Predicting an epidemic based on syndromic surveillance

Alex Skvortsov
HPP Division
DSTO
Australia
alex.skvortosv@dsto.defence.gov.au

Branko Ristic
ISR Division
DSTO
Australia
branko.ristic@dsto.defence.gov.au

Chris Woodruff
HPP Division
DSTO
Australia
chris.woodruff@dsto.defence.gov.au

Abstract − *Early detection and prediction of the size and the peak time of an epidemic outbreak (malicious or natural) is of crucial importance for a timely medical response (quarantine, vaccination, etc). A conventional approach to this problem is based on large scale agent-based computer simulations. This paper proposes an alternative framework formulated in the context of stochastic nonlinear filtering. The framework is based on the stochastic SIR epidemiological model of infection dynamics, with syndromic (often non-medical) observations of the number of infected people (e.g. visits to pharmacies, sale of certain products, absenteeism from work/study etc.). The unknown parameters of the SIR epidemic model are estimated via the sequential Monte Carlo method, with the prediction based on the dynamic model. The numerical results indicate that the proposed framework can provide useful early prediction of the epidemic peak if the uncertainty in prior knowledge of model parameters is not excessive.*

Keywords: Epidemics, bio-terrorism, mathematical biology, state estimation, particle filter.

1 Introduction

Epidemics can impose significant challenges on modern societies. Not only can they affect the health of the general population, but also they can cause negative trends in the economy by increasing the costs of immunisation, follow-up medical treatments, absenteeism from work, preparedness efforts, missed business opportunities, etc. The ongoing epidemics of AIDS, tuberculosis and the recent outbreaks of SARS and H1N1 (swine flu) provide some revealing examples. In the absence of a cure against many diseases, the best approach to mitigate an epidemic outbreak (malicious or natural) resides in the development of capability for its early detection and for prediction of its further development [6], [12], [3]. Such a capability would allow making any countermeasures (quarantine, vaccination, medical treatment) much more effective and less costly [25]. Conventional approach to early detection and prediction of an epidemic is often based on massive scale agent-based simulation programs, like EpiSim [23],[24].

One of the promising methods for early detection of epidemics, which has recently attracted significant attention by scientists and practitioners, is the so-called method of syndromic surveillance (specific patterns of symptomatic behavior), for more details see [12] and [5]. The main idea of this method rests on the reasonable assumption (in some cases validated by field studies) that an increase in the number of infected people in the community is usually associated with the changes of some other indicators, which can be easily measured/estimated by "non-medical" means (e.g. number of visits to pharmacies, sales of a particular product, number of hits of a particular web site, absenteeism from work/school etc).

In general, in order to plausibly predict the existence and further development of an epidemic based on syndromic surveillance data, the latter should be somehow assimilated into a feasible, preferably analytical, epidemiological model. Then from this model the amplitude (the size) and the peak time of the epidemic (the two most important parameters for medical response to the epidemic) can be estimated by employing a suitable estimation algorithm. Unfortunately, a straightforward application of this framework to the traditional epidemiological (or compartmental) models runs into problems. Firstly, the traditional (compartmental) epidemiological models, such as the SIS (susceptible-infectious-susceptible) model, SIR (susceptible-infectious-removed) model and SEIR (susceptible-exposed-infectious-removed) model are based on the so-called well-mixed population assumption. This is, however, a highly idealised assumption in which population is perfectly mixed to a uniform state after each social contact [8]. A reconciliation of a syndromic surveillance data obtained from the realis-

tic scenarios (i.e. for communities with inhomogeneous mixing) with such idealised epidemiological models results in poor estimates of the amplitude or the peak time of an epidemic [21], [20]. Secondly, quantitative characteristics of social mixing (i.e. interactions of individuals in a community) are strongly non-universal (i.e. may significantly vary for different communities) and are very rarely known beforehand. This makes the calibration of the epidemiological models for a given community quite a challenging task. Moreover, when some characteristics of social interactions are known (or can be initially postulated with some degree of uncertainty) they still may change significantly after an epidemic outbreaks (as a result of self-isolation, school closures, cancellation of pubic gatherings, etc). As a consequence, a "generic" estimation algorithm based on syndromic surveillance data ideally should be able to deal with the intrinsic uncertainty in intensity of social interactions as well as with its possible variations.

A simple way to overcome the first issue was proposed in the recent paper [23] (for further references see also [14]), where a new "inhomogeneity" parameter was heuristically introduced into the SIR epidemiological model. It was suggested [23] and also validated by numerical simulations that with the new parameter, the modified epidemiological model can capture the inhomogeneity of social interactions. The model reduces to the conventional SIR model for a particular value of the "inhomogeneity" parameter (details below).

The second issue was addressed in [17], where it was proposed to incorporate the "mixing" parameter in the state vector and to estimate it recursively in conjunctions with other unknown parameters of the epidemiological model. The assumption was that we have a relatively tight prior probability density function on the "mixing" parameter. In [17] we showed using a deterministic SIR model and a theoretical Cramér-Rao bound based analysis that the peak and the size of an epidemic can be predicted, with accuracy dependent on the prior knowledge of unknown model parameters and the time span of available observations of infected people. Also, the SIR dynamic model was verified against the agent-based simulation data.

As an extension of the initial work presented in [17], this paper presents a framework for early detection and prediction of an epidemic, where a stochastic SIR epidemic model is used to describe the dynamics of the disease spread. Assuming the measurements are syndromic observations, a particle filter is developed to estimate the state of the dynamic model and to predict the size and the peak of an epidemic.

2 Model Formulation

This section introduces two mathematical models. The first is the dynamic model of an epidemic; the second is the observation model of symptomatic behavior resulting from the epidemic outbreak.

2.1 Epidemic Outbreak Model

To describe the dynamics of an epidemic outbreak we employ the stochastic SIR epidemic model [9], [8], [1], [7]. According to this model, the population can be subdivided into sets of distinct compartments in relation to the disease: susceptible (S), infectious (I) and recovered (R). Susceptible individuals have never come into contact with the disease. They are able to catch the disease and thus to move to the I compartment. Eventually the infectious individuals recover and thus move into the R compartment.

An outbreak of an epidemic is usually far more rapid than the vital dynamics of a population. Hence we can neglect the birth-death process to state that $S+I+R = P$, where P is the (constant) number of people in the population (assumed to be known). For simplicity and without loss of generality we can consider a normalised system where $s = S/P$, $i = I/P$ and $r = R/P$. The stochastic SIR model can be expressed by two stochastic differential equations [2], [9], [8]:

$$\frac{ds}{dt} = -q + \sigma_q \xi, \tag{1}$$

$$\frac{di}{dt} = q - \beta i - \sigma_q \xi + \sigma_\beta \zeta, \tag{2}$$

$$r = 1 - s - i, \tag{3}$$

where the last equation is simply due to the "conservation" law for the population. Here $q \equiv q(s,i)$ is a nonlinear mixing term, describing social contacts between individuals; $\beta = const$ is the recovery rate (a disease specific parameter); ξ, ζ are two uncorrelated white Gaussian noise processes, both with zero mean and unity variance. The terms $\sigma_q \equiv \sigma_q(s,i)$ and $\sigma_\beta \equiv \sigma_\beta(s,i)$ are introduced to capture the so-called stochasticity of the real social network (random variations in the contact rate q and in the recovery time β between individuals), for details see [9], [8].

In the deterministic SIR model with homogeneous mixing $q(s,i) = \alpha\, i\, s$, with $\alpha = const$, [13], [7]. A simple phenomenological extension of the SIR model to the non-homogeneous mixing case was proposed in [23] with

$$q(s,i) = \alpha\, i\, s^\nu. \tag{4}$$

Here a parameter, ν, describes a mixing inhomogeneity, with a particular value $\nu = 1$ corresponding to the uniform mixing scenario. In general, ν can be treated as another fitting parameter of the model (for a theoretical derivation of ν see [14]). Indeed the behavior of the modified model is very sensitive to the variations of the inhomogeneity parameter ν, so it is reasonable to expect that PDF of its observable values still has a decent peak around its "well-mixed" value $\nu = 1$. We will use this fact as prior information in our estimation

framework (see Sec. 4). Note that an epidemic (effectively a chain reaction due to the interaction of people) will happen only if the ratio $\rho = \alpha/\beta$ is greater than unity[1] [1], [13].

The amplitude of noise terms can be established from a scaling law of Gaussian fluctuations generated by the random contact rate q and recovery rate βi. Thus for a dynamical system (1) consisting of a large number of individuals P we can write (for details see [9], [8])

$$\sigma_q(s,i) = \sqrt{\frac{q(s,i)}{P}}, \quad (5)$$

$$\sigma_\beta(s,i) = \sqrt{\frac{\beta i}{P}}. \quad (6)$$

Although stochastic Eqs (1) – (6) form a closed system, which is sufficient to formulate an estimation problem, they are still too complex for development of plausible estimation algorithms for operational applications and require further simplifications. To achieve computational efficiency and peak performance, which are critical for the syndromic surveillance systems, we propose the following approximation to the model (1) – (6)

$$\sigma_q(s,i) \approx \sigma_q(s_0,i_0) = \sqrt{\frac{q(s_0,i_0)}{P}} = const, \quad (7)$$

$$\sigma_\beta(s,i) \approx \sigma_\beta(s_0,i_0) = \sqrt{\frac{\beta i_0}{P}} = const. \quad (8)$$

With further reasonable assumptions $s_0 \approx 1, r_0 \approx 0, i_0 \approx 1/P$, where s_0, i_0, r_0 initial values of s, i, r and $\nu \approx 1$ Eqs.(7),(8) can be reduced to

$$\sigma_q(s,i) \approx \frac{\sqrt{\alpha}}{P}, \quad \sigma_\beta(s,i) \approx \frac{\sqrt{\beta}}{P}. \quad (9)$$

The effect of this approximation will be discussed in Sec.3.

2.2 Model of Syndromic Observations

For syndromic observations we employ a linear model. This means that each syndrome (number of visits to pharmacies, calls to "hot lines", sales of a particular product, visits of particular web sites, etc [5], [12]) is a linear function of the number of infected people. The observation model is then

$$z^j = b^j\, i + \sigma_j \eta^j, \quad (10)$$

where z^j is the observable syndrome index $j = 1, \ldots, N_z$; $b^j, \sigma_j = const$ (different for different syndromes); η^j is zero-mean, unit variance white Gaussian noise (since $z^j \geq 0$, η^j is actually a truncated Gaussian), associated with syndrome j; η^j is assumed to be uncorrelated to other syndromes and noises ξ and ζ.

[1] In epidemiology ρ is called the basic reproduction number [1], [13].

For other, possibly more complex and nonlinear functional forms of z^j, we assume that model (10) still holds at least as the first (linear) approximation. This implies that more complex forms of z^j at least can be linearized and expressed in terms of (10) for small values of i. Note that at initial stages of an epidemic (when detection and prediction are most important), we indeed deal with small values of i.

An important advantage of model (10) is that it preserves its functional form in the case of symptoms with delays. Indeed for delayed symptoms we can write $z^j(t) \approx b^j\, i(t-\tau^j) + \sigma_j \eta^j$ or

$$z^j(t) \approx b^j\, i(t) - b^j\, \tau^j \frac{di}{dt} + \sigma_j \eta^j \approx c^j\, i + \omega_j \eta^j, \quad (11)$$

where $c^j = b^j(1 - \tau^j(\alpha - \beta)) = const$, τ^j are time delays (symptom specific parameters), ω_j is the "renormalised" noise component. We observe that syndromic observations in this case still follow a linear model. For derivation of (11) we used (2), (4) and the assumptions of short delays (i.e. τ^j are shorter than time of epidemic outbreak) and low fraction of initially infected population i_0 (i.e. $s_0 \approx 1$).

It is worth noting that with the introduced normalisation for i ($i = I/P$) we have the obvious constraint on the values of parameters b^j in (10): $b^j \leq 1$, since the number of observable cases of a particular syndrome should not be greater than the number of infected individuals (at the same time one individual still can show many different symptoms). This constraint is used later in our numerical simulations.

3 Early detection and prediction

The problem of early detection and prediction of an epidemic will be formulated in the framework of sequential Bayesian estimation for stochastic dynamic systems. We adopt the state-space approach and for the purpose of estimation define the state vector and its initial (prior) PDF. Finally using the time-discretised dynamic model, we estimate sequentially the state vector, as the measurements become available. The estimated state vector is eventually predicted for future times using the dynamic model.

3.1 Adopted framework

The state vector is adopted based on the epidemic model (1)-(2). We assume that the process noise statistics are known and adopt the state vector as follows:

$$\mathbf{x} = \begin{bmatrix} i & s & \alpha & \beta & \nu \end{bmatrix}^\intercal \quad (12)$$

where $^\intercal$ denotes matrix transpose. Neglecting for the moment the process noise terms, the evolution of the epidemic state can be written according to (1)-(2),(4) as $\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x})$ where $\mathbf{g}(\mathbf{x}) = \begin{bmatrix} (\alpha s^\nu - \beta) i & -\alpha i s^\nu & 0 & 0 & 0 \end{bmatrix}^\intercal$. The nonlinear differential equation governing the evolution of the state

cannot be solved in the closed-form. The Euler method provides a simple approximation valid for small integration interval $\tau > 0$: $\mathbf{x}(t+\tau) \approx \mathbf{x}(t) + \tau \mathbf{g}(\mathbf{x}(t))$. The state-evolution in discrete-time t_k can then be expressed as:

$$\mathbf{x}_{k+1} \approx \mathbf{f}_k(\mathbf{x}_k) + \mathbf{w}_k \tag{13}$$

where $k = t_k/\tau$ is the discrete-time index. Transition function $\mathbf{f}_k(\mathbf{x}_k)$ in (13) is given by

$$\mathbf{f}_k(\mathbf{x}_k) = \begin{bmatrix} \mathbf{x}_k[1] + \tau \mathbf{x}_k[1]\left(\mathbf{x}_k[3]\mathbf{x}_k[2]^{\mathbf{x}_k[5]} - \mathbf{x}_k[4]\right) \\ \mathbf{x}_k[2] - \tau \mathbf{x}_k[3]\mathbf{x}_k[1]\mathbf{x}_k[2]^{\mathbf{x}_k[5]} \\ \mathbf{x}_k[3] \\ \mathbf{x}_k[4] \\ \mathbf{x}_k[5] \end{bmatrix} \tag{14}$$

In this notation $\mathbf{x}_k[j]$ represents the jth component of vector $\mathbf{x}_k$. Process noise $\mathbf{w}_k$ in (13) is approximated by zero-mean white Gaussian noise with a diagonal covariance matrix $\mathbf{Q} = \text{diag}[\sigma_i^2,\ \sigma_s^2,\ 0,\ 0,\ \sigma_\nu^2]$. Process noise on component ν of the state vector in general should be non-zero in order to capture the possible time variation in population mixing (people tend to change their behaviour in the presence of an epidemic).

Fig.1 shows ten (random) realisations of the proportion of infected people $i(t_k) = \mathbf{x}_k[1]$ using the model in (13) with the following parameters: $\alpha = 0.2444$, $\beta = 0.1066$, $\nu = 1.2042$. The initial values were $i(0) = 0.002$ and $s(0) = 0.998$ (hence $r(0) = 0$). The integration step used was $\tau = 0.0052$ days. Process noise parameters were selected as $\sigma_i = \sigma_s = 5 \cdot 10^{-5}$ and $\sigma_\nu = 10^{-5}$. Note that in the numerical implementation of (13) one has to make sure that conditions $0 \le i, s, r \le 1$ and $i+s+r = 1$ are always satisfied. The red line in Fig.1 shows the $i(t_k)$ curve in the absence of process noise $\mathbf{w}_k$.

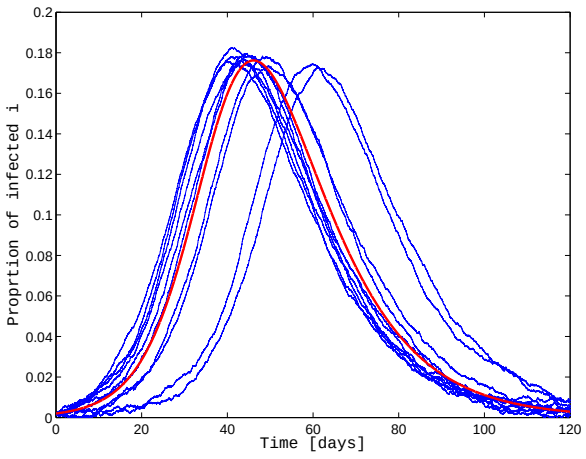

Figure 1: Ten realisations of the stochastic SIR model (in blue) The red line corresponds to the deterministic model (no process noise).

Early detection and prediction refers to the time instant in the first half of the time interval before the peak of $i(t_k)$. We have validated the approximation (9) numerically and generally found that it is in good agreement with the solutions of the full system (1), (4), (5),(6) for a wide range of parameters, during the early stages of an epidemic (although the approximation becomes less accurate as the time progresses).

The syndromic observations are assumed to arrive at irregular intervals. Let z_ℓ^j denote an observation available at time t_ℓ from a syndromic source j. We will assume for convenience that t_ℓ is a multiple of τ (a reasonable approximation for very small integration time τ), that is $\ell = t_\ell/\tau$. The measurement is in a linear relationship with the state vector, i.e.

$$z_\ell^j = \mathbf{H}^j \mathbf{x}_\ell + v_\ell^j \tag{15}$$

where $\mathbf{H}^j = [b^j\ 0\ 0\ 0\ 0]$ and $v_\ell^j = \sigma_j \eta^j(t_\ell)$. The set of all observations (from all sources of syndromic surveillance), accumulated from time 0 to $\ell\tau$ is denoted by $z_{0:\ell}$.

3.2 Algorithm

In the sequential Bayesian estimation/prediction framework [11],[16] the goal is to estimate the probability density function (PDF) $p(\mathbf{x}_k|z_{0:\ell})$, with $\ell \le k$. This is done in two steps, namely *prediction* and *update*. Suppose the posterior PDF at time $t'_\ell = \ell'\tau$ is denoted by $p(\mathbf{x}'_\ell|z_{0:\ell'})$. This PDF is predicted to future time $t_\ell = \ell\tau$, with $\ell > \ell'$, in $(\ell - \ell')/\tau$ prediction steps:

$$p(\mathbf{x}_{k+1}|z_{0:\ell'}) = \int p(\mathbf{x}_{k+1}|\mathbf{x}_k)\, p(\mathbf{x}_k|z_{0:\ell'})\, d\mathbf{x}_k \tag{16}$$

where $t_k = \tau k$ and $k = \ell', \ell'+1, \cdots, \ell-1$. If at time $t_\ell = \ell\tau$ an observation z_ℓ^j becomes available, then the predicted PDF is updated via:

$$p(\mathbf{x}_\ell|z_{0:\ell}) = \frac{p(z_\ell^j|\mathbf{x}_\ell)\, p(\mathbf{x}_\ell|z_{0:\ell'})}{\int p(z_\ell^j|\mathbf{x}_\ell)\, p(\mathbf{x}_\ell|z_{0:\ell'})\, d\mathbf{x}_\ell} \tag{17}$$

For the purpose of sequential Bayesian estimation via (16) and (17), one requires the initial (prior) PDF $p(\mathbf{x}_0)$, the transitional PDF $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ and the likelihood $p(z_\ell^j|\mathbf{x}_\ell)$. The transitional PDF and the likelihood can be easily worked out from (13) and (15), respectively. The initial PDF is adopted as follows:

$$p(\mathbf{x}_0) = p(i_0)\, p(s_0)\, p(\alpha)\, p(\beta)\, p(\nu_0) \tag{18}$$

The PDFs $p(i_0)$ and $p(s_0)$ will be constructed using the first available measurement z_0^j, $j \in \{1, \ldots, N_z\}$ and assuming that initially $r_0 \approx 0$. We adopt the truncated Gaussian PDF for this purpose (truncation in interval $[0,1]$): $i_0 \sim \mathcal{N}_{[0,1]}(z_0^j, \sigma_j^2)$, and $p(s_0) = \mathcal{N}_{[0,1]}(1 - z_0^j, \sigma_j^2)$. The remaining initial PDFs are adopted as uniform densities, i.e. $p(\alpha) = \mathcal{U}[\alpha_1, \alpha_2]$, $p(\beta) = \mathcal{U}[\beta_1, \beta_2]$ and $p(\nu_0) = \mathcal{U}[\nu_1, \nu_2]$.

The estimation-prediction in the sequential Bayesian framework is implemented using the sequential Monte

Carlo method or particle filter [11],[16]. In the particle filter framework the PDF $p(\mathbf{x}_\ell|z_{0:\ell})$ is approximated by a set of N weighted random samples (or particles) $\{w_\ell^i, \mathbf{x}_\ell^i\}_{i=1}^N$, where the weights w_ℓ^i sum up to one. The initial set of particles is generated by sampling N times from $p(\mathbf{x}_0)$,

$$\mathbf{x}_0^i \sim p(\mathbf{x}_0) \quad (i = 1, \ldots, N)$$

and the weights are uniform, $w_0^i = 1/N$. The pseudo-code of a single cycle of the *bootstrap* particle filter is given in Algorithm 1.

Algorithm 1 Pseudo-code of the bootstrap particle filter for early detection/prediction of an epidemic

1: **Input:**
2: • New measurement (time and value): t_ℓ; z_ℓ^j;
3: • Previous time particle set: $t_{\ell'}$; $\{1/N, \mathbf{x}_{\ell'}^i\}_{i=1}^N$
4: **% Prediction to** t_ℓ:
5: **for** $k = \ell' + 1, \ldots, \ell$ **do**
6: Propagate all particles according to (13)
7: **end for**
8: %The resulting set of particles is $\{1/N, \mathbf{x}_\ell^{i*}\}$
9: **% Update:**
10: **for** $i = 1, \ldots, N$ **do**
11: Compute unnormalised weights $\tilde{w}_\ell^i = p(z_\ell^j|\mathbf{x}_\ell^{i*})$
12: **end for**
13: **for** $i = 1, \ldots, N$ **do** ▷ Normalise weights
14: $w_\ell^i = \tilde{w}_\ell / \sum_{i=1}^N \tilde{w}_\ell$
15: **end for**
16: Resample N times from $\{w_\ell^i, \mathbf{x}_\ell^{i*}\}_{i=1}^N$
17: % The resulting set of particles $\{1/N, \mathbf{x}_\ell^i\}$ is used in the next cycle
18: % Predict the epidemic peak (time and amplitude)
19: **for** $i = 1, \ldots, N$ **do**
20: Predict forward particle $\mathbf{x}_\ell^i$ using (13)
21: **if** peak exists **then**
22: Find peak amplitude $A_{\max}^i$ and timing $t_{\max}^i$
23: **end if**
24: **end for**

The input to the algorithm is the set of particles at the previous update time $t_{\ell'}$ and the received observation z_ℓ^j at the new (update) time t_ℓ. The prediction of particles is described in lines 4 to 7 of Algorithm 1. The update based on the new measurement is described in lines 9 to 17. After the update, the particles are predicted forward to find the peak of the epidemic (size and timing), see lines 19-24 in Algorithm 1 (if there is no peak, then there is no epidemic). This method of computation of the peak amplitude $A_{\max}$ and timing $t_{\max}$ can be computationally intensive. An alternative is to derive analytic expressions. From asymptotic formulas for the epidemic curve of well-mixed population [18], we can write in the limit $s_0 \approx 1$:

$$A_{\max} = 1 - \rho + \rho \log(\rho) \quad (19)$$

$$t_{\max} = -(\alpha - \beta)\log(i_0) + G(\rho), \quad (20)$$

where $G(\rho)$ is a known function of $\rho = \alpha/\beta$ (tabulated in [18]) with the maximum $G \approx 0.4$ at $1 \le \rho \le 2$ and the saturation limit $G \approx 0.25$ as $\rho \to \infty$. From analysis of (1) – (3) it can be shown that as a first approximation the mixing inhomogeneity can be incorporated in (19), (20) by a simple substitute $\beta \to \beta^{1/\nu}$.

We found approximation (19) quite reasonable for a wide range of parameters. It is most useful when applied in the two-stage approach to our estimation problem. At the first stage of this approach we provide the fast estimates based on approximation (19) which are then refined at the second stage with the "forward particle predictions". This two-stage approach offers a flexible balance of accuracy and performance efficiency in the context of the problem under consideration. The accuracy and performance efficiency are two of the most important parameters for design and evaluation of any operational system for syndromic surveillance and approximations like (19) provide a revealing example of trade-off strategy in this domain.

Since the proposed estimation framework produces the posterior PDF $p(\mathbf{x}_\ell|z_{0:\ell})$ we can easily extract the PDF of the marginals of the state vector $\mathbf{x}_\ell$ at any point of time. The most important marginal PDF for epidemiologists and medical practitioners is PDF of infected people $p(i(t_k)) \equiv p(\mathbf{x}_\ell[1])$. It is worth emphasizing that by continuous estimation of $p(i(t_k))$ our approach essentially provides a so-called "situation awareness" capability for a syndromic surveillance system. This capability includes computing run-time estimates of the probabilities such as $Pr\{i(t_k) > \gamma\}$ and based on that making conscious decisions on more extensive (and, perhaps, more intrusive) data collection and appropriate mitigation actions (e.g. quarantine, vaccination, etc). Such contingent actions are typically agreed beforehand and are based on several critical γ values (or agreed thresholds) for different interventions. For example, a small value of γ can be selected for a mild action such as "special blood tests", while a higher value of γ can be adopted for a more drastic measure such as "bulk vaccination".

4 Numerical Results

4.1 Experimental data set

The prediction of epidemic peak will be carried out using an experimental data set obtained using a large-scale agent based simulated population model [20], [21] of a virtual town of $P = 5000$ inhabitants, created in accordance with the Australian Census Bureau data. The agent based model is rather complex and takes a long time to run. It includes typical age/sex breakdown and family-household-workplace habits with the realistic day-to-day people contacts for a disease spread. The blue line in Fig.2 shows the number of people of this town infected by a fictitious disease, reported once per day during a period of 154 days (only first 120 days shown). The dashed red line represents the deterministic SIR model fit (using the entire batch of 154 data

points, and setting $\mathbf{w}_k = 0$ in (13)), with estimated parameters $\hat{\alpha} = 0.2399$, $\hat{\beta} = 0.1066$, $\hat{\nu} = 1.2042$. The parameter estimates were obtained using the progressive correction algorithm [15]. Fig.2 serves to confirm that the modified SIR model, although very simple and fast to run, is remarkably accurate in explaining the data obtained from a very complex simulation system (for further details see [23]).

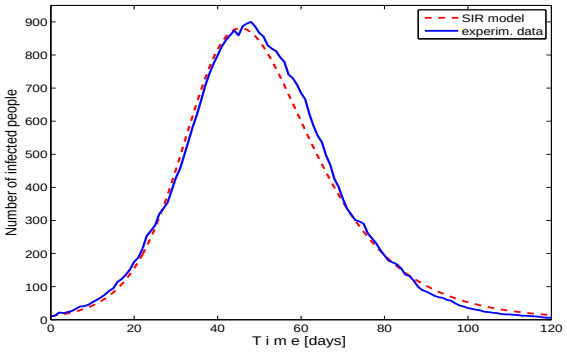

Figure 2: Measured number of infected people $I_k = 5000\, i_k$ from the agent based simulation population model (blue line) versus the modified SIR model fit

4.2 Results and discussion

In predicting the peak and the amplitude of the epidemic, we used the prior PDF with the following parameters: $\alpha_1 = 0.2$, $\alpha_2 = 0.5$, $\beta_1 = 0.1$, $\beta_2 = 0.15$, and $\nu_1 = 0.9$, $\nu_2 = 1.3$. Note that this prior will lead to an epidemic, because $\alpha/\beta > 1$ (hence in this example we are not interested in detection of an epidemic, only in its prediction). A synthetical dataset for syndromic observations (four sources) was generated based on algorithm (10) and $i(t)$ provided by our agent-based model. Parameters b^j and σ_j took values from $\{0.3, 0.5, 0.7, 0.9\}$ and $\{0.005, 0.006, 0.007, 0.008\}$ for $j = 1, 2, 3, 4$ correspondingly. Then this dataset was used to validate the proposed estimation/prediction framework. The observations were available on a daily basis (every 6 hours one source would report its observation count). Fig.3.(a) shows the histograms of the particle filter estimated values from α, β and ν, after using the data collected in the first 25 days of syndromic surveillance. The particle filter used $N = 5000$ particles. Fig.3.(b) depicts the estimated histogram (pdf) for the infected individuals in the community at $t = 25$ days. Fig.3.(c) shows: 20 overlayed predicted epidemic curves (corresponding to randomly chosen 20 particles, shown with blue lines); the "true" proportion of infected people (in fact the experimental curve from Fig.2, shown here in red) and the measurements collected over the surveillance period (green dots).

The numerical results in Fig.3.(a) indicate that parameters α and β can be estimated fairly accurately using the observations collected during the early stages

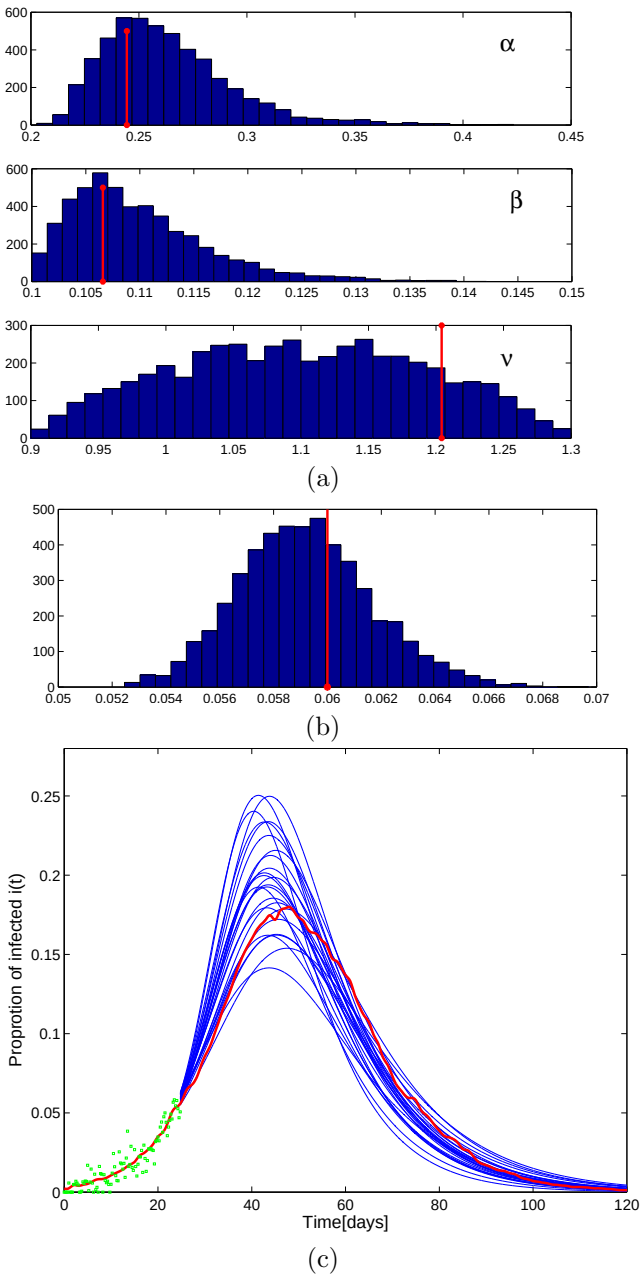

Figure 3: Estimation/prediction results from the particle filter after processing the data collected over 25 days of surveillance: (a) the histograms of estimated parameters α, β and ν (true values indicated by vertical red lines); (b) the histogram (pdf) of infected people after $t = 25$ days where the red line indicates the threshold value $\gamma = 0.06$ corresponding to $I = 300$ infected people; (c) Prediction results for a random sample of 25 particles (blue lines); red line is the experimental curve from Fig.2; green dots correspond to observations (i.e. z_ℓ^j / b^j)

of the epidemic. However, it appears that parameter ν cannot be estimated and therefore for its value one has to rely purely on prior knowledge. While this is unfortunate, it does not appear to be a serious problem since the prior on ν in practice is tight ($\nu \approx 1$). As a consequence, the prediction results shown in Fig.3.(b) are quite useful. The timing of the peak appears to be very accurate, while the amplitude is approximately in the interval from 0.15 to 0.22 (the true amplitude is 0.18).

In studying an epidemic outbreak it is often of interest to know when the epidemic actually started. Using the SIR dynamic model one could easily estimate this parameter via "retradiction" (or "back-tracking"). Indeed the uncertainty in the estimate will correspond to the amount of process noise $\mathbf{w}_k$ in (13).

In summary, it is worth noting that the mathematical models like (1) with non-linear "mixing" and "network" noise are quite general and are very common for the systems with the "supply and demand" constraitns resulting in the "logistic" growth. In addition to the celebrated examples from epidemiology, it is widely used to describe cooperative phenomena and collaborative behavior in complex physical and biological systems (computer worm propagation, wireless sensor network, phase transition, coupled chemical reactions, ecological competition, quorum sensing, tumor growth, community response to significant social events, etc, see [13], [10], [19], [4], [22]). With a straightforward change of notation the proposed algorithm can be easily employed for estimation problems in all these areas.

5 Conclusions

The paper studied the problem of predicting the dynamics of an epidemic (in particular the timing and the size of its peak). Typically the prediction of an epidemic is carried out using large scale agent-based simulations, which are rather costly to develop and run. In this paper we proposed an alternative framework formulated in the context of stochastic nonlinear filtering. This framework is based on a highly nonlinear SIR epidemic model with imprecisely known model parameters. The measurements of the number of infected people are very noisy, assumed to be collected by non-medical (syndromic only) means. The numerical results suggest that the adopted approach can provide very useful early predictions about the epidemic. Further work is required to verify the method using various experimental data sets.

6 Acknowledgement

The authors would like to thank Peter Dawson and Russell Connell for helpful discussions and for assistance in the numerical simulations.

References

[1] R. M. Anderson, C. Fraser, and A. C. Ghani. Epidemiology, transmission dynamics and control of SARS: the 2002-2003 epidemic. *Philos. Trans. R. Soc. B Biol. Sci.*, 359:1091–1105, 2004.

[2] R. M. Anderson and R. M. May. Population biology of infectious diseases: Part 1. *Nature*, 280:361–367, 1979.

[3] C. Fraser at al. Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science*, 324(5934):1557 – 1561, 2009.

[4] C.C. Zou at al. The monitoring and early detection of internet worms. *IIEEE/ACM Transactions on Networking*, 13(5):961–974, October 2005.

[5] S. M. Babin. Using syndromic surveillance systems to detect pneumonic plague. *Epidemiol. Infect.*, 138(1):18, 2010.

[6] S. Cauchemez and N. M. Ferguson. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *J. R. Soc. Interface.*, 5(25):885897, 2008.

[7] D.J. Daley and J. Gani. *Epidemic Modelling*. Cambridge Univ. Press, 1996.

[8] C. E. Dangerfield, J. V. Ross, and M. J. Keeling. Integrating stochasticity and network structure into an epidemic model. *J. R. Soc. Interface.*, 6(38):761–774, 2009.

[9] C. Dargatz. A diffusion approximation for an epidemic model. Technical report, Ludwig-Maximilian Universitat Munchen, 2007.

[10] H.P. de Vladar and I. Pen. Determinism, noise and spurious estimations in generalised model of population gropwth. *Physica A: Statistical Mechanics and its Applications*, 373(1):477–485, January 2007.

[11] A. Doucet, J. F. G. de Freitas, and N. J. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

[12] L.Dailey, R. E. Watkins, and A. J. Plant. Timeliness of data sources used for influenza surveillance. *J. Am. Medical Inf Ass.*, 14(5):177185, 2007.

[13] J. D. Murray. *Math. Biology*. Springer, USA, 2002.

[14] A. S. Novozhilov. On the spread of epidemics in a closed heterogeneous population. *Math. Bioscienses.*, 215(2):177185, 2008.

[15] N. Oudjane and C. Musso. Progressive correction for regularized particle filters. In *Proc. 3rd Int. Conf. Information Fusion*, Paris, France, 2000.

[16] B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman filter*. Artech House, 2004.

[17] B. Ristic, A. Skvortsov, and M. Morelande. Predicting the progress and the peak of an epidemic. In *Proc. IEEE Inter. Conf. Acosutic, Speech and Signal Processing (ICASSP 2009)*, pages 513–516, Taipei, Taiwan, April 2009.

[18] I. Sazonov, M. Kelbert, and M. B. Gravenor. The speed of epidemic waves in a one-dimensional lattice of SIR models. *Math. Model. Nat. Phenom.*, 3(4):28–47, 2008.

[19] F. Schweitzer and R. Mach. The epidemics of donations: Logistic growth and power-laws. *PLoS One*, 3(1):1458–1463, 2008.

[20] A. Skvortsov, R. Connell, P. Dawson, and R. Gailis. Epidemic modelling: Validation of agent-based simulation by using simple mathematical models. In *International Congress on Modelling and Simulation (MODSIM 2007)*, pages 657–662, Christchurch, New Zealand, December 2007.

[21] A. Skvortsov, R. Connell, P. Dawson, and R. Gailis. Epidemic spread modeling: Alignment of agent-based simulation with a simple mathematical model. In *Proc. Int. Conf. Bioinformatics & Comput. Biology*, pages 487–890, Las Vegas, USA, June 2007. CSREA Press.

[22] A. Skvortsov, B. Ristic, and M. Morelande. Networks of chemical sensors: a simple mathematical model for optimisation study. In *5th Int. Conf. on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 2009)*, pages 60–66, 2009.

[23] P. D. Stroud, S. J. Sydoriak, J. M. Riese, J. P. Smith, S. M. Mniszewski, and P. R. Romero. Semi-empirical power-law scaling of new infection rate to model epidemic dynamics with inhomogeneous mixing. *Math. Biosciences*, 203:301–318, 2006.

[24] Z. Toroczkaia and H. Guclub. Proximity networks and epidemics. *Physica A: Statistical Mechanics and its Applications*, 378(1):68–75, May 2007.

[25] J. Walden and E. H. Kaplan. Estimating time and size of bioterror attack. *Emerging Infectious Diseases*, 10(7):1202–1205, July 2004.