Information Epidemics: Outbreak and Decay of Viral Videos

By Melissa Flores and Stephen Slater

May 8, 2014

Abstract

How does the spread of a viral YouTube video differ from the spread of an epidemic? How might one model a viral video using the SIR model? The spread of viral videos and the spread of infections share many common traits. In a fixed population with different subsets of susceptible people, infected people, and immune people, one can model the flow of people from group to group (as the infection grows and dies) with a system of differential equations. Likewise, one can model the views/day of a viral video using a system of differential equations. However, the situations are a bit different: the decay pattern of a viral video is much slower than that of an infection, and hence the SIR is not a good fit for the video views/day; furthermore, people can become "infected" without any physical contact, and any individual can contribute to the views/day "infected" group more than once.

In this paper we present three different models of these flow equations and seek to determine the best fit for the spread of views/day of a viral video. The most prominent feature of the viral video views/day pattern is that the decay in popularity does not approach zero in the near future, unlike an infection, which rapidly decays to zero in a typical SIR model. Instead, the views/day follow a slower decay with an exponential factor that brings views/day to a rough baseline popularity level. We model the views/day of three popular YouTube videos that have gone viral in recent years, and find that our Model 3, the SEIR model that includes an exponential decay term, serves as the most accurate model in fitting the outbreak and decay of the views/day of viral YouTube videos.

1 Introduction

The basic SIR model we studied in class inspired us to expand the model's application and try to explain the outbreak and decay pattern of viral YouTube videos. The basic SIR model includes three distinct populations in modeling a disease or epidemic: people susceptible to the disease, people infected with the disease, and a removed population who have either been cured or have passed away. These distinct populations and the differential equations we use to model the spread of epidemics are also effective in modeling the spread of information, or more specifically, the viewing audience of viral YouTube videos. In this study we have expanded the basic SIR model to three different sets of differential equations in order to model the change in views/day of three YouTube videos that have recently gone viral. Instead of modeling the number of people in each group, we have determined population I to be the views/day of a YouTube video, and hence the other populations must be in terms of views/day as well such that we can have a constant N equal to the total potential views/day of a given video. We assume that this N is equal to S, the number of views/day that may occur for a given video, plus I, the number of views/day that do occur, plus R, the lost views/day from loss of interest.

1.1 Model 1 Background

Our first model is a simple modification of the basic SIR model in that we have categorized the spread of information: the "infection" of viewing the video can spread by word of mouth or by online sharing, and we expect these rates to be different. Specifically, we expect the online sharing rate to have a larger effect on the views/day because an online share can reach more viewers than an in-person share. Mathematically, however, one can add the talking and online sharing constants in this model to get the basic SIR model. With this basic set of differential equations, we model the populations of views/day as if they were populations of people.

1.2 Model 2 Background

Our second set of differential equations accounts for aspects of the spread of information that differ from the spread of disease. We do this with another extension to the SIR model that we briefly learned in our study of epidemiology: we included populations that are at high risk of contracting and spreading the infection. We have modelled the second set of differential equations after the idea that there is a highly connected population of Internet users and a population with an average number of connections. This is based on the data that the median number of Facebook friends that adult Facebook users have is 200, with an average of 338 (Smith). The fact that the average is so much larger than the median implies that there are some users that are much more connected online; for example, 15% of adult Facebook users have over 500 Facebook friends (Smith). We account for this by using different constants for the interactions between the highly connected susceptible, highly connected infected, regular susceptible, and regular infected populations.

1.3 Model 3 Background

Finally, our third model is an expansion of SIR that includes an "exposed" population E as well as an exponential decay term in the derivative of population I. The authors of one paper we studied note that viral YouTube videos either decline steadily over time, or decline rapidly at first until they reach a somewhat low and constant level (Broxton, Interian, Vaver, & Wattenhofer, 2010). This is very different from the models we have studied of infections, which all decay somewhat rapidly to zero over time.

In Brauer's paper that reviews extensions to the SIR model, he discusses the usefulness of the SEIR model in modeling diseases with this "exposed" population. Brauer notes that the analysis of this model is the same as the analysis of the basic SIR model, but with I replaced by E + I. That is, instead of using the number of infected people as one of the variables we use the total number of infected members, whether they are simply exposed to the disease by viewing it or if they are also capable of transmitting infection via sharing. Brauer notes that this model is useful because "some diseases have an asymptomatic stage in which there is some infectivity rather than an exposed period". Thus, for our third set of differential equations we have utilized the SEIR model to include a population who are exposed to the viral YouTube video, but have not yet begun sharing the video and "infecting" others. This is due to the fact, as Broxton et al. conclude, that "these videos demonstrate the power of sharing, and its role in shaping video viewing habits" (Broxton et al., 2010). With this set of differential equations, we have made the assumption that there is a period of time between when people become "infected" by viewing the video, and when people begin sharing the video and "infecting" others.

Using these three sets of differential equations, which we designed to fit the circumstances described above, we proceeded to model the views per day of three different viral YouYube videos:

Gangnam Style by PSY, Call Me Maybe by Carly Rae Jepsen, and The Fox (What Does the Fox Say?) by Ylvis.

2 Methods

2.1 Model 1: SIR with different ways of sharing (talk and online)

This is similar to the original SIR model except that we have categorized the spreading of the "infection" between in-person sharing and online sharing. Here is the system of differential equations:

Figure 1: Model 1 equations

$\beta = spread\ by\ talking$

$\alpha = spread\ by\ sharing\ online$

$\gamma = loss\ of\ relevance$

$N = total = S + I + R$

$S = haven't\ seen\ the\ video\ /\ day$

$I = views\ /\ day$

$R = lost\ views\ /\ day$

$$\frac{dS}{dt} = -\beta SI - \alpha SI$$

$$\frac{dI}{dt} = \beta SI + \alpha SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

$Units$

$$\frac{dS}{dt} = \frac{haven't\ seen\ the\ video}{day * day} = \frac{views}{day * day}$$

$$\frac{dI}{dt} = \frac{views}{day * day}$$

$$\frac{dR}{dt} = \frac{lost\ views}{day * day} = \frac{views}{day * day}$$

$Therefore, because\ \frac{dS}{dt} = \frac{views}{day * day},$

$$\frac{views}{day * day} = \beta(\frac{views}{day})(\frac{views}{day}) - \alpha(\frac{views}{day})(\frac{views}{day})$$

$$\beta\ units = \frac{1}{views}$$

$$\alpha\ units = \frac{1}{views}.$$

$Because\ \frac{dR}{dt} = \frac{views}{day * day},$

$$\frac{views}{day * day} = \gamma(\frac{views}{day})$$

$$\gamma\ units = \frac{1}{day}$$

2.2 Model 2: Normal and highly-connected groups

This model contains different-risk populations for being infected. This is because some people are more socially connected than others are (online, for example, the highest-connected group has more friends on social media than the lowest-connected group), so there are different constants of sharing

online depending on the populations that are interacting such that

$$\alpha_1 < \alpha_2 < \alpha_3.$$

.
Hence, the Model 2 equations show dS/dt, dI/dt, dR/dt, dSh/dt, dIh/dt, and dRh/dt:

Figure 2: Model 2 equations

β = *spread by talking*

α_1 = *spread by sharing online between 2 low risk groups*

α_2 = *spread by sharing online between one high and one low risk group*

α_3 = *spread by sharing online between 2 high risk groups*

γ = *loss of relevance*

$N = total = S + I + R$

S = *haven't seen the video / day*

I = *views / day*

R = *lost views / day*

$$\frac{dS}{dt} = -S(\beta I + \alpha_1 I + \alpha_2 I_H)$$
$$\frac{dI}{dt} = S(\beta I + \alpha_1 I + \alpha_2 I_H) - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$
$$\frac{dS_H}{dt} = -S_H(\beta I + \alpha_2 I + \alpha_3 I_H)$$
$$\frac{dI_H}{dt} = S_H(\beta I + \alpha_2 I + \alpha_3 I_H) - \gamma I_H$$
$$\frac{dR_H}{dt} = \gamma I_H$$

2.3 Model 3: SEIR with exponential decay in dI/dt

This model contains 4 groups that make up the total potential views/day (analogous to the total population in the SIR model for infections). This model expands the SIR version to include an exposed population, which in this case is made of people who have seen the video today, but are not yet exposing others to the video. Our model assumes that people in the infected population share the video rate with rates alpha and beta, but now the views/day flow from susceptible to exposed before moving to the infected population (which shares the video) with rate mu. Another feature of our SEIR model is an exponential decay factor. This ensures that as time increases,

the second term in the differential equation for I reduces to zero. We see this in the graphs below when the slope of population I becomes less and less negative. This slower exponential decay is a key difference between modeling diseases (as with the SIR model) and viral videos: the decay of infections in the SIR model (and in the shares/day of a YouTube video; more on this later) reduces to zero much more quickly, whereas the views/day of a popular video decay slowly and appear to level off to a baseline popularity level, at least in the near future.

Figure 3: Model 3 equations

$\beta = spread\ by\ talking$

$\alpha = spread\ by\ sharing\ online$

$\mu = \text{exp}osed\ person\ begins\ to\ share$

$\gamma = loss\ of\ relevance$

$N = total\ potential\ views\ /\ day = S + E + I + R$

$S = haven't\ seen\ the\ video\ /\ day$

$E = views\ /\ day\ (\text{exp}osed, but\ not\ sharing\ with\ others)$

$I = views\ /\ day\ (sharing\ video\ with\ others)$

$R = lost\ views\ /\ day$

$$\frac{dS}{dt} = -\beta SI - \alpha SI$$

$$\frac{dE}{dt} = \beta SI + \alpha SI - \mu E$$

$$\frac{dI}{dt} = \mu E - Ie^{-\gamma t}$$

$$\frac{dR}{dt} = Ie^{-\gamma t}$$

Units

$$\frac{dE}{dt} = \frac{views\ from\ those\ \text{exp}osed\ but\ not\ sharing\ yet}{day * day}$$

$$\frac{dE}{dt} = \frac{views}{day * day}$$

$$\frac{dE}{dt} = \beta SI + \alpha SI - \mu E$$

$$\frac{views}{day * day} = (\frac{1}{views})(\frac{views}{day})(\frac{views}{day}) + (\frac{1}{views})(\frac{views}{day})(\frac{views}{day}) - \mu(\frac{views}{day})$$

$$\mu\ units = \frac{1}{day}$$

2.4 Methodology

YouTube displays cumulative and daily statistics below each video (views, subscriptions, and shares). In order to model our differential equations against these video statistics (specifically the views/day graphs), we needed to acquire the data from YouTube. However, this data is not publicly available. To solve this problem, we took screenshots of the graphs of views/day in YouTube and analyzed the images as matrices in Python (see youtube.py uploaded to the course website). In doing so we scanned down each column of the picture, found the pixel that corresponds to the colorful datapoint in the graph (see example below), and stored the height of that pixel in that image. With a list of these points, we can recreate the views/day data from YouTube in a graph in which we also plot our expansions of the SIR model. We have decided not to transform the pixel height of the image to match the numbers from the YouTube graphs to minimize error from estimating the right transformation based on the imprecise YouTube graph.

For example, one of the viral videos we model in this paper is Gangnam Style by PSY. These are the steps we took to fit our model to the YouTube data:

1. Find the video statistics below the video on YouTube.
2. Crop image of views/day graph.
3. Use our program youtube.py to recreate the YouTube image and graph it in the same window as the integrated system of differential equations. Adjust the constants in the model to fit the "outbreak" of the viral video.

Below is an example of this methodology using Gangnam Style by PSY and our Model 3. In Figure 6 we fit Gangnam Style views/day (populations E + I in our 3rd model). We have expressed "population" (views/day) on the y-axis in terms of the pixel height of the image. To get the real number of views/day would require a transformation (such as multiplying the y-values in the graph to be in the range of 0 to 15,000,000) as they are in the YouTube daily views plot shown in Step 1.

Figure 4: Gangnam Style statistics.

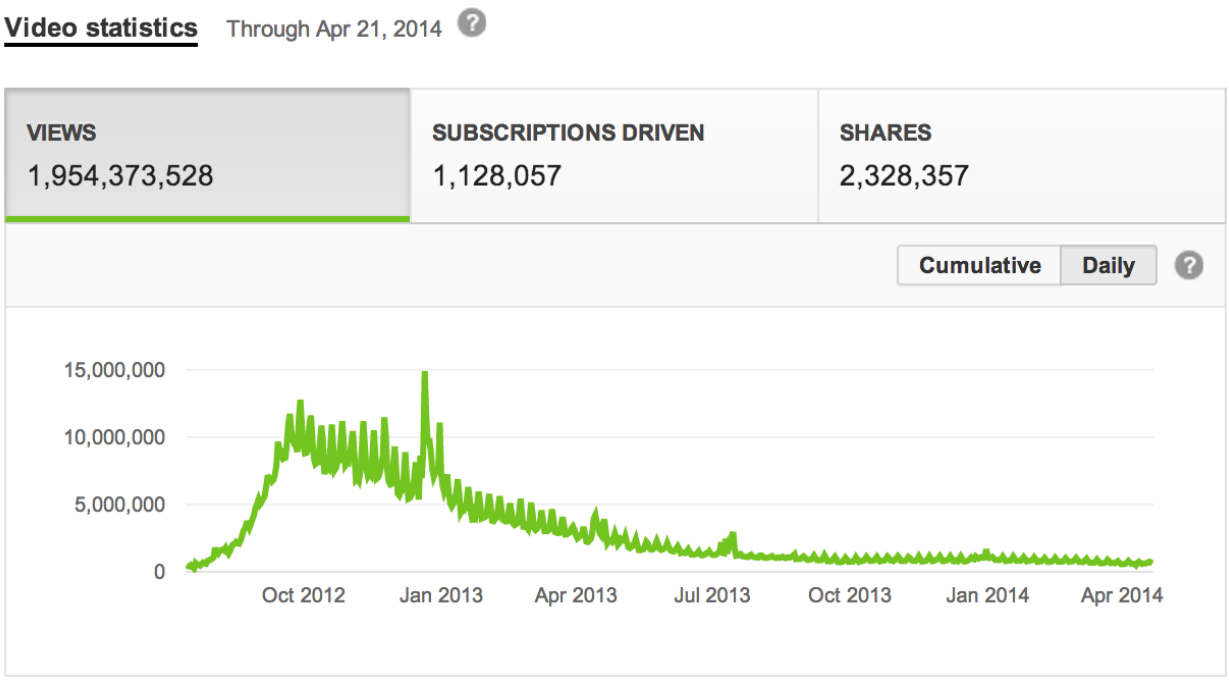
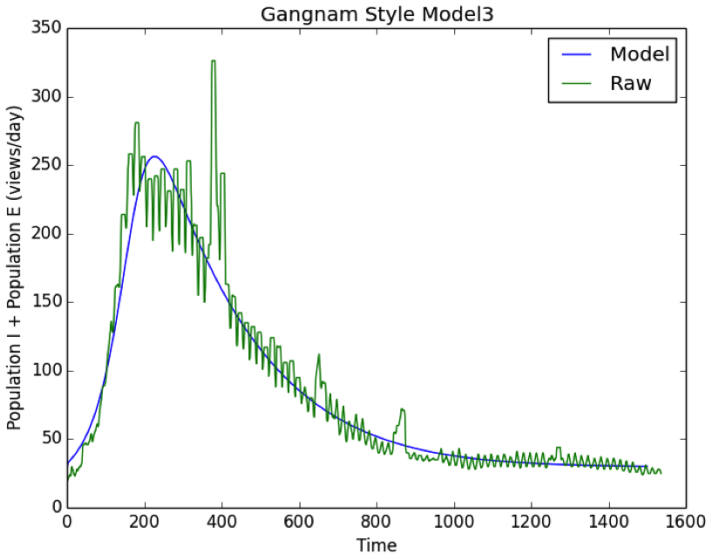


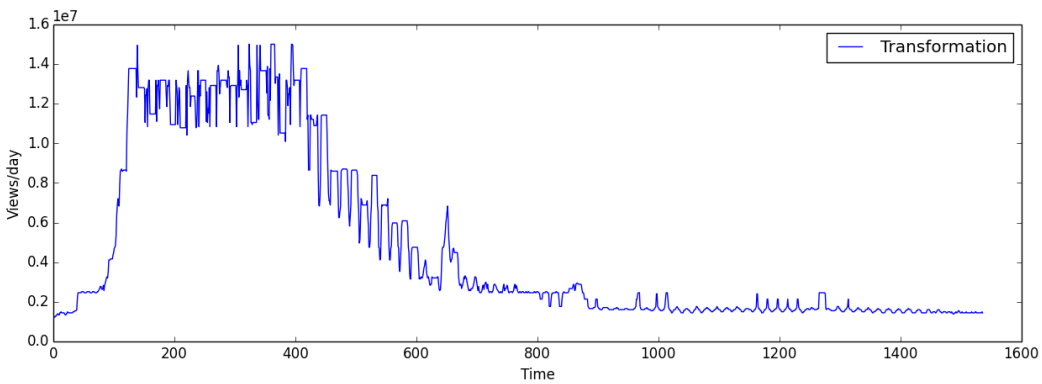
Figure 5: Screenshot of the views/day graph by itself.

Figure 6: Gangnam views/day fit with Model 3



We have chosen to fit our model against the pixel height of the image (related to the views/day on the YouTube graph in Step 1) because a transformation of the pixel values leads to some uncertainty and disclarity. For example, below is a transformation of the pixel height based on the maximum height of the colored pixel array ("raw" in the python file) and the maximum height, 15,000,000 views/day, in the YouTube graph in Step 1. It is evident that the pixel-height images we have chosen to use are cleaner than the transformation:

Figure 7: Transformation of Gangnam Style views/day to match YouTube data.



3 Results & Discussion

3.1 Gangnam Style

The following 3 graphs are fits of the Gangnam Style views/day data using Models 1, 2, and 3, respectively. See Appendix for values of constants for each fit.

Figure 8: Gangnam views/day fit with Model 1

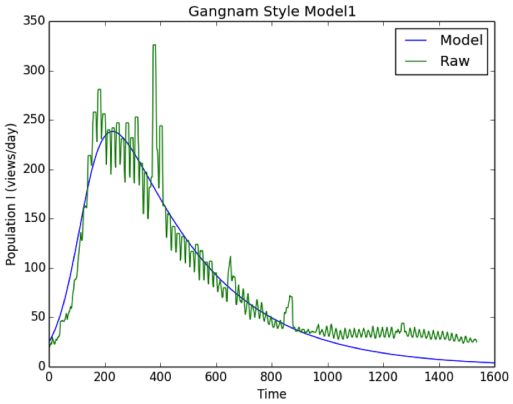


Figure 9: Gangnam views/day fit with Model 2

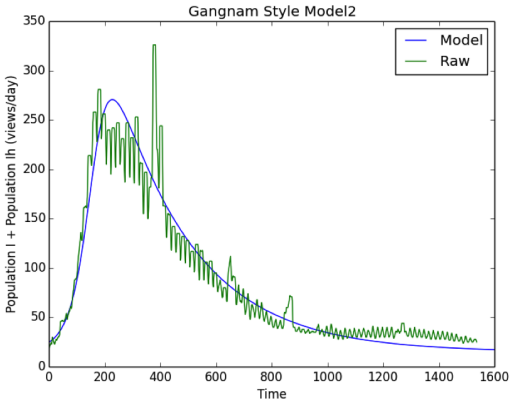
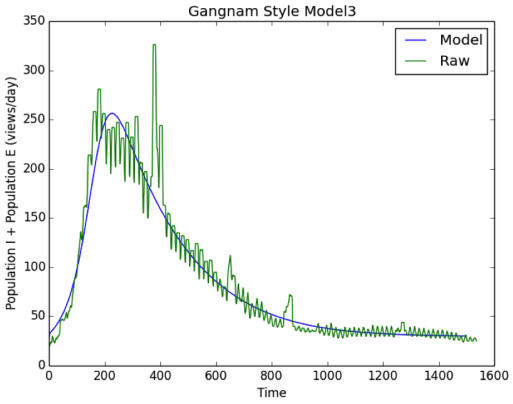


Figure 10: Gangnam views/day fit with Model 3



3.2 Call Me Maybe

The following 3 graphs are fits of the Call Me Maybe views/day data using Models 1, 2, and 3, respectively. See Appendix for values of constants for each fit.

Figure 11: Call Me Maybe views/day fit with Model 1

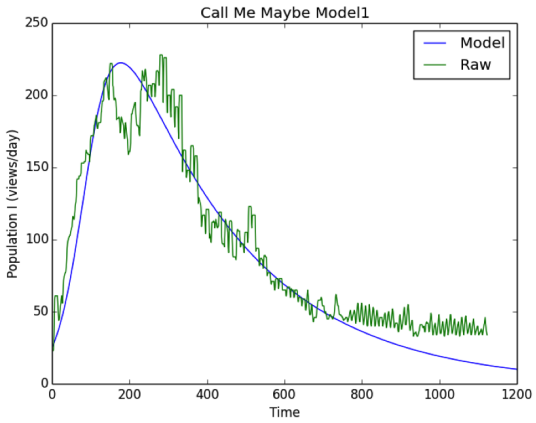


Figure 12: Call Me Maybe views/day fit with Model 2

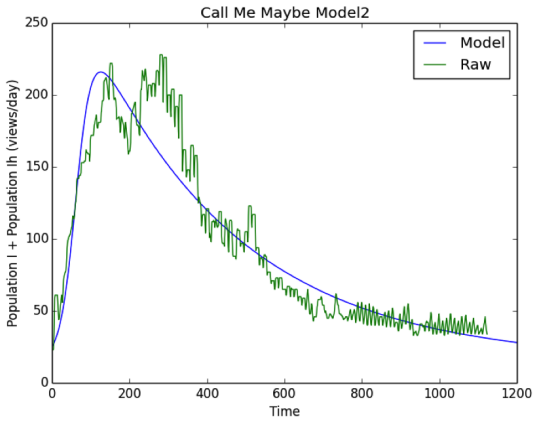
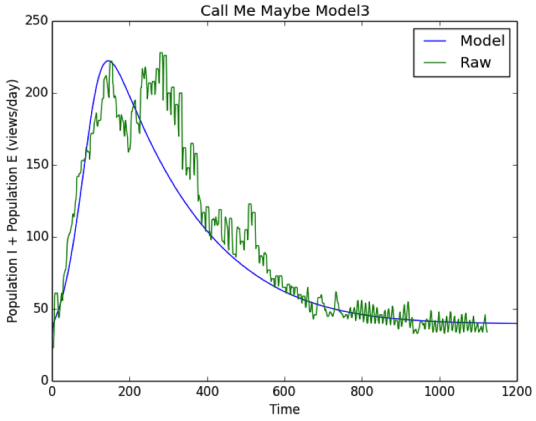


Figure 13: Call Me Maybe views/day fit with Model 3



3.3 The Fox (What Does The Fox Say?)

The following 3 graphs are fits of The Fox (What Does The Fox Say?) views/day data using Models 1, 2, and 3, respectively. See Appendix for values of constants for each fit.

Figure 14: The Fox views/day fit with Model 1

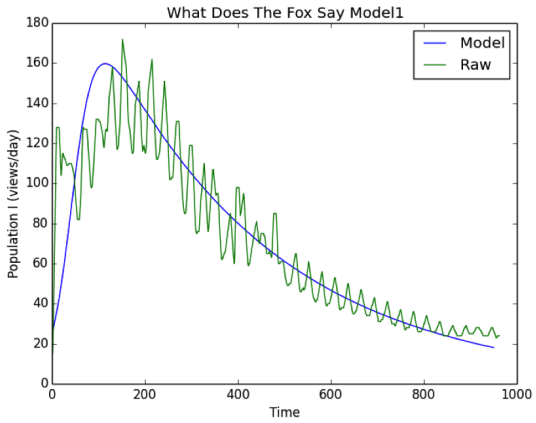


Figure 15: The Fox views/day fit with Model 2

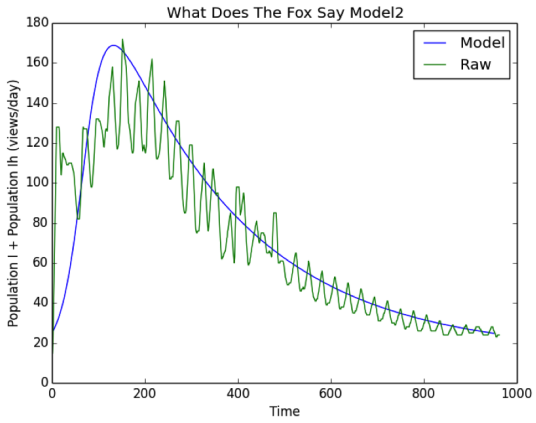
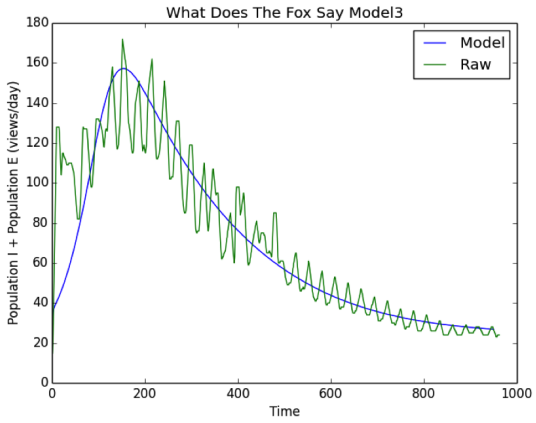


Figure 16: The Fox views/day fit with Model 3



3.4 Discussion

The simplest extension of the SIR model (Model 1 above) is a reasonable fit for modeling the views/day of viral YouTube videos, but it could be better. Our model captures the growth patterns and maximum peak of the actual data, but does not properly capture the decay pattern. In Model 1, though, we can solve for the starting value of S needed such that views/day initially grow, and hence that the viral video will go viral. One can calculate this value of S, as well as the value of S at the peak of I, by setting the differential equation for population I to zero:

$$\frac{dI}{dt} = \beta SI + \alpha SI - \gamma I$$

$$0 = I(\beta S + aS - \gamma)$$

$$I = 0$$

$$\beta S + \alpha S - \gamma = 0$$

$$S(\beta + \alpha) = \gamma$$

$$S = \frac{\gamma}{\beta + a}. \textit{ This is S at peak value of I}$$

Therefore, for I to grow,

$$S > \frac{\gamma}{\beta + a}.$$

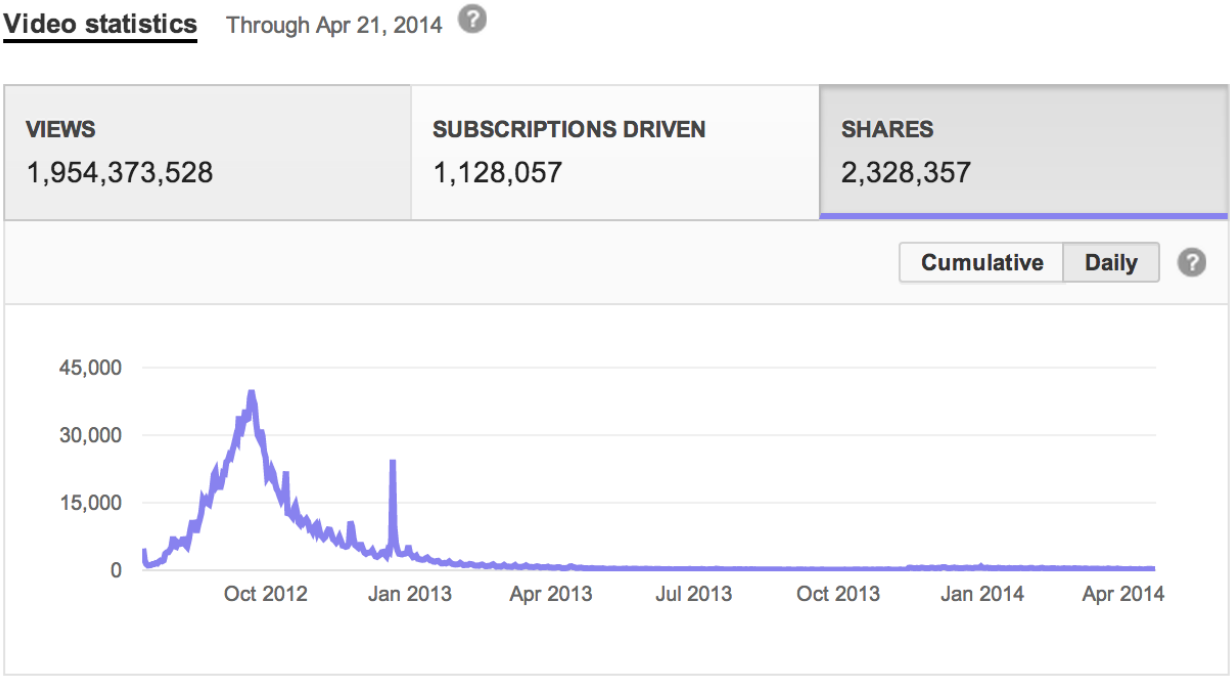
The next extension of the SIR model (Model 2 above) is a set of 6 differential equations that include highly connected susceptible, infected, and removed populations. It is a better model for the spread of viral YouTube videos than the basic SIR model, but still does not fit the data perfectly. It captures the growth patterns, the maximum peak, and better captures the decay than the first model; however, it does not exactly follow the decay pattern, as seen in the 3 graphs above.

The final extension to the SIR model (Model 3 above) is an SEIR model with an exponentially decaying constant to represent people losing interest in the viral YouTube videos. It is the best fit of the data of views/day and almost perfectly depicts the growth, peak, and decay of the infected population (E+I).

There is a notable second peak in views/day in each of the videos above. In Gangnam Style and Call Me Maybe, our models tried to match the first peak while fitting the rest of the data, and in The Fox (What Does The Fox Say?) our model matched the second peak. The fact that these views/day graphs have two peaks is a result of seasonality. In Figure 17 below, there is a noticeable spike in shares/day at the end of the 2012 calendar year. This likely comes from end-of-year sharing, such as "Best of 2012" posts on social media. Hence, from the spike in shares/day we expect the second peak in views/day for each the YouTube views/day graphs above. It is also notable that the shares/day follow a faster decay similar to the SIR model for infections. The intuition here is that a person share the video only once, but that share will reach several other people (increase in views/day), and the model does not explicitly account for how many times a person will watch the

video after receiving the share (example: person A shares the video today, and person B watches it three times today and 2 times tomorrow). As a result, the views/day graphs have a slower decay in popularity than the decay in shares/day:

Figure 17: Gangnam Style shares/day



3.5 Limitations

The first area of limitation in our model and results is the fundamental assumption that SIR models will be just as effective in modeling the spread of information as they are in modeling infectious diseases. Similar differential equations have been shown in the past to have good predictive power in modeling diseases, and we are now trying to apply them to another nonidentical situation. By doing this, we are assuming that the spread of viral videos has the same underlying mechanism as the spread of disease, and that there is total fixed population N of potential views/day (not seen, seen, and lost). We have assumed that the susceptible population matches the population of people who have not viewed the video, that the infected population matches the people who have viewed the video, and that the removed population (dead or cured) matches the people who have lost interest in the video. With this analogy, we can model our populations in terms of how they contribute to population I, the total views/day.

We are also only able to determine views/day from the YouTube data, which does not tell us the actual "infected" population (people who have viewed the video). People who watch it more than once cannot be infected more than once. From an advertising standpoint, it may also be useful to predict the total number of viewers/day, but this would require some assumptions about how many views/day come from any given individual (the person could watch the video once or five times that day). However, it probably benefits the advertising company regardless of the number of views/day that come from one viewer; if two people see the video, the company may be more likely to make the sale, but if one person sees the video twice, then that one person may be more likely to buy the product.

Additionally, we do not have data to model people losing interest in the video (related to gamma), and we would need to find a way to collect this data in the future. We also cannot account for people losing interest in the video and regaining interest again (moving from removed to infected), which is a possibility we may need to consider with YouTube video views.

The second area of limitation in our model and our results is the quality of fit of our model, and how to measure this fit. As seen in the graphs above, our model does not take into account seasonality or spikes during the end of the year. In each of the videos above, it is apparent that there is a second peak in views/day that the model does not capture; this is likely a result of end-of-year online sharing (for example, "Best of 2012"). On YouTube, this spike in views/day at the end of a calendar year coincides with a spike in the shares/day (see graph in Discussion). Furthermore, our study does not account for the size of deviations in the views/data curve (the up and down fluctuations may come from higher views on the weekends, for example). When measuring the spread of infectious disease, there are significantly less daily and even yearly fluctuations to account for. Although we can view these problems in the graph, we are also limited in that we do not have a way to measure goodness of fit of our curves to the data. We are estimating the goodness of fit by inspection but we have not measured any statistical accuracy versus the data itself. This could be measured in the future by measuring the distance of each data point from the predicted data point by the curve.

The last area of limitation in our study is the inability to apply our model to future behavior of viral YouTube videos, which would hopefully be the ultimate goal of this research. In the future, we might find a way to predict constants to fit the model, rather than use more of a trial and error approach. We would have to research data to estimate our parameters: alpha, beta, gamma, and mu. Another possible limitation would be how much our predictive results would vary with parameters, or in other words, how sensitive our future predictions would be to the parameter estimates. This is important because in finding parameters for our second model, which included highly-connected populations, the model was very sensitive, and slight changes in the parameters resulted in large changes in the predictions and our model curve. Unlike in the second model, we found that the third model (SEIR with exponential decay) was much less responsive to changes in the parameters, so Model 3 would likely be more reliable and robust in predicting views/day, given that it is more accurate and stable under a wider range of parameter constants.

4 Conclusion

In this study we found that SIR models can accurately predict the past behavior of the outbreak and decay of viral YouTube videos. We designed three different sets of differential equations, and we found that the SEIR model with an exponential decay constant is the best fit for modeling the views/day of a viral video. As the video loses popularity, it seems to decay to a baseline popularity that is still a positive number of views/day. This is unlike the SIR model of an infection, in which the infected population term I decreases to zero (see Model 1 system of equations and graphs above). Although our models appear to be very accurate, there could be further work in implementing a statistical measure of their goodness of fit to the data. In the future, we could explore other extensions of the SIR model that might better model the spread of information. This might include using a combination of the highly connected population, the SEIR model, and an exponential decay term in the differential equation of population I. We could also use our differential equations to model other spreads of information, such as non-viral YouTube videos or views of advertisements, which has implications in economics and politics. Lastly, as discussed above, we could use our models to predict the general behavior of viral YouTube videos in the future.

5 References

- Brauer, Fred, Pauline Van Den Driessche, Jianhong Wu, and Linda J. S. Allen. "Compartmental Models in Epidemiology." Mathematical Epidemiology. Berlin: Springer, 2008. N. pag. Web.
- Broxton, Tom, Yannet Interian, Jon Vaver, and Mirjam Wattenhofer. "Catching A Viral Video." Journal of Intelligent Information Systems 40.2 (2013): 241-59. Web.
- Smith, Aaron. "6 New Facts About Facebook." Pew Research Center RSS. N.p., n.d. Web. 06 May 2014

6 Appendix

6.1 Gangnam Style

6.1.1 Model 1

- alpha=0.00002
- beta=0.00004
- gamma=.0032
- time=np.linspace(0,1600,8000)
- xinit=np.array([370,25,0])

6.1.2 Model 2

- alpha1=0.0001
- alpha2=0.01
- alpha3=0.043
- beta=0.017
- gamma=0.0026
- time=np.linspace(0,1600,8000)
- xinit=np.array([370,10,0,62,15,0])

6.1.3 Model 3

- alpha=0.0005
- beta=0.0093
- mu=0.0035
- gamma=0.0061
- time=np.linspace(0,1500,8000)
- xinit=np.array([360,22,3,9])

6.2 Call Me Maybe

6.2.1 Model 1

- alpha=0.00005
- beta=0.00004
- gamma=.0032
- time=np.linspace(0,1200,8000)
- xinit=np.array([310,25,0])

6.2.2 Model 2

- alpha1=0.0001
- alpha2=0.01
- alpha3=0.043
- beta=0.00016
- gamma=0.0026
- time=np.linspace(0,1200,8000)
- xinit=np.array([240,10,0,122,15,0])

6.2.3 Model 3

- alpha=0.0005
- beta=0.02
- mu=0.0035
- gamma=0.0078
- time=np.linspace(0,1200,8000)
- xinit=np.array([280,22,3,9])

6.3 What Does The Fox Say

6.3.1 Model 1

- alpha=0.00002
- beta=0.0002
- gamma=.0027
- time=np.linspace(0,950,8000)
- xinit=np.array([180,25,0])

6.3.2 Model 2

- alpha1=0.00001
- alpha2=0.00002
- alpha3=0.00003
- beta=0.0002
- gamma=0.0035
- time=np.linspace(0,950,8000)
- xinit=np.array([200,10,0,122,15,0])

6.3.3 Model 3

- alpha=0.0005
- beta=0.024
- mu=0.0035
- gamma=0.0073
- time=np.linspace(0,950,8000)
- xinit=np.array([200,22,3,9])