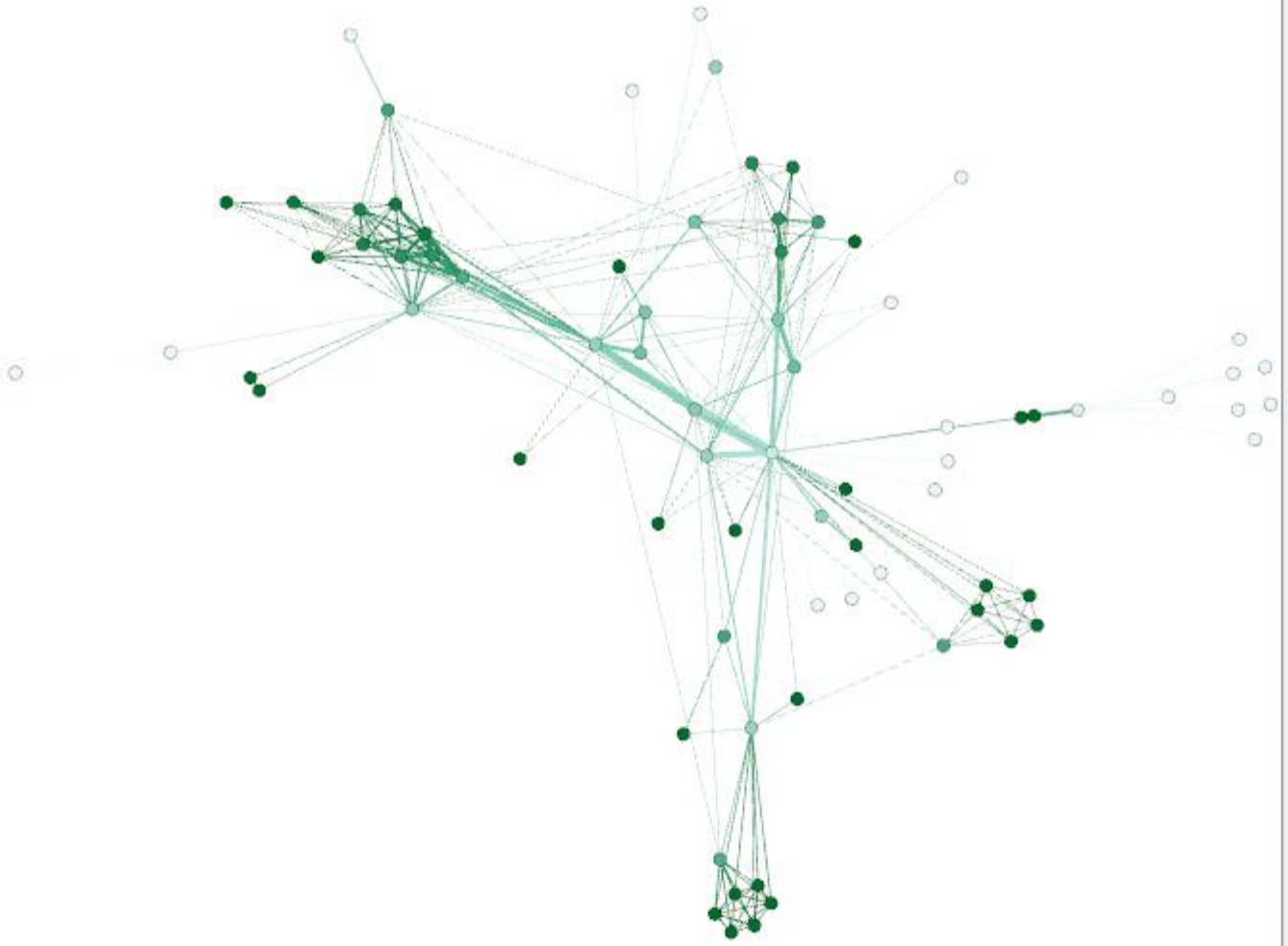


社交网络分析-Demo-可视化节点运作方式





按照聚类系数给节点染色，颜色越深度数越大



德以明理　学以精工