For office use only	Team Control Number	For office use only
T1 _____	**0000**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**D**	F4 _____

2018
MCM/ICM
Summary Sheet

How does the Information Flow

Summary

In the era of web 2.0, tons of information flow through social networks, bringing opportunities and challenges at the same time. Based on an improved-SDIR model, we explore the society's network and predict the information flow according to its structure.

Our work initially focuses on data analyses. Based on five periods, we investigate the evoulution of society's networks. Mass media evolved in types and they transfromed from advertiser-generated modes to consumer-generated modes. Then we propose an algorithm to evaluate the inherent value of information based on Analytical Hierarchy Process(AHP) and validate it by evaluating the values of three pieces of information. Next we define the qualifications of news from two perspectives: inherent value and propagation speed. And we sucessfully calculate the values of a hot issue.

As for the information diffusion, we introduce an improved-SDIR model after a brief introduction of epidemic dynamics model. We simulate the structure of of a blog network and verify the degree distribution fitted by the power-law distribution. Additionally, we visualize the propagation time of the information, and we find the life span of information is mainly dependent on the total number of nodes and the nodes in the state of D. Besides, we do the sensitivity analyses,which shows the scale-free network, like CatalogBlog, is more likely to generate explosive public issues.

Next, we introduce Ammerican annual population to predict the information diffusion. We simulate the corresponding social network in 5 periods. Since these periods vary from population and the evolution of mass media,we totally control three parameters: the number of nodes,the number of edges and the infection probability function. We predict the social information network nowadays and compare it with the current situation to validate our model. Meanwhile we forecast the social information network in 2050. The gap of the influence power will be reduced between mass-media and self-media.

Finally, we figure out how public opinions can be changed through information networks. By monitoring the number of influential nodes and common nodes through the time, we find people are more inclined to convey the message after the influential ones do.

Contents

1 Introduction **3**

1.1 Background 3

1.2 Previous Work 4

1.2.1 Stochastic Network Model 4

1.2.2 Small World Network Model 4

1.2.3 Scale-free Network Model 5

1.3 Our Work 6

2 General Assumptions **6**

3 Data Analysis **7**

3.1 The Evoulution of Society's Networks 7

3.2 The Inherent Value of Information 9

3.2.1 Model Description 9

3.2.2 Model Testing 10

3.3 The Qualifications of News 11

3.3.1 Model Description 11

3.3.2 Model Testing 12

4 Models of Information Diffusion **13**

4.1 Epidemic Dynamics Model 13

4.2 a SDIR-Improved Model 14

4.3 Model Testing 15

4.3.1 The structure of a social network 15

4.3.2 Simulation of the information diffusion 16

4.4 Sensitivity Analyses 17

4.4.1 Sentivity Analysis of the number of D nodes in the initial state . . . 17

4.4.2 Sentivity Analysis of the Total Number of Nodes 17

5 Prediction of Current Social Information Network & Forecast of Social Information Network in 2050 **18**

5.1 Evolution of the social information network in different periods 18

5.2 Validation of Model 20

5.3 Analysi & Prediction 20

6 What change people's minds **21**

6.1 Model Decription 21

6.2 Model Testing 21

7 Strengths and weaknesses **22**

7.1 Strengths 22

7.2 Weaknesses 22

1 Introduction

1.1 Background

Everyday we check the e-mais in time,browse the latest news on phones or keep track of the social lives of entertainers.We have already been accustomned to the overwhelming information around us.But have you ever thought about how does the information flow through the world? How does the society's networks influence the public opinions? These intriguing questions lure numerous people to study the rules of the information diffusion.

Information diffusion is defined as the transmission of information among individuals,groups and organizations in order to convey knowledge,attitudes and emotions.It has the characters of immortality,infinite repeatability, directivity and timeliness.Information can be distributed in various forms,such as text,video and pictures.Studying information diffusion has a profund influence in capturing business opportunities, monitoring public opinions,guaranteeing the network safety and so om.

The models of information diffusion can be mainly divided into two categories:the explaination model and the prediction model.Explaination model observes the active nodes in a social network and infers the nodes' potential propagation path.Prediction model is designed to predict a specific process of communication on the basis of a given network structure.One type of the prediction model is the cascaded model based on the graph theory.It includes the independent casades model(IC) and linear threshold(LT).

Complex network model and infectious disease model are two typical information diffusion models.The complex network model is a kind of statistical analysis.It studies a large scale of nodes and investigates their contacts to represent different network structures.Complex network is a scale-free network in a small-world model.The typical terms of complex network are as follows:

- **Degree Distribution**
  Degree is the most common indicator the describe a single node.The degree of a node $k_i$means the total number of nodes which are connected to the node i.In a directed network,the in-degreee of a node is the number of edges pointing to it and the out-degree of a node is the number of edges starting from it.

  The average degree of a socail network is calculated as

$$\overline{k} = \frac{1}{N} \sum_i k_i \tag{1}$$

- **Average Path Length**
  The distance between of two nodes $i$and$j$ is defined as the minimum length of path $l(i,j)$between them.The average path length of a network is calculated as

$$I = \frac{1}{\frac{1}{2}N(N-1)} \sum_{i<j} l(i,j) \tag{2}$$

  where N is the total number of nodes in a network.

- **Cluster Coefficient**
  Cluster coefficient is a standard to measure the level of connection between two nodes.The cluster coefficient C_i of a node i is

$$C_i = \frac{2e_i}{k_i(k_i - 1)} \tag{3}$$

  where e_i is the number of edges among the nodes which are connected to the node$i$

1.2 Previous Work

1.2.1 Stochastic Network Model

Random graph theory is the most popular theory in the study of complex networks. This theory was established by two Hungarian mathematicians Erdos and Renyi in the 1950s. A random graph is a graph generated by a random process. The biggest difference between a graph and a classical graph is the introduction of a random method, which connects the given nodes in a random manner according to a certain random pattern.

If the random probability between two random nodes is p, the constructed random graph model is called $Erdos - Renyi$ random graph (ER random graph). There are two ways to construct ER random graph:

- For $G(N, M)$, first determine N points and then M edges between the N points;
- For $G(N, p)$, we also determine N points first. The probability of the edge between any two different nodes is p, and the connectivity of each point is judged.

If two random ER maps have the same p, then their properties are almost always the same:

- Average Path Length: $L_{ER} \propto \frac{\ln N}{\ln <k>}$
- Clustering coefficient: $C = p = \frac{<k>}{N} << 1$
- Degree distribution: $k \sim P(< k >)$

1.2.2 Small World Network Model

Many biological, technical and social networks lie between these two extremes. Here, we explore some simple network models through which these networks can be adjusted: Regular network "rewiring" to introduce more and more confusion. We found that these systems can be highly clustered like a lattice, but have a small characteristic path length, just like a random graph.

In the ER stochastic graph model, the average path length increases logarithmically, which is the nature of a classical small-world network. The network enhances signal propagation speed, computing power and synchronization, increasing contagion.

The small-world network is between the regular network and the random network. It is a model established by Watts and Strogatz in 1998, which is a general term of a class

of networks with shorter average path length and higher clustering coefficient. A small-world network starts with a circular regular network, keeping one endpoint of each edge unchanged while the other endpoint has the probability of p connected to a randomly selected node in the network. There can be at most one edge between any two different nodes, and each node can not have an edge connected to itself.

Due to the possibility that the randomization process in the WS-world model construction algorithm may undermine the connectivity of the network, Newman and Watts proposed the NW-small-world network model by replacing the WS-world network model with the random reconnection.

All small-world network models have the following properties:

- Average Path Length: $dist_c = \frac{2}{N(N+1)}\Sigma_{i\leq N}\Sigma_{j\geq i}dist(i,j) \propto \ln N$, which N refers to the quantity of nodes
- Clustering coefficient: When $K < \frac{2}{3}$, $C(i) \approx \frac{3}{4}$

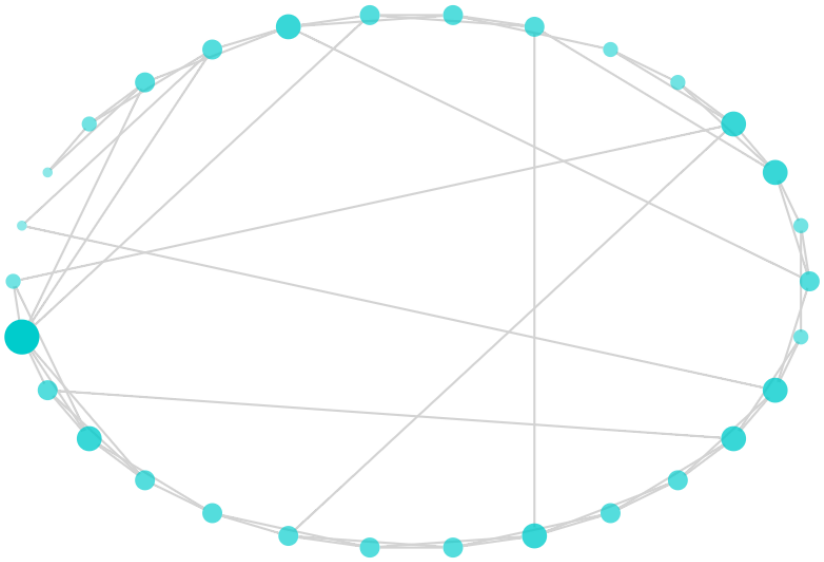

Figure 1: The classic model of Small-World Network

1.2.3 Scale-free Network Model

Scale-free networks are complex networks whose distributions conform to power-law distributions, with no obvious feature length. Most of the complex networks exposed in life are scale-free networks.

Albert and Barabasi proposed BA scale-free network model. Compared to other stochastic networks, scale-free network models will continue to grow and there will be a "Matthew Effect" that tends to be more connected to larger nodes.

When building a map of the BA network, each time a new node is introduced. Assuming that the degree of each existing node is k, the probability between each node and the new node is $\Pi(k_i) = \frac{k_i}{\Sigma_j k_j}$.

In many large-scale scale-free networks in reality, the degree distribution approximates the power-law distribution with a power index of three.

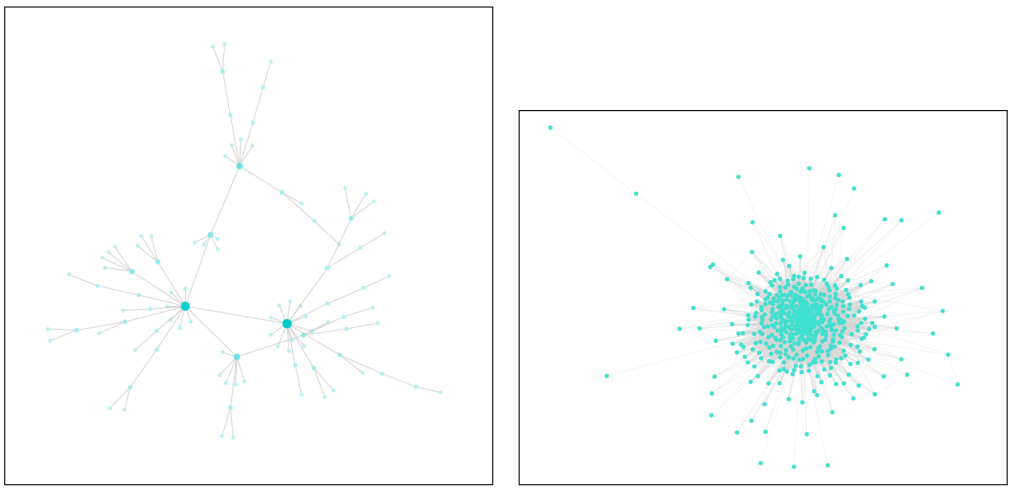

Figure 2: The classic model of Scale-free Network

1.3 Our Work

We tackle five main subproblems:

- investigate the evoulution of societyâĂŹs networks
- evaluate the inherent value of information and pick news out of information
- figure out how information flows through social networks
- predict the information diffusion currently and in 2050
- figure out how public opinions can be changed through information networks.

2 General Assumptions

To simplify the problems,our models are based on the following assumptions.

- **The total number of users is constant.**
  We assume the structure of the network is stable.In other words,the circle of a user's friends does not change during the flow of a certain piece of information.
- **We simplify the network as an undirected graph.**
  We assume the relationships between friends are mutual.Each user must follow his followers.So the graph of the network is undirected.
- **There isn't any isolated node in the network.**
  Since an isolated node is unable to receive or convey information,it does not have any effect on the information flow.Naturally,we ignore the existence of isolated nodes.

- **The network only has one propagation node in the initial state.**
  We assume a piece of information flows from a single source and that node is randomly selected.

3 Data Analysis

3.1 The Evoulution of Society's Networks

The society's information networks have experienced huge changes throughout history,enriching the lives of people in numerous aspects.According to the types of mass media,the evolution of information flow can be basically divided into five periods.

- In the early history,there were limitted ways to exchange information.It could only be spread by word of mouth.Then'print revolution' occurred in Europe in the fifteenth century and print media stepped into people's lives in the 1870s.Newspapers were delivered by trains and stories were passed by telegraph.
- The wireless telegraph machine was the forerunner to radio broadcasting.Lee de Forest—'the father of radio' managed to send the music and speech wirelessly and the radios became a household necessity in the 1920s.
- Soon,the invention of television captured the attention of people.From 1940s to 1970s,television was in its 'golden age',dominating the visual medium market.
- In the 1990s,the introduction of internet brought about earth-shaking changes to the social networks.People tend to gain easy access to the internet increasingly.
- Nowadays,in the 2010s,people are accustomed to connecting to the world on the phones.

In this media-saturated world,a variety of old and new media coexist.New media rarely completely replace old media,but new media do limit the growth of old media.For instance,Figure 3 depicts the downfall of newspaper paid circulation from 1940 to 2006 in America.This can be explained by the massive television users since 1955(Figure 4).

In addition,Figure 5 shows where Americans get news from 1991 to 2002.The shares of newspapers and radio drop continuously through these two decades while digital news surpasses them in a rapid speed.Television ranks as the public's top daily news source in America.But it also has a powerful competitor.The number of internet users rises dramatically from 2004,tiggering the proprotion of television eventually stablizes at 55% in 2012 after a little fluctuations. Above,we can see the rise of a medium often accompanies the decline of the others.

Apart from the evolution of types,mass media also transform in the operating modes at the same time.Traditional mass media were the scarce resource that connected content,consumers and advertises.Comsumers received information,to a large extent,dependent on the promotion of advertisers and they do not have much choices.

However,as the new media emerge,we step from the era of web 1.0 to web 2.0.The media model gradually transforms from advertiser-generated-media to consumer-generated-

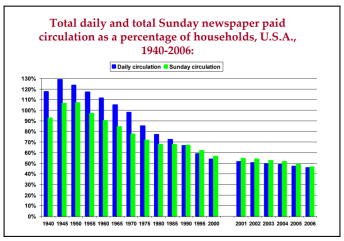

Figure 3: newspaper paid circulation of American households,1940-2006

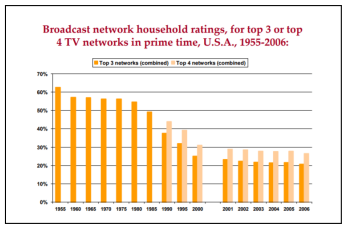

Figure 4: broadcast network household ratings,for top 3 or top 4 TV networks in prime time,USA,1955-2006

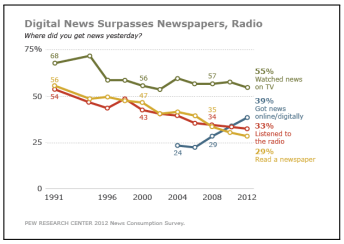

Figure 5: the trends of different media from 1991 to 2012 in America

media.Increasing people seize the initiative to gain access to thier desired information,such as searching online.

Specifically,Figure 6 visualizes the video continuum value chain of television in early 21st century.Information is diffused among content,consumers and advertises.The activities of people are heavily dependent on social networks.They are the core of this complex system ,playing a prominent role in the flow of information especialy in the consumer-generated era.

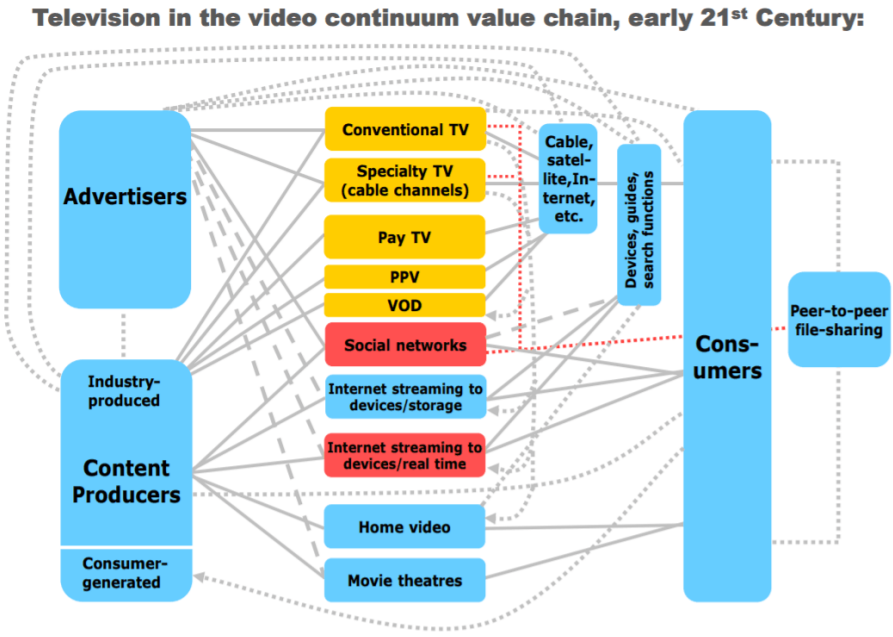

Figure 6: television in the video continuum value chain,early 21st century

3.2 The Inherent Value of Information

3.2.1 Model Description

To determine the inherent value of a piece of informaion,we set a series of parameters,considering the value of content and the quality of compilation respectively.We evaluate the value of content from five perspectives:authenticity,importance,timeliness,acuracy and completeness.Meanwhile we also evuluate the quality of compilation from five perspectives:whether it has a clear topic,thoughtful thinking, acurate description ,concise expression or comprehensive elements.

We determine the weights via Analytical Hierarchy Process(AHP)[Saaty 1982].For the value of content and the quality of compilation respectively,we build two5×5 reciprocal matrixs $G_1$and G_2 by pair comparison:

$$\mathbf{G_1} = \begin{pmatrix} 1 & 2 & 1 & 3 & 5 \\ 1/2 & 1 & 1 & 2 & 1 \\ 1 & 1/3 & 1 & 3 & 5 \\ 1/3 & 1/2 & 1/3 & 1 & 1 \\ 1/5 & 1 & 1/5 & 1 & 1 \end{pmatrix}$$

$$\mathbf{G_2} = \begin{pmatrix} 1 & 4 & 1/2 & 1 & 1 \\ 1/4 & 1 & 1/3 & 1/4 & 1/4 \\ 2 & 3 & 1 & 1 & 2 \\ 1 & 4 & 1 & 1 & 1 \\ 1 & 4 & 1/2 & 1 & 1 \end{pmatrix}$$

The meaning of the number in each cell is explained in Table 1.The numbers themselves are based our own subjective decisions.

Table 1: The multiplication table

Intensity of Value	Interprtation
1	Requirements i and j have equal value.
3	Requirement i has a slightly higher value than j.
5	Requirement i has a strongly higher value than j.
7	Requirement i has a very strongly higher value than j.
9	Requirement i has an absolutely higher value than j.
2,4,6,8	Intermediate scales between two adjacent judgments.
Reciprocals	Requirement i has a lower value than j

We then put the matrix into Matlab program that calculates the weight w_i of each factor,as given in 2.

Given a piece of information,we can estimate its inherent value from the parameters above.Indeed we think the value of content is somehow more important than the quality of compilation,so we set different weights for them on our subjective judgement.We give

Table 2: APH-derived weights

General factors	Specific factors	Weight
Value of content	Authenticity	0.3428
	Importance	0.1800
	Timeliness	0.2843
	Acuracy	0.0958
	Completeness	0.0971
Quality of compilation	Clear topic	0.2029
	Thoughtful thinking	0.0629
	Acurate description	0.2991
	Concise expression	0.2321
	Comprehensive elements	0.2029

out the general formulas as follows:

$$\begin{cases} \varphi_{val} = \sum_{i=1}^{5} w_i \mu_i \\ \varphi_{com} = \sum_{i=6}^{10} w_i \mu_i \\ \varphi_c = 0.6 \times \varphi_{val} + 0.4 \times \varphi_{com} \end{cases} \tag{4}$$

Where w_i is the weight of above 10 factors respectively and μ_i is the value of the corresponding indexes, ranging from 1 to 5. φ_{val} is the value of content and φ_{com} is the quality of compilation.φ_c is the value of a piece of information.

We test the consistency of the preferences for this instance of the AHP.For good consistency[Alonso and Lamata 2006,446-447];

- **Principal Eigenvalue**
  The principal eigenvalue $\lambda_{max}$of the matrix should be close to the number n of alternatives,here 5;we get λ_{max1}=5.1344 for G_1 and λ_{max2}=5.1208 for G_2.
- **Consistency Index** CI
  The consistency index $CI = \frac{\lambda_{max}-n}{n-1}$should be close to 0 For the value of content,we get $CI = 0.0336$. For the quality of compilation,we get $CI = 0.0302$.
- **Consistency Ratio** CR
  The consistency ratio$CR = CI/RI$(where RI is the average value of CI for random matrices)should be less than 0.1. For the value of content,we get $CR = 0.0300$. For the quality of compilation,we get $CR = 0.0270$.

Hence,our decision method displays perfectly acceptable consistency and the weights are reasonable.

3.2.2 Model Testing

To validate the reliability of our model,we use a sample of there pieces of news below.

- **News 1**
  Hiddleston broke up with Taylor Swift after 3 months of date.(True news)

- **News 2**
  US President Abraham Lincoln was assassinated on April 14,1865.(True news)
- **News 3**
  Adonis,a Syria poet,won the Nobel Prize in literature in 2016.(False news)

Based on the expert scoring method,we rate the ten indexes above for each news. The simulation results are shown in Table 3.

Table 3: Simulation of expert ranking for the given news

Value \ News / Index	1	2	3
Authenticity	4	5	1
Importance	2	5	4
Importance	5	3	1
Timeliness	4	4	1
Acuracy	3	3	1
Completeness	5	5	1
Clear topic	3	4	2
Thoughtful thinking	4	3	1
Acurate description	3	4	1
Concise expression	3	3	1
Comprehensive elements	4	4	2
Final Value	3.7783	3.9652	1.3492

Our results show the inherent values of News 1 and News 2 are obviously higher than that of the News 3.We can consider the first two news as official and newsworthy.On the other hand,although the content of the last news is attractive with great significance,we determine it as fake news according to its brief content,ambiguous expression,etc.In fact,the Nobel Prize winner is the Bob Dylan,an American folk singer, the last news is fake indeed. So our model is reliable under reasonable conditions,as can be seen from the testing above.

3.3 The Qualifications of News

3.3.1 Model Description

In the above analyses,we are able to determine the inherent value of information.However,when we try to filter the news out of endless information,the inherent value should not be the only standard.In our common sense,if a piece of information spreads in a slow speed and could only flow in limitted areas in a certain time,it can not be decribed as news regradless of its high value.So the speed of the propagation is an important index as well.

We propose the following formula to evaluate the quality of the newsφ_{news}.

$$\varphi_{news} = w_1\varphi_c + w_2 t_c \tag{5}$$

where φ_c is the inherent value of the information and its life span is t_c.$w_1$andw_2 are their weights.Obviously, w_1 is positive and w_2 is negtive.

In this way,we can quantify the value of news.Then we can set a thresholdφ_0,if$\varphi_{news} > \varphi_0$,the information could be considered as news literally.Otherwise,it is only informarion instead of news.

3.3.2 Model Testing

We test our model on Baidu Index,a data sharing platform based on the search volumn of netizens in Baidu.It introduces the term of search index,which is defined as the weighted sum of search frequency for a given keyword.

We take the news of AlphaGo as a sample to test our model.In March,2016,AlphaGo won the world champion Li Shishi,seizing the attention of people all over the world.The trend of its search volumn is vividly reflected in Figure 7.The average search index soars from 633 in February to its peak in March,finally falling down and stablizing at about 1838 in April.The maximum of its search index is about 64654 in the second week of March.We briefly take the life span of the news as the length of time from the peak search volumn to its lowest point. So the life span of this news is about 2 months.

We define $w_1 = 0.6$ and $w_2 = -0.4$ on our subjective.Then we calculate the inherent value of alphago,which is 3.8824,so the value of this issue is 1.5294.

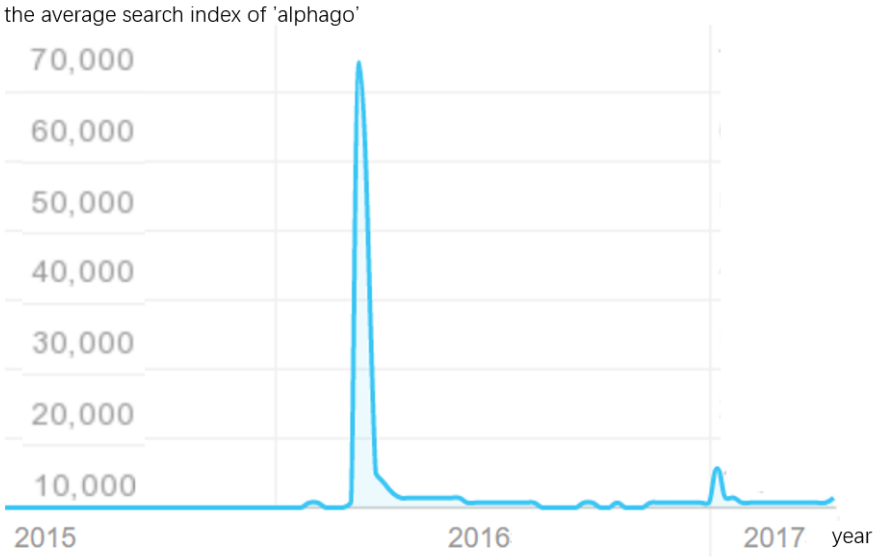

Figure 7: trends of the search volumn of alphago on Baidu

4 Models of Information Diffusion

4.1 Epidemic Dynamics Model

Epidemic dynamics model is a typical model to explore the information diffusion which is similar with virus transmission.Based on our problems,we initially bulid the Susceptible-Inflected-Recovered (SIR) model.The states of nodes fall into three categories.The susceptible node(S) represents the user who has not touched the information;the infected node(I) indicates the user who is transmitting the information;the recovered node(R) means the user will no longer transmit the message.

The rules of information flow in a SIR model are reflected in Figure 8.

The SIR model can be bulit by the following differential equations:

$$\begin{cases} \frac{ds(t)}{dt} = -\lambda s(t)i(t) \\ \frac{di(t)}{dt} = \lambda s(t)i(t) - \eta i(t) \\ \frac{dr(t)}{dt} = \eta i(t) \end{cases} \tag{6}$$

Based on the SIR model,we further build an improved model(SDIR)to explore the information flow more accurately.SDIR is the abbreviation of Susceptible-Disguising-Infectious-Recovered model.It introduces another state of nodes:Disguiser(D).Disguiser indicates the user who has received the information but has not transmitted it yet.He may spread the message after many of his friends do.In other words,the state of a node is dependent on its neighbor nodes.

The rules of information flow in a SDIR model are reflected in Figure 9.

The SDIR model can be bulit by the following differential equations:

$$\begin{cases} \frac{dS(t)}{dt} = -P_1 S(t)I(t) - P_2 S(t)D(t) \\ \frac{dD(t)}{dt} = P_2 S(t)D(t) - P_4 D(t) - P_3 D(t) \\ \frac{dI(t)}{dt} = P_1 S(t)I(t) + P_3 D(t) - P_5 I(t) \\ \frac{dR(t)}{dt} = P_4 D(t) + P_5 I(t) \end{cases} \tag{7}$$

Where P_i is the propagation probability between different states.(i=1,2,3,4,5.)

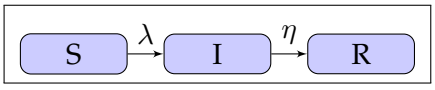

Figure 8: Nodes in the SIR model

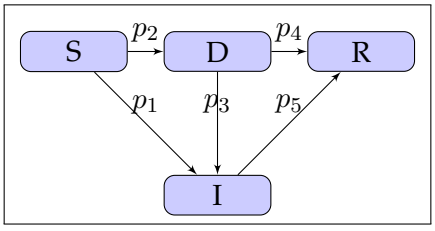

Figure 9: Nodes in the SDIR model

4.2 a SDIR-Improved Model

People's attitudes will change in the influence of interpersonal relationship,so we use the SDIR model to consider the effect of neighbor nodes and influential nodes.However,in the SDIR model,the propagation probability(P_i) between different states are constant.As for this limitation,we introduce a SDIR-improved model,successfully determining the propagation probability under different situations. First,we define three propagation probability functions.

- **Infection Probability Function**
  Infection probability function represents the probability of conveying a message instantly for a susceptible(S) node.The infection probability function$F_{SI}(g_i, g_j)$ pointing from the node i to the node j is defined as follows:

$$F_{SI}(g_i, g_j) = \begin{cases} \frac{2F(g_i,g_j)}{F(g_i,g_j)+F(g_j,g_i)} P_1 & , F_{SI}(g_i, g_j) < 1 \\ 1 & , F_{SI}(g_i, g_j) \geq 1 \end{cases} \tag{8}$$

  where k_g is the degree of the node g.We difine$\Gamma(g)$as the set of the neighbor nodes of g.$F(g_i, g_j$ is the relative weight of$g_i$ong_j,which indicates the influence of node g_i on nodeg_j.The specific formulas of$F(g_i, g_j and F(g_j, g_i)$are defined as follow:

$$F(g_i, g_j) = \frac{k(g_i)}{\sum\limits_{g_m \in \Gamma(g_j)} k(g_m)} \tag{9}$$

$$F(g_j, g_i) = \frac{k(g_j)}{\sum\limits_{g_m \in \Gamma(g_i)} k(g_m)} \tag{10}$$

- **Transition probability function**
  For a node in the state of D,it will transfer to the state of I or R.We introduce the transition probability function $F_{DI}(g)$ to describe the changes of states for the node g,which can be calculated below.

$$F_{DI}(g) = \begin{cases} \frac{3F_I(g)}{\sum\limits_{N \in (S,R,D)} F_N(g)} P_3 & , F_{DI}(g) < 1 \\ 1 & , F_{DI}(g) \geq 1 \end{cases} \tag{11}$$

  The specific formulas of $F_I(g)$ and$F_N(g)$ are defined as follows.

$$F_I(g) = \frac{k(g)}{\sum\limits_{g_m \in \Gamma_I(g)} k(g_m)} \tag{12}$$

$$F_N(g) = \frac{k(g)}{\sum\limits_{g_m \in \Gamma_N(g)} k(g_m)} \tag{13}$$

  Where $\Gamma_I(g)$ is the set of modes in the state of I among the neighbor nodes of g.

  As we can see from the equations above,among the neighbor nodes of g,if there are more nodes in the state of I than the state of R,the node g is more likely to transfer to the state of I

- **Survival function**
  The survival function $F_{IR}(g)$ is used to measure the propagation time of the information starting from the node g.Naturally,the node with a larger weight can spread the information in a longer period.Below,we present the definition of $F_{IR}(g)$.

$$F_{IR}(g) = \begin{cases} \frac{k(g)}{(\sum_{g_j \in \Gamma(g)} k(g_j))/N_{\Gamma(g)}} P_4 & , F_{DR}(g) < 1 \\ 1 & , F_{DR}(g) \geq 1 \end{cases} \tag{14}$$

Where $N_{\Gamma(g)}$ is the number of the neighbors of the node g.

4.3 Model Testing

4.3.1 The structure of a social network

We use the dataset from BlogCatalog in July,2009.BlogCatalog is a social blog directory which manages the bloggers and their blogs.All the contents are organized in CSV file format.We totally use 88784 nodes and there are 4186390 edges in the network.

Based on the improved-SDIR model,we simulate the degree distribution of a blog network (Figure 10).The degree of the nodes is reflected in the x-axis and the number of nodes with the corresponding degree is reflected in y-axis. After taking logarithm of them,,they show the linear relationship.In other words,the degree distribution can be fitted by the power-law distribution.To our relief,this finding is consistent with previous research in the literature.

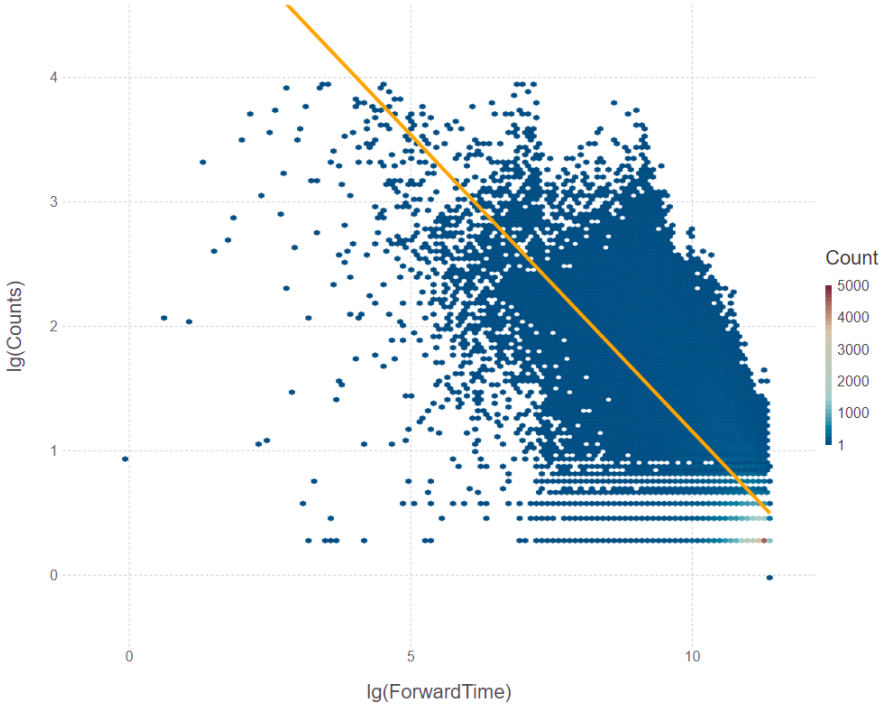

Figure 10: the degree distribution of a blog network

Besides,we simulate the overall structure of the blog network(Figure 11).We can vividly see the complex interpersonal relationship of this network.The nodes centralize at the source in a large density and they spread dispersedly far from the source.

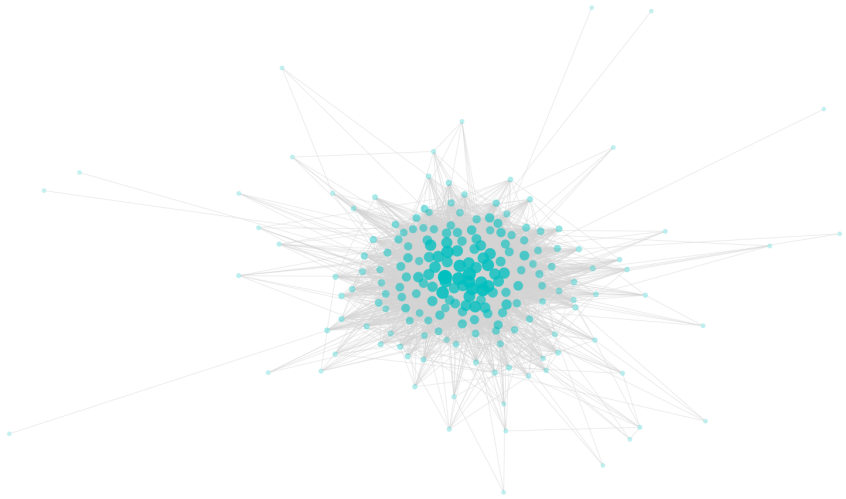

Figure 11: the overall structure of the blog network

4.3.2 Simulation of the information diffusion

We visulize the propagation time of a piece of information.In Figure 12,we use different colors to present various lengths of time.Specifically,the warmer color indicates the shorter time of information diffusion.Generally,we can find it takes longer time for the information to spread to the nodes far from its source.

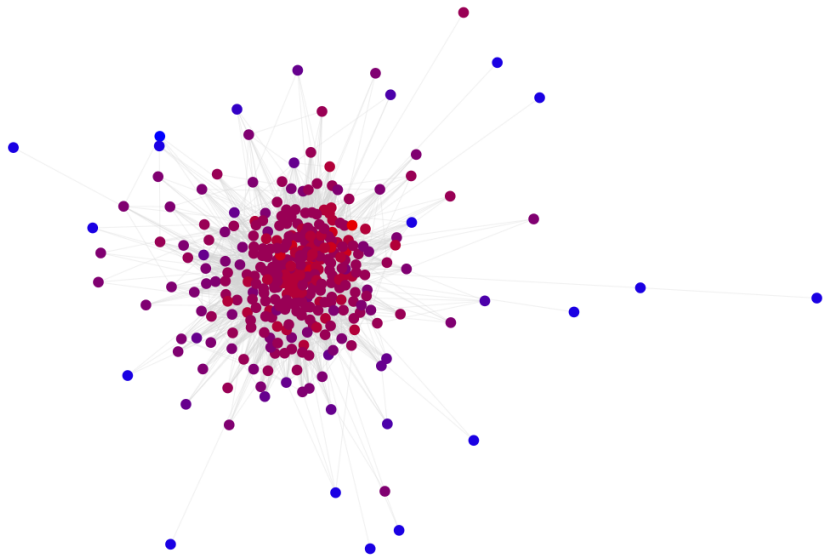

Figure 12: propagation time of the information in a blog network

4.4 Sensitivity Analyses

Based on the improved-SDIR model,two main features are the number of D nodes in the initial state and the total number of nodes.

Here are 4 initial parameters and 6 probabiliy constants in the $SDIR$ model. Because the probabiliy constants is influenced mainly by network structures, we will only analyze initial parameters in this section.

Firstly, we set $N = 81565, D = 1$ as the control group, where N means the number of people and D means the initial trend for the Disguiser.

For the first parameter D, we set it as 100 and 0.01. In the first group of figures, when D decreases, we can see the speed of propagation will be delayed, and the final number of influenced people will decreases, and vice versa.

For the second parameter N, we set it as 70000 and 100000. In the second group of figures, when N decreases, we can see the speed of propagation will be accelerated, and the final number of influenced people will increases, and vice versa.

Further, more disguisers lead to faster information spread in the network, the density of susceptible user "S" begin to decline shortly after propagating and the small number of disguisers cause dissemination of information lag, i.e. the number of S start falling fast spread over quite a period of time. On the other hand, when the network is large, the threshold of information dissemination is small. It shows that scale-free network, like CatalogBlog, is more likely to generate explosive public events.

4.4.1 Sentivity Analysis of the number of D nodes in the initial state

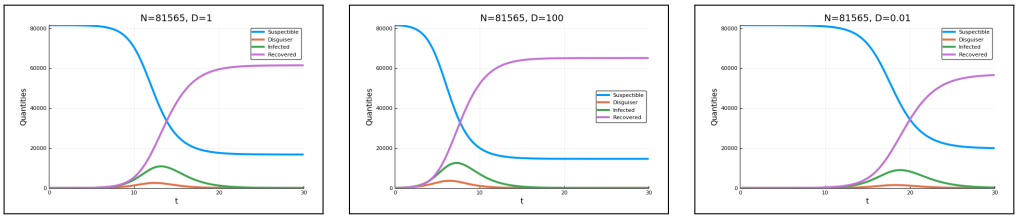

Figure 13: Sentivity Analysis of the number of D nodes in the initial state

4.4.2 Sentivity Analysis of the Total Number of Nodes

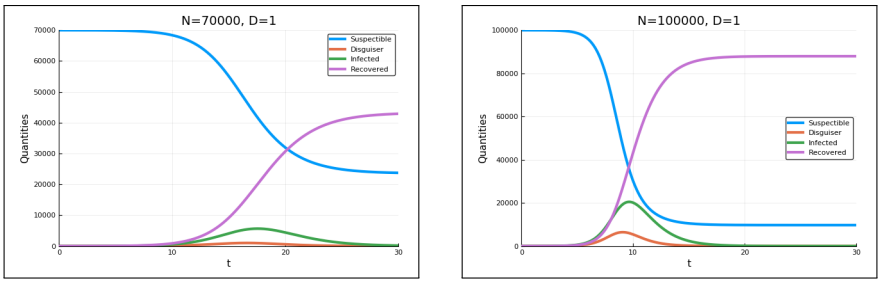

Figure 14: Sentivity Analysis of the total number of nodes

5 Prediction of Current Social Information Network & Forecast of Social Information Network in 2050

The social information network steadily change and grow over years. First, we observe the birth and popularization period of telecommunication network technology and visualize it as 4 After that, through the improved SDIR model, we simulate the dissemination of information flow in different periods, make predictions, validate our model and make several conclusions.

5.1 Evolution of the social information network in different periods

The new social information network technology between 1870 and 2010 is divided into 5 stages:

- The first stage in 1870s: the newspaper sent by train;
- The second stage in 1920s: the radio became more common household products;
- The third stage in 1970s: the emergence of television in most families;
- The fourth stage in 1995s: families are connected to the inchoate Internet;
- The fifth stage in 2010: people use smart mobile phone and Internt to connect with the world.

In order to explore the birth and popularity of new technologies for social information networks, more clearly, we visualize these important time points and get Figure 15.

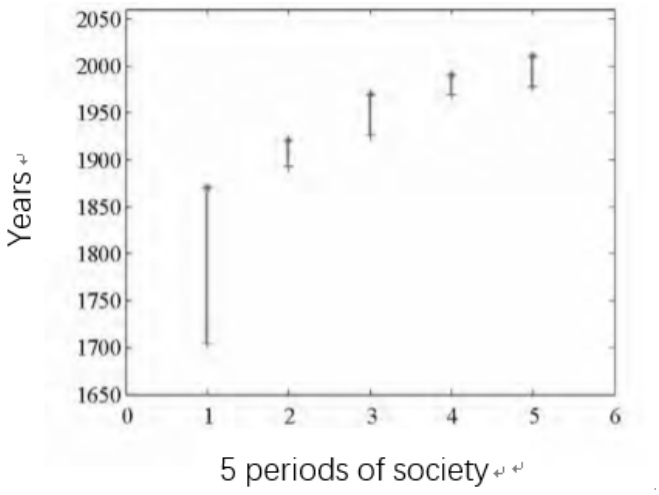

Figure 15: 5 stages of the social information network technology

As figure15 shows, the birth and popularization of a new social information network technology are getting faster and faster. Concretely, based on the relationships that:

1. The population in different periods reflects that the majority of information receivers.
2. The capacity of big nodes reflects the charactistics of media.
3. The infection probalitiy function F that varies with the degree of nodes, reflects the improvement of different propagation tools: newpaper, radio, television, inchoate Internet, smart phone and Internet.

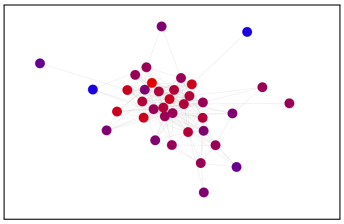

Figure 16: In 1870

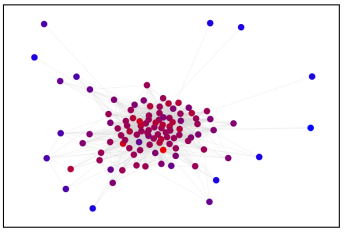

Figure 17: In 1920

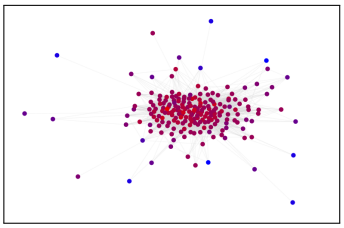

Figure 18: In 1970

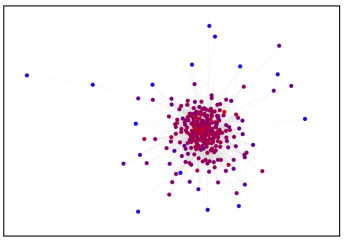

Figure 19: In 1995

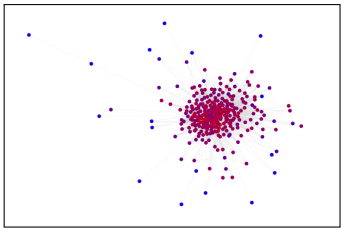

Figure 20: In 2010

We simulate the network in the periods mentioned above. The first stage in 1870s: the means of social information flow were poor, which mainly spread through newspapers and oral communication. The first phase of the social information network as shown in 16 the initial flourish of social information network took place in this stage, it exists some isolated nodes, four media nodes (red nodes) and few user nodes in the network. which is sparse and single-style. It means that random communication between user nodes is few and the ability of media posting information is weak. This period formed a preliminary star stucture.

In 1920s, as shown in figure17 the number of nodes increase and the spread ability of network go strong as some users far away from the media centre can get information. However, the spread efficiency is low as the most of outside points are blue(means getting the information slowly).

In 1970s, as shown in figure18 the network stucture(multipoint radiation) is similar with the previous period as televisions and radio are both based on wireless signal, but the number of users and medias increase considerably as the population rises rapidly during this period.

In 1995, as shown in figure19 the point distribution shows that the adjacent nodes tend to connect and exchange information close, also more and more media centres emerge at the less information-dense area, which means the initial form of individual media.

In 2010, as shown in figure20 the ability of connections and information spread present "ring" shape, which means the points show similar ability are located in the same cycle. The power of self-media expand to more remote areas.

5.2 Validation of Model

Through the comparison of simulation and reality among the indexes as table4 we can find that our model achieve a reasonable explanation.

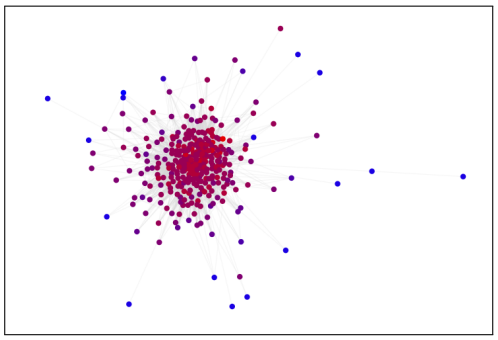

Figure 21: The simulation of information flow in 2010

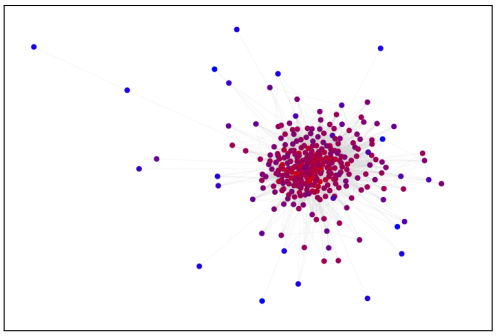

Figure 22: The reality of information flow in 2010 (data from BlogCatalog)

Table 4: Key feature of the networks

Types	Nodes	Edges	Maximum degree	Average Degree	Cluster Coefficient
Simulation	88705	4186012	9378	14.08	0.680
Reality	88740	4186390	9486	14.15	0.683

5.3 Analysi & Prediction

Based on the analysis of the 5 stages of social information network in the past 140 years(1870-2010), we make analysis and prediction as follows:

1. Due to the population growth and the further increase of people's information needs, the information network media nodes and user nodes will increase, and the fixed edges (fixed information propagation paths) between nodes will also increase.

2. The increase in the amount of information causes more people to have the opportunity to pay attention to the information in a particular field, and people shared similar interests and hobbies form a larger group scale clusters, the number of randomly connected edges (non-fixed information propagation paths) between nodes increases significantly, and appears a more pronounced cluster effect

3. In the near future, it is very likely that a new telecommunication technology will be born and popularized so that more user nodes have opportunity to disseminate information, so as to empower the function of nodes;

4. In 2050, first, the random connection of social information network and cluster effect will become more obvious. Second, the gap between users nodes and media nodes will reduce further, self-media becomes considerably strong, every user is capable of disseminate information to thousands of people by diverse tools. Importantly, both media nodes and user nodes tend to connect to nodes whose information interest is close to each other , it may trigger some other network effects, such as the old media connection decreases by a certain percentage.

6 What change people's minds

6.1 Model Decription

Definition 6.1 (Cluster Effect)**.** *The concentration of a certain area in a particular area of the division of labor and cooperation between the different sizes of enterprises and their development-related institutions.*

Definition 6.2 (Matthew Effect)**.** *A phenomenon sometimes summarized by the adage that "the rich get richer and the poor get poorer."*

In fact, the cluster effect is co-generated with the social information network, but with the development of information and communication technology, this property is more obvious. This property describes that both media nodes and user nodes tend to be close to the information interest (Shown in the figure as the closeness of information exchange between nodes).

The newly generated nodes are more likely to be nodes with higher connectivity / larger amount of information. If the newly generated node is a user node, it is more inclined to connect media nodes with higher outgoing / outgoing information volume; if it is a media node, it is more likely that the user node with higher connection / receiving amount of information .

6.2 Model Testing

In the modern society network, our opinions are mainly changed by the opinion leaders(central nodes), which are defined as the point whose degrees are larger than 1000.

When the Central Nodes propagate the information, the speed of spread will be accelerated. From the models above, we get the influence time in the whole network. Thus, we want to find the relationship between the influence of central nodes and the total nodes, as the figure shows below.

In this figure, we can see the relationship obeys the exponential distribution, which shows us the positive correlation of central nodes and total nodes. And it also refers:

1. Opinion Leaders have strong impact on the shift in people's perspectives.
2. The newly generated nodes are more likely to be connected with nodes whose have larger connectivity. If the newly generated node is a user node, it is more inclined to connect media nodes with larger out-degree; if it is a media node, it is more likely to connect users nodes with larger in-degree.

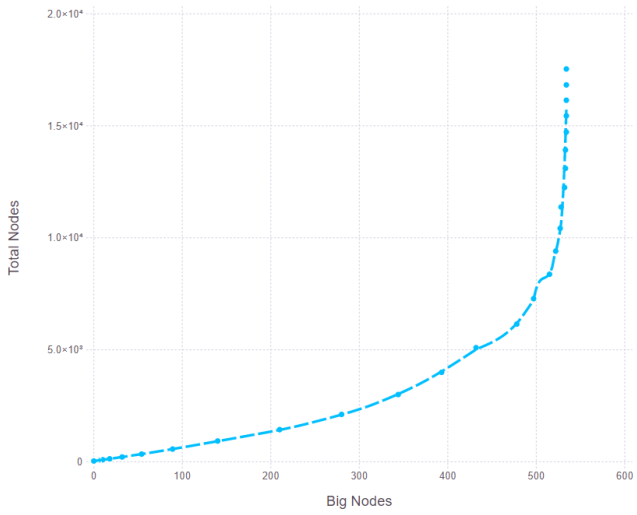

Figure 23: The relationship between central nodes and total nodes

7 Strengths and weaknesses

7.1 Strengths

- **Our model is roubst based on quantitative analyses.**
  Our model contains no arbitary parameters.Our prediction process is based on a rich supply of data so it is both objective and efficient.

- **Our results are understandable in our common sense and it conforms to the theory we found in the literature.**
  We simulate the degree distribution of a blog network and it is fitted by the power-law distribution as many researhes mentioned before.

- **Our model is flexible which can be applied widely.**
  We successfully present the process of information flow and quantify its speed and volumn.It is worhty to mention that our model can be applied to different structures of social network.This helpful and crucial work has a high value in business,such as helping us design a strategy to promote goods.Additionally,our model plays an essential role in monitoring public opinions and guaranteeing the network safety.

- **Our model provides an easy way for journalists to evaluate the values of news.**
  We propose an algorithm to evaluate the valus of information and judge whether a piece of information meet the qulifications of news.This can help the journalists do the self-assessment quickly and process better news in the future.

7.2 Weaknesses

- **Our model can only be applied to an undirected graph.**
  We assume the interpersonal relationship is mutual between friends.However,in many platporms such as Weibo and Twitter,a user does not have to follow his

fans.That means we may improve and perfect our model by intrducing the directed graph.

- **Our model does not consider the emerging new users in the network.**
  We consider the total number of users in the sensitivity analyses.However,the social network is rapidly changing.The number of users vary in almost every minute.It is necessary to investigate the dynamic changes of the number of users before different states of nodes stablize.

- **Our model is based on a single source of information.**
  To simplify the problem,we assume the network only has one propagation node in the initial state.Indeed,sometimes a piece of information starts from several sources.The number and the states of the source nodes may have effect on our model.

References

[1] Dang Keller. Research on prediction algorithm of social network information transmission based on neural network [D]. Nanjing University of Posts and Telecommunications, 2016.

[2] Sun Ling. Research on the mode of information and influence transmission in social networks [D]. Beijing Jiaotong University, 2017.

[3] HUANG Hong-Cheng, SUN Xin-Ran, HU Min.Social Network Information Propagation Model Based on Node Attitude [J]. Engineering Science and Technology, 2018 (01): 1-6 [2018-01-23] .https: /doi.org/10.15961/j.jsuese.201700093.

[4] ZHAO Jian-hua, WEN Ke-wen.Study on Social Network Public Opinion Propagation Dynamics Model Based on Information Dissemination Model-SIR Epidemic Model [J] .Journal of Information Science, 2017,35 (12): 34-38.

[5] Zhu Haitao, Zhao Pengwei, Qin Chunxiu.An improved SEIR model for mobile social networks [J] .Journal of Information Science, 2016,34 (03): 92-97.