微信复杂网络与应用

Randyling（凌国惠） 2016/04

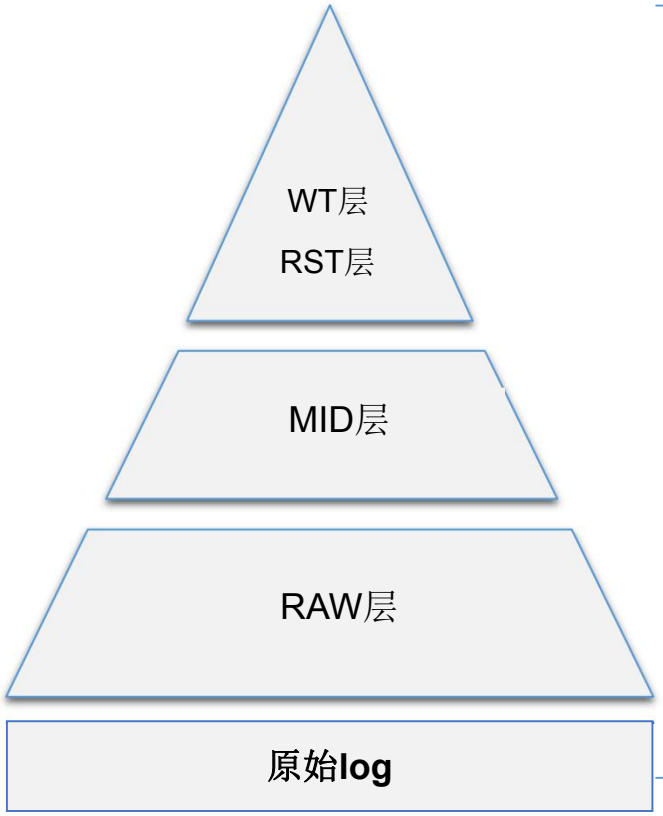
Question

对腾讯来说，什么数据最珍贵？

来自业务的需求

- 喜欢听周董歌曲的用户
- 有送儿童节礼物需要的用户
- 高端，守信用的用户

挖掘过程：数据体系&用户画像



基础属性： 年龄、性别、地域.....

行为： 朋友圈活跃度、文章阅读.....

状态： 育儿、在校、旅行.....

兴趣： 金融、互联网、汽车.....

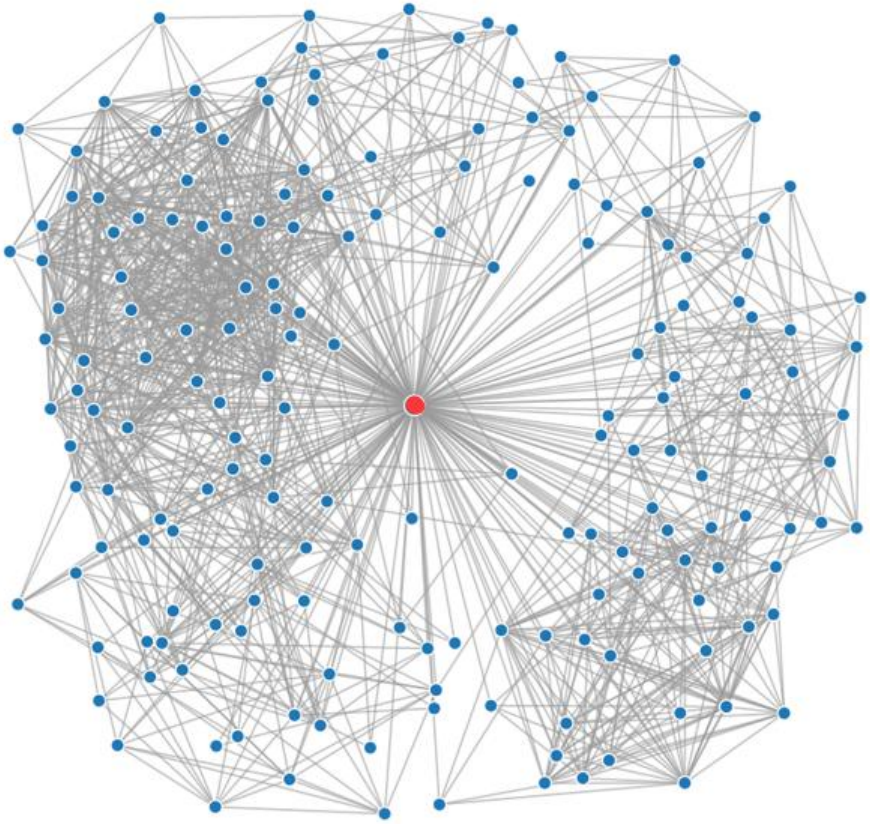
其它.....

量的问题

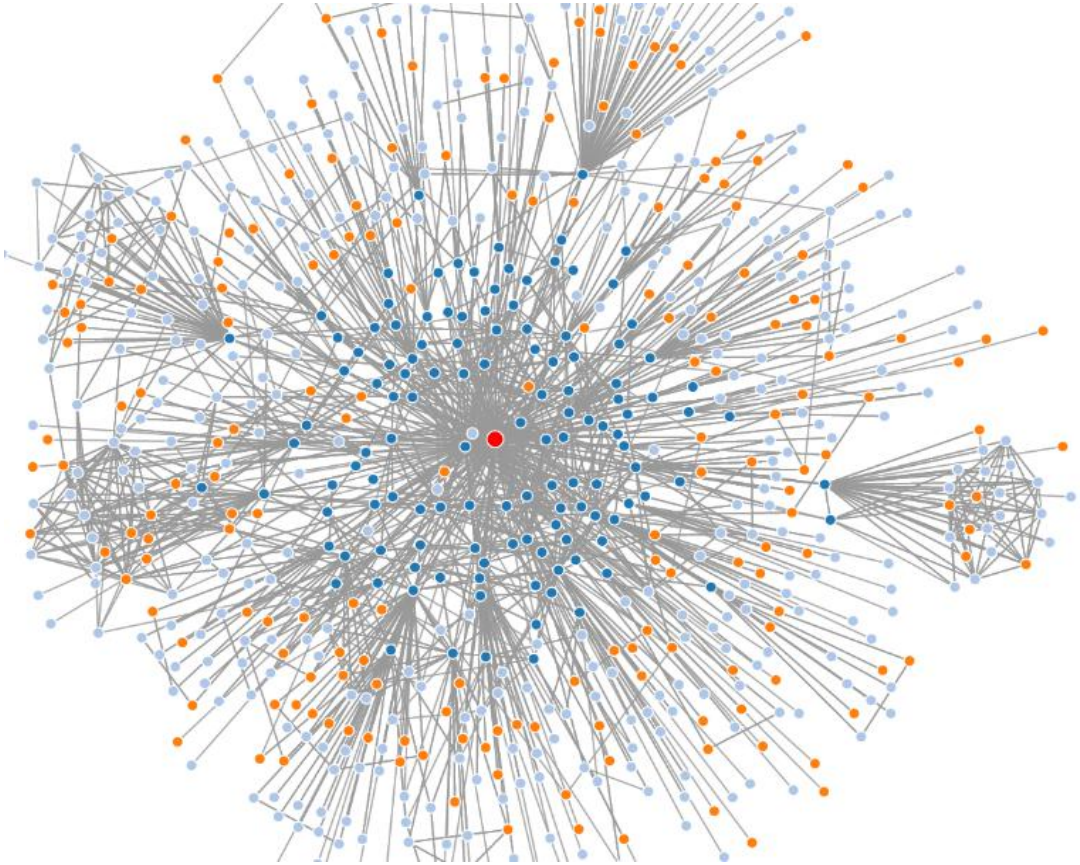
纳尼？才一百万用户？我要二十倍！

分析共同点：社交性质

微信一度好友



微信社交网络（局部图）



社团识别算法简介

算法名称	时间复杂度	实现原理
GN	$O(m^2n)$	基于边介数进行分割
LPA	$O(m)$	基于邻居信息进行迭代决策
CNM	$O(mdlogn)$	基于模块度增益进行的贪心算法
K-派系	$O(dn3^{d/3})$	基于最大团的邻接矩阵

K-派系算法

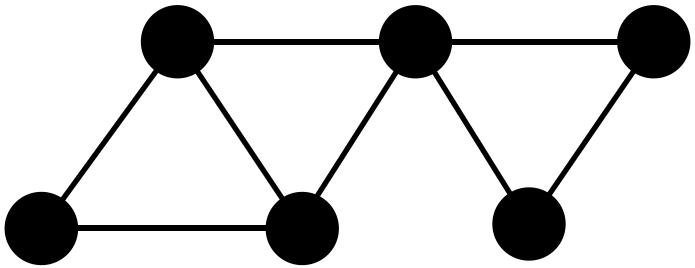
相邻的 k-cliques

- 两个 k-cliques 有共同的 **k-1** 节点

k-clique 社区

- k-cliques 的集合: 集合中的 clique 可以通过一系列相邻的 k-clique 相互到达

k = 3



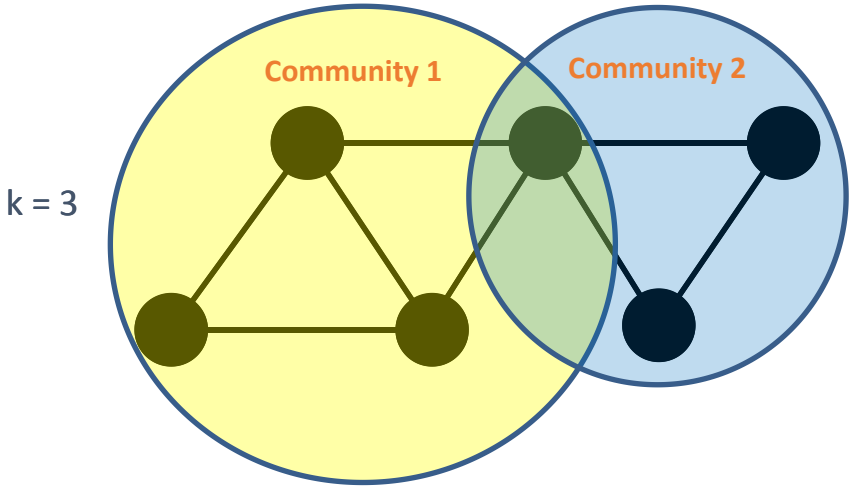
K-派系算法

相邻的 k-cliques

- 两个 k-cliques 有共同的 k-1 节点

k-clique 社区

- k-cliques 的集合: 集合中的 clique 可以通过一系列相邻的 k-clique 相互到达



CNM算法(Clauset、Newman、Moore)

1. 模块度 Q 的定义:

给定一个实际网络，假设找到一种社团划分，那么所有社团内部边数的总和为:

$$Q_{real} = \frac{1}{2}\sum_{ij} a_{ij}\delta(C_i, C_j)$$

其中$\boldsymbol{A} = (a_{ij})$为实际网络的邻接矩阵

C_i 和 C_j 表示节点i 和j所属的社团; 如果i 和 j 属于同一社团则δ为 1；否则 δ 取值为 0

引入一个相同规模的零模型，和 $\boldsymbol{A}$ 有相同的度序列,其他随机。用相同的社团划分有:

$$Q_{null} = \frac{1}{2}\sum_{ij} p_{ij}\delta(C_i, C_j)$$

其中p_{ij}是零模型网络中节点i 和节点j之间连边数的期望值.

网络的一种社团划分的模块度定义为:

$$Q = \frac{Q_{real} - Q_{null}}{M} = \frac{1}{2M}\sum_{ij}(a_{ij} - p_{ij})\delta(C_i, C_j)$$

其中M为网络的边数，$p_{ij} = \frac{k_i k_j}{2M}$，$k_i$和$k_j$ 分别为节点i 和j的度

CNM算法(Clauset、Newman、Moore)

2. cnm 算法初始时每个节点都表示一个社团，初始化增量矩阵 ΔQ 和辅助向量 $\boldsymbol{a}$：

$$a_i = \frac{k_i}{2m} \quad ; \quad \Delta Q_{ij} = \begin{cases} \frac{1}{2m} - \frac{k_i k_j}{(2m)^2} & if\ i,j\ are\ connected \\ 0 & otherwrise \end{cases}$$

3. 选取最大的 ΔQ_{ij} ，执行社团 $i \rightarrow j$的合并，合并后模块度为： $Q = Q + \Delta Q_{ij}$ ；同时更新 ΔQ 和 $\boldsymbol{a}$ ：

$$\Delta Q'_{jk} = \begin{cases} \Delta Q_{ik} + \Delta Q_{jk} & k \text{ is connected to both } i \text{ and } j \\ \Delta Q_{ik} - 2a_j a_k & k \text{ is connected to } i \text{ but not } j \\ \Delta Q_{jk} - 2a_i a_k & k \text{ is connected to j but not } i \end{cases}$$

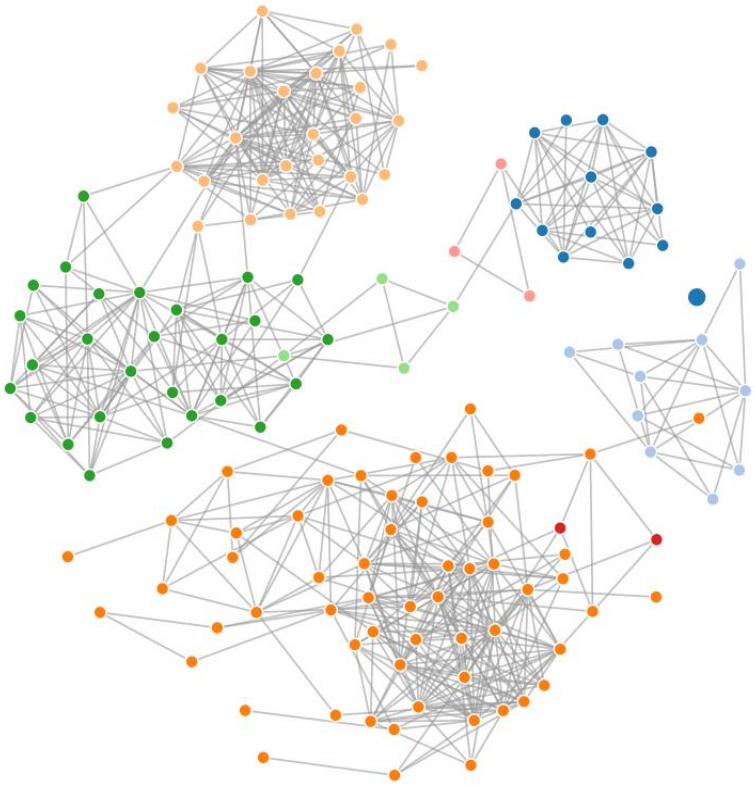
$$a'_j = a_j + a_i \ ; \ a_i = 0$$

4. 重复 3 直至 Q 出现首次下降，ΔQ 中所有元素都为负值，记录 Q_{max} ，此时社团合并的状态 便为 CNM 算法的社团划分结果;

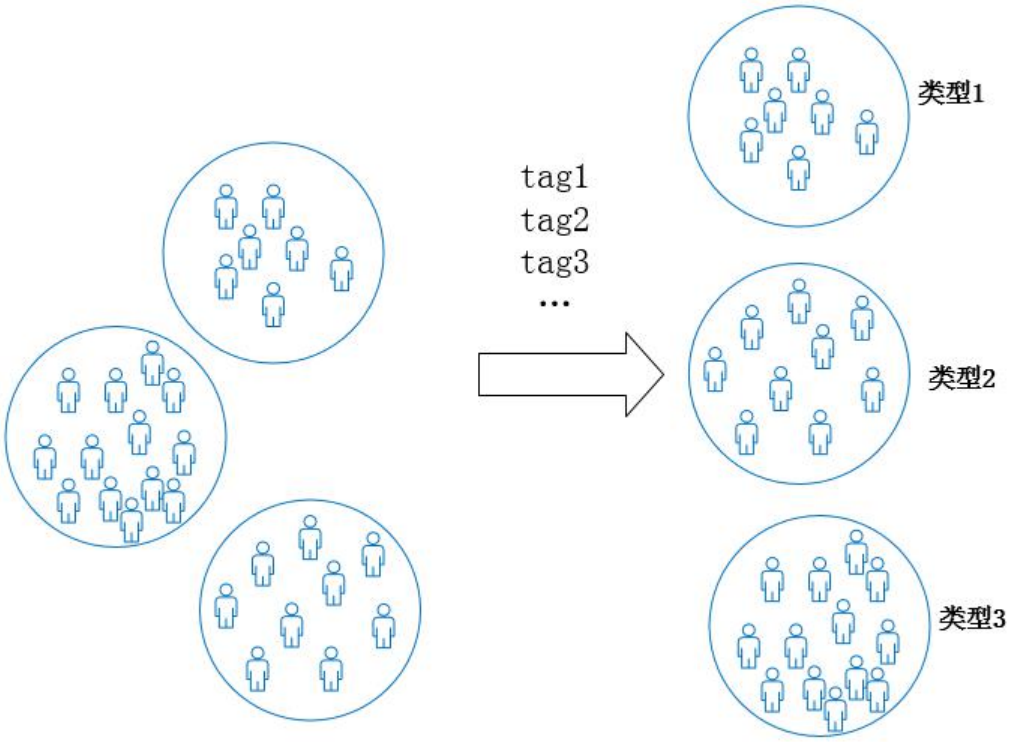
社团识别算法优化

- 算法差异
  - K派系精准，复杂度高
  - CNM效果差些，复杂度低些
- 存在问题
  - 不是同一个圈被拉进去
  - 该合并的没有被合并
- 深度优化
  - 综合运用各种算法
  - 叠加关系链之外的数据
  - 特定条件下合并

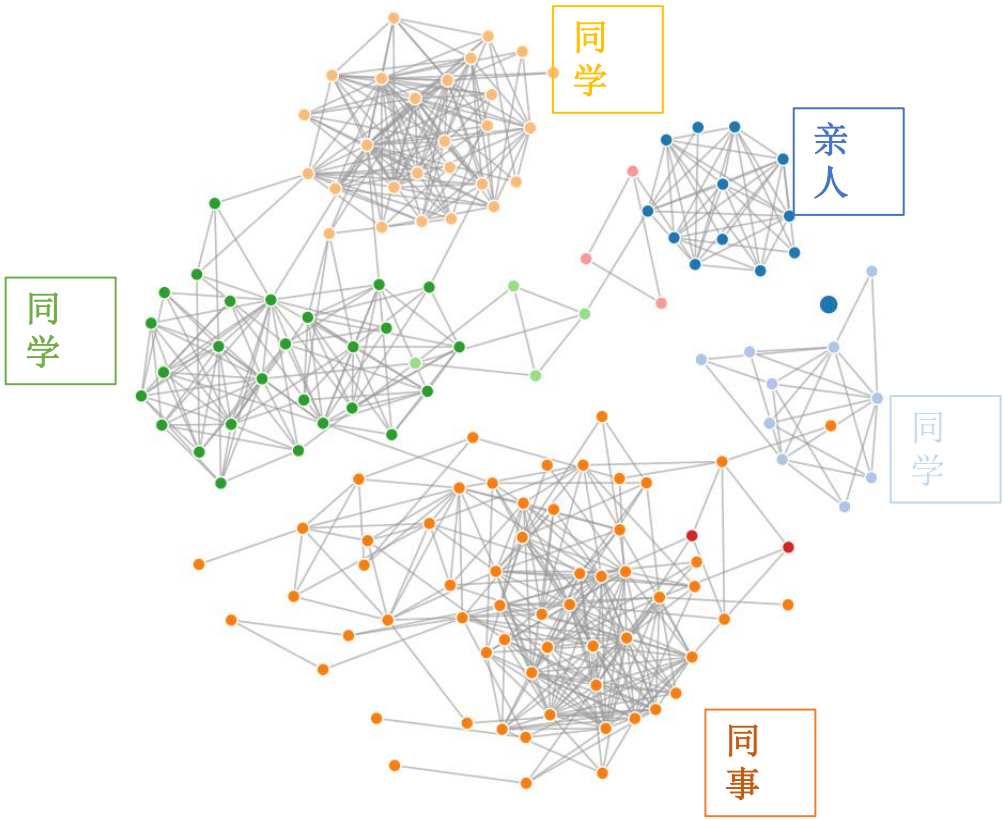
社团识别结果



社团分类模型



社团分类

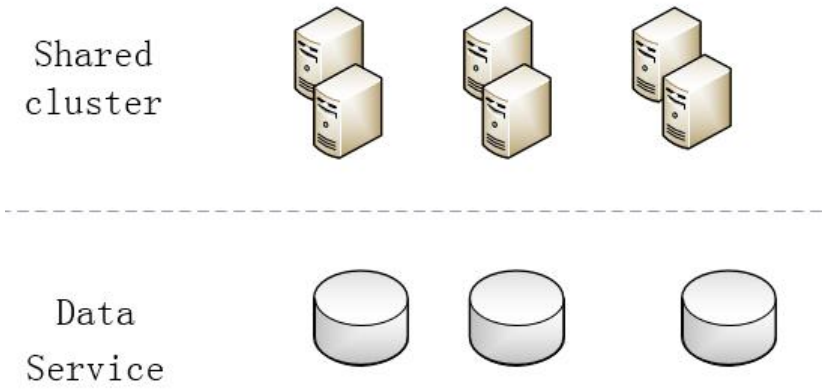


计算难点

- 规模大
  - 点：6.97亿+
  - 边：千亿+
- 复杂度高

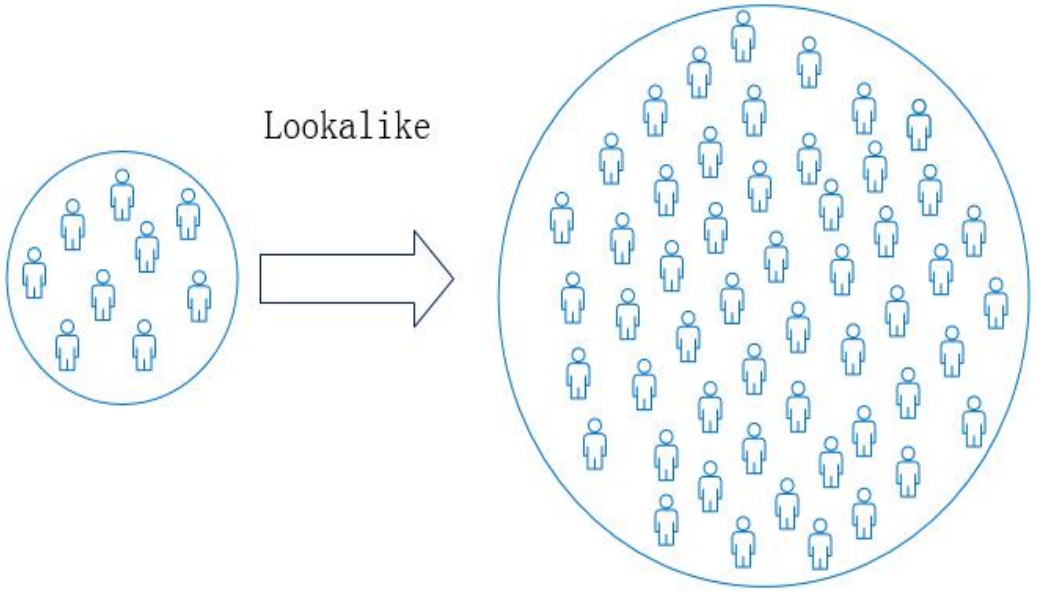
计算框架——微信资源调度系统

- 微信后台强大
  的服务器集群
- 空闲时跑挖掘计算
- 白天是你的，
  晚上是我的

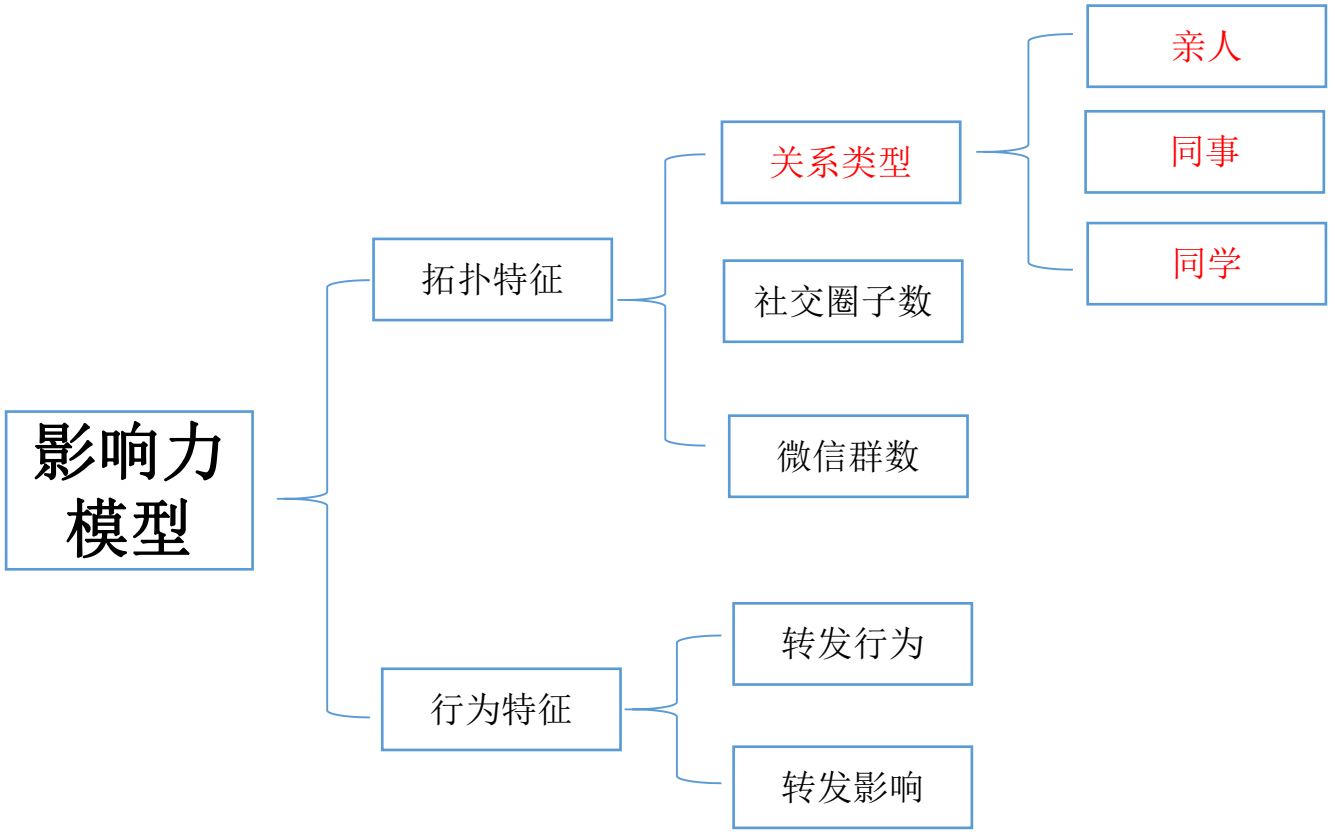


复杂网络应用——广告

- 10多倍扩散
  效果依然很好
- 任意标签社交
  Lookalike
- 动态投放&闭环处理

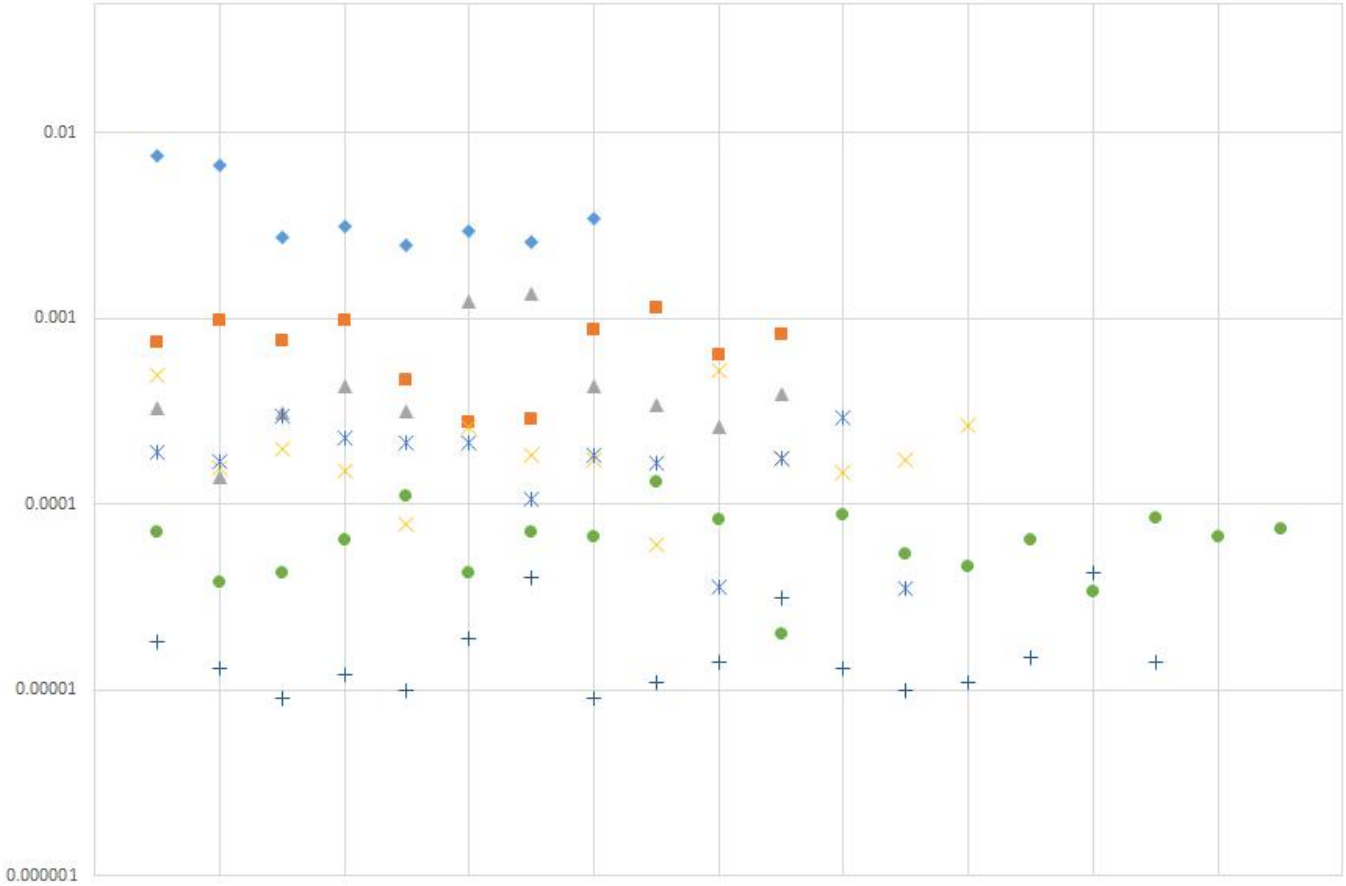


复杂网络应用——社交模型



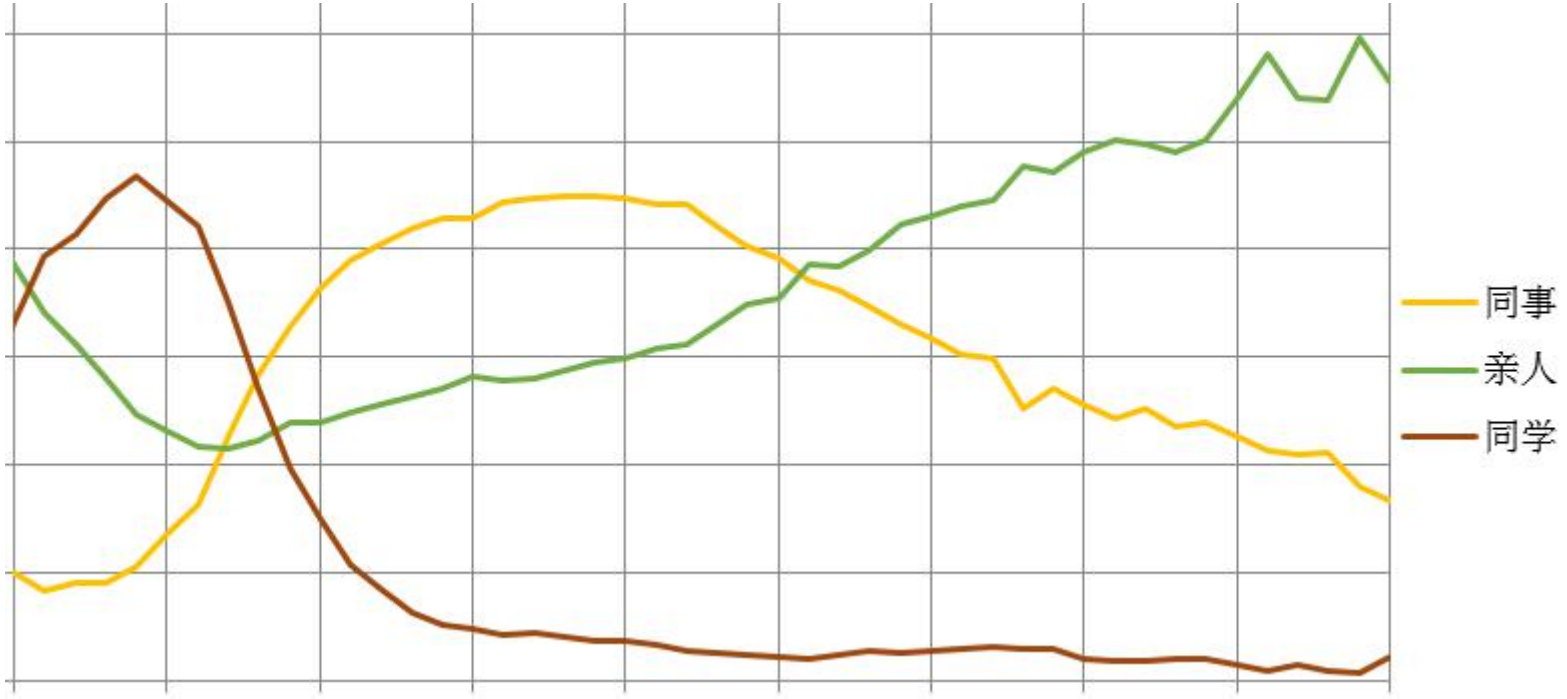
复杂网络应用——People Rank

- 网络+业务 → 用户分层



复杂网络应用——分析

- 微信人生——各关系类型随年龄变化



复杂网络应用——其它

- 征信、游戏、搜索、推荐.....

深入方向

- 社会层次关系
- 全局计算
- 更多业务应用

谢谢！