第五届“认证杯”数学中国

数学建模网络挑战赛

承 诺 书

我们仔细阅读了第五届“认证杯”数学中国数学建模网络挑战赛的竞赛规则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的, 如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

我们允许数学中国网站(www.madio.net)公布论文，以供网友之间学习交流，数学中国网站以非商业目的的论文交流不需要提前取得我们的同意。

我们的参赛队号为：1981

参赛队员 (签名) ：

队员 1：项曾哲

队员 2：冯诚诚

队员 3：李熊俊喜

参赛队教练员 (签名)： 万华舰

参赛队伍组别：本科组

第五届“认证杯”数学中国

数学建模网络挑战赛

编 号 专 用 页

参赛队伍的参赛队号：（请各个参赛队提前填写好）：

竞赛统一编号（由竞赛组委会送至评委团前编号）：

竞赛评阅编号（由竞赛评委团评阅前进行编号）：

2012 年第五届“认证杯”数学中国
数学建模网络挑战赛

题　　目　　碎片化趋势下的奥运会商业模式

关 键 词 *Jarque-Bera* 法正态性检验　SIR 微分方程　IDM(α,β)模型　整数规划

摘　　要：

本文研究的是信息传播类问题，主要关于信息传播的预测和传播成本的优化。首先对所给的数据进行处理，并统计出总人数为 465023 人。利用 *Jarque-Bera* 法对粉丝数的分布进行正态性检验，然后根据所得出的数据特性建立起以下五个数学模型。

针对问题一，首先建立基于经典网络信息传播规则的 SIR 微分方程模型，定义接触率和不转发率，借助信息传播变化率的微分关系建模求解。然后考虑时序的离散不连续性特点，建立离散数学模型从概率期望的角度对预计人数作出预测。最后基于社交网络拓扑结构对信息传播的影响，建立传播仿真 IDM(α,β)模型进行仿真模拟预测。若以 2 亿用户为总体，三种模型最终的预测结果如下表：

模型	SIR 微分方程模型	离散数学模型	传播仿真 IDM(α,β)模型
预测人数	1.1196 亿	2 亿	0.7521 亿

对以上结果进行分析，SIR 微分方程模型和传播仿真 IDM(α,β)模型结果更符合实际情况。

针对问题二的第一小问，以专业推广者的人数为设计变量，以广告宣传能覆盖 2 亿潜在用户的 40%为约束条件，以专业推广者的人数最少为目标函数，建立人数最优化模型。用 Lingo 软件编程求解得至少需要 3 个专业推广者。

针对问题二的第二小问，先定义接受率 η，再以雇员（专业推广者和兼职宣传者）的人数和工作天数为设计变量，以广告宣传能覆盖 2 亿潜在用户的 40%为约束条件（即分享的次数与接收率的乘积大于 2 亿潜在用户的 40%），以总成本最少为目标函数建立用人方案最优化模型。用 Lingo 软件编程求解得到最优用人方案如下表：

雇员	工作天数	日工资（元）	总工资（元）	分享次数	覆盖人数
第一个专业推广者	100	500	50000	50550000	80102473
第二个专业推广者	100	500	50000	50550000	
第三个专业推广者	92	500	46000	42826000	
总和	292		146000	143926000	

专业推广者人数 p=3 人，其工作天数分别 n_1=100 天，n_2=100 天，n_3=92 天，宣传兼职者人数 q=0 人，　宣传成本 f=146000 元。

参赛队号 1981

参赛密码 ________

（由组委会填写）

所选题目 C 题

Abstract

Information Dissemination is studied in this paper, mainly concerned on the prediction of Information Dissemination and the optimization of Information Dissemination Cost. First, we deal with the given data, and statistics shows the total number of 465,023 people. According to the Jarque–Bera test, we do normality test on the distribution of the number of fans, then establish the following five mathematical models by the characteristics of the resulting data.

Regarding Question One, we establish the SIR differential equation model based on the classic network information dissemination rules firstly, defining the transmit rate and stay-still rate and solving it by differential equation model of information transmission's rate of change. Then consider the characteristics of the Discrete Fourier Transform, Discrete Mathematical Model established to forecast the expected number via Probability Expectations. Finally, based on the social network topological structure's influence on Information Dissemination, we do the simulation forecast by establishing the Simulation IDM(α,β) Model. The three models' final predict results showed in the following table:

Model	SIR Differential Equation Model	Discrete Mathematical Model	Simulation IDM(α,β) Model
Predicted Number	1.1196	2	0.7521

Analyzing the results above, the SIR Differential Equation Model and the Simulation IDM(α,β) Model come to the same result, more in line with the actual situation and accuracy higher.

Regarding Question Two's first asked question, set the number of professional promoter as design variables, set possibility to cover 40% of the 200 million potential customers as constraints, the minimum of professional promoters as Objective Function to establish a Optimization Model of People's amount. Lingo software solving it and said using three professional promoters is at least.

For Question Two's second asked question, defining the acceptance rate as η, the number and work days of employees (the professional promoters and part-time advocates) as design variables, the possibility that the advertising could cover 40% of the 200 million potential customers as constraints (the product of the transmit rate and stay-still rate is greater than 40% of the 200 million potential customers)the least total cost of as Objective Function to establish a Optimization Model of best employing scheme. The following table showed the best employing scheme solving by Lingo software:

Employee	Work days	Daily wages(yuan)	Total payment(yuan)	Share Times	People Covered
First professional promoter	100	500	50000	50550000	80102473
Second professional promoter	100	500	50000	50550000	
Third professional promoter	92	500	46000	42826000	
Summation	292		146000	143926000	

Professional promoters: p=3, Days each one worked for: n_1=100days,n_2=100days,n_3=92days, pluralists: q=0, Cost of Promotion: f=146000yuan.

1 问题重述

1. 1 问题背景

自从 1984 年美国洛杉矶奥运会以来，奥运会开始拥有其独特的商业模式。由于它是一台全球瞩目的体育盛会，因此成为了商业社会里企业最重要的展示舞台。每个行业中的奥运全球合作伙伴只有一家，而且赞助费用尤其是宣传费用将绝大多数企业排除在了奥运会之外。他们不甘心错过奥运会这个吸引大众眼球的巨大宣传机会，试图寻找新的新闻传播渠道。随着科技日新月异地进步，数字化将一切碎片化，快捷、自由、方便的社交网络登上了商业宣传的舞台，为更多的企业提供了在奥运期间宣传展示自己的机会。

1. 2 需要解决的问题

一家企业想利用社交网络在奥运会期间进行企业宣传。假设现在奥运会开幕还有 100 天，一个社交网络的专业推广者平均每天可以新增 500 个粉丝，普通网络用户平均每天可以新增 20 个粉丝，些粉丝会把推广者发布的和奥运会相关的所有信息都分享给自己的粉丝们。附件中为 Twitter 社交网站用户之间的链接关系的数据，用于发现用户组，以及分析 Twitter 用户的链接分布。

问题 1：建立数学模型，预测奥运会开始后，一条含有企业广告的奥运会新闻可以被多少人观看？

问题 2：假设企业产品潜在用户大约有 2 亿人，他们都在使用社交网络，企业希望广告宣传覆盖其中 40%的人群，至少需要雇佣几名专业社交网络推广者才能实现？假设专业推广者每天工资是 500 元。还可以从网络上雇佣兼职宣传者，每天仅需要付 50 元的工资，但是他们每天新增的粉丝数仅为 35 人，考虑到成本，请为企业制定一份合理的用人方案。

2 模型的假设及符号说明

3.1 模型假设

假设 1：社交网络上的用户总数在所研究的短时间段内保持不变。

假设 2：待预测的社交网络中的用户总数为 2 亿。

假设 3：已转发者不会再次转发相同的新闻，即用户只看一遍新闻。

假设 4：附件中的数据具有一般的代表性，可以作为对整体预测的依据。

假设 5：不考虑在所研究的时间段内用户粉丝数的动态减少。

假设 6：第二问中社交网络用户全为 2 亿潜在用户。

假设 7：新闻在社交网络中呈星型辐射式传播而不会形成传播闭合回路。

假设 8：专业推广者和兼职宣传者初始粉丝数为 0 人

3.2 符号说明

符号	说明
N	某社交网络中的总用户数
$S(t)$	未观看者在总用户数种所占比例
$I(t)$	已观看且传播者在总用户数种所占比例
$R(t)$	已观看但不传播者在总用户数种所占比例
λ	已观看者的转发率
μ	已观看者的不转发率
X_n	第n个单位时间C中观看到新闻的用户数即观看者数
M	第n个和第$n+1$个单位时间间隙内新增的观看者数
k	转发系数
$E(X_n)$	在第n个单位时间已观看者的数目的数学期望

3 问题分析

此题属于信息传播类问题，与当今社会的发展联系得十分紧密。一个社交网络有着庞大的用户群体，对于这个复杂网络结构的研究分析是解决此题的关键。因此必须首先对题目附件中给出的连接关系数据进行处理分析和挖掘，进而建立起适当的模型进行求解。

3.1 问题一的分析

此问属于预测类问题，在一个庞大的网络系统中对信息的传播是一个典型的复杂系统的演化过程，需要将个体的相互作用和网络的结果进行综合考虑才能更准确地描述信息在网络中的传播并作出合理的预测。由于对于传播的预测方法有很多，根据其动态变化的特性可以建立起微分关系通过微分方程进行建模预测。从另外一个角度考虑社交网络系统中的用户虽然是有链接关系但可看作离散的点，时序按照天数计算也是非连续的，因此也可从离散数学的角度去建模求解。同时基于信息在社交网络上的传播规则，并结合社交网络拓扑结构对信息传播的影响建立仿真模型进行模拟，也是一种预测的方法。

3.2 问题二的分析

（1）求至少需要雇佣几名专业社交网络推广者。

在现实社交网络中，有可能出现几个传播者将一条相同的信息分享给一个粉丝的情况，故定义接受率为广告宣传覆盖的用户数与广告分享的次数之比。欲使广告宣传能覆

盖 2 亿潜在用户的 40%，须分享的次数与接收率的乘积大于需要覆盖的用户数。欲求最少需要几名专业推广者，因现在距离奥运会开幕还有 100 天，可在每名专业推广者工作 100 天的情况下计算满足需要的专业推广者的人数。解不等式，可得所需人数。

在求得所需专业推广者的人数之后，为使三名专业推广者的工作天数之和最小并确定具体的用人方案，故以各个专业推广者工作的天数为设计变量，以广告宣传能覆盖 2 亿潜在用户的 40%为约束条件，以三名专业推广者的工作天数之和最小为目标函数，建立一个整数规划性质的优化模型。

（2）求在考虑宣传成本情况下的合理用人方案。

此问与前一问的不同之处在于还可以从网络上雇佣兼职宣传者，并且其日工资低于专业推广者，但其每天新增的粉丝数少于专业推广者。考虑到成本，以雇员（专业推广者和兼职宣传者)的人数和工作天数为设计变量,以广告宣传能覆盖 2 亿潜在用户的 40%为约束条件（即分享的次数与接收率的乘积大于 2 亿潜在用户的 40%），以总成本最少为目标函数建立一个整数规划的优化模型。

4 数据处理与分析

4.1 数据的预处理

参考数据给出了 Twitter 社交网站用户之间的链接关系（follow 关系），一共有 835541 行，其中 835424 行数据为两列，第一列表示传播者，第二列表示其粉丝；而一共有 117 行数据出现异常情况：只有一列，其中第 47390 行数据为 3ActspfMurder，根据前后行数据对比可知，该数据应为：3Actspf Murder，表示 Murder 为 3Actspf 的粉丝，而对于剩余的 116 行，对比前后数据，均找不出其规律，故将其删除。

表1.参考数据中出现异常情况（只有一列数据）的行数序号

47390	117841	117851	117861	117871	117881	117891	117901	117911	117921	117931	574475
117832	117842	117852	117862	117872	117882	117892	117902	117912	117922	117932	626149
117833	117843	117853	117863	117873	117883	117893	117903	117913	117923	117933	663625
117834	117844	117854	117864	117874	117884	117894	117904	117914	117924	117934	759891
117835	117845	117855	117865	117875	117885	117895	117905	117915	117925	117935	760094
117836	117846	117856	117866	117876	117886	117896	117906	117916	117926	144258	760102
117837	117847	117857	117867	117877	117887	117897	117907	117917	117927	267534	785956
117838	117848	117858	117868	117878	117888	117898	117908	117918	117928	343072	
117839	117849	117859	117869	117879	117889	117899	117909	117919	117929	353288	
117840	117850	117860	117870	117880	117890	117900	117910	117920	117930	498548	

4.2 人数的统计及分析

将附件中给出的 Twitter 社交网站用户之间的链接关系数据导入 Matlab 软件进行编程统计得出各列人数及总人数结果如下表：

表 2

第一列人数	第二列人数	总人数
2589	465017	465023

从表中结果得出可以作为传播者的人数有 2589 人，仅占总人数的 0.55674％。可作

为粉丝的有 465017 人与总人数 465023 仅相差 6 人，说明在所有人中只有 6 个人能充当传播者而不充当粉丝。

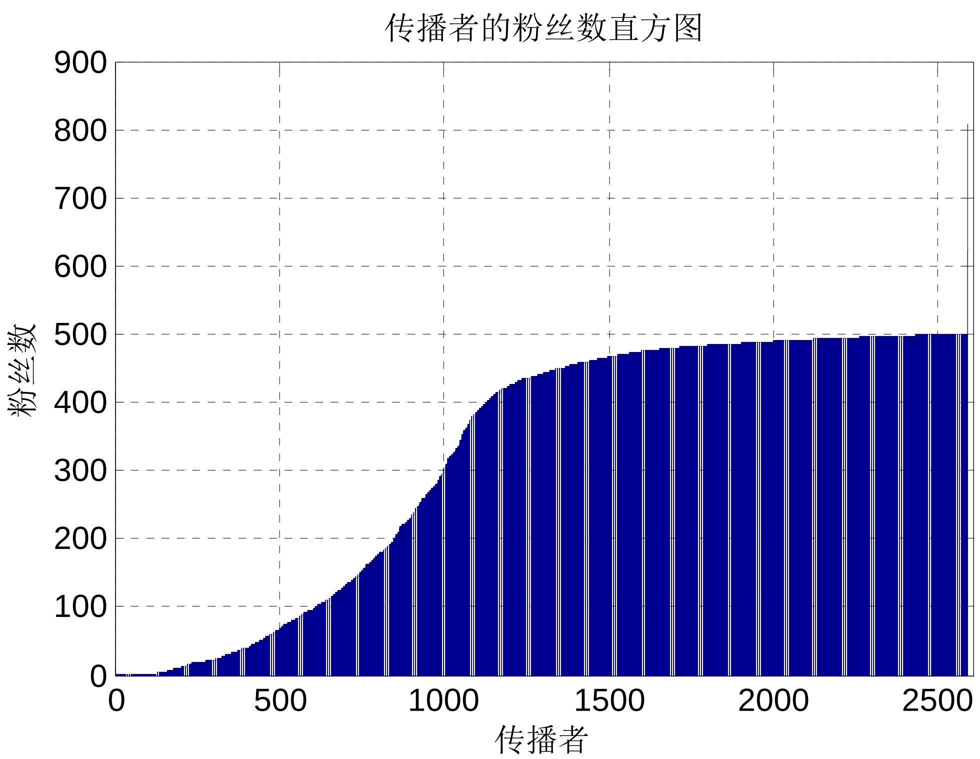
4.3.1 传播者对应的粉丝数的统计及分析

用 Matlab 软件将 2589 个传播者对应的粉丝数进行统计，得到的具体对应粉丝人数和所占总总粉丝数的百分比结果如下表：

表 3

传播者	粉丝数	百分比	传播者	粉丝数	百分比
agnieszkasshoes	491	0.058764322	corcarrasco	250	0.029920734
agreer	107	0.012806074	davidcohen	289	0.034588368
AKGovSarahPalin	41	0.004907	motionmigs	175	0.020944514
algore	10	0.001196829		⋮	
alyankovic	115	0.013763538			
amandaeyreward	492	0.058884004	CaptainLockheed	7	8.38E-04
Annewhitfield	366	0.043803955	HellBlog	484	0.057926541
BarackObama	480	0.057447809	companybahadur	498	0.059602102
barrylyga	81	0.009694318	OpusOneWinery	489	0.058524956
BenjaminFolds	9	0.001077146	ryzgo	128	0.015319416

从上表结果中得出不同的传播者对应的粉丝数不同，甚至有的差别巨大。为了更加直观地反映出传播者与粉丝数之间的关系，将上表中的 2589 个传播者按照粉丝数的多少进行排序作出传播者的粉丝数直方图如下：



图一

该直方图中只有第 2588 和 2589 号的粉丝数超过了 500 人，分别为 624 人和 807 人，其余传播者的粉丝数均小于或等于 500 人。对该直方图进行各项数据特性分析得到结果如下表：

表 4

传播者粉丝数的数据特性分析							
最大值	最小值	均值	中位数	标准差	方差	偏度	峰度
807	1	322.7273	441	192.9903	37245.2378	-0.5888	1.6240

从表中结果可以看出，最大值比最小值多出 806，差值巨大。平均人数为 322.7273 但标准差和方差分别为 192.9903 和 37245.2378，数值很大说明数据极不稳定，粉丝数波动很大。偏度和峰度值用于接下来的 *Jarque-Bera* 法正态性检验。

4.3.2Jarque-Bera 法正态性检验

对得出的传播者对应的粉丝数进行正态性检验，这里的检验对象为传播者对应的粉丝数。$x_1, x_2, \cdots, x_n$ 是来自 2589 个传播者对应的粉丝数样本，$\bar{x}$ 和 SD 分别为样本平均数和标准差，则样本偏度 *skew* 和峰度 *kurt* 定义为

$$skew = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{SD^3}$$

$$kurt = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{SD^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Jarque-Bera 统计量的计算公式为

$$JB = n[skew^2 + \frac{kurt^2}{4}]/6$$

如果变量服从正态分布则 *JB* 的值为零，如果显著性水平 $p<0.05$，可拒绝正态分布的假设。

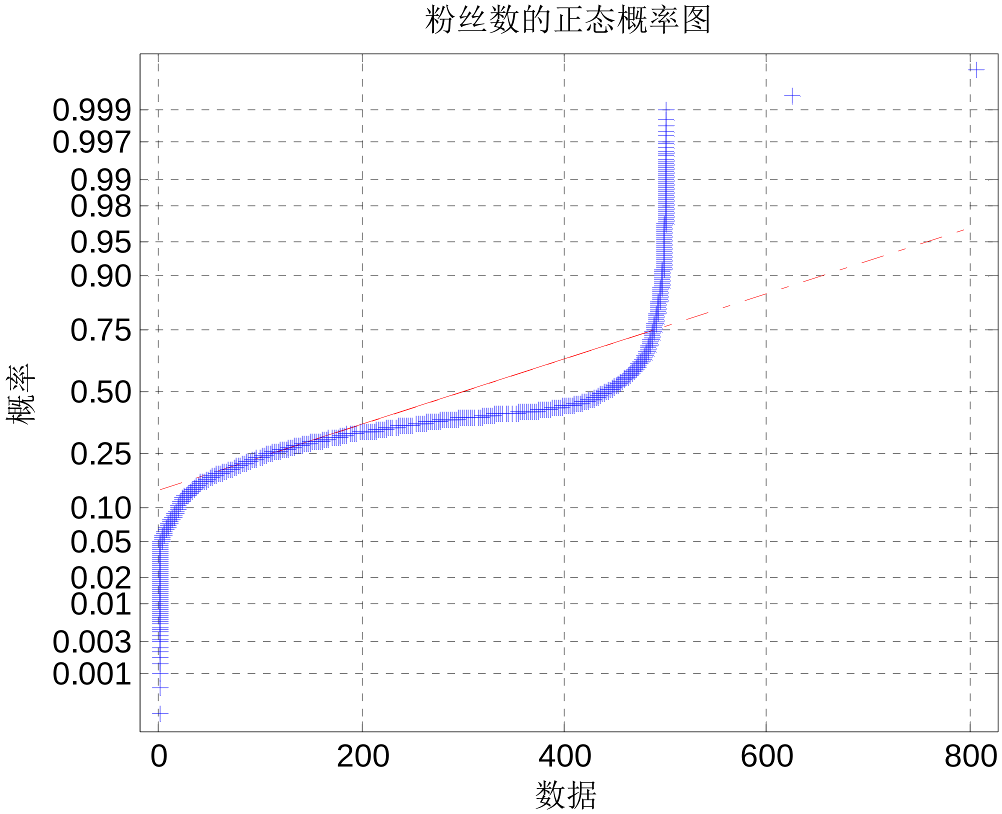
将传播者对应的粉丝数进行上述 *Jarque-Bera* 正态性检验

表 5

Jarque-Bera 正态性检验结果(显著水平 0.05)					
平均值	标准差	偏度 *skew*	峰度 *kurt*	*JB* 值	著性水平 *p*
322.7273	192.9903	-0.5888	1.6240	434.1017	0.00

由表中结果知 $JB \neq 0, p < 0.05$，因此传播者对应的粉丝数不符合正态分布假设。

用 Matlab 做出粉丝数的正态概率图如下：



图二

由上图中的图像更加直观地反映出播者对应的粉丝数不符合正态分布假设。

5 问题一的解答

5.1.1 SIR 微分方程模型的建立

某社交网络上的总用户数为 N，在短期内保持不变，则

(1) 用户分为未观看者，已观看者和转发者（退出传播），这三类用户在总用户数 N 中得比例分别为 $S(t)$、$I(t)$和 $R(t)$，三者关系为

$$S(t)+I(t)+R(t)=1$$

(2) 单位时间内一个已看广告的人能传播的人数与当时未看广告者人数比成正比，比例系数为 λ（接触率）。单位时间内已看广告但未传播的人数与当时的已看广告人数比成正比，比例系数 μ（不转发率）。

$$N\frac{dI(t)}{dt}=\lambda NS(t)I(t)-\mu NI(t)$$

(3) 对于退出传播者而言有

$$N\frac{dR(t)}{dt}=\mu NI(t)$$

(4) 对于未观看广告者人数而言有

$$N\frac{dS(t)}{dt}=-\lambda NS(t)I(t)$$

(5)初始时刻未观看者和已观看者的比例分别为S_0和I_0

$$S(0)=S_0$$
$$I(0)=I_0$$

综上建立的 SIR 微分方程模型为

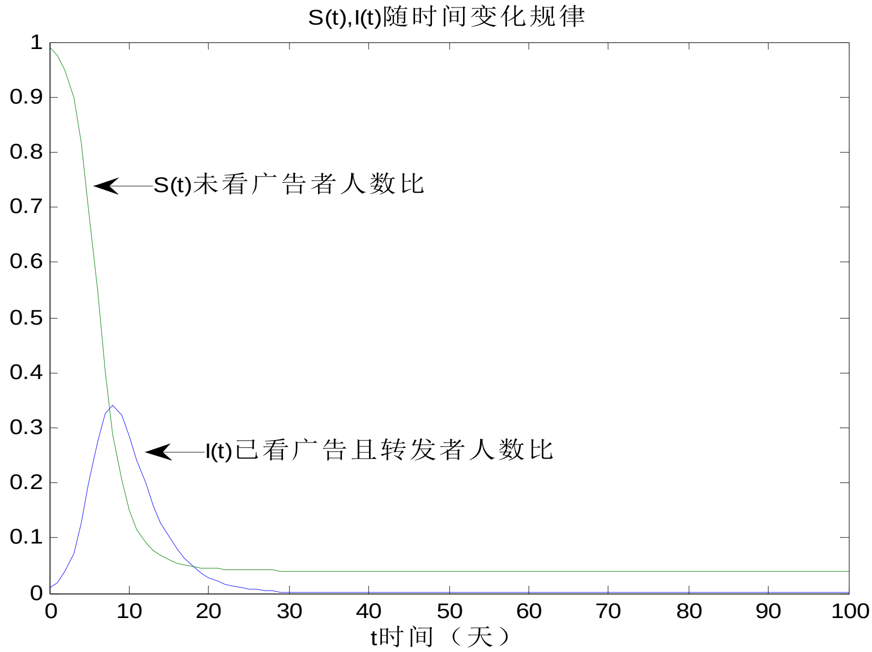
$$\begin{cases}S(t)+I(t)+R(t)=1\\ N\dfrac{dI(t)}{dt}=\lambda NS(t)I(t)-\mu NI(t)\\ N\dfrac{dR(t)}{dt}=\mu NI(t)\\ N\dfrac{dS(t)}{dt}=-\lambda NS(t)I(t)\\ S(0)=S_0,I(0)=I_0\end{cases}$$

5.1.2SIR 微分方程模型的求解

取$\lambda=1,\mu=0.3,I(0)=0.01,S(0)=0.99$利用 Matlab 软件计算出部分值。

表 6

t	0	1	2	3	4	5	6	7	8	9	10
I(*t*)	0.0100	0.0198	0.0385	0.0720	0.1261	0.1995	0.2752	0.3273	0.3416	0.3231	0.2856
S(*t*)	0.9900	0.9759	0.9488	0.8991	0.8159	0.694	0.5468	0.4034	0.2878	0.206	0.1519
t	14	18	22	26	30	34	38	42	44	48	52
I(*t*)	0.1605	0.0618	0.0225	0.008	0.0029	0.001	0.0004	0.0001	0.0001	0	0
S(*t*)	0.078	0.0514	0.0441	0.0417	0.0409	0.0406	0.0405	0.0404	0.0404	0.0404	0.0404



图三

显然由图像与数据易知两函数的变化规律，当$t \to \infty$时$I(t) \to 0, S(t) \to 0.404$；在$t = 8$附近，$I(t)$取得最大值 0.3416。要弄清$I - S$的关系必须找到两者的关系方程。记$\varepsilon = \dfrac{\mu}{\lambda}$，对原方程组消去$dt$得：

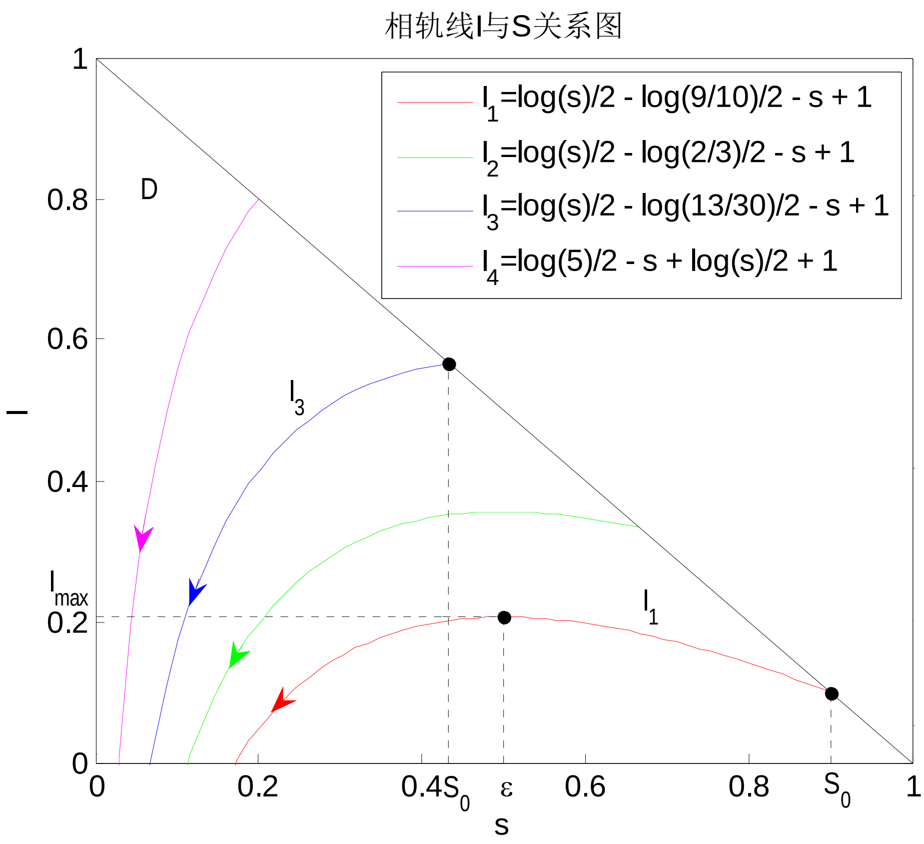
$$\begin{cases} \dfrac{dI}{dS} = \dfrac{\varepsilon}{S} - 1 \\ I\big|_{s=s_0} = I_0 \end{cases}$$

利用Matlab软件求解得到二维微分方程解的相轨线方程[1]，即为模型的解：

$$I(S) = (S_0 + I_0) - S + \varepsilon \ln \frac{S}{S_0}\text{。}$$

而相轨线$I(S)$的定义域$D = \{(S, I) \big| S \ge 0, I \ge 0, S + I \le 1\}$

取$\varepsilon = 0.5$，利用 Matlab 软件在D内作多个相轨线$I(S)$的图形，进行分析。



图四

相轨线图像明显的反映了阀值 ε 与 S_0 的关系影响 $I(t)$ 曲线变化规律。

当 $\varepsilon > S_0$ 时 $I(t)$ 单调减少直到 0，即广告传播量越来越少直到停止

当 $\varepsilon < S_0$ 时 $I(t)$ 先增，在 $S(t)=\varepsilon$ 时 $I(t)$ 取最大值 $I(t)_{\max}=S_0+I_0-\varepsilon(1+\ln\frac{S_0}{\varepsilon})$ 然后减少到 0，即广告传播量先增后减最后停止。

估计已经看到广告的人数：

最终未看到广告的人数比为 S_∞，记观看到广告的人数比例为 $x=S_0-S_\infty$。

当 $t\to\infty$ 时，有 $I_\infty=S_0+I_0-S_\infty+\varepsilon\ln\frac{S_\infty}{S_0}=0$。由 $I_0\approx 0, S_0\approx 1$ 得到 $x+\varepsilon\ln(1-\frac{x}{S_0})\approx 0$。

取对数函数 Taylor 展开的前两项有 $x(1-\frac{1}{s_0\sigma}-\frac{x}{2{s_0}^2\sigma})\approx 0$。记 $s_0=\varepsilon+\delta$，δ 可视为该社交网络用户受众比例超过阈值 ε 的部分。

当 $\delta\le\varepsilon$ 时 $x\approx 2s_0\sigma\left(s_0-\frac{1}{\sigma}\right)\approx 2\delta$，即估计已经看到广告人数比为 2δ。

以 Twitter 社交网络为例，由数据处理知，每天每人新增关注数 1.7 人，每人每天新

增粉丝数 1.0 人。

$$\lambda = \frac{1.0}{1.7} = 0.5882\text{，}\quad \mu = \frac{1.7-1.0}{1.7} = 0.4118\text{，}\quad \varepsilon = \frac{\mu}{\lambda} = 0.7001$$

设 $S_0 = 0.98$, $\delta = S_0 - \varepsilon = 0.2799$,则最终看到广告人数比例为 $2\delta = 0.5598$ 。则当社交网络中用户总数为 2 亿时，预测得到观看者人数为 1.1196 亿。

5.1.3 SIR 微分方程模型的结果分析

(1) 无论 S_0, I_0 如何，广告终将停止传播，即 $I_\infty = 0$ 。

首先知道 $\frac{dS}{dt} \le 0$ ，而 $0 \le S(t) \le 1$，由单调有界必有极限知 S_∞ 必定存在；又 $\frac{dR}{dt} \ge 0$ ，而 $0 \le R(t) \le 1$，同理知 R_∞ 必定存在；又因为 $S(t)+I(t)+R(t)=1$，故 I_∞ 必定存在。

其次从相轨线图上看，无论相轨线从哪里出发，它都终将与 S 相交。

(2) 被传播人数比例可估计且趋于定值。

这个结果表明，观看到广告的人数比例约为 δ 的 2 倍。对于一则广告，当该社交网络受众对待广告的喜恶水平不变，即 δ 不变时，这个比例就不会改变。而当阈值 ε 提高时，δ 减小，于是这个比例就会降低。

5.2.1 离散数学模型的建立

某社交网络上的总用户数为 N，在短期内保持不变，对这 N 个用户进行编号：$C_1, C_2, \cdots, C_N$ 并定义集合 $C = \{C_1, C_2, \cdots, C_N\}$ 。新闻进入社交网络开始 $(n=0)$ 时，此时 N 个用户中有一个发布了该条新闻，设其为 C_1 ；令 X_n 是第 n 个单位时间 C 中观看到新闻的用户数即观看者数，则 $\{X_n \ge 1\}$ 构成一个离散随机过程[2]。

(1) 在第 n 个单位时间，已观看者的数目为

$$E(X_n)$$

未观看者数目为

$$N - E(X_n)$$

(2) 在第 n 个和第 $n+1$ 个单位时间间隙内新增的 M 个观看者中，转发者的期望

$$\frac{M}{N}E(X_n)$$

不转发者的数学期望为

$$M-\frac{M}{N}E(X_n)$$

(3)新增加观看者的数学期望有如下关系

$$\frac{N}{M}E(X_n)(1-\frac{E(X_n)}{N})=kE(X_n)(N-E(X_n))$$

其中k为定义的转发系数

$$k=\frac{M}{N^2}$$

(4)结合(1)、(2)将(3)中关系式变形整理后得

$$\frac{E(X_{n+1})-E(X_n)}{(n+1)-n}=kE(X_n)(N-E(X_n))$$

(5)把$E(X_n)$视为一连续函数$f(x)$在$x=n$点的离散化数值，可列出$f(x)$的微分方程[3]

$$\frac{df(x)}{dx}=kf(x)(N-f(x))$$

(6)由初值$f(x)=0$将(5)中等式分离变量并离散化得

$$E(x)=\frac{N}{1+(\frac{N}{X_0}-1)e^{-nkN}}$$

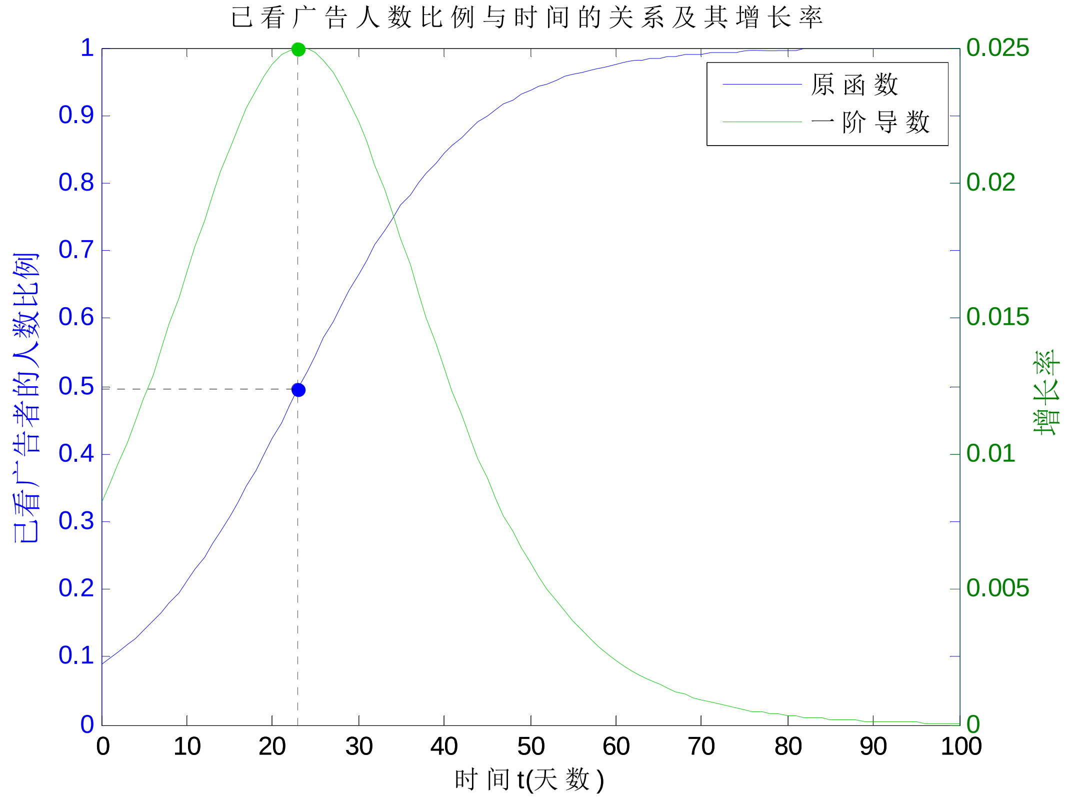
综上建立的离散数学模型为

$$E(x)=\frac{N}{1+(\frac{N}{X_0}-1)e^{-nkN}}$$

5.2.2 离散数学模型的求解

对于离散模型的表达式，为更好的研究其性质和规律，将离散连续化。$E(X_n)=f(t)\cdot N=\frac{1}{1+(\frac{1}{b}-1)e^{-t\lambda}}\cdot N$，$b=\frac{X_0}{N}$为初始传播者占终用户数的比例。$\lambda=kN=\frac{M}{N}$为单位时间观看率。欲研究$E(X_n)$的性质和规律即研究函数$f(t)$的性质和规律。$f(t)$的意义即为已看广告者的人数比例。

取 $b=0.09,\lambda=0.1$ 利用 Matlab 软件画出 $f(t)$ 的函数图象如下



图五

由图象易知当已看广告人数达到总用户数的一半时，其增长率达到最大值，即传播速率最快；当时间 $t\to\infty$ 时，即天数大于 100 天时，已看广告人数将趋于总用户数，但这是不符合实际的。则将 2 亿代入得观看者人数为 2 亿，不符合实际情况。

5.3.1 传播仿真 IDM(α,β)模型的建立

新闻在某社交网络上的传播过程可以简述如下：

(1) 用户 u 发布含有企业广告的奥运新闻 I

(2) u 的粉丝获知含有企业广告的奥运新闻 I

(3) 对新闻 I 感兴趣的粉丝 v 对该条新闻进行转发，不感兴趣的粉丝则不转发

(4) 如此重复 (2) 和 (3) 过程

将社交网络Twitter上的传播机制和网络拓扑结构相结合[4]。

定义消息敏感度系数

$$\alpha\in[0,1],$$

其值是由消息的来源，内容及其表现形式决定的。

定义转发新闻的方式系数

$$\beta\in\{-1,0,1\}$$

其值用来决定用户的转发方式。

a). 若 $\beta=1$，选择粉丝数多的用户对新闻进行转发；

b). 若 $\beta=-1$，选择粉丝数少的用户对新闻进行转发；

c). 若 $\beta=0$，随机选择用户对新闻进行转发。

社交网络中的用户只可能处于三种 S 型状态[5]：

a). 未激活，即不知道奥运新闻 I，记为 S_0；

b). 被奥运新闻 I 激活但选择不转发，记为 S_1；

c). 被奥运新闻 I 激活后对新闻进行转发，记为 S_2。

时间步序是一个离散的等间隔时间序列，则模型定义如下：

(1) 初始时刻，将网络中所有节点即用户的初始状态初始化为 S_0。在时间步 $T=0$，随机选择一个节点 v 作为发布新闻 I 的源节点。节点 v 的出度是 $dout(v)$。

(2) 在时间步 $T=T+1$，关注节点 v 的 $dout(v)$ 个节点被激活，其中转发新闻 I 的节点数量为 M $(M\le dout(v))$ 且 $M=dout(v)\bullet\alpha$。设 P^T 是时间步 T 内选择转发的节点的集合，即 $P^T=\left\{u_j\right\}$，其中 u_j 的状态为 S_2。

(3) 对于关注节点 v 且不属于集合 P^T-1 的 n $(n\le dout(v))$ 个节点的出度进行排序，按照从小到大的顺序记为 $Q=\left\{u(1),u(2),\cdots,u(dout(v))\right\}$。

a). 若 $\beta=1$，顺序选择序列 Q 中前 M 个节点，将它们状态设为 S_2，放入集合 P^{T+1}

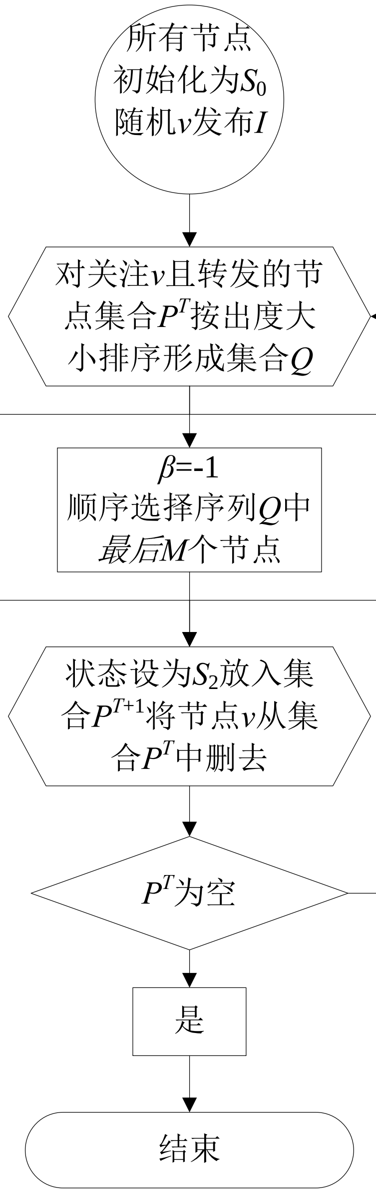
b). 若 $\beta=-1$，顺序选择序列 Q 中最后 M 个节点，将它们状态设为 S_2，放入集合 P^{T+1}

c). 若 $\beta=0$，从序列 Q 中随机选出 M 个节点，将它们状态设为 S_2，放入集合 P^{T+1}

将节点 v 从集合 P^T 中删去。

(4) 对每一个节点 $v\in P^T$ 递归执行 (2) 和 (3)，直到集合 P^T 为空。

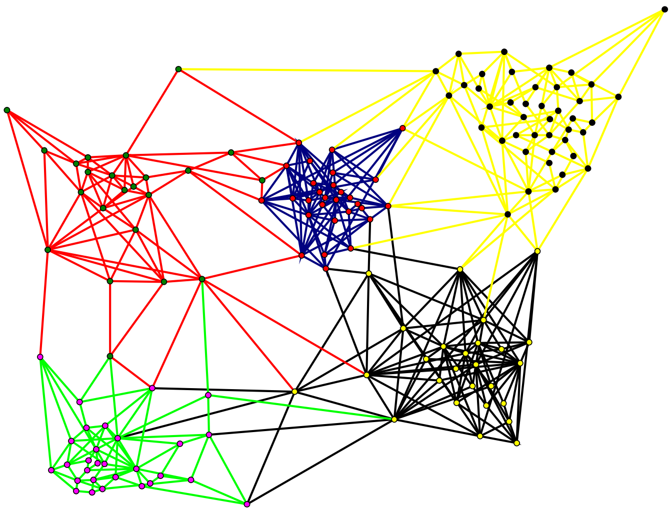
综上建立的传播仿真 IDM(α, β)模型为



图六

5.3.2 传播仿真 IDM(α, β)模型的求解

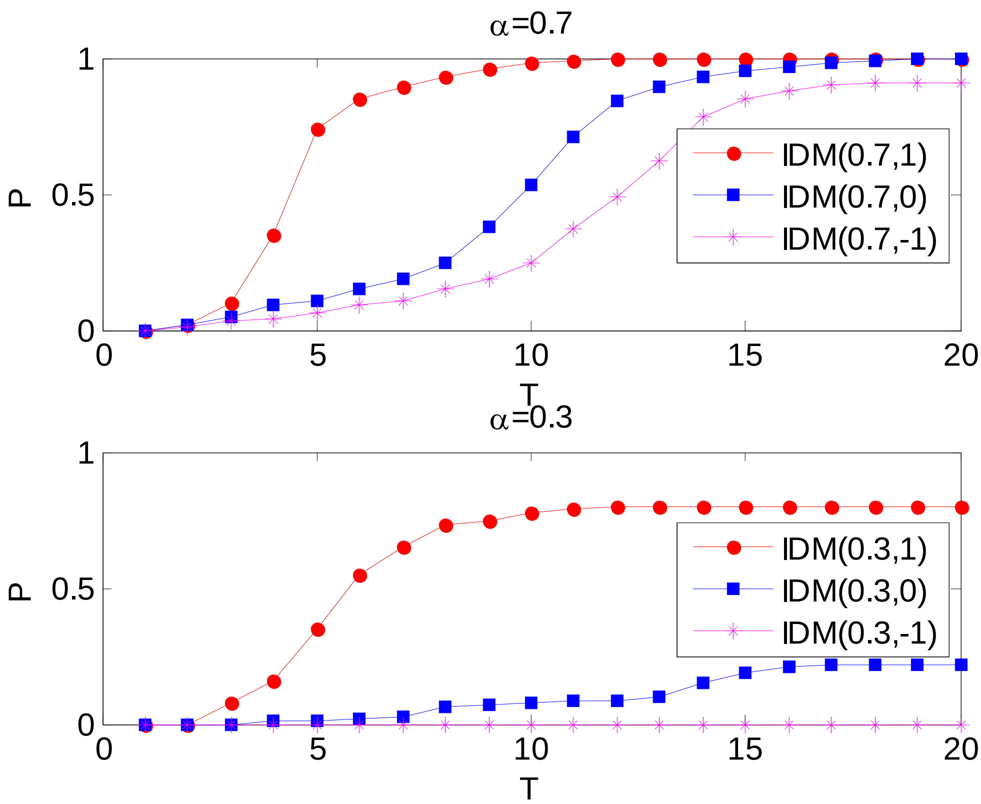
将附件中给出的链接数据作为已知网络如下图，该网络是个有向图，记为 $G=(V,E)$，其中节点 $v \in V$ 表示一个用户，有向边 $(u,v) \in E$ 表示用户 v 是用户 u 的粉丝，信息从 u 流向 v。



图七

该网络的节点数为 465023，平均入度 $\overline{din} = 1.6724$，平均出度 $\overline{dout} = 322.7273$。

将此网络作为已知数据库[6]（Twitter社交网站用户数据）按照仿真流程图对参数 α, β 进行了 100 次仿真计算，将各个参数的平均结果得到绘制成曲线。



图八

从图中可以明显看出，当信息敏感度（广告吸引力）较小时如$\alpha=0.3$时，明星类用户对广告的传播起推动作用；但当信息敏感度较大时如$\alpha=0.7$时，明星类用户对广告的传播速率的提高和传播人数的扩大的作用越来越小。

对于 2 亿用户，因为实际情况广告信息敏感度不高，假设消息敏感度为 0.4。IDM(α,β)模型最终看到广告人数比例结果见下表

表 7

IDM(0.4,1)	IDM(0.4,0)	IDM(0.4,-1)
0.8237	0.2851	0.0193

平均最终看到广告人数比例为 0.3760
即最终看到广告人数为 0.7521 亿人。

综上三种模型的求解结果

对于假定的 2 亿用户三种模型的预测结果如下：

表 8

模型	SIR 微分方程模型	离散数学模型	传播仿真 IDM(α,β)模型
预测人数	0.7521 亿	2 亿	1.1196 亿

6 问题二的解答

6.1 人数最优化模型的建立

6.1.1 确定目标函数

雇佣的专业社交网络推广者数 p 最少

$$\min f=p$$

6.1.2 定义接受率 η

X：附件数据中 Twitter 社交网站的传播者人群（第一列）的集合
Y：附件数据中 Twitter 社交网站的粉丝人群（第二列）的集合
Z：附件数据中 Twitter 社交网站的总人群集合，则依题意知：

$$Z=X\bigcup Y$$

n_{share}：附件数据中 Twitter 社交网络中传播者（第一列）分享的次数之和，即附件中数据的行数，即$n_{share}=835541$。

用 Matlab 编程统计得到各人群的人数列于下表：

表 9

人群	X	Y	Z
数量 n	2589	465017	465023

由表中数据可知$n_X+n_Y\neq n_Z$， $n_Z-n_Y=6$，即在Twitter社交网站中只有 6 个人只做

传播者而不做其他人的粉丝，其余的传播者则同时兼为粉丝。因 $n_Z < n_{share}$ ，即Twitter社交网站中的总人数少于该信息被分享的次数，即有的人重复接收到同一条信息。

定义接收率

$$\eta = \frac{n_z}{n_{share}}$$

其表示接受到信息的总人数与该信息被分享的次数。由表中数据可算得 $\eta = 0.5566$ 。

6.1.3 确定约束条件

(1)所有专业推广者须使得发出的广告覆盖 *2* 亿潜在用户的 *40%*，则

$$\left[\Delta x \bullet \Delta z \sum_{i=1}^{p} \frac{(1+n_i)n_i}{2} + \Delta x \sum_{i=1}^{p} n_i\right] \times \eta \ge 2 \times 10^8 \times 40\%$$

其中 Δx 为专业推广者每天新增的粉丝数，Δz 为普通网络用户每天新增的粉丝数， p 为专业推广者的人数，n_i 为第 i 个专业推广者工作的天数， p 为专业推广者的人数，η 为接收率。

(2)由于距离奥运会开幕只有 100 天，故所有专业推广者的工作天数均应小于 100 天

$$n_i \le 100\,;\, n_i \in N^* \quad i = 1, 2, \cdots\cdots, p$$

综上建立的人数最优化模型为

$$\min\ f = p$$

$$s.t.\begin{cases} \left[\Delta x \bullet \Delta z \sum_{i=1}^{p} \frac{(1+n_i)n_i}{2} + \Delta x \sum_{i=1}^{p} n_i\right] \times \eta \ge 2 \times 10^8 \times 40\% \\ n_i \le 100\,;\, n_i \in N^* \quad i = 1, 2, \cdots\cdots, p \end{cases}$$

6.1.4 人数最优化模型求解

用 Lingo 软件编程求解结果为至少需要雇佣 3 名专业推广者，且 3 名专业推广者的工作业绩如下表：

表 10

雇员	工作天数	分享的次数	覆盖的人数
第一个专业推广者	100	50550000	80102473
第二个专业推广者	100	50550000	
第三个专业推广者	92	42826000	
总和	292	143926000	

综上雇佣专业社交网络推广者的人数为 $p = 3$

6.1.5 人数最优化模型结果分析

题中要求广告宣传覆盖的潜在用户的人数为 $2 \times 10^8 \times 40\% = 80000000$ ，由上表知：3 个专业推广者能使广告宣传覆盖的潜在用户数为 3 个专业推广者分享次数之和与接收率之积，即为：$143926000 \times \eta = 80102473$ ，显然 $80102473 > 80000000$ ，即该方案满足要

求。至少雇佣的 3 名专业推广者几乎是 100 天满负荷工作，增大了雇佣效率。

6.2 用人方案最优化模型的建立

6.2.1 确定目标函数

企业付给专业推广者和兼职宣传者的工资之和最小

$$\min\ f = C_x \sum_{i=1}^{p} n_i + C_y \sum_{j=1}^{q} n_j$$

其中 C_x 为每天付给专业推广者的工资； C_y 为每天付给兼职宣传者的工资； p,q 分别为专业推广者和兼职宣传者的人数； n_i,n_j 分别为专业推广者和兼职宣传者的工作天数。

6.2.2 确定约束条件

(1)所有专业推广者和所有兼职宣传者须使得发出的广告覆盖 *2* 亿潜在用户的 *40%*

$$(a+b)\eta \geq 2\times 10^8 \times 40\%$$

其中 a 为所有专业推广者分享信息的次数：

$$a = \Delta x \Delta z \sum_{i=1}^{p} \frac{(1+n_i)n_i}{2} + \Delta x \sum_{i=1}^{p} n_i$$

b 为所有兼职宣传者分享信息的次数：

$$b = \Delta y \Delta z \sum_{j=1}^{q} \frac{(1+n_j)n_j}{2} + \Delta y \sum_{j=1}^{q} n_j$$

(2)由于距离奥运会开幕只有 100 天，故所有专业推广者和兼职宣传者的工作天数均应小于 100 天，即：

$$n_i \leq 100\,;\, n_i \in N^* \quad i = 1,2,\cdots\cdots,p$$

$$n_j \leq 100\,;\, n_j \in N^* \quad j = 1,2,\cdots\cdots,q$$

综上建立的用人方案最优化模型为

$$\min\ f = C_x \sum_{i=1}^{p} n_i + C_y \sum_{j=1}^{q} n_j$$

$$s.t.\begin{cases} a = \Delta x \Delta z \sum_{i=1}^{p} \frac{(1+n_i)n_i}{2} + \Delta x \sum_{i=1}^{p} n_i \\ b = \Delta y \Delta z \sum_{j=1}^{q} \frac{(1+n_j)n_j}{2} + \Delta y \sum_{j=1}^{q} n_j \\ (a+b)\eta \geq 2\times 10^8 \times 40\% \\ n_i,n_j \leq 100\text{；}\ n_i,n_j \in N^* \quad i=1,2,\cdots\cdots,p \quad j=1,2,\cdots\cdots,q \end{cases}$$

6.2.2 用人方案最优化模型求解

用 Lingo 软件编程求解结果为需要 3 个专业推广者，不需要兼职宣传者，具体的用人方案见下表：

表 11

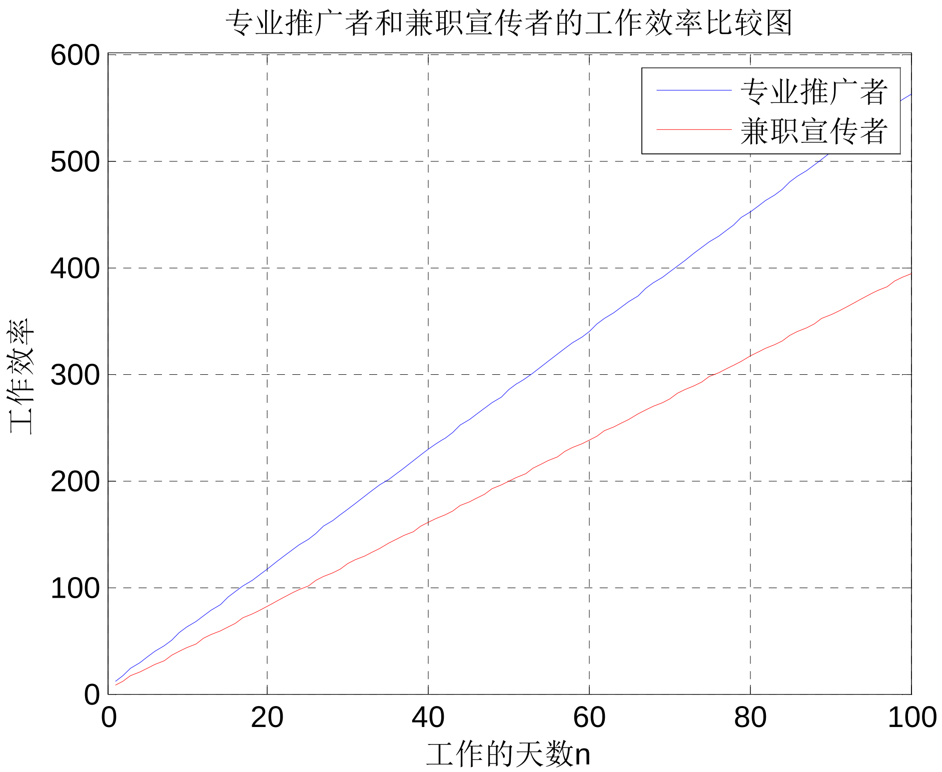
雇员	工作天数	日工资（元）	总工资（元）	分享次数	覆盖人数
第一个专业推广者	100	500	50000	50550000	80102473
第二个专业推广者	100	500	50000	50550000	
第三个专业推广者	92	500	46000	42826000	
总和	292		146000	143926000	

专业推广者人数 $p=3$ 人，其工作天数分别 $n_1=100$ 天， $n_2=100$ 天， $n_3=92$ 天，宣传兼职者人数 $q=0$ 人， 宣传成本 $f=146000$ 元。

6.2.3 用人方案最优化模型结果分析

题中要求广告宣传覆盖的潜在用户的人数为 $2\times10^8\times40\%=80000000$，由上表知：三个专业推广者能使广告宣传覆盖的潜在用户数为三个专业推广者分享次数之和与接收率之积，即为：$143926000\times\eta=80102473$，显然 $80102473>80000000$，即该方案满足要求。宣传成本为所有雇员的工资之和，由上表得宣传成本为：146000 元。

为分析该用人方案中只雇佣专业推广者而不雇佣兼职宣传者的原因，故定义工作效率为一个雇员工作 n 天一共能使广告宣传覆盖的人数与其工作 n 天所需的工资之比。



图九

图中代表专业推广者工作效率的曲线在工作天数[0 100]范围内均高于代表兼职宣传者工作效率的曲线，即不论工作几天，专业推广者的工作效率均高于兼职宣传者的工作效率，这也印证了用 Lingo 软件编程求得的结果：企业在进行广告宣传时，为使宣传

成本最少，只雇佣专业推广者而不雇佣兼职宣传者。

7 模型的评价，改进及推广

7.1 模型的评价

模型优点：

(1) 第一问的三个模型的原理不同，可用于对人数进行综合性预测，减少了片面性。
(2) 传播仿真 IDM(α,β)模型将个体的相互作用和网络的结构进行了综合考虑。
(3) 第二问的两个最优化模型简洁明了，可操作性强。

模型缺点：

(1) SIR 微分方程模型将信息传播设定为连续性变化，没有考虑其时序的离散性。
(2) 离散数学模型具有一定的随机性，会影响到预测精度，且结果不符合实际。
(3) 问题二的模型假设过于理想，致使该模型在解决实际问题时会影响到方案设计的最优性。

7.2 模型的改进

对于传播仿真 IDM(α,β)模型应该进行更多次数的仿真，在仿真过程中对于敏感度系数的取值跨度应该明显。这样可以使得结果更加有说服力。

对于第二问中的人数最优化模型和用人方案最优化模型考虑每个雇员的初始粉丝数，可以使得模型所处理的情况更贴近现实问题。

7.3 模型的推广

SIR 微分方程模型还可以用于流行病的传播、计算机网络病毒传播的预测。人数最优化模型和用人方案最优化模型可以用于解决人力资源的分配及调度等实际生产问题。

参考文献

[1]查淑玲、陈冲.　SIR 模型的研究. 传播理论研究者，2003(04):12-15
[2]韩中庚.　数学建模方法及其应用. 北京：高等教育出版社，2005
[3]化存才.　微分方程学习、设计与建模应用导引. 成都：西南交通大学出版社 2011.3
[4]郑蕾、李生红.　基于微博网络的信息传播模型. 通信技术，2012(02):39-41
[5]袁立庠.　微博的传播模式与传播效果. 湖北社会科学，2011(03):22-23
[6]陈浩、陈颖.　复杂网络模型的研究与应用. 科技时代，2009(04):17-20

附录

第一问

```
cma=2;y0=linspace(0.15,0.8,4);s0=linspace(0.85,0.35,4);
ys=['r','g','b','m','k'];flag=cell(1,length(y0));
for i=1:length(y0);
    t=subs('y(s0)=y0',{'s0','y0'},{s0(i),y0(i)});
    t=char(t);
    F=dsolve('Dy=1/cma/s-1',t,'s');
    f=subs(F);
    h=ezplot(f);
    set(h,'color',ys(i))
    flag(i)={char(f)};
    hold on
end
fill([0 1 1],[1 1 0],'w')
legend(['f_1=',flag{1}],['f_2=',flag{2}],['f_3=',flag{3}],['f_4=',flag{
4}])
%画第一条曲线的顶点及顶点的垂线与水平线
qdd=sym(flag{1});
hs=inline(-qdd);
x=fminbnd(hs,0,1);
y=-hs(x);
i=linspace(0,x,150);
j=linspace(0,y,50);
plot(i,y*ones(1,length(i)),'k:','linewidth',1)
plot(x*ones(1,length(j)),j,'k:','linewidth',1)
plot(x,y,'ko','markerfacecolor','k','markersize',5)
%画第三条曲线端点的垂线
jg3=solve(['-s+1=',flag{3}]);
x3=double(vpa(jg3));
y3=-x3+1;
j3=linspace(0,y3,100);
plot(x3*ones(1,length(j3)),j3,'k:')
plot(x3,y3,'ko','markerfacecolor','k','markersize',5)
%画第一条曲线端点的垂线
jg1=solve(['-s+1=',flag{1}]);
x1=double(vpa(jg1));
y1=-x1+1;
j1=linspace(0,y1,100);
plot(x1*ones(1,length(j1)),j1,'k:')
plot(x1,y1,'ko','markerfacecolor','k','markersize',5)
axis([0 1 0 1])
xlabel('s');ylabel('f');title('图形')
```

双坐标轴图形

```
clear,clc,clf
a=0.1;b=0.09;
h=dsolve('Dy=a*y*(1-y)','y(0)=b','t');
f=subs(h);f0=vectorize(f);
t0=0:1:100;y0=1./(exp(log(91./9) - t0./10) + 1);
hold on
f1=diff(f);f1=vectorize(f1);
t1=0:1:100;
y1=exp(log(91./9) - t1./10)./(10.*(exp(log(91./9) - t1./10) + 1).^2);
[ax h1 h2]=plotyy(t0,y0,t1,y1);
set(h1,'color','b');set(h2,'color',[0 0.8 0])
legend('原函数','一阶导数')
ydd1=max(y1);xdd1=t1(y1==ydd1);
plot(xdd1,ydd1*40,'ro','markerfacecolor',[0 0.8 0],...
    'markeredgecolor',[0 0.8 0],'markersize',5)
jdd1=linspace(0,ydd1*40,100);
plot(xdd1*ones(1,length(jdd1)),jdd1,'k:')
xdd0=t0(y1==ydd1);ydd0=y0(y1==ydd1);
plot(xdd0,ydd0,'bo','markerfacecolor','b','markersize',5)
idd0=linspace(0,xdd0,100);
plot(idd0,ydd0*ones(1,length(idd0)),'k:')
xlabel('时间t(天数)')
set(get(ax(1),'ylabel'),'string','已看广告者的人数比例')
set(get(ax(2),'ylabel'),'string','增长率')
title('已看广告人数比例与时间的关系及其增长率')
```

数据处理及计算接收率

```
[m n]=strtok(socialgraph);%m为传播者（第一列）的集合
n=strtrim(n);%n为粉丝（第二列）的集合
abc1=tabulate(m);%统计各个传播者和其粉丝数及其百分比
abc2=tabulate(n);%统计各个粉丝和其传播者数及其百分比
jh=union(m,n);%该附件数据中包含的人群集合
save all abc1 abc2 jh m n%保存数据到all.mat
save abc1 abc1;save abc2 abc2;save jh jh;save m m;save n n%分别保存数据
load all%导入all.mat的数据
load abc1;load abc2;load jh;load m;load n%分别导入数据
n_cbz=length(abc1)%第一列传播者的数量
n_fs=length(abc2)%第二列粉丝的数量
jsl=length(jh)/length(n)%接受率
n_cbz2fs=cell2mat(abc1(:,2));%将元胞数组（各个传播者的粉丝数）转化为矩阵
n_cbz2fsmax=max(n_cbz2fs)%最大值
n_cbz2fsmin=min(n_cbz2fs)%最小值
```

```
n_cbz2fsmean=mean(n_cbz2fs)%均值
n_cbz2fsmedian=median(n_cbz2fs)%中位数
n_cbz2fsstd=std(n_cbz2fs)%标准差
n_cbz2fsvar=var(n_cbz2fs)%方差
n_cbz2fsskewness=skewness(n_cbz2fs)%偏度
n_cbz2fskurtosis=kurtosis(n_cbz2fs)%峰度
jb_cbz2fs=length(abc1)*(n_cbz2fsskewness^2+n_cbz2fskurtosis^2/4)/6
bar(sort(n_cbz2fs))
axis([0,length(n_cbz2fs)+15,0,900])
xlabel('传播者');ylabel('粉丝数');title('传播者的粉丝数直方图');grid on
normplot(n_cbz2fs)%分布的正态性检验
[muhat,sigmahat,muci,sigmaci]=normfit(n_cbz2fs)%参数估计
[h,sig,ci]=ttest(n_cbz2fs,3.227273078408652e+002)%假设检验
```

第二问第一小问

求至少需要多少个专业推广者

```
eta=0.5565531793173524865991907972784247249364852905273437 5;!eta为接收率;
 (500*20*(100+1)*100/2+500*100)*p*eta>=2*10^8*0.4;!专业推广者应使广告宣传覆盖2亿潜在用户中40%的人群;
@gin(p);!限制p为整数;
min=p;!目标为使专业推广者的人数最少;
```

计算结果为所需专业推广者的人数 p=3;

求三个专业推广者的工作方案，目标函数为使三个专业推广者工作的天数之和最小。

```
sets:
zytgz/1..3/:Nz;!5为专业推广者（zytgz）的人数上限;
endsets
a=500*20*@sum(zytgz:Nz^2+Nz)/2+500*@sum(zytgz:Nz);!a为所有专业推广者分享的次数;
eta=0.5565531793173524865991907972784247249364852905273437 5;!eta为接收率;
a*eta>=2*10^8*0.4;
@for(zytgz:@gin(Nz));!限制专业推广者的人数p为整数;
@for(zytgz:Nz<=100);!限制每个专业推广者工作的天数小于100;
min=@sum(zytgz:Nz);!目标为使三个专业推广者工作的天数之和最小;
```

第二问第二小问

```
sets:
!优化结果中Nz(i)为第i个专业推广者工作的天数,不为0的Nz(i)的个数即为实际雇佣的专业推广者的人数，不为0的Nz(i)对应的Value即为该专业推广者工作的天数
!优化结果中Nj(j)为第j个专业推广者工作的天数,不为0的Nj(j)的个数即为实际雇佣的兼职宣传者的人数，不为0的Nj(j)对应的Value即为该兼职宣传者工作的天数
!改变程序中专业推广者(zytgz)和兼职宣传者(jzxcz)的人数上限，使其人数不会构成紧约束（在优化结果中表现为至少有一个Nz(i)和一个Nj(j)为零），即有足够
 的专业推广者个兼职宣传者可供选择来完成任务，就可优化使总花费最小，即目标值最小;
```

```
zytgz/1..5/:Nz;!5为专业推广者（zytgz）的人数上限;
jzxcz/1..25/:Nj;!25为兼职宣传者（jzxcz)的人数上限;
endsets
a=500*20*@sum(zytgz:Nz^2+Nz)/2+500*@sum(zytgz:Nz);!a为所有专业推广者分享的次数;
b=35*20*@sum(jzxcz:Nj^2+Nj)/2+35*@sum(jzxcz:Nj);!b为所有兼职推广者分享的次数;
eta=0.55655317931735248659919079727842472493648529052734375;!eta为接收率;
(a+b)*eta>=2*10^8*0.4;
@for(zytgz:@gin(Nz));!限制每个专业推广者工作的天数为整数;
@for(zytgz:Nz<=100);!限制每个专业推广者工作的天数小于100;
@for(jzxcz:@gin(Nj));!限制每个兼职宣传者工作的天数为整数;
@for(jzxcz:Nj<=100);!限制每个兼职宣传者工作的天数小于100;
min=500*@sum(zytgz:Nz)+50*@sum(jzxcz:Nj);!目标函数为使总花费最小;
```