Going Multi-viral: Synthedemic Modelling of Internet-based Spreading Phenomena

Marily Nika
Department of Computing
Imperial College London
SW7 2AZ, UK
marily@imperial.ac.uk

Thomas Wilding
Department of Computing
Imperial College London
SW7 2AZ, UK
tw610@imperial.ac.uk

Dieter Fiems
Ghent University, Dep. TELIN
St-Pietersnieuwstraat 41
9000 Gent, Belgium
dieter.fiems@UGent.be

Koen De Turck
Ghent University, Dep. TELIN
St-Pietersnieuwstraat 41
9000 Gent, Belgium
koen.deturck@UGent.be

William J. Knottenbelt
Department of Computing
Imperial College London
SW7 2AZ, UK
wjk@imperial.ac.uk

ABSTRACT

Epidemics of a biological and technological nature pervade modern life. For centuries, scientific research focused on biological epidemics, with simple compartmental epidemiological models emerging as the dominant explanatory paradigm. Yet there has been limited translation of this effort to explain internet-based spreading phenomena. Indeed, single-epidemic models are inadequate to explain the multimodal nature of complex phenomena. In this paper we propose a novel paradigm for modelling internet-based spreading phenomena based on the composition of multiple compartmental epidemiological models. Our approach is inspired by Fourier analysis, but rather than trigonometric wave forms, our components are compartmental epidemiological models. We show results on simulated multiple epidemic data, swine flu data and BitTorrent downloads of a popular music artist. Our technique can characterise these multimodal data sets utilising a parsimonous number of subepidemic models.

Keywords

Epidemiology, synthedemic model, spreading phenomena

1. INTRODUCTION

Human existence has always been driven by interactions between humans and between humans and their environment. Spreading processes of various kinds arise as an inevitable consequence of these interactions. Where the spreading is rapid and widespread, the resulting outbreak is termed an *epidemic*. Epidemics occur in and impact on almost every domain from biology (e.g. infectious diseases) to technology (e.g. computer viruses and social networks). For this reason, the study of epidemics and spreading processes has been a vital scientific endeavour throughout history.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
VALUETOOLS 2014, December 09-11, Bratislava, Slovakia
Copyright © 2015 ICST 978-1-63190-057-0
DOI 10.4108/icst.valuetools.2014.258221

Since physiological well-being is one of the most basic human needs [15], it is natural that the study of spreading processes focused for many centuries on disease propagation and biological epidemics in populations. The last two centuries witnessed the emergence of the evidence–based scientific study of disease we know today as *epidemiology*. Over the same time period, increased industrialization, mass transit and technological developments have increased not only the potential for the activation of a broad class of spreading processes but also the rates of transmission, increasing the likelihood that they manifest themselves as epidemics.

It has long been recognised that it is not only diseases that are subject to spreading processes. Amongst others, Dawkins has suggested the theory of memes i.e. ideas that spread like "mind-viruses" [8]. A key assumption of our present research is that there are many similarities between the way diseases spread and the way internet-based spreading mechanisms – such as tweeting and sharing of online content – operate. That is, an outbreak of interest starts with a few susceptible individuals who are exposed to an originating event and some become "infected". These individuals then interact with others, passing on the "infection". Eventually the infected individuals recover/lose interest and the outbreak dies out. We consequently adopt epidemiological models to describe the dynamics of internet–based phenomena.

A model of a *single* epidemic is inadequate to characterise the multimodality that emerges from many complex internet-based spreading phenomena. We speculate that this is because mono-epidemic-based modelling efforts cannot account for the potential influence of multiple underlying spreading mechanisms, each of which may initiate at a different time. Consider for example YouTube video views. Views may be due to sharing of the content on social media platforms such as Facebook and Twitter, links on other websites, the content being featured and/or recommended in a news article or by YouTube itself, notifications to channel subscribers etc. Ideally we require a model that is able to adapt to the sudden activation of any of these mechanisms, rapidly updating itself to enable near-term predictions of reasonable quality, without detailed knowledge of the underlying spreading mechanisms involved.

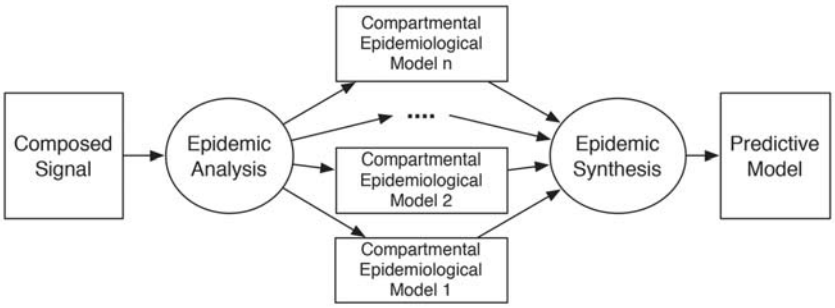

Figure 1: The proposed synthedemic modelling and prediction framework.

We propose a novel modelling and prediction framework based on the analysis and synthesis of multiple epidemic models as shown in Figure 1. Given a composed signal which is presumed to represent the aggregated observable manifestation of multiple underlying epidemics, the framework breaks down the incoming signal into its fundamental components and selects the disease spreading models that best explain each component. These models are resynthesized in order to predict the future evolution of the signal.

Our approach is inspired by Fourier analysis, but instead of trigonometric wave forms our components are compartmental epidemiological models. There are several challenges inherent in our approach, not least in determining the number of epidemics to be fitted, and in selecting appropriate epidemiological models and parameters for each component.

The remainder of this paper is organised as follows. Section 2 presents relevant background of the theory and analysis of disease propagation, and an overview of contemporary epidemiology-based social network analysis. Section 3 lays out our synthedemic decomposition algorithm which is implemented in a prototype version of our framework. Section 4 presents our case study results on simulated multiple epidemic data, swine flu data and BitTorrent download activity from artists that went viral in recent years. Lastly, Section 5 concludes and considers avenues for future work.

2. BACKGROUND

Disease Propagation Theory and Predictive Analytics

In ancient times, and throughout the Dark and Middle ages, the predominant explanations for disease propagation included the supernatural, superstition and miasma theory, which held that diseases were caused by "bad air". A notable exception was Hippocrates who correctly identified the role of human behaviours and environmental factors [1].

Progress towards a more scientific and data–based approach began to be made from 1600 onwards with the collection of the first public health statistics, by John Graunt (1620–1674) [12] and others. One of the most famous studies now regarded as the foundation of this discipline was by John Snow of the 1854 London Cholera epidemic [20] in which he identified a particular water pump on Broad Street as the likely source of the outbreak.

Predictive mathematical models for epidemics were relatively slow to develop, despite their tremendous utility in understanding, managing and forecasting the progression of epidemics. One of the earliest was in 1766 by Daniel Bernoulli who carried out a study of the effects of smallpox vaccination. However, arguably the most significant breakthrough in this context was that of compartmental disease models based on Ordinary Differential Equations (ODEs), as proposed by Kermack and McKendrick in 1927 [14].

A variety of compartmental disease models are currently used in practice [22]. The most well-known of these, the Susceptible-Infected-Recovered (SIR) model features a closed population of individuals divided into three evolving subpopulations: $S(t)$ tracks the number of individuals who are susceptible to become infected by the disease at time t, $I(t)$ tracks the number of individuals who are infected by the disease with rate β and $R(t)$ tracks the number of individuals who have recovered from the disease at rate γ.

Epidemiology-based Social Network Analysis

Goffman and Newill were the first to bring a social context to epidemiology, with their mathematical model for the spreading of rumours [13]. The rise of the internet, particularly search engines and Online Social Networks (OSNs), led to two classes of studies: those designed to augment conventional epidemiology (e.g. [6, 9]), and those applying epidemiological or diffusion process principles to model the dissemination of information (e.g. [2, 3, 5, 11, 16]). The former includes detection of real physical disease outbreaks by assuming a relationship between online searches and the real number of infected individuals [9]. The latter includes the work of Tweedle and Smith, who applied an SIR-inspired model to pop star Justin Bieber's popularity based on *Google Trends* data [21]. Very recently, Coviello et al. published a controversial study which measured the contagion of emotional expression amongst Facebook users [7].

A recent study explored the potential for epidemiology to explain certain outbreaks of internet-based information spreading [18]. The authors were able to progressively fit and to parameterise simple epidemiological models from single data traces of BitTorrent downloads and YouTube views. Subsequently they investigated confidence intervals on the outbreak parameter values as the outbreak unfolded over time [19]. Another work explored the dynamics of interacting epidemics in multiple overlapping populations [17].

3. METHODOLOGY

The synthedemic[1] methodology is designed to fit composed epidemic models to outbreak datasets that are regularly augmented with new observations (so as to facilitate potentially real-time operation). We start with a small truncated data set and at each step we add one new data point to the truncated dataset until we reach the end of the time frame to be considered. Initially we start by fitting no epidemics, and dynamically incorporate more epidemics when it becomes necessary to improve the fit.

It is clearly important to chose a set of compartmental model types which are appropriate for the context within which the synthedemic framework is deployed. It transpires that followers of online phenomena have noticed that there appear to be two types of content diffusion: *growth*, characterised by organic spreading of content in communities (initially by influencers), and *spike*, which represents a sudden "explosion" of sharing activity sparked by some (mass-media) event that is then followed by a gradual decay [10]. Here we propose to model the former by an SIR process, and the latter by an IR (Infected-Recovered) process consisting of an initial impulse followed by exponential decay. That is,

- An SIR epidemic starting at time t_0 is characterised by the initial number of infected individuals I_0, the initial number of susceptible individuals S_0, the initial number of recovered individuals, the infection rate β and the recovery rate γ. The SIR model dynamics are:

$$\begin{aligned} S'(t) &= -\beta I(t) S(t), \\ I'(t) &= \beta I(t) S(t) - \gamma I(t) \\ R'(t) &= \gamma I(t) \end{aligned}$$

  for $t > t_0$ with $[S(t_0), I(t_0), R(t_0)] = [S_0, I_0, R_0]$ and with $I(t) = R(t) = S(t) = 0$ for $t < t_0$.

- An IR (spike) epidemic starting at time t_0 is characterised by the initial number of infected individuals I_0 and the decay rate γ. The IR model dynamics are:

$$I'(t) = -\gamma I(t)\,,$$

  for $t > t_0$ with $I(t_0) = I_0$ and with $I(t) = 0$ for $t < t_0$.

Synthedemic Methodology Overview

Let $\mathcal{M}$ be the class of subepidemic models that we are considering and let $\mathcal{M}^{(k)}$ be the set of vectors with k subepidemics. Generally the set $\mathcal{M}$ can contain any type of epidemic model but here we restrict ourselves to the 2 types of epidemics introduced above. In view of the parameter sets of these processes, elements of $\mathcal{M}$ take the form,

$$\texttt{sir}(t_0, I_0, S_0, \beta, \gamma) \quad \text{or} \quad \texttt{ir}(t_0, I_0, \gamma)\,.$$

Note that we do not include the initial number of recovered individuals in the parameter set of the SIR, as the number of recovered individuals does not influence the evolution of the number of infected individuals. For further use, we also introduce the type $\texttt{base}(t_0, I_0) \doteq \texttt{ir}(t_0, I_0, 0)$, which corresponds to a constant infection level I_0 starting at t_0.

[1]A portmanteau term from *synthesised epidemic*

For any $m \in \mathcal{M}$, let $f_m(t)$ be the number of infected individuals at time t of model m. With a slight abuse of notation and assuming that epidemics are additive, we associate with every vector $\mathbf{E}$ of elements of $\mathcal{M}$, the multiple epidemic,

$$f_{\mathbf{E}}(t) = \sum_{E \in \mathbf{E}} f_E(t)\,.$$

Let y_i be the ith data point which is collected at time t_i, and let $\mathbf{t}$ and $\mathbf{y}$ be the vectors with elements t_i and y_i, respectively. Moreover, let $\mathbf{t}_i$ be the vector with elements t_1 to t_i. $\mathbf{y}_i$ is defined likewise. We aim to find a sequence of vectors of subepidemics $\{\mathbf{E}(i) : \mathbf{E}(i) \subset \cup_k \mathcal{M}^{(k)}\}$ such that $\mathbf{E}(i)$ maximizes the coefficient of determination for the data up till time t_i, whereby the number of subepidemics is upper-bounded. The bound is chosen such that a target coefficient of determination r^2_{target} can be attained. The coefficient of determination for a vector of epidemics $\mathbf{E}$ and data points $\mathbf{y}$ collected at epochs $\mathbf{t}$, is defined as,

$$r^2(\mathbf{E}, \mathbf{y}, \mathbf{t}) = 1 - \frac{|\mathbf{y} - f_{\mathbf{E}}(\mathbf{t})|^2}{|\mathbf{y} - \bar{\mathbf{y}}|^2}$$

where $|\cdot|$ and $\bar{\mathbf{y}} = \frac{1}{\ell(\mathbf{y})} \sum_{k=1}^{\ell(\mathbf{y})} y_k$ denote Euclidean distance and sample mean, respectively. We also introduce the notation $\ell(\mathbf{y})$ for the number of elements in a vector $\mathbf{y}$ and the vector $f_{\mathbf{E}}(\mathbf{t})$ with elements $f_{\mathbf{E}}(t_i)$ for ease of notation.

The general optimisation problem can be formulated as,

$$\mathbf{E}(i) = \underset{\mathbf{F} \subset \mathcal{M}^{(k_i^-)}}{\operatorname{argmax}}\ r^2(\mathbf{F}, \mathbf{y}_i, \mathbf{t}_i) = \underset{\mathbf{F} \subset \mathcal{M}^{(k_i^-)}}{\operatorname{argmin}}\ |\mathbf{y}_i - f_{\mathbf{F}}(\mathbf{t}_i)|^2$$

with

$$k_i^- = \min\left\{ k \in \mathbb{N}_0 \;\middle|\; \exists \mathbf{F} \in \mathcal{M}^{(k)} : r^2(\mathbf{F}, \mathbf{y}_i, \mathbf{t}_i) \geq r^2_{\text{target}} \right\}.$$

The bound k_i^- on the number of subepidemics allows for achieving r^2_{target} with a parsimonious model. Without such bound the optimisation problem would be trivial. In that case, the optimal fit is to have a spike with infinite (or very large) decay rate at every data point. As the formulated optimisation problem is numerically involved, we formulate a heuristic optimisation approach in the next section.

Practical Implementation Issues

In order to improve the speed and stability of our online fitting procedures, we constrain the search space for finding $\mathbf{E}(i)$ as follows:

- We add or subtract at most one epidemic at each t.
- In updating the vector of epidemics at time t_i, the start times and types of all currently-fitted subepidemics are assumed to be fixed. Other parameters of subepidemics are free and can be updated.
- If an epidemic is added, we use a heuristic to determine its type based on the residual process prior to adding this epidemic.
- SIR-type processes are assumed to start with a single infected individual. Henceforth this parameter is suppressed in the notation.

In view of the former assumptions, let $\mathcal{N}_\delta(E)$ denote the δ-neighbourhood of subepidemic E. For a SIR process, this neighbourhood is defined as,

$$\mathcal{N}_\delta(\texttt{sir}(t_0, S_0, \beta, \gamma)) = \{\texttt{sir}(t, s_0, b, g) \mid t \in (t_0 - \delta, t_0 + \delta), s_0 > 0, b > 0, g > 0\},$$

whereas for the IR and baseline, the neighbourhood is,

$$\mathcal{N}_\delta(\texttt{ir}(t_0, I_0, \gamma)) = \{\texttt{ir}(t, i_0, g) \mid t \in (t_0 - \delta, t_0 + \delta), i_0 > 0, g > 0\},$$

and,

$$\mathcal{N}_\delta(\texttt{base}(I_0)) = \{\texttt{base}(i_0) \mid t = 0, i_0 > 0\},$$

respectively. With a slight abuse of notation, the neighbourhood of vector of epidemics is defined as

$$\mathcal{N}_\delta([E_1, E_2, \ldots, E_k]) = [\mathcal{N}_0(E_1), \mathcal{N}_0(E_2), \ldots, \mathcal{N}_0(E_{k-1}), N_\delta(E_k)].$$

Notice that we only allow changes of the start time for the epidemic which was added last and keep the start time of the preceding epidemics fixed.

Our practical experience to date is that a value of $\delta = 20$ yields good results; this corresponds to a large enough window to provide start time flexibility while maintaining computational feasibility.

With the notation introduced above, our heuristic online fitting algorithm is shown in Algorithm 1. Here **ite**(*cond*,*a*, *b*) is an if-then-else function that returns *a* when *cond* is true and *b* otherwise. Informally, the algorithm can be described as follows. First, as there is insufficient information if only the first few data points are known, we set

$$\mathbf{E}(0) = \{\texttt{base}(0)\}.$$

For each additional data point, we do the following.

1. We first check if the target coefficient of determination can be attained by parametrising the current set of epidemics. The optimal set is

$$\hat{\mathbf{E}}(i) = \operatorname*{argmax}_{\mathbf{F} \in \mathcal{N}_\delta(\mathbf{E}(i-1))} r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i),$$

and the corresponding coefficient of determination is

$$\hat{r}^2(i) = r^2(\hat{\mathbf{E}}(i), \mathbf{t}_i, \mathbf{y}_i)$$

2. If $\hat{r}^2(i) \geq r^2_{\text{target}}$, we try to reduce the number of epidemics. Therefore, we try to attain the target coefficient of determination without the last epidemic (provided that there is more than one epidemic).

$$\tilde{\mathbf{E}}(i) = \operatorname*{argmax}_{\mathbf{F} \in \mathcal{N}_\delta(\mathbf{E}_{\ell(\mathbf{E}(i-1))-1}(i-1))} r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i),$$

and the corresponding coefficient of determination is

$$\tilde{r}^2(i) = r^2(\tilde{\mathbf{E}}(i), \mathbf{t}_i, \mathbf{y}_i).$$

If $\tilde{r}^2(i) \geq r^2_{\text{target}}$, we can reduce the number of epidemics and set $\mathbf{E}(i) = \tilde{\mathbf{E}}(i)$. If not then we set $\mathbf{E}(i) = \hat{\mathbf{E}}(i)$ and move on to the next data point.

3. If $\hat{r}^2(i) < r^2_{\text{target}}$, we consider adding an epidemic. To determine the type of the epidemic (`sir` or `ir`), we first calculate the residual vector

$$\mathbf{z}_i = \mathbf{y}_i - f_{\hat{\mathbf{E}}(i)}(\mathbf{t}_i).$$

Let $\mu(i)$ be the sample mean of $\mathbf{z}_i$ and let $\sigma(i)$ be the sample standard deviation of $\mathbf{z}_i$. As the new epidemic should be located at the end of the residual, let $\mathbf{z}_{i-\kappa+1:i}$ be the last κ data points in $\mathbf{z}_i$.

- The new epidemic type is `sir` if the minimum value in $\mathbf{z}_{i-\kappa:i}$ exceeds $\mu(i) + 2\sigma(i)$. In our experiments, we found that $\kappa = 2$ yields good results.
- The new epidemic type is `ir`, if the most recent residual exceeds $\mu(i) + 6\sigma(i)$.
- if neither `ir` nor `sir` are detected, we set $\mathbf{E}(i) = \hat{\mathbf{E}}(i)$, and issue a warning that r^2_{target} can not be attained at t_i due to no epidemic being detected.

If an epidemic is detected, we extend $\hat{\mathbf{E}}(i)$ with the detected epidemic $E^{(d)}$ started at time t_i, and let $\breve{\mathbf{E}}(i)$ be the optimal vector of epidemics in the neighbourhood of this extended vector,

$$\breve{\mathbf{E}}(i) = \operatorname*{argmax}_{\mathbf{F} \in \mathcal{N}_\delta([\hat{\mathbf{E}}, E^{(d)}])} r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i).$$

The corresponding coefficient of determination is

$$\breve{r}^2_i = r^2(\breve{\mathbf{E}}(i), \mathbf{t}_i, \mathbf{y}_i).$$

If $\breve{r}^2(i) > r^2_{\text{target}}$ then we add the new epidemic and set $\mathbf{E}(i) = \breve{\mathbf{E}}(i)$. If not then $\breve{r}^2_i$ is below the target value; however, we may still be able to improve on the current fit. So we check if $\breve{r}^2_i > \hat{r}^2$, if this is the case we set $\mathbf{E}(i) = \breve{\mathbf{E}}(i)$ and issue a warning that r^2_{target} could not be attained at time t_i, even though an epidemic has been added. Finally, if $\breve{r}^2_i \leq \hat{r}^2$, then the new epidemic did not improve the fit; in this case we set $\mathbf{E}(i) = \hat{\mathbf{E}}(i)$ and issue a warning that r^2_{target} could not be attained at time t_i.

4. RESULTS

We demonstrate our technique's applicability on a synthetic data trace (derived by composing two time-shifted SIR model traces to produce a double epidemic model) and real data including swine flu data and BitTorrent downloads of music artist Robin Thicke. The BitTorrent download data were retrieved by the *MusicMetric API* (an online artist analytics toolbox that contains detailed information on fan trends for particular artists).

Synthetic Double Epidemic Model (2 SIR models)

This dataset was created by the superposition of two time-shifted stochastic simulation trajectories of SIR epidemics with known parameters:

$$\beta^{(1)} = 0.001,\ \gamma^{(1)} = 0.05,\ S_0^{(1)} = 400,\ I_0^{(1)} = 1$$
$$\beta^{(2)} = 0.001,\ \gamma^{(2)} = 0.01,\ S_0^{(2)} = 400,\ I_0^{(2)} = 1$$

The combined epidemic was then used as input to a prototype implementation. As observed in Figure 2, on day 56 the fit of the epidemic component is proceeding well ($r^2 = 0.999$)

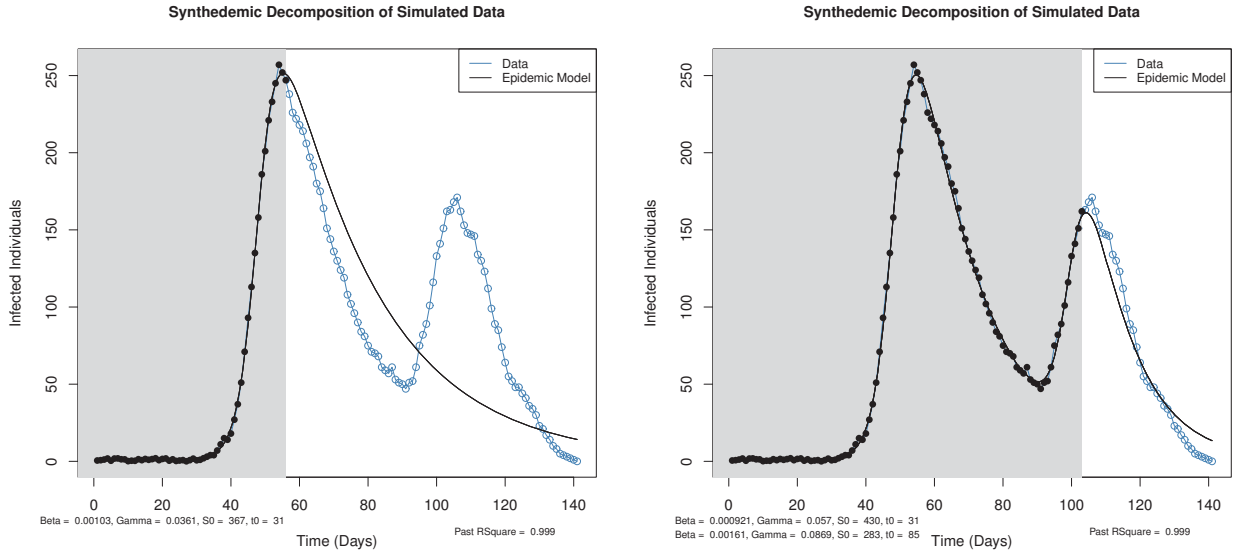

Figure 2: Synthedemic fit at days 56 and 103 to synthetic data with 2 subepidemics ($r^2_{\text{target}} = 0.99$).

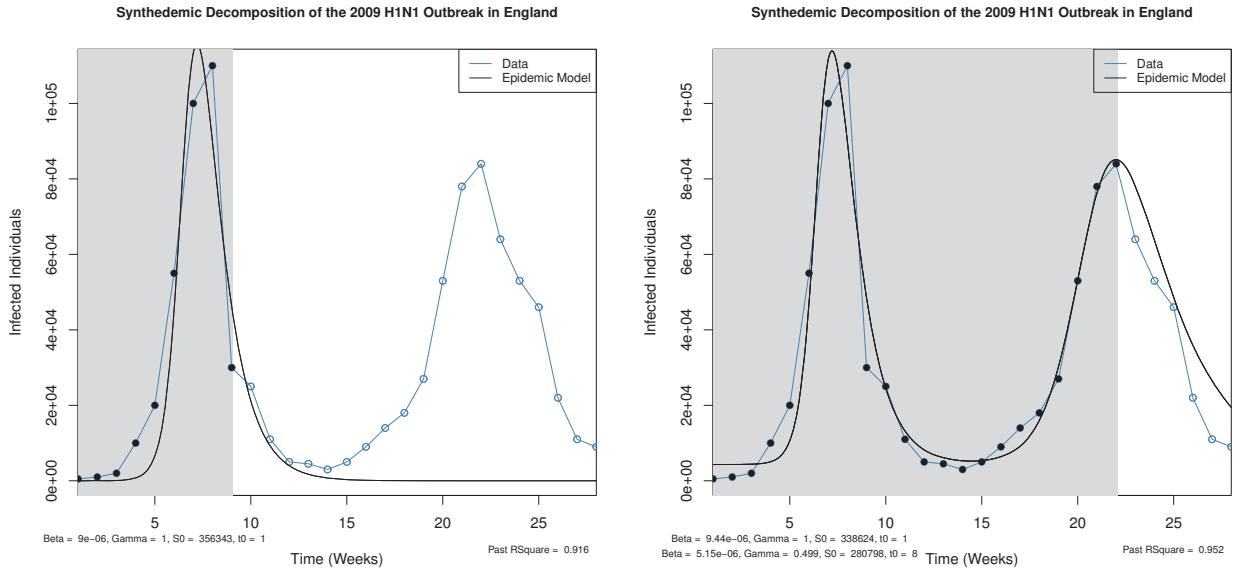

Figure 3: Synthedemic fit at weeks 9 and 22 to weekly Swine Flu reported cases in England during 2009 ($r^2_{\text{target}} = 0.9$).

Algorithm 1 Fitting Process

function ONLINEFITTING($\mathbf{t}$, $\mathbf{y}$)
  $\mathbf{E} \leftarrow [\texttt{base}(0)]$
  for i = 1 to ℓ(y) **do**
    $\hat{\mathbf{E}} \leftarrow \text{argmax}_{\mathbf{F} \in \mathcal{N}_\delta(\mathbf{E})}\, r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i)$
    $\hat{r}^2 \leftarrow r^2(\hat{\mathbf{E}}, \mathbf{t}_i, \mathbf{y}_i)$
    if $r^2 \geq r^2_{\text{target}}$ **then**
      $\tilde{\mathbf{E}} \leftarrow \text{argmax}_{\mathbf{F} \in \mathcal{N}_\delta(\mathbf{E}_{\ell(\mathbf{E})-1})}\, r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i)$
      $\tilde{r}^2 \leftarrow r^2(\tilde{\mathbf{E}}, \mathbf{t}_i, \mathbf{y}_i)$
      $\mathbf{E} \leftarrow \textbf{ite}(\tilde{r}^2 \geq r^2_{\text{target}}, \tilde{\mathbf{E}}, \hat{\mathbf{E}})$
    else
      $\mathbf{z} \leftarrow \mathbf{y}_i - f_{\hat{\mathbf{E}}}(\mathbf{t}_i)$
      $\mu \leftarrow \frac{1}{i} \sum_{k=1}^{i} z_k$
      $\sigma \leftarrow \sqrt{\frac{1}{i-1} \sum_{k=1}^{\ell(\mathbf{z})} (z_k - \mu)^2}$
      $\breve{r}^2 \leftarrow 0$
      if $\min(\mathbf{z}_{i-\kappa:i}) \geq \mu + 2\sigma$ **then**
        $\breve{\mathbf{E}} \leftarrow \text{argmax}_{\mathbf{F} \in \mathcal{N}_\delta([\hat{\mathbf{E}}, \texttt{sir}(t_i, 1, 1, 1)])}\, r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i)$
        $\breve{r}^2 \leftarrow r^2(\breve{\mathbf{E}}, \mathbf{t}_i, \mathbf{y}_i)$
      else if $z_i > \mu + 6\sigma$ **then**
        $\breve{\mathbf{E}} \leftarrow \text{argmax}_{\mathbf{F} \in \mathcal{N}_\delta([\hat{\mathbf{E}}, \texttt{ir}(t_i, 1)])}\, r^2(\mathbf{F}, \mathbf{t}_i, \mathbf{y}_i)$
        $\breve{r}^2 \leftarrow r^2(\breve{\mathbf{E}}, \mathbf{t}_i, \mathbf{y}_i)$
      end if
      if $r^2 \geq r^2_{\text{target}}$ **then**
        $\mathbf{E} \leftarrow \breve{\mathbf{E}}$
      else
        print r^2_{target} not attained at time t_i
        $\mathbf{E} \leftarrow \textbf{ite}(\breve{r}^2 > \hat{r}^2, \breve{\mathbf{E}}, \hat{\mathbf{E}})$
      end if
    end if
    print t_i, $\mathbf{E}$
  end for
end function

and the short term prediction quality is good as the model matches the forthcoming decay of the epidemic. The estimated parameters are close to the known parameters. The estimated parameters become even closer to the actual parameters as the data points move towards the introduction of the second epidemic on day 91. Our framework realises the need for a second epidemic and begins the fitting procedure again. On day 103 the quality of the fit to past data is good (as $r^2 = 0.999$) and the model predicts the downward trend. By the end of the epidemic, the model fitted the data and estimated the parameters of both epidemics well.

Swine flu 2009 reported cases in the UK

In 2009, there was a global outbreak of a new strain of influenza A virus subtype H1N1 (colloquially called *swine flu*) which was termed a pandemic by the World Health Organization. We use weekly reported swine flu cases in 2009 in England as provided by the Health Protection Agency [4], and successfully fit a double epidemic. On week 9 in Figure 3 the model has detected and fitted the swine flu data with past $r^2 = 0.916$. On week 22, the model detects with a good precision the subsequent downwards evolution of the second outbreak. Investigating the biological interpretation of our model would be an interesting exercise, but one which is beyond the scope of the present paper.

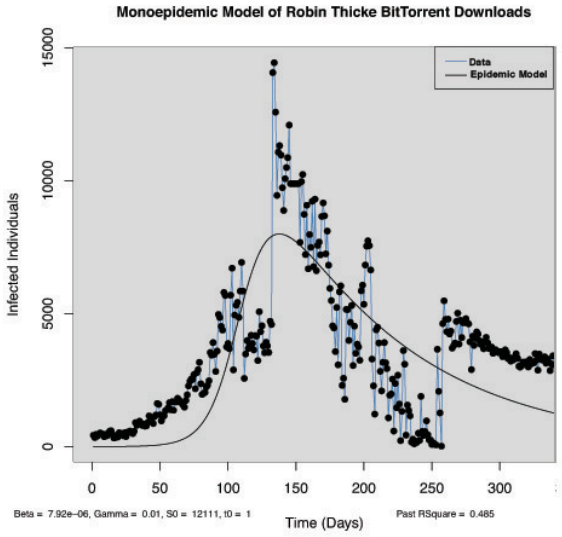

Figure 4: Monoepidemic SIR model fit to Robin Thicke BitTorrent download data

Robin Thicke BitTorrent Downloads

This dataset begins with the release of Robin Thicke's album *Blurred Lines*, the title track of which was the best-selling song of 2013 in the UK and the second best-selling song of 2013 in the US. Each data point represents the number of daily downloads of Robin Thicke's songs, and it is these downloads that we presume are the manifestation of a number of underlying epidemic spreading processes.

Figure 4 presents a monoepidemic fit in the style of the work described in [18] which clearly demonstrates the inability of a single-epidemic model to reflect adequately the complexity of this kind of data set. Indeed the r^2 value is just 0.485.

As illustrated in Figure 5, the synthedemic model fit with $r^2_{\text{target}} = 0.9$ fares much better. On day 94 the model not only fits the historical data extremely well but also predicts the peak of the initial growth phase accurately, albeit that the model tends to overestimate near-term future download counts. On day 135 a sudden peak in the data is observed corresponding to Robin Thicke's performance of *Blurred Lines* on the TV show Jimmy Kimmel Live. The model not only detects this as a second epidemic, but also predicts the downward trend with good accuracy. On day 206, we observe a new short and sharp outbreak which is detected and fitted well. This corresponds to the infamous live performance of *Blurred Lines* by Robin Thicke and Miley Cyrus at the 2013 MTV Video Music Awards. Last but not least, our model has successfully detected the outbreak on day 254 corresponding to Robin Thicke's live performance on the X Factor results show.

We note there is clear potential to improve historical fit and prediction quality through the application of appropriate residual refinement techniques e.g. use of an autoregressive (AR) process to characterise the variability of residuals.

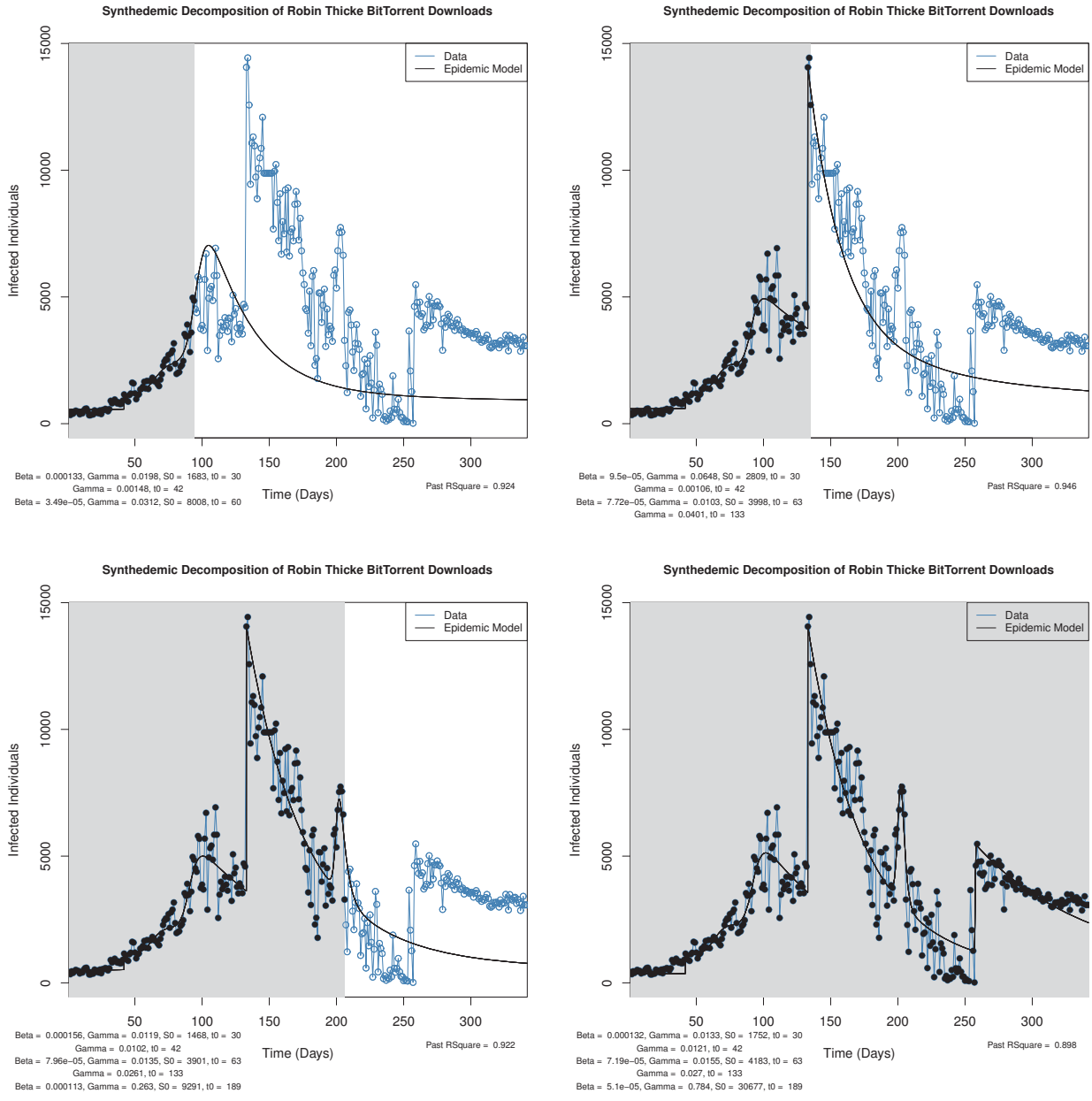

Figure 5: Synthedemic fit at days 94, 135, 206 and 342 to Robin Thicke's BitTorrent downloads ($r^2_{\text{target}} = 0.9$)

5. CONCLUSION

This paper has proposed a novel framework for quantitative modelling and prediction based on the analysis and synthesis of multiple epidemic models. Using a surprisingly low number of synthesised epidemics, a prototype implementation of this framework is able to adequately characterise the evolution of an artificially-generated data set and real-world data sets based on swine flu data and daily BitTorrent downloads. Model fitting can be performed in an online manner as an outbreak of interest unfolds, and the short-term model predictions are generally pleasing although they have not been subjected to rigorous quantitative analysis in the present work and there is scope to improve predictions using residual refinement techniques.

There are several possible directions for future work. Firstly, although we have focused on internet-based phenomena, we anticipate that our methodology may be readily applied in many other domains that might arguably be driven by underlying epidemic-like phenomena, such as computer viruses and retail sales. To facilitate this, we believe our prototype implementation could be extended in order to support epidemic model selection from a broader range of candidate models, as well as to support negative epidemic terms.

We also plan to incorporate event prediction with prior knowledge of upcoming events (where applicable) in order to improve predictive ability. We also plan to investigate how appropriate confidence intervals can be computed on synthedemic model predictions. Last but not least, we would like to explore the range of potential dependencies between epidemics and their host populations and what implications these may have for the synthedemic paradigm.

6. REFERENCES

[1] F. Adams and E.C. Kelly. *The Genuine Works of Hippocrates*. Kessinger Publishing, 2006.

[2] K. Avrachenkov, K. De Turck, D. Fiems, and B. J. Prabhu. Information dissemination processes in directed social networks. In *International Workshop on Modeling, Analysis and Management of Social Networks and their Applications (SOCNET)*, 2014.

[3] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proc. ACM WWW 2012*, 2012.

[4] E. Brooks-Pollock and K. Eames. Pigs didn't Fly, but Swine Flu. *Mathematics Today*, 47:36–40, 2011.

[5] M. Cha et al. Measuring influence on twitter: The million follower fallacy. In *4th International Conference on Weblogs & Social Media*, 2010.

[6] N. Christakis, J. Fowler, and H. James. The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*, (4):370–379, July 2007.

[7] N. A. Christakis and J. H. Fowler. Detecting emotional contagion in massive social networks. *PLoS ONE*, 5(9):e12948, 09 2010.

[8] R. Dawkins. *The Selfish Gene*. Oxford University Press, Oxford, UK, 1976.

[9] A. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, and R. Rothman. Influenza forecasting with Google flu trends. *Online Journal of Public Health Informatics*, 5(1), 2013.

[10] Facegroup. How stuff spreads: How videos go viral part I. Available at `http://www.facegroup.com/how-videos-go-viral.html`.

[11] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the Twitterers — predicting information cascades in microblogs. In *Proc. 3rd Workshop on Online Social Networks*, June 2010.

[12] J. Gaunt. *Natural and Political Observations Mentioned in a following index, and made upon the Bills of Mortality*. 1662. Available online at `http://www.neonatology.org/pdf/graunt.pdf`.

[13] W. Goffman and V. A. Newill. Generalization of epidemic theory: An application to the transmission of ideas. *Nature*, 204:225–228, October 1964.

[14] W. Kermack and A. McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proc. Royal Society of London. Series A*, 115(772):700–721, 1927.

[15] A. H. Maslow. A theory of human motivation. *Psychological Review*, 50(1):370–396, 1943.

[16] S. Myers and J. Leskovec. Clash of the Contagions: Cooperation and Competition in Information Diffusion. In *Proc. IEEE International Conference on Data Mining (ICDM 2012)*, 2012.

[17] M. Nika, D. Fiems, K. De Turck, and W. J. Knottenbelt. Modelling interacting epidemics in overlapping populations. In *Proc. 21st International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA)*, 2014.

[18] M. Nika, G. Ivanova, and W. J. Knottenbelt. On celebrity, epidemiology and the internet. In *Proc. 7th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS 2013)*, Turin, Italy, December 2013.

[19] T. Wilding R. Danila, M. Nika and W.J. Knottenbelt. Uncertainty in on-the-fly epidemic fitting. In *Proc. 11th European Performance Engineering Workshop (EPEW 2014)*, Florence, Italy, 2014.

[20] J. Snow. *On the Mode of Communication of Cholera*. John Churchill, 1855.

[21] V. Tweedle and R. J. Smith. A mathematical model of Bieber fever: The most infectious disease of our time. In S. Mushayabasa and C.P. Bhunu, editors, *Understanding the dynamics of emerging and re-emerging infectious diseases using mathematical models*. Transworld Research Network, 2012.

[22] E. Vynnycky and R. White. *Introduction to Infectious Disease Modelling*. Oxford University Press, 2010.