For office use only	Team Control Number	For office use only
T1 ________	**47471**	F1 ________
T2 ________		F2 ________
T3 ________	Problem Chosen	F3 ________
T4 ________	**D**	F4 ________

2016 Interdisciplinary Contest in Modeling (ICM) Summary Sheet

Modeling Our Information Network

Abstract

In this article, we build 2 models respectively to simulate the information network and evaluate the influence of media on public opinion.

We start with defining what qualifies as news. By using a dataset of more than 39,000 entries of different levels of news, we set up a *Heat Standard* to evaluate value of news, and implement 5 different learning algorithms (*Logistic Regression, SVN, Random Forest, K-Nearest Neighbor, C4.5 algorithm)* to do the learning prediction. *Random Forest* algorithm has the best performance and is hence used as our news filter.

Next, we build up a *5-layer Grid Information-flow Model* to simulate the real-world information network. *Cellular Automata* and *Computational Simulation* are used to analyze the speed and broadness of information transmission in different historical stages. Every grid unit in different layer can represent individual, group, city or nation. And every piece of news is an information unit that have defined dynamics to interact with every gird. This strategy lays a solid foundation in simulating geographic limit and difference in transmission capacity. Stratification also allow us to maintain the essential topology feature within individual network as well as network between individual and media. Our result shows accuracy and stability in fitting current information communication situation. A prediction is made about the situation in 2050.

Finally, we discuss three famous theories of media influence on public and build our second model to investigate the media effects on public opinion. *BA Scale-free Network* and *Epidemic SIER Model* are used to represent current information flow through Internet. We find that the intensity of media coverage and the number of media involved in the transmission are two key factors on the number of people accepting the idea proposed by public media. Greater coverage intensity and larger number of media lead to deeper influence on public opinions.

Key words: Heat Standard; Random Forest; 5-layer Grid Information-flow Model; Cellular Automata; BA Scale-free Network; Epidemic SIER Model

Contents

1 Question Restatement

We are now living in a world with myriad information flooding around us. With the rapid development of news media and the enrichment of information network, the speed of information transmission is increasing along with broader range of influence. At present, news of different levels of importance can transmit rapidly in the information network, which is quite different from the information communication situation 200 years ago. Besides tradition information propagation media such as newspaper, radio and television, now we have handheld devices like mobile phones and tablets and hence a much easier access to various levels of information worldwide. However, the easy Internet access brings about different problems unlike the past, such as how we can filter out the trivial information and find out the real news. My teammates and I are designated to analyze the relationship between the flows of information vs. its inherent value. We are given a historical range of five period to do our study.

Considering the complexity of this study, we would like to separate it into four specific questions and to tackle them one by one.

a. How to filter the information and keep only those qualified as news?

b. How to model the information flow in different layers of communication network and through different transmission media? Whether our model has the reliability of correctly reflect the current information communication situation?

c. What is the communication situation in the year 2050?

d. Based on theories and concepts, how does current information network influences people's interest and opinion?

2 Assumptions

In order to streamline our model, we have made several key assumptions:

1. People do not share news which they find dubious about its credibility.
2. People are not likely to share news with low inner value.
3. People are more willing to share news with more inner value.

3 Symbols

Symbols	Explanation
x_i	Eigenvalue defined in the dataset
y	Class tag in news classification
$F(j)$	Fish score (Fisher criterion)

L_i	Layer tag in 5-layer information-flow network
N	Total number of individual unit in the model
S_i	Stage of transmission by different media
P_{ijk}	The probability of successful transmission for type j news in ith stage under kth condition
W_i	Information coverage of news i
Q_i	Total number of information unit of news i
$S(x,e)$	Coefficient of sensitivity

Model One

4 Question One

News filter and basic idea of model establishment

4.1 Data Source

Our data set is found from a research article A Proactive Intelligent Support System for Predicting the Popularity of Online News[1], the author of this article used an Intelligent Decision Support System to predict whether an article will become popular, based on a broad set of extracted features (e.g. keywords, digital media content etc.) . In general, the author collected 39,000 articles from the Mashable website and extracted about 60 features of each piece of news. We use this dataset as our start and draw out the Heat Standard to evaluate the importance (inner value) of news.

The dataset consists of about 40 thousand pieces of news with about 60 eigenvalue corresponding to each piece of news.

Table 1. List of attributes by category

Feature	Type	Feature	Type
Words		**Keywords**	
Number of words in the title	Number	Number of keywords	Number
Number of words in the article	Number	Worst keyword (min./avg./max. shares)	Number
Average word length	Number	Average keyword (min./avg./max. shares)	Number
Rate of non-stop words	Ratio	Best keyword (min./avg./max. shares)	Number
Rate of unique words	Ratio	Article category	Nominal

Rate of unique non-stop	Ratio	**Natural language processing**	
Links		Closeness to top 5 LDA topics	Ratio
Number of links	Number	Title subjectivity	Ratio
Number of Mashable article link	Number	Article text subjectivity score and its absolute difference of 0.5	Ratio
Min. Ave. and Max. number of shares of Mashable links	Number	Title sentiment polarity	Ratio
Digital media		Rate of positive and negative words	Ratio
Number of images	Number	Pos. words rate among non-neutral words	Ratio
Number of videos	Number	Neg. words rate among non-neutral words	Ratio
Time		Polarity of positive words (min./avg./max)	Ratio
Day of the week	Nominal	Polarity of negative words(min./avg./max)	Ratio
Published on a weekend?	Bool	Article text polarity score and its absolute difference to 0.5	Ratio
		Target	
		Number of article Mashable shares	number

4.2 News Identification

In order to determine what qualifies as news, we first define a four-level Heat Standard, corresponding to non-news, unimportant news, normal news and important news. Then we come up with a classification strategy based on machine learning algorithm. This strategy starts with a real-life news data learning process, followed by prediction based on the learning result.

4.3 Heat standard

We use Share Count to measure the four levels of news. According to our model assumption, the more inner value a piece of news has, the quicker, wider will it transmit and the more people will be influenced. Hence, it is reasonable for us to use Share Count as the classification standard.

Table 2. Classification Table

Share Count	Class
<900	0 (Non-news)
900~2000	1 (Unimportant news)
2000~6600	2 (Normal news)
>6600	3 (Important news)

4.4 Create Training Set

We tag the original dataset with the four classification (Table 1.) and create the training set.

4.5 Feature Selection

1. We calculate *the mutual information* $MI(x_i, y)$ between features and class labels to be the score to rank features. It is also referred to as the Kullback-Leibler (KL) divergence:

$$MI(x_i, y) = KL(p(x_i, y) \| p(x_i)p(y))$$

2. We use a second feature ranking method as the *Fisher criterion*. The Fish score for jth feature is given by:

$$F(j) = \frac{(\overline{x}_j^1 - \overline{x}_j^2)^2}{(s_j^1)^2 + (s_j^2)^2}$$

Where

$$(s_j^k)^2 = \sum_{x \in X^k} (x_j - \overline{x}_j^k)^2$$

The numerator indicates the discrimination between popular and unpopular news, and the denominator indicates the scatter within each class. Larger F-score indicates more discriminative feature. Then a cross-validation (with logistic regression) is used to find a feature size of k=10, which gives us the best performance over other values of k and mutual-information criterion. Therefore, we use this feature set as the eigenvalue of our training set.

Table 3. Feature Rank

Feature	Rank	Feature	Rank
Avg. keyword (avg. shares)	1	Best keywords	6
Avg. keyword (max. shares)	2	Avg. shares of Mashable links	7
Closeness to top 3 LDA topics	3	Closeness to top 2 LDA topic	8
Article category (Mashable data channel)	4	Worst keyword (avg. shares)	9
Min. shares of Mashable links	5	Closeness to top 5 LDA topic	10

Next we use different classification algorithms in learning prediction. Totally 5 learning algorithms for classification are used as below.

Table 4. Algorithms

Learning Prediction Algorithm
Logistic Regression
SVM (d=9 Poly Kernel)
Random Forest (500 Trees)
K-nearest Neighbors (k=5)
C4.5 Algorithm

For each algorithm, we conduct a 10-fold cross validation for inferring different levels of news. We evaluate the results using the following measurements in terms of precision and recall.

Table 5. Precision and Recall of Different Algorithms

Algorithm	**Precision**	**Recall**
Logistic regression	0.70	0.67
SVM	0.58	0.52
Random forest	**0.78**	**0.71**
K-nearest Neighbors	0.62	0.50
C4.5 algorithm	0.64	0.59

In general, we have both relative high precision and recall values, suggesting we manage a good selection of eigenvalues. In addition, the random forest method shows the best performance among five algorithms.

4.6 Summary

1. We have chosen the appropriate eigenvalues to best represent the feature of Mashable news.
2. The methods we use to select the useful eigenvalues is effective and we are able to maintain the precision of the prediction.
3. We obtain relative good prediction result using different algorithms, among which Random Forest is the best with 0.78 precision and 0.71 recall.

4.7 Model Establishment—basic concept

We have come up with a model concept based on real-life with different levels of communication network. The lowest level is one individual. The second level, as the group level, is an egocentric network centered at the individual. The third level, city level, consists of different groups distributed in the city. Different groups are treated as independent ones, in other words, they can only have one representative connection with any other group. Moreover, we assume that the groups of people is constrained in the city they live in, hence news can only be flowing into another city when the original news reach the city level. The national level and the global level follow the same principle. They consists of different With our basic concept of communication network, we are able to establish a manipulable network to trace the flow of news.

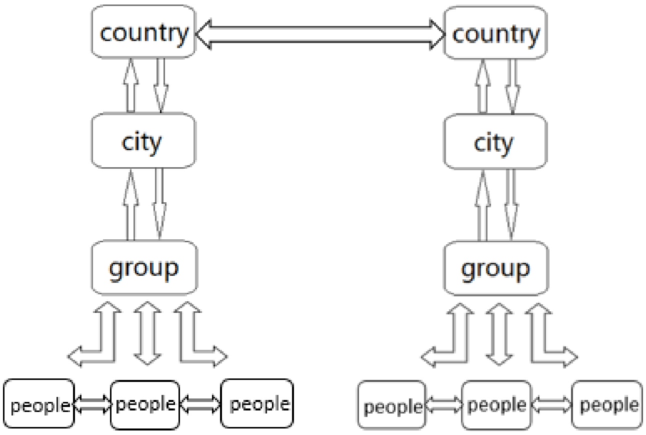

Figure 1. Basic concept of network model

5 Question Two

5.1 Model Building

Based on the basic concept of our communication network, we can draw a *5-layer grid information- flow model*, with Global layer, National layer, City layer, Group layer and Individual layer. Each layer is built on the lower layer, while individual is referred to as an independent information collector and processor. After constructing the Grid Information-flow Model, we define specific behavior of information-flow and corresponding probability for each grid unit. The behavior and probability are used to simulate different grid's (i.e. individuals, groups, cities, nations) influence on information flow.

Next we use cellular automata to simulation the information flow in this 5-layer networks. Information of news is specified as information block, which is flowing in and between each network.

Last, we use uniform standard to measure information-flow capacity in different stages of history.

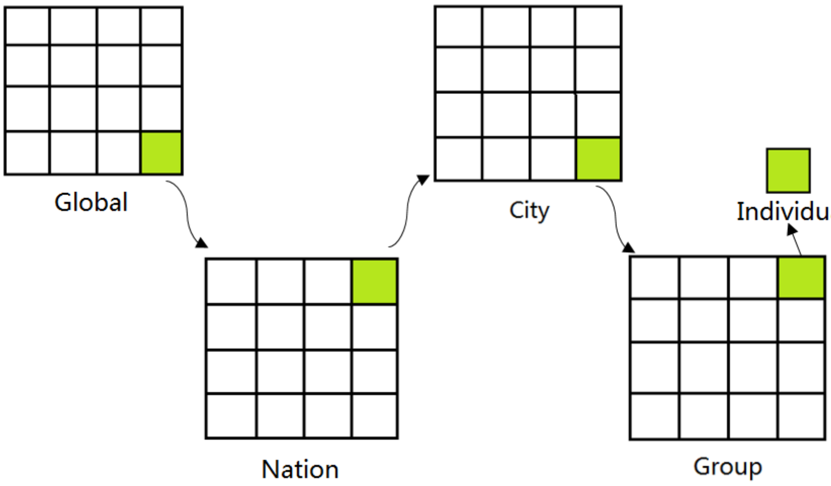

Figure 2. Structure of 5-layer grid information-flow model

5.2 Specify the Model

Quantifying and gridding the 5-layer network, we use grid unit to construct each layer of network.

Table 6. Network layer description

Network Layer	Description
Global Layer	With about 220 recognized nations in the world, we use 15*15 grid network to describe the abstract global network. Each grid unit represent a nation.
National Layer	We assume every nation has around 100 cities and use a 10*10 grid network to represent each nation and its cities.
City Layer	We assume every city has around 2025 groups and use a 45*45 grid network to represent each city and its groups.
Group Layer	According to the 'Rule of 150'*, we use 12*12 grid network to represent each group and every individual inside.
Total	The total individual number worldwide is 225*100*2025*144≈6.5billion, approximately the real total number.

*In 1993, British Scholar Robin. I. M. Dunbar and his research group showed that the scale of an animal population is correlated with the scale of neocortex relative to its whole brain. After scientific calculation, the suitable group of people should be about 150, i.e. one individual can maintain his/her relationship with at most 150 people. This is the famous biological 'Rule of 150'.

Table 7. Connecting Rules

Network Layer	Rule
Global Layer	The global network contains different nations, and news information can flow between each national grid unit.
National Layer	Each national network contains different cities, and is connected with every city unit. Information can flow into national unit from sub-layer city unit, or flow out of national unit into city units (not every city unit).
City Layer	Each city network contains different groups, and is connected with every group unit. Information can flow into city unit from sub-layer group unit, or flow out of city unit into group units (not every group unit).
Group Layer	A group of individual is a group in which people have relative frequent contact with each other. They are connected with each other and may exchange information from time to time.

Table 8. Flow capacity of different media in different network layers

Media Type	Flow Capacity
Newspaper	Transmission among groups / between up-layer city and groups
Radio	Transmission among groups / between up-layer city and groups / between up-layer nation and cities
Television	Transmission among groups / between up-layer city and groups / between up-layer nation and cities
Internet	Transmission among groups / between up-layer city and groups / between up-layer nation and cities / among nations
Mobile	Transmission among groups / between up-layer city and groups / between up-layer nation and cities / among nations

Table 9. Transmission dynamics of different layers of network and different media

Grid Unit	Transmission Dynamic Description
National Unit	For every specific information, possession of news comes along with the exact moment the news is transmitted into the unit. A constant probability is set to let the information flow into sub-layer city units in the subsequent time unit, as well as the same layer other national unit.
City Unit	For every specific information, possession of news comes along with the exact moment the news is transmitted into the unit. A constant probability is set to let the information flow into sub-layer group units in the subsequent time unit, as well as the same layer other city unit.
Group Unit	For every specific information, possession of news comes along with the exact moment the news is transmitted into the unit. A constant probability

	is set to let the information flow into sub-layer individual units in the subsequent time unit, as well as the same layer other group unit.
Individual Unit	For every specific information, we separate the individual into two cohorts, known or unknown. Individual in the known cohort are capable of transmitting information, and will transmit information to unknown individuals (of the same group) in the subsequent time unit with a constant probability.

5.3 Simulation Methods

We use cellular automata method to establish and simulate our information-flow model. Each piece of news information are treated as an information unit that can move and transmit through the 5-layer network, following the Transmission Dynamics (Table 9). In addition, each information unit are classified into 3 classes, important news, normal news, and unimportant news. For each unit in the 5-layer information-flow network, possessing the information unit means knowing the news, then in the subsequent time unit follows the transmission dynamics described in Table 9, i.e. the replication and movement of cell in the cellular automata.

Table 10. Parameter description for different media stage in the model

*Abbreviation: TP as transmission probability; I as individual; G as group; C as city; N as national; O as global

Stage	Importance	TP from I to G	TP from I to I	TP from G to I	TP from G to C	TP from C to G	TP from C to N	TP from N to C	TP from N to N
1870s Newspaper	**Important**	0. 15	0. 08	0. 13	0. 09	0. 21	0. 06	0. 15	0. 09
	Normal	0. 12	0. 05	0. 10	0. 07	0. 17	0. 04	0. 13	0. 07
	Unimportant	0. 10	0. 03	0. 06	0. 04	0. 15	0. 02	0. 08	0. 04
1920s Radio	**Important**	0. 18	0. 12	0. 16	0. 14	0. 23	0. 13	0. 23	0. 15
	Normal	0. 15	0. 10	0. 12	0. 12	0. 20	0. 09	0. 20	0. 13
	Unimportant	0. 13	0. 07	0. 09	0. 08	0. 18	0. 07	0. 16	0. 10
1970s Television	**Important**	0. 20	0. 15	0. 18	0. 18	0. 29	0. 19	0. 29	0. 20
	Normal	0. 17	0. 12	0. 14	0. 16	0. 27	0. 17	0. 26	0. 18
	Unimportant	0. 15	0. 09	0. 11	0. 14	0. 24	0. 14	0. 22	0. 15
1990s Internet	**Important**	0. 26	0. 19	0. 23	0. 20	0. 35	0. 24	0. 34	0. 23
	Normal	0. 22	0. 15	0. 20	0. 16	0. 32	0. 22	0. 32	0. 21
	Unimportant	0. 19	0. 13	0. 16	0. 14	0. 30	0. 18	0. 27	0. 18

2010s Mobile	**Important**	0.34	0.27	0.29	0.30	0.46	0.28	0.40	0.31
	Normal	0.31	0.24	0.27	0.28	0.43	0.25	0.37	0.27
	Unimportant	0.25	0.22	0.24	0.25	0.41	0.23	0.35	0.25

5.4 Evaluation Criterion

Based on our model and reality, we define 2 evaluation criteria.

1. Information coverage: the ratio of individual unit known the information relative to the total individual units.

2. Information amount: the amount of information units in our 5-layer information-flow model.

5.5 Experiment Procedure

Due to the large amount of individual unit in our model (about 6.5 billion), it is almost impossible to simulate the information flow with personal computer in such a short period of time, hence we scale down the model and apply our model in a 5-layer network with 10 million individual units.

For every historical stage, we initialize our model with three classes of news information unit, corresponding to important, normal and unimportant. We make 10 copies for each class of news and put totally 30 information units into our initial model, then we let it iterate for 100 time unit, and retrieve the evaluation criterion.

5.6 Experiment Result

Table 11. Experiment result

Stage	Importance	Information Coverage	Information Amount
1870s Newspaper	Important	0.5%	5.2×10^{5}
	Normal	0.4%	4.2×10^{5}
	Unimportant	0.2%	2.1×10^{5}
1920s Radio	Important	1.6%	1.7×10^{6}
	Normal	1.4%	1.5×10^{6}
	Unimportant	0.7%	7.5×10^{5}
1970s Television	Important	9.0%	9.5×10^{6}
	Normal	7.8%	8.2×10^{6}
	Unimportant	4.5%	4.8×10^{6}
1990s Internet	Important	18.0%	1.9×10^{7}
	Normal	14.9%	1.6×10^{7}
	Unimportant	8.6%	9.0×10^{6}
2010s Mobile	Important	27.0%	2.8×10^{7}
	Normal	18.7%	2.0×10^{7}
	Unimportant	13.6%	1.4×10^{7}

5.7 Verification of our model

Using experiment data from the early four stages, we are able to draw a scatter plot and fit the data with polynomial regression method, the regression line intercept x=2010 and we can get the evaluation value. By comparing the evaluation value and the evaluation value from the cellular automata, we are able to determine the effectiveness and accuracy of our model.

5.8 Experiment Plot

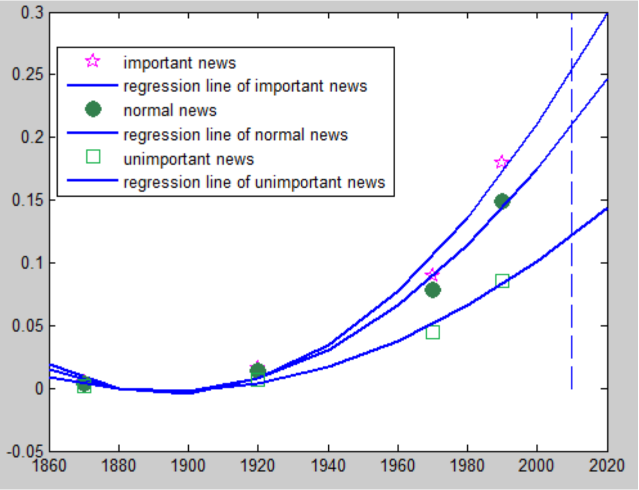

Figure 3. Verification of our model

By comparing the estimated evaluation value and the cellular automata evaluation value in 2010s, we find that they are relative equal to each other, and hence verify the effectiveness and accuracy of our model.

6 Question Three

Based on the five stages experiment data, we fit the regression line and get the intercept coordinate with x=2050. The evaluation value of 2050s as follow.

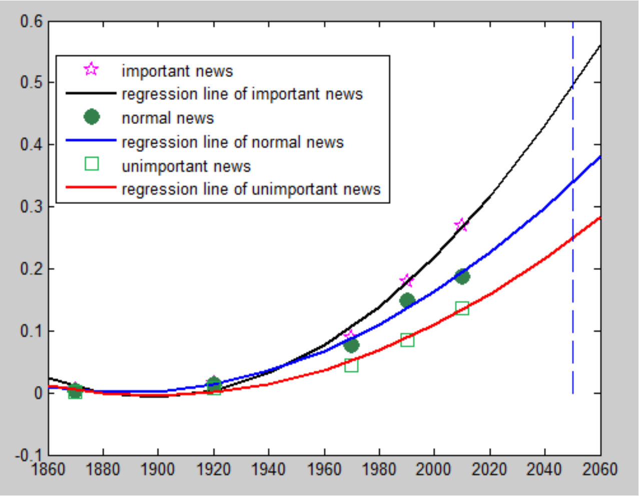

Figure 4. Prediction of year 2050

Table 12. Experiment result

Stage	Importance	Information Coverage	Information Amount
2050s	**Important**	49. 9%	5. 3× 10^7
	Normal	34. 1%	3. 6× 10^7
	Unimportant	24. 9%	2. 6× 10^7

Model Two

7 Question Four

Media are often referred as 'watchdog' for the government, because they often play an important role in keeping government from taking too much power from the people. Aside from the neutral media that are of equality and self-justice, there are other types of media that are too much polarized. For those controlled by the state regime, the media is much more like a 'lapdog' when they are too cozy with a politician or other public figure. Those 'lapdogs' might lead to unauthentic reports or information that distract people from the real issue. Another type of problematic media is called 'attack-dog', for their role is primarily to sneak out the possible scandal of public figure and exaggerate it. For neutral media, its role is merely transmit objective information to public and to express its opinion without preference, which are not very likely to cause significant influence on public interest and opinion. However, the polarized media, which are other for or against its target, will somehow have impact on their audience, either in rapid or chronic way.

We have put our eyes on three most famous theories on media effects.

The first is called the hypodermic needle approach[2] or bullet theory. It is a basic thinking with flaws. The main idea is that certain stimuli (such as a product is on sale) is connected to certain behaviors[3] (such as people's reaction to buy the product they may not otherwise want or need).

The second is the theory of media effects. The theory claim that it is true for media to directly and intentionally influence audience members. Additional claim is that media messages have no little power over viewers. Combining two claims can conclude that media messages do affect viewers but that viewers also have some agency to identify with, reject or reinterpret a message[4].

The last theory is cultivation theory, created by George Gerbner. It states that media exposure, specifically to television, shapes our social reality by giving us a distorted view on the amount of violence and risk in the world, and viewers identify with certain values and identities that are presented as mainstream on television even though they do not actually share those values or identityes in their real lives.

On the foundation of these three theories, we are reasonable to build up a model to consider the effect to information network on public opinion through specific perspective.

As the main carrier of information, Internet media plays a more important role in providing public a broader, faster and real-time information platform than traditional media. With this common notion, we build up a BA scale-free network and epidemic transmission SIER model to analyze media impact on public from two different aspects, the total number of internet media and the intensity of media coverage.

7.1 Model Building

Internet network is a complex network, with no specific length for the connection of each nodes. We set up our initial model with m0 nodes randomly connected with each other. And we continue inputting new nodes. A new node will be connected with m existing nodes, here m m−0. We set m0=4, m=3. The connecting probability for existed node to connect to the new node is expressed as:

$$\Pi(k_i) = \frac{k_i}{\sum_{j=1}^{N-1} k_j}$$

Where k_i the number of edge of an existing node *I; N* is the number of nodes in the network.

We define the dynamics rule for each node to input and output information. First we set up five stages of nodes (i.e. four stages of people + media), the unknown node, the potential node, the transmission node, the immune node and the media.

Table 13. Node Stage, Denotation and Description

Node Stage	**Denotation**	**Description**
Unknown node	S	People who do not know the information
Potential node	E	People who know but do not tell yet
Transmission node	I	People who tell others
Immune node	R	People who are immune to the same information
Media	M	Media that help transmit the information

- Rule One: Unknown node *S* is affected by its neighbors, with α_1 probability to receive the information and becomes potential node *E*; also, unknown node S is affected by surrounding media, with α_2 probability to become potential node E.

- Rule Two: Potential node *E* has a probability of β_1 to become transmission node I; potential node E has a probability of β_2 to become immune node *R*. $\beta_1 + \beta_2 = 1$.
- Rule Three: transmission node *I* has a probability of γ to lose its interest and become immune node *R*.

Here is a schematic diagram to show the structure of the transmission model.

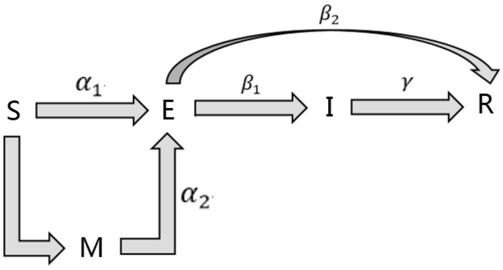

Figure 5. Schematic Diagram

Based on the transmission model and epidemic transmission SIER model, we come up with the following equations.

$$\begin{cases} \dfrac{\partial s_k(t)}{\partial t} = -\alpha_1 k s_k(t)\Theta_k(t) - \alpha_2 P_t(m \mid k) s_k^m(t) \\ \dfrac{\partial e_k(t)}{\partial t} = \alpha_1 k s_k(t)\Theta_k(t) + \alpha_2 P_t(m \mid k) s_k^m(t) - (\beta_1 + \beta_2) e_k(t) \\ \dfrac{\partial i_k(t)}{\partial t} = \beta_1 e_k(t) - \gamma i_k(t) \\ \dfrac{\partial r_k(t)}{\partial t} = \beta_2 e_k(t) + \gamma i_k(t) \end{cases}$$

We focus our study on $\alpha_2 = \theta - \theta e^{-\delta n}$

Where

- α_2 is the probability that a node change from S to E
- θ is the credibility of media in public view. Based on Pareto's Law[5], we take θ =0.8 for there is assumed to be 20% audience casting doubt on the media.
- δ is the intensity of media coverage, with range of [0, 1]
- n is the number of media involved in the transmission

Additional assumption: the larger probability people tend to transmit the information, the larger influence the internet media has on public opinion.

According to additional assumption, the α_2 probability is linked to the influence of internet media on public opinion. From this equation, we infer that the more intensity of media coverage, the larger influence of internet media. Moreover, the larger number of media in the transmission, the larger influence of internet media. Based on this two inference, we want to investigate the influence of the intensity of media coverage and the number of media participants separately on the public opinion. The public opinion is quantified by the number of immune node R or the ratio of immune node and total node, because the immune node is the final stage of information transmission.

Factor One: Intensity of media coverage. By changing the media coverage δ , we are able to compare different levels of coverage and their effects on the ratio of immune node and total node, using computational simulation.

Result: we find that as the coverage intensity increases, the responding time of curve reaching stable status decreases while the ratio of immune node and total node increases.

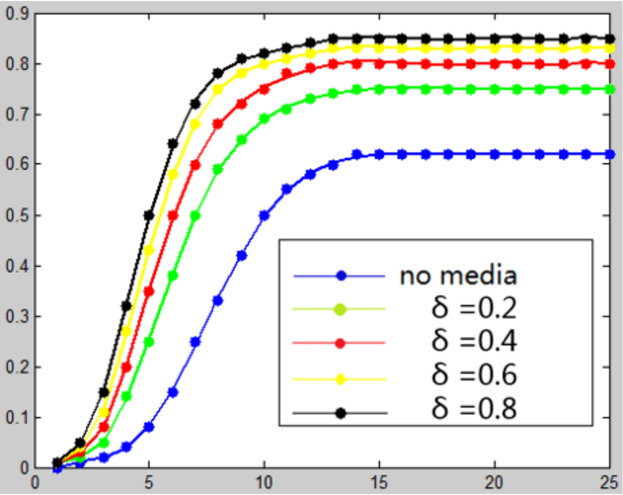

Figure 6. Intensity of media coverage

The result corresponds with our prior experience that when the media increase the time and frequency to report a piece of news, people are more tending to remember and discussion them in their friends circle. Separate estimation results from simulation or regression draw approximately the same conclusion.

Factor Two: Number of media involved. We also alter the number of media involved n and compare different media numbers and their effects on the ratio of immune node and total node.

Result: As the number of media n increases, the final ratio of immune node and total node also increases, with less responding time to stable status.

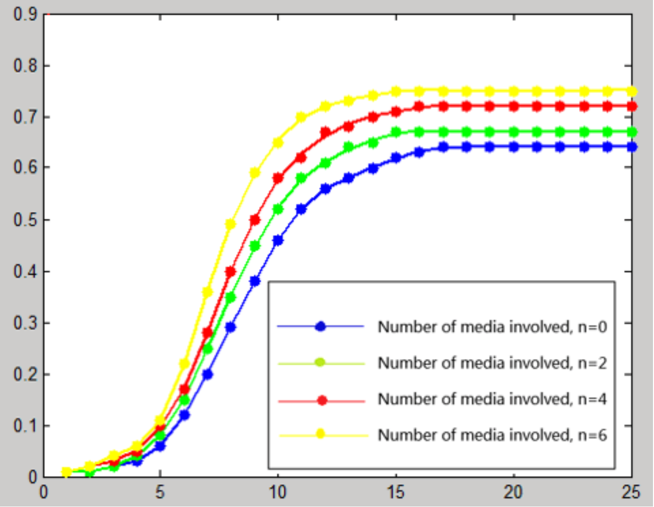

Figure 7. Number of media involved

This result correspond with our first and second media theory on how media shapes public's view and behavior. Separate estimation results from simulation or regression draw approximately the same conclusion.

8 Error Analysis

8.1 News Filter

In the evaluation of learning model, it is critical for us to do the prediction error analysis. We use ten-fold cross validation to test the error of the news filter. The reason why we choose ten-fold is because based on a large number of dataset and various learning techniques, ten-fold is the best choice for error analysis. We separate the dataset into ten groups, take in turn nine of them as training set and one as testing set. Every trial gives an accuracy rate. After ten trials, we calculate the accuracy rate and conclude that the total 10 accuracy rate is stable and no apparent fluctuation is detected, indicating good control of error.

8.2 Information-flow Model

The main source of error comes from the model assumption. In order to facilitate calculation, we use 5-layer information flow network to simulate the real world information flow. In this simplified model, we ignore the difference in nations, cities and groups and use universal parameters to simulate. This simplification strategy will definitely cause deviation and error. Nevertheless, original intention to build the model is to simulate the general structure of information network and general tendency in information flow, the error in geology or local development can be reasonably excluded from our model. Hence, it is acceptable to use this information-flow model to simulate real world information network.

9 Sensitivity Analysis

9.1 Sensitivity Analysis in Information-flow Model

We define 'coefficient of sensitivity' to calculate sensitivity. Formula as follow:

$$S(x,e)=\frac{\Delta e}{\Delta x}$$

Where Δx is percentage change for uncertainty parameter, Δe is percentage change in model calculation

In our information-flow model, P_i is denoted as uncertainty parameter, hence we can measure the sensitivity through calculating $S(p_i,e)$. The result shows that coefficient of sensitivity are all within the acceptable range, indicating a small sensitivity for our model.

10 Strength and Weakness

10.1 Strength

1. In news filtering process, we use a dataset of more than 39,000 entries of news data. The abundance of data provide enough training set for our model and increase the model stability and accuracy.
2. Considering the large number of eigenvalues defined by data, we first use F-score to filter out the irrelevant eigenvalue and leave those most represent the news. Eigenvalue extraction effectively decreases our model dimension, saves our resource and excludes the inference factor.
3. We implement five different learning algorithms to do learning prediction and select the best algorithm as our final strategy.
4. We use stratified network structure to simulate the information-flow in the real world. Our model simulates the geographic limitation of media and keep the essential feature of connection within individuals and between individual and media. In general, we succeed in simplify the model while keeping the feature of information transmission.
5. Cellular automata and computational simulation methods are used to do the simulation and calculation in our model. After assessing the accuracy and stability of our model, we

conclude that it has a good performance in reflecting the capacity and feature of information transmission in different stages.

6. In order to investigate the influence of media on the public interest and opinion, we use BA scale-free network and epidemic SIER model to simulate the process. The finding is that compared with old times, current information network enable public media (especially internet media) to exert a bigger influence on public opinion.

10.2 Weakness

1. The news filter is based on a relative simple index to reflect inherent value. It might cause inaccuracy in classification when we do not consider other features of news.
2. In our information flow model, we do not consider the difference in local development and inter-region communication, which causes error in our simulation.
3. Simplification of information flow process cannot reflect the whole picture of information network and information flow dynamics.
4. BA scale-free network and SIER model has their inherent flaws and errors.

11 Reference

[1] K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015- Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.

[2] Berger, A. A. (1995). Essentials of Mass Communication Theory. London: SAGE Publications.

[3] Charles C. Self, Edward L. Gaylord, and Thelma Gaylord, “The Evolution of Mass Communication Theory in the 20th Century,” The Romanian Review of Journalism and Communication 6, no. 3 (2009): 34.

[4] Denis McQuail, McQuail’s Mass Communication Theory, 6th ed. (Thousand Oaks, CA: Sage, 2010), 465.

[5] THE APPLICATION OF THE PARETO PRINCIPLE IN SOFTWARE ENGINEERING. Ankunda R. Kiremire 19th October, 2011