社交网络大数据建模的框架探索

腾讯公司 社交网络运营部 数据中心

岳亚丁（博士）， 2014-10-23

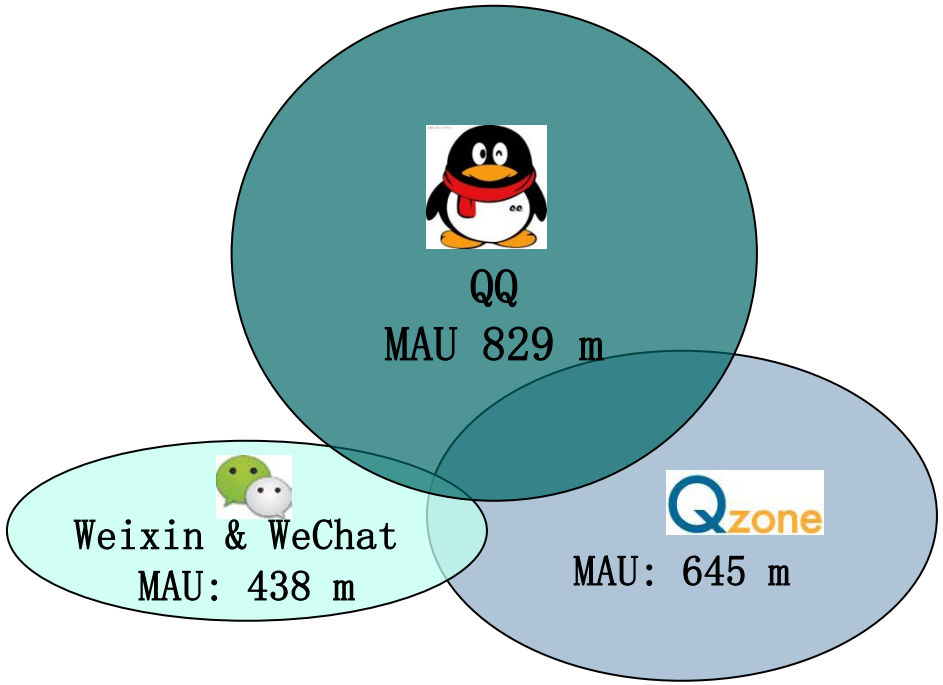
Outline

- 腾讯社交网络的研究内容
- 遇到的问题、解决思路
- 模型框架
- 展望

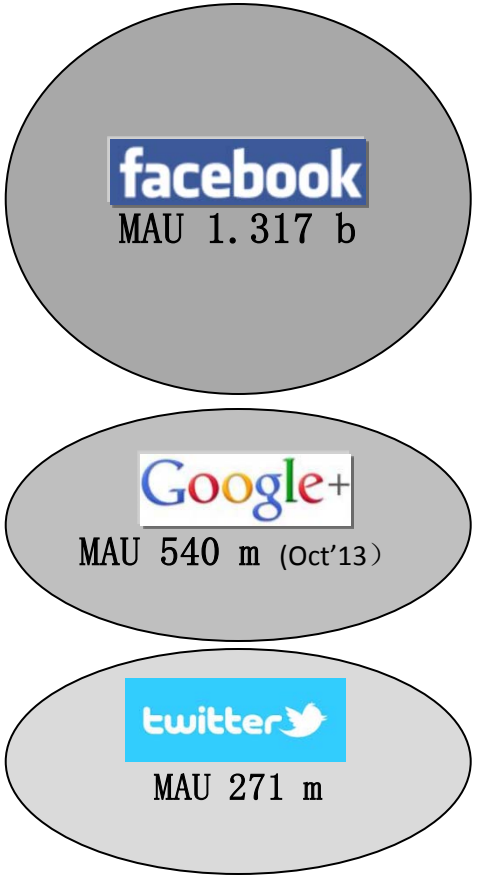
Outline

- 腾讯社交网络的研究内容
- 遇到的问题、解决思路
- 模型框架
- 展望

社交平台规模：2014 Q2



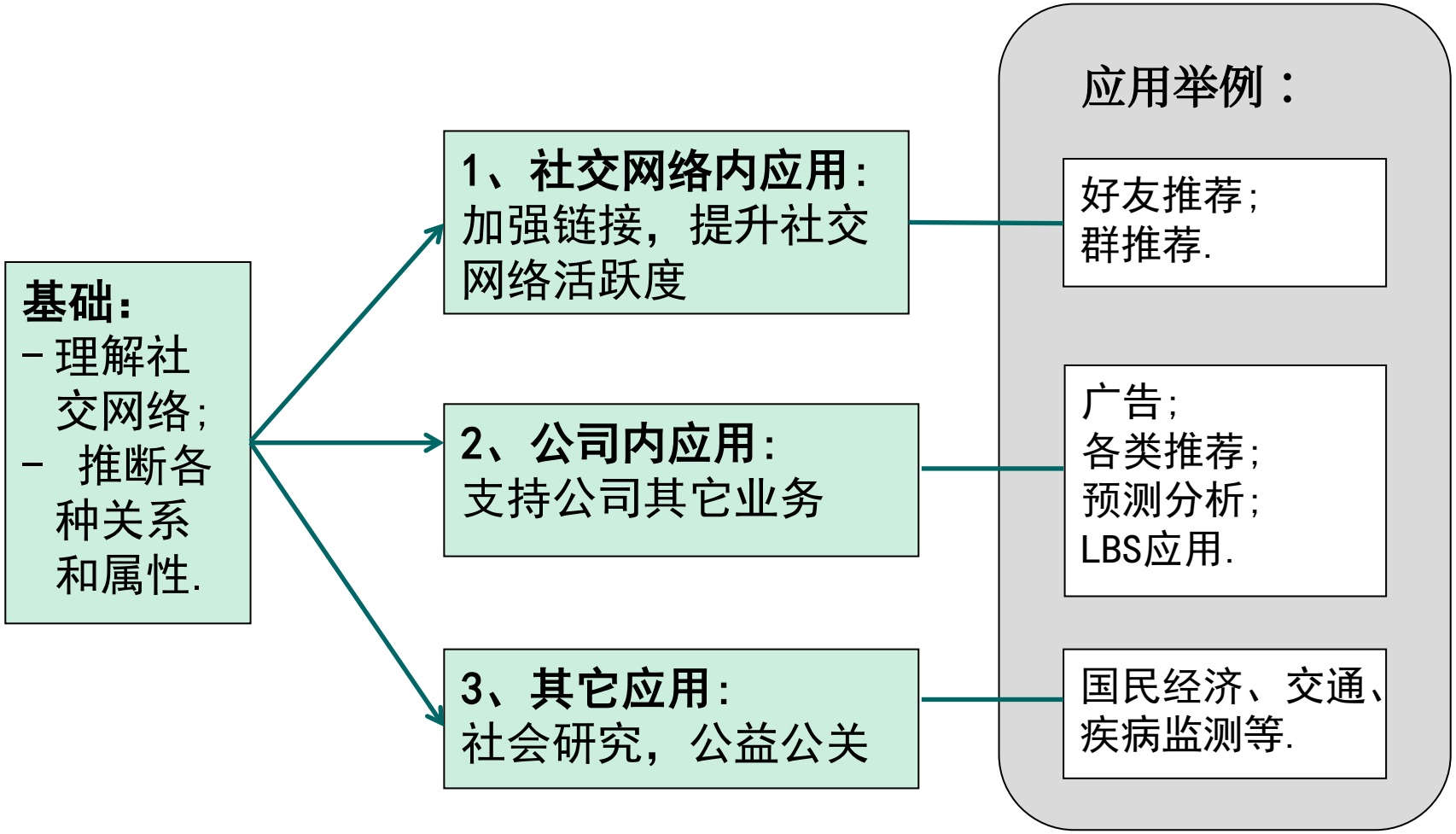
MAU: Monthly active user accounts



Data sources:
http://www.tencent.com/zh-cn/content/at/2014/attachments/20140813.pdf
http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/
http://en.wikipedia.org/wiki/Google%2B
http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

研究方向：基础 + 应用

目的：帮助用户高效地社交，并支持人、信息、实物之间的高效流动。



基础研究

用户的基础属性及行为特征

- 年龄、兴趣…
- 多QQ_uin ≡ 1人
- 常去地点及其属性
- IP所在的城市、区县、街道
- “小数据”（用户自身行为数据汇总）

用户和群组的关系结构及其属性

- 关系链性质：同学、同事、家人等
- 实体圈划分：线下实体组织
- Location-based Social Network (LBSN)：
  用户网络
  + 地点网络

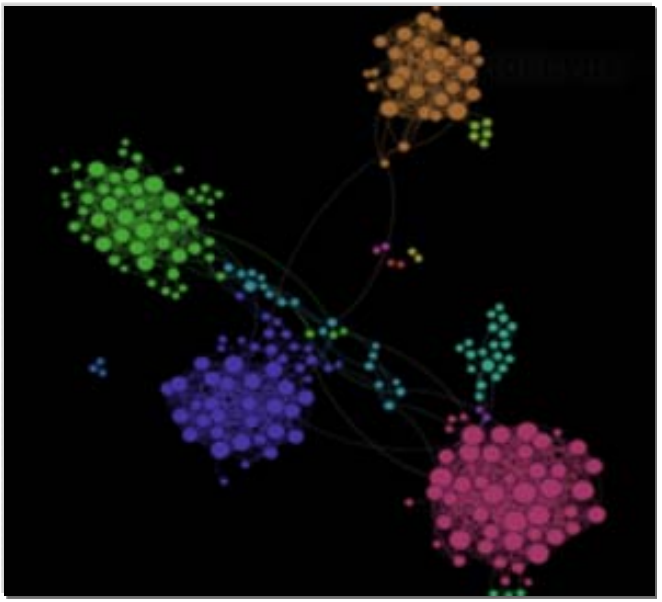
社会化影响和传播

- 群体趋同性（group homophily）
- 信息、行为、兴趣的影响和传播
- 推断群组的主题（炒股、育儿、旅游、羽毛球…）

QQ圈子

- 功能

  线上好友、潜在好友（好友的好友）
  → 现实生活中的关系
  （例如：找到多年不联系的同学）



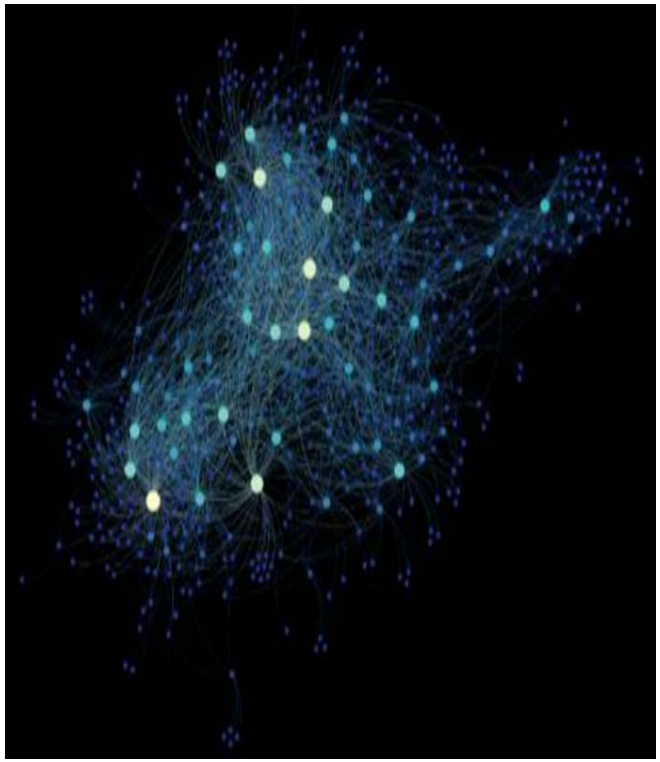
- 算法：

  圈内紧密、圈外松散
  对圈子以及圈友的类型进行识别

- 用途：

  找回老朋友，认识新朋友，…

关系链类型判断



- 好友关系类型：
  - 亲戚？
  - 同事？
  - 同学？
  - 一般性朋友？
  - 合作伙伴？
  - …

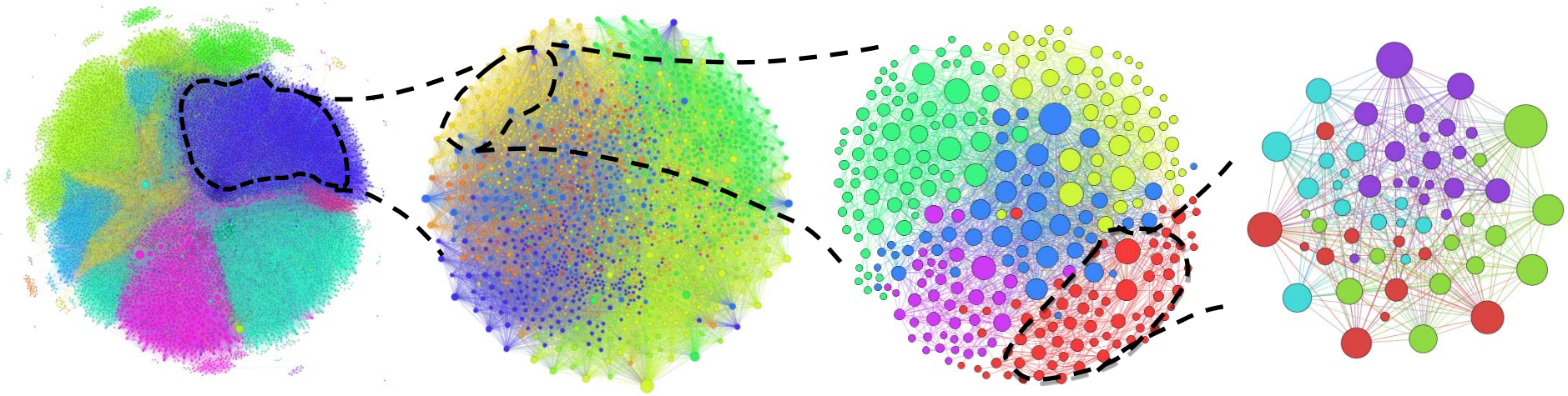
- 用途：
  - 广告
  - 推荐
  - 商务
  - …

第一步：“父母－子女” 的关系判断

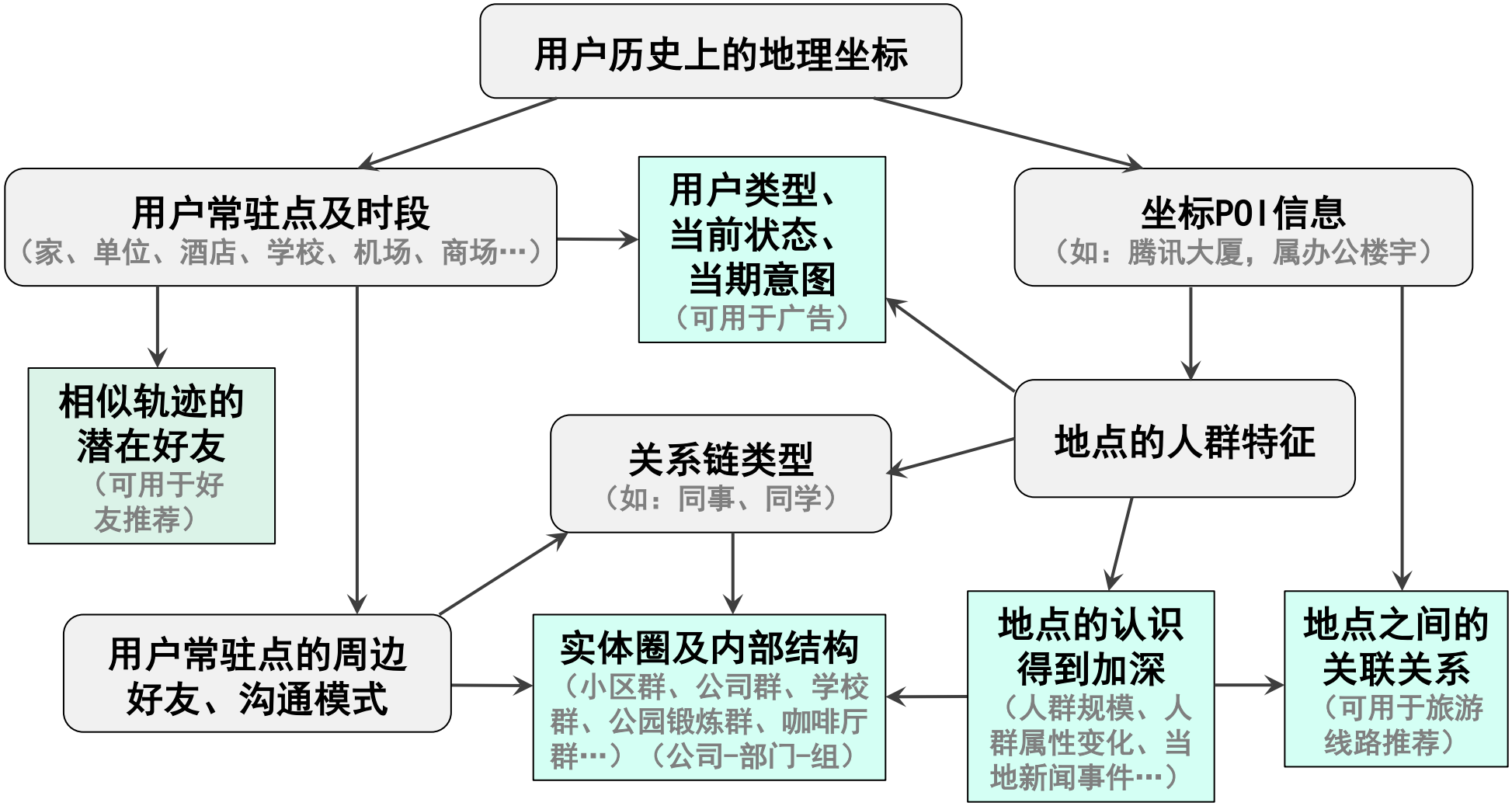
实体圈划分

- **实体圈：线下社团组织**
  - 公司、学校、班级、小组、住宅小区 ...
- **准确率、覆盖率 比较满意**

（基于QQ群以及腾讯实体圈进行验证）



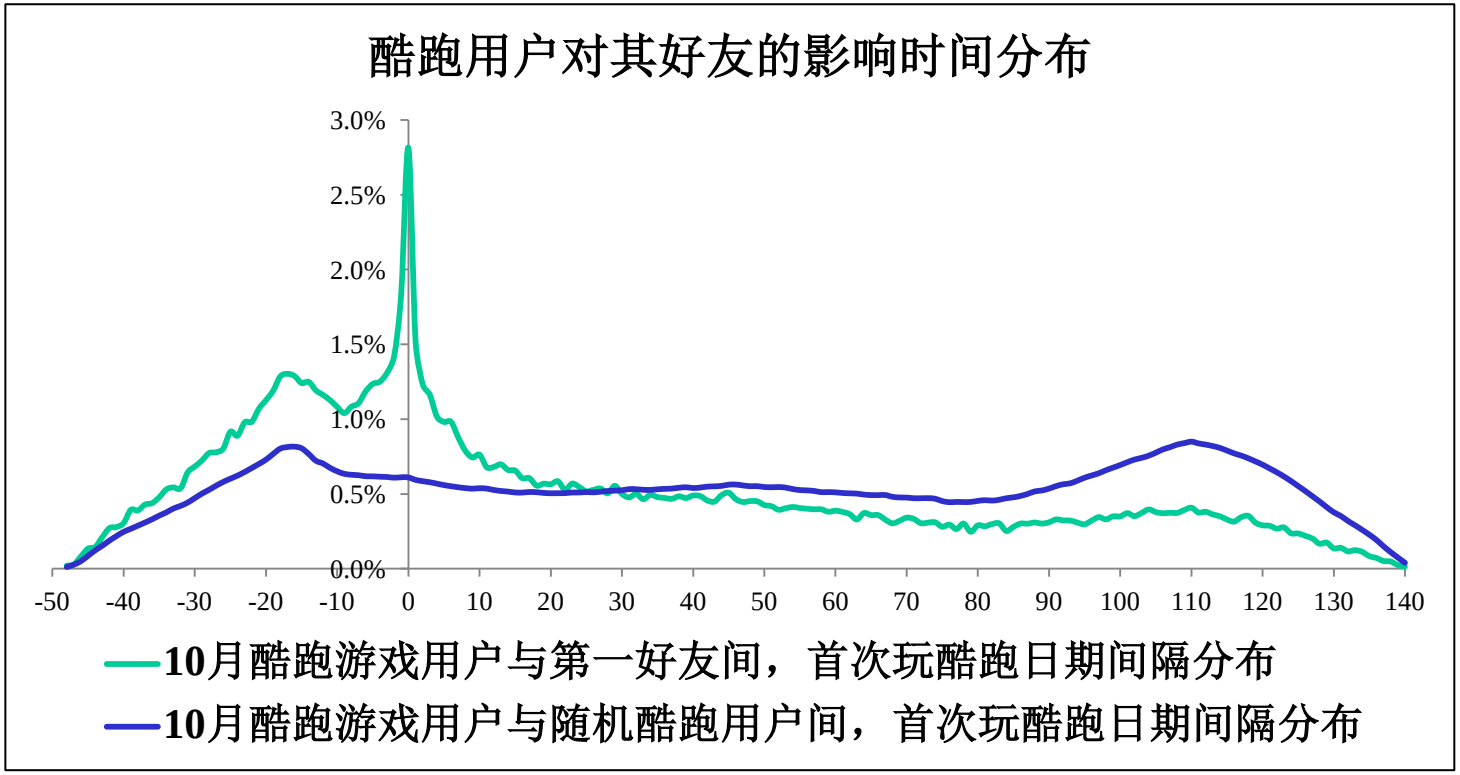
LBSN （Location-based Social Networking）



根据：paulhe(贺鹏) 《LBSN技术峰会2.1》

用户之间、用户与群体之间的行为相关

游戏app传播受关系链的影响：



第一好友：
聊天消息量
最大的好友

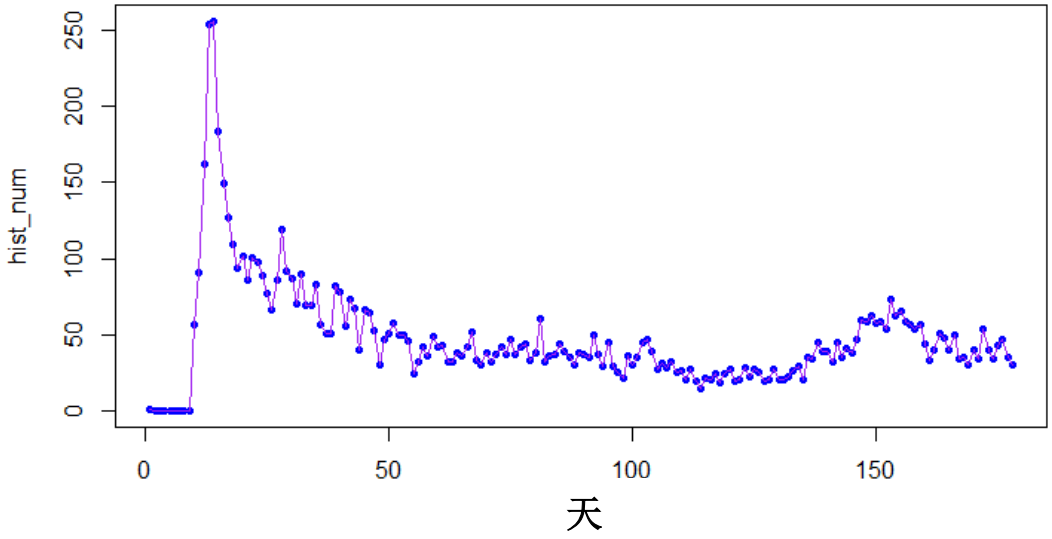
好友间玩酷跑的时间间隔小于陌生人（随机用户）
—— 好友关系链的拉动作用 ！

用户之间、用户与群体之间的行为相关

游戏app传播受局部网络结构的影响：

- **三角关系：3个人互为好友，已有2人在玩“天天酷跑”，则第三人玩“天天酷跑”的概率会更高，如果：**
  - 前两人进入游戏的时间间隔长度越小；
  - 前两人属于游戏早期用户；
  - 三角关系中的消息量越多。

- **“天天酷跑”在某大学的传播速率：**



Jessiexiong(熊祎) 《 APP在关系链上的传播（二）》

群组的主题推断

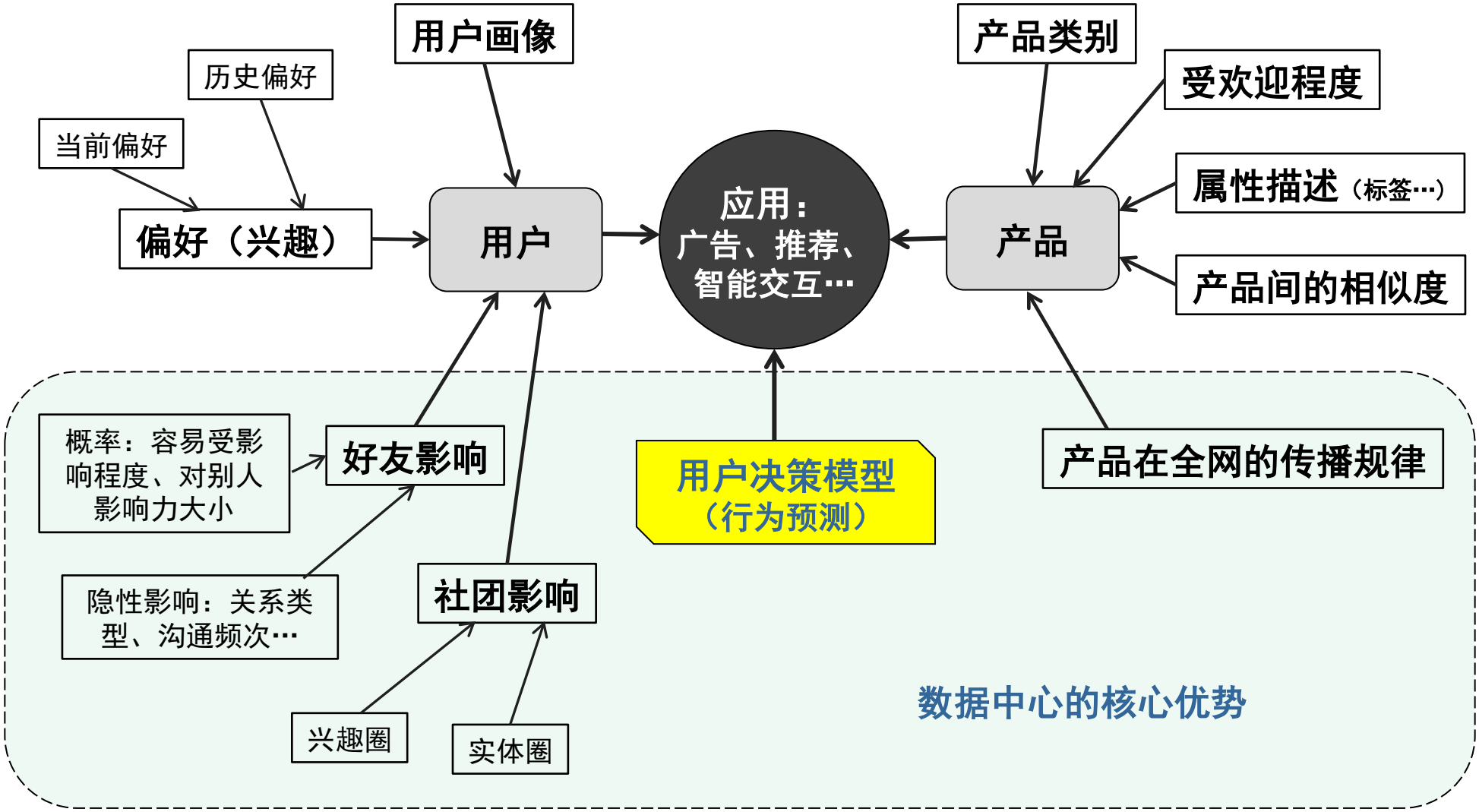
- **群名称、群简介等短文本信息，无聊天内容 → 群组主题**
  - 标签传播

奶粉：价格 同行 图片 专业 代理 经验 进口 货运 税
业主：楼 羊 团购 装修 日记
二手车：买卖 行业 市场 交易 过户 牌 机动车
食品：资源 行业 西餐 原料
地区：城市 昵称 技术 互联网 微博
厨师：菜 品牌 精华 厨艺 文化
汽车：电子眼 同行 违章 车辆 单 代办 全省

应用

社交网络内	公司内	其它
• 好友推荐 • 好友推荐 • 平台安全	• 基础数据用于广告业务 • 个性化推荐：app、音乐、书籍... • 各类分析（现状评估、趋势预测、模式探查等） • 吃喝玩乐等 LBS 应用	• 评估监测：国民经济水平，灾情，交通状况，世界杯观看，情感（幸福指数）等 • 预测：股市，疾病暴发，电影票房等

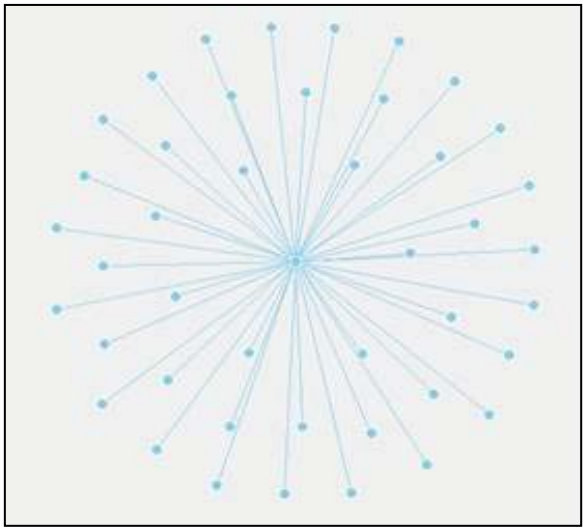
社交网络中受影响下的用户行为预测

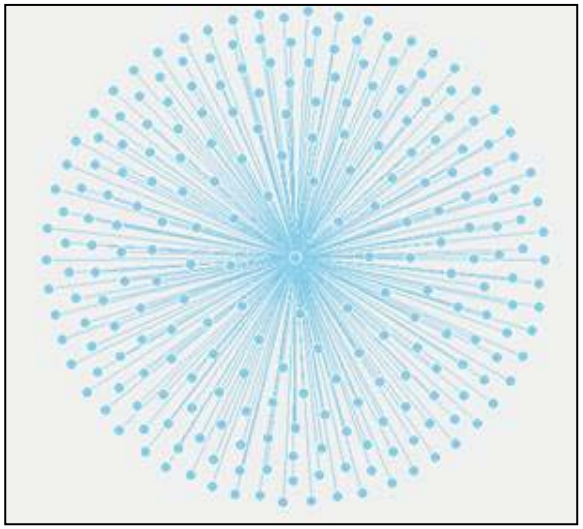


参考：chrisyi(易玲玲)，《APP推荐（社会化推荐）综述》

恶意行为识别

- 某些行为容易伪造，社会关系和沟通不易伪造
- 例：某类恶意用户的特征
  - 较低的局部聚类系数（clustering coefficient ）；

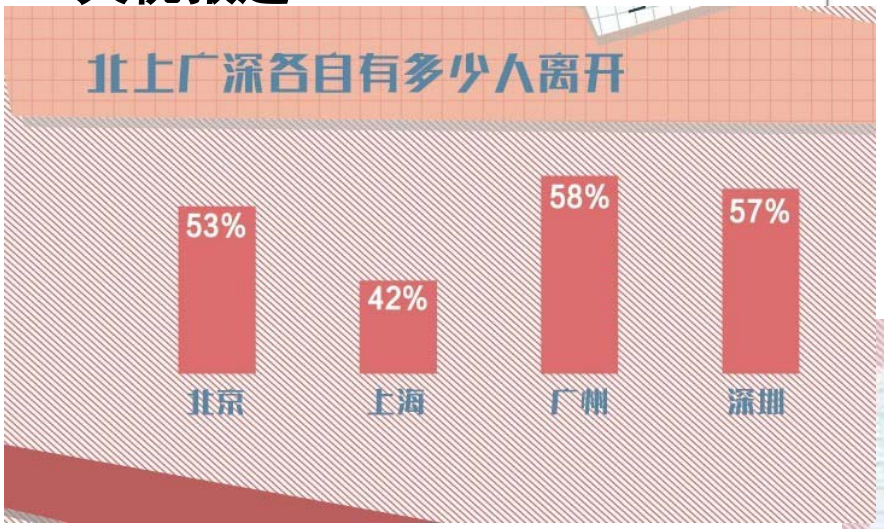




  - 与“好友”的消息数很少；
  - 快速频繁的登录。

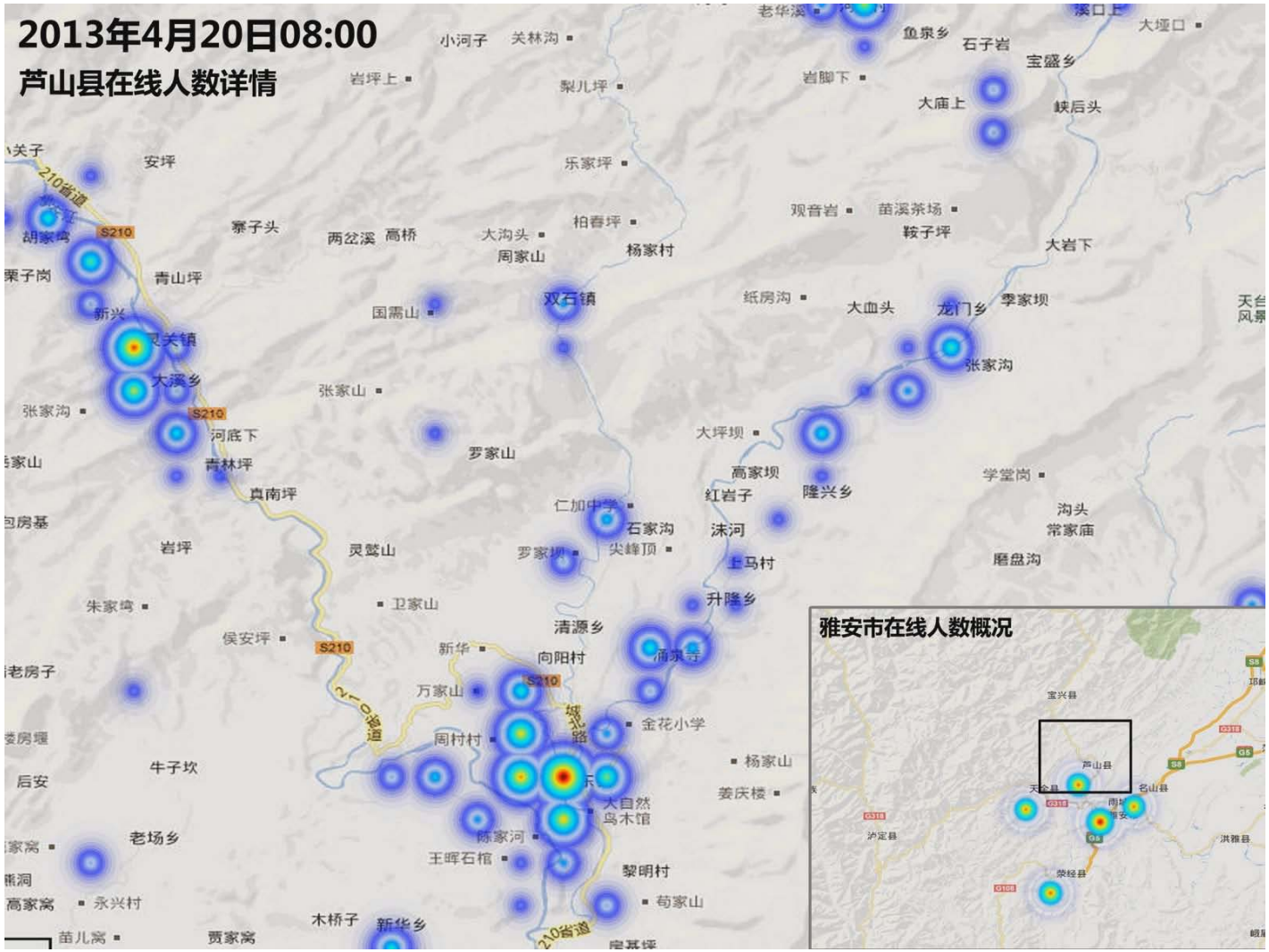
春运人群迁徙分析

央视报道



yoyozeng(曾宇宇)，《春运北上广深人群流动情况分析》

公益应用：汶川地震，优先到哪里救人？



根据下线用户数、重新上线用户数，判断局部受损程度

Outline

- 腾讯社交网络的研究内容
- **遇到的问题、解决思路**
- 模型框架
- 展望

应用中遇到的问题 1：建模过程

- 1、建模过程
  - 辛辛苦苦建立的复杂模型，有时 打不过简单经验规则。
  - 如何“招安”经验规则？
  - 如何构造更多有用的特征？
  - 对同一个问题，模型种类繁多。每一个都去试一下？还是看看别人用啥，我就跟着用啥？
  - …

期望：

- 理论的指导，实用的方法。
- 与应用领域无关的通用方法？
  （domain-independent universal solver）

应用中遇到的问题 1：建模过程

- 1、建模过程
  - 辛辛苦苦建立的复杂模型，有时 打不过简单经验规则。
  - 如何“招安”经验规则？
  - 如何构造更多有用的特征？
  - 对同一个问题，模型种类繁多。每一个都去试一下？还是看人用啥，我就跟着用啥？

规则式模型

- DT
- AR
- 1R、PRISM
- FOIL、RIPPER、SLIPPER …
- Rough Set
- Fuzzy Rules + DST
- ILP

特征构造

- 降维：PCA、LPP、ICA；MDS、kernel PCA …
- 构造：DL, Decision tree related (FRINGE、CITRE…），AR-related，GP-related，ILP-based，Annotation-based…

模型选择

- 角度：概率、统计力学、系统识别
- 通用：自组织建模 (GMDH、SOM、…), Evolutionary (NEAT、GNARL、…)

$$
\begin{aligned}
Y^* = & \left((\cos(x)\,\mathrm{sqrt}(abs(0.713614))) * (\cos(x)x)\right) \\
& + \cos(\mathrm{sqrt}(abs(\log(abs(10\hat{}(\cos(\cos(0.755058))))))) + x \\
& + \cos((\mathrm{sqrt}(abs((0.198157 * 0.518540))) + \sin(\cos(x)))) + x;
\end{aligned}
$$

Parikshit Sondhi， Feature Construction Methods: A Survey, *sifaka.cs.uiuc.edu/~sondhi1/survey3.pdf*

应用中遇到的问题 2：模型应用

- 2、模型应用
  - 上线测试的效果 有时不如 离线测试的效果 。
  - 业务部门不全盘接受模型结果，而是增加自己的策略。

    (business rules override model results)

    业务部门的理由：
    - 考虑合作伙伴的关系；
    - 考虑当前推广重点的特定产品（例如应用市场上的某款新app需要尽快抢占市场）；

期望：

- 拓展模型范围，把业务策略考虑进来。

应用中遇到的问题 2：模型应用

- 2、模型运营
  - 上线测试的效果 有时不如 离线测试的效果 。
  - 业务部门不全盘接受模型结果，而是增加自己的策略。
    (business rules override model results)

- **在线学习的模型**
  - Hoeffding Tree, VFDT, CVFDT and its variants (CVFDTNBC, SCRIPT, aCFVDT, SL_CVFDT), UFFT…； FIRM, FIRT-DD, SAIRT, …； REA, LEARN++.NSE, LEARN++.smote, LEARN++.NIE, …
  - 或，模型参数 = f1 (时刻、场景)，f1 用 PSO、ACO 等按时间迭代求得。
- **在线模型效果 = f_2(离线模型效果)**

- **模型中增加业务策略变量**
  - 类比：自适应最优控制

$$\dot{x} = f(x) + g(x)u(x)$$

  - 其中 x 是可测量的系统状态，f(x) 是系统迁移动力学，g(x) 是系统输入动力学，u(x) 是控制输入（对应于业务策略）

（详见下页）

应用中遇到的问题 2：模型应用

- 2、模型运营
  - 上线测试的效果 有时不如 离线测试的效果 。
  - 业务部门不全盘接受模型结果，而是增加自己的策略。
    (business rules override model results)

模型中增加业务策略变量（一种方案）

- 目的：单位时间收益 $V \equiv (V_K/K)$ 最大，其中：
- V_K：截止到 K 时刻的累计总收益，如总收入、总下载量等。 $V_K = \Sigma_{(k=1,\ldots,K)} r_k$， r_k：从k-1到k时间内的收益，$r_k = \psi(u_{k-1}, p)$， ψ 是一个模型，p 是参数。
- u_k：k时刻的控制变量，即业务策略变量，依赖于状态 x_k： $u_k = \phi(x_k, q)$， ϕ 是一个模型，q 是参数。
- x_k：第k时刻的状态，如 用户属性、场景信息、当前点击率或转化率等，根据从 k-1 到 k 之间的记录或统计计算得到。

单位时间收益 V =
F3 (状态 x, 业务策略 u, 模型参数p、q)

V 最大化的必要条件：

$$\partial V / \partial p = 0$$
$$\partial V / \partial q = 0$$

据此设计学习算法。

应用中遇到的问题 3：推荐系统

3、推荐系统

- 模型僵硬：刚买了一个微波炉，又接二连三地推荐微波炉？我的微波炉没这么快坏掉的。
- 个性化不足：为什么推荐这些给我？它们都不是我想要的。我想要的在哪里？
- 退化成搜索：给我一个长长的列表，我得一个个地去看，每一个都似是而非，还是拿不定主意。这叫什么推荐？
- 缺少互动：猜我喜欢什么？为啥不问我喜欢什么？
- …

期望：

回归到推荐本质，不要瞎蒙；

Maximize：E［推荐成功率 / 用户成本(user's effort)］

应用中遇到的问题 3：推荐系统

- 3、推荐系统
  - 模型僵硬：刚买了一个微波炉，又接二连三地推荐微波炉？我的微波炉更新没有这么快的。
  - 个性化不足：为什么推荐这些给我？这些都不是我想要的。我想要的在哪里？
  - 退化成搜索：给我一个长长的列表，我得一个个地去看，每一个都似是而非，还是拿不定主意。这叫什么推荐？
  - 缺少互动：猜我喜欢什么？为啥不问我喜欢什么？

- **当前做法本质是贪婪的**
  - 贪婪：预测用户对什么评价最高，或最可能点击/下载什么，就推荐什么；
  - 用户迟早会遇上“最”，即使在没有推荐系统工作时；
  - 如果不推荐“最”，而是推荐“次最”，全局会否更优？

- **重视个体用户的历史行为**
  - 事件序列的预测模型；
  - 预测用户下一个动作。

- **交互式推荐**
  - 用户对当前推荐做出实时反馈；
  - 系统实时更新推荐列表；
  - 多次迭代，用户找到想要的东西。

交互式推荐

- 例：推荐餐馆（*Adaptive Place Advisor*）
  - 用户: 你觉得我今晚应该到哪里去吃饭？
  - 系统: 你喜欢吃哪种类型的？
  - 用户: 有哪些类型呢？
  - 系统: 有中餐、印度菜、地中海餐....
  - 用户: 哦，可能找一个便宜的印度餐馆吧 .
  - 系统: 你想把车停在哪里？
  - 用户: 我无所谓，只要是在 Palo Alto 里面.
  - 系统: 抱歉，那个地方还没有这样的餐馆。你想先查查价格吗？
  - 用户: 不了，要不我还是选中餐吧 .
  - 系统: 那么 Ramona 路 420 号的 Mandarin Gourmet 餐馆怎么样？
  - 用户: 不好。你还有别的推荐吗？
  - 系统: Emerson 街 443 号的 Jing-Jing Szechwan 湘菜馆呢？
  - 用户: 好啊，听起来不错 .

http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume21/thompson04a-html/node4.html

交互式推荐

基本假设

- 系统对用户当前兴趣的把握程度：

  实时表达或选择 > 最近发生的准实时行为 > 历史行为

特点

- 模拟现实生活中的推荐过程；
- 多次交互，逐步逼近用户当前的真正兴趣，缩小候选集范围；
- 回归到推荐本质！

待研究

- 交互方式。（不一定是自然语言）
- 减少交互次数，快速收敛。
- 与强化学习结合，长期记录用户的响应，极大化一个累积回报函数，使得用户长期受益最大。

应用中遇到的问题 4：SNS 的长期演化

- 4、社交网络的长期演化方向
  - QQ 会不会成为第二个 MySpace、Friendster 轰然倒下？
  - 产品的 UI、结构、功能，对用户行为及市场的影响

期望：

  - Leading indicators, crises/crash forecasting
  - 产品的长期运营

应用中遇到的问题 4：SNS 的长期演化

- 4、社交网络的长期演化方向
  - QQ 会不会成为第二个 MySpace、Friendster 轰然倒下？
  - 产品的 UI、结构、功能，对用户行为及市场的影响

系统全局性指标

- David Garcia et al.,：低的付出收益比率和弱的k核心分布；
- 用户对产品兴趣下降的程度：用户使用产品总时长分布对应 first passage time distribution 的隐含随机过程的时变参数，反映出用户单位时间内使用产品的冲动次数。

产品运营的支持性模型

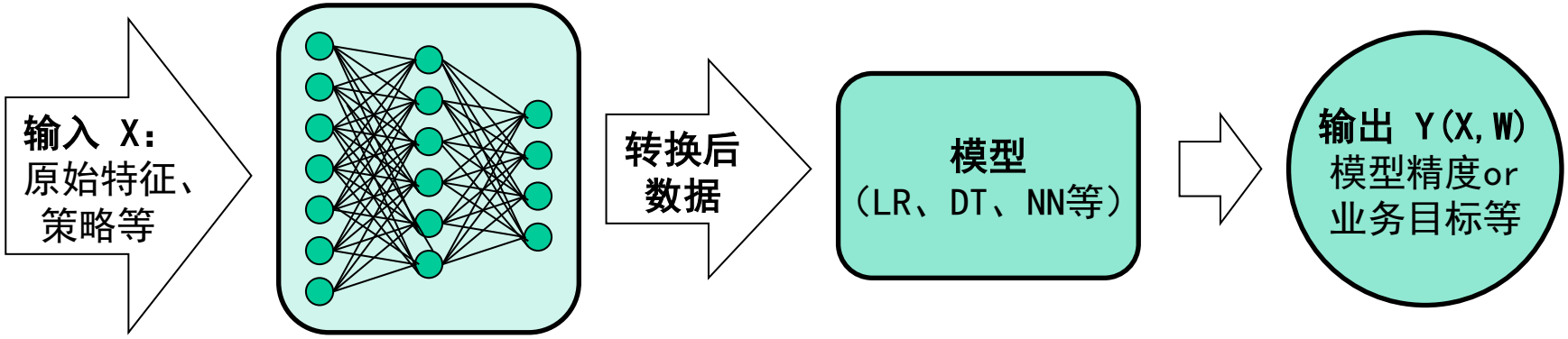
- 产品的定量化描述：包括对其属性、菜单、界面、工作流、功能、可用性等方面的抽象和定量描述方法；
- 在外界因素影响下的产品业务指标（KPI）的预测模型；
- 用户与产品交互作用的系统仿真模型，指导制订最优的产品运营策略。

David Garcia et al., Social Resilience in Online Communities: The Autopsy of Friendster, http://arxiv.org/abs/1302.6109
http://select.yeeyan.org/view/383585/350280
D. Lamper et al., Predictability of Large Future Changes in a Competitive Evolving Population, Physical Review Letters Vol 88, No. 1, 7 Jan. 2002
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.261.1677

Outline

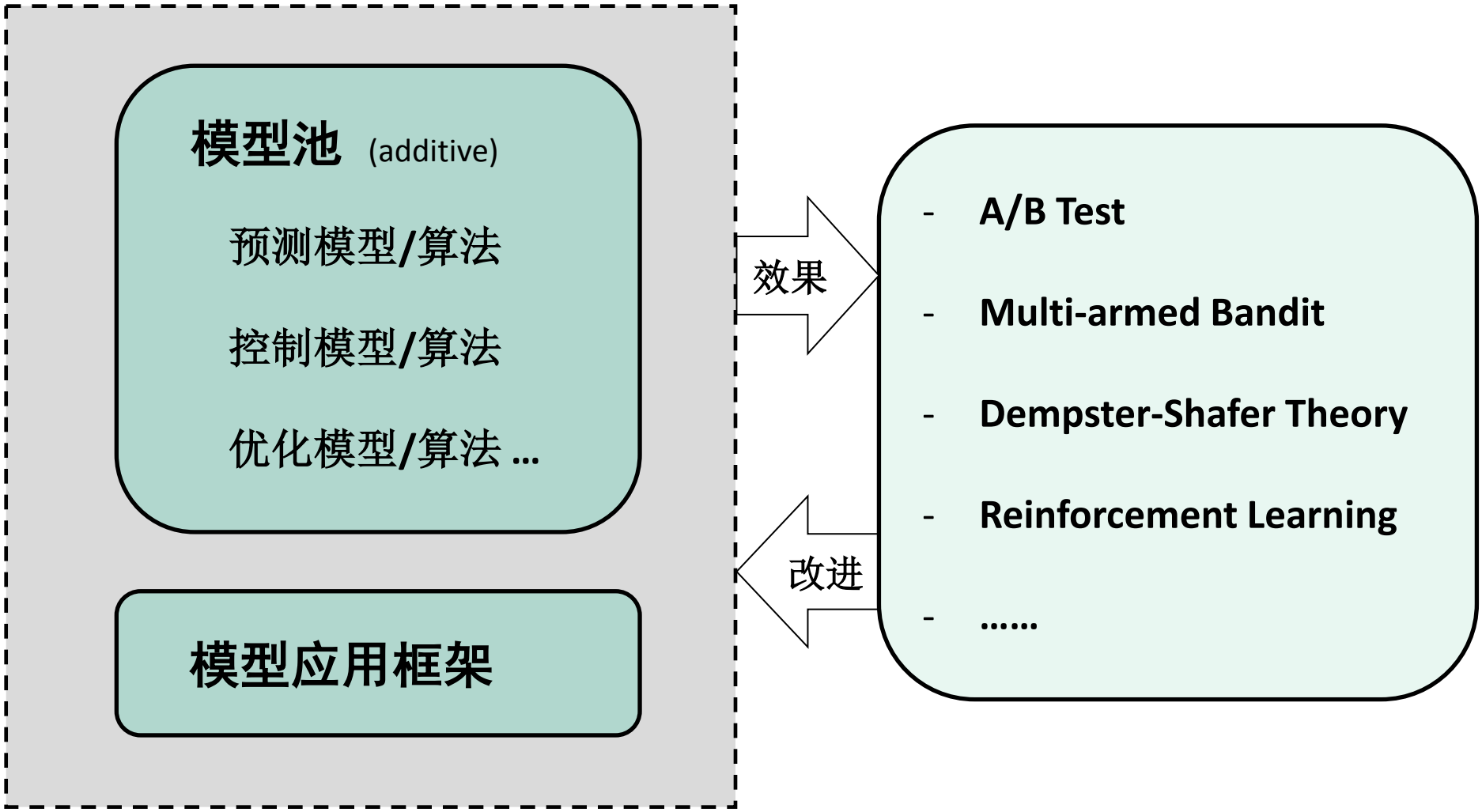
- 腾讯社交网络的研究内容
- 遇到的问题、解决思路
- **模型框架**
- 展望

一种框架建议：整合特征构造和业务目标优化



- 求 $argmax_W\ Y(X,W)$
  - 可用 PSO、ACO等直接搜索法（M/R上可实现）。 $\Theta = \Theta(W)$ 。
- 好处：
  - 省心：不需人为构造特征；不再有离线测试效果与上线效果的差异。
  - 直接确保商业目标的最大化。

框架探索方式：最优迭代改进



Outline

- 腾讯社交网络的研究内容
- 遇到的问题、解决思路
- 模型框架
- **展望**

社交网络的未来模型的狂想

- 探索途径
  - Heuristics + Principled Approaches
  - Well-conceived postulations of local mechanisms
  - Universal approximator / solver
- 真正的智能
  - 非目前流行的 Intelligent Personal Assistant （ Google Now、Microsoft Cortana、Apple Siri）
  - 基于更少用户反馈、更大数据集背景的 推理机制
- 人是群居的动物
  - SNS 永远存在，交流模式继续演进，探索永无止境。

致谢

社交网络运营部 数据中心：
Chuanchen(陈川)、Paulhe(贺鹏)、
Chrisyi(易玲玲)、Alangao(高瀚)、
jinpogao(高金坡)、Jessiexiong(熊祎) 、
yoyozeng(曾宇宇)

特别感谢
Nofeeling（邱跃鹏）

谢谢大家!

Q & A …