Parameter inference in small world network disease models with approximate Bayesian Computational methods

David M. Walker\textsuperscript{a,}\textsuperscript{*}, David Allingham\textsuperscript{a}, Heung Wing Joseph Lee\textsuperscript{b}, Michael Small\textsuperscript{c}

\textsuperscript{a} Centre for Complex Dynamic Systems and Control, School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW 2308, Australia

\textsuperscript{b} Department of Applied Mathematics, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

\textsuperscript{c} Department of Electronic & Information Engineering, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 5 January 2009
Received in revised form 22 September 2009
Available online 6 October 2009

PACS:
02.50.-r
05

Keywords:
Small world networks
Approximate Bayesian Computation
Stochastic simulation & inference
Epidemiological models

ABSTRACT

Small world network models have been effective in capturing the variable behaviour of reported case data of the SARS coronavirus outbreak in Hong Kong during 2003. Simulations of these models have previously been realized using informed “guesses” of the proposed model parameters and tested for consistency with the reported data by surrogate analysis. In this paper we attempt to provide statistically rigorous parameter distributions using Approximate Bayesian Computation sampling methods. We find that such sampling schemes are a useful framework for fitting parameters of stochastic small world network models where simulation of the system is straightforward but expressing a likelihood is cumbersome.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

In the Chinese province of Guandong during November 2002, the Severe Acute Respiratory Syndrome (SARS) appeared before spreading to Hong Kong SAR (Special Administrative Region) in early 2003. The duration of the epidemic in Hong Kong was 105 days. This paper focuses on modelling the transmission of SARS within Hong Kong during this period. International travel networks facilitated the further spread of the virus to other cities. The effect of the SARS coronavirus was damaging both economically and socially and was indeed tragic for some individuals. The impact and widespread concern about a virus such as SARS causes it important for government and officials to make informed decisions regarding measures to deal with the virus and so reassure the public. Such measures might include quarantine, and questions about what type of quarantine require answers and risk assessment. Furthermore, as the outbreaks in the Prince of Wales Hospital [1,2] showed, even hospitalization cannot be considered an effective quarantine, nor could demanding people stay in their buildings, as the peculiarities of the virus spread in the Amoy Gardens Housing Estate demonstrated [3, 1].

Mathematical models can help inform about the spread and the rate of transmission of the disease throughout the population, as well as the likely effectiveness of control measures. Models, however, are only as good as their underlying assumptions allow. For example, standard epidemiological models [4] such as susceptible-infected-removed (SIR) equations assume a constant infection and recovery rate across all individuals at any given time. That is, each infected individual has...
an equal chance of infecting any susceptible in the population regardless of its social structure. People know and come into contact with a varying number of individuals – the social secretary of a football club would meet many more individuals than a housebound person, for example – and so it makes sense to consider different infection rates for different individuals or assume some kind of (social) network structure within models.

It is possible to imagine more complicated models in which the infection rate for individuals is allowed to vary based on physiological differences, but we do not consider such heterogeneity in this paper. Two reasons for not doing so are that the number of parameters to be estimated would greatly increase, and our focus is on presenting a method for parameter inference. Nonetheless, it is clear how our methods can be readily generalized to include more complicated heterogeneous models. To focus our discussion we adopt an existing model [5]. This epidemiological model [5,6], in which an infection is mediated via a contact network with so-called “small world” properties [7], has been shown to exhibit many of the features characteristic in the SARS epidemic spread in Hong Kong. In particular, a small number of individuals showed a great propensity to infect others, so-called “super-spreaders” (e.g., Ref. [8]).

The data such models aim to be consistent with are the daily new SARS cases reported in Hong Kong between 15 February 2003 and 15 June 2003. The model and parameters are discussed more fully in Section 2, but in the above studies, the consistency between data and model outputs has been realized using educated guesses for plausible parameter values [5].

In this report we aim to do better, and report a generic method whereby more statistically rigorous fits to the model parameters can be found.

The method of fitting is a form of Markov chain Monte Carlo (MCMC) [9] sampling. The use of MCMC to fit stochastic epidemics is not new but since Small’s small world disease model is probabilistic, MCMC appears to be a most useful methodology to attempt parameter fitting. For such network models as are employed here, however, the likelihood functions are complicated and often intractable, and so standard MCMC algorithms, which require explicit likelihood functions, cannot be applied. Instead, we turn to a form of MCMC called Approximate Bayesian Computation (ABC) [10] which instead relies on model simulation and realization. Summary statistics are used to compare simulated data from the model with the observed data; when the summary statistic is sufficient and the values from the simulated data exactly match those from the observed data, samples are obtained from the exact posterior distributions of the model parameters. When no sufficient statistic is available, acceptable parameter values are drawn from an approximation of this distribution.

The application of this technique to network models is very general, and can be applied to any model for which simulations can be performed. This offers hope that more complex and realistic network models can be considered in the future. The version of ABC which we use here is the particle rejection control (PRC) algorithm of Ref. [11], which can reduce the number of model simulations required to achieve a reasonable number of samples for the approximated posterior distribution.

The outline of the remainder of this paper is as follows: in Section 2 we reprise the small world disease model used by Small and Tse [6] and introduce the methodology of ABC parameter fitting in Section 3. The data and ABC results are presented and analyzed in Section 4 before a concluding discussion.

2. Small world network model

The conventional SEIR (susceptible-exposed-infected-removed) compartment models, both deterministic and stochastic, suppose that all individuals have an equally small chance (probability) of being infected [4]. In this section, we describe an alternative model structure where infection can only occur along certain predefined paths [6] which link subsets of individuals.

We arrange the population in a regular grid and suppose each individual (node) can infect its four immediate neighbours (horizontally and vertically), as well as a random number of remote nodes (long-range links). The motivation for this model is that the near-neighbour links provide a model of infection within households and between adjacent households. In a dense urban environment such as Hong Kong, it provides a model for the spread of contagion among families based on residential units. However, in an effort to simplify (and homogenise) the model structure we do not build separate family units into our model, rather, we assume that individuals residing in close proximity to one another are likely to spread an infection between them.\(^1\) In addition to spread in residential environments, we build a random number of long-range links into the model. These long-range links account for the fact that individuals are free to move within in the community and will therefore have direct contact with others in different residential locations. We feel justified in using a random and long-tailed distribution for these connections (rather than the fixed number associated with the regular grid) because the number (and diversity) of contacts an individual will have will vary greatly between individuals. Moreover, the probability of infecting near neighbours (considered to be members of the same household) is likely to be distinct from the probability of infecting remotely-linked neighbours (say, daily acquaintances).

The local near-neighbour links ensure that individual nodes are highly clustered, whereas the random long-range links will cause the average path length (or, equivalently, network diameter) to be small. Combined, these two factors will

---

\(^1\) Evidence from the SARS outbreak of 2003 supports this. The infection did spread within families, buildings and entire housing estates based on geographical proximity, across multiple scales.
generate a (clustered) small-world network [7]. With this network structure we propose a discrete-time model with a time step of one day — this is the resolution of data available to us.

We consider the model with four distinct groups (compartments) of individuals, S, P, I and R, corresponding to those that are susceptible, prone (exposed), infected, and removed. No re-infection was observed, and so the removed group corresponds to individuals that have recovered, expired or been isolated from the community. The structure of this model (which is a discrete-time version of the usual SEIR model [12]) is depicted in Fig. 1. Let \( p_1 \) denote the daily probability of a node infecting each of its nearest neighbours and let \( p_2 \) denote the probability of a node infecting its remotely-linked neighbours on one day. Suppose that there are \( n_1 \) near neighbours and \( n_2 \) remote neighbours. We allow \( n_2 \) to be fixed for each node but randomly distributed across nodes, that is, node \( i \) has \( k_i \) links (so that \( E(k_i) = n_2 \)) which, in order to explicitly model the small world network structure, are distributed according to an exponential distribution with mean \( \mu \).

Hence, although the near-neighbour links are bi-directional, the remote-neighbour links are only one-way. The induced small world network structure implies that the number of links between any two nodes in the model is small, and, the probability that a node has a given (large) number of links is exponentially distributed. The probability of a transition from state \( P \) to state \( I \) on a given day is \( r_0 \), and the probability of a transition from \( I \) to \( R \) is \( r_1 \). We always take \( n_1 = 4 \) to indicate the immediate horizontal and vertical neighbors if the population is imagined on a regular grid.

The model above is further adjusted to mimic executive decisions which were made during the period of the epidemic in the Hong Kong SAR. Three phases of the epidemic are captured by including change-points for the infection rate \( p_2 \) and the removal rate \( r_1 \), reflecting stages of intervention in the crisis. In the first phase SARS spread freely and so we have a relatively high \( p_2 \) and low value for \( r_1 \). As rudimentary hygiene measures were introduced, the value of \( p_2 \) will fall, and as a quicker response time in hospitalizing suspected cases developed \( r_1 \) would rise in value. In a third phase, where very strict control measures and increased public awareness were prevalent, we consider an even smaller value of \( p_2 \) and a higher value of \( r_1 \).

The actual changepoint models for \( p_2 \) and \( r_1 \) used by Small and Tse [6] were

\[ p_2 = p_{21} I(t \leq 30) + p_{22} I(30 < t \leq 60) + p_{23} I(t > 60) \]
\[ r_1 = r_{11} I(t \leq 20) + r_{12} I(20 < t \leq 40) + r_{13} I(t > 40) \]

where \( t \) is measured in days and \( I(\cdot) \) is an indicator function, \( p_{21}, p_{22} \) and \( p_{23} \) are values for the parameter \( p_2 \) in different time periods, and \( r_{11}, r_{12} \) and \( r_{13} \) are similarly related to \( r_1 \). We note that in our inference experiments we shall only attempt to estimate plausible distributions of the rate parameters while preserving the changepoint (time) values as used by Small and Tse [6]. In principle, ABC-PRC can be used to also estimate these, for example, Ref. [13] considers a changepoint model for analysis of DNA segmentation.

We use MCMC ABC-PRC methodology to provide plausible estimates of the transmission probabilities \( (p_1, p_{21}, p_{22}, p_{23}, r_0, r_{11}, r_{12}, r_{13}) \) and the exponential parameter \( \mu \) which induces the small world structure of the model. MCMC sampling in a Bayesian framework has been used to provide estimates of parameters of an extended SEIR model, which includes hospitalization and death compartments to describe the spread of the SARS virus in Shanxi province of China between February 2003 and late May 2003 [14], but they had a richer data set which facilitated the use of likelihoods.

3. Approximate Bayesian computation

The method of MCMC is particularly potent when used with the Bayesian paradigm [9]. Here, unknown model parameters and data are assumed to be random variables and one is interested in the posterior distribution of the parameters given the observed data. This is achieved by appealing to Bayes’ rule which expresses the posterior as (proportional to) the product of the likelihood – the probability of the data given a particular set of parameters – and prior distributions describing our knowledge or assumptions about the parameters before observing the data. If, for notational convenience, we represent the model parameters as \( \alpha \) and the data as \( D \) then mathematically the posterior is given by

\[ \pi(\alpha | D) \propto L(D | \alpha) \pi(\alpha). \]
We can use the posterior distribution to obtain quantities of interest. For example, the mean value of the parameters is given by
\[ \hat{\alpha} = \int \alpha \pi(\alpha | D) d\alpha. \]  
(2)

Such integrals, however, are analytically intractable for most applications, and so must be numerically approximated. An effective method of approximation is to draw samples, \( \alpha_n \), from the posterior in such a way that \( \sum_{n=1}^{N} \alpha_n / N \) converges to the integral for large enough \( N \). Markov chain Monte Carlo is an effective way of sampling from the posterior using only the terms on the rhs of (1), i.e., without knowing the constant of proportionality, and algorithms such as Metropolis–Hastings (MH) dictate update rules. Specifically, a sequence of samples from the posterior can be generated as follows:

An MH algorithm
1. Start with an initial value for \( \alpha_0 \). Set \( k = 1 \).
2. Propose a new value of \( \alpha' \) from the prior or other proposal distribution \( q(\alpha', \alpha_{k-1}) \).
3. Set \( \alpha_k = \alpha' \) with probability
   \[ \min \left\{ 1, \frac{L(D|\alpha') \pi(\alpha') q(\alpha_{k-1}, \alpha')}{L(D|\alpha_{k-1}) \pi(\alpha_{k-1}) q(\alpha', \alpha'_{k-1})} \right\} \]
   otherwise \( \alpha_k = \alpha_{k-1} \).
4. Advance \( k = k + 1 \) and goto step 2 unless convergence realized.

In the above, \( q(\cdot, \cdot) \) is a proposal probability distribution. A judicious choice of \( q(\cdot, \cdot) \) can improve the mixing properties of the Markov chain and so explore the parameter space better.

The above inference scheme, although very powerful, has the drawback that an explicit likelihood function is needed. This can be a difficult and onerous task especially for dynamic network models of the type we consider here. (It is far easier to simulate realizations of the model than to consider the complexities of a possibly intractable likelihood.) This is where the method of Approximate Bayesian Computation opens the door to reliable parameter inference of complex dynamic models. Here, summary statistics of model realizations and the data are compared using a suitable distance function \( \rho \). Parameters which give rise to realizations that are close to the data are accepted as samples from an approximation of the posterior distribution. An MCMC-like ABC algorithm proposed by Marjoram et al. [10] is as follows:

An ABC algorithm
1. Start with an initial value for \( \alpha_0 \). Set \( k = 1 \).
2. Propose a new value of \( \alpha' \) from some proposal density \( q(\alpha', \alpha_k) \).
3. Generate a data set \( x^* \sim f(x|\alpha') \).
4. Set \( \alpha_k = \alpha' \) with probability
   \[ \min \left\{ 1, \frac{\pi(\alpha') q(\alpha_{k-1}, \alpha')}{\pi(\alpha_{k-1}) q(\alpha', \alpha_{k-1})} 1(\rho(S(x^*), S(D)) \leq \epsilon) \right\} \]
   otherwise \( \alpha_k = \alpha_{k-1} \).
5. Advance \( k = k + 1 \) and goto step 2 unless convergence realized.

In the above, \( x^* \sim f(x|\alpha') \) is a simulated realization of our (stochastic) small world network disease model. A commonly chosen and useful proposal distribution is the random walk model [9]
\[ \alpha' = \alpha_{k-1} + \omega' \]  
(5)
where \( \omega' \) is often chosen to be a zero-mean Gaussian with a small variance \( v \). An advantage of such a choice is that \( q(\cdot, \cdot) \) is symmetric and the proposal probability ratio in step 4 is unity. Furthermore, if we consider uniform prior distributions then the acceptance probability is unity for all realizations that have a statistic score \( S \) within a tolerance distance of \( \epsilon \) of the data score, i.e.,
\[ \rho(S(x^*), S(D)) \leq \epsilon. \]  
(6)

An adjustment of \( q(\cdot, \cdot) \) and step 2 allow the easy addition of any desired parameter constraints (such as integer-valued parameters, positivity or multiparameter inequality constraints) to the model.

The result of such an algorithm is a series of highly correlated samples from the distribution \( \pi(\alpha | \rho(S(x), S(D)) \leq \epsilon) \). If \( \epsilon \) is sufficiently small and the summary statistics \( S(\cdot) \) are near-sufficient then this distribution is a reasonable approximation of \( \pi(\alpha | D) \).

As can be seen no evaluation of complicated likelihoods are necessary; only the ability to generate realizations of the model is required. The subtleties, however, are now in the setting of a realizable tolerance \( (\epsilon) \), a good choice of summary statistic \( \rho(S(\cdot)) \) and a suitable distance function \( \rho \). We choose \( \rho \) to be the Euclidean distance and we discuss it, \( \epsilon \) and the chosen summary statistic later, after first introducing a more efficient algorithm.
Fig. 2. The revised case data of the SARS epidemic in Hong Kong (blue). We take this data as a proxy for the number of new infectious individuals in the population per day. The red curve is a moving average with window size 10 of the blue line. This red curve is used to compare simulated realizations and the data in the ABC-PRC algorithm. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

It is clear that demanding as small an $\epsilon$ as possible is better in terms of the closeness of the approximate distribution to the true posterior distribution but having too small an $\epsilon$ can lead to many realizations being simulated per accepted parameter sample. This can become prohibitively time consuming. Fortunately, there is a variant to the above MCMC ABC algorithm which has the potential to greatly reduce the number of model evaluations per accepted parameter sample. It is the particle rejection control algorithm ABC-PRC proposed by Sisson et al. [11]. A decreasing sequence of tolerances is considered and $N$ samples are generated for each tolerance. Initially, the parameters are sampled from their priors or some other suggested distribution. The sampled parameters at subsequent (lower) tolerances are generated by sampling from those previously accepted and applying a perturbation. A simplified version of their algorithm which we use is the following:

An ABC-PRC algorithm

1. Initialize $\epsilon_1, \epsilon_2, \ldots, \epsilon_T$, a decreasing sequence of tolerances and specify an initial sampling distribution $\nu_1$ for the parameters $\alpha$. Set population indicator $t = 1$.
2. Set particle indicator $i = 1$ (first of $N$ particles/samples); note that sample number is $t$ with tolerance $\epsilon_t$.
   2a. If $t = 1$, sample $\alpha^{**} \sim \nu_1(\alpha)$ independently from $\nu_1$. If $t > 1$, sample $\alpha^*$ from the previous population $\{\alpha^{(i)}_{t-1}\}$ uniformly and perturb the particle using a (symmetric) Gaussian distribution (à la the random walk model). Generate a data set $x^{**} \sim f(x|\alpha^{**})$. If $\rho(S(x^{**}), S(D)) \geq \epsilon_t$ goto 2a.
   2b. Set $\alpha^{(i)} = \alpha^{**}$. If $i < N$ increment $i = i + 1$ and goto 2a.
3. If $t < T$, increment $t = t + 1$ and goto 2.

We note that the ABC-PRC algorithm of Ref. [11] is more complicated than the above in that it allows for non-uniform (weighted) sampling of the previous population and non-Gaussian, non-symmetric perturbations. We have found the above simplified version adequate for our purposes in terms of reducing the number of model simulations required to obtain $N$ samples of the parameter population at a given tolerance level. One major advantage of this algorithm is that no test for convergence is required, as the particles at every stage are uncorrelated [11].

4. Analyzing the reported SARS data

The data we have available for use to estimate the unknown parameters of the small world network model are the number of reported new cases per day in Hong Kong during the period of the epidemic (see Fig. 2). This information is neither the number of removed nor number of infected individuals, i.e., one of the compartments in our model. We must thus construct a reasonable combination of model compartments to compare with the data. We take the following output from the model as a good proxy for the number of new infectives per day $x(t)$ to compare to the case data:

$$x(t) = \Delta I(t) + \Delta R(t) = I(t) - I(t - 1) + (R(t) - R(t - 1))$$

where $I(t)$ are the number of model infectious individuals each day and $R(t)$ denotes the number of removals.

In Ref. [6] the total number of infected individuals throughout (i.e., integrated over) the period of the epidemic was used as a filter to identify model realizations which were similar to the data. This suggests that such a quantity is a good candidate to choose as the summary statistic to compare to the data in the ABC algorithms. It turns out, however, that such a choice does not result in inferred parameters whose simulated realizations are similar to the data. Instead we typically obtain
model realizations which have high \( p_2 \) probabilities (long-range infections) and after \( t = 60 \), a removal rate of \( r_{13} \) very close to one.

It also seems natural to directly compare \( x(t) \) with the observed case data but since the actual data appears quite noisy, with large deviations from a smooth, deterministic curve, the comparison of such a data set with similar stochastic realizations in the ABC-PRC algorithm can produce large errors, and so have a very low acceptance rate. Since the system is such that \( \Pr(x^* = D) \approx 0 \), we must use a summary statistic, and so choose one which allows for such stochastic variation whilst preserving the gross features of the data which we wish to represent. Consider, for example, the case when the peak is offset by only one or two days while other aspects of the data are well-represented, a case in which we may wish to accept the generating parameters. To allow for such variation, we have chosen \( S \) to be a moving-average filter, thus allowing comparisons to be made between smoothed versions of the observed and simulated data sets. Fig. 2 shows the actual data in blue and a smoothed version with a window size of 10 in red. This is the \( S(D) \) we compare our realizations \( S(x^*) \) to using \( \rho \) as Euclidean distance. This choice of \( S \) and \( \rho \) allows us to accept parameter values which produce data sets matching the trajectory of the data without being overly sensitive to the exact realization obtained via simulation. While no sufficient statistic for this model is known to us, the moving-average filter preserves important features of the data, as discussed below.

The use of the Euclidean distance restricts accepted parameter values to those whose realizations are close to \( D \) across the whole data set. The length of the smoothing window determines the scale of the resulting errors in the ABC-PRC algorithm. If the window length is too short, we return to the problem above, of comparing relatively noisy data sets and obtaining large errors. To accept even very good parameters we would need a large \( \epsilon \), thus also accepting many bad parameter sets. Conversely, if the window length is too long, \( S \) approximates the mean of the data, and so information about important local features, such as regions of rise and fall and turning points, is lost. We have chosen 10 as a moderate window length relative to the size of the data. Provided that the above extremes are avoided, an increase in the window length is balanced by a smaller value of \( \epsilon \) and vice versa, and the ABC-PRC algorithm provides us with an automatic, systematic reduction in \( \epsilon \) while maintaining a reasonable computation time.

We note that \( \rho(0, S(D)) \) is slightly higher than 190 and that in the ABC-PRC it is only when the tolerance \( \epsilon_1 \) is lower than this value do we begin to see plausible estimates of the posterior distribution revealing themselves. It is, however, better to start with an \( \epsilon_1 \) much greater than this value to find an initial set of particles (sampled parameter sets) as finding subsequent particle sets at lower tolerances takes fewer model realizations than going straight to low tolerances to begin with. As such we consider the following decreasing set of tolerances in our ABC-PRC procedure:

\[
\epsilon = \{1000, 500, 250, 190, 150, 125, 100, 75, 50, 45, 40\}.
\]

In MCMC, the use of uninformative priors means that the posterior distributions are entirely informed by the observed data, and this principle was followed here. For the network parameter \( \mu \), a wide uniform prior distribution was chosen, mainly enforcing positivity, with \( \mu \sim \text{Unif}[0, 10] \). The remaining parameters, the transition probabilities, used uninformative \( \text{Unif}[0, 1] \) priors. A stronger reliance could be placed on existing parameter values by the use of more informative priors.

Table 1 summarizes the number of model realizations taken to find the 250th particle for each tolerance. In the perturbation part of step 2a of the ABC-PRC algorithm we use a zero-mean Gaussian with a standard deviation of 5% of the range of the sampled parameter values.

Given the statistic, distance function, and tolerances above we use the ABC-PRC algorithm to generate 250 samples for each tolerance of the nine parameters \( \mu, p_1, p_{21}, p_{22}, p_{23}, r_0, r_{11}, r_{12} \) and \( r_{13} \). The improvement gained in the resolution of the sampled distributions can be seen by comparing Fig. 3 with Fig. 4 which are obtained using tolerances \( \epsilon = 1000 \) and \( \epsilon = 40 \) respectively. We see in Fig. 4 that the sampled parameters obtained at the stricter tolerance start to reveal informative, unimodal distributions. We summarize the means of these estimated distributions in Table 2. In the paper of Ref. [6] the parameters shown in row one of Table 2 are the ones used. In the second row of the table we present the mean of the samples obtained using our strictest tolerance \( \epsilon = 40 \).

---

**Table 1**

| \( \epsilon \) | Number of realizations |
|----------------|-----------------------|
| 1000           | 789                   |
| 500            | 518                   |
| 250            | 692                   |
| 190            | 5,127                 |
| 150            | 4,200                 |
| 125            | 35,670                |
| 100            | 58,063                |
| 75             | 11,512                |
| 50             | 90,600                |
| 45             | 67,848                |
| 40             | 112,110               |

A comparison of the number of model realizations required to obtain the last particle (sampled parameter set) at the given tolerance. We see that as the tolerance becomes stricter the number of realizations increases significantly, however, this is still much smaller than the number we would expect if we were to begin with the strictest tolerance of all.
Fig. 3. The histograms and scatter plots of 250 accepted samples of the ABC-PRC algorithm at the slackest tolerance $\epsilon = 1000$. The prior distributions of each parameter were uniform as is evident from the samples.

Fig. 4. The 250 accepted samples of the parameters at our strongest tolerance level of $\epsilon = 40$. We can see recognizable distributions developing as the tolerance lowers but at the expense of many more model realizations. There still appears to be some correlations between the sampled parameters; lowering the tolerance would clarify whether these are a true feature of the model.

We mention three aspects of Table 2 together with Fig. 4. The first point to note is that apart from the parameter $r_0$, the means of the estimated posterior distributions are quite different from the values used by Small and Tse [6], which lie in the posterior distributions’ tails.
Table 2
A comparison of the value of the parameters used in Ref. [6] and the mean value of the ABC-PRC inferred sampled distributions. Histograms and scatter plots of the sampled parameters are shown in Fig. 4.

|     | $\mu$ | $p_1$ | $p_{21}$ | $p_{22}$ | $p_{23}$ | $r_0$ | $r_1$ | $r_2$ | $r_3$ |
|-----|-------|-------|----------|----------|----------|-------|-------|-------|-------|
| Small & Tse | 2.4   | 0.02  | 0.1      | 0.02     | 0.01     | 1/6.4 | 0.025 | 0.1   | 0.333 |
| Mean ABC   | 0.55  | 0.16  | 0.47     | 0.20     | 0.05     | 1/5.5 | 0.19  | 0.52  | 0.96  |

Fig. 5. A histogram of the accepted errors at tolerance $\epsilon = 40$. This is the shape we expect of the errors in an ABC simulation. It suggests that the tolerance can be lowered further to improve the results displayed in Fig. 4.

We do, however, observe the same general trend about the assumptions of the spread of the disease. That is, early on in the epidemic we note high values for the rate of long-range transmission together with a lower removal rate. As we proceed through the course of the epidemic the long-range transmission rate falls and the removal rate rises. A final point of observation is the apparent correlation between the transmission rate $p_1$ and the network parameter for acquaintances $\mu$. There appears to be a negative correlation between these two parameters, however, whether or not this suggestion is fact requires further simulation at ever lower tolerances, with correspondingly longer computation times. We can see from Fig. 5 that the distribution of accepted errors at $\epsilon = 40$ suggests we can lower the tolerance still further but at the expense of ever more simulated realizations of the model.

Finally, in Fig. 6 the number of infected individuals at different times during the course of the outbreak are shown for all accepted realisations at the lowest tolerance level, together with the data. We can see that the realizations have the capability to capture many features of the observed (revised) case data, only failing at the peak of the revised case numbers. This suggests that as good as the model is, more work is required, such as investigating other social network structures, perhaps scale-free distributions rather than exponential, in an effort to attain this peak value.

5. Conclusion

We have considered the problem of estimating plausible values and distributions of parameters of a small world network disease transmission model. The model in question has been shown to be capable of capturing the behaviour of the observed case data seen in the spread of the SARS-CoV outbreak in Hong Kong during 2003 [5]. The model is probabilistic both in terms of the mode of transmission between individuals and the connections between individuals. In particular, individuals are assumed to be connected, and hence capable of transmitting the virus, to a fixed number of close neighbours (relatives) and a small number of more distant individuals, the latter number being identified from an exponential distribution. This leads to a small world network model of the connections between individuals and consequently the possible transmission vectors for the disease.

Since the model is probabilistic, the method of Markov chain Monte Carlo integration by sampling is particularly amenable to fitting parameters of the model. In particular, given the complexities of a network model and the complication of expressing the likelihood in a simple form, the method of ABC using particle rejection control proved to be a most effective way of carrying out such inference. We obtained estimates of the small world network parameter $\mu$ and the various transmission probabilities between the compartments in the population, i.e., $r_0$ the probability at which exposed (prone) individuals become infectious, $r_1$ the probability at which an infected individual is removed from the population through effective quarantine, recovery or death, $p_1$ the probability at which an infectious individual infects a close susceptible neighbour, and $p_2$ the probability of an infectious individual infecting a susceptible long-range linked neighbour.
We found that plausible estimates of parameter distributions are obtainable using MCMC ABC. Furthermore, our estimated distributions were consistent with parameter values used by Small and Tse [6] chosen based on medical opinion, although those values were located in the tails of our distributions, indicating that they were not the optimal values.

In summary, we have shown that parameters of probabilistic models can be fitted to observed data using MCMC ABC methods even when the data is stochastic and the models complex. The use of ABC circumvents any difficulties in expressing a computable likelihood and, provided that a suitable statistic is used, in our case a moving-average of the stochastic data, informative and plausible distributions of model parameters can be inferred to satisfaction.

Acknowledgments

This work was supported in part by a Hong Kong Polytechnic University Research Grant G-YE08. DMW would like to thank the University of Newcastle for accommodation and support during a recent visit.

References

[1] C.A. Donnelly, A.C. Ghani, C.M. Leung, A.J. Hedley, C. Fraser, S. Riley, L.J. Abu-Raddad, L.M. Ho, T.Q. Thach, P. Chau, K.P. Chan, T.H. Lam, L.Y. Tse, T. Tsang, S.H. Liu, J.H. Kong, E.M. Lau, N.M. Ferguson, R.M. Anderson, Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. The Lancet 361 (2003) 1761–1766.
[2] B. Tomlinson, C. Cockram, SARS: Experience at the Prince of Wales hospital, Hong Kong. The Lancet (commentary) 361 (2003) 1486–1487.
[3] D. Cyranoski, A. Abbott, Apartment complex holds clues to pandemic potential of SARS, Nature (news article) 423 (2003) 3–4.
[4] L.J.S. Allen, An Introduction to Stochastic Processes with Applications to Biology, Pearson Prentice Hall, 2003.
[5] M. Small, P. Shi, C.K. Tse, Plausible model for propagation of the SARS virus, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences E87-A (9) (2004) 2379–2386.
[6] M. Small, C.K. Tse, Small world and scale free model of transmission of SARS, International Journal of Bifurcation and Chaos 15 (5) (2005) 1745–1755.
[7] S.H. Strogatz, Exploring complex networks, Nature 410 (2001) 268–276.
[8] M. Small, C.K. Tse, D.M. Walker, Super-spreaders and the rate of transmission of the SARS virus, Physica D 215 (2006) 146–158.
[9] D. Gamerman, Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Chapman & Hall, 1997.
[10] P. Marjoram, J. Molitor, V. Plagnol, S. Tavaré, Markov chain Monte Carlo without likelihoods, Proceedings of the National Academy of Sciences of the United States of America 100 (23) (2003) 15324–15328.
[11] S.A. Sisson, Y. Fan, M.N. Tanaka, Sequential Monte Carlo without likelihoods, Proceedings of the National Academy of Sciences of the United States of America 104 (6) (2007) 1760–1765.
[12] J.D. Murray, Mathematical Biology, 2nd ed., in: Biomathematics Texts, vol. 19, Springer, 1993.
[13] D. Allingham, R.A.K. King, D. Nur, K.L. Mengersen, Further DNA segmentation analysis using approximate bayesian computation, in: Proceedings of IASC2008, December 5–8, Yokohama, Japan, 2008.
[14] E.S. McBryde, G. Gibson, A.N. Pettitt, Y. Zhang, B. Zhao, D.L.S. McElwain, Bayesian modelling of an epidemic of severe acute respiratory syndrome, Bulletin of Mathematical Biology 68 (4) (2006) 869–917.