Understanding viral video dynamics through an epidemic modelling approach

Rahil Sachak-Patwa, Nabil T. Fadai, and Robert A. Van Gorder*

Mathematical Institute, University of Oxford, Andrew Wiles Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, UK
**Robert.VanGorder@maths.ox.ac.uk*

Abstract

Motivated by the hypothesis that the spread of viral videos is analogous to the spread of a disease epidemic, we formulate a novel susceptible-exposed-infected-recovered-susceptible (SEIRS) delay differential equation epidemic model to describe the popularity evolution of viral videos. Our models incorporate time-delay, in order to accurately describe the virtual contact process between individuals and the temporary immunity of individuals to videos after they have grown tired of watching them. We validate our models by fitting model parameters to viewing data from YouTube music videos, in order to demonstrate that the model solutions accurately reproduce real behaviour seen in this data. We use an SEIR model to describe the initial growth and decline of daily views, and an SEIRS model to describe the long term behaviour of the popularity of music videos. We also analyse the decay rates in the daily views of videos, determining whether they follow a power law or exponential distribution. Although we focus on viral videos, the modelling approach may be used to understand dynamics emergent from other areas of science which aim to describe consumer behaviour.

Keywords: viral videos; epidemic model; discrete delays; delay differential equations

1. Introduction

Since the emergence of video sharing sites such as YouTube [1], the ability for both individuals and organisations to create, share, and promote videos on the internet has lead to the phenomenon of viral videos. As of August 2017, 76 videos on YouTube have received over 1 billion views, with the top two videos having been viewed over 3 billion times [2]. Beyond entertainment purposes, viral videos have had significant impacts on society, regularly emerging from political campaigns [3]. Companies and individuals have also exploited the power of viral videos for financial purposes [4]. Understanding how videos gain their views would be useful to many parties, including content generators, online advertisers, and internet search engines.

The viewing rate of a video depends on a multitude of factors, including the popularity of the uploader, the quality of the video, the power of media outlets to promote the video, and social trends. All these factors influence the behaviour of individuals to take action to watch and share a video. The viewing rates of user-generated content, particularly of YouTube videos, have been analysed in many papers. In [5], it was found that the popularity of individual videos shows substantial variation, especially within the first six weeks of being uploaded, such that there is little correlation between the number of added views in successive weeks when videos are young. It was shown that roughly three quarters of videos reach their peak popularity within these first six weeks, leading to the fact the set of videos which receive the top 1% or 10% percent of weekly views varies significantly. The importance of popularity, content, and social features which influence the evolution of a video's popularity was analysed in [6]. Specifically, the impact of referrers (incoming links to videos) was studied. It was found that on average, YouTube videos have long lifetimes, with around 80% of videos watched each day being older than a month, although the most popular videos tend to be recently uploaded [7].

Copyright protected videos on the whole have high viewing rates when they are young, and these views tend to decrease as the video gets older [8]. Groups of duplicate videos were analysed in [9], where it was shown that duplicate videos often receive a significantly different number of views due to the popularity of the uploader, the title, and tags associated with the video. Viewing rates of YouTube videos is strongly dependent on referrals from social media sites, such as Facebook and Twitter [3]. An algorithm in [6] was used to extract patterns from a dataset of video

views, which clusters videos into groups based on their distribution of their viewing rates. The viewing activity of videos is described as being exogenously perturbed when the video is featured by YouTube editors (an outside force), or endogenously perturbed when a video appears on the YouTube "most viewed today" list [10]. Viral videos are defined as those whose popularity is driven by endogenous effects where the population is "ripe" for the particular video. Their dynamics show "precursory word-of-mouth growth resulting from epidemic like propagation through a social network" [10]. Two simple models were presented in [11, 12], which predicted the future popularity of internet content. In [5], a three-phase characterisation is used, where videos were categorised as being before, at, or after their peak popularity. A model was proposed where videos move between the three phases and the model predicts key components of popularity dynamics, such as the distribution of weekly views, the distribution of total views, and the relative popularity of videos, reasonably accurately. A model based on extremely randomized ensemble trees was used to predict popularity trends of viral videos in [6].

Epidemic models have been used in various contexts to model internet content. Diffusion of messages on bulletin boards have been predicted with an SIR model [13], while a modified SIR model was proposed in [14] to describe the adoption and abandonment of online social networks. Various stochastic epidemic models were used in [15] to describe the spread of memes (content such as images, text, and videos which are shared online), with an SIRS model found as the optimal model to match the data. A rumour propagation SIR model, adapted from [16], was used in [17] to model the daily views of the video "Gangnam Style". An epidemic analogy was used in [18], where the "attention dynamics" of viral videos is modelled via an SIR Markov process. When the willingness of an individual to adopt a new product depends on whether their acquaintances have adopted it or not, epidemic models suggest that the transition towards large social contagion becomes "explosive", such that the product catches on quickly [19].

We are motivated by the above studies to take an epidemic modelling approach to the study of viral videos. In Section 2, we formulate an SEIRS epidemic model consisting of classes of susceptible, exposed, infected, and recovered individuals to model of the popularity of viral videos. Time-delays are incorporated into the model in order to accurately describe the virtual contact process between individuals, and the temporary immunity of individuals to videos after they have grown tired of watching them. We validate our models using daily viewing data from popular YouTube music videos in Section 3. We use an SEIR model to describe the initial growth and decline of daily views and an SEIRS model to describe the long term behaviour of the popularity of music videos. We also analyse the decay rates in the daily views of videos, determining whether they follow a power law or exponential distribution. We compare the ability of our SEIR model and an SIR model proposed in [17] to describe the initial viewing data for the video "Gangnam Style". Finally, in Section 4, we summarise our results and discuss the useful findings.

2. Single-Group Epidemic Models

We compartmentalize the population of relevant internet users into susceptible, exposed, infected, and recovered classes. Susceptible individuals are those who are not watching or sharing the video. Members of the exposed class are those who are currently in the latent period between becoming infected and being infectious themselves; that is to say, they have not watched the video yet but soon will do due to social pressure. Infected individuals are those who are watching and are actively promoting the video, while individuals in the recovered class are those who have become disinterested in the video and have stopped viewing and promoting it in any manner, albeit only temporarily. This compartmentalization leads us to the formulation of an SEIRS model, where individuals move through the susceptible, exposed, infected and recovered classes and back to the susceptible class, where the cycle repeats itself. We assume that the process of watching and sharing a video occur together, as during or straight after watching a video is when individuals are most likely to have the motive to share it. We define an individual as becoming infected when they have first watched and shared the video, as they then have a mechanism to infect susceptibles (such as a shared link to the video). We think of the sharing process as somewhat akin to that of a disease vector.

Time-delayed SEIRS models and their limiting cases have been studied in many instances, although it should be noted that the specific formulations of the models previously studied varies widely depending on the process the model was attempting to describe. Integral equation SIS models incorporating time-delays were analysed in [20] and [21] with the disease either dying out or approaching an endemic state. A time-delayed SIS population awareness epidemic model was investigated in [22], where it was shown that there exists a critical time-delay parameter value at which a Hopf bifurcation takes place. The stability of an SEIS integral equation model was analysed in [23], and the inclusion of delays did not change the nature of thresholds or asymptotic stability. In [24], it was shown that periodic

solutions may exist in an SIRS integral equation model and [25] showed that the inclusion of delays can destabilise the endemic equilibrium in an integro-differential equation SIRS model. In [26], the global behaviour of the SEIRS model proposed in [27] is analysed, and conditions for the asymptotic stability of the endemic equilibrium are given.

We build upon the SEIRS model proposed in [27], although in our case, we shall consider a closed system with a constant total population $N = S(t) + E(t) + I(t) + R(t)$, where S, E, I, and R represent the density of susceptible, exposed, infected and recovered classes, respectively. We define α as the rate of infection, which determines the ability of an infected individual to promote the popularity of a video. The force of infection is the rate of infection multiplied by the fraction of the population that is infected, namely $\alpha I/N$. Individuals leave the susceptible class at a rate $\alpha S I/N$. We assume that there exists a constant latent period, τ_1, from when an individual is exposed to the infection until they become infectious and a constant period of temporary immunity, τ_3, as in [27]. The recovery rate, γ, determines how quickly individuals become disinterested in the video. We assume that the period of time individuals remain infected is exponentially distributed ($e^{-\gamma t}$), with mean $1/\gamma$. We also include a constant latent period, τ_2, representing the time it takes for a shared link to a video to be seen. This is motivated by a vector transmission disease model [28], where an analogous delay is included to represent the time between the initial exposure of the vector carrier to the disease and the time when the vector carrier becomes infective. The inclusion of this parameter builds upon the model in [27] and leads to an SEIRS integro-differential equation.

$$\frac{dS(t)}{dt} = -\frac{\alpha}{N}S(t)I(t-\tau_2) + \gamma I(t-\tau_3), \tag{1}$$

$$E(t) = \int_{t-\tau_1}^{t} \frac{\alpha}{N}S(u)I(u-\tau_2)\,du, \tag{2}$$

$$\frac{dI(t)}{dt} = \frac{\alpha}{N}S(t-\tau_1)I(t-\tau_1-\tau_2) - \gamma I(t), \tag{3}$$

$$R(t) = \int_{t-\tau_3}^{t} \gamma I(u)\,du, \tag{4}$$

which holds for $t > 0$. We can differentiate (2) and (4) to obtain a delay differential equation system

$$\frac{dS(t)}{dt} = -\frac{\alpha}{N}S(t)I(t-\tau_2) + \gamma I(t-\tau_3), \tag{5}$$

$$\frac{dE(t)}{dt} = \frac{\alpha}{N}S(t)I(t-\tau_2) - \frac{\alpha}{N}S(t-\tau_1)I(t-\tau_1-\tau_2), \tag{6}$$

$$\frac{dI(t)}{dt} = \frac{\alpha}{N}S(t-\tau_1)I(t-\tau_1-\tau_2) - \gamma I(t), \tag{7}$$

$$\frac{dR(t)}{dt} = \gamma I(t) - \gamma I(t-\tau_3). \tag{8}$$

A complete list of variable and parameter definitions for our model can be found in Table 1. We assume non-negative histories $S(t) \geq 0$ for $t \in [-\tau_1, 0]$, $E(t) \geq 0$ for $t \in [-\tau_1, 0]$, $I(t) \geq 0$ for $t \in [-\bar{\tau}, 0]$, and $R(t) \geq 0$ for $t \in [-\tau_3, 0]$, where $\bar{\tau} = \max\{\tau_1 + \tau_2, \tau_3\}$. The solution of the integro-differential system (1)-(4) satisfies the delay differential system (5)-(8), provided that E and R are differentiable functions. Conversely, let $S(t)$, $E(t)$, $I(t)$, $R(t)$ be a solution to (5)-(8) with non-negative histories on the intervals stated above. Following on from [27], if additionally

$$E(0) = \int_{-\tau_1}^{0} \frac{\alpha}{N}S(u)I(u-\tau_2)\,du \quad \text{and} \quad R(0) = \int_{-\tau_3}^{0} \gamma I(u)\,du, \tag{9}$$

then the solution satisfies the integro-differential system (1)-(4). Moreover, for all $t \geq 0$, there exists a unique solution to (5)-(8) with $S(t) \geq 0$, $E(t) \geq 0$, $I(t) \geq 0$ and $R(t) \geq 0$. This result is proved in [27] and our addition of the parameter τ_2 does not alter the proof.

The importance of the integral conditions (9) should not be understated. It is shown in [28] that the neglect of these conditions may lead to solutions of a system that behave significantly differently to solutions of the same system restricted to obey them. For instance, periodic solutions may only occur when the integral conditions are neglected. When dealing with systems of delay differential equations, the non-negativity of solutions is not assured even if the

Symbol	Definition
$N(t)$	Total population density
$S(t)$	Susceptible population density
$E(t)$	Exposed population density
$I(t)$	Infected population density
$R(t)$	Recovered population density
α	Rate of infection (days^{-1})
γ	Rate of recovery (days^{-1})
τ_1	Individual latent period (days)
τ_2	Transmission latent period (days)
τ_3	Period of immunity (days)

Table 1: Variable and parameter definitions of the SEIRS model (5)-(8).

equivalent non-delayed system ensures it [29]. For example, the inclusion of a time-delay in an epidemic model in [30] results in a system in which solutions may become negative in finite time, which is not biologically feasible.

We non-dimensionalise the system (5)-(8) by making use of the non-dimensional parameters $S = S_0\hat{S}$, $E = E_0\hat{E}$, $I = I_0\hat{I}$, $R = R_0\hat{R}$, $t = \theta\hat{t}$, $\tau_1 = \theta\hat{\tau}_1$, $\tau_2 = \theta\hat{\tau}_2$, $\tau_3 = \theta\hat{\tau}_3$. We choose $\theta = 1/\gamma$ (the average infectious period) and $S_0 = E_0 = I_0 = R_0 = N$ so that $\hat{S} + \hat{E} + \hat{I} + \hat{R} = 1$, where $\hat{S}, \hat{E}, \hat{I}, \hat{R} \in [0, 1]$. Substituting our non-dimensionalised parameters into (5)-(8) and dropping the hats, we arrive at the non-dimensional system

$$\frac{dS(t)}{dt} = -\beta S(t)I(t-\tau_2) + I(t-\tau_3), \tag{10}$$

$$\frac{dE(t)}{dt} = \beta[(S(t)I(t-\tau_2) - S(t-\tau_1)I(t-\tau_1-\tau_2)], \tag{11}$$

$$\frac{dI(t)}{dt} = \beta S(t-\tau_1)I(t-\tau_1-\tau_2) - I(t), \tag{12}$$

$$\frac{dR(t)}{dt} = I(t) - I(t-\tau_3), \tag{13}$$

where $\beta = \alpha/\gamma > 0$ is the ratio of the infection and recovery rates. The integral conditions are given in non-dimensional form by

$$E(0) = \int_{-\tau_1}^{0} \beta S(u)I(u-\tau_2)\,du \quad \text{and} \quad R(0) = \int_{-\tau_3}^{0} I(u)\,du. \tag{14}$$

The non-dimensionalised infective and susceptible history functions are

$$I(t) = \begin{cases} 0, & \text{for } t < 0, \\ \mathcal{I}_0, & \text{for } t = 0, \end{cases} \qquad S(t) = \begin{cases} 1, & \text{for } t < 0, \\ 1 - \mathcal{I}_0, & \text{for } t = 0, \end{cases} \tag{15}$$

where $\mathcal{I}_0 \leq 1$ is the initial infective density. These history functions obey the integral conditions (14), as (15) implicitly implies that $E(t) = 0$ and $R(t) = 0$ for $t \leq 0$. The motivation for these history functions is that for $t < 0$, there should be no infected individuals as the video has not been released yet. When $t = 0$ and the video is released, a small number of individuals become infected and watch and share the video.

2.1. Disease-Free and Endemic Equilibria

As $E(t)$ and $R(t)$ are determined by $S(t)$ and $I(t)$, we initially need only consider (10) and (12) when determining the equilibrium solutions of the system. An equilibrium (S^*, I^*) of (10) and (12) satisfies

$$-\beta S^* I^* + I^* = 0, \tag{16}$$

$$S^* + I^* + E^* + R^* = 1. \tag{17}$$

Either, $I^* = 0$ satisfies (16), corresponding to the disease-free equilibrium (DFE), or $I^* > 0$ and $S^* = \beta^{-1}$ satisfies (16), corresponding to the endemic equilibrium (EE), in which case the disease persists. For the DFE we find $E^* = 0$ and $R^* = 0$, giving $S^* = 1$. For the EE, we find $E^* = \tau_1 I^*$ and $R^* = \tau_3 I^*$, implying $\beta^{-1} + I^* + \tau_1 I^* + \tau_3 I^* = 1$ from which we have $I^* = \frac{\beta-1}{\beta(1+\tau_1+\tau_3)}$. To summarise, the two equilibrium values of the system are

$$\text{DFE:} \quad S^* = 1, \qquad I^* = 0, \tag{18}$$

$$\text{EE:} \qquad S^* = \beta^{-1}, \quad I^* = \frac{\beta - 1}{\beta(1 + \tau_1 + \tau_3)}, \tag{19}$$

where the endemic equilibrium exists if and only if $\beta > 1$, i.e. if $\alpha > \gamma$. This corresponds to when the rate of infection is greater than the rate of recovery.

2.2. Linear Stability Analysis

To determine the linear stability of the DFE and EE, we consider linear perturbations $S(t) = S^* + \epsilon S_1 e^{\lambda t} + O(\epsilon^2)$ and $I(t) = I^* + \epsilon I_1 e^{\lambda t} + O(\epsilon^2)$, where $|\epsilon| \ll 1$ and $\lambda \in \mathbb{C}$. This gives the linear system

$$\begin{pmatrix} \lambda + \beta I^* & \beta S^* e^{-\lambda\tau_2} - e^{-\lambda\tau_3} \\ -\beta I^* e^{-\lambda\tau_1} & \lambda + 1 - \beta S^* e^{-\lambda(\tau_1+\tau_2)} \end{pmatrix} \begin{pmatrix} S_1 \\ I_1 \end{pmatrix} = 0. \tag{20}$$

As $S_1, I_1 = O(1)$ are arbitrary and non-zero, for (20) to have a solution we must have

$$\det \begin{pmatrix} \lambda + \beta I^* & \beta S^* e^{-\lambda\tau_2} - e^{-\lambda\tau_3} \\ -\beta I^* e^{-\lambda\tau_1} & \lambda + 1 - \beta S^* e^{-\lambda(\tau_1+\tau_2)} \end{pmatrix} = 0, \tag{21}$$

which gives the characteristic equation for a general equilibrium of the system

$$(\lambda + \beta I^*)(\lambda + 1 - \beta S^* e^{-\lambda(\tau_1+\tau_2)}) + \beta I^* e^{-\lambda\tau_1}(\beta S^* e^{-\lambda\tau_2} - e^{-\lambda\tau_3}) = 0, \tag{22}$$

where the eigenvalues λ_k are solutions of (22). If the spectrum of eigenvalues $\{\lambda_k\}$ satisfies $\max\{\mathrm{Re}(\lambda_k)\} \leq 0$, then the equilibrium (S^*, I^*) is stable. Conversely, if there exists a single eigenvalue with $\mathrm{Re}(\lambda_k) > 0$, then the equilibrium is unstable.

Theorem 1. *The disease-free equilibrium ($S^* = 1$, $I^* = 0$) of* (10)-(13) *is stable if and only if $\beta \leq 1$.*

Proof. Substituting $(S^* = 1, I^* = 0)$ into (22) gives $\lambda(\lambda + 1 - \beta e^{-\lambda(\tau_1+\tau_2)}) = 0$ and so $\lambda = 0$ is an eigenvalue and the remaining eigenvalues solve

$$\lambda + 1 - \beta e^{-\lambda(\tau_1+\tau_2)} = 0. \tag{23}$$

If $\lambda \in \mathbb{R}$, define $T = \tau_1 + \tau_2 \geq 0$. We can make the substitution $y = (\lambda + 1)T$ to give $ye^y = \beta T e^T$, which implies $y = W\left(\beta T e^T\right)$ where W is the Lambert-W function. After substituting $y = (\lambda + 1)T$, we obtain

$$\lambda = -1 + \frac{1}{T} W(\beta T e^T). \tag{24}$$

Define $f(T, \beta) = \frac{1}{T} W(\beta T e^T)$. One may prove that for $T > 0$, $f(T, \beta) < 1$ for all $\beta \in (0, 1)$, $f(T, 1) = 1$, and $f(T, \beta) > 1$ for all $\beta > 1$. We then deduce from (24) that the disease-free equilibrium is unstable for $\beta > 1$, as there exists a real positive eigenvalue λ. Although we have shown that the real eigenvalues are non-positive for $\beta \leq 1$, we must consider $\lambda \in \mathbb{C}$ to determine the stability for $\beta \leq 1$ as we need to show there exists no eigenvalues with positive real part in order to demonstrate stability.

Writing $\lambda = \mu + i\omega$ with $\omega \neq 0$, (23) becomes $\mu + i\omega + 1 - \beta e^{-\mu T} e^{-i\omega T} = 0$, where $T = \tau_1 + \tau_2$ as before. Splitting this into real and imaginary parts yields $\mu + 1 - \beta e^{-\mu T} \cos(\omega T) = 0$ and $\omega + \beta e^{-\mu T} \sin(\omega T) = 0$, which after rearranging, squaring, and adding gives the relation $(\mu + 1)^2 + \omega^2 = \beta^2 e^{-2\mu T}$. If $\beta \leq 1$, then $(\mu + 1)^2 + \omega^2 \leq e^{-2\mu T}$. Now assume $\mu > 0$ such that $(\mu + 1)^2 + \omega^2 < 1$. This yields a contradiction, as this disc is bounded in the left half of the complex plane where $\mu < 0$. Hence, along with our analysis for the degenerate case where $\lambda \in \mathbb{R}$, we can conclude that the disease-free equilibrium is stable if and only if $\beta \leq 1$. Recall that $\beta \leq 1$ corresponds to the case where the disease-free equilibrium is the unique equilibrium of the system. □

Substituting the endemic equilibrium $\left(S^* = \beta,\ I^* = \frac{\beta-1}{\beta(1+\tau_1+\tau_3)}\right)$ into (22) yields

$$\lambda(\lambda + 1 - e^{-\lambda(\tau_1+\tau_2)}) + \frac{\beta-1}{1+\tau_1+\tau_3}\left(\lambda + 1 - e^{-\lambda(\tau_1+\tau_3)}\right) = 0. \tag{25}$$

In general, (25) cannot be solved analytically in a generic manner, and hence, the stability of the EE must be determined numerically for fixed parameter values. In the next subsection, we will consider several specific special cases, in order to better understand loss of stability of the EE.

2.3. Stability Loss of the EE in Limiting Cases

As (25) cannot be solved analytically in a generic manner, the stability of the EE cannot be determined in general, as was the case for the DFE. That said, analytical progress is possible in certain limiting cases, which we shall now explore.

To begin, we write $\lambda = \mu + i\omega$, and partition (25) into real and imaginary parts, obtaining

$$\mu^2 - \omega^2 + \mu - e^{-\mu(\tau_1+\tau_2)}[\mu\cos(\omega(\tau_1+\tau_2)) + \omega\sin(\omega(\tau_1+\tau_2))] + \frac{\beta-1}{1+\tau_1+\tau_3}[\mu + 1 - e^{-\mu(\tau_1+\tau_3)}\cos(\omega(\tau_1+\tau_3))] = 0, \tag{26}$$

$$2\mu\omega + \omega - e^{-\mu(\tau_1+\tau_2)}[\omega\cos(\omega(\tau_1+\tau_2)) - \mu\sin(\omega(\tau_1+\tau_2))] + \frac{\beta-1}{1+\tau_1+\tau_3}[\omega + e^{-\mu(\tau_1+\tau_3)}\sin(\omega(\tau_1+\tau_3))] = 0. \tag{27}$$

Setting $\mu = 0$ to find the Hopf bifurcation boundary gives

$$-\omega^2 - \omega\sin(\omega(\tau_1+\tau_2)) + \frac{\beta-1}{1+\tau_1+\tau_3}[1 - \cos(\omega(\tau_1+\tau_3))] = 0, \tag{28}$$

$$\omega - \omega\cos(\omega(\tau_1+\tau_2)) + \frac{\beta-1}{1+\tau_1+\tau_3}[\omega + \sin(\omega(\tau_1+\tau_3))] = 0. \tag{29}$$

In general, (28) and (29) cannot be solved analytically, which motivates us to consider some special cases for the time-delay parameters τ_1, τ_2 and τ_3. From this, we determine stability conditions and generate bifurcation diagrams numerically which we consider in the following sections.

2.3.1. The $\tau_2 = \tau_3$ Limiting Case

Let us first consider the case where $\tau_2 = \tau_3 = 0$ and $\tau_1 \neq 0$, in which (25) becomes

$$\left(\lambda + \frac{\beta-1}{1+\tau_1}\right)\left(\lambda + 1 - e^{-\lambda\tau_1}\right) = 0, \tag{30}$$

thus $\lambda = -\frac{\beta-1}{1+\tau_1}$ an eigenvalue and the remaining eigenvalue solves

$$\lambda + 1 - e^{-\lambda\tau_1} = 0. \tag{31}$$

This implies that

$$\lambda = -1 + \frac{1}{\tau_1}W(\tau_1 e^{\tau_1}) = 1 - \frac{\tau_1}{\tau_1} = 0. \tag{32}$$

Hence the only non-zero eigenvalue is given by $\lambda = -\frac{\beta-1}{1+\tau_1}$, so assuming that the endemic equilibrium exists (i.e., $\beta > 1$), the endemic equilibrium is stable for all $\tau_1 > 0$, provided that $\tau_2 = \tau_3 = 0$.

More generally, if $\tau_2 = \tau_3 = \tau$, then (25) becomes

$$\left(\lambda + \frac{\beta-1}{1+\tau_1+\tau}\right)\left(\lambda + 1 - e^{-\lambda(\tau_1+\tau)}\right) = 0, \tag{33}$$

and we can use the same argument to show the endemic equilibrium is stable for all $\tau_1, \tau > 0$. Similarly if all delay parameters are equal such that $\tau_1 = \tau_2 = \tau_3 = \tau$, with the substitution $\bar{\tau} = \tau_1 + \tau$ we obtain the same stability results.

2.3.2. The $\tau_1 = \tau_3 = 0$ Limiting Case

When $\tau_1 = \tau_3 = 0$ and $\tau_2 \neq 0$, (25) becomes

$$\lambda(\lambda + \beta - e^{-\lambda\tau_2}) = 0. \tag{34}$$

Firstly assuming $\lambda \in \mathbb{R}$, the non-zero eigenvalues are given in terms of the Lambert-W function by

$$\lambda = -\beta + \frac{1}{\tau_2} W(\tau_2 e^{\tau_2\beta}). \tag{35}$$

If we let $\tilde{\tau} = \tau_2\beta$, then (35) becomes

$$\lambda = \beta\left(-1 + \frac{1}{\tilde{\tau}} W(\beta^{-1}\tilde{\tau}e^{\tilde{\tau}})\right). \tag{36}$$

We can then use an analogous argument as in the proof of Theorem 1 to show that if we define $g(\tau,\beta) = \frac{1}{\tau} W(\beta^{-1}\tau e^{\tau})$, then

$$g(T,\beta) > 1 \quad \text{for all } \beta \in (0\,,1), \tag{37}$$
$$g(T,1) = 1, \tag{38}$$
$$g(T,\beta) < 1 \quad \text{for all } \beta > 1. \tag{39}$$

Hence, $\lambda < 0$ if and only if $\beta > 1$.

Now, considering $\lambda \in \mathbb{C}$ such that $\lambda = \mu + i\omega$, splitting the factorised expression in (34) into real and imaginary parts yields

$$\mu + \beta - e^{-\mu\tau_2}\cos(\omega\tau_2) = 0, \tag{40}$$
$$\omega + e^{-\mu\tau_2}\sin(\omega\tau_2) = 0, \tag{41}$$

which after rearranging, squaring, and adding gives

$$(\mu + \beta)^2 + \omega^2 = e^{-2\mu\tau_2}. \tag{42}$$

If $\beta > 1$, then

$$(\mu + 1)^2 + \omega^2 < e^{-2\mu\tau_2}. \tag{43}$$

Now assume $\mu > 0$ such that

$$(\mu + 1)^2 + \omega^2 < 1, \tag{44}$$

which yields a contradiction as this disc is bounded in the left half complex plane where $\mu < 0$. Hence, if $\beta > 1$ such that the endemic equilibrium exists, then it is stable for all $\tau_2 > 0$, provided that $\tau_1 = \tau_3 = 0$.

2.3.3. The $\tau_1 = \tau_2 = 0$ Limiting Case

In the case where $\tau_1 = \tau_2 = 0$ and $\tau_3 \neq 0$, (25) simplifies to

$$\lambda^2 + \frac{\beta - 1}{1 + \tau_3}\left(\lambda - e^{-\lambda\tau_3} + 1\right) = 0. \tag{45}$$

Splitting λ into real and imaginary parts and by substituting $\tau_1 = \tau_2 = 0$ into (28) and (29), we find that the Hopf bifurcation boundary is determined by the equations

$$-\omega^2 + \frac{\beta - 1}{1 + \tau_3}[1 - \cos(\omega\tau_3)] = 0, \tag{46}$$
$$\omega + \sin(\omega\tau_3) = 0, \tag{47}$$

Rearranging and squaring (46) and (47) gives

$$\omega^2\left[\omega^2 + \frac{\beta - 1}{1 + \tau_3}\left(\frac{\beta - 1}{1 + \tau_3} - 2\right)\right] = 0, \tag{48}$$

and so the non-zero values of ω are determined by a quadratic, which gives real solutions for

$$\frac{\beta-1}{1+\tau_3}\left(\frac{\beta-1}{1+\tau_3}-2\right)<0, \tag{49}$$

implying

$$1<\beta<3+2\tau_3, \tag{50}$$

which is clearly a subset of $\beta>1$. Without loss of generality, we can take the positive root, namely,

$$\omega=\frac{\sqrt{(\beta-1)(3+2\tau_3-\beta)}}{1+\tau_3}, \tag{51}$$

and from (47) we obtain the implicit equation

$$\tau_3=\frac{-1+\tau_3}{\sqrt{(\beta-1)(3+2\tau_3-\beta)}}\arcsin\left(\frac{\sqrt{(\beta-1)(3+2\tau_3-\beta)}}{1+\tau_3}\right), \tag{52}$$

which determines the Hopf bifurcation boundary.

This boundary is shown in Figure 1, where we see that the endemic equilibrium is stable for approximately $\tau_3<4.6$ for all $\beta>1$. For $\tau_3>4.6$, as β increases for fixed τ_3, the endemic equilibrium switches from being stable to unstable and back to stable again. Recall that for $\beta\leq 1$ the endemic equilibrium does not exist; this corresponds to the black unfeasible region.

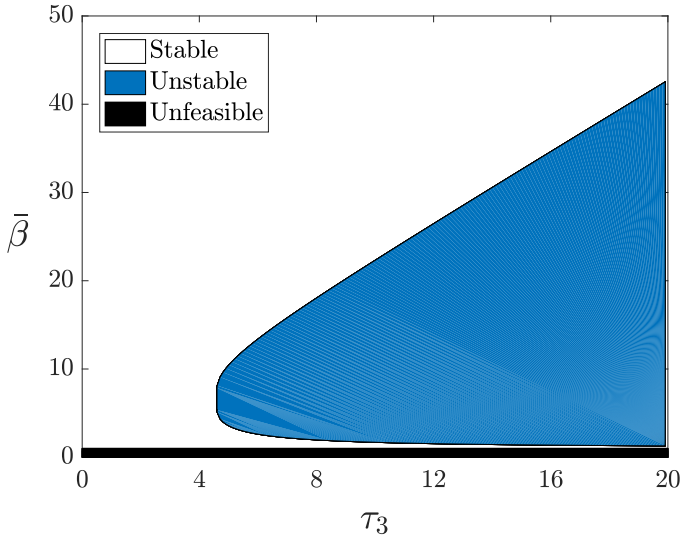
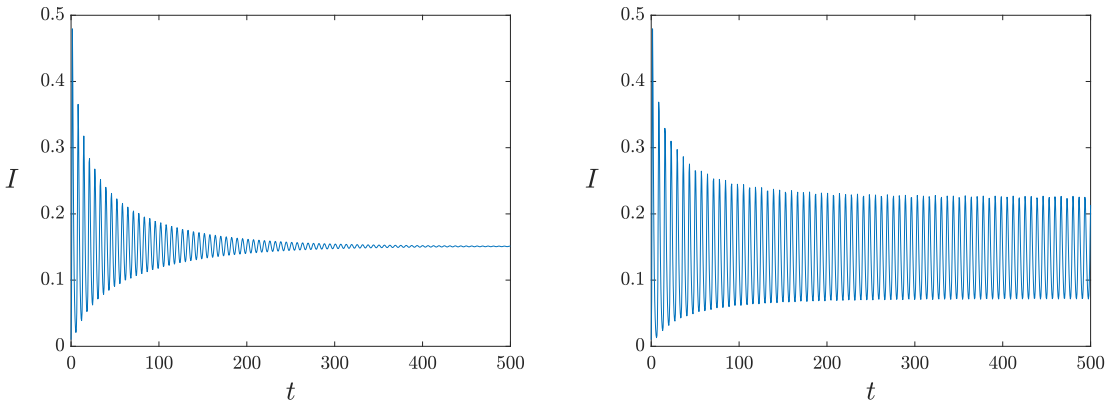


Figure 1: Bifurcation diagram in the τ_3-β plane of the endemic equilibrium of (10)-(13) with fixed $\tau_1=\tau_2=0$.



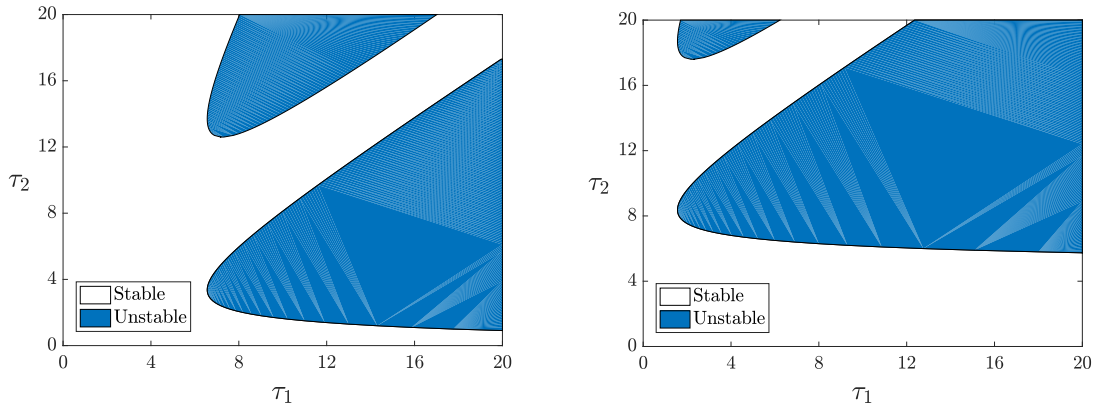
(a) $\beta = 5\,, \tau_3 = 4.3$ (b) $\beta = 5\,, \tau_3 = 4.8$

Figure 2: Plots of the non-dimensionalised infective density solutions I of (10)-(13), corresponding to points on either side of the Hopf bifurcation boundary for $\tau_1 = \tau_2 = 0$ in Figure 1, displaying a stable and unstable endemic equilibrium.

In Figure 2, we have plotted the solutions to the system (10)-(13) for the case of $\tau_1 = \tau_2 = 0$ on either side of the bifurcation boundary (52) with discontinuous history (15) and $I_0 = 0.01$. Note that we need only consider the differential equations for the susceptible and infected classes, namely (10) and (12), when carrying out numerical simulations, as E and R are determined by S and I. For $\beta = 5$ and $\tau_3 = 4.3$ in Figure 2 (a), the system converges to a stable endemic equilibrium, while for $\beta = 5$ and $\tau_3 = 4.8$ in Figure 2 (b), the system is unstable, and oscillates about the endemic equilibrium.

2.3.4. The Fixed τ_3 Limiting Case

We now consider limiting cases in which only one time-delay parameter is fixed. We generate the bifurcation diagrams for these cases by solving (28)-(29) numerically. We first consider the τ_3 delay parameter fixed. Note that $\tau_3 = 0$ is the case in which there is no temporary immune period, and infected individuals immediately become susceptible when they recover, resulting in an SEIS model.



(a) $\tau_3 = 0$ (b) $\tau_3 = 5$

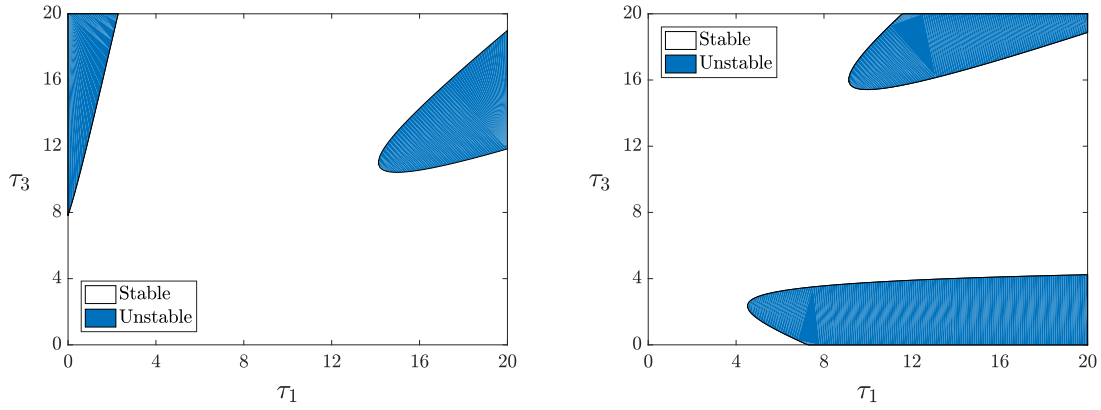
Figure 3: Bifurcation diagram in the τ_1-τ_2 plane of the endemic equilibrium of (10)-(13) with fixed τ_3 and $\beta = 2$.

The $\tau_3 = 0$ limit is essentially the model proposed in [28], although the authors do not analyse the stability of the endemic equilibrium apart from stating that periodic solutions may exists for a certain range of parameter values.

The bifurcation diagram in the τ_1-τ_2 plane for the endemic equilibrium can be seen in Figure 3 with $\beta = 2$ and $\tau_3 = 0$ and $\tau_3 = 5$. These two values of τ_3 represent the two cases when the the immunity related delay parameter is "off" and "on" respectively. In both Figure 3 (a) and (b) we see two distinct regions in the τ_1-τ_2 plane where the endemic equilibrium is unstable. From Figure 1 we expected that the origin in both plots in Figure 3 corresponding to $\tau_1 = \tau_2 = 0$ gives a stable equilibrium. Furthermore, in both Figure 3 (a) and (b), the τ_1 and τ_2 axes correspond to a parameter space where the endemic equilibrium is stable.

2.3.5. The Fixed τ_2 Limiting Case

Secondly, we consider the case where τ_2 is fixed. When $\tau_2 = 0$, there is no delay in the transmission of the video via a shared link or message. The endemic equilibrium of a similar model with two delays, but which also includes births and deaths, is analysed in [26]. It was proved that the disease is uniformly persistent if the basic reproduction number (the expected number of secondary cases produced by a single infection in a completely susceptible population) is greater than one. The corresponding bifurcation diagrams in the τ_1-τ_3 plane for $\tau_2 = 0$ and $\tau_2 = 5$ can be seen in Figure 4, again with $\beta = 2$. Similarly to the previous limiting case, in both Figure 4 (a) and (b) we see two distinct regions in the τ_1-τ_3 plane where the endemic equilibrium is unstable. In Figure 4 (a) when $\tau_2 = 0$, there exists an unstable region on the τ_3 axis where $\tau_1 = 0$ and τ_3 is approximately 8. We can see that this corresponds to the bifurcation diagram Figure 1 for fixed $\tau_1 = \tau_2 = 0$, as when $\beta = 2$ in Figure 1, the Hopf bifurcation boundary value of of τ_3 is approximately 8 as well.



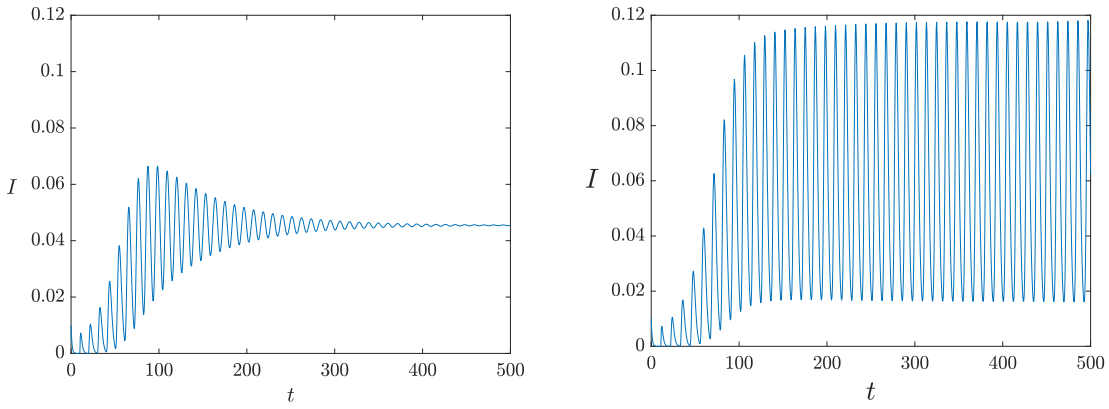
(a) $\tau_2 = 0$ (b) $\tau_2 = 5$

Figure 4: Bifurcation diagram in the τ_1-τ_3 plane of the endemic equilibrium of (10)-(13) with fixed τ_2 and $\beta = 2$.

In Figure 5, we have plotted solutions to the system (10)-(13) for $\beta = 2$ and $\tau_2 = 5$ with discontinuous history (15) and $I_0 = 0.01$. In Figure 5 (a), the system converges to a stable endemic equilibrium for $\tau_1 = 5$ and $\tau_3 = 5$, while in Figure 5 (b), the system exhibits oscillations about the endemic equilibrium for $\tau_1 = 6$ and $\tau_3 = 2$. We can see from Figure 4 (b) that these parameter values relate to the stable and unstable regions of parameter space respectively. Also note that Figure 5 (a) has all delay parameters equal ($\tau_1 = \tau_2 = \tau_3 = 5$) which we proved earlier always results in a stable endemic equilibrium (assuming that $\beta > 1$ and the endemic equilibrium exists).

2.3.6. The Fixed τ_1 Limiting Case

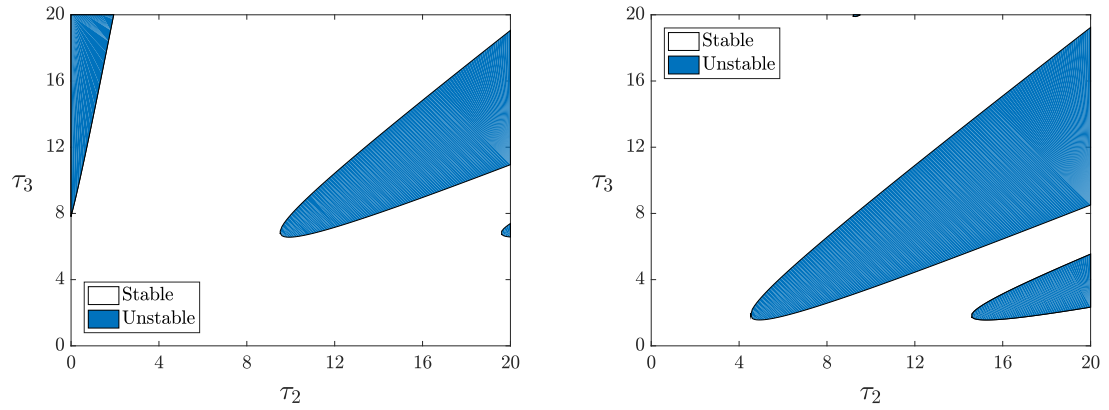
Finally, we consider the τ_1 delay parameter fixed. The limiting case $\tau_1 = 0$ results in an SIRS model which assumes that individuals instantaneously watch and share a video once they have become infected, so that there is no delay in the response to the stimuli of receiving a link to the video or seeing it appear on a Facebook News Feed. Again, bifurcation diagrams are plotted in Figure 6 for analogous parameter values as in the previous limiting cases. We observe that Figure 6 (a) looks similar to Figure 4 (a), which is not unexpected as in both plots one of the contact process delay parameters (either τ_1 or τ_2) is set to zero, and the other varied. In Figure 6 (b) where τ_1 is turned "on",



(a) $\tau_1 = 5\,, \tau_3 = 5$ (b) $\tau_1 = 6\,, \tau_3 = 2$

Figure 5: Plots of the non-dimensionalised infective density solutions I of (10)-(13), corresponding to points on either side of the Hopf bifurcation boundary for fixed $\beta = 2$ and $\tau_2 = 5$ in Figure 4 (b), displaying a stable and unstable endemic equilibrium.

we no longer see the unstable parameter region on the τ_3 axis, but the other two unstable regions close in towards the origin. Note that in both Figure 6 (a) and (b) the line $\tau_2 = \tau_3$ is stable as proven earlier.



(a) $\tau_1 = 0$ (b) $\tau_1 = 5$

Figure 6: Bifurcation diagram in the τ_2-τ_3 plane of the endemic equilibrium of (10)-(13) with fixed τ_1 and $\beta = 2$.

2.4. Reduction to a SEIR Model

We can also consider the limit of $\tau_3 \to \infty$, where recovered individuals remain immune for all time so that there is no feedback into the susceptible class. This results in an SEIR model, which can be written as

$$\frac{dS(t)}{dt} = -\beta S(t) I(t - \tau_2), \tag{53}$$

$$\frac{dE(t)}{dt} = \beta[(S(t) I(t - \tau_2) - S(t - \tau_1) I(t - \tau_1 - \tau_2)], \tag{54}$$

$$\frac{dI(t)}{dt} = \beta S(t - \tau_1) I(t - \tau_1 - \tau_2) - I(t), \tag{55}$$

$$\frac{dR(t)}{dt} = I(t). \tag{56}$$

There exists a continuum of disease-free equilibria ($S^* \in [0,1]$, $I^* = 0$) where the equilibrium value S^* is determined by the history. The non-zero eigenvalues of this disease-free equilibrium for a given $S^* \in [0,1]$ solve $\lambda + 1 - \beta S^* e^{-\lambda(\tau_1+\tau_2)} = 0$. We deduced earlier, when determining the stability of the disease-free equilibrium of the SEIRS model, that a solution λ of $\lambda + 1 - \beta e^{-\lambda(\tau_1+\tau_2)} = 0$ with $\mathrm{Re}(\lambda) > 0$ exists if and only if $\beta > 1$. Here, an eigenvalue λ satisfying $\mathrm{Re}(\lambda) > 0$ exists if and only if $\beta S^* > 1$. Hence, a disease-free equilibrium $(S^*, 0)$ is stable if $S^* \leq 1/\beta$ and unstable if $S^* > 1/\beta$. Note that if $\beta \leq 1$, the continuum of equilibria $S^* \in [0,1]$ is stable since $1/\beta \geq 1$.

3. Validation of the Models Against Data

We shall now fit our epidemic models to viewing data from YouTube to determine how well the model solutions describe the popularity evolution of viral videos. Due to the stochasticity seen in the viewing data, we cannot hope to obtain perfect fits. However, we wish to see whether our models can qualitatively describe the defining characteristics of the data. Videos which we can describe via our models must become popular through predominantly endogenous means. That is to say, they must obtain their views by individuals sharing the video to one another, rather than through exogenous means, such as being chosen to be featured on media websites. With the limited data that we have access to, it is difficult to tell exactly how a video has gained its popularity and whether exogenous effects have come into play. We shall assume that large spikes in viewing rates are caused by these outside forces which our models do not account for.

The data that we use comes from music videos which have received a large number of views. We choose to fit our models to data from these videos as, on the whole, the daily views of music videos follows some common trends, suggesting that there is an underlying deterministic process governing their viewing rates. It is mentioned in [8] that copyright protected videos such as music videos gain their popularity in an epidemic-like propagation process, while in [18], it was found that the most infectious videos tended to be music videos. Additionally, nearly all of the eighty most viewed videos on YouTube are music videos [2].

We also have to make assumptions regarding the time scale which the models remain valid for. It is reasonable to suggest that our epidemic models may be only valid for a fixed period of time from the date the video was originally uploaded, as once a video has reached a certain age, it may sustain its popularity by means other than the virtual contact process which our models describe.

We fit the infective density I in our models to the YouTube viewing data, as we assumed that only members of the infective class watch the video. We assume discontinuous infective histories $I(t) = 0$ for $t < 0$ and $I(0) = \mathcal{I}_0$ as in (15), where $\mathcal{I}_0$ is obtained from the data as the number of views the video received on the day it was uploaded. We shall use dimensionalised versions of our models and thus treat α, γ, N, τ_1, τ_2, and τ_3 as fitting parameters, where relevant. We choose N as a fitting parameter, as we do not know *a priori* the size of the population that may be interested in a video.

3.1. Data Collection and Parameter Fitting

We collect data from YouTube. For most videos, YouTube displays the viewing statistics of the video as a graph, displaying both the daily and cumulative views of a video. As YouTube does not make its raw viewing data available, to obtain the data, we took a screenshot of the daily views graph and imported it into WebPlotDigitizer [31], where the data was extracted using its "X step w/ Interpolation" algorithm in one day intervals. It should be noted that, as we are collecting data from a graph, the data we use is not the exact viewing data of each video, but we assume that it is very close to the exact raw data (which itself again is unavailable for download). Each dataset which we obtained from YouTube can be found in the Supplemental Material.

To smooth the data and reduce its noise, we take a moving average of each dataset. We wish to smooth the data enough to eliminate random fluctuations in views, but not to over-smooth it so as to make the error between the smoothed and original data too large and miss important features in the data. In Table 2, we have calculated the n-day moving average errors for the viewing data for eight videos, for different values of n. The moving averages were calculated using MATLAB's built-in `movemean` function. The error in each case was normalised with respect to both time and population size as given by

$$\text{error} = \frac{\sum_{i=1}^{m} |y_i - x_i|}{\max\limits_i x_i \times t_{\max}}, \tag{57}$$

Video	n-day Moving Average Error						
	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 100$
(a) [32]	0.0148	0.0268	0.0260	0.0278	0.0308	0.0326	0.0409
(b) [33]	0.0240	0.0406	0.0369	0.0383	0.0431	0.0443	0.0553
(c) [34]	0.0175	0.0293	0.0270	0.0290	0.0331	0.0352	0.0526
(d) [35]	0.0170	0.0281	0.0255	0.0263	0.0295	0.0311	0.0483
(e) [36]	0.0074	0.0133	0.0138	0.0164	0.0192	0.0210	0.0306
(f) [37]	0.0131	0.0231	0.0239	0.0259	0.0299	0.0328	0.0551
(g) [38]	0.0208	0.0357	0.0326	0.0349	0.0395	0.0426	0.0716
(h) [39]	0.0199	0.0371	0.0345	0.0359	0.0392	0.0407	0.0536

Table 2: n-day moving average errors for the daily viewing data of the eight YouTube videos we consider. Video labels (a)-(h) correspond to [32]-[39].

where $\{y_i\}$ are the original data points, $\{x_i\}$ are the smoothed data points, and $t_{\max}$ is the largest time value the data was collected at in days. The original data and 50-day moving averages for the eight videos we consider are shown in Figure 1 of the Supplemental Material. We choose to use a 50-day moving average, as we consider the error small enough and also the data sufficiently smooth.

In order to fit our epidemic models to the smoothed 50-day moving average daily viewing data, we need to estimate model parameters. We do so by using an ordinary least squared (OLS) method proposed in [40] and also a similar log least squares (LLS) method. We assume that a particular choice of parameters (which we shall denote by θ_0), exactly describes the viewing rates of videos, but that the m observations X_i of daily views from YouTube are affected by noise, which cause random deviations from the underlying deterministic processes described by our model. Explicitly, we have

$$X_i = I(t_i, \theta_0) + \epsilon_i, \quad i = 1, ..., n, \tag{58}$$

where $I(t_i, \theta_0)$ is the model infective density at time t_i with model parameters θ_0. The errors ϵ_i are assumed to be independent and identically distributed random variables with zero mean ($E[\epsilon_i] = 0$) and uncorrelated across time so that $\text{var}(\epsilon_i) = \sigma_0^2$, where σ_0^2 is finite. We have that the mean of the data is given by $E[X_i] = I(t_i, \theta_0)$, with longitudinally constant variance $\text{var}(X_i) = \sigma_0^2$. For a set of observation $X = (X_1, ..., X_m)$, we define the estimator θ_{OLS} as

$$\theta_{OLS} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \left[X_i - I(t_i, \theta_0)\right]^2, \tag{59}$$

where Θ represents the feasible region for parameter values of our models. For our parameters, the only condition is that they must each be non-negative. The estimator θ_{OLS} is a random variable, which minimises the distance between the data and the model prediction. In this formulation, each observation is treated as having equal importance.

We also define the estimator θ_{LLS} as

$$\theta_{LLS} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \left[\log(X_i) - \log(I(t_i, \theta_0))\right]^2, \tag{60}$$

where θ_{LLS} is a random variable which minimises the distance of the between the natural logarithm of the data and the natural logarithm of the model prediction.

If $\{x_i\}$ is a realization of $\{X_i\}$, and $\{y_i\}$ is the 50-day moving average of $\{x_i\}$, we define the cost functions

$$J_{OLS}(\theta) = \sum_{i=1}^{n} \left[y_i - I(t_i, \theta_0)\right]^2, \tag{61}$$

$$J_{LLS}(\theta) = \sum_{i=1}^{n} \left[\log(y_i) - \log(I(t_i, \theta_0))\right]^2. \tag{62}$$

Video	Model	α	γ	N	τ_1	τ_2	τ_3	Error
(c) [34]	SIR	0.0434	0.0056	5.4165×10^6	N/A	N/A	N/A	9.4765×10^{-4}
	SEIR	0.0600	0.0051	5.0547×10^6	7.6490	0	N/A	5.3258×10^{-4}
	SEIRS	0.0522	0.0056	5.3198×10^6	2.3145	2.3535	399.4400	3.0029×10^{-4}
(d) [35]	SIR	0.0520	0.0024	4.1293×10^6	N/A	N/A	N/A	5.2655×10^{-3}
	SEIR	0.0950	0.0023	4.0235×10^6	11.6786	0	N/A	4.0882×10^{-3}
	SEIRS	0.0859	0.0030	4.1636×10^6	11.1300	0.7473	364.6297	2.3059×10^{-3}

Table 3: Best fit parameter values and resulting errors of the SIR, SEIR, and SEIRS model solutions shown in Figure 7.

The solutions

$$\hat{\theta}_{OLS} = \arg\min_{\theta \in \Theta} J_{OLS}(\theta), \tag{63}$$

$$\hat{\theta}_{LLS} = \arg\min_{\theta \in \Theta} J_{LLS}(\theta), \tag{64}$$

are realizations of the random variables θ_{OLS} and θ_{LLS} respectively.

We solve the ordinary and log least squares optimization problems (63) and (64) using MATLAB's built-in `fminsearch` function, which employs the Nelder-Mead simplex algorithm as described in [41]. For both the ordinary and log least squared methods, we normalise the error of the objective function by defining

$$\text{error}_{OLS} = \frac{J_{OLS}(\hat{\theta}_{OLS})}{(\max_i y_i)^2 \times t_{\max}}, \tag{65}$$

$$\text{error}_{LLS} = \frac{J_{LLS}(\hat{\theta}_{LLS})}{(\max_i \log(y_i))^2 \times t_{\max}}. \tag{66}$$

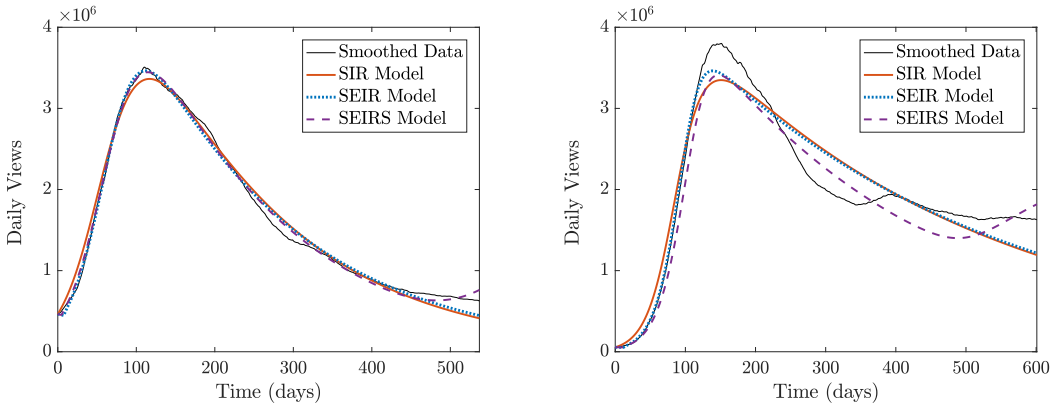
3.2. Model Comparison

In this section, we compare how well the SIR, SEIR, and SEIRS models describe the smoothed daily viewing data of two YouTube videos. In the SIR and SEIR models, recovered individuals do not become susceptible again, so there is no time-delay parameter τ_3. In the SIR model, there is no exposed class, so $\tau_1 = 0$, and we also set $\tau_2 = 0$ so that it is an ordinary differential equation system. It was noticed that in the SEIR model, τ_1 and τ_2 have nearly the exact same effect on the system. That is, in nearly all cases, only the combined value $\tau_1 + \tau_2$ (the sum of the individual and transmission latent periods) determines the effect of the time-delay on the model, so it suffices to set $\tau_2 = 0$ and treat τ_1 as the only free time-delay parameter. By making this selection, MATLAB can minimise the objective function more quickly and with greater success, thus arriving at a better fit for the SEIR model with a smaller error.

For each model, the parameters were fitted to the data from two YouTube videos using the ordinary least squares method. The best fit parameters and errors for each model, for the two videos considered, are shown in Table 3, while the smoothed data and best fit models are shown in Figure 7. We see in Table 3 that for both videos the SEIRS model fits best, as it yields the smallest error, followed by the SEIR model, and then the SIR model. Figure 7 suggests that the SEIR model may be able to describe the initial growth and decline in daily views very well. As shown in Section 2, the SEIRS model can reach an endemic equilibrium, and so it may be better suited to modelling the long term behaviour (over several years) of viewing data in situations where the daily views a video receives may rise back up after it declines.

3.3. SEIR Model Validation

We now fit the SEIR model to the smoothed daily viewing data of eight different videos using the ordinary least squares method. We wish to accurately describe the initial growth and decline in daily views for each video, and thus, we only fit the model to the data for a time interval less than the lifetime of the video (apart from "Caroline" and "Cheap Thrills", which do not show a resurgence in daily views). Hence, we neglect describing any later growth in the viewing rates the videos we consider. As in the previous section, we set $\tau_2 = 0$, so that τ_1 is the only free time-delay parameter.

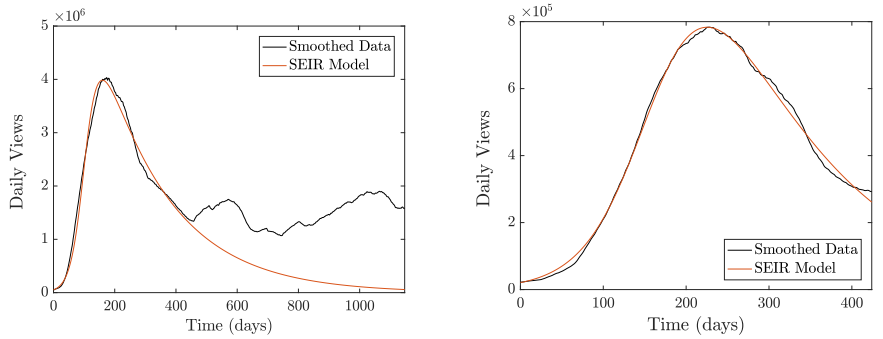


(a) Cheap Thrills (b) Faded

Figure 7: SIR, SEIR, and SEIRS models fitted to the 50-day moving averages of daily viewing data of the YouTube videos "Cheap Thrills" and "Faded" using parameter values in Table 3.

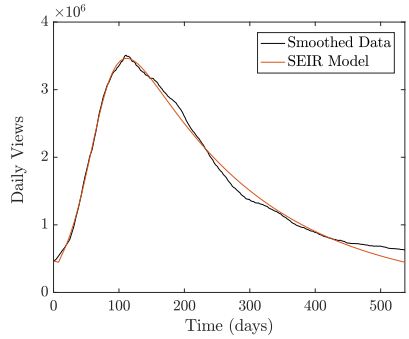
Video	α	γ	N	τ_1	$\mathcal{I}_0$	Fit Time	Error
(a)	0.0494	0.0044	5.7323×10^6	0.0000	5.1687×10^4	457	1.1008×10^{-3}
(b)	0.0354	0.0089	1.9721×10^6	1.6093	2.1828×10^4	424	2.9928×10^{-4}
(c)	0.0600	0.0051	5.0547×10^6	7.6490	4.6474×10^5	537	5.3258×10^{-4}
(d)	0.0692	0.0050	5.3679×10^6	6.3021	5.5135×10^4	295	1.0505×10^{-4}
(e)	0.0761	0.0077	1.5149×10^7	5.2707	8.5365×10^5	333	1.1328×10^{-3}
(f)	0.0575	0.0074	6.3304×10^6	1.7455	9.6519×10^5	277	1.3591×10^{-4}
(g)	0.0837	0.0147	3.2977×10^6	19.4936	3.0299×10^4	347	6.7051×10^{-4}
(h)	0.0547	0.0052	7.4397×10^6	11.5827	1.1185×10^6	278	1.7880×10^{-4}

Table 4: Best fit parameter values, fit times, and errors of the SEIR model solutions shown in Figure 8.

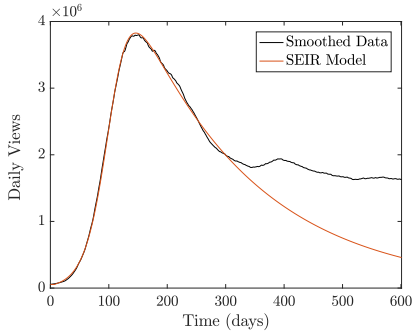


(a) All About That Bass

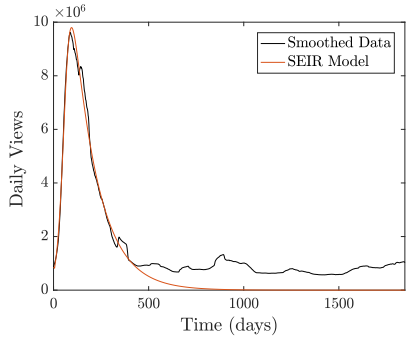
(b) Caroline



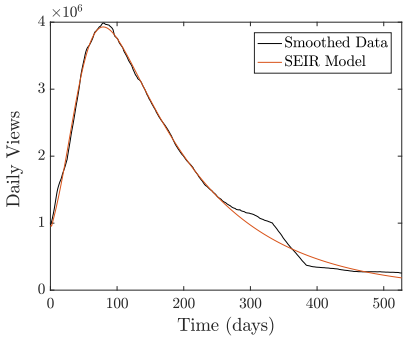
(c) Cheap Thrills



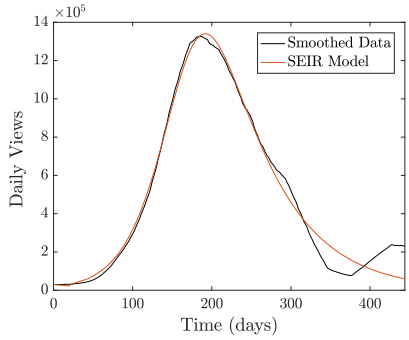
(d) Faded



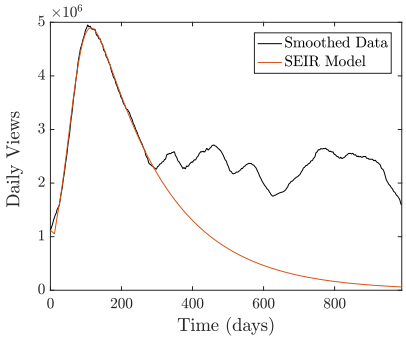
(e) Gangnam Style



(f) I Took A Pill In Ibiza

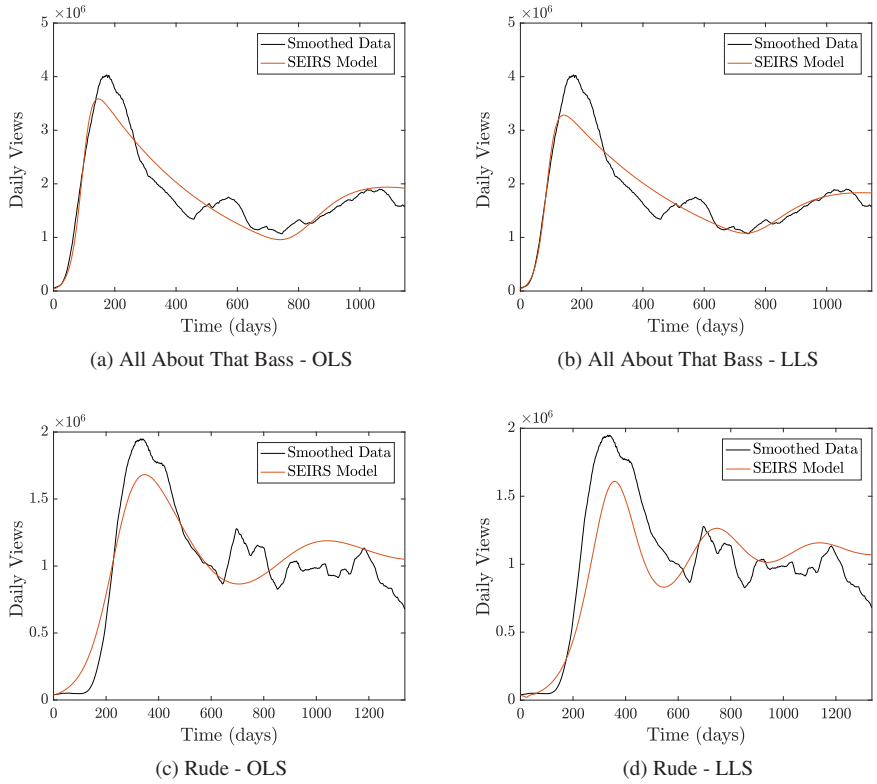


(g) OOOUUU



(h) Uptown Funk

Figure 8: SEIR model fitted to the 50-day moving averages of daily viewing data for eight different YouTube videos using parameter values in Table 4.



(a) All About That Bass - OLS

(b) All About That Bass - LLS

(c) Rude - OLS

(d) Rude - LLS

Figure 9: SEIRS model fitted to the 50-day moving average of viewing data from two YouTube videos using parameter values in Table 5. The ordinary least squares (OLS) fitting method is used in (a) and (c), while the log least squares (LLS) fitting method used in (b) and (d).

The best fit parameters and errors for each video are shown in Table 4, as well as the Fit Time, defined as the number of days the model was fitted to the data for. Note that for "Caroline" and "Cheap Thrills", the fit time is equal to the lifetime of each video. The corresponding best fit SEIR models and smoothed data are shown in Figure 8. We see that for each of these videos the model describes the initial growth and decline in daily views very well. It should be noted that the best fit recovery rate parameters, γ, are very small and are roughly an order of magnitude smaller than the infection rate parameters, α. This suggests that individuals remain infected for a long period of time, which in turn implies that these videos obtain most of their views through loyal fans. The best fit time-delay parameters range between 0 and 20 days, which appears to be perfectly reasonable as this represents the sum of the individual and transmission latent periods.

3.4. SEIRS Model Validation

We now consider the SEIRS model and fit it to the smoothed daily viewing data of the YouTube videos "All About That Bass" [32] and "Rude" [42] using both the ordinary least squares (OLS) and log least squares (LLS) methods. We use the SEIRS model to attempt to describe the long term viewing data of a video, especially when the daily views persist and do not die out. We treat N, α, γ, τ_1, τ_2, and τ_3 as free parameters. For each video, and both fitting methods, we fit the SEIRS model to the complete dataset, that is the lifetime viewing data for each video.

The best fit parameters and errors for the OLS and LLS fitting methods for each video are shown in Table 5, with the corresponding best fit models and smoothed data shown in Figure 9. We note that the OLS and LLS errors are calculated differently, so should not be directly compared to one another for the same video. We can see that for both videos, the OLS method most accurately captures the largest peak of the data, while the LLS best describes the viewing data in the long term, after the first peak. Table 5 suggests that the best fit τ_3 parameter values are very large. Assuming that the SEIRS model accurately describes the views of each video, this suggests than individuals remain

Video/Fit	α	γ	N	τ_1	τ_2	τ_3	$\bar{I}_0$	Error
(a)	0.0919	0.0024	4.2139×10^6	2.4331	8.7969	639.6338	5.1687×10^4	4.1962×10^{-3}
(b)	0.0688	0.0020	3.8302×10^6	3.3918	1.2147	629.8794	5.1687×10^4	6.6650×10^{-5}
(c)	0.0237	0.0034	2.9650×10^6	6.0019	0.8006	389.7554	3.9812×10^4	5.2655×10^{-3}
(d)	0.0653	0.0266	1.1256×10^7	21.3355	4.2290	168.2511	3.9812×10^4	3.9045×10^{-4}

Table 5: Best fit parameter values and errors for the SEIRS model solutions shown in Figure 9.

immune to each video for many months before they become susceptible again and get infected and re-watch them. This is plausible, as it can be expected that fans become tired of a video after a while, only to view it again a long time later.

3.5. Daily Views Decay Distribution

The data for each music video we have collected shows an initial growth and decay, which our SEIR model describes well. In this section, we focus on describing this decay in views after a video has reached its peak popularity, determining whether it follows either a power law or exponential distribution for large time.

We plot the natural logarithm of the smoothed daily viewing data against the natural logarithm of time in days and determine a line of best fit. If the data lies close to a straight line, this suggests that the decay rate follows a power law distribution

$$V = kt^{-a}, \tag{67}$$

where V is the number of daily views, t is time, a is the absolute value of the slope of the line of best fit, and k is a scaling constant. We will also plot the natural logarithm of the smoothed daily viewing data against time and again determine a line of best fit. In this case, if the data lies close to a straight line, the decay rate is assumed to follow an exponential distribution

$$V = ke^{-at}, \tag{68}$$

where V, t, a, and k are defined as for (67). For each of these plots, we calculate the line of best fit using MATLAB's `polyfit` function with a polynomial of degree one. We calculate the error in each case by

$$\text{error} = \frac{\sum_{i=1}^{m} |p_i - \log(y_i)|}{\max_i \log(y_i) \times t_{\max}}, \tag{69}$$

where $\{p_i\}$ is the set of points on the line of best fit.

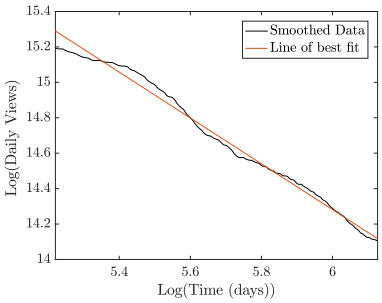
Suppose that the decay in daily views follows an exponential distribution. Assuming that the SEIR model accurately describes the data, the exponential decay rate should be close in value to $\text{Re}(\lambda_{\max})$, where $\lambda_{\max}$ is the eigenvalue with maximum real part that solves the characteristic equation

$$\lambda + \gamma - \frac{\alpha}{N} S^* e^{-\lambda(\tau_1+\tau_2)} = 0. \tag{70}$$

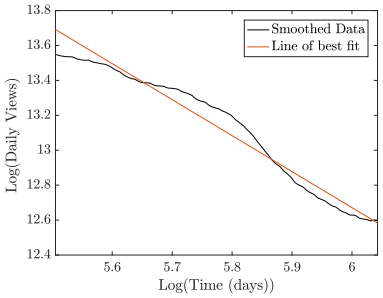
Here α, γ, N, τ_1, and τ_2 are the best fit parameters from Table 4 and the susceptible equilibrium, S^*, is calculated numerically by solving the system (53)-(56) with these parameter values.

Figures 10 and 11 show the plots for the smoothed viewing data for each of the eight videos previously modelled in Figure 8. Table 6 shows the errors and decay rates for both the power law and exponential distributions, as well as the start and end time (in days) where we judge the initial decay in daily views of each video to take place. The value $\text{Re}(\lambda_{\max})$, as calculated in equation (70) with parameters given in Table 5, is also shown and is compared with the decay rates for the exponential distribution.

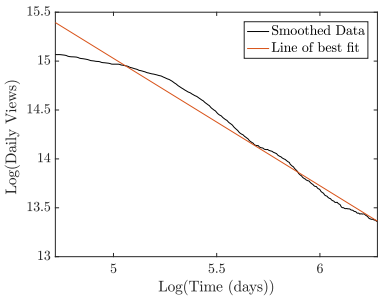
We can see in Figure 10 that none of the eight videos fit a power law distribution very well. In Figure 11 however, we observe that that the decay in views in general follows an exponential distribution more closely. In particular, the data in Figure 11 (e), (f), and (h) nearly follows a straight line when the natural logarithm of daily views is plotted against time, suggesting that their decay is exponential. We see in Figure 8 that the SEIR model does indeed model the initial decline in daily views of these videos very well. For these three videos, $\text{Re}(\lambda_{\max})$ is close to the the exponential decay rate in Figure 11 (e), (f), and (h). Note that $\text{Re}(\lambda_{\max})$ is also close to the exponential decay rate for many of the other videos, suggesting that even when the SEIR model cannot accurately describe the decay in views, the best fit model is still doing as well as possible to describe the data.



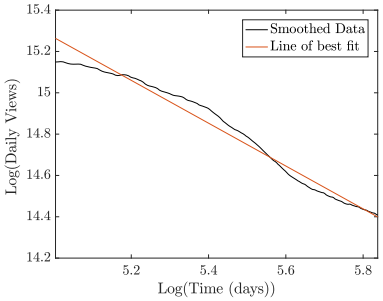
(a) All About That Bass - PL



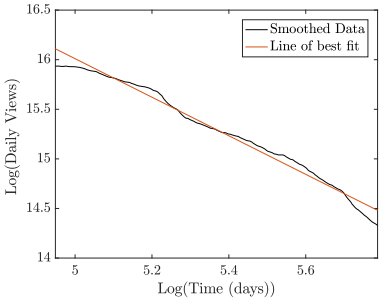
(b) Caroline - PL



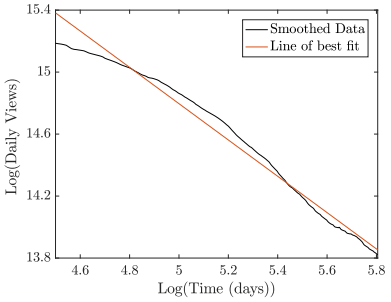
(c) Cheap Thrills - PL



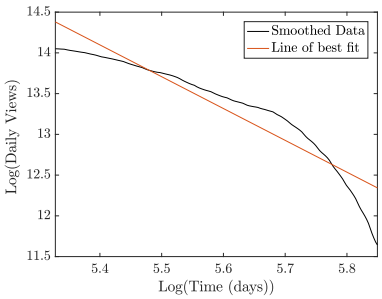
(d) Faded - PL



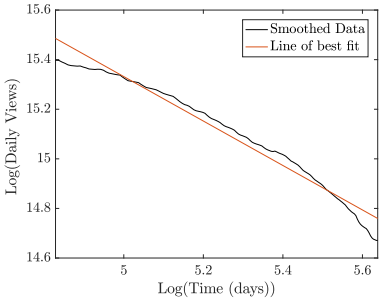
(e) Gangnam Style - PL



(f) I Took A Pill In Ibiza - PL

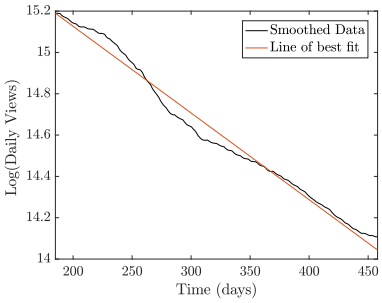


(g) OOOUUU - PL

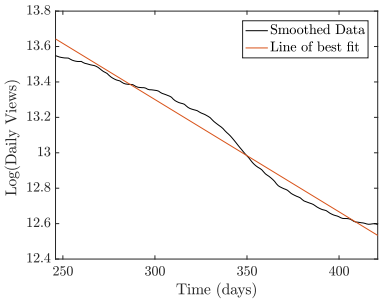


(h) Uptown Funk - PL

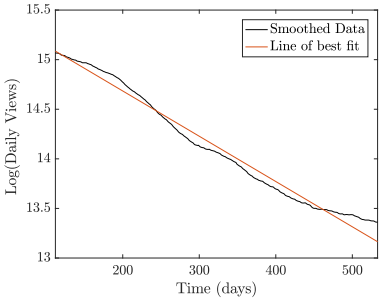
Figure 10: The natural logarithm of the decay period of the smoothed daily viewing data plotted against the natural logarithm of time to determine whether the data fits a power law (PL) distribution.



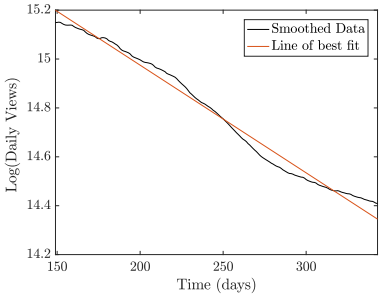
(a) All About That - Exp



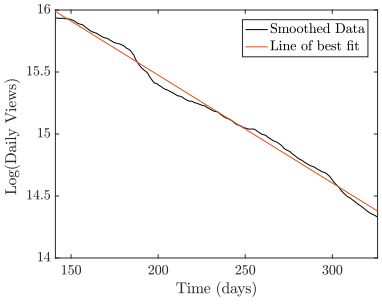
(b) Caroline - Exp



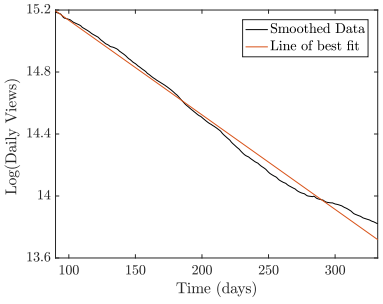
(c) Cheap Thrills - Exp



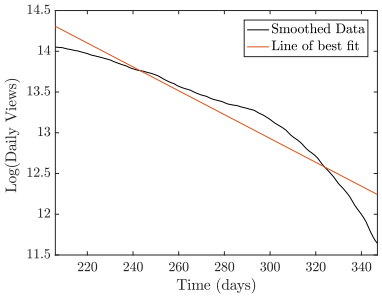
(d) Faded - Exp



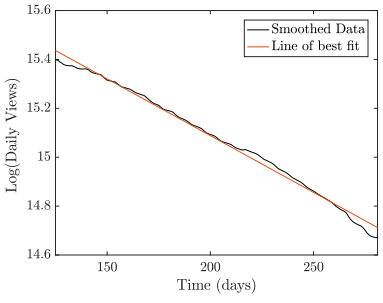
(e) Gangnam Style - Exp



(f) I Took A Pill In Ibiza - Exp



(g) OOOUUU - Exp



(h) Uptown Funk - Exp

Figure 11: The natural logarithm of the decay period of the smoothed daily viewing data plotted against time to determine whether the data fits an exponential (Exp) distribution.

Video	PL Error	PL Dec Rate	Exp Error	Exp Dec Rate	Re($\lambda_{\max}$)	Start	End
(a)	1.8107×10^{-3}	-1.2934	2.3497×10^{-3}	-4.1969×10^{-3}	-4.4224×10^{-3}	185	458
(b)	4.1114×10^{-3}	-2.0621	3.2581×10^{-3}	-6.3392×10^{-3}	-9.1799×10^{-3}	246	421
(c)	4.4054×10^{-3}	-1.3021	4.5043×10^{-3}	-4.5647×10^{-3}	-5.1359×10^{-3}	112	533
(d)	2.2167×10^{-3}	-1.0365	1.7519×10^{-3}	-4.4041×10^{-3}	-4.8909×10^{-3}	149	343
(e)	3.2258×10^{-3}	-1.9437	2.2369×10^{-3}	-8.7060×10^{-3}	-8.0720×10^{-3}	141	326
(f)	3.3430×10^{-3}	-1.1715	2.2599×10^{-3}	-6.1007×10^{-3}	-7.4174×10^{-3}	90	332
(g)	1.2491×10^{-2}	-3.9051	1.0341×10^{-2}	-1.4639×10^{-2}	-1.5358×10^{-2}	206	347
(h)	2.0846×10^{-3}	-0.8964	9.0045×10^{-4}	-4.6430×10^{-3}	-5.1769×10^{-3}	125	281

Table 6: The best fit power law (PL) and exponential (Exp) distribution decay rates and errors corresponding to plots in Figure 10 and Figure 11. We also give Re($\lambda_{\max}$), as calculated with parameters given in Table 5. The start and end time (in days) of the decay in the smoothed daily viewing data is also provided.

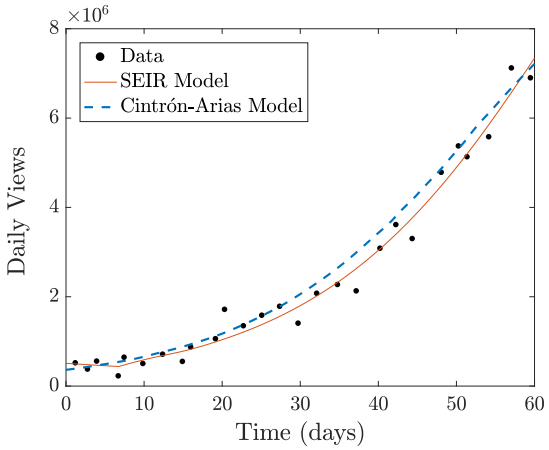


Figure 12: Comparison of our SEIR model and the model shown in [17] model when fitted to the first 60 days of daily viewing data of "Gangnam Style".

3.6. Comparison with the Work of Cintrón-Arias

Cintrón-Arias [17] proposed a variation of an ordinary differential equation SIR model adapted from [16] to describe the growth of viral video views. The primary differences between our SEIR model and that of [17] is that we include a time-delay and that the rate of change of recovered individuals in our model is linear. In Figure 4 of [17], their model was fitted to the daily views of "Gangnam Style" for the first 60 days of the video's existence. We use WebPlotDigitizer to extract the data points of the daily views of "Gangnam Style" from that figure, as well as points on their best fit curve. The points on their curve are interpolated using MATLAB's `spline` function. The error between the Cintrón-Arias model and the data points is calculated using the formula

$$\text{error} = \frac{\sum_{i=1}^{n} [x_i - C_i]^2}{(\max_i x_i)^2 \times t_{\max}}, \tag{71}$$

where $\{x_i\}$ are the daily views data points and $\{C_i\}$ are the interpolated Cintrón-Arias model infective density values. This error formula is the same as the ordinary least squares error shown in (65).

We now fit our SEIR model using the ordinary least squares method to the same data extracted from [17]. It should be noted that the "Gangnam Style" data obtained from [17] is slightly different from the data for the same video which

α	γ	N	τ_1	$\mathcal{I}_0$	SEIR Error	Cintrón-Arias Error
0.1207	0.0225	2.9146×10^7	6.7507	5.09506×10^5	6.9277×10^{-4}	1.2220×10^{-3}

Table 7: Best fit parameter values of the SEIR model solution and the SEIR and Cintrón-Arias [17] model solution errors corresponding to Figure 12.

we used earlier on in this section, which was obtained directly from YouTube. We also do not smooth the data from [17] when fitting the SEIR model, in order to have a fair comparison between our model and that of [17]. The best fit paramater values and error of our SEIR model, as well as the error for the model of [17] are given in Table 7. We see that our SEIR model performs better as it yields a smaller error. This can be seen in Figure 12, where the data and the two best fit models are plotted.

4. Discussion

We have formulated a novel SEIRS delay differential equation epidemic model to describe the popularity of viral videos in terms of views over time. As is common in epidemic models, a single threshold parameter, given by the ratio of the infection and recovery rate, determines whether the disease (or, in our context, a viral video) dies out or spreads throughout a population. We were able to validate our model using data from YouTube music videos. We demonstrated that our SEIR model describes the initial growth and decline in views of various videos accurately, and that the SEIRS model was able to capture the overall long-term viewing trends of some videos. We compared our SEIR model to a model proposed in [17] and showed that our model outperformed that of [17] in terms of error minimisation relative to data.

We recognize that the popularity of a video is influenced by a variety of factors and that we cannot hope to capture all of these using a simple model. Our model was unable to account for exogenous forces that may have caused a resurgence in popularity of the videos we studied. To remedy this, future work could be to include a self-excited point process such as a Hawkes Process [43] into our model, with self-excitations occurring when a media outlet or social influencer shares a video, in turn resulting in more viewing and sharing. One could also incorporate white noise into the model with a stochastic differential equation formulation, with the hope of appropriately describing random fluctuations in the data. When a video is uploaded onto YouTube, it becomes part of a huge network formed by millions of videos. Being able to better capture how videos interact with each other is key to being able to understanding their popularity evolution. To be able to discern this in more detail, specific data about how individuals discover videos (such as what videos link to each other) would be helpful for improving our models.

Our work, although focused on modelling the popularity of videos, could be extended to many different applications in the social sciences to describe the behaviour of consumers. For instance, a similar model could be used to describe the popularity of songs on music streaming platforms, such as Spotify, or to describe the download rates of eBooks from Amazon. Our work may indeed be applicable in a range other contexts when describing popularity trends.

References

[1] YouTube, `https://www.youtube.com/`, accessed: 2017-09-5.

[2] List of most viewed YouTube videos, `https://en.wikipedia.org/wiki/List_of_most_viewed_YouTube_videos`, accessed: 2017-09-5.

[3] T. Broxton, Y. Interian, J. Vaver, M. Wattenhofer, Catching a viral video, Journal of Intelligent Information Systems 40 (2) (2013) 241–259.

[4] L. Jiang, Y. Miao, Y. Yang, Z. Lan, A. G. Hauptmann, Viral Video Style: A Closer Look at Viral Videos on YouTube, in: Proceedings of International Conference on Multimedia Retrieval, ICMR '14, ACM, New York, NY, USA, 2014, pp. 193:193–193:200.

[5] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, A. Mahanti, Characterizing and modelling popularity of user-generated videos, Performance Evaluation 68 (11) (2011) 1037–1055.

[6] F. Figueiredo, On the Prediction of Popularity of Trends and Hits for User Generated Videos, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13, ACM, New York, NY, USA, 2013, pp. 741–746.

[7] M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahn, S. Moon, I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System, in: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, IMC '07, ACM, New York, NY, USA, 2007, pp. 1–14.

[8] F. Figueiredo, F. Benevenuto, J. M. Almeida, The Tube over Time: Characterizing Popularity Growth of Youtube Videos, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11, ACM, New York, NY, USA, 2011, pp. 745–754.

[9] T. Rodrigues, F. Benevenuto, V. Almeida, J. Almeida, M. Gonalves, Equal but different: a contextual analysis of duplicated videos on YouTube, Journal of the Brazilian Computer Society 16 (3) (2010) 201–214.

[10] R. Crane, D. Sornette, Viral, Quality, and Junk Videos on YouTube: Separating Content from Noise in an Information-Rich Environment., in: AAAI Spring Symposium: Social Information Processing, 2008, pp. 18–20.

[11] H. Pinto, J. M. Almeida, M. A. Gonalves, Using Early View Patterns to Predict the Popularity of Youtube Videos, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13, ACM, New York, NY, USA, 2013, pp. 365–374.

[12] G. Szabo, B. A. Huberman, Predicting the popularity of online content, Communications of the ACM 53 (8) (2010) 80–88.

[13] M. Kubo, K. Naruse, H. Sato, T. Matubara, The Possibility of an Epidemic Meme Analogy for Web Community Population Analysis, in: Intelligent Data Engineering and Automated Learning - IDEAL 2007, Springer, Berlin, Heidelberg, 2007, pp. 1073–1080.
[14] J. Cannarella, J. A. Spechler, Epidemiological modeling of online social network dynamics, arXiv preprint arXiv:1401.4208.
[15] C. Bauckhage, Insights into Internet Memes., in: ICWSM, 2011, pp. 42–49.
[16] D. J. Daley, D. G. Kendall, Stochastic Rumours, IMA Journal of Applied Mathematics 1 (1) (1965) 42–55.
[17] A. Cintrón-Arias, To Go Viral, arXiv:1402.3499 [physics].
[18] C. Bauckhage, F. Hadiji, K. Kersting, How Viral Are Viral Videos?, in: ICWSM, 2015, pp. 22–30.
[19] J. Gómez-Gardeñes, L. Lotero, S. N. Taraskin, F. J. Pérez-Reche, Explosive Contagion in Networks, Scientific Reports 6 (2016) 19767.
[20] K. L. Cooke, J. A. Yorke, Some equations modelling growth processes and gonorrhea epidemics, Mathematical Biosciences 16 (1) (1973) 75–101.
[21] J. M. Greenberg, F. Hoppensteadt, Asymptotic Behavior of Solutions to a Population Equation, SIAM Journal on Applied Mathematics 28 (3) (1975) 662–674.
[22] G. O. Agaba, Y. N. Kyrychko, K. B. Blyuss, Time-delayed SIS epidemic model with population awareness, Ecological Complexity 31 (2017) 50–56.
[23] H. W. Hethcote, H. W. Stech, P. van den Driessche, Stability analysis for models of diseases without immunity, Journal of Mathematical Biology 13 (2) (1981) 185–198.
[24] H. W. Hethcote, M. A. Lewis, P. van den Driessche, An epidemiological model with a delay and a nonlinear incidence rate, Journal of Mathematical Biology 27 (1) (1989) 49–64.
[25] H. W. Hethcote, H. W. Stech, P. van den Driessche, Nonlinear Oscillations in Epidemic Models, SIAM Journal on Applied Mathematics 40 (1) (1981) 1–9.
[26] W. Wang, Global behavior of an SEIRS epidemic model with time delays, Applied Mathematics Letters 15 (4) (2002) 423–428.
[27] K. L. Cooke, P. van den Driessche, Analysis of an SEIRS epidemic model with two delays, Journal of Mathematical Biology 35 (2) (1996) 240–260.
[28] S. Busenberg, K. L. Cooke, The effect of integral conditions in certain equations modelling epidemics and population growth, Journal of Mathematical Biology 10 (1) (1980) 13–32.
[29] M. Bodnar, The nonnegativity of solutions of delay differential equations, Applied Mathematics Letters 13 (6) (2000) 91–95.
[30] S. Kovács, Dynamics of an HIV/AIDS model–the effect of time delay, Applied Mathematics and Computation 188 (2) (2007) 1597–1609.
[31] WebPlotDigitizer, `http://arohatgi.info/WebPlotDigitizer/app/?`, accessed: 2017-09-5.
[32] Meghan Trainor - All About That Bass, `https://www.youtube.com/watch?v=7PCkvCPvDXk`, accessed: 2017-09-5.
[33] Aminé - Caroline, `https://www.youtube.com/watch?v=3j8ecF8Wt4E`, accessed: 2017-09-5.
[34] Sia - Cheap Thrills, `https://www.youtube.com/watch?v=nYh-n7EOtMA`, accessed: 2017-09-5.
[35] Alan Walker - Faded, `https://www.youtube.com/watch?v=60ItHLz5WEA`, accessed: 2017-09-5.
[36] PSY - Gangnam Style, `https://www.youtube.com/watch?v=9bZkp7q19f0`, accessed: 2017-09-5.
[37] Mike Posner - I Took A Pill In Ibiza, `https://www.youtube.com/watch?v=foE1mO2yM04`, accessed: 2017-09-5.
[38] Young M.A. - OOOUUU, `https://www.youtube.com/watch?v=gVf_4Ns3qLU`, accessed: 2017-09-5.
[39] Mark Ronson - Uptown Funk ft. Bruno Mars, `https://www.youtube.com/watch?v=OPf0YbXqDm0`, accessed: 2017-09-5.
[40] A. Cintrón-Arias, C. Castillo-Chávez, L. M. A. Bettencourt, A. L. Lloyd, H. T. Banks, The estimation of the effective reproductive number from disease outbreak data, Math Biosci Eng 6 (2) (2009) 261–282.
[41] J. C. Lagarias, J. A. Reeds, M. H. Wright, P. E. Wright, Convergence properties of the nelder-mead simplex method in low dimensions, SIAM Journal on optimization 9 (1) (1998) 112–147.
[42] MAGIC! - Rude, `https://www.youtube.com/watch?v=PIh2xe4jnpk`, accessed: 2017-09-5.
[43] A. G. Hawkes, D. Oakes, A cluster process representation of a self-exciting process, Journal of Applied Probability 11 (3) (1974) 493–503.