Epidemiological Modeling of News spreading on Twitter

Yunran Chen
MSS, Duke University
yunran.chen@duke.edu

Junwen Huang
MSS, Duke University
junwen.huang@duke.edu

Abstract

We aim to characterize how information spreads on social platforms like Twitter.[1] *Our analysis is based on dynamic retweet network related to Higgs boson discovery. In the first part we presented a thorough exploratory data analysis on the information spreading processes before, during and after the announcement of the discovery of Higgs boson. Specifically, how network statistics and structure vary with time. In the second part, we use epidemiological models (SI model, SIS model and SEIZ model) to capture information diffusion in this event. And we make predictions on the characteristics of the information spreading process such as transition rate and so on. The compartmental models in epidemiology performs well at fitting the data under different situations.* [2]

1. Introduction

1.1. literature review

Compartmental models in epidemiology provide a classical approach to study how information diffuses. The basic idea is to divide the total population into different compartments, which reflects different status of an individual. Within each compartment, individuals share the same characteristics. These models, which are basically ordinary differential equations, are originally used in mathematical modelling of infectious disease. The simplest is the SI model, which has two states S(susceptible) and I(infectious). As extensions, researchers have developed SIR (susceptible , infectious, recovered) model, SIS (susceptible, infectious, susceptible) model, SEIZ (susceptible, exposed, infected, skeptic) model and so on. Many studies further developed epidemiological models and applied them on studying information diffusion in networks. (Newman *et al.* [4], Zhao *et al.* [5] and so on)

Here our study is mainly based on Jin *et al.*'s work.[3] We applied SI model, SIS model and SEIZ model on Higgs Boson Twitter[2] dataset to charaterize how information spread on social platforms like Twitter. The result shows the compartmental models in epidemiology perform well on capturing the diffusion for the event.

1.2. Dataset

We use the Higgs Twitter Dataset, which is built by monitoring the spreading processes on Twitter before, during and after the announcement of the discovery of a new particle with the features of the elusive Higgs boson on 4th July 2012. To be more specific, the dataset includes messages posted in Twitter about this discovery between 1st and 7th July 2012 with at least one of the following keywords or hashtags: lhc, cern, boson and higgs.

Based on the time points of the two major announcements, Domenico *et al.* divide the time frame into four different periods:

1. Before the announcement on 2^{nd} July, there were some rumors about the discovery of a Higgs-like boson at Tevatron;

2. On 2^{nd} July at 1 PM GMT, scientists from CDF and D0 experiments, based at Tevatron, presented results indicating that the Higgs particle should have a mass between 115 and 135 GeV/c^2 (corresponding to about 123 to 144 times the mass of the proton);

3. After 2^{nd} July and before 4^{th} of July there were many rumors about the Higgs boson discovery at LHC;

4. The main event was the announcement on 4^{th} July at 8 AM GMT by the scientists from the ATLAS and CMS experiments, based at CERN, presenting results indicating the existence of a new particle, compatible with the Higgs boson, with mass around 125 GeV/c^2. After 4^{th} July, popular media covered the event.

The final amount of tweets in this dataset is 985,590. The corresponding social network is consist of 456,631 nodes and 14,855,875 directed edges, with nodes corresponding to the authors of the tweets as well as edges

[1] Our report is reproduction of Jin *et al.*'s paper.

[2] Please refer to Github: https://github.com/YunranChen/HiggsBoson.git

represent the relationships between follower and followee/retweet/reply/mention relationships. The retweet network consists 256,491 nodes and 328,132 edges. Our analysis is mainly based on retweet network.

2. Models

2.1. SI model

The simplest model, SI, divides population into two compartments: susceptible (S) and infected (I). Once an individual is infected, he or she would be infected forever. In our setting, SI model divides the Twitter user into two classes: At any given time period t, N denotes the total population size, $I(t)$ is the size of the population that has retweeted about the topic of interest, while $S(t)$ is the remaining population size. Here $N = I(t) + S(t)$. Note, for this model, I is absorpting state. This means all the people involved would be infected finally. So this model is suitable for fitting the network within which users retweeted at least once.

Mathematically, the SI model can be represented by the following system of ordinary differential equations:

$$\frac{d[S]}{dt} = -\frac{\beta SI}{N}$$
$$\frac{d[I]}{dt} = \frac{\beta SI}{N}$$

Here, β is contact rate, it takes into account the probability of getting the disease in a contact between a susceptible and an infectious individual.

2.2. SIS model

Similar to the SI model, the SIS model divides the Twitter user into exactly the same classes. Compared with SI model, individuals in I state can transfer back to S state with transition rate α (called recovery rate). This difference can be illustrated clearly by Figure 1 and Figure 2. This model is suitable for fitting the total friendship network of the users who has retweeted instead of the purely retweeted network.

Mathematically, the SIS model can be represented by the following system of ordinary differential equations:

$$\frac{d[S]}{dt} = -\frac{\beta SI}{N} + \alpha I$$
$$\frac{d[I]}{dt} = \frac{\beta SI}{N} - \alpha I$$

2.3. SEIZ model

SEIZ model is proposed by Bettencourt *et al.*[1] The model is developed from epidemiological models but for capturing the spread of ideas. It incorporate four different states to represent different reaction of people to an idea. Here, susceptible(S) represents the users who have not heard about the news yet; Infected(I) denotes the users who have tweeted about the news; Skeptic(Z) denotes users who have heard about the news but chooses not to tweet about it; Exposed (E) represents users who have received the news via a tweet but has taken some time(i.e. exposure delay) to post.SEIZ model is an improved version of the SIS model for that not all user will retweet at the moment they are exposed to the news.

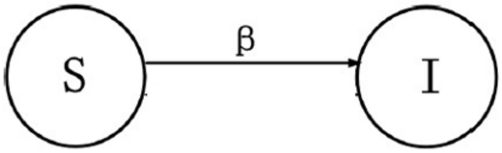

Figure 1. SI model

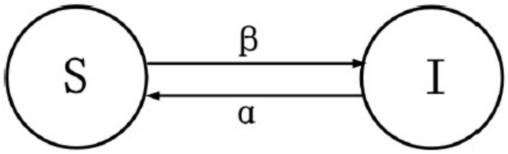

Figure 2. SIS model

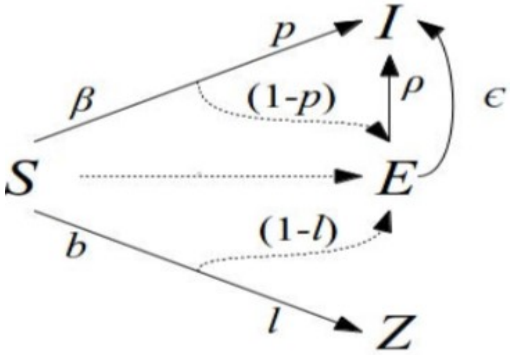

Figure 3. SEIZ model

The SEIZ model is mathematically represented by the following system of ODEs:

$$\frac{d[S]}{dt} = -\beta S\frac{I}{N} - bS\frac{Z}{N}$$
$$\frac{d[E]}{dt} = (1-p)\beta S\frac{I}{N} + (1-l)bS\frac{Z}{N} - \rho E\frac{I}{N} - \epsilon E$$
$$\frac{d[I]}{dt} = p\beta S\frac{I}{N} + \rho E\frac{I}{N} + \epsilon E$$
$$\frac{d[Z]}{dt} = lbS\frac{Z}{N}$$

Parameter	Definition
β	S-I contact rate
b	S-Z contact rate
ρ	E-I contact rate
ϵ	Incubation rate
$1/\epsilon$	Average Incubation Time
bl	Effective rate of S -> Z
$\beta\rho$	Effective rate of S -> I
b(1-l)	Effective rate of S -> E via contact with Z
$\beta(1-p)$	Effective rate of S -> E via contact with I
l	S->Z Probability given contact with skeptics
1-l	S->E Probability given contact with skeptics
p	S->I Probability given contact with adopters
1-p	S->E Probability given contact with adopters

2.4. Parameter Estimation

We applied nonlinear least squares fit (based on Levenberg-Marquardt algorithm) to related network data. For SI model and SIS model, we minimize $(I(t)-\hat{I}(t))^2+(S(t)-\hat{S}(t))^2$ to get optimal parameter sets. For SEIZ model, we minimize $(I(t)-\hat{I}(t))^2+(E(t)-\hat{E}(t))^2+(S(t)-\hat{S}(t))^2+(Z(t)-\hat{Z}(t))^2$ instead. In addition, the ODE systems were solved with a forward Euler function.

3. Results

3.1. EDA

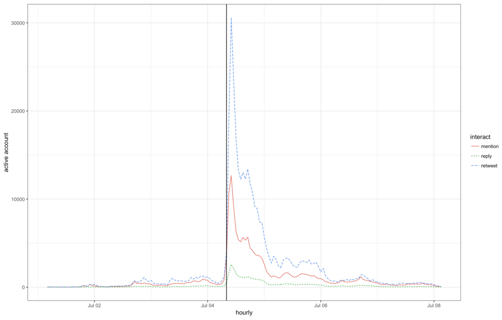

Figure 4. Number of active accounts per hour

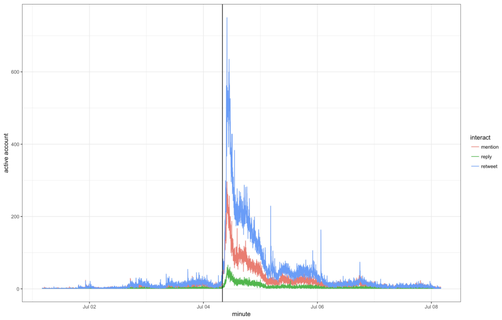

Figure 5. Number of Active accounts per minute

As is shown in Figure 4 and Figure 5, three types of activities all shows the similar pattern. However, retweet network is more fluctuated compared to mention and reply network. This corresponds to the results of related scientific studies. Since most people are not familiar with the research area, they would prefer to retweet rather than write something by themselves. Therefore, here we would focus on the retweet network to study the spread of information.

After the announcement on 4^{th} July at 8 AM GMT, the number of active accounts experienced a sharp linear increase, reaching the peak in 1 hour and 55 minutes, with the number of active users (per 15 min) 44 times the number of active users at the same time period on 3^{th} July. The increase is followed by an exponential decay, the number of active accounts went back to normal on 7^{th} July. 256491 people in total were involved in the retweet network. Epidemiological models will be used to analylze this later.

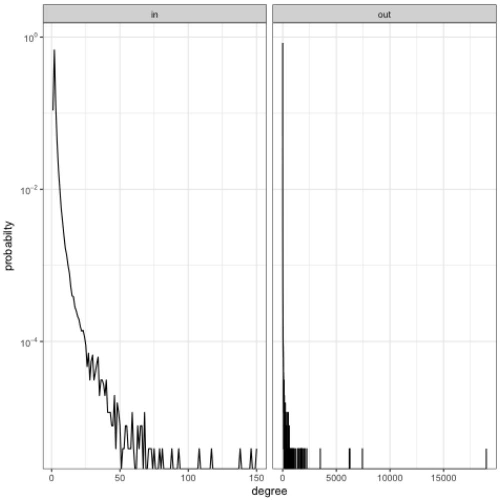

Figure 7. Degree Distribution for Retweet Network

In Figure 7 we show the distributions of the in-degree, out-degree and of the users that retweeted about the Higgs boson. The in-degree and out-degree distribution looks like power-law distribution.

Notice we define the direction of the edge according to the direction for how information spread. The scale of out-degree is much more larger than in-degree. Since normally very few users would retweet multiple twitter while a twitter come from influential user may be retweeted much. Because very few users have more than thousands followers, the distribution of the out-degree become relatively sparse when larger than 1000.

In Figure 8 we show the number of the edges, 2-in-stars, 2-out-stars and triangles. Comparing them to Figure 4 and Figure 5 we can see that they share similar trend, although their magnitude is quite different. The in-2-stars are much

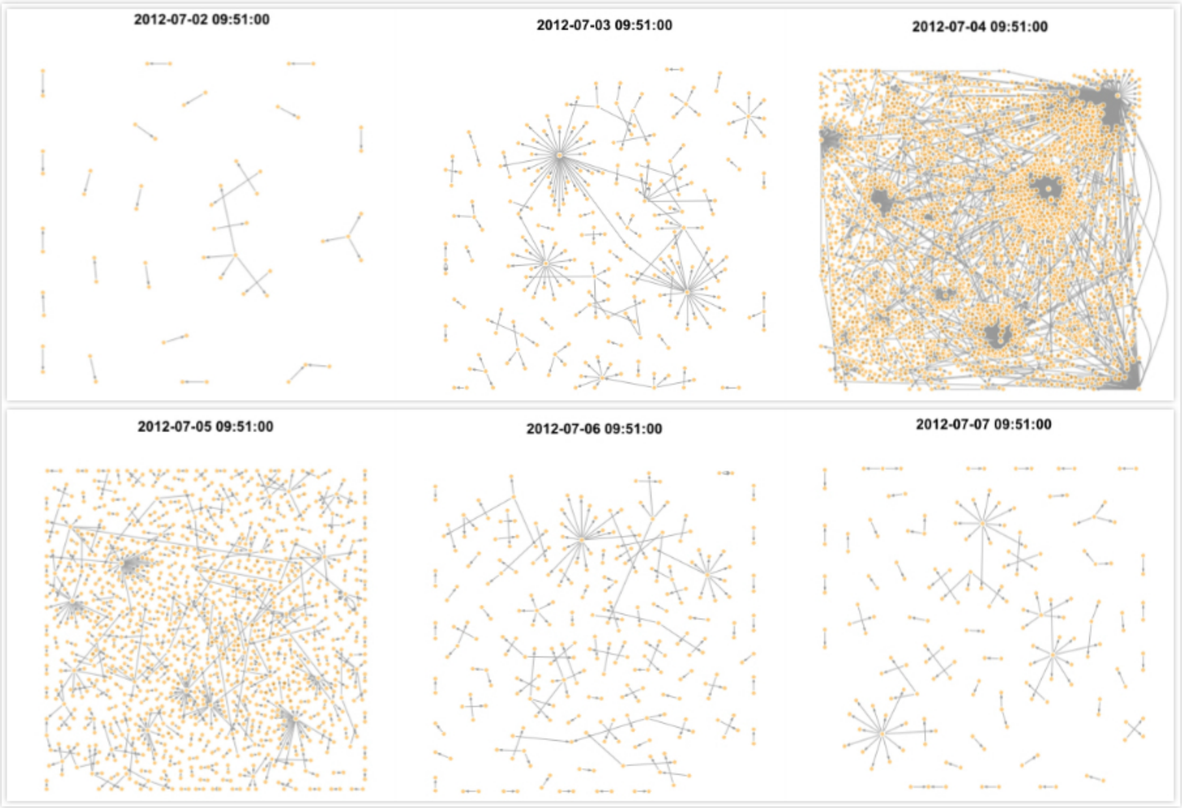

Figure 6. Network Structure Evolution

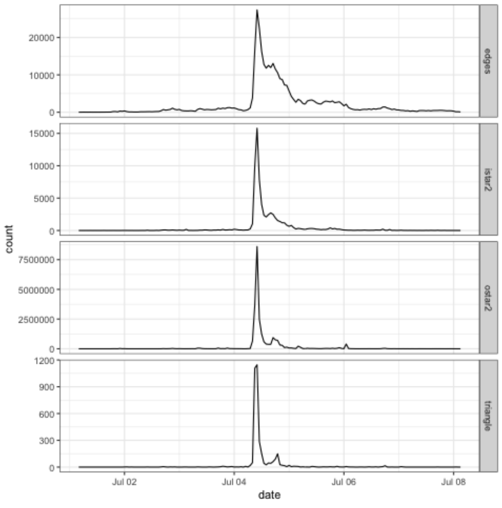

Figure 8. Number of Edges, 2-in-star, 2-out-star and Triangles

smaller than out-2-stars, which makes sense since one tweet are likely to be retweeted by many users but one user is unlikely to retweet hundreds of times. It is worth noting that number of edges, unlike other statistics, does not experience sharply drop from the peak. This is due to a large quantities of ”node-to-node” pattern after the drop out of influential users. See the Figure 6 for more details.

We get a snapshot of the structure of the retweet network at $09:51-10:06$ on each day from 2^{nd} of July to 7^{th} of July. Notice that the third network on $4th$ July is only based on 40% of data. We can find several interesting things as follows:

1. The attendance and dropout of several large influential hubs may account for the sharp increase and decay of retweets. In reality, these nodes are likely to be mass media. The degree can be considered as influence of the individual(and thus a valuable resource). In this way Figure 7 accords with the 80/20 rule.

2. After the dropout of large influential hubs, the network is mainly consist of isolated small groups, we call it a ”node to node” pattern. This can be thought of as the small-scale discussion huge news bring to public.

3.2. Modeling

3.2.1 SI Model for Retweet Network

Retweet network is formed by the users who has retweeted at least once. Given that the I state in the SI model is an absorbing state, SI model is suitable for the retweet network.

Based on the assumption of SI model, once a user retweet, he or she would stay in infected state forever. In the end, all the users in the network would get infected.

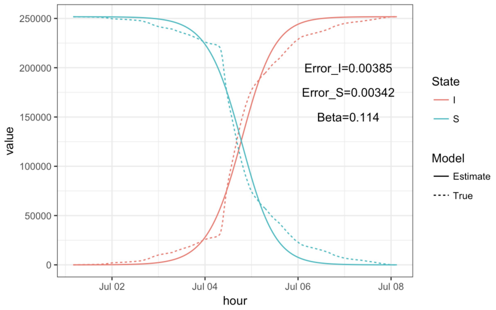

Figure 9. SI Model for Retweet Network

The estimate for the contact rate(β) of the retweet network is 0.114, which means on average one infected user infects 0.114 user in the next time period per hour. Figure 9 indicates that we are doing a good job in terms of the goodness of fit: The relative error is pretty low and the fitted trends are close to the real ones.

3.2.2 SI Model for Friendship Network

In reality, we care more on the characteristic of information spreading on friendship network instead of retweet network. The friendship network is formed by the users who retweet and the followers of these users.

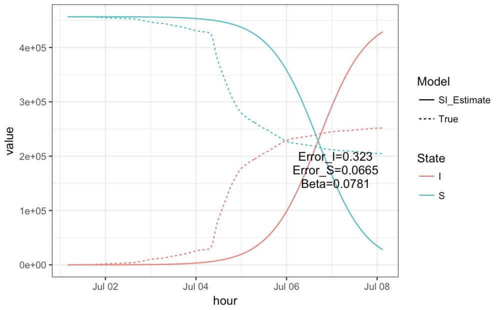

Figure 10. SI Model for the Friendship Network

Here the contact rate(β) for the retweet network is estimated to be 0.0781, which means on average 7.8% of a user's friends will retweet his or her tweets related to Higgs boson at the next time period.

From Figure 10 it is obvious that the SI model is not performing well here: It fails to capture the rapid increase of I and the rapid decrease of S timely. It also overestimate the number of I and the underestimate the number of S when the sizes become stable. This is because the model assumption for SI model is not suitable for the data. We assume I here to be an absorbing state so eventually all users should be infected but in reality this is not the case. A large quantity of users were not infected.

The failure of the SI model on the friendship network leads us to think about more sophisticated networks such as SIS and SEIZ.

3.2.3 SIS Model for Friendship Network

Given the fact that there is always a fraction of users who don't retweet tweets about Higgs boson, SIS is more appropriate. SIS model consider the I as transient state instead of absorbing state. Users who get infected can go back to susceptible state.

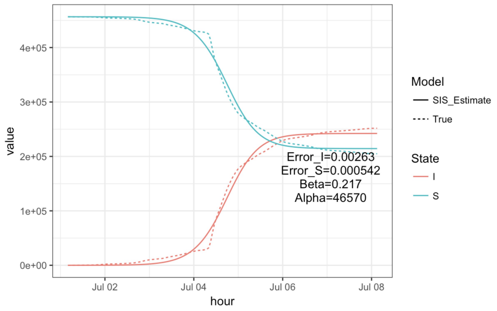

Figure 11. SIS Model for Friendship Network

The ODEs for SIS have two parameters: α and β. In epidemic settings α is considered as the recover rate – how fast do the patients recover from a disease, whereas β is considered as the contact rate – on average how many people a single patient infects in each time period. Here the estimates for α and β are 46579 and 0.217. We can interpret α as the increment of deactivated users and β as on average how many followers will retweet a certain user's tweets related to Higgs boson. Here the α and β are relatively high, showing that the news would spread fast among people on twitter, but people's interest on the news fades quickly too.

From Figure 11 we can see that the SIS model performed well: The overall trends match and the relative errors are small. However, SIS model is less interpretable. Because it is hard to correspond the transition from the infected state to susceptible state to the reality, we tend to believe there exists some users who never retweet or do not retweet as soon as exposed to the news. SEIZ model is alternative choice for better interpretation.

3.2.4 SEIZ Model for Friendship Network

SEIZ model is proposed by Bettencourt *et al.*[1] and aims at capturing the spread of ideas. We consider SEIZ model to be more suitable here. SEIZ model added two more states, skeptic(Z) and exposed (E). Skeptic(Z) denotes users who have heard about the news but chooses not to tweet about it; Exposed (E) represents users who have received the news via a tweet but has taken some time to react. With susceptible (S) and infected (I), these four state represents four different attitudes toward news.

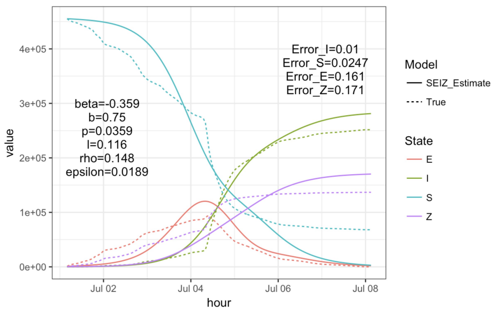

Figure 12. SEIZ Model for Friendship Network

Figure 12 shows the performance of the SEIZ model. We can see that it has similar performance compared to the SIS model in terms of relative error and fitted trend. The estimates for the parameters are: $\beta = -0.359$, $b = 0.750$, $p = 0.0359$, $l = 0.116$, $\rho = 0.148$ and $\epsilon = 0.0188$. Here we get a negative β, which is kind of weird. A reasonable explanation is that many users remained skeptical about the rumors before the major announcement on the 4^{th} of July; But after the official announcement the size of S experienced a steep decrease since the authenticity of the news has been confirmed. From the differential equation, $\frac{d[S]}{dt}$ needs to be negative and relatively large in terms of absolute value. However, note that the before $4^t h$ of July the decrease in S is relatively small, and β is forced negative to drag the derivative back during this period.

It is worth mentioning that the decrease in E, the steep increase in I almost happen at the same time, shortly after the major announcement. This suggests that most users are alert and don't want to tweet about rumors unless they are proven true. We can also look at the ratio $\frac{\epsilon}{\rho} \approx 0.127$, which suggests that the exposed users(E) became infected more so due to direct contact with the infected users(I) and not so much from information incubation and self-adoption. This somewhat suggest users retweet tweets related to Higgs boson mainly because their friends(the people they follow) have tweeted about it.

Besides this, we can construct some index to capture different aspect of the evolving network based on the parameters in the compartmental models. In Jin *et al.*'s paper[3], they defined $R_{SI} = \frac{(1-p)\beta+(1-l)b}{\rho+\epsilon}$ to distinguish news and rumor. Here $R_{SI} \approx 1.900$ is relatively large, we can conclude that this is a true news instead of rumor, which corresponds to the reality.

4. Conclusion

In this report, we visualize how the retweet network evolve before, during and after the announcement of breaking news. Then we applied compartmental models in epidemiology to capture the characteristics of the evolving network, which is based on Jin *et al.*'s paper. We have three mainly interesting finding:

1. The peak and decay in information spread largely rely on the attendance and dropout of a few large influential users (official media). As these influential users dropout, they would leave a large quantity of "node-to-node" patterns, which can be considered as a broad local discussion to the public.

2. The information spread on Twitter can be accurately captured by compartmental models in epidemiology. The speed of information spreading is relatively high (contact rate is relatively high). Meanwhile the fading speed of information spreading is relatively high (recovery rate is relatively high).

3. The SEIZ model are more interpretable compared to traditional compartmental models. A large quantity of users are skeptical to the news. The outbreak of twitters is partly due to the official announcement resolved the skeptical views, and thus transferring a large part of exposed or skeptical users to infected users.

5. Discussion

1. All the compartmental models in epidemiology smooth the process, leading to bad performance on capturing some important changing points. A weighted version that emphasize the fitting of the changing points or adding some noise term would be better, if we focus on this aspect.

2. The shape of the true data are different from that of the fitted value. For example, for the SI model for retweet network, the true data for $I(t)$ is convex while the fitted data is concave. Some other ODE equations should be considered to improve the performance.

3. Under SI, SIS and SEIZ model, one basic assumption is that the population at each time point is constant. This assumption is pretty strict and not suitable in reality, a more flexible version with increasing $N(t)$ should be proposed.

References

[1] L. M. Bettencourt, A. Cintrón-Arias, D. I. Kaiser, and C. Castillo-Chávez. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models, 2006.

[2] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi. The anatomy of a scientific rumor, 2013.

[3] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan. Epidemiological modeling of news and rumors on twitter, 2013.

[4] M. E. Newman. Spread of epidemic disease on networks, 2002.

[5] L. Zhao, H. Cui, X. Qiu, X. Wang, and J. Wang. Sir rumor spreading model in the new media age. *Physica A: Statistical Mechanics and its Applications*, 392(4):995–1003, 2013.