Video Viralization: a YouTube Case Study

Professional and Social Aspects of Informatics Engineering
- 2018/19 -

Author: Filipe Pires [85122]
Supervising Professor: Manuel de Oliveira Duarte

Abstract **– The study approach described in this report has the intetion of answering three main questions regarding viral phenomena: what happens during a viralization? why do viralization occur? and how do YouTube videos become viral? My methods focus on applying population based models to the question at hands and presenting a set of factors that greatly influence the success of video popularization. Along with this report, I deliver my work done in MATLAB that models the behaviour of any video popularity throughout a period of time from data collected from Google Trends.**

Index Terms **– YouTube, viralization, viral videos, epidemiological model, information cascade, MATLAB simulations, analysis tool**

Table of Contents

Introduction

Attempts to model population behaviour and topology behaviour in order to increase effectiveness and productivity in a certain field is not an uncommon subject amongst researchers. The applications of better understanding a target audience and the operations of any sort of platform that serves as an interface between that audience and the people interested in engaging them are many and a lot is to be gained from such valuable knowledge.

This report on how to create viral content on YouTube was written for the subject of "Aspetos Profissionais e Sociais de Engenharia Informática" of the bachelor's degree in Informatics Engineering at the University of Aveiro and describes a case study covering *what* happens when online content becomes highly popular, *why* do viral phenomena happen and *how* do YouTube videos reach millions of views nowadays.

If the reader is looking for a way to become rich by following a series of steps when creating a video and publishing it on YouTube, this paper is not recommended. My approach is a more didactical one, basing myself on theoretical experiments and population behaviour models, and offering a sort of analysis tool for readers interesting in making a research of their own on the topic.

I begin by comparing online information spreading to viral epidemics, then I explain the idea behind information cascades and correlate these concepts with the factors found to be the most connected to success within the video-sharing platform. Finally I execute several simulations to model successful cases of viral videos and draw some conclusions from the study.

1. Viral Phenomena

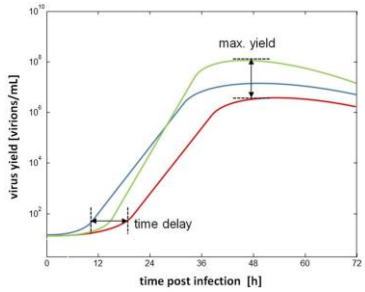
The definition of viral (or virality) has suffered a change throughout the past decades with the emergence of the social media technologies and the global networks of information. What once was an adjective applied to diseases relating to / caused by a biological virus (1) (2) (3), is now mostly addressed at as the "tendency of an image, video, or piece of information to be circulated rapidly and widely from one Internet user to another" by the same references and practicaly any other source you may find on a search engine (4). However, in this case, Wikipedia (5) seems to have the best references and give the most complete definition:

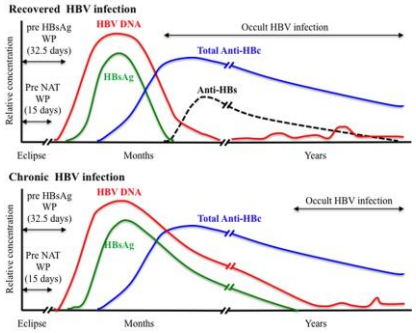
"Viral phenomena are objects or patterns that are able to (...) convert other objects into copies of themselves when these objects are exposed to them. (...) This has become a common way to describe how thoughts, information, and trends move into and through a human population. 'Viral media' is another common term whose popularity has been fueled by the rapid rise of social network sites. Different from the 'spreadable media', 'viral media' uses viral metaphors of 'infection' and 'contamination', which means that audiences play as passive carriers rather than an active role to 'spread' contents."

Let us consider this definition to address the question proposed in our study. From what I can gather, it seems that the application of this concept on the field of information propagation comes from the assumption that it resembles the way viruses propagate (hence the name "viral"). But is this true indeed? Can we use words like "infection" appropriately? Well, it is safe to respond "yes" if the behaviour of these two distinct situations are similar throughout time, so the first step must be to verify this claim.

1.1. Concept and Terminology

After doing a bit of research about virus behaviour on biological tissues, immunity amongst cells and the evolution process, a pattern could be distinguished: infection seems to draw a line S shaped, meaning that the initial propagation is a slow process, followed by an incredibly quick growth and stagnating in a specific percentage (bellow 100%) of the entire population; this percentage is defined by the number of cells immune to the virus, and the behaviour after this stagnation depends on whether there is a cure process or not; if it is observed a regression on the number of infected cells, the virus is curable and the S curve falls into the initial values tending to 0%. This can be more easily seen in figures 1,2 and 3 from three different articles related to "Multiscale Modeling of Influenza Virus Replication" (6), "Hepatitis B Virus Blood Screening" (7) and "Influenza Virus in Swine: Transmissibility Within and Between Populations" (8), respectively:





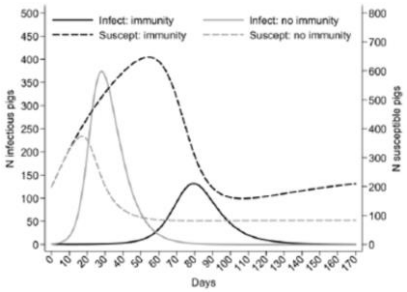
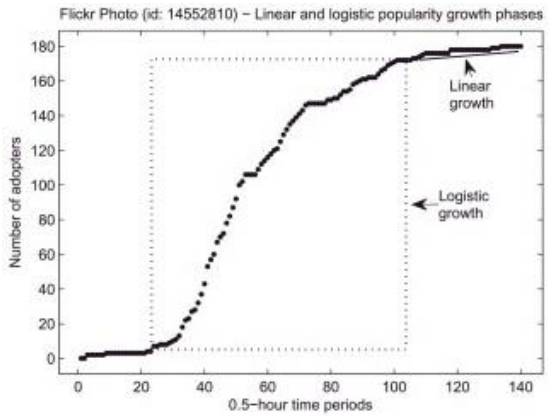


Figure 1: "Variability of maximum virus yield and the time delay before cells start to release (...) virus particles."
Figure 2: "Viral plasma markers in recovered and chronic hepatitis B infection."
Figure 3: "Expected circulation of influenza virus based on a deterministic model (...)."

*Influenza: commonly known as flu.

The research on information spreading behaviour throughout time focused on news, image and video, and showed interesting results. Apparently the evolution curve over time is similar to the one in the previous examples. Figures 4 and 5 show this behaviour in viral image (9) and articles (10), respectively.
This pattern is not only seen on content propagation (of images, articles and videos (11) (12) (13)), but also in interactions over the same content (e.g. reposts, likes, etc.) (14) and it appears to be reaching levels of contamination never before seen (in the music industry (15) and perhaps in other fields as well). The reason we do not see the decay after reaching the maximum values (the so called recovery phase) is because of the visualization metric chosen: views. As this metric only increases, it does not provide the information relative to the decay in popularity of the subject being analysed. However, this decay exists and is implied on the articles mentioned. Making the adaptation from total number of views to number of views per hour or day would be an alternative solution.



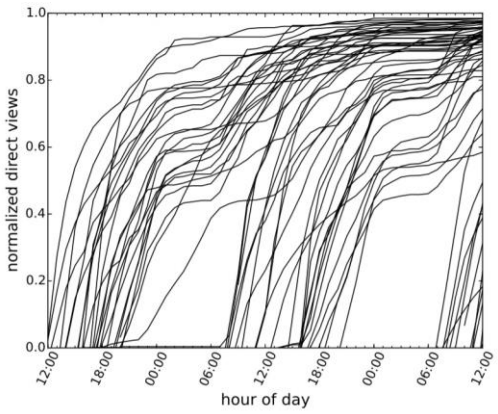


Figure 4: "Popularity evolution of Flickr photos through linear and logistic growth phases."
Figure 5: "Number of views in function of time for articles published between June 10 (...) and June 12 (...)."

The conclusion that can be drawn from this exercise is that, not only can we use the keywords "contamination" and "infected subjects" when talking about information viralization, but we can also apply the same models used in biology to track these phenomena.
The Princeton University's online course over "Networks: Friends, Money and Bytes" (16) goes through, amongst the many subjects discussed on the course, this topic of information spreading and defends the existance of 3 major characteristics of the online content gone viral observable after the phenomenon:

1. **High Peak** - the occurrence of a high peak of infection, where the contamination rate drastically accelerates and stabilizes (reduces to zero) on the top;
2. **Large Volume** - the existance of many infected subjects, in comparison to the total population;
3. **Short Time to Rise** - the difference between times of rise and drop of the propagation, being the infection time really small relative to the recovery time.

1.2. Epidemiological Model

But after all, how exactly do these curves work? And how can they be applied to a study on content viralization over the YouTube platform? The article written by J. Cannarella and J. A. Spechler on "Epidemiological Modeling of Online Social Network Dynamics" (17) stands as a valuable asset on understanding these social phenomena, as the authors explain in detail the math behind it and present an application of a model capable of plotting the curve of real data and predicting the evolution of viralizations in progress.

This is called the SIR model, a standard mathematical model of epidemics. SIR stands for Susceptible, Infected (or Infectious) and Recovered, which represent the 3 states in which we can find a subject of the population being studied. It is consisted of 3 ordinary differential equations, one for each subset of the population N (split by state), that govern the rates at which these subsets evolve during an outbreak. SIR works with the assumption that the population remains constant in number during the outbreak, as it states that S + I + R = N, independent of time (t).
I present the system of equations that translates the dynamics of online information spreading, along with the variable names in table 1, which compares the epidemiological interpretation of the variables to the equivalent interpretation on our case study. The additional variables β and γ are the contagious and recovery coefficients. The β (or beta) rate defines how fast is the disease transferred through interactions between the susceptible and the infected population. The γ (or gamma) controls the recovery speed at which the infected population recovers, without requiring any sort of interaction between population compartments. Beta is influenced by two other variables δ (or delta) and τ (or tau), which correspond to the contact rate among individuals per time iteration and the probability of a contact between a susceptible and an infected that results in disease transmission (or, in our case, in a new video viewer), and beta's value is given by: β = δτ.

$$\begin{cases} \frac{dS(t)}{dt} = -\beta \frac{SI}{N} \\ \frac{dI(t)}{dt} = \beta \frac{SI}{N} - \gamma I \\ \frac{dR(t)}{dt} = -\gamma I \end{cases}$$

System 1: System of differential equations that govern the rates at which the subsets evolve.

The first equation states that the rate at which the susceptible population turns contaminated is proportionate to the fraction of infected population $\frac{I}{N}$, the infection rate β, and the susceptible population S. The third equation states that the rate at which the infected population recovers is proportionate to the recovery rate γ and the infected population I. As it is seen in the second equation, there is a need for setting initial conditions for each of the population comparments: S(0) = So; I(0) = Io; R(0) = Ro. And Io, the size of the initial outbreak, must be non-zero for the infection to begin to spread. This and other details regarding the model are explained in greater detail in article (17). Other studies show that the applicability is transversal to several fields such as Marketing, as seen in article (18).

Symbol	Units	Disease Model Parameter	Equivalent Model Parameter
S	People	Susceptible	Potential video viewers
I	People	Infected	Video viewers
R	People	Recovered	Population no longer interested
β	$People \times Time^{-1}$	Infection Rate	Rate at witch people watch the video
γ	$People \times Time^{-1}$	Recovery Rate	Video Abandonment Rate
δ	$People \times Time^{-1}$	Contact Rate	Contact Rate
τ	-	Transmition Probability	New video viewer Probability

Table 1: Summary of parameters in the epidemiological model and their equivalent video spreading analogs. This table was taken from the article (17), increased and adapted to our case study.

In order to simulate the basic function of the model, a script was written in MATLAB, with the routine ode45 (19). The population was converted into a ratio of popularity, where 100 is equivalent to the entire population (N) and 0 to no subject from the population. This was useful further in the study as no accurate values were known about the total number of subjects nor the potential video viewers.
The 3 plots in figure 6 were accomplished after reading about solutions for the SIR model implemented in MATLAB (20), and they show how the epidemic develops starting from 1 unit of infected and continuing during 50 time iterations. The values for beta and gamma were changed for each simulation and the analysis of these parameters variation focuses on the infected compartment, in order to see their effect on the transmission curve.
The first simulation had defined $\beta = 0.01$, $\gamma = 0.1$ and the infection reached a rate of near 70 by the end of the 7th iteration. On the second plot, I increased the infection rate to 0.02 and, as the recovery rate maintained, the contamination process accelerated and by the 5th iteration it had already reached 80 on the ratio of popularity. The third and final simulation had defined $\beta = 0.01$, $\gamma = 0.2$, resulting in a peak of popularity happening in the same iteration as the first plot but reaching only a popularity rate of 50.

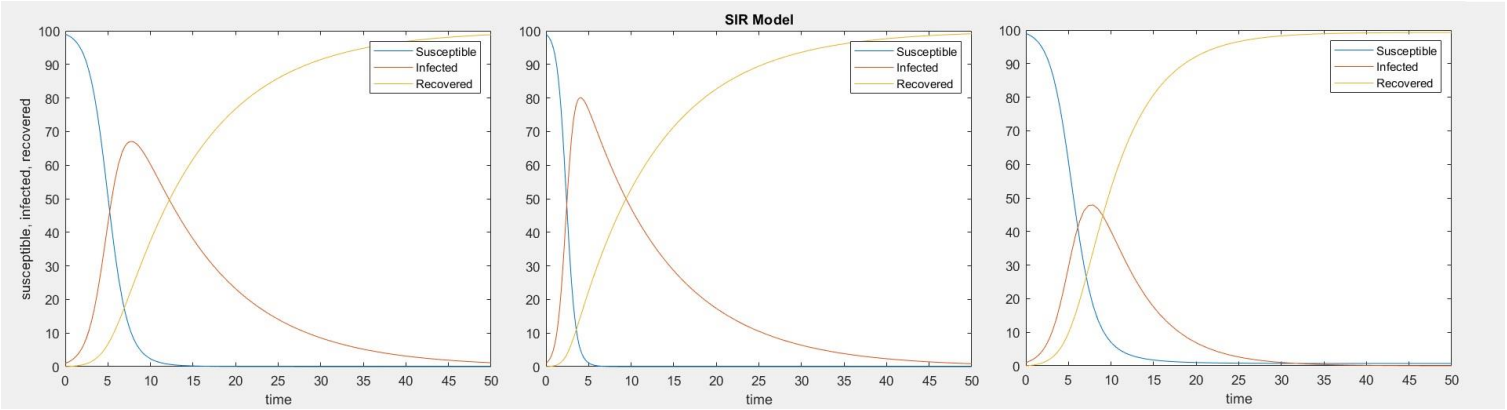


Figure 6: Simulations executed in MATLAB with different values for the beta and gamma parameters.

We can model the growth in popularity of YouTube videos with a fair accuracy by feeding the program with data from the subject being modeled (e.g. number of views per day for N days); I can do this without going too deep on how exactly does the platform deal with viralization by adjusting the model with a random variable.

1.3. Information Cascade

Now that we understand what happens to a population when an outbreak occurs through the use of the epidemiological model, we can start looking for the factors that influence the model's parameters in favour of a video publisher aiming at the viralization of his/her upload on YouTube. But before that, there is another matter to take in consideration.
The SIR model gives us an understanding of ***what*** happens when a video goes viral. However, we still do not understand ***why*** does it happen. This is surely to be important in order to reach conclusions on how to make a video go viral. So, in this section, I discuss the theory behind a related phenomenon called "Information Cascade" and how can it be applied to the analysis on online content viralization.

Once again, the Princeton University's online course (16) was a valuable reference to understanding what exactly is information cascade starting by explaining an experiment conducted by M. J. Salganik and D. J. Watts in 2005 (21). This controlled experiment envolved over 14000 participants and 48 songs from unknown bands and it consisted on showing these songs in different orders to the participants and analyse the popularity of the songs based on the number of downloads made in 4 different scenarios, depending on 2 different criteria:

- Ordering – the authors showed the songs either randomly or in a descending order by the current download number.
- Visibility – for each of the ordering options, they either showed the download number or hid it.

The results led to the conclusion that going from random order to descending and from hiding the download number to showing it not only increased the peer influence amongst participants but also greatly increased the number of total downloads as well. It is important to notice here that the authors also concluded that the unpredictability of success of a band also increased as the social influence power expanded.

The course then evolved into answering the question "Is there a general theory about these phenomena (including the impact of knowing what others have done in front of you)?" by building a simple model on sequential decision making in a crowd. Practically speaking, the model can be thought of as a line of people where each element can observe what happened with the previous elements and make a decision for the remaining to observe. On every iteration of the model execution, each person receives a private signal (unknown to other elements) and releases a public action.
This, in consequence, creates an information dependence amongst the population as each action depends on the observation of the previous actions. And it is this dependence that leads to what we call a cascade, where, no matter what the private signals are, people will always follow the same actions as the crowd before them have taken.

In order to understand this last paragraph, let us consider a though experiment that can be generalized to our case study of video popularization with some error margin as the experiment assumes that people react in a rational way and, as we know, this is not always the case. This time we will analyse the experiment as a series of steps and then confront the reality:

1. A set of people is formed, lined up and each receive a piece of paper without reading it
2. The paper has either 0 or 1 written in it, and that number is considered the correct one
3. Each person receives a private signal corresponding to a probabilistic version of the paper number, with a probability P of being the correct number, where $P>0.5$
4. As each person i has a private signal, let us assume that $Pi = P, \forall i$
5. Each person then guesses what is the correct number and writes it on a public display

When executing a simulation of this experiment, assuming that you are the first person and that you have received the private signal "1" (X1=1), it is logical to choose 1 as the correct answer (see step 3) and write it down (Y1=X1=1). The second person sees the previous guess "1" (or "0" if you had recieved that private signal). So she can assume what you have previously assumed and establish that Y1=X1.

Now if the second private signal is equal to the first (X2=X1), then logically she will write down the same number (Y1=Y2). But what if the private signals are different? Then Y2 can either be "0" or "1" with the same probability. The following probabilities help understanding the experiment at this stage.

$$P(Y2 = 1 \mid X2 = 1, Y1 = 1) = 1$$
$$P(Y2 = 1 \mid X2 = 1, Y1 = 0) = 0.5$$
$$P(Y2 = 0 \mid X2 = 1, Y1 = 0) = 0.5$$

Equations 1, 2 and 3: Probabilities for the decision of the second person from the thought experiment.

It is in the third person that cascades start. If Y1 and Y2 are different, this person can assume that X1 is different from X2 and so we go back to the same situation as the first person and ignore the first two guesses as they recieved different private signals. But what if Y1=Y2? In this case, if the third person's private signal is also equal (X3=Y1=Y2), then Y3 will clearly be equal to the remaining guesses. However, if X3 is different from the initial guesses – and this is the most interesting scenario – , the person is faced with two public signals that tell him one thing and one private signal that tells the opposite. Which one should he choose?

$$P(1 \mid (1,1,0)) = \frac{P(1) \times P((1,1,0)|1)}{P((1,1,0))} = \frac{P(1) \times P((1,1,0)|1)}{P(1) \times P((1,1,0)|1) + P(0) \times P((1,1,0)|0)}$$
$$= \frac{\frac{1}{2}(P^2(1-P) + \frac{1}{2}P(1-P)^2)}{\frac{1}{2}((P^2(1-P) + \frac{1}{2}P(1-P)^2) + ((1-P)^2P + \frac{1}{2}P^2(1-P)))}$$
$$= \frac{\frac{1}{2}(P + \frac{1}{2}(1-P))}{\frac{1}{2}(P + \frac{1}{2}(1-P) + (1-P) + \frac{1}{2}P)} = \frac{\frac{1}{2}(1+P)}{\frac{3}{2}} = \frac{1+P}{3}$$

Equation 4: Probability of the two public signals being correct and the private signal being wrong.

By calculating the probability of the true number being "1" given that Y1=Y2=1 and X3=0, we can learn the choice of the third person: if P(1|(1,1,0))>0.5 then he will choose "1". Through equation 4, and taking in consideration that P>0.5 (again, see step 3), then we come to the following conclusion:

$$\begin{cases} P(1 \mid (1,1,0)) = \frac{1+P}{3} \\ P > 0.5 \end{cases} \equiv \begin{cases} P(1 \mid (1,1,0)) > \frac{1+0.5}{3} \\ P > 0.5 \end{cases} \equiv \begin{cases} P(1 \mid (1,1,0)) > 0.5 \\ P > 0.5 \end{cases}$$

System 2: System applied to equation 4 in order to find out if P(1|(1,1,0))>0.5.

Consequently, Y3 will be equal to "1", the third person will have ignored its own private signal and an information cascade will have just been initated from that point on. Any of the remaining people will follow the crowd for the same reason as the third person did and this would happen wether the first two had written "1" or "0".

Another interesting conclusion that can be derived from these calculations is that, as the number of people *n* belonging to the experience increases, the probability of occurring an information cascade tends to 1:

$$P(no\_cascade) = P(1-P)$$
$$P(no\_cascade\_after\_2n\_users) = (P(1-P))^{n}$$
$$\lim_{n\to\infty}((P(1-P))^{n}) = 0$$

Equations 4, 5 and 6: Probability of not occurring cascade at the third user; Probability of not occurring cascade after n pairs of users; Probability of not occurring cascade after n pairs of users when n→∞.

The lecturer of the course proceeds on explaining the existance of correct and incorrect information cascades, as well as how to break them in the simplified experience previously explored. But for the purpose of our study what we have learned so far suffices to understand the importance of having access to information about others' decisions when making our own. Peer influence is indeed a great factor on the spread of information and online content such as videos. The exercise on binary decisions of a population allowed us to understand that, in theory, it is powerful enough to make rational people stop considering their own knowledge and choose to follow the crowd's decisions.
However, in practice, humans are not that simple or binary. Many are the factors that play a part on a person's action of watching a YouTube video or even having it appear on his/her feed or recommended tabs. Information cascades teach us that, whatever it is that the authors of viral videos do, it leads to the appearance of some sort of "third person's situation" where either a susceptible person actually did have interest in watching their video or it did not but still was subconsciously influenced to watch it anyway. Understanding all this helps us in our case study by providing a basis to where must the efforts be applied on when making a video "go viral".

2. Reaching the Trending Tab

Finally our focus shifts entirely towards YouTube as a platform that seems to be the heaven for viralizing multimedia content. In this chapter I assume that the reader knows what is YouTube and has a basic understanding of how it works. Nevertheless, I leave a reference to one of my previous works that covers exactly that (22).

I will present a contextualization on the main factors that influence the success of a video on the platform. I will then try to correlate the parameters from the SIR model of the MATLAB script I developed with these new factors; then I will attempt to conduct an analysis on existing successful cases where I managed to collect data and model it using the script. This conduct is ment to achieve a plausible understanding of ***how*** YouTube videos go viral.

2.1. Understanding the Context

As you might guess, the population is not the only reason why videos become viral. Information cascades are described by population models with great value, but topology has also an immense impact on the outcome of a viralization process.

Going back to Princeton's course, the lecturer also tackles this point when explaining topology based influence models. He goes over the problem with the help of graph networks and explaining how to measure the importance of a node when influencing other nodes. He states that there are several notions that help quantifying the node importance:

- Degree – the number of connections that a node has to others
- Dumbar's Number – an estimate, used by sociologists, of the largest number of nodes considered "friends" in a social network an average node can keep, which is around 150
- PageRank and Eigenvector Centrality – the importance of the nodes connected to a node influence the importance of that node (i.e. more important nodes connected means the node importance will increase); the most important nodes are considered the dominant eigenvectors and it is from them that the remaining nodes are ranked
- Closeness – how much is the distance between the node and its connections
- Betweenness Centrality – a quantification of how many nodes does one node link as an intermediary and how closely does it link these node pairs

The choice of using one of these notions depends on our purpose, but we can understand that by positively influencing the values of one or more of these importance quantifications, we will be contributing to the effectiveness of the topology on influencing the success of a video.

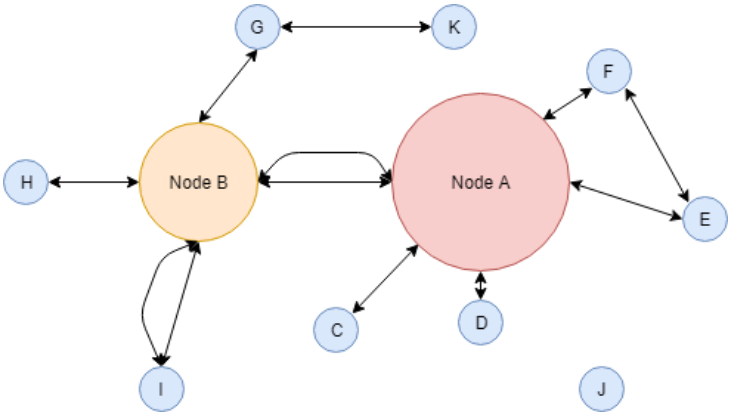


Figure 7: Visualization of an influence graph of 11 nodes with different characteristics.

To understand the importance quantifiers I will make a quick analysis on the example graph I made in figure 7 with the help of the tool Draw.io (23). The degree is easy to calculate for any node, for example degree(A) = 6 (and, yes, 6 not 5 as two nodes might be connected in more than one way). Dumbar's number does not quite apply here as it is a simple example. As for the eigenvectors, the simplest way to explain it is by stating that any of the small nodes connected to A will be considered more important than any of the nodes connected to B since node A is clearly larger (more important). But what is this of being "larger"? Well the node sizes in this visualization gives us the amount of influence power each node has (for external reasons or unknown). As for the closeness, one can understand that node D will be more influenced by node A than node C as it is closer to it. Finally we have the betweenness centrality that is easy to explain by mentioning that, although node G does not have many connections, its importance will be great as it is the only link between node K and any other node.

My strategy starts with looking for the attributes and features of YouTube and then map those with greater impact to these quantifications. What attributes am I talking about? Brian Dean, a researcher on SEO (Search Engine Optimization), presents a study on the optimization of YouTube's search engine (24) and the factors that most inflence the results that helps me answer this question. Dean and two other researchers analyzed 1.3 million YouTube videos and discovered that there is a list of 10 factor that most tightly correlate with successful YouTube videos. I present the list below in a descending order of "Spearman Correlation" (25) used by the authors to measure the impact of each factor:

1. View Count - while views aren't as important as they once were, they're still used to rank videos.
2. Video Likes - YouTube may use likes as a direct ranking factor; or it could be that heavily-liked videos generate other signals that YouTube truly cares about.
3. Video Shares - although Google, the owner of YouTube, has denied the fact that social signals play a role in their algorithm, it has not done so and from what the authors gathered is seems that shares have a strong correlation with higher rankings.
4. Video Comments - the more comments a video has, the higher it tends to rank.
5. New Subscribers - "Subscriptions driven", where uploaders seek to create videos that generate new subscribers, has a reasonably strong correlation with higher YouTube rankings
6. Channel Subscribers - channels with lots of subscribers have an advantage in YouTube; however, videos from smaller channels consistently outrank videos from popular channels.
7. Video Length - the average length of a video ranking on the first page of YouTube is 14 minutes and 50 seconds; Google has a patent for an algorithm that uses 'watch time' as a ranking signal.
8. Image Quality - HD is significantly more common than SD on the 1st page of the search results.
9. Keywords in Tags - including your target keyword as a tag may help with rankings; but the overall impact of tags appears to be small.
10. Keywords in Description - these do not improve YouTube's ranking but writing keyword-rich descriptions is recommended to make videos rank for related terms (suggested tab, etc.).

Do not be confused by the presence of factors such as view count. Yes, that is what this study is trying to help achieve – an increased number of viewers of a video – but here I am only explaining the factors that influence the topology of YouTube's network and the way its algorithms act. Dean's study only presents the raw facts, it does not attempt to reach any conclusions about them and I have not yet presented what I learned from his study and other sources around the matter. The view count factor is indeed inapropriate for our study, however the remaining factors can be linked to the node importance quantifiers and to the SIR model's rate variables, as it is done in table 2, along with an explanation also present in the table.

Success Factor	Node Importance Quantifier	SIR model Parameter	Observations
Views	-	-	Not applicable.
Likes	Eigenvector Centrality	δ	Although all factors are somehow connected to the eigenvector centrality quantifier, this one is the most logical one, as YouTube gives great importance to nodes (videos) with a large number of likes (positive feedback). These, in return also increase the contact rate as the main reason for viewers to share content is because they actually enjoyed it and desire for others to feel the same way or learn the same thing.
Shares	Degree, Betweenness	δ, γ	Shares and comments are powerful tools to spread information and flourish engagement with online content, as you can imagine. The degree of a video as a node is increased by these two factors since they are a form of connecting a susceptible person to a video through an infected viewer (either directly and taking advantage of other platforms such as Facebook or Email, or with the use of the tag feature on the comment section).
Comments			
New Subscribers	Degree, Dumbar's Number, Closeness	τ	Subscriptions are the most stable influencial tool, as it assures YouTube users that the channel viewers will most likely be exposed to new content from the same channel. Subscribers increase the degree because they automatically give connections to a new video without it having to prove its quality. Dumbar's number is also influenced here because YouTube helps users to keep track of their "friends" (a.k.a. subscriptions).
Subscribers			
Length	-	γ	Videos within the average length previously mentioned and/or a high image quality tend to maintain popular videos trending for a longer period of time, as people keep engaged and are less likely to get tired of them.
Quality			
Tag Keywords	Closeness	τ	Keywords are responsible for placing videos in categories. This reduces the distance (in terms of similarity) not only between videos, but also between potential users, resulting in a higher probability of propagation amongst people interested in categories that a video belongs to.
Descripttion Keywords			

Table 2: Mapping of the success factors to the importance quantifiers and the SIR parameters, with useful observations explaining the mapping of each.

It is important to be aware of several aspects of our study until so far: first of all, the reader must keep in mind that the models used have an error margin and may not be able to explain much of the reality; the same happens to the studies such as the one from B. Dean, as they do not have direct access to YouTube's algorithms and are works of reverse engineering conducted in order to better comprehend how the platform acts in different cases; the observations present in table 2 are of my own authorship, as well as the associations between columns, meaning that they may not be entirely correct and are merely a result of a personal analysis on the subject; a lot more sources alter the popularity of videos that are not described on this report as it is ment to maintain a certain level of simplicity and provide didactic content to any reader interested in the YouTube viralization process.

2.2. Successful Infections

I can now proceed to applying all that we have learned so far in real-case scenarios and extact conclusions from them. Having table 2 as a guide, it is possible to study the success of viral videos and detect which factors had the greatest impact on the outcome.
In this section I made several adaptations to the MATLAB script in order for it to be capable of receiving the data of each subset (S, I and R) for a number of time iterations and calculate the values for beta and gamma from the input data. The way the software calculates these values is through a simple implementation of the least squares method, which approximates the two parameters by computing the linear polynomial coefficients (you can learn more about it in MathWorks' Documentation (26)). This new code is also available at the end of the report.

The difficulty in the tasks presented below was to find data appropriate to feed the script. My research on analysis tools for YouTube and databases with information about the development of videos throughout time proved to be futile, as YouTube is not interested in sharing too much information as such, and the few services that could provide the intended information charged for that data and did not even have content about many of the videos currently available at YouTube.
My solution was to follow the path of other studies already mentioned in this report and take advantage of Google Trends' service (27) to have access to temporal data regarding the evolution of video popularity within the YouTube platform. This, however, proved to have many limitations. Google Trends only offers the ability to search for keywords; this means that the studies I present below include data that might not exactly match the video I reffer to – for example, in the first video I present one specific interview, but the data collected is from the keywords "<PersonName>" + "Debate", which means that it might reffer to other debates other than the intended one. Another problem was that the results are not presented in number of views or some similar metric, rather they are presented in a popularity unit (from 0 to 100) that only reffers to the time period I select; this is a limitation because, although I can still make an independent analysis of each video, I cannot compare the results amongst different videos as they are in different scales (i.e. a popularity of 100 in the first video might mean 1 million views whereas on the second it can mean 100 million views). The adaptation on the MATLAB code to this "popularity" metric was relatively easy, but this scale problem could not be solved. A third and fourth problems are that it is only possible to collect daily data (not hourly) and many of the keywords I attempted to search for did not return any results.
Nevertheless I proceeded with my objectives and divided the case studies by the keywords searched for on Google Trends: "Jordan Peterson Debate"; "Avengers Endgame"; "Notre Dame Fire"; "GoldieBlox". The reason I chose these four was to make a simple analysis on videos from several natures and different reasons for being viral.

Jordan Peterson Debate

The first successful infection I will present regards Jordan B. Peterson, a canadian "professor of psychology at the University of Toronto, a clinical psychologist and the author of the (...) bestseller '12 Rules for Life: An Antidote to Chaos'" (28). His work, online lectures and latest book have created what one might call an "Internet Boom" along with a lot of controversy around him and what he defends. My analysis focuses on one specific interview he attended for Channel 4 in 2018 where he talks about sensitive topics; this video (29) has reached over 16 million views and is a good example of a viral video concerning topics of sociology and online discussion.

When applying the SIR model to the data collected from Google Trends corresponding to the time period between the 8th of January and the 4th of March (when the video went viral), the infection curve had present a very easily visible S curve, followed by a slower descent in popularity, as seen in figure 8. The software began with $\beta = 0.01$ and $\gamma = 0.1$ (these are the default initial values for all simulations) and the output was $\beta = 0.0048$ and $\gamma = 0.\ 0672$. The reason why it maintained such as slow decrease in popularity is partially explained by the number of comments on the video (over 100 thousand); as we have previously seen, this factor is very much connected to high levels of engagement from the viewers, which in return results in high levels of popularity. Other factors like the image quality and video length might also have had some impact on this engagement. But in this particular case the most likely reason to why the video became so popular was the way he dealt with the attacks from the interviewer and the infrequent event of leaving a tv host speachless.

Avengers Endgame

Marvel Studios' "Avengers: Endgame - Official Trailer" (30) has emerged as a viral trailer video amongst the usually highly viewed Hollywood trailers due to many reasons as well. Curious enough, its peak only happened once the actual film was released, which explains the rapid growth of the infected subset. This video has reached over 120 million views but its slow recovery rate cannot be explained by the number of comments (only twice as much as the number of comments from J.P.'s debate), rather it is probably due to the nature of films themselves – one of today's greater sources of entertainment and, therefore, intrinsicaly very likely to be engaging to a large audience.

The input data corresponded to the time period between the 19th of April and the 27th of May of 2019 and the execution returned an output of $\beta = 0.0071$ and $\gamma = 0.0469$. The output plot is seen in figure 9.

Notre Dame Fire

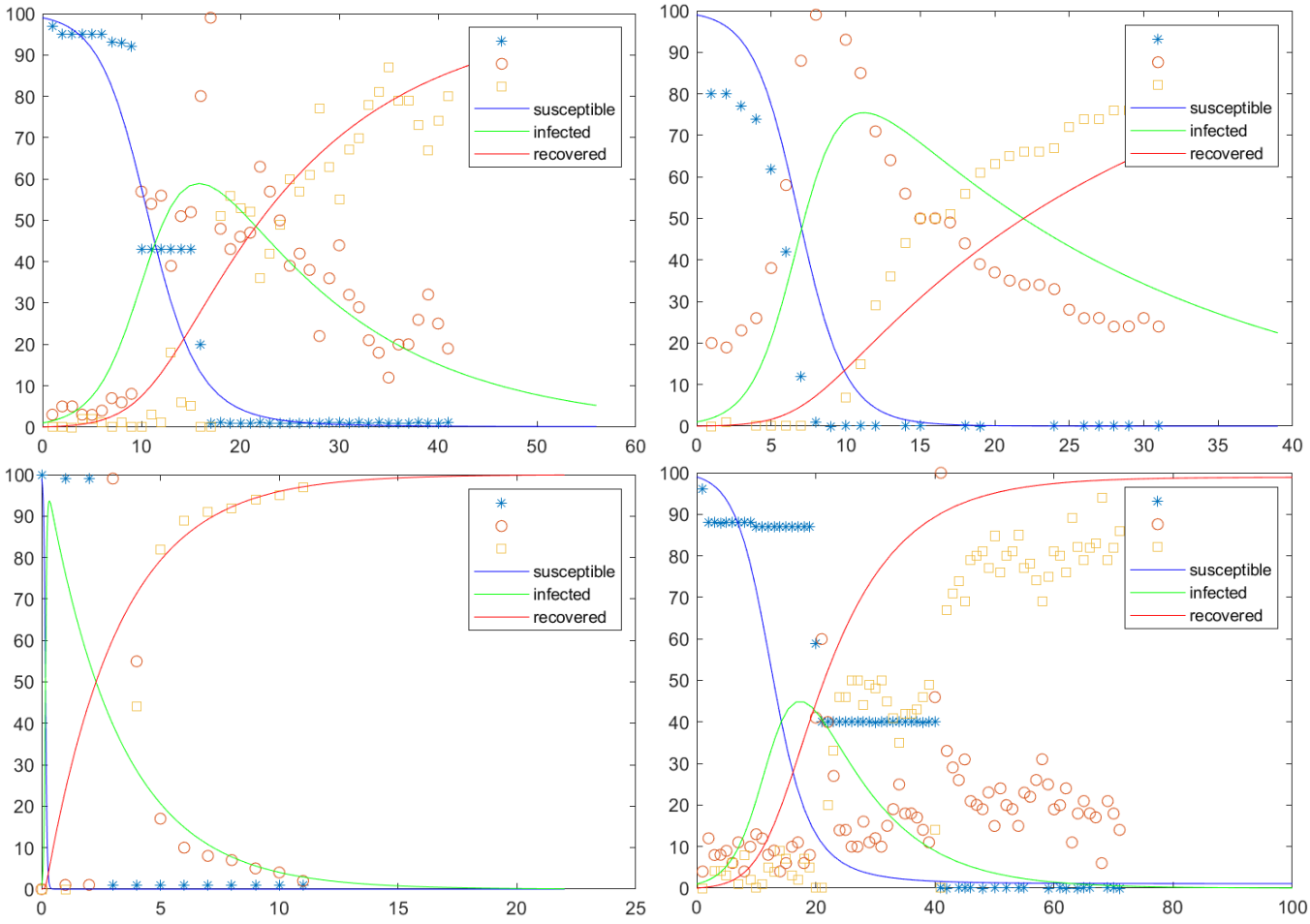
Catastrophes like Notre Dame's fire in Paris are one of the worst reasons for a video to go viral. The live footage of this event (31) was covered by several news channels throughout the world and reached millions of views in less than 4 hours. Figure 10 clearly shows how quickly was the infection amongst the susceptible population and it is in such cases that we more clearly understand the limitations of the data source, as the S curve would only be visible if the time iterations were in hours or even minutes.

The input data corresponded to the time period between the 13th of April and the 4th of May of 2019 and the execution returned an output of $\beta = 0.2918$ and $\gamma = 0\ 3288$. The fast decrease in populariy of the video is explained by its nature and not so much by its properties; viral news usually do not engage audiences for a long period of time. Nevertheless, a few observations can be made concerning the video publisher: BBC News. As it is a well known source of news and worldwide events, the company's YouTube channel might have had an advantage for creating such viral video due to its large number of subscribers and famous reputation. Also, the use of appropriate keywords might have also helped directing susceptible people towards the video.

GoldieBlox

Finally I present "This is Your Brain on Engineering (GoldieBlox PSA)" (32), a promotional video published by the toy company GoldieBlox that reached over 30 million views and promotes the increase of women in the fields of engineering.

The input data corresponded to the time period between the 18th of October 2014 and the 28th of January of 2015. The popularity evolution of this video is seen in figure 11 and the output values for beta and gamma were $\beta = 0.0044$ and $\gamma = 0.0972$. This is a good example of a viral video concerning topics of marketing and call-outs to action. It engages the audience mainly through its content, but also by applying efforts on generating new subscribers, making the production incredibly positive and, in consequence, likable and shareable, and uploading the video with great image quality. All these proved to be effective, as the company continued doing similar videos ever since.



Figures 8, 9, 10 and 11: Simulations executed in MATLAB with the datasets from J. B. Peterson's debate, Avengers Endgame trailer, Notre Dame's fire and GoldieBlox's promotional video, respectively.
The symbols *, ○, □ correspond to the values from the datasets of susceptible, infected and recovered populations.

2.3. Analysis Tool

The analysis done on the four viral cases were merely illustrative of the power of the script for studying a market. The real purpose of the software is to simulate the development of viral cases in one specific field (e.g. marketing videos), extract patterns and detect the efforts made in each that had a greater impact on the result.
The reason I did not choose videos all from one area of interest was to show the reader the generalized applicability of the software and its limitations in terms of time range as we saw on the Notre Dame's case, and to make a simple non-focused demonstration of the software's use.

This analysis tool written in MATLAB becomes more interesting when used by video producers aiming at increasing their chances of success in terms of number of views compared to their rivals. It should not be a unique source of information, as I have already presented many of its limitations, but it has the potential to be a valuable asset to such ends.
The software user should normalize all cases by executing the script for time periods with the same length and weight each case according to the total number of views they received. To improve the accuracy of the analysis, one should increase the number of successful infections on the field as much as possible; this way, patterns could start to emerge and more valuable conclusions could be extracted from them. Table 2 should also go along with the script as it helps understanding the practical connections between the model's parameters calculated when executed and the success factors intrinsic to YouTube as a video sharing platform.

Conclusion

My goals with this case study on YouTube viralization can be summarized into a list of three. First of all, I intended to provide the reader a solid theoretical documentation on what happens during the occurrence of viral phenomena and why do they happen, while leaving a margin for deeper studies conducted by more curious readers. Secondly, it was my desire to focus on the strongest factors of success of the platform in order to avoid creating unnecessary complexities and lose the reader's engagement on the topic; but I wished to tackle these factors in a mathematical way instead of studying the reasons for their impact on crowds and potential viewers. Finally I did not want to attempt to provide some sort of mathematical formula for success, as it would be very much likely inaccurate, rather I wished to develop an analysis tool that could help any reader with the desire to be successful on the production of videos of specific subjects with specific groups of susceptible audiences.

Having said this, I believe that my objectives were completed with success, even though my access to appropriate datasets was limited and it was not possible for me to present data from different sources in order to avoid the existance of the possibility that the one used was altered by the provider, Google itself. This possibility exists as they are the owners of YouTube and it is easy to imagine that a company with such influence might be interested in maintaing a level of ignorance regarding some aspects of its operations.

With regards to my methods, it is important to state that much of my work was based on the references mentioned throughout the report and that, without them, the quality of this work would not reach my expectations. For a future work as a continuation of this study, I would recommend applying the analysis tool in the way I mention in section 2.3 and present stronger correlations between model parameters and success factors, along with a description of several strategies that took advantage of these correlations and perhaps an experimental exercise with a test video.

Bibliography

1. Definition of Viral. *Oxford Lexico.* [Online] https://www.lexico.com/en/definition/viral.

2. Viral: Other Words. *Merriam-Webster Dictionary.* [Online] https://www.merriam-webster.com/dictionary/viral#other-words.

3. Viral (Disease). *Cambridge Dictionary.* [Online] https://dictionary.cambridge.org/dictionary/english/viral.

4. Definition of Virality. *Oxford Lexico.* [Online] https://www.lexico.com/en/definition/virality.

5. Viral Phenomenon. *Wikipedia.* [Online] https://en.wikipedia.org/wiki/Viral_phenomenon.

6. **Rudiger, Daniel.** Stochastic Multiscale Modeling of Influenza Virus Replication in Cell Cultures. *MPI-Magdeburg.* [Online] Max Planck Institute. https://www.mpi-magdeburg.mpg.de/3015798/Stochastic_Multiscale_Modeling_of_Influenza_Virus_Replication_in_Cell_Cultures.

7. **Laperche, Syria and Candotti, Daniel.** Hepatitis B Virus Blood Screening: Need for Reappraisal of Blood Safety Measures? *Frontiers in Medicine.* [Online] https://www.frontiersin.org/articles/10.3389/fmed.2018.00029/full.

8. Influenza virus in swine: Transmissibility within and between populations. *Engormix.* [Online] https://en.engormix.com/pig-industry/articles/influenza-virus-swine-transmissibility-t36226.htm.

9. Predicting the popularity growth of online content: Model and algorithm. *Science Direct.* [Online] https://www.sciencedirect.com/science/article/pii/S0020025516305242.

10. Modeling and Predicting the Popularity of Online. *Biblio UGent.* [Online] https://biblio.ugent.be/publication/8547204/file/8547206.pdf.

11. Characterizing and Predicting the Popularity of Online Videos. *IEEExplore.* [Online] https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7450136.

12. Modeling Dynamics of Online Video Popularity. *Research Gate.* [Online] https://www.researchgate.net/figure/CDF-of-entropy-for-each-video-type_fig2_269339482.

13. Understanding, Modeling and Predicting the Popularity of Online Content on Social Media Applications. *GitHub.* [Online] https://flaviovdf.github.io/papers/figueiredo2015-dissertation.pdf.

14. Model of information spread in social networks. *Research Gate.* [Online] https://www.researchgate.net/publication/307598597_Model_of_information_spread_in_social_networks.

15. Number of Days Until Music Video Hits 1 Billion Views. *Digital Music News.* [Online] https://www.digitalmusicnews.com/wp-content/uploads/2016/02/e714a27e-f3d6-424e-9c30-f710bbedd464.gif.

16. **Chiang, Mung and Brinton, Christopher.** Networks: Friends, Money and Bytes. *Coursera.* [Online] Princeton University. https://www.coursera.org/learn/friends-money-bytes.

17. **Cannarella, John and Spechler, Joshua A.** *Epidemiological modeling of online social network dynamics.* Princeton, NJ, USA : Princeton University, 2014. arXiv:1401.4208v1.

18. *Viral marketing as epidemiological model.* **Rodrigues, Helena and Fonseca, Manuel.** s.l. : 15th International Conference.

19. ode45. *MathWorks.* [Online] The MathWorks, Inc. https://www.mathworks.com/help/matlab/ref/ode45.html.

20. *MATLAB Programming for Simulation of an SIR Deterministic Epidemic Model.* **Agrawal, Ankit, Tenguria, Abha and Modi, Geeta.** 1, Bhopal, India : International Journal of Mathematics Trends and Technology, 2017, Vol. 50. 10.14445/22315373/IJMTT-V50P509.

21. **MJ, Salganik, PS, Dodds and DJ, Watts.** Experimental study of inequality and unpredictability in an artificial cultural market. *PubMed.* [Online] 2005. https://www.ncbi.nlm.nih.gov/pubmed/16469928.

22. **Pedrosa, André, et al., et al.** *Analysis of an Internet Service: YouTube.* Aveiro : DETI - University of Aveiro, 2019.

23. User Documentation. *Draw.io.* [Online] https://about.draw.io/tag/user-documentation/.

24. **Dean, Brian.** We Analyzed 1.3 Million YouTube Videos. Here's What We Learned About YouTube SEO. *Backlinko.* [Online] https://backlinko.com/youtube-ranking-factors.

25. YouTube Ranking Factors Study: Methods & Results. *Backlinko.* [Online] https://cdn-backlinko.pressidium.com/wp-content/uploads/2017/02/YT_Ranking_Factors_Study_Methods-2.pdf.

26. Documentation: Least-Squares Fitting. *MathWorks .* [Online] https://www.mathworks.com/help/curvefit/least-squares-fitting.html.

27. Explore what the world is searching. *Google Trends.* [Online] Google. https://trends.google.com/trends/.

28. About Jordan Peterson. *JordanBPeterson.* [Online] https://www.jordanbpeterson.com/about/.

29. Jordan Peterson debate on the gender pay gap, campus protests and postmodernism. *YouTube.* [Online] Google. www.youtube.com/watch?v=aMcjxSThD54.

30. Marvel Studios' Avengers: Endgame - Official Trailer. *YouTube.* [Online] https://www.youtube.com/watch?v=TcMBFSGVi1c.

31. Notre Dame: Blaze engulfs medieval icon - BBC News. *YouTube.* [Online] https://www.youtube.com/watch?v=rcGjykjs2Kk.

32. This is Your Brain on Engineering (GoldieBlox PSA). *YouTube.* [Online] https://www.youtube.com/watch?v=ArNAB9GFDog.

Appendix

Below is the code written in MATLAB responsible for creating the SIR models plotted in figures 6, 8, 9, 10 and 11. The first two files create a SIR model given a time period and the default values for the model's parameters. The remaining files are responsible for the processing to data corresponding to the development of the subsets S, I and R and returning the values for the model's parameters calculated with the help of the least squares method.

sir.m

```
clear; clc;
to = 0;          % starting time
tf = 50;         % finishing time
yo = [99 1 0];   % population, where yo(1,1) is the susceptible population,
yo(1,2) is the infected population and yo(1,3) is the recovered population

% calculate and plot graph
[t y] = ode45('ypsir',[to tf],yo);  % Matlab command used to appoximate the
solution of our system of differential equation
plot(t,y(:,1),t,y(:,2),t,y(:,3))    % Matlab command used to generate the
graphs of the three population groups
title('SIR Model')
legend('Susceptible', 'Infected', 'Recovered')
xlabel('time')
ylabel('susceptible, infected, recovered')
```

ypsir.m

```
function ypsir =ypsir(t,y)
    beta = .01;     % "contagious" coefficient - defines how fast is the
infection process
    gamma = .1;     % "recovery" coefficient - defines how fast is the
recovery process
    ypsir(1) =-beta*y(1)*y(2);
    ypsir(2) = beta*y(1)*y(2)-gamma*y(2);
    ypsir(3) = gamma*y(2);
    ypsir = [ypsir(1) ypsir(2) ypsir(3)]';
```

sirid.m

```
function [t y] = sirid(tf)
    yo = [99 1 0];
    to = 0;
    [t y] = ode45('ypsirid',[to tf],yo);
```

ypsirid.m

```
function ypsirid = ypsirid(t,y)
    global old_beta old_gamma
    ypsirid(1) = -old_beta*y(1)*y(2);
    ypsirid(2) = old_beta*y(1)*y(2) - old_gamma*y(2);
    ypsirid(3) = old_gamma*y(2);
    ypsirid = [ypsirid(1) ypsirid(2) ypsirid(3)]';
```

sir_parid.m

```
clear; clf(figure(1)); clc;
global old_beta old_gamma;
old_beta = 0.010; old_gamma = 0.100;

% Jordan Peterson Debate
td = [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53
54 55 56];
Id = [3 5 5 3 3 4 7 6 8 57 54 56 39 51 52 80 99 48 43 46 47 63 57 50 39 42 38
22 36 44 32 29 21 18 12 20 20 26 32 25 19 14 11 12 18 18 13 13 10 19 10 17 20
6 11 19];
Rd = [0 0 0 2 2 1 0 1 0 0 3 1 18 6 5 0 0 51 56 53 52 36 42 49 60 57 61 77 63
55 67 70 78 81 87 79 79 73 67 74 80 85 88 87 81 81 86 86 89 80 89 82 79 93 88
80];
numdata = 56; meas = 40; tf = 56;

% Avengers Endgame
% Notre Dame Fire
% GoldieBlox This is Your Brain on Engineering
% (...)

rvec = rand(1,numdata);
Id(2:numdata) = Id(2:numdata) + .1*rvec(1,2:numdata) - .05;

rvec = rand(1,numdata);
Rd(2:numdata) = Rd(2:numdata) + .1*rvec(1,2:numdata) - .05;
Sd = 100 - Id - Rd;
%
for i = 2:1:numdata-1
     ii = (i-1)*3;
     d(ii) = (Sd(i+1) - Sd(i-1))/(td(i+1) - td(i-1));
     d(ii+1) = (Id(i+1) - Id(i-1))/(td(i+1) - td(i-1));
     d(ii+2) = (Rd(i+1) - Rd(i-1))/(td(i+1) - td(i-1));
     A(ii,1) = -Sd(i)*Id(i); A(ii,2) = 0;
     A(ii+1,1)= Sd(i)*Id(i); A(ii+1,2) = -Id(i);
     A(ii+2,1)= 0.0; A(ii+2,2) = Id(i);
end

m = 3*meas + 1;
x = A(2:m,:)\d(2:m)';

[old_beta old_gamma]
[x(1) x(2)]
plot(td(1:1:meas+1),Sd(1:1:meas+1),'*',td(1:1:meas+1),Id(1:1:meas+1),'o',...
 td(1:1:meas+1),Rd(1:1:meas+1),'s')%,td,Sd,'x',td,Id,'x',td,Rd,'x')
old_beta = x(1);
old_gamma = x(2);
[t y] = sirid(tf);
hold on
plot(t, y(:,1), 'b',t, y(:,2), 'g',t,y(:,3), 'r')
legend('', '', '', 'susceptible', 'infected', 'recovered')
ylim([0 100])
```