

# 2012 年第五届“认证杯”数学中国 数学建模网络挑战赛

题 目 基于 SNS 的消息推广效应及规划模型

关 键 词 社交网络 消息传播模型 分布密度函数

## 摘 要:

本文以碎片化趋势下的奥运会商业模式为背景,研究了 SNS 社交网络中的消息传播模式,用户对消息的反应行为,分析了消息传播过程与机制。结合本文结合传播理论、网络拓扑结构将消息传播问题抽绎成社交网络中的信息流动模型,对网络节点及其间的消息传播过程进行了划分,并在此基础上基于微分方程模型、图论、传染病动力学构造了一个社交网络消息传播模型。基于总体功能与架构,将整个模型划分为二次传播模型,消息蔓延模型,规划模型三个子模型。

通过构建二次传播中消息转发比率演化模型,考虑了传播过程中的边际转发率递减规律,结合微分方程、传染病动力学等,研究了消息未知用户,传播用户,仅接收用户分布密度增长加速度。通过节点度的引入,用 SPSS 分析处理出网络结构特征数据,带入模型中结合 C++编程得到中间数据,再导入 MATLAB 中分析,拟合出了传播节点、未知节点、只读节点,三类节点分布密度随时间的变化趋势图。

消息蔓延模型以二次传播模型为基础,从传播过程和时间两个维度对模型进行分解,得出基于不同基数的求解公式。并通过 C++分层模块化实现,拟出基于粉丝基数、粉丝增加数,参数化的实现过程,并进行迭代求总的蔓延效应。通过对不同粉丝基数和粉丝增加数的对比分析,发现粉丝基数对于消息蔓延效应的影响大于粉丝增加数。本部分还通过 SPSS, excel 数据分析,统计计算出其粉丝重复度,带入模型消除重复情况对结果的影响。

基于上述两个子模型,提出了规划模型,带入数据用 MATLAB 求解出最优方案。此外文章还提出了利用幂律分布、BA 模型,均匀分布等对模型进行改进的方案,以及随时间变化的消息蔓延模型的思路。最后文章基于分析结果,对如何利用社交网络平台进行宣传 and 营销提出了建议。

# 基于 SNS 的消息推广效应及规划模型

## 一、 问题重述

随着网络的不断开放和技术的不断先进，数字化已成为当今潮流，它让传播渠道、受众的注意力、品牌营销方式甚至一切都碎片化了，观众不在只关注电视，他们利用社交网络可以获得更加丰富的比赛信息和网友的评论。这也为更多的企业提供了在奥运期间宣传自己的机会。

一家企业想利用社交网络在奥运会期间为企业做宣传，在距离奥运会开幕的100天，一个专业推广者新增粉丝的平均速度为500个/天，这些粉丝会把消息转发分享给自己的粉丝们，普通网络用户新增粉丝的平均速度为20个/天。

第一个需要解决的问题是建立相关的数学模型以及分析Twitter用户的链接分布来预测奥运会开始后，一条含有企业广告的奥运会消息可以被观看到的人数。

第二个需要给企业制定一份合理的用人方案。假设企业产品在社交网络中潜在用户大约有2亿人，要求广告宣传覆盖潜在用户的40%，需要雇佣的专业社交网络推广者人数；假设专业推广者工资是500元/天，网络上可以雇佣兼职宣传者，工资为50元/天，相应的新增粉丝的平均速度为35个/天，综合考虑经济效益，制定出最佳用人方案。

## 二、 符号说明

符号	说明
$P_1$	未知节点转发消息，转化为传播节点的概率；
$1-P_1$	未知节点转化为只读节点的概率；
$P_{ww}^j$	表示节点 j 在 $[t, t + \Delta t]$ 时段内处于未知状态的概率；
$P_{wc}^j$	表示节点 j 从未知状态转移到转发状态的概率；
$P_{wg}^j$	表示节点 j 从未知状态转移到只读状态的概率；
H	表示在 t 时刻节点 j 的邻接节点中转发节点数量；
$L(m, t)$	表示在 t 时刻从具有 m 条边的未知节点连接到一个转发节点的概率；
$p(C_m   W_m)$	表示一个拥有 $m'$ 条边的节点在它连接到一个度为 m 的未知节点的情况下，处于转发状态的概率；

$p^c(m',t)$	表示在 t 时刻度为 $m'$ 的转发节点的密度;
$N(m,t)$	在 t 时刻社交网络中度为 m 的节点总数;
$w(m,t)$	在 t 时刻网络中度为 m 的未知节点的数量;
$c(m,t)$	在 t 时刻网络中度为 m 的转发节点的数量;
$g(m,t)$	在 t 时刻网络中度为 m 的只读节点的数量;
m	普通网络用户和兼职宣传员的原始粉丝基数;
$m'$	专业推广者的原始粉丝基数;
$\Delta S_t$	第 t 天由普通网络用户引起的新增消息转发者;
$\Delta K_t$	第 t 天由普通网络用户引起的新增只读者;
$\Delta T_t$	由第 t 天新增的 500 粉丝在剩余的 $100-t+1$ 天内总的消息接收者;
$P(m')$	度为 $m'$ 的节点的分布函数
J	总的普通网络用户接收者的数目;
X	招聘专业推广者人数;
Y	兼职宣传者人数

### 三、问题假设

1. 由转发引起的时间延迟可忽略, 用户接收到携带广告的奥运新闻后可以及时将此消息转发, 且他的粉丝可以及时收到该条新闻并看到其中的广告;
2. 假设粉丝消息接收方式仅为通过状态更新看到奥运新闻广告, 即不考虑进入收听者主页, 通过查看其历史状态记录等方式。
3. 普通网络用户转发消息当天, 新增粉丝行为发生在转发消息之前, 即当天新增粉丝都可收到其动态 (转发消息)。
4. 假设所有网络普通用户的转发行为都发生在第一次看到该条新闻之后, 此后看到同一条新闻, 用户将不再转发此新闻。因此, 消息转发的时间都为首次看到新闻的时间。专业推广者和兼职宣传者因任务要求, 需要每天都发布或者转发该条新闻。
5. 假设专业推广者的粉丝均为普通网络用户。
6. 假设企业具有选择偏好, 都偏向于雇佣具有较多粉丝的专业推广者。专业推广者分布于粉丝较多的网络用户群中。兼职宣传者粉丝基数和普通网络用户一样; 专业推广者, 兼职宣传者, 普通网络用户的差异表现在原有粉丝基数和粉丝增长速度。

## 四、问题分析

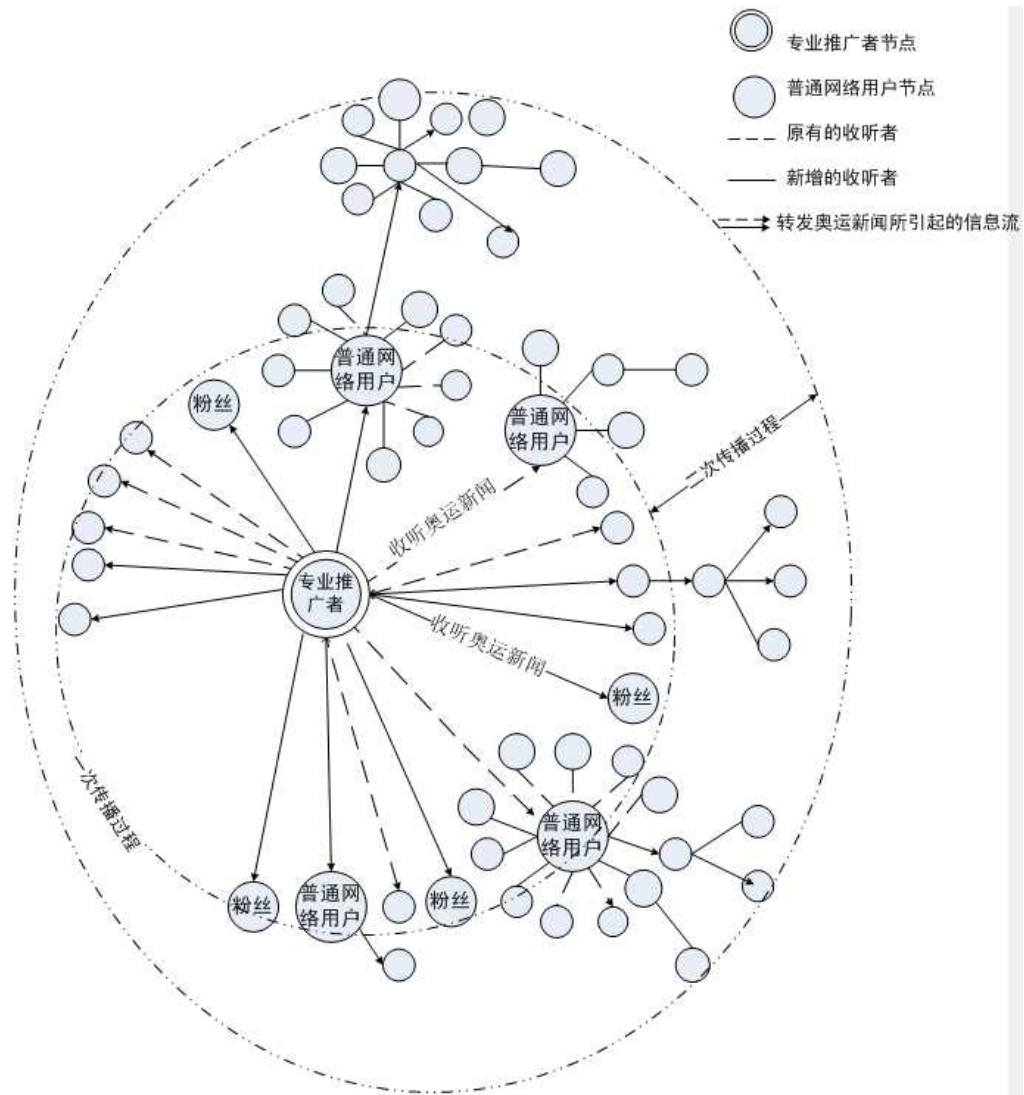
### 4.1 社交网络中信息传播模式分析

携带广告的奥运新闻在社交网络中的传播模型可以抽象为，信息在社交网络中节点（专业推广者，普通网络用户）间的流动过程。社交网络中的节点分为专业推广者节点和普通网民节点。消息的传播方式也有两类。

①一次传播。以专业推广者为中心，由专业推广者发布或者转发奥运新闻。从而将广告信息直接推向其粉丝（普通网民）。这种奥运新闻及广告由专业推广者直接到普通网民的流动过程，叫做消息的一次传播过程。专业推广者普通网络用户。

②二次传播。普通网民通过转发该条新闻，从而将广告信息推给其粉丝，即由专业推广者引起的，经普通网民转发的普通网民对消息的接收，叫做二次传播过程。

专业推广者  $\cdots\cdots$  普通网络用户  $\longrightarrow$  普通网络用户。



## 社交网络传播模式

上图专业推广者一次转发所引起的传播效益图。图中“虚线”代表专业推广者或者普通用户与其原有粉丝的收听关系。“实线”代表专业推广者或普通用户与其新增粉丝的新增收听关系。信息的推动过程表现为新闻的首次发布或者转发引起的动态更新。

专业推广者的首次发布信息过程,其接收者为所有粉丝(粉丝基数+新增粉丝),转发者考虑范围也为所有粉丝,而在之后的重复转发过程中,其接受者仍为所有粉丝,但转发者考虑范围仅为新增粉丝(基于假设4)。

### 4.2 社交网络中节点群演化模式

在社交网络中,一个人发布的消息会被其粉丝看到,并以一定的概率转发、传播。同时,若其粉丝对其内容不感兴趣则只阅读新闻并不转发消息,此条信息流到此终止。本文把 SNS 网络上的所有相关用户抽象为节点,用户之间的收听关系则可以抽象地用节点之间的边来表示,消息沿着边传播。根据用户节点是否阅读到携带广告的奥运新闻,以及是否转发该条新闻,将社交网络中的所有用户节点分为三类。

①未知节点,该节点还没有接收过某条信息,并有可能接收消息。

②只读节点,该节点已经接收了来自于其邻接节点转发的消息,但是不准备转发消息。

③转发节点,该节点接受了来自其收听节点的消息,并通过转发将此消息传播给他的邻接节点;

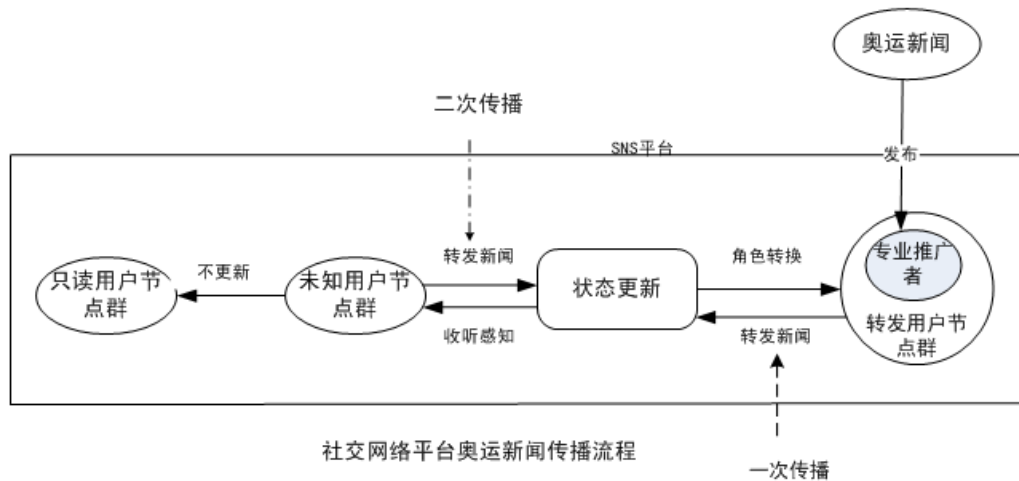
这三类节点分别对应三类用户群。

①未知用户节点群。该集合中的元素为没有阅读到带广告的奥运新闻的用户。

②只读用户节点群。该集合中的所有用户均阅读过新闻,但不转发新闻。

③转发用户节点群。该群中所有用户均阅读过并且转发了携带广告的新闻。

初始时,企业刚刚在奥运新闻中植入广告,SNS(社交关系网络)中所有用户均属于未知用户群。企业通过专业推广员在线发布新闻,使其进入 SNS 平台。新闻以专业推广员动态更新的方式传送到其粉丝(即一次传播过程)。专业推广员的粉丝在收听感知新闻后,状态激活。若该粉丝不转发此新闻则其由原来的未知状态转为只读状态,进入只读用户群。若该粉丝转发此新闻,将新闻推送给他自己的粉丝(二次传播过程)后,其状态变为转发态,其角色也转化为转发用户进入转发用户群。



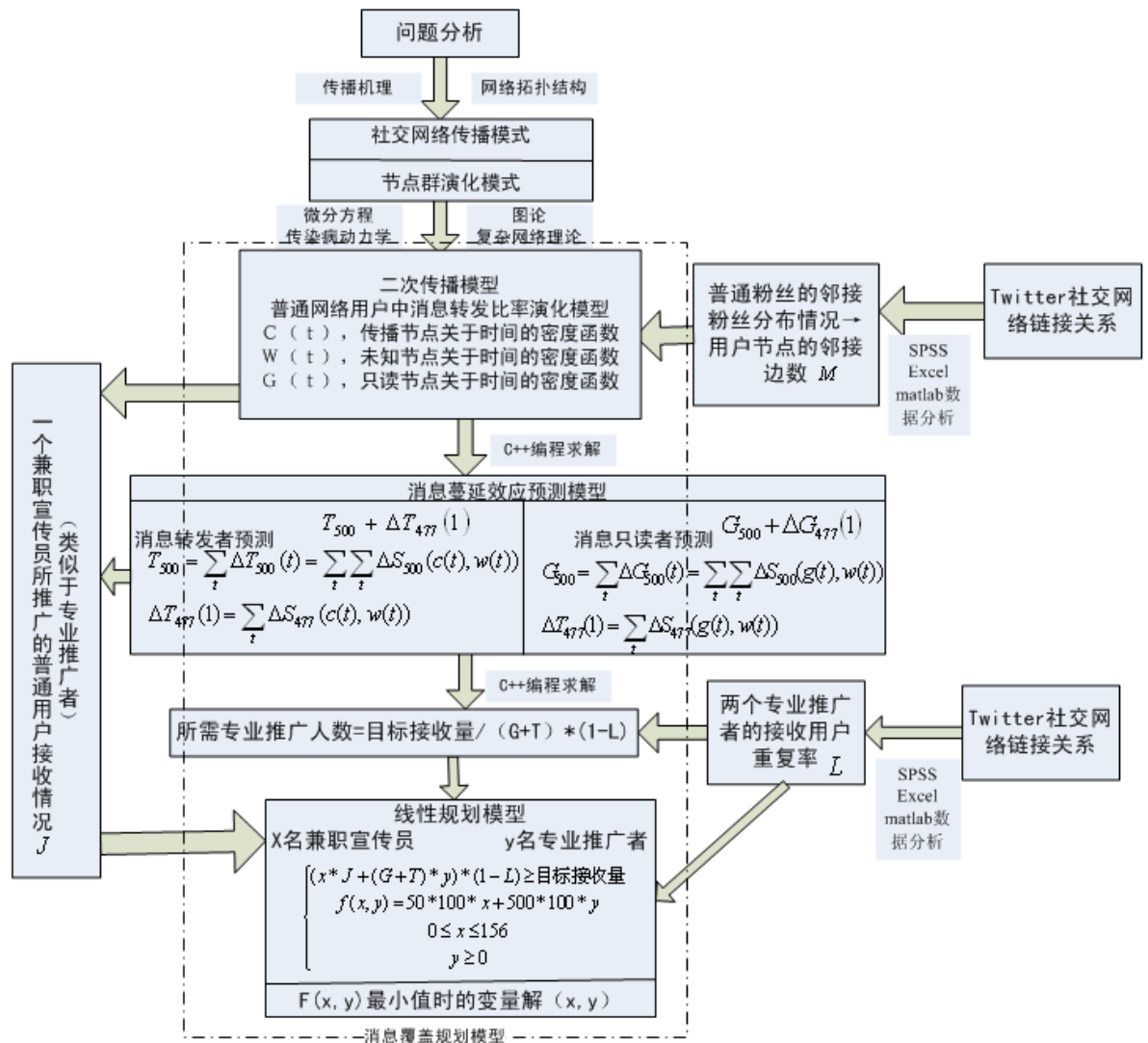
### 4.3 模型选择与解题思路

在前述社交网络信息传播模式和节点群演化模式的基础上，考虑到节点间的相互转化关系，信息的流动过程类似于传染性疾病的传染模型（SIR 模型）。以转发为方式的消息传播类似于以感染为方式的传染病传播模型。未知用户群可以看做未感染人群，具有感染疾病（接收到消息）和传播疾病（转发消息）的可能。只读用户群类似于免疫人群，他们感染过疾病（接受到过消息）但具有免疫力不会在传播疾病（转发消息）。转发用户群类似于传染人群，他们通过转发将消息传播给临近节点。

在 SIR 模型的基础上考虑到节点度（用户链接关系数）和传播机理（边际转发效率递减），结合网络拓扑结构和传染病动力学，构建出了消息传播的动力学方程组。在此方程组的基础上通过 matlab、C++编程等模拟出未知节点、只读节点和转发节点关于时间  $t$  的密度函数  $W(t)$ 、 $G(t)$ 、 $C(t)$ ，即二次传播过程中消息转发比率演化模型，简称二次模型。

基于此二次模型，以一次传播过程的基数为依据，将一次传播分为原有粉丝引发的二次传播和新增粉丝引发的二次传播，并对他们分别代入消息蔓延效应预测模型中求出消息收听者的蔓延情况。由于消息转发密度  $C(t)$  和只读密度  $G(t)$  演化情况不同，消息蔓延效应模型又分为消息转发者预测模型和消息接受者预测模型。到此已求出以推广能力和原有粉丝数为基数的消息蔓延模型。

接着将专业推广者和兼职宣传者的推广能力和粉丝基数代入模型中求解出单个推广人员或者兼职宣传者分别引起的消息蔓延规模。最后再以单个消息蔓延规模为基础，结合粉丝重复率，成本约束条件等构建线性规划模型求解最优组合。此过程称作消息覆盖规划模型。



求解程序总流程图

## 五、模型建立

### 5.1 二次传播中消息转发比率演化模型（简称二次传播模型）

#### 5.1.1 模型的建立

##### 5.1.1.1 概念界定

**节点**：二次传播过程中的普通网络用户。分为未知节点，传播节点，只读节点。

**邻接/边**：节点之间的收听关系。分为节点与其收听者的主动收听关系和与粉丝间的被动收听关系。

**度**：粉丝节点的邻接数量。包括与收听者的消息入度和与粉丝的消息出度。

传播概率：未知节点接受到消息后，转发消息，变成转发节点的概率。

二次传播模型：即二次传播中消息转发比率演化模型，此模型主要研究普通用户间消息传播过程（即二次传播过程）中，传播比率随时间的演化关系。

消息传播规则：一个转发节点将消息转发给一个未知节点后，未知节点会以  $P_1$  的概率转发此条消息，转化为传播节点，以  $1-P_1$  的概率转化为只读节点。

### 5.1.1.2 建立模型

节点  $j$  在  $t$  时刻处于未知状态， $P_{ww}^j$  表示节点  $j$  在  $[t, t + \Delta t]$  时段内处于未知状态的概率， $P_{wc}^j$  表示节点  $j$  从未知状态转移到转发状态的概率， $P_{wg}^j$  表示节点  $j$  从未知状态转移到只读状态的概率，且  $P_{wc}^j = 1 - P_{wg}^j$ ，其中

$$\square P_{ww}^j = (1 - \Delta t p_1)^h \quad (1)$$

$h = h(t)$  表示在  $t$  时刻节点  $j$  的邻接节点中转发节点数量。节点  $j$  含有  $m$  条边， $h$  是具有如下二项式分布的随机变量：

$$\Pi(h, t) = \binom{m}{h} L(m, t)^h (1 - L(m, t))^{m-h} \quad (2)$$

$L(m, t)$  表示在  $t$  时刻从具有  $m$  条边的未知节点连接到一个转发节点的概率，

$$L(m, t) = \sum_{m'} p(m' | m) p(c_{m'} | w_m) \approx \sum_{m'} p(m' | m) p^c(m', t) \quad (3)$$

$$P(m' | m) = \frac{m' P(m')}{\bar{m}}$$

$P(m')$  是度为  $m'$  的节点的概率分布函数， $\bar{m}$  为样本网络空间的节点平均度。

$P(m' | m)$  为度相关函数，表示度为  $m$  的节点邻接度为  $m'$  的节点的条件概率。

$p(c_{m'} | w_m)$  表示一个度为  $m'$  的节点邻接到一个度为  $m$  的未知节点条件下，转化为转发节点的概率， $p^c(m', t)$  表示在  $t$  时刻度为  $m'$  的转发节点的概率密度。

度为  $m$  的未知节点在  $[t, t + \Delta t]$  时段内处保持未知状态的平均转移概率  $p_{ww}^-(m, t)$ 。

$$p_{ww}^-(m,t) = \sum_{h=0}^m \binom{m}{h} (1-p_1\Delta t) L(m,t)^h * (1-L(m,t))^{m-h} = (1-p_1\Delta t L(m,t))^m$$

上面两式联合得到:

$$p_{ww}^-(m,t) = \left(1 - p_1\Delta t \sum_{m'} p(m'|m) p^c(m',t)\right)^m \quad (4)$$

设  $N(m,t)$  为在  $t$  时刻社交网络中度为  $m$  的节点密度,  $w(m,t)$ ,  $c(m,t)$ ,  $G(m,t)$  分别为在  $t$  时刻网络中度为  $m$  的未知节点、转发节点和只读节点的概率密度。  $\Delta w(m,t), \Delta c(m,t), \Delta g(m,t)$  分别表示相应密度的变化量。

$$w(m,t) + c(m,t) + G(m,t) = N(m,t) \quad (5)$$

$$\Delta w(m,t) = \Delta c(m,t) + \Delta G(m,t)$$

网络中度为  $m$  的未知节点的数量在  $[t, t + \Delta t]$  时段内的变化情况如下

$$\begin{aligned} w(m,t + \Delta t) &= w(m,t) - w(m,t)(1 - p_{ww}^-(m,t)) \\ &= w(m,t) - w(m,t) * \left[1 - \left(1 - p_1\Delta t * \sum_{m'} p^c(m',t) p(m'|m)\right)^m\right] \end{aligned} \quad (6)$$

以此类推, 可得到度为  $m$  的转发节点和只读节点的数量在  $[t, t + \Delta t]$  时间段内的变化情况, 如下:

$$\begin{aligned} C(m,t + \Delta t) &= C(m,t) + p_1 * W(m,t)(1 - p_{ww}^-(m,t)) \\ &= C(m,t) + p_1 * W(m,t) * \left[1 - \left(1 - p_1\Delta t * \sum_{m'} p^c(m',t) p(m'|m)\right)^m\right] \end{aligned} \quad (7)$$

$$\begin{aligned} G(m,t + \Delta t) &= G(m,t) + (1 - p_1) * W(m,t)(1 - p_{ww}^-(m,t)) \\ &= G(m,t) + (1 - p_1) * W(m,t) * \left[1 - \left(1 - p_1\Delta t * \sum_{m'} p^c(m',t) p(m'|m)\right)^m\right] \end{aligned} \quad (8)$$

由 (5), (6) 公式得到,

$$\frac{W(m, t + \Delta t) - w(m, t)}{N(m, t)\Delta t} = -\frac{W(m, t)}{N(m, t)\Delta t} \left[ 1 - \left( 1 - p_1 \Delta t * \sum_{m'} p^c(m', t) p(m'|m) \right)^m \right] \quad (9)$$

当  $\Delta t$  趋近于 0 时, 将 (9) 号公式右侧进行泰勒公式展开得到, 未知节点的密度随时间的变化关系  $w(t)$ :

$$\frac{\partial p^w(m, t)}{\partial t} = -mp_1 p^w(m, t) \sum_{m'} p^c(m', t) p(m'|m), \quad (10)$$

以此类推, 联立 (5), (7) 公式在  $\Delta t$  趋近于 0 可以得到, 转发节点的密度随时间变化的关系  $c(t)$ :

$$\frac{\partial p^c(m, t)}{\partial t} = mp_1^2 p^w(m, t) \sum_{m'} p^c(m', t) p(m'|m), \quad (11)$$

联立 (5), (8) 公式在  $\Delta t$  趋近于 0 可以得到, 只读节点的密度随时间变化的关系  $g(t)$ :

$$\frac{\partial p^g(m, t)}{\partial t} = mp_1 * (1 - p_1) p^w(m, t) \sum_{m'} p^c(m', t) p(m'|m) \quad (12)$$

公式 (10), (11), (12) 分别描述了未知节点密度函数  $W(t)$ , 转发节点密度函数  $C(t)$ , 只读节点密度函数  $G(t)$  随时间的变化情况。这三个公式称为节点的传播动力学演化方程组, 用于刻画整个网络中节点的演化力度

## 5.1.2 模型的求解

### 5.1.2.1 数据支撑

根据 twitter 社交网络的连接关系数据, 对第一列中的用户账号和其对应的粉丝账号导入 SPSS 中进行描述统计分析, 得出其频数数据。并以此为基础对频数的分布进行统计处理。spss 和 excel 对参考数据中的 835541 条记录进行分析, 统计出用户 ID 数和粉丝频数分布图等, 部分数据截图如下, 部分数据见附件。

## Frequencies

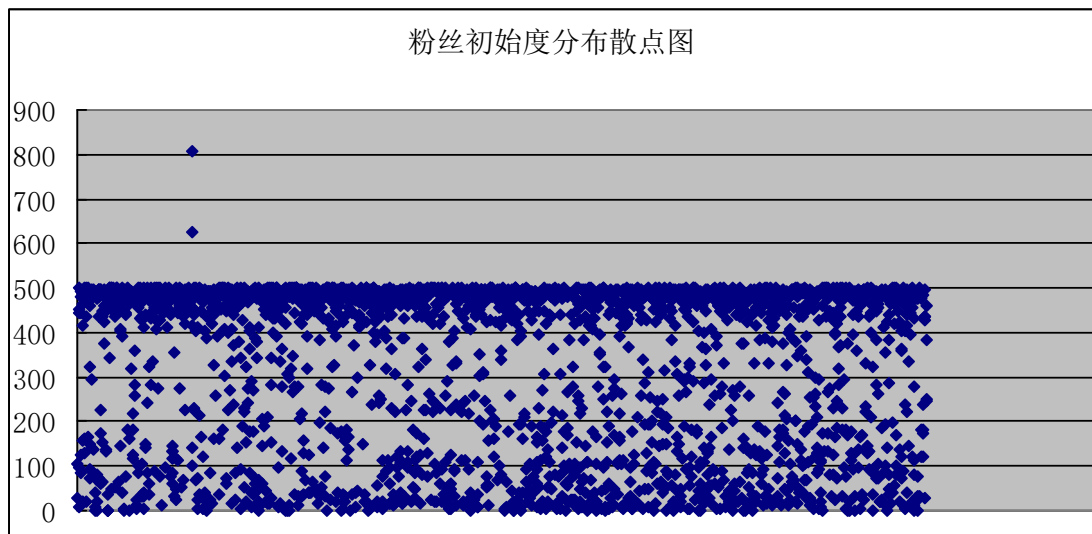
### Statistics

V1

N	Valid	835541
	Missing	0

粉丝数范围	频数分布
0~100	522
101~200	241
201~300	150
301~400	130
401~500	1458
600以上	2
总计:	2503

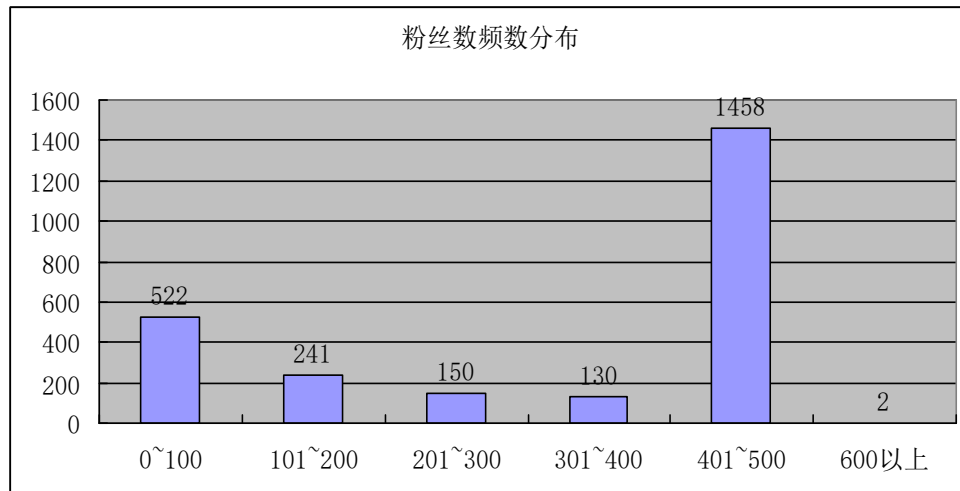
	Frequency	Percent	Valid Percent
Valid	104	.0	.0
_Breeanne	27	.0	.0
_winter	8	.0	.0
0_0kimjy	96	.0	.0
00__00	500	.1	.1
000webhosting	443	.1	.1
001dabrown	453	.1	.1
005dabrown	444	.1	.1
005matt	481	.1	.1
007brasil	85	.0	.0
007joectms	498	.1	.1
007Lifestyle	127	.0	.0
007LouiseOB	491	.1	.1
009onthebeach	466	.1	.1
00aidenrich	417	.0	.0
00joe	488	.1	.1
00SVT	20	.0	.0
00sw	157	.0	.0
01countrygirl01	131	.0	.0
01Natalie	481	.1	.1
01watch	500	.1	.1
03dentchamp	161	.0	.0
040Milad	485	.1	.1
07BondGirlXX	435	.1	.1
07forcada	22	.0	.0
099c	471	.1	.1



基于以上处理结果，对粉丝账号的整体分布情况求均值，得每个节点的平均粉丝度  $m = 835541/2503 = 333.8158 \approx 334$ ，此为普通网络用户和兼职宣传者的粉丝基数。考虑到企业的选择偏好（假设 7），企业偏向于雇佣粉丝数较多的专业推广者，故专业推广者主要分布于粉丝高频段集中段。从粉丝频数分布图可看出，频数为 401~500 的高频集中段用户数量高达 1458，占整个用户的 58.3%。故可将受雇佣的专业推广者化为归于此部分，求的其平均粉丝度为  $m=477$ 。此部分数据即为  $P(m')$ （度为  $m'$  的节点的分布函数）的函数数据，以二维数组（ $m'$ ，

$P(m')$  的形式为后面编程求解所用。

每个用户名 ID 进行粉丝数分段点频数分布统计如下表：



### 5.1.2.2 函数模拟

每个节点首次受到消息后转发概率  $P_1$ ，根据问题情景中海怪号（Mar Mostro）帆船发布所引起的 2.5 万互动信息中，60%转发了这张照片，此处  $P_1$  近似取值 0.6。  $P(m')$ （度为  $m'$  的节点的分布函数）的函数数据，为上述二维数组  $(m', P(m'))$ ，  $\bar{m}$  为数组中  $m'$  的平均值。对于  $m$ ，若为专业推广者则取上述高频分布段均数 477，普通用户和兼职宣传员则取全部分布段 334。

基于以上基础数据通过 C++编程，求的  $L(m,t)$ ，  $p^w(m,t)$ ，  $\frac{\partial p^w(m,t)}{\partial t}$ ，

$\frac{\partial p^c(m,t)}{\partial t}$  等中间函数的对应数据。再将上述中间数据导入 MATLAB 中，迭代运

行模拟出  $C(t), W(t), g(t)$  随时间的变化情况，即不同节点密度随时间的变化关系：  
 ( $W(t)$  为未知节点的密度，  $C(t)$  为转发节点的密度，  $g(t)$  为只读节点的密度)。（部分代码见附件）

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/847146164134006141>