



# 网络数据挖掘

## 第二部分：图数据挖掘

沈华伟

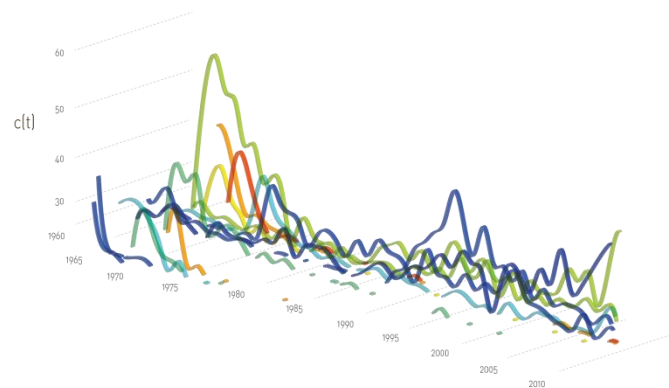
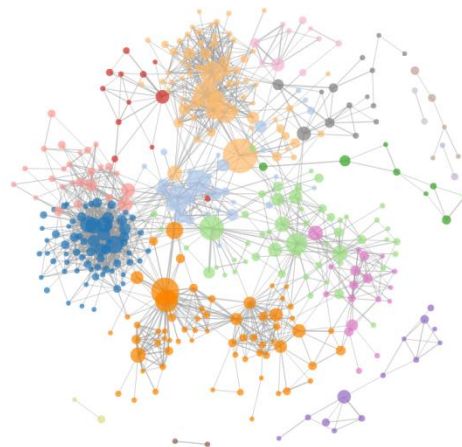
[shenhuawei@gmail.com](mailto:shenhuawei@gmail.com)

中国科学院计算技术研究所

2016.11.15

# 图数据挖掘

- 第一讲：图排序(11月1日)
  - 复杂网络
  - 图排序
- 第二讲：图挖掘(11月8日)
  - 图聚类
  - 社区发现
- 第三讲：图预测(11月15日)
  - 网络推断
  - 传播预测

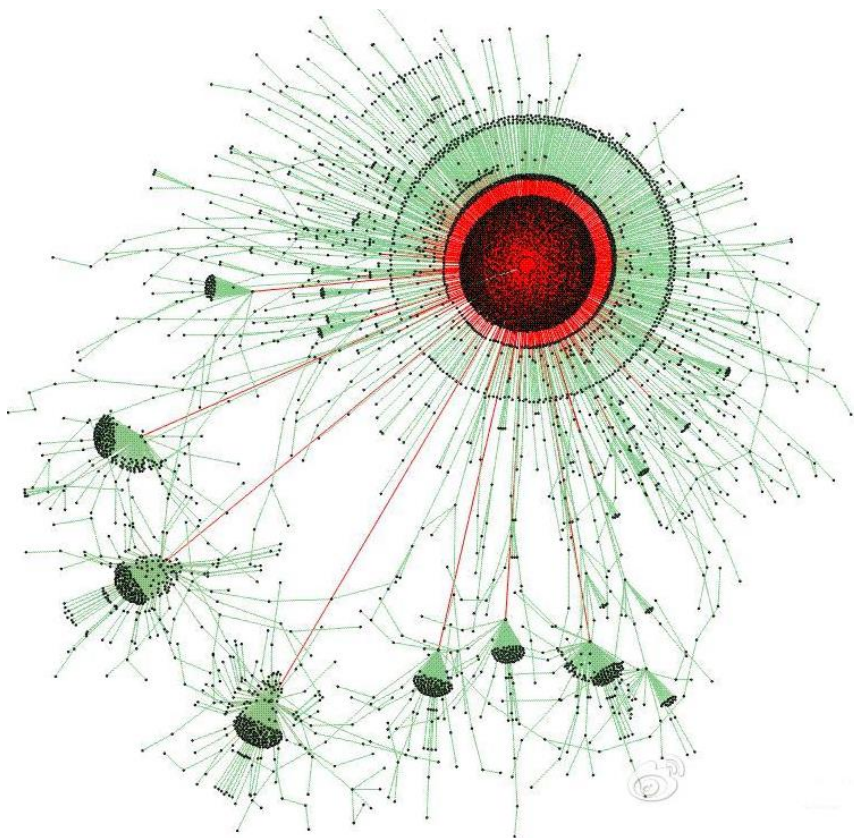


# 第三讲：图预测

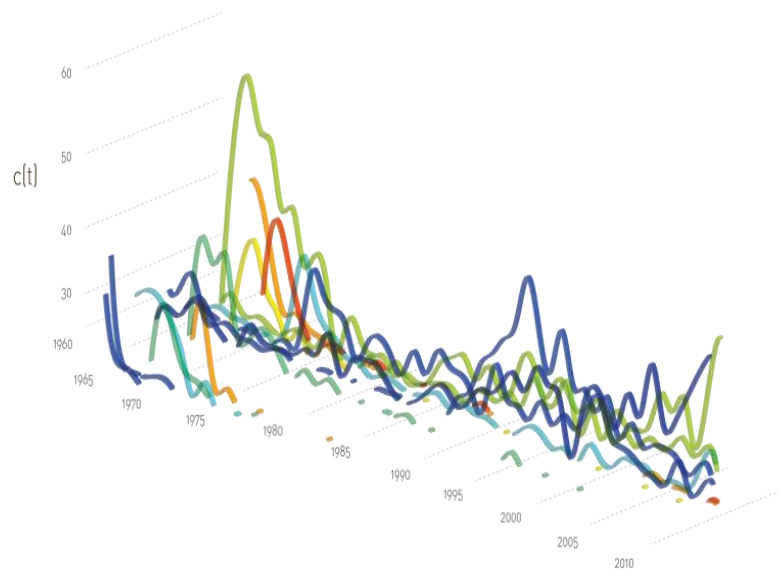
# 内容提纲

- 网络信息传播
  - 信息传播模型
  - 影响最大化
  - 传播网络推断
  - 流行度预测

# 网络信息传播



一条微博的传播树



论文引用次数的时序图

# 网络信息传播的特点

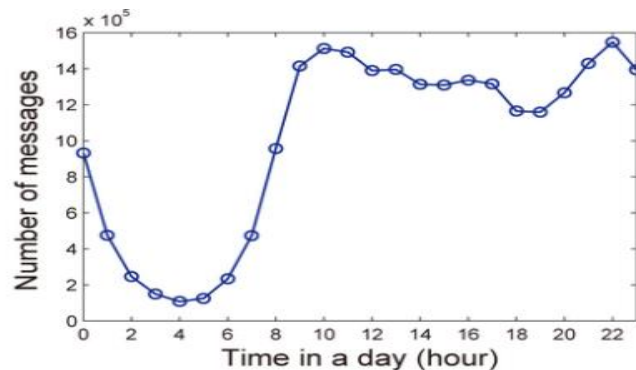
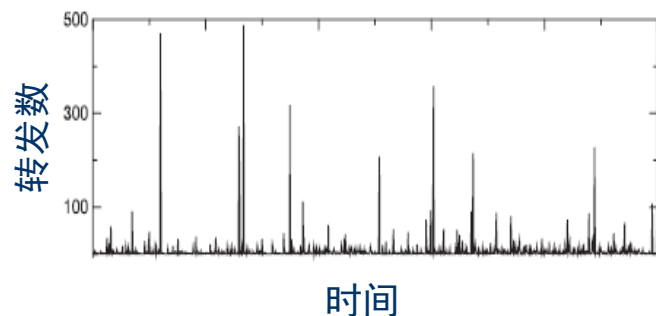
## ■ 网络效应

- 传播以社会关系网络为媒介进行，网络结构的不均匀使得网络信息传播呈现出突发涌现的特点

- 星星之火可以燎原
- 网络是放大器

## ■ 阵发性

- 传播在时间上呈现出阵发性
- 多次传播
- 活动规律（昼夜、工作日-非工作日）



用户活跃时间

# 信息传播模型

- 个体间的社会关系表示为社会网络  $G = (V, E)$
- 信息在社会网络  $G$  上传播
  - **二元性**：传播过程中，每个节点的状态要么是激活的（active），要么是未激活的（inactive）
  - **不可逆**：节点一旦被激活，就不会再变成未激活状态
  - **单调性**：对于一个节点  $v$ ，其邻居节点中处于激活状态的节点越多，其被激活的可能性越大

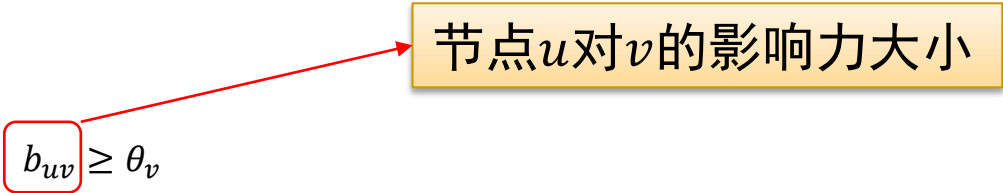
# 两类信息传播模型

- 阈值模型（Threshold Model）
  - 线性阈值模型（LT: Linear Threshold）
- 级联模型（Cascade Model）
  - 独立级联模型（IC: Independent Cascade）



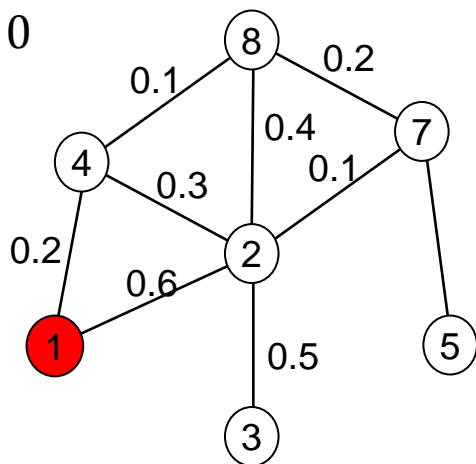
# 线性阈值模型

- 每个节点 $v$ 有一个阈值，记为 $\theta_v$ ，传播开始前随机产生
- 初始时处于激活状态的节点集合记为 $A_0$
- 时间步 $t = 0, 1, 2, \dots$ 
  - 每个未被激活的节点 $v$ ，根据其邻居节点的激活情况决定自己是否被激活
    - 节点 $v$ 激活的条件为
$$\sum_{u \text{ 是 } v \text{ 的活跃邻居}} b_{uv} \geq \theta_v$$

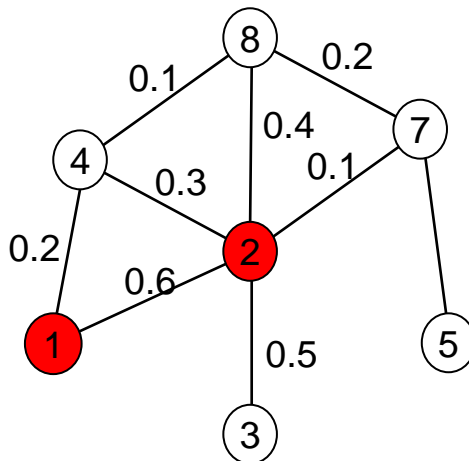

  - 如果同一个时间步中没有任何节点被激活，传播过程结束
- 性质
  - 有记忆：根据邻居节点情况来判定当前节点是否被激活

# 线性阈值模型示例

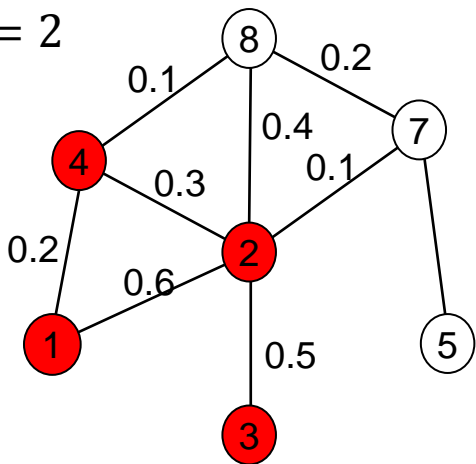
$t = 0$



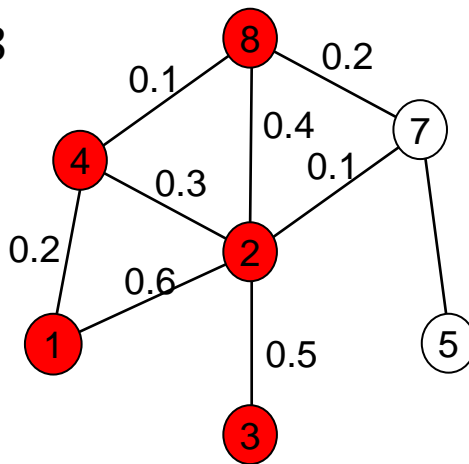
$t = 1$



$t = 2$



$t = 3$



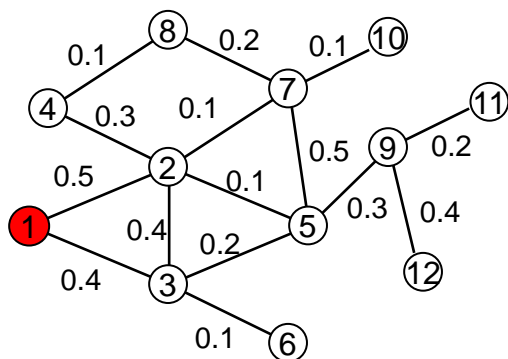
● 激活 ○ 未被激活

假设所有节点的激活阈值都是0.5

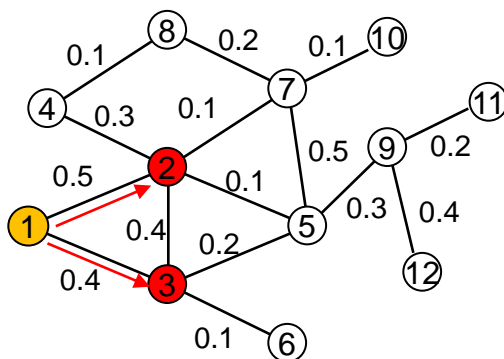
# 独立级联模型

- 初始时处于激活状态的节点集合记为 $A_0$
- 时间步 $t = 0, 1, 2, \dots$ 
  - 在时间步 $t$ 被激活的每个节点 $u$ ，**有且只有一次机会**去尝试激活其未被激活的邻居节点 $v$ ，成功激活的概率为 $p_{uv}$
  - 如果 $v$ 被成功激活，将其放到节点集合 $A_{t+1}$ （ $t + 1$ 步被激活的节点集合）
  - 当 $A_t$ 为空集时，传播过程结束
- 性质
  - **顺序无关**：有多个节点尝试激活同一个节点时，按照任意顺序进行
  - **无记忆性**：节点 $u$ 成功激活 $v$ 的概率只和 $p_{uv}$ 有关，和历史上有有多少节点尝试激活节点 $v$ 无关

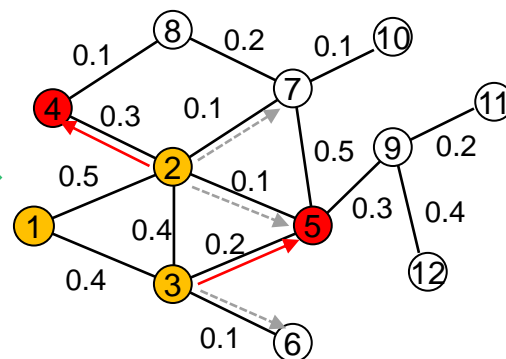
# 独立级联模型示例



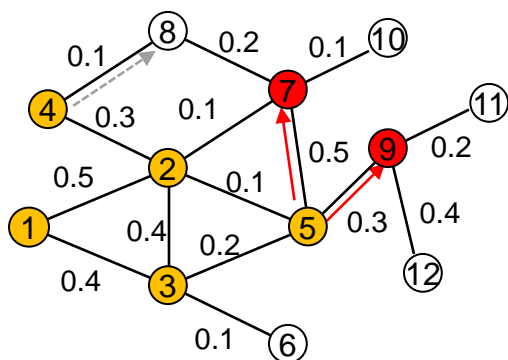
$t = 0; A_0 = \{1\}$



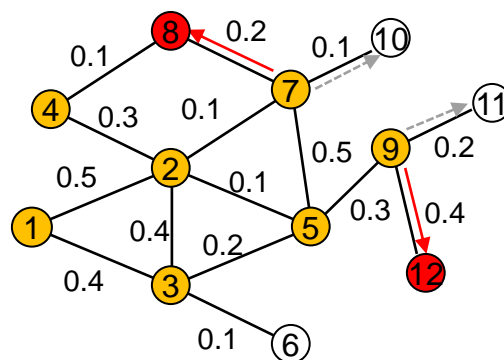
$t = 1; A_1 = \{2, 3\}$



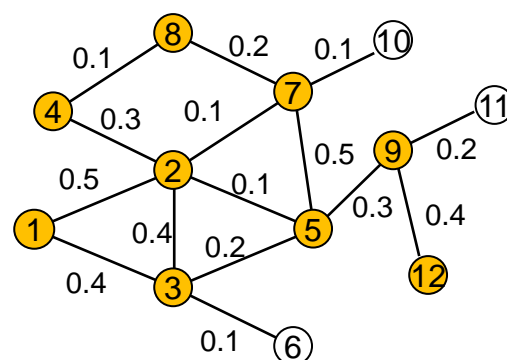
$t = 2; A_2 = \{4, 5\}$



$t = 3; A_3 = \{7, 9\}$



$t = 4; A_4 = \{8, 12\}$



$t = 5; A_5 = \emptyset$

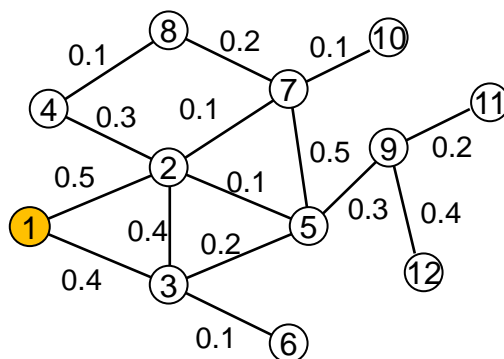
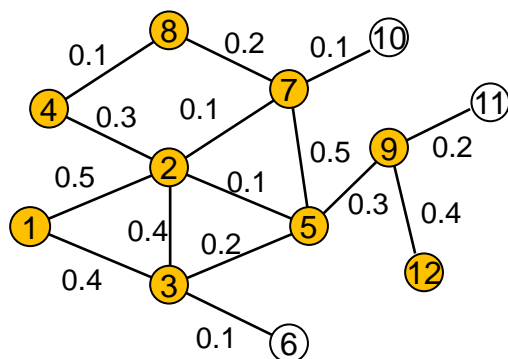
● 处于激活态，且具备激活邻居节点的能力

○ 未被激活

● 处于激活态，但不具备激活邻居节点的能力

# 节点影响范围

- 传播具有很大的随机性
  - 从同样的初始激活节点开始，每次传播范围也不同



独立级联模型为例： $A_0 = \{1\}$ 的两次不同传播过程对应的传播范围

- 节点的影响范围
  - 通过蒙特卡罗模拟得到多次传播的范围，**取平均值**

缺点：计算不同节点的影响范围时，对于每个节点需要重新进行蒙特卡罗模拟

# 节点传播范围

## ■ 独立级联模型的传播范围计算

- 独立级联模型中，各条边上的传播概率是彼此独立的
- 可以事先通过抛硬币的方式确定每条边是否存在，从而得到传播过程的一个快照网络
- 节点的影响范围则为其在快照网络中可达节点集合

## ■ 传播范围计算

- 按照上述方式获得多个快照网络，计算可达节点集合大小
- 取平均值得到传播范围

优点：计算不同节点的影响范围时，可以复用同样的快照网络

缺点：产生单个快照的代价高于单次蒙特卡罗模拟的代价

# 内容提纲

- 信息传播预测
  - 信息传播模型
  - 影响最大化
  - 传播网络推断
  - 流行度预测

# 影响最大化问题

## ■ 应用背景：病毒式营销

□ 核心问题：如何选择一组种子节点，获得最大的影响范围？

■ 例子：Hotmail

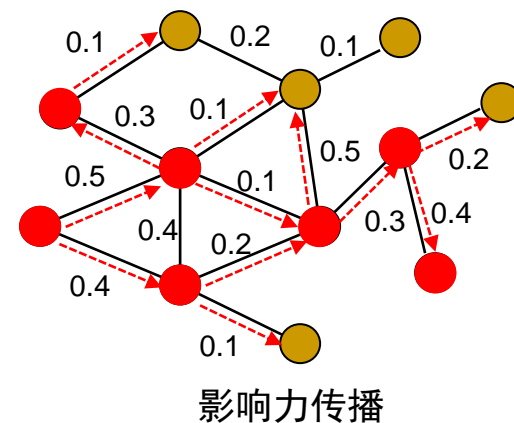
## ■ 问题定义：Influence Maximization

— 输入：

- 社交网络 $G=(V,E)$ ：节点是用户，边是用户间的关系
- 影响力扩散模型：级联模型或阈值模型
- $k$ ：种子节点的数量

— 输出：种子集合 $S$ ,  $|S| \leq k$

— 目标：最大化影响力传播范围  $\sigma(S)$





# 影响最大化问题

- 影响最大化问题是一个NP-hard问题 [Kempe et al. KDD 2003]
  - 蛮力计算不可行
    - 复杂度:  $O\left(\binom{n}{k}T\right)$
    - $n$ 是节点个数,  $T$ 是估计节点影响范围的平均代价
- 设计近似算法成为影响最大化问题求解的关键
  - 贪心算法
    - 精度有保障, 速度较慢
  - 启发式算法
    - 速度快, 精度无保障

# 影响最大化问题的性质

## ■ 非负性

- 对于任意节点集合 $S$ ，其影响范围 $\sigma(S) \geq 0$

## ■ 单调性

- 对于两个节点集合 $S$ 和 $T$ ，如果 $S \subseteq T$ ，有 $\sigma(S) \leq \sigma(T)$

## ■ 次模性：边际效益递减

- 对于两个节点集合 $S$ 和 $T$ ，满足 $S \subseteq T$ ，假设节点 $v \notin T$ ，有 $\sigma(T \cup \{v\}) - \sigma(T) \leq \sigma(S \cup \{v\}) - \sigma(S)$

# 影响最大化的贪心算法

## ■ 贪心算法 (Greedy)

- 初始时  $S = \emptyset$
- 逐个选择边际效益最大的节点  $v^*$  加入到  $S$  中

$$v^* = \arg \max_v (\sigma(S \cup \{v\}) - \sigma(S))$$

## ■ 贪心算法的性质

- 精度不小于  $1 - \frac{1}{e} - \varepsilon$
- 假定最优解为  $S^{OPT}$ ，贪心算法的解为  $S^*$ ，有

$$\sigma(S^*) \geq \left(1 - \frac{1}{e} - \varepsilon\right) \sigma(S^{OPT})$$

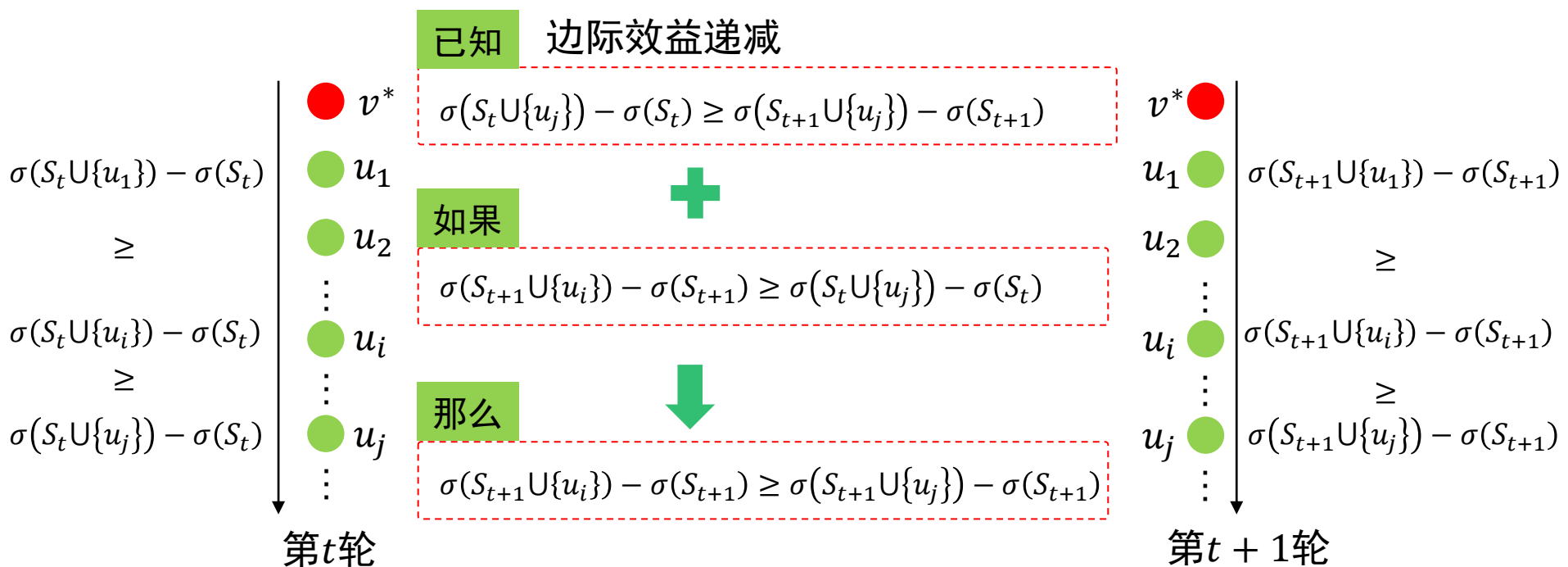
- $\varepsilon$  是一个非常小的整数，其值依赖于蒙特卡罗模拟带来的随机误差

# 影响最大化贪心算法的加速

## ■ CELFGreedy [Leskovec et al. KDD 2007]

- 思路：避免不必要的计算，减少计算影响范围 $\sigma$ 的次数

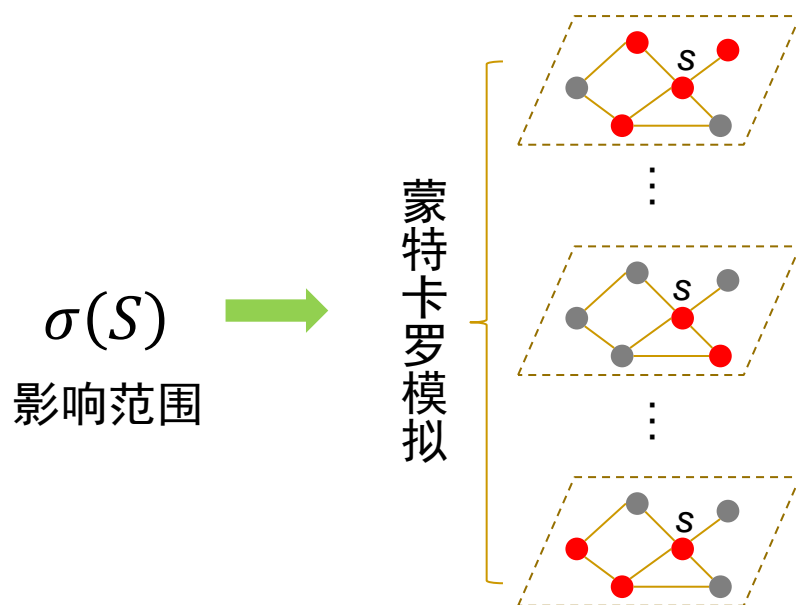
$$S_{t+1} = S_t \cup \{v^*\}$$



启示：每轮只对部分节点通过蒙特卡罗模拟计算其边际效益（必要时才计算）

# 影响力最大化贪心算法的困境

## ■ 精度和速度的矛盾(Accuracy-Scalability dilemma)



### 精度和速度的矛盾

- ✓ 贪心算法  $1-1/e$  的精度保障，取决于  $\sigma(S)$  的次模性和单调性，这需要对  $\sigma(S)$  进行尽可能精确的估计
- ✓ 提高模拟次数，精度提高，速度下降
- ✓ 降低模拟次数，速度提高，精度下降

## ■ 矛盾产生的原因：蒙特卡罗模拟带来的随机性

# 单调性和次模性的破坏

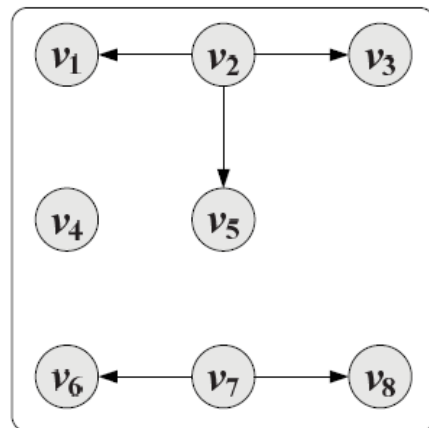
## ■ 单调性的破坏

$$\sigma_1(\{v_2\}) > \sigma_2(\{v_2, v_5\})$$

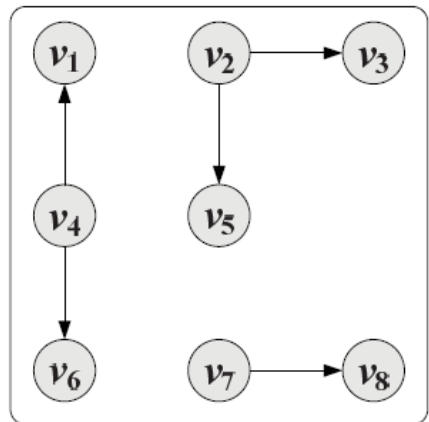
## ■ 次模性的破坏

$$\sigma_1(\{v_4\}) - \sigma_1(\emptyset) = 1$$

$$\sigma_2(\{v_2, v_4\}) - \sigma_2(\{v_2\}) = 3$$



第一轮的传播快照



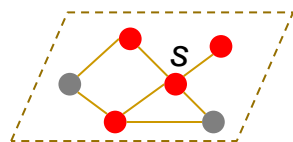
第二轮的传播快照

注： $\sigma_1$ 基于第一轮的传播快照计算， $\sigma_2$ 基于第二轮的传播快照计算

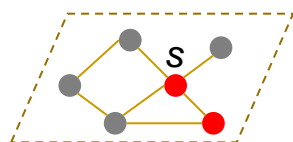
# 影响最大化贪心算法困境的解决方案

## ■ 解决思路

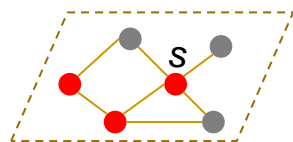
- StaticGreedy: 在贪心算法的迭代过程中, 复用蒙特卡罗模拟的结果



⋮

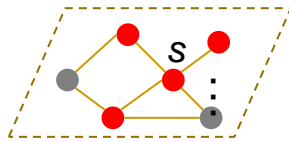


⋮

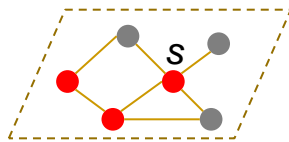


所有可能的蒙特卡罗模拟

采样  
→



⋮



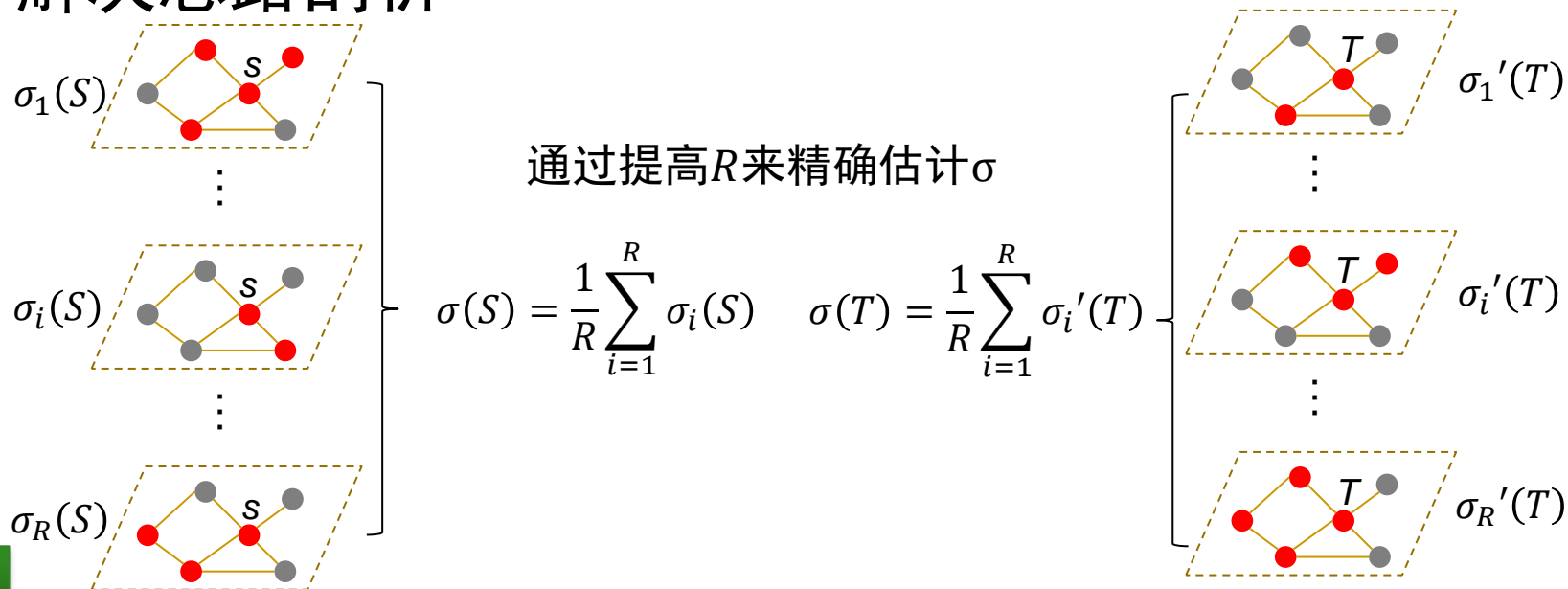
用于贪心算法的  
蒙特卡罗模拟

### 算法对照

- ✓ 传统贪心算法  
每次迭代进行一组采样  
每组采样的个数要足够大, 才能保证子模性和单调性  
→ 系统误差
- ✓ 静态贪心算法  
只进行一组采样, 迭代过程中复用这组采样  
通过少量采样严格保证次模性和单调性  
→ 随机误差

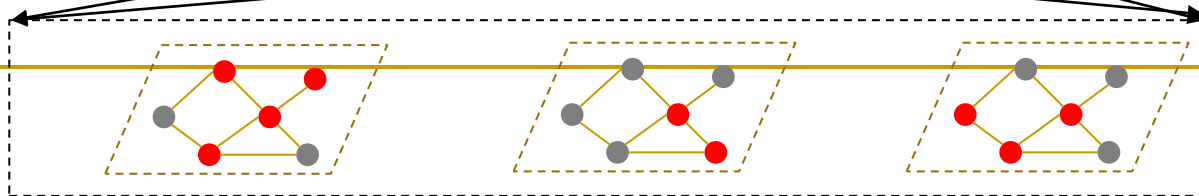
# 影响最大化贪心算法困境的解决方案

## ■ 解决思路剖析



## 策略2

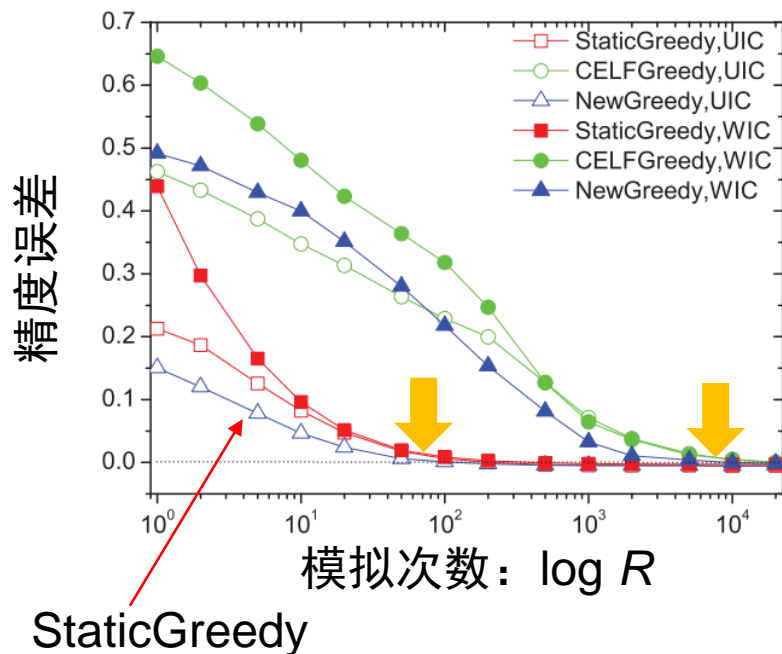
$$\sigma(S) = \frac{1}{R} \sum_{i=1}^R \sigma_i(S) \quad \sigma(T) = \frac{1}{R} \sum_{i=1}^R \sigma_i(T)$$



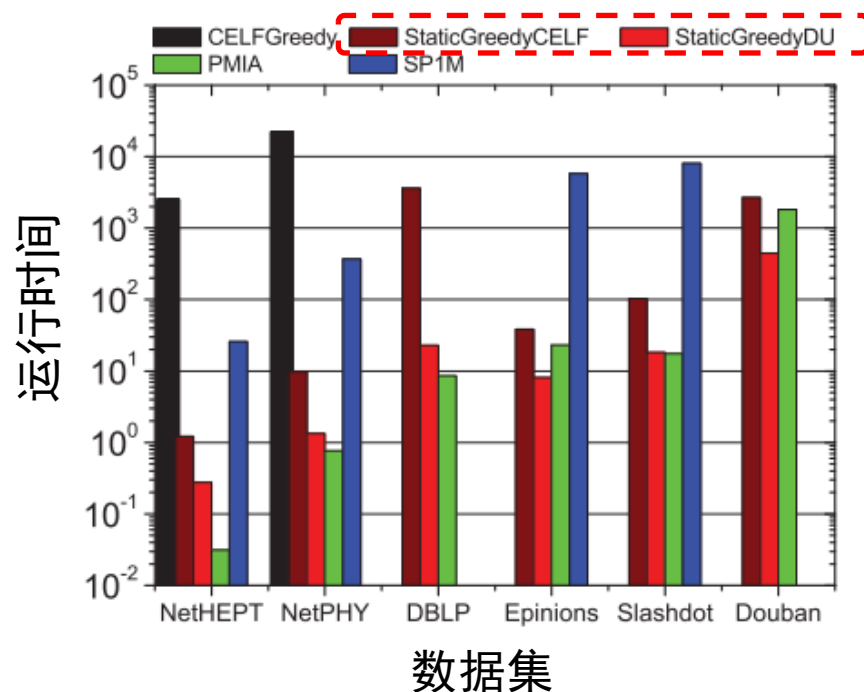


# StaticGreedy的效果

同样精度情况下，蒙特卡罗模拟次数从20000降低到100，降低了2个数量级



运算速度比经典的CELFGreedy算法快了1000倍



# 影响最大化问题小结

- 影响最大化问题具有重要的应用前景
  - 广告投放
  - 病毒式营销
  - .....
- 目前主要关注算法优化
  - 近几年研究热度在降低
  - 未来将侧重于和active learning的结合

# 课间休息



# 内容提纲

- 信息传播预测
  - 信息传播模型
  - 影响最大化
  - 传播网络推断
  - 流行度预测

# 网络推断问题

## ■ 问题描述

- 根据信息传播记录（information cascade），推断背后的传播网络

- 输入

  - cascade

$C = \{\langle u, t \rangle\}$       节点 $u$ 在 $t$ 时刻被激活

- 输出

  - 节点之间的传播概率 $p_{uv}$

# 网络推断：点对点模型

## ■ 基本思路

- 模型参数 $p_{uv}$ ： $n \times n$ 个参数
  - 各个 $p_{uv}$ 是彼此独立的
- $p_{uv}$ 的值依赖于
  - $u$ 在 $v$ 之前被激活的次数：次数越多， $p_{uv}$ 越大
  - 每个cascade中， $u$ 被激活的时刻 $t_u$ 和 $v$ 被激活的时刻 $t_v$ 之间的时间间隔大小 $t_v - t_u$ ：间隔越小， $p_{uv}$ 越大

## ■ 模型学习过程

- 最大化模型生成cascade的似然，估计出各个参数

# 点对型模型的缺陷

## ■ 过表达

- 各个 $p_{uv}$ 是彼此独立的
- 同一个用户对不同人的影响彼此独立

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

## ■ 过拟合

- $p_{uv}$ 的学习仅依赖于 $u$ 和 $v$ 的信息
- 参数多，模型易过拟合

# 点对型模型的改进

- 每个用户采用两个低维( $k$ )向量表达
  - $I$  : 表示节点的影响力 (influence)
  - $S$  : 表示节点的易感度 (susceptibility)

- 用户间的人际影响力建模为

$$p_{uv} = I_u S_v$$

- 好处
  - 模型参数由 $n^2$ 降到了 $2nk$
  - 克服了点对型模型的过表达和过拟合问题



# 网络推断小结

- 网络推断是目前的研究热点之一
  - 效率较低，不适用于大规模的网络
  - 精度依赖于传播模型的设计
- 未来方向
  - 表达学习将成为主流
  - 稀疏模型是发展方向

# 内容提纲

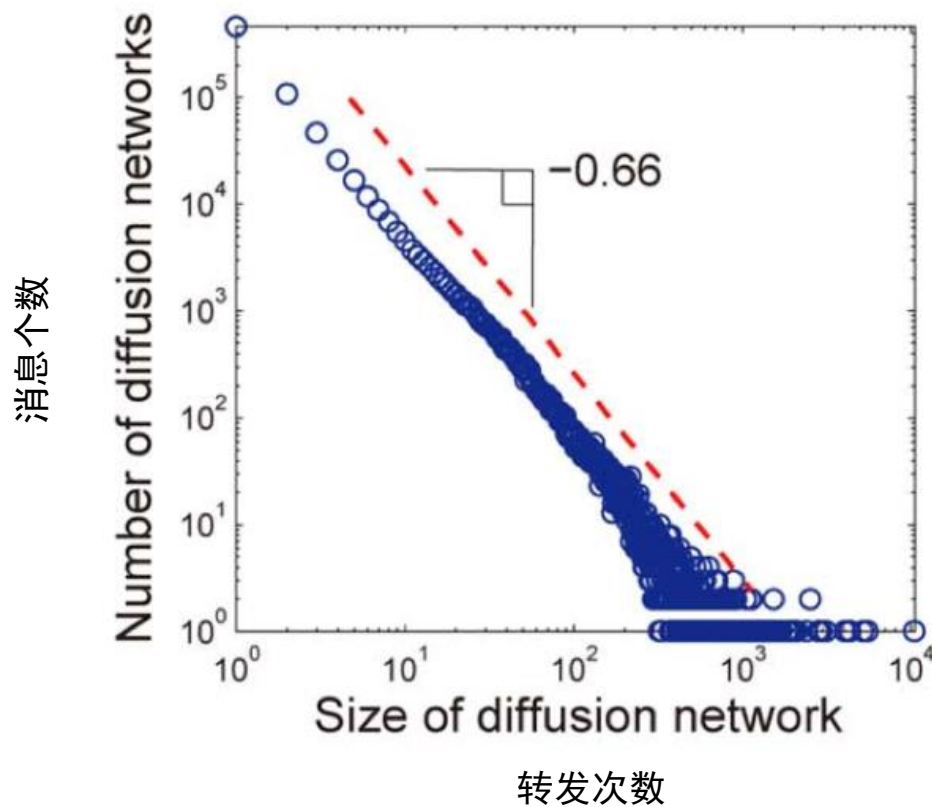
- 信息传播预测
  - 信息传播模型
  - 影响最大化
  - 传播网络推断
  - 流行度预测

# 流行度预测问题

## ■ 问题描述

- 给定一个对象一段时间内 ( $t_i$ ) 的群体关注情况, 预测其最终流行度
  - 微博消息转发次数预测
  - 学术论文引用次数预测
  - 网页导入链接数预测
  - 电影票房预测
  - .....

# 流行度的幂率分布



# 流行度预测问题

## ■ 三类数据

□  $t_1, t_2, \dots, t_i, \dots, t_n$

■ 页面访问, 搜索日志, .....

□  $\langle u_1, t_1 \rangle, \langle u_2, t_2 \rangle, \dots, \langle u_3, t_3 \rangle, \dots, \langle u_n, t_n \rangle$

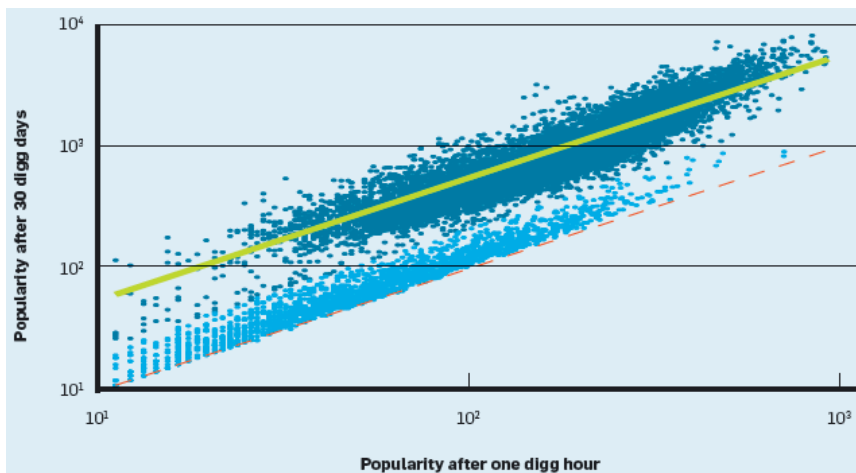
■ 流行病、微信, .....

□  $\langle u_1, v_1, t_1 \rangle, \langle u_2, v_2, t_2 \rangle, \dots, \langle u_3, v_3, t_3 \rangle, \dots, \langle u_n, v_n, t_n \rangle$

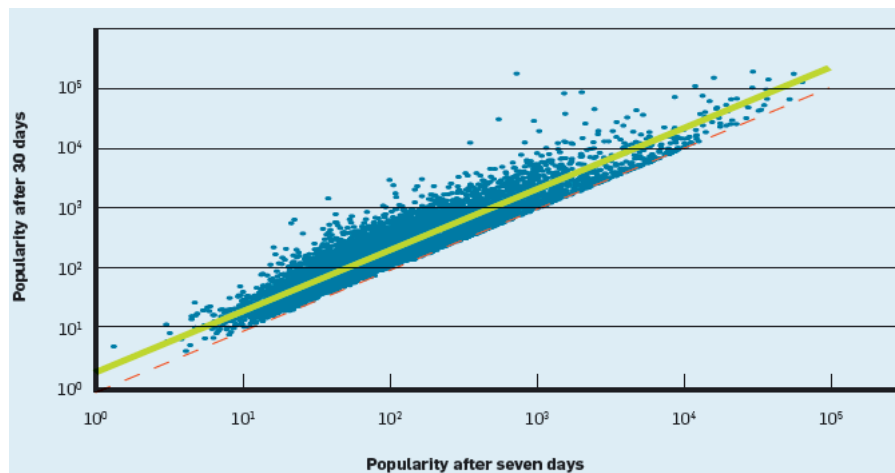
■ 微博、论文引用, .....

# 基于时序分析的预测

## ■ 流行度在时间上呈现对数自相关性



Digg



Youtube

## ■ 预测模型

### □ 乘性模型

$$\ln N(t_r) = \ln N(t_0) + \sum_{\tau=t_0}^{t_r} \eta(\tau)$$

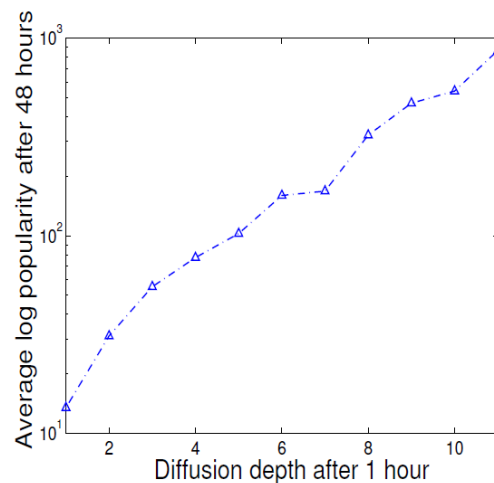
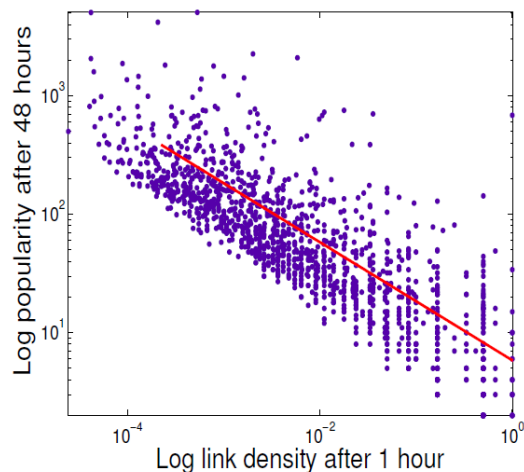
影响转发次数的因素

$t_r$ 时刻的累积转发次数

$t_0$ 时刻的累积转发次数

# 基于结构多样性的预测

- 利用早期传播者间的社会关系网络（G）或者传播树（T）的结构特性进行预测
  - G的连边密度：密度越低，流行度越大
  - T的深度：深度越大，流行度越大



## ■ 预测模型

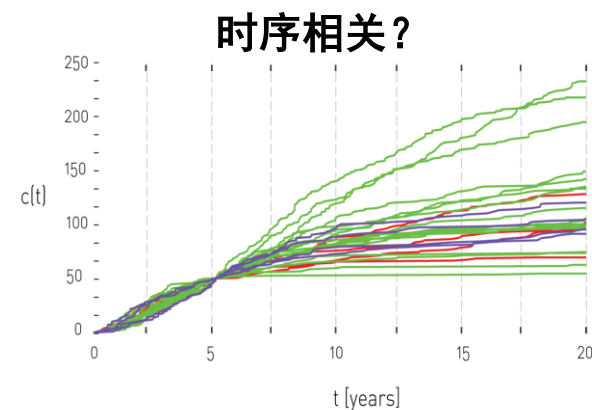
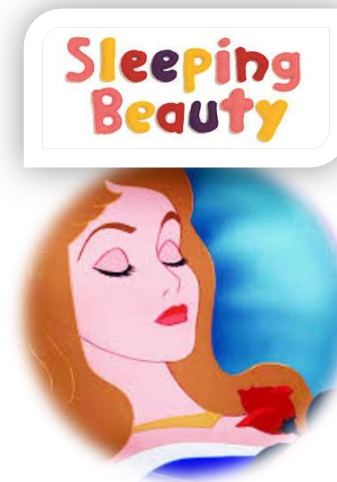
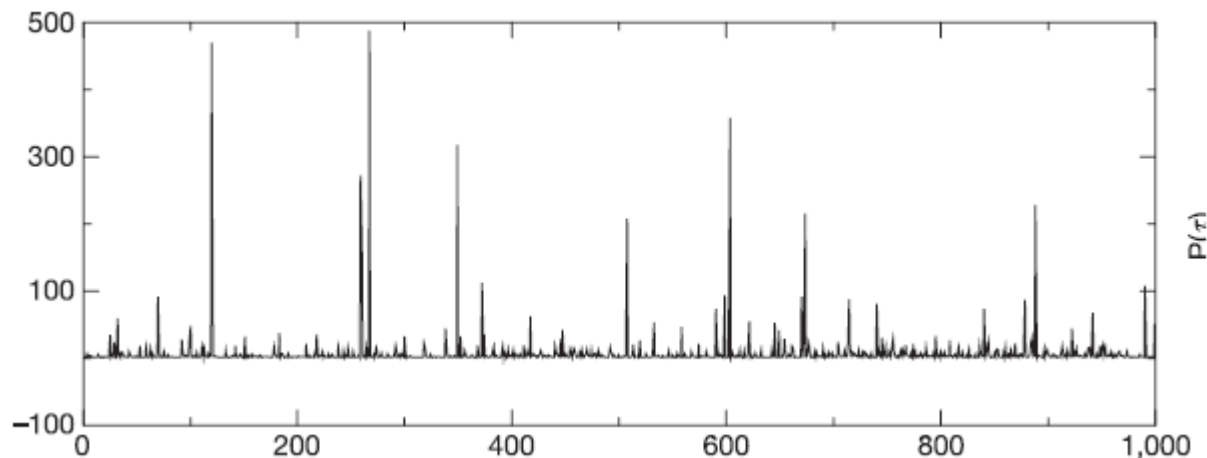
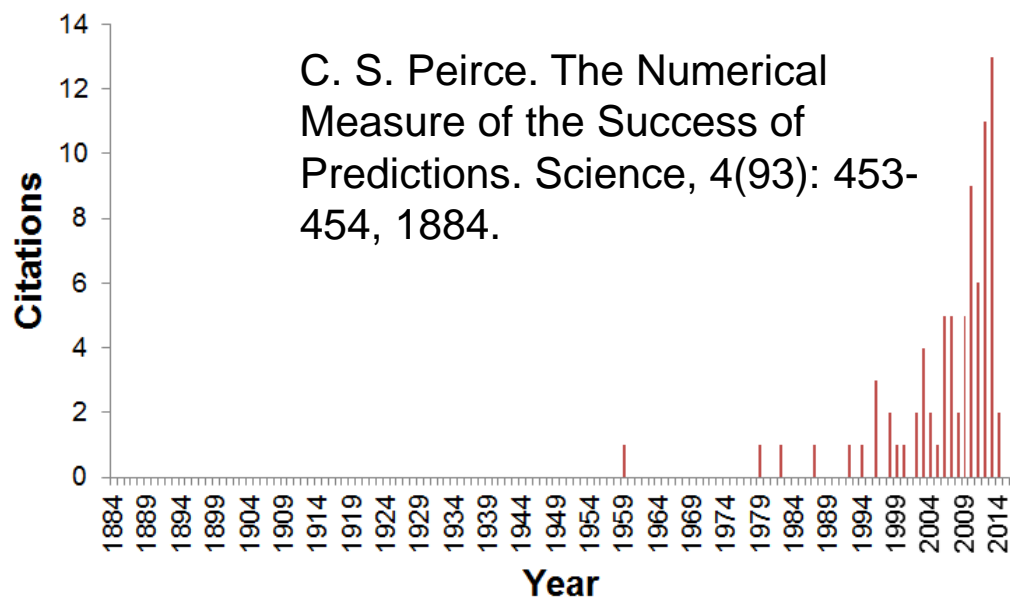
$$\ln \hat{p}_k(t_r) = \alpha_1 \ln p_k(t_i) + \alpha_2 \ln \rho_k(t_i) + \alpha_3$$

— G的连边密度

—  $t_r$ 时刻的累积转发次数

—  $t_0$ 时刻的累积转发次数

# 流行度预测问题



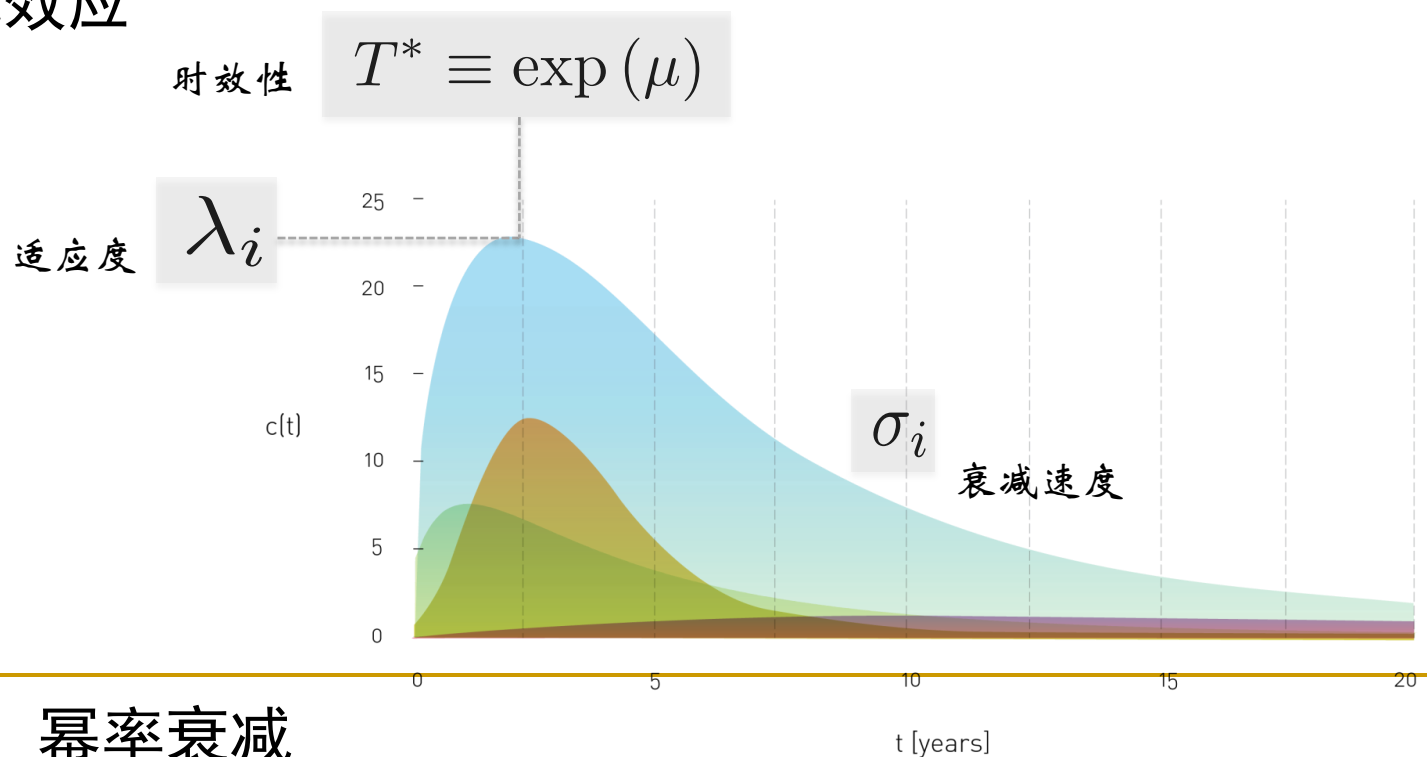
阵发：时序无尺度



# 建模传播过程进行流行度预测

## ■ 基于自增强泊松过程的流行度预测

- 富者愈富
- 适者生存
- 老化效应



# 自增强泊松过程

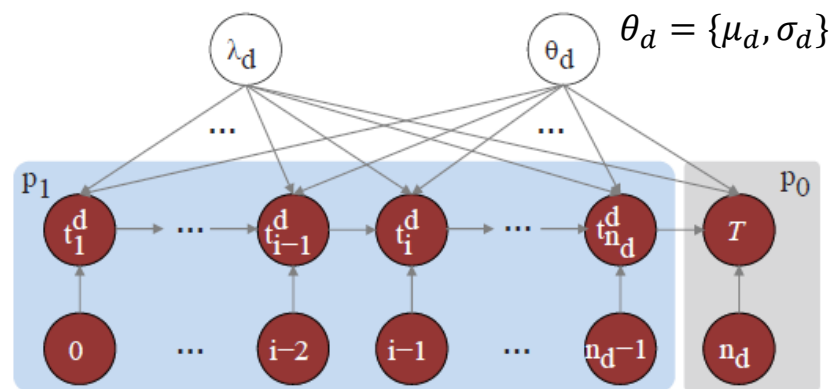
## ■ 自增强泊松过程

$$x_d(t) = \lambda_d f_d(t; \theta_d) i_d(t)$$

消息自身的吸引力

当前转发的次数

时效性：随时间衰减



$$i_d(t) = m + i - 1$$

## 最大似然参数估计

$$\begin{aligned} \mathcal{L}(\lambda_d, \theta_d) &= p_0(T|t_{n_d}^d) \prod_{i=1}^{n_d} p_1(t_i^d|t_{i-1}^d) \\ &= \lambda_d^{n_d} \prod_{i=1}^{n_d} (m + i - 1) f_d(t_i^d; \theta_d) \times \\ &\quad e^{-\lambda_d((m+n_d)F_d(T;\theta_d) - \sum_{i=1}^{n_d} F_d(t_i^d; \theta_d))} \end{aligned}$$

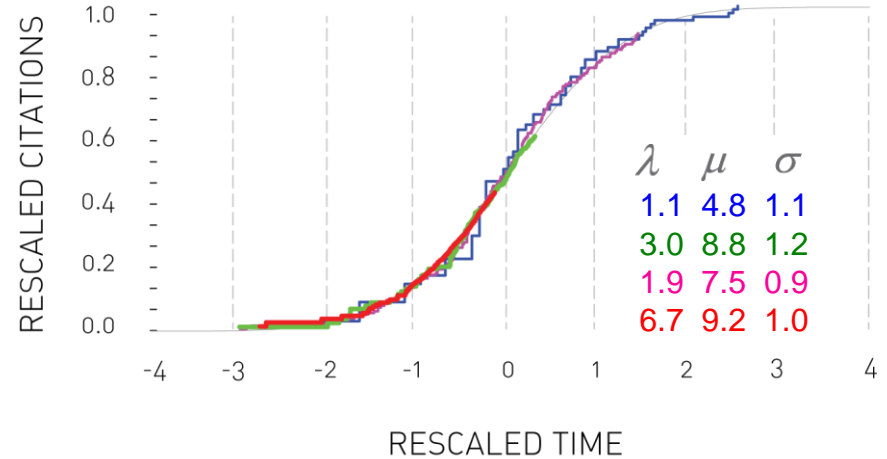
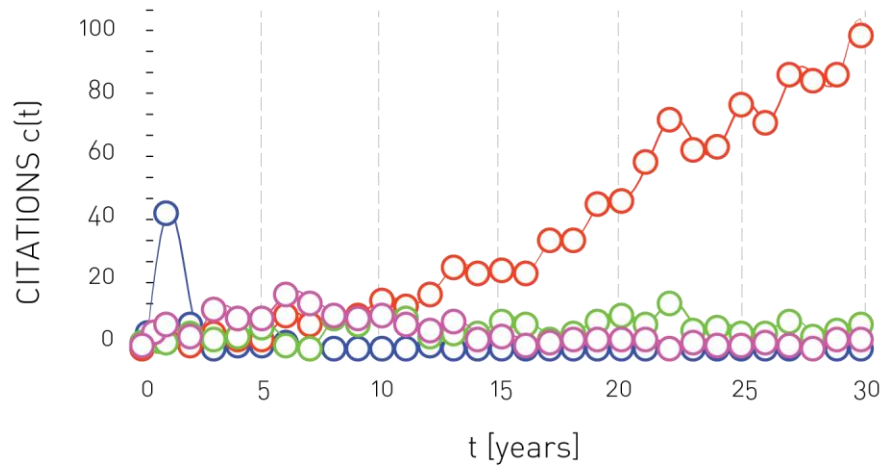
## 预测函数

$$\frac{dc^d(t)}{dt} = \lambda_d f_d(t; \theta_d) (m + c^d(t))$$

$$c^d(t) = (m + n_d) e^{\lambda_d^* (F_d(t; \theta_d^*) - F_d(T; \theta_d^*))} - m$$

# 自增强泊松过程

## 案例



$$\begin{aligned} \tilde{t} &\equiv (\ln t - \mu_i) / \sigma_i \\ \tilde{c} &\equiv \ln(1 + c_i^t / m) / \lambda_i \end{aligned} \longrightarrow \tilde{c} = \Phi(\tilde{t})$$

Bonner & Fisher, *Linear magnetic chains with anisotropic coupling*, Physical Review (1964)

Hohenberg & Kohn, *Inhomogeneous electron gas*, Physical Review (1964)

Bardakci et al. *Intrinsically Broken  $U(6) \otimes U(6)$  Symmetry for Strong Interactions*, Physical Review Letters (1964)

Berglund & W.F. Spicer, *Photoemission studies of copper and silver: Theory*, Physical Review (1964)

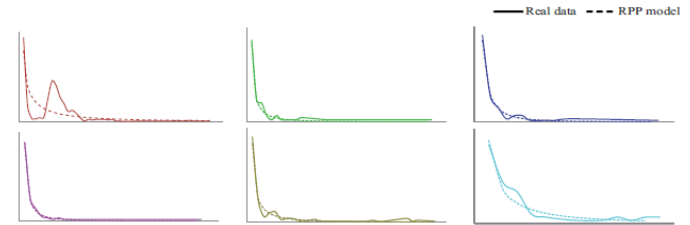
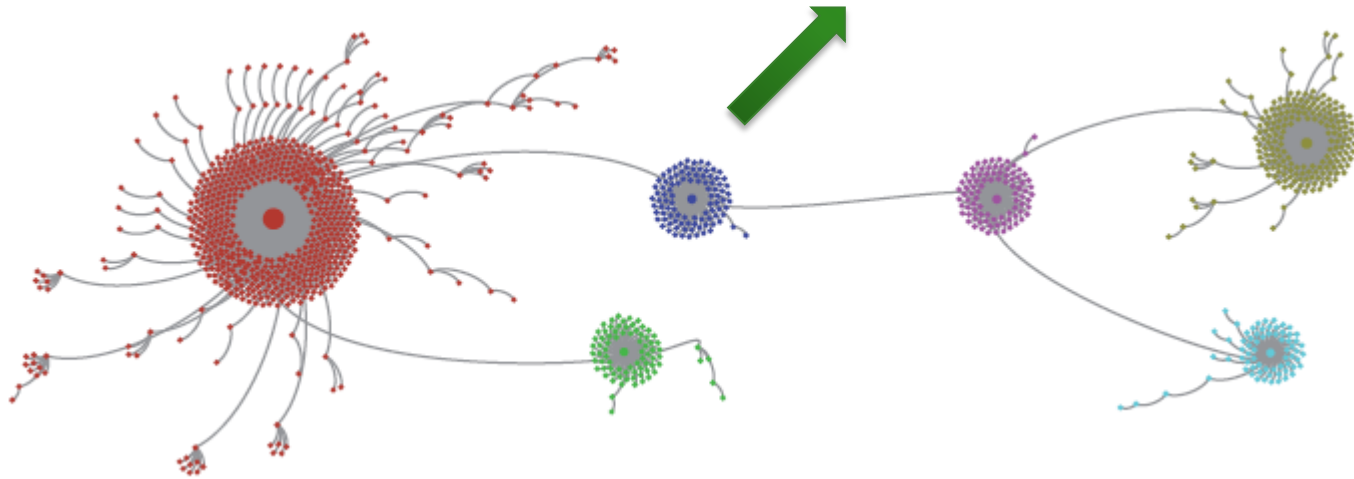
# 自增强泊松过程的扩展

$$x_d(t) = \lambda_d f_d(t; \theta_d) i_d(t)$$

- Replace the relaxation function with other form of functions  
[Gao et al., WSDM 2015]  
e.g., log normal  $\rightarrow$  exponential or power law function
- Replacing the “rich-gets-richer” mechanism with observed visibility [Zhao et al., KDD 2015]  
e.g., Number of retweeters  $\rightarrow$  Follower count of each retweeter
- Mixture of RPP to model multiple diffusion  
[Gao et al., WWW 2016]

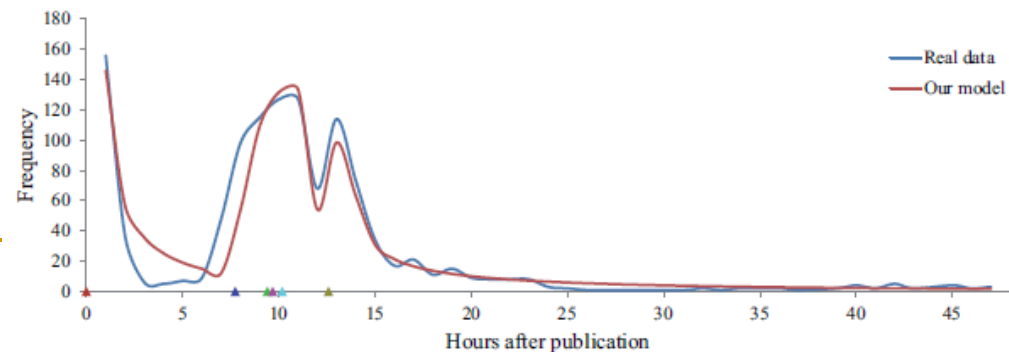
# 混合自增强泊松过程

Diffusion process with multiple stages



$$c(t) = n * \exp \left( \int_T^t \sum_{l=1}^k \lambda_l f(s - \tau_l; \theta_l) ds \right)$$

Each component is a RPP model

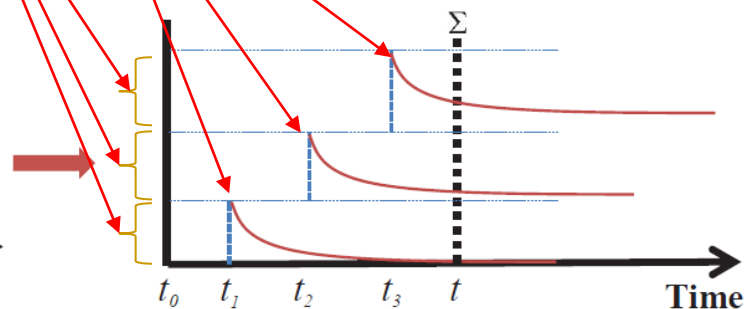
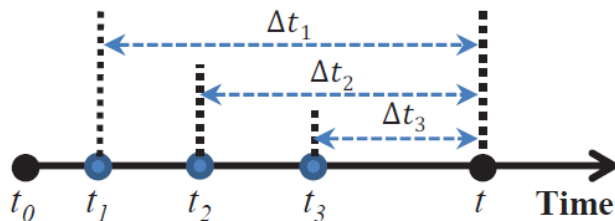


# 自激励Hawkes过程

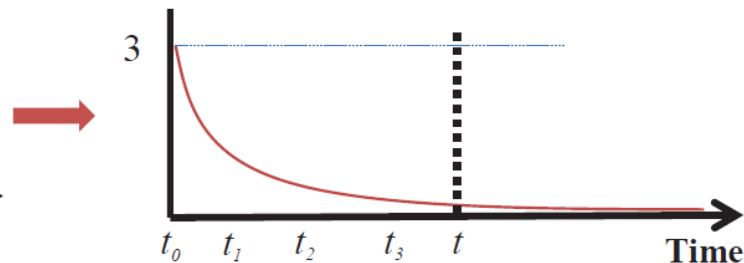
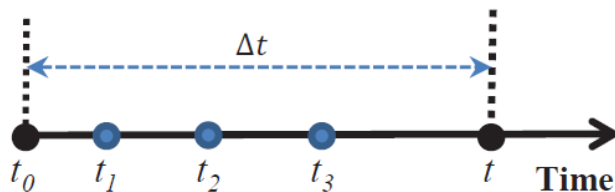
Attractiveness or infectiousness of message  $\hat{p}_t = \frac{R_t}{\sum_{i=0}^{R_t} n_i \int_{t_i}^t \phi(s - t_i) ds}$

Rate function:  $\lambda_t = p_t \cdot \sum_{t_i \leq t, i \geq 0} n_i \phi(t - t_i), \quad t \geq t_0$

Self-Excited Hawkes Process

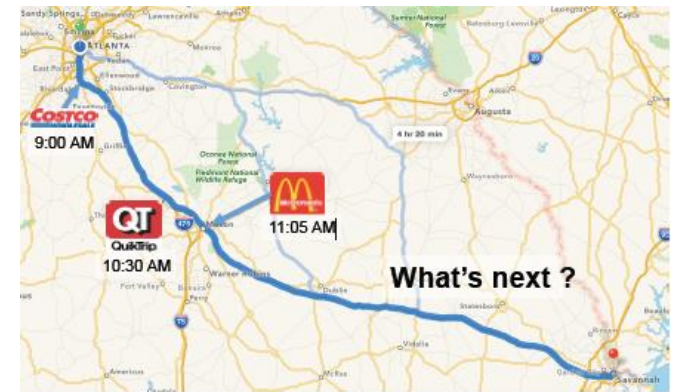
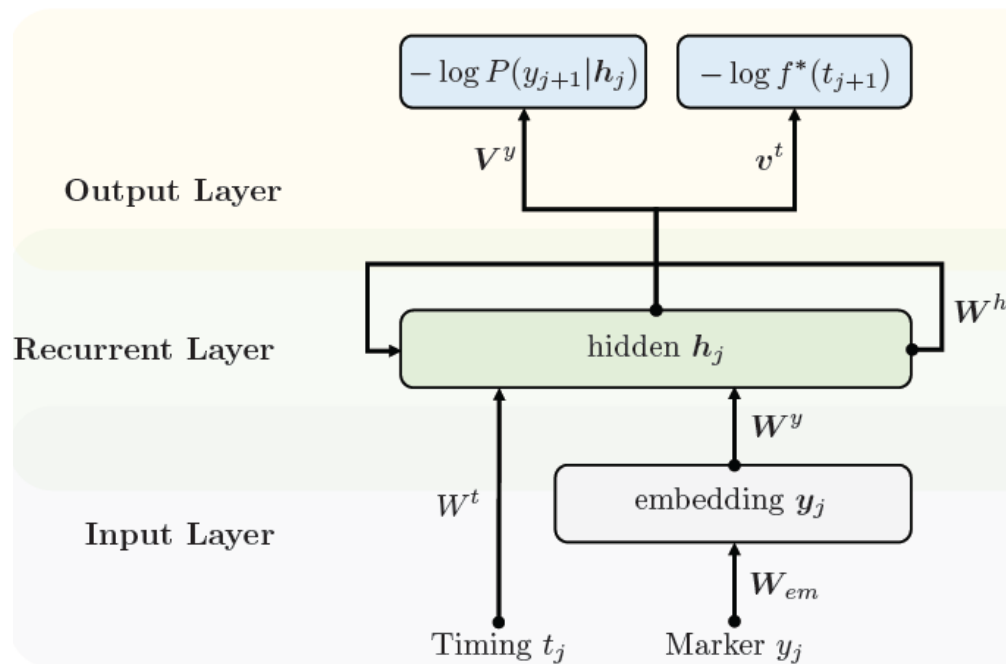


Reinforced Poisson Process



# RNN: Recurrent Neural Network

Idea: Learning the rate function from data, instead of human-defined rate function



- ✓ Embedding event history to vector
- ✓ Learn rate function within the framework of marked point process

$$\lambda^*(t) = \exp \left( \underbrace{v^{t\top} \cdot h_j}_{\text{past influence}} + \underbrace{w^t(t - t_j)}_{\text{current influence}} + \underbrace{b^t}_{\text{base intensity}} \right)$$

# 流行度预测小结

- 网络信息传播的热点
- 涉及因素多
  - 传播模型
  - 网络结构
  - 群体行为
  - 外部因素
  - 信息之间的相互影响
  - .....



# 参考文献

- D. Kempe, J. Kleinberg, E. Tardos. Maximizing the spread of influence through a social network. KDD 2003.
- J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance. Cost-effective outbreak detection in networks. KDD 2007.
- S. Cheng, H. Shen, J. Huang, G. Zhang, X. Cheng. StaticGreedy: Solving the Scalability-Accuracy Dilemma in Influence Maximization. CIKM 2013.
- G. Szabo, B. A. Huberman. Predicting the popularity of online content. Communications of ACM, 53: 80-88, 2010.
- J. Ugander, L. Backstrom, C. Marlow, J. Kleinberg. Structural diversity in social contagion. PNAS, 109: 5962-5966, 2012.
- L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan. Group formation in large social networks: membership, growth, and evolution. KDD 2006.
- J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, A. Vespignani. Characterizing and modeling the dynamics of online popularity. Physical Review Letters 105:158701, 2010.

谢谢各位同学  
三周来的陪伴！

