TO GO VIRAL

ARIEL CINTRÓN-ARIAS[1]
[1]DEPARTMENT OF MATHEMATICS AND STATISTICS
EAST TENNESSEE STATE UNIVERSITY
JOHNSON CITY, TN 37614

Abstract. Mathematical models are validated against empirical data, while examining potential indicators for an online video that went viral. We revisit some concepts of infectious disease modeling (e.g. reproductive number) and we comment on the role of model parameters that interplay in the spread of innovations. The dataset employed here provides strong evidence that the number of online views is governed by exponential growth patterns, explaining a common feature of viral videos.

1. Introduction

In recent years the phrase *"it went viral"* is commonly coined to denote popularity and visibility on the internet. It is almost exclusively reserved for videos, but also applies to other documents that can be navigated online.

An example of a viral video is Gangnam Style, where Korean singer Psy leads choreographies with equestrian-inspired moves. This video is enhanced by music that is remarkably catchy, along with multiple colors in scenarios induced by an odd sense of humor. Gangnam Style was uploaded to YouTube[1] on July 15, 2012 and since then it has been viewed 1.9×10^9 times. This video is without a doubt the most watched video on YouTube and is also the first to achieve views on the order of 1 billion [26].

For mathematicians this type of phenomenon is intrinsically associated with one of the most eloquent functions in mathematics: the exponential function. The same function that connects some of the best-known mathematical constants ($e^{i\pi} = -1$), and one that is typical in describing phenomena with remarkable growth patterns (such as bacterial growth).

In this article we discuss how something may *go viral*, from the point of view of mathematical modeling, with particular emphasis on exponential growth. We revisit some concepts that are traditionally invoked in models of infectious diseases, such as reproductive number and final epidemic size, and we comment on the role played by model parameters that are used to described propagation.

2. The mathematics of contagion

Infectious diseases become established, in part, due to the mobilization ability of a pathogen that finds its way from host to host. One may argue this dynamic process bears resemblance with way in which people may pass on information, from one person to another. The most basic mathematical model of epidemics, developed in 1927 by Kermack and McKendrick [13], was the inspiration, for

---

Date: February 13, 2014.

[1] `http://www.youtube.com`

Daley and Kendall in 1965 [7], to formulate a simple model for the spread of unverified information, also known as rumors or gossip. Here we derive an adaption of the latter that applies to viewership of videos posted online.

Suppose we consider a closed population of constant size, in which individuals mix homogeneously at random. In other words, there is an equal chance for individuals to encounter one another. Let us focus on a scenario where a new video is posted online and some initial viewers (i.e., early adopters) take on the quest of making it available to others. This hypothetical new video can be considered a hip innovation such as *Gangnam Style*[2], *Baby*[3], *Call Me Maybe*[4], and *Get Lucky*[5], which so far have cumulative number of views with orders of magnitude between 10^7 and 10^9. Under such circumstances the closed population can be divided into three main groups. Naive individuals are those who have never watched the video, and their size at time t is denoted by $u(t)$. Gladwell [10] re-defined mavens as individuals who are ahead of the curve in identifying fads (likely to trend) and who disseminate them with the aid of the so-called connectors. In this model mavens and connectors are lumped together in a group of size $v(t)$. Mavens (also referred to as spreaders) are those individuals who have watched the video and are actively involved in promoting viewership by means of word of mouth, email, social media (e.g. Facebook, Twitter, Google+, Reddit, Tumblr, Pinterest), etc. The third group is denominated stiflers, denoted by $w(t)$, these are individuals who stopped being interested in watching the video.

The following nonlinear ordinary differential equations model the time evolution of these three groups:

$$\frac{du}{dt} = -buv \tag{1}$$

$$\frac{dv}{dt} = buv - cv(v+w) \tag{2}$$

$$\frac{dw}{dt} = cv(v+w) \tag{3}$$

$$1 = u + v + w. \tag{4}$$

The interactions between naive and mavens give rise to more mavens, as suggested by the term buv in equations (1)–(2), where the parameter $b > 0$ denotes how such interactions take place, and is reminiscent of a rate of transmission. Moreover, the loss of mavens is assumed to occur proportional to the contacts between individuals who already know about the video, such that repeated encounters with those who have seen it may simply discourage spreaders, as they learn it is no longer an innovation. Spreaders may feel the video loses appeal once it is perceived as main stream. The term $cv(v + w)$ denotes the loss of mavens, where the parameter $c > 0$ is called the halting rate.

Figure 1 portrays numerical solutions of equations (1)–(3). As expected the naive population size is monotonically decreasing, while the stifler group increases, an implication of negativity and positivity in rates of change, clearly evidenced in equations (1) and (3), respectively. The mavens population size, $v(t)$, is displayed in Figure 1(b) and exhibits the shape of an outbreak curve.

---

[2] Total number of views to date 1,891,963,283; `http://www.youtube.com/watch?v=9bZkp7q19f0`

[3] Total number of views to date 976,385,021: `http://www.youtube.com/watch?v=kffacxfA7G4`

[4] Total number of views to date 527,469,035: `http://www.youtube.com/watch?v=fWNaR-rxAic`

[5] Total number of views to date 47,704,256: `http://www.youtube.com/watch?v=h5EofwRzit0`

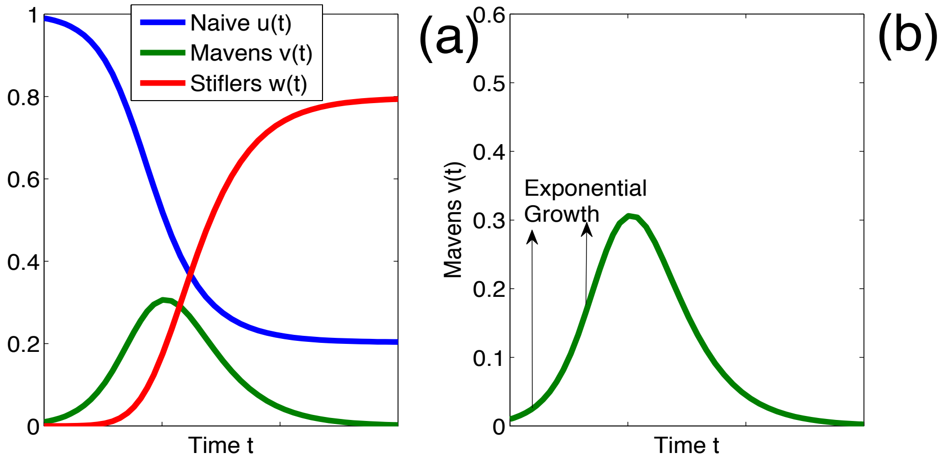

Figure 1. Panel (a): Numerical solutions of equations (1)–(3) versus time. In Panel (b) the curve $v(t)$ is displayed as a function of time t, where an exponential growth pattern can be observed, for an initial time interval. Initial conditions and parameter values: $u(0) = 0.99$, $v(0) = 0.01$, $w(0) = 0$, $b = c = 0.1$.

3. Final size

The total number of individuals that were at some point mavens can be quantified by solving a transcendental equation. First, let us divide equation (2) by equation (1) to obtain an expression for dv/du. Second, using the substitution $v + w = 1 - u$ we integrate in both sides over the interval $[t_0, t]$, while assuming $u(t_0) = 1$ and $v(t_0) = 0$, which reduces to

$$v = (-1)(u-1)\left[1 + \frac{c}{b}\right] + \frac{c}{b}\ln u. \tag{5}$$

Over the long run the population of mavens vanish while the naive population size approaches a horizontal asymptote (see Figure 1(a)). Thus, let us assume that $\lim_{t\to\infty} v(t) = 0$ and $\lim_{t\to\infty} u(t) = u_\infty$. Taking the limit as $t \to \infty$ in both sides of equation (5) yields

$$u_\infty = e^{-\mathcal{R}(1-u_\infty)}, \tag{6}$$

where $\mathcal{R} = b/c + 1$. Solutions to the transcendental equation (6) in closed form cannot be obtained. However, numerical approximations can be computed, provided that a value for the parameter $\mathcal{R}$ is known. Additionally, it is clear from equation (6) that in the extreme scenario when $\mathcal{R}$ is sufficiently large, i.e., when $\mathcal{R} \to \infty$, then $u_\infty \to 0$, implying that the entire naive population eventually transitions into becoming mavens. In the literature of infectious disease modeling, analogous quantities to $\mathcal{R}$ are typically referred to as reproductive numbers [1, 4, 8, 11, 21]. In classic epidemic models with simple dynamics these reproductive numbers serve as thresholds that separate two main qualitative regimes [11]: an outbreak taking place (whenever $\mathcal{R} > 1$) versus not having enough critical mass to kickoff an epidemic (if $\mathcal{R} < 1$).

Let $\bar{u}$ denote a solution to equation (6). In view of the conservation of mass assumption given by equation (4), it then follows that at the end of an outbreak the quantities $\bar{u}$ and $1 - \bar{u}$ denote

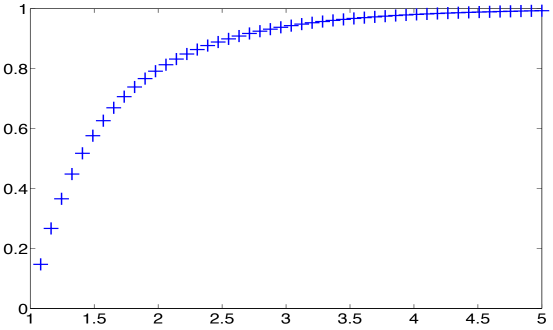

FIGURE 2. Final size $1-\bar{u}$ displayed as a function of $\mathcal{R}$, where $\bar{u}$ is a solution to the transcendental equation (6). Fifty equally spaced values of $\mathcal{R}$ are used, such that $1 \leq \mathcal{R} \leq 5$. Numerical solutions were computed by implementing a root-finding algorithm.

the fractions of those who never watched the video and those who actively did so, respectively. It is common to refer to $1-\bar{u}$ as the final size of an outbreak.

Figure 2 portrays numerical solutions of $1-\bar{u}$ versus $\mathcal{R}$. For each value of $\mathcal{R}$ a solution to equation (6) is computed by implementing an algorithm that finds roots of nonlinear functions[6]. Figure 2 illustrates what was previously noted in relation to how $\bar{u}$ changes when $\mathcal{R}$ surges. More specifically, as $\mathcal{R}$ increases in value the final size approaches unity, i.e., as $\mathcal{R} \to 5$ then $1-\bar{u} \to 1$.

The model defined by equations (1)–(4) differs from its counterpart epidemic model mainly in the lack of a dynamic threshold. Because $\mathcal{R} = b/c + 1$ always exceeds unity an outbreak curve is always guaranteed. In contrast, for a single-outbreak Susceptible-Infective-Recovered (SIR) model there is a threshold given by the ratio of transmission to recovery rates, say for example, β/γ, and if such threshold is below unity then the infective population decays to zero, while when $\beta/\gamma > 1$ then an outbreak curve exists [11]. In the next section we carry out a linearization and elaborate more on this feature.

4. EXPONENTIAL GROWTH

The rate of change of the mavens population size can be linearized to elucidate understanding on the growth patterns at initial times. More concretely, consider equation (2) with the substitution $v + w = 1 - u$, which is implied by equation (4):

$$\frac{dv}{dt} = v\left[bu - c(1-u)\right] = cv\left[\left(\frac{b}{c}+1\right)u - 1\right] = f(u,v).$$

The linearization relies on $\partial f/\partial v$ when $(u,v) \to (1,0)$ which becomes b. Thus, the linearization of dv/dt under the limit $(u,v) \to (1,0)$ is given by

$$\frac{dv}{dt} \approx bv,$$

---

[6] MATLAB (Mathworks, Inc.) built-in function `fzero` employs a combination of bisection, secant and inverse quadratic interpolation methods.

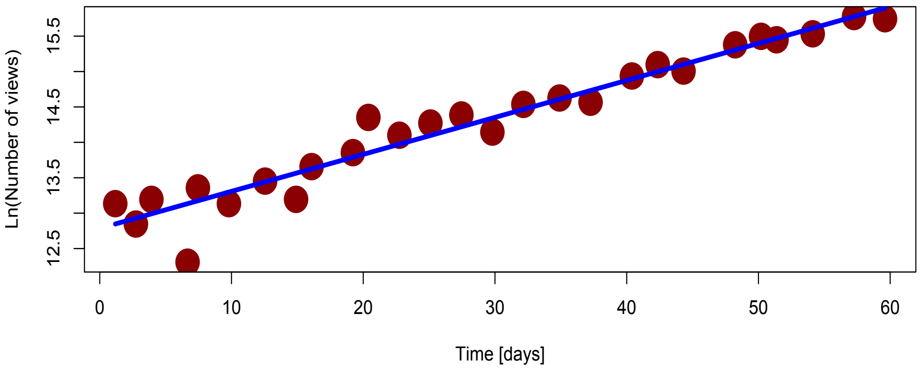

FIGURE 3. Number of views (circles) versus time and best-fit regression line for *Gangnam Style* video in YouTube. Data and regression line displayed in logarithmic scale.

which in turn is equivalent to

$$v(t) \approx v(0)e^{bt}. \tag{7}$$

Equation (7) is an approximation in the so-called *invasion limit*, i.e., $(u, v) \to (1, 0)$, where the word invasion finds its roots in the theoretical ecology literature with the context of having a species invade another species [5, 9, 14, 18, 19, 24]. Here invasion refers to a population that is initially made of all naive individuals having only an infinitesimal presence of mavens.

The equivalent calculation leading to equation (7) for an SIR model (see Appendix) yields $\beta-\gamma$, instead of b. Thus, the faith of the infective population is determined by the positivity or negativity of $\beta - \gamma$. More specifically, when $\beta/\gamma > 1$ the infective population initially grows exponentially, otherwise it is destined to exponential decay. The fact that $b > 0$ implies a guaranteed exponential take off is one the main distinctions between a single-outbreak SIR model and the model defined by equations (1)–(4).

Let us suppose the number of mavens at time t is given by $v(t) \approx x(t) = Ae^{bt}$, where the parameter A denotes the initial condition, i.e., $A = v(0)$. Whenever a temporal dataset is available then it is straightforward to estimate the parameters A and b. It is common to transform the data and model using a natural logarithm transformation. The model becomes

$$y(t) = \ln x(t) = \alpha + bt,$$

where $\alpha = \ln A$. A log-transformed dataset $\{(t_1, \ln x_1), (t_2, \ln x_2), \dots, (t_n, \ln x_n)\}$ can used to compute point-estimates, denoted as $\hat{\alpha}$ and $\hat{b}$, of the parameters α and b while applying linear regression formulas (see [22] and Appendix for details).

Viewership data for the music video *Gangnam Style*, that is, data on the number of views versus time, were available in the blog known as YouTube Trends [26]. These temporal observations are displayed in Figure 3 (circles) in logarithmic scale. Moreover, the best-fit regression line $y(t) = \hat{\alpha}+\hat{b}t$

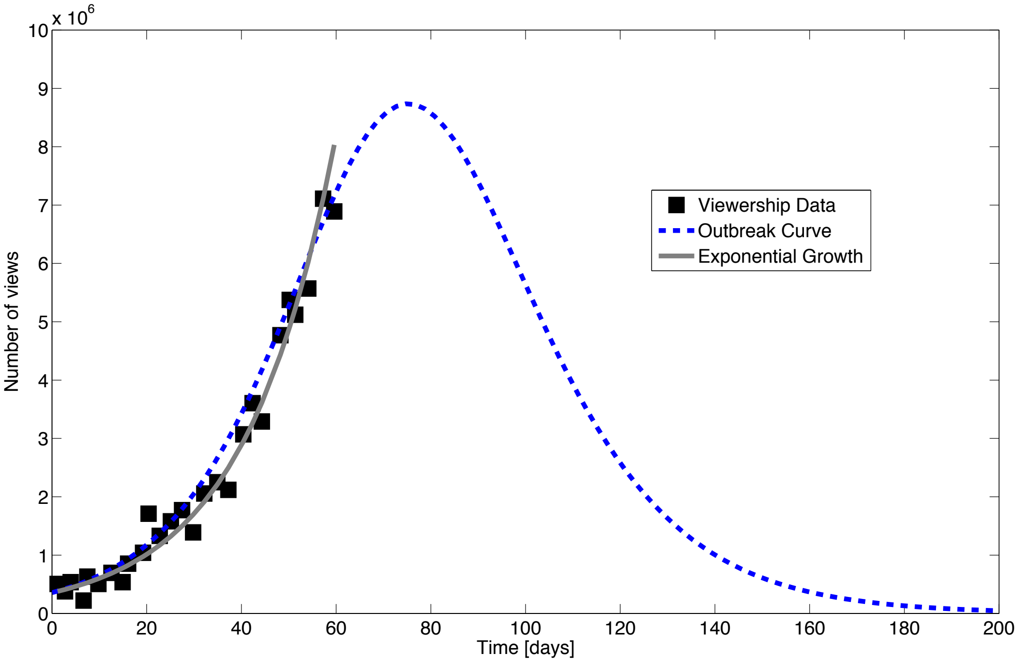

FIGURE 4. Longitudinal observations on the number of views of *Gangnam Style* are depicted with squares. Best-fit exponential model, defined in equation (7), is displayed as a gray solid curve. Best-fit outbreak curve, $v(t)$ in equations (1)–(3) scaled by a total population size $N = 7.0 \times 10^7$, appears as a blue dashed curve. Initial conditions: $v(0) = 3.6 \times 10^5/N$, $u(0) = 1 - v(0)$, $w(0) = 0$. Model parameter estimates and standard errors are given in the text.

also appears in Figure 3. Parameter estimates within one standard error[7] are:

$$\hat{b} = 0.0522 \pm 0.0026 \tag{8}$$

$$\hat{\alpha} = 12.7856 \pm 0.0883 \tag{9}$$

The Pearson correlation coefficient (see [22] and Appendix) for the data displayed in Figure 3 equals 0.9700, implying a strong correlation between the $\ln x_i$'s and t_i's. Moreover, the coefficient of determination is then $0.9700^2 = 0.9409$, meaning that 94% of the variability in the data can be explained by a linear model [22], in other words, a linear model is justifiably well suited. The strong correlation in the logarithmic scale translates into solid evidence of exponential growth patterns for the viewership data, in the particular case considered here. In other words, there is strong evidence that the number of views for *Gangnam Style* obeys exponential growth. It is precisely this feature what is commonly associated with the phrase *"it went viral"*.

5. THE MOST INFLUENTIAL PARAMETER

The same dataset that was used in the previous section, for a linear regression analysis, is now employed to validate the mathematical model defined by equations (1)–(3). This model is validated against empirical data by applying ordinary least squares (OLS) methods for inverse problems (see [3, 6] and references therein for additional details).

The model parameters are the transmission rate b and the halting rate c, and their OLS estimates are denoted by $\hat{b}$ and $\hat{c}$. The numerical solution to equations (1)–(3) obtained with the parameter estimates is usually denominated the best-fit solution. Figure 4 displays the temporal data (squares) and the best-fit solution (dashed curve). The OLS parameter estimates within one standard error are[8]

$$\hat{b} = 0.0623 \pm 0.0020 \tag{10}$$

$$\hat{c} = 0.2102 \pm 0.0413. \tag{11}$$

Because data is given in raw quantities and not in percentages or fractions, as the mathematical model was formulated (see equation (4)), it was necessary to also estimate a total population size, which was used as a scaling factor. For this quantity, only a point estimate (without uncertainty bounds) was computed, namely, $N = 7.0359 \times 10^7$.

For the sake of comparison, the exponential model defined in equation (7) was also fitted to the longitudinal dataset. The best-fit exponential solution to equation (7) appears displayed in Figure 4 as a solid curve. The exponential model has two parameters: $A = v(0)$ and b. Estimates plus minus one standard error for these parameters are given by

$$\hat{A} = 4.3590 \times 10^5 \pm 4.1370 \times 10^4 \tag{12}$$

$$\hat{b} = 4.7750 \times 10^{-2} \pm 1.8510 \times 10^{-3}. \tag{13}$$

Whenever mathematical models are validated against empirical data, it is customary to address the role that parameters play in state variables. In other words, how variability in model parameters manifest in the output of the model. One can rule dependence on model parameters by computing sensitivity functions. For example, let us consider equation (7), where we write $x(t) = Ae^{bt}$. The partial derivatives $\partial x/\partial A$ and $\partial x/\partial b$ are denominated traditional sensitivity functions [2]. They are time functions addressing the rate of change with respect to parameter variability. A re-scaled version of these functions is useful in circumventing issues of magnitude and scale. Define

$$\nu_A(t) = \frac{A}{x(t)}\frac{\partial x}{\partial A}(t) \equiv 1$$

and

$$\nu_b(t) = \frac{b}{x(t)}\frac{\partial x}{\partial b}(t) = bt.$$

The functions $\nu_A(t)$ and $\nu_b(t)$ are called relative sensitivity functions. For the exponential growth model, equation (7), both of these functions are linear, one of them with zero slope while the other one has slope $b > 0$. On one hand, when the relative sensitivity functions approach zero one concludes the rate of change with respect to that parameter vanishes, meaning that no dependence on that parameter exists. On the other hand, whenever the relative sensitivity functions are away from zero, then monotonic dependence on the parameters can be ruled in consistency with the

---

[7] The bulit-in function `nls` of the open-source statistical package R was implemented to compute standard errors.

[8] Numerical solutions of OLS inverse problems can obtained with implementations in MATLAB that include the Global Optimization toolbox. Built-in functions `ga` and `patternsearch` were employed here.

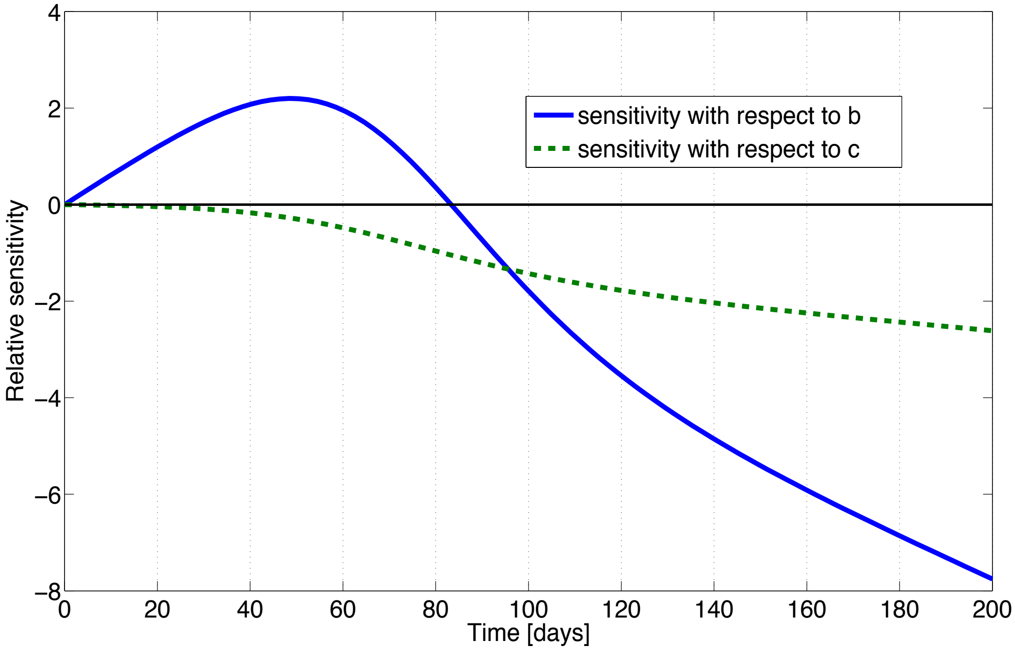

FIGURE 5. Relative sensitivity functions with respect to model parameters b and c, versus time.

sign: negativity would imply decreasing dependence, while positivity would suggest increasing dependence. More specifically, the exponential growth model has $\nu_A(t) > 0$ and $\nu_b(t) > 0$ implying that $x(t)$ increases when A and b increase. Furthermore, if $t > 1/b$ then $\nu_b(t) > \nu_A(t)$, meaning that b is the most influential parameter in $x(t)$.

Relative sensitivity functions cannot be computed in analytic form for the state variables of equations (1)–(3). Because this is a nonlinear system for which closed form solutions are not available. Instead numerical solutions can be computed by using the parameter estimates to solve an extended system that includes auxiliary equations. The latter equations are known as the *forward sensitivity equations*, where the state variables of the original system act as time-dependent coefficients for a linear (auxiliary) system, with unknowns equal to the partial derivatives of the state variables with respect to model parameters. For additional details in numerical solutions of forward sensitivity equations, the reader is referred to [2, 3, 6] and references therein.

Let us define the relative sensitivity functions for the mavens population size $v(t)$, from equations (1)–(3), as follows:

$$\phi_b(t) = \frac{b}{v(t)} \frac{\partial v}{\partial b}(t)$$

and

$$\phi_c(t) = \frac{c}{v(t)} \frac{\partial v}{\partial c},$$

where $\partial v/\partial b$ and $\partial v/\partial c$ are computed numerically. Figure 5 displays $\phi_b(t)$ (solid curve) and $\phi_c(t)$ (dashed curve) versus t. It is clear from Figure 5 that for $t \in [0, 30]$ we have that $\phi_c(t) \approx 0$, implying the halting rate c does not influence $v(t)$ over this initial time interval. The model parameters play an expected role for $30 \leq t \leq 83$ where $v(t)$ increases with the transmission rate b (because $\phi_b(t) > 0$) and $v(t)$ decreases with the halting rate c (due to $\phi_c(t) < 0$). Surprisingly, the role of parameter b is reversed when $t > 83$, in the sense that $\phi_b(t) < 0$. Furthermore, towards the end of the outbreak, as $t \to 200$, we have that $\phi_b(t) < \phi_c(t) < 0$, suggesting that b is more influential than c in the ending phase, because $\phi_b(t)$ is farthest away from zero. In summary, the transmission rate b is the most influential parameter.

We also note that having $\phi_c(t) \approx 0$ and $\phi_b(t) > 0$ for $0 \leq t \leq 30$, also confirms that during this time interval the curve $v(t)$ is only influenced by the transmission parameter b, but it is also heavily dominated by exponential growth. The latter is confirmed in Figure 4 by comparing the best-fit curves (solid versus dashed), where reasonable agreement with the pattern displayed by the longitudinal data is observed for $t \in [0, 30]$. Temporal patterns of exponential growth are, to the best of our knowledge, footprints of going viral.

6. FINAL REMARKS

The website known as YouTube was launched in early 2005 with the innovative goal of providing a user-friendly means to share videos. Over nearly a decade, it is clear this website is revolutionizing online video streaming. A technology that is replacing video stores by online rentals (e.g. Amazon Instant Video, Netflix).

In the context of our contemporary era of social networking sites, powered by digital media, traditional terms in epidemiology of communicable diseases (such as transmissibility, reproductive number, final size, etc) find new meanings, when viewed through the lens of marketing and digital strategy. In reverse sense of what we would aim with infectious diseases (most efforts target eradication), digital media would intend to establish trends. More importantly, it is of great interest to determine what factors (e.g. age group, demographics, education level, gender, etc) would favor the likelihood of establishment. Mathematical, statistical and computational methods (e.g. data science and predictive modeling) are merging together to address questions in the latter context, however the challenge remains paramount.

In this paper, we have revisited a mathematical model proposed by Daley and Kendall [7], in the context of rumor propagation by word of mouth. Most of the active development of theory centered around this model relates to stochastic modeling [12, 15, 16, 17, 20, 23, 25, 27], including explorations of social landscapes equivalent to complex heterogeneous networks. However, to the best of our knowledge, most of this theoretical work remains yet to be validated against empirical data.

In this study we were successful in validating Daley-Kendall's model against longitudinal observations. Nevertheless, the parameter estimates must be considered with a degree a caution, because they only constitute first order approximations conditional on the dataset. The longitudinal counts on the number of views of Gangnam Style are aggregate, as they inevitably include repeated counts from individuals who watched the video more than once. The model parameter estimates and their uncertainty bounds, can only be accepted by making the simplifying assumption that the viewership data approximate the actual number of mavens at each time point.

More important than the estimated parameter values is how the process of dissemination is readily verified with simple mathematical models. A contagion model, equations (1)–(4), can successfully describe the surge and fall of mavens, as it is clearly seen in Figure 4. A better dataset to validate a contagion model would include observations at the peak and in the downfall phase of the outbreak curve, not just at the beginning.

A central point for our analysis here is to articulate in mathematical terms what it means when a video *"went viral"*. Figures 3 & 4 confirm that to go viral is the equivalent of exponential growth. More specifically, we find that strong positive correlation (with a Pearson coefficient of 0.97) in logarithmic scale is an indicator of exponentially growing patterns. The transmission rate b denotes the ability to promote viewership, and it is the most influential parameter for both the contagion and the exponential models. One of the features of exponentially growing processes is that they can easily climb up orders of magnitude within a short time window, a feature that is well documented for viral videos. For example, Gangnam Style has views jumping from order 10^5 (on day 0) to 10^6 (on day 20).

On a closing note, mathematicians can only hope that some their publications in peer-reviewed journals *"went viral"*. Perhaps the closest metric of such degree of visibility is the number of citations. This author would be thrilled if any of his publications are ever cited with order of magnitude 10^2. But yet it is still a long way to go, because his most cited paper has to date 71 citations (according to Web of Knowledge[9]), and when considered over a period of 8 years, it leaves much to be desired.

Acknowledgments

The author thanks Tara Gancos Crawford, Rachel Fovargue, Amanda Traud, Jian Zhang, Virginia Ingram, and the National Institute for Mathematical and Biological Synthesis, for helpful discussions on this topic.

References

[1] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, New York, 1992.

[2] H. T. Banks, S. Dediu and S. L. Ernstberger, *Sensitivity functions and their uses in inverse problems*, J. Inverse Ill-Posed Probl. **15** (2007), 683–708.

[3] H.T. Banks and H.T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, Taylor & Francis, Boca Raton, 2009.

[4] F. Brauer and C. Castillo-Chávez, *Mathematical Models in Population Biology and Epidemiology*, 2nd Edition, Springer, New York, 2011.

[5] H. Caswell, *Matrix Population Models: Construction, Analysis, and Interpretation*, Sinauer Associates, Sunderland, 2001.

[6] 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** (2009), 261–282.

[7] D. J. Daley and D. G. Kendall, *Stochastic rumours*, J. Inst. Math. Appl. **1** (1965), 42-55.

[8] O. Diekmann and J.A.P Heesterbeek, *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*, John Wiley, Chichester, 2000.

[9] L. Edelstein-Keshet, *Mathematical Models in Biology*, Society for Industrial and Applied Mathematics, Philadelphia, 1988.

[10] M. Gladwell, *The Tipping Point: How Little Things Can Make a Big Difference*, Bay Back Books, Boston, 2002.

[11] H. Hethcote, *The mathematics of infectious diseases*, SIAM Rev. **42** (2000), 599-653.

[12] V. Isham, S. Harden, M. Nevokee, *Stochastic epidemics and rumours on finite random networks*, Phys. A **389** (2010), 561–576.

[13] W. Kermack and A. McKendrick, *Contributions to the mathematical theory of epidemics I*, Proc. R. Soc. Lon. Ser. A Math. Phys. Eng. Sci. **115** (1927), 700–721.

---

[9] `http://wokinfo.com`

[14] M. Kot, *Elements of Mathematical Ecology*, Cambridge University Press, Cambridge, 2001.

[15] C. Lefevre and P. Picard, *Distribution of the final extent of a rumour process*, J. Appl. Probab. **31** (1994), 244–249.

[16] D. Maki and M. Thompson, *Mathematical Models and Applications*, Prentice-Hall, Englewood Cliffs, 1973.

[17] Y. Moreno, M. Nekovee, A.F. Pacheco, *Dynamics of rumor spreading in complex networks*, Phys. Rev. E **69** (2004), 066130.

[18] J. Murray, *Mathematical Biology I: An Introduction*, Springer, New York, 2002.

[19] S. P. Otto and T. Day, *A Biologist's Guide to Mathematical Modeling in Ecology and Evolution*, Princeton University Press, Princeton, 2011.

[20] C. E. M. Pearce, *The exact solution of the general stochastic rumour*, Math. Comput. Modelling **31** (2000), 289–298.

[21] L. Sattenspiel and A. Lloyd, *The Geographic Spread of Infectious Diseases: Models and Applications*, Princeton University Press, Princeton, 2009.

[22] E. Seier and K. H. Joplin, *Introduction to Statistics in a Biological Context*, CreateSpace, Charleston, 2011.

[23] A. Sudbury, *The proportion of the population never hearing a rumour*, J. Appl. Probab. **22** (1985), 443–446.

[24] H. R. Thieme, *Mathematics in Population Biology*, Princeton University Press, Princeton, 2003.

[25] R. Watson, *On the size of a rumour*, Stochastic Process. Appl. **27** (1987), 141–149.

[26] YouTube Trends, *Gangnam Style vs Call Me Maybe: A Popularity Comparison*, website: `http://youtube-trends.blogspot.com/2012/09/gangnam-style-vs-call-me-maybe.html`. Accessed on January 23, 2014.

[27] D. H. Zanette, *Dynamics of rumor propagation on small-world networks*, Phys. Rev. E **65** (2002), 041908.

7. Appendix

SIR Model. A single outbreak Susceptible-Infective-Recovered model is defined by the nonlinear system of equations (see [11] and references therein for additional details):

$$\begin{aligned}\frac{ds}{dt} &= -\beta si\\ \frac{di}{dt} &= \beta si - \gamma i\\ \frac{dr}{dt} &= \gamma i\\ 1 &= s + i + r,\end{aligned}$$

where the epidemic parameters are β and γ. A linearization of the infectives equation, around $s = 1$ and $i = 0$, yields

$$\frac{di}{dt} \approx (\beta - \gamma)i,$$

which implies,

$$i(t) \approx i(0)e^{(\beta-\gamma)t}.$$

Linear Regression. Given the observations $(t_1, \ln x_1), (t_2, \ln x_2), \ldots, (t_n, \ln x_n)$, then point-estimates for the parameters of the linear model

$$y(t) = \ln x(t) = \alpha + bt,$$

are computed as follows [22]:

$$\begin{aligned}\hat{b} &= \frac{\sum_{i=1}^{n} (t_i - \bar{t}\ )(\ln x_i - \bar{y})}{\sum_{i=1}^{n} (t_i - \bar{t}\ )^2}\\ \hat{\alpha} &= \bar{y} - \hat{b}\bar{t}\end{aligned}$$

where

$$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

and

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i.$$

The sample standard deviations for the t_i's and $\ln x_i$'s are denoted by s_t and s_y, respectively, and are computed using the sample averages $\bar{t}$ and $\bar{y}$, namely [22]:

$$s_t = \sqrt{\frac{\sum_{i=1}^{n}(t_i - \bar{t})^2}{n-1}} \qquad \text{and} \qquad s_y = \sqrt{\frac{\sum_{i=1}^{n}(\ln x_i - \bar{y})^2}{n-1}}.$$

The Pearson correlation coefficient is defined in terms of the sample average and sample standard deviation [22]:

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{t_i - \bar{t}}{s_t}\right)\left(\frac{\ln x_i - \bar{y}}{s_y}\right).$$