Proceedings of the 2007 Winter Simulation Conference
S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds.

DETERMINING EFFICIENT SIMULATION RUN LENGTHS FOR REAL TIME DECISION MAKING

Russell Cheng

Highfield
School of Mathematics
University of Southampton
Southampton, SO17 1BJ, U.K.

ABSTRACT

Suppose that there are a number of alternative ways of operating a system, and a performance measure is available for comparing them. Simulation runs can be carried out to estimate this measure for different alternatives, but there are too many for all to be examined because there is a strict limit to the time available for simulations. If the accuracy with which the performance measure can be estimated increases with run length, then a balance has to be made between making few runs where the performance measure is accurately estimated and making a large number of runs but with the performance measure poorly estimated. We analyse how the best run length can be selected to ensure that an alternative is found with a good performance measure. This problem can arise in real time decision making, and we give a real example arising in the provision of fire service emergency cover.

1 INTRODUCTION

We consider the use of discrete event simulation in situations where quick decision making is needed for a combinatorially hard problem. Such decisions may need to be taken frequently, so their number can, in time, be large. But each decision is a 'one-off' reaction to a random event, and so each can be studied independently. We suppose that the simulation model has to be detailed to properly capture the logical complexity of the behaviour of the system and so is expensive to run. The limited total run-time therefore places a serious restriction on how many runs can be made.

In this paper we focus on one particular form that each 'one-off' decision can take and discuss a problem of simulation run length design that it gives rise to. We describe this decision problem first, and then give examples of where it might arise.

We suppose that the core decision problem is essentially one of optimization. The decision is simply one where in response to some, possiblly external, triggering event, we have to adjust the system. However this adjustment can be done in a large number of alternative ways. We would like to choose the best alternative, as measured by some performance measure. However for each alternative we can only examine the performance of the system by carrying out a simulation run to evaluate the consequences of choosing that alternative. The simulation is subject to stochastic variation so that the accuracy of the measure of performance is dependent on the length of the simulation run – the longer the run the more accurate the estimate of the performance measure.

We shall assume that little is known about the relative merits of the different alternatives, so we therefore have to examine a number in trying to select the best.

The question that we address is the following: If there is a known, but limited, amount of time available for making the simulation runs, how long should each run be? There is clearly a balance to be made between making long runs, where an accurate measure of performance can be obtained, but only for a few alternatives, and making a large number of short runs, which allows a large number of alternatives to be considered, but none very accurately.

We could treat this problem within the context of an numerical optimization procedure, say a search or stochastic optimization algorithm. However this would involve a a potentially complicated study of the characteristics of such an optimization process and how these influence the simulation run length problem just described.

Instead we consider the considerably simpler situation, where we are willing to look at just a number of *randomly* selected alternatives. The 'optimization' process then reduces to simply selecting the alternative which yields the best estimate of performance measure. The specific question we address is: How good a solution is given by this very simple procedure?

Before considering this problem in detail we first give a number of examples where the decision making scenario just described might arise.

One example occurs in air traffic management. The Eurocontrol Central Flow Management Unit (EUROCONTROL CFMU, 2002) manages slot allocation using near real-time simulation. The alternatives are the different slot allocations that might be made, and simula-

tions are needed to to determine the likely delays for each possible slot pattern.

Another example is in the study of military engagements (Wrigley and Taylor, 2003). A battle typically hinges on a number of key decisions of how to deploy forces, taken at critical moments in the battle. A real time analysis of possible alternatives is in principle possible if simulation runs could be conducted fast enough made during the course of an engagement.

The final example involves the provision of fire service emergency cover. We discuss this example in detail in the following Section, as it will provide the motivation for the formulation and analysis of the more general problem carried out in Sections 3 and 4.

An extended abstract of this paper was presented at the INFORMS Simulation Workshop held at INSEAD in July 2006.

2 AN EXAMPLE

2.1 Fire Service Emergency Cover

To give our general problem some focus and motivation, we describe a very specific, genuine, example. We describe the construction of a DES model of fire service emergency cover (FSEC) provided by regional fire brigades in the UK. The work is ongoing and involves the active participation and financial support of a UK Government department, formerly the Office of the Deputy Prime Minister (ODPM), but recently reconstituted as the Department of Communities and Local Government (DCLG).

Regional Fire Brigades in the UK already possess a very sophisticated tool for gathering and analysing incident data in a very comprehensive way. This information is used for planning and to provide operational statistics to the UK Government. It is realised by brigade management that the data could be used to inform day to day management decisions.

The specific question of interest was whether it would be possible to develop a simulation model that would run sufficiently fast to be used to evaluate risk in real time. Such a model might then be deployed as an operational tool to provide real-time advice to brigade officers in responding to actual incidents.

The speed at which such a model can be run is a determining factor of its practically. In our case the model was able to simulate a year's operation in about three seconds.

2.2 The 'Cover-Moves' Problem

An operational problem of particular interest is the *Cover-Moves Problem*. This occurs when a fire brigade responds to a large incident (one that needs a large, say 8 or more, number of fire appliances to attend). The incident controller then usually repositions a small number of vehicles not involved in the large incident in what are called *cover-moves*, to try to minimize risk in the remainder of the region. Here risk can be clearly defined, either as the expected fatality rate in the region, or as an overall cost that takes into account both costed expected fatalities and brigade operating costs. In the rest of the paper, we shall take as our performance measure the expected fatality rate (as measured by the expected number of fatalities over a given period of time, under the conditions of the large incident).

The choice of a worthwhile cover-move combination (CMC) is an example of a problem in combinatorial optimization. It is not usually possible, certainly in real time, to identify the best solution. The real question is whether a worthwhile operational solution can be found.

We consider the kind of cover-moves solution achievable. In one example of a typical large incident, consideration was given to selecting 3 vehicles for cover-moves out of 16 available vehicles located in 11 stations. The 3 vehicles were to be sent to 3 out of the 6 stations that had supplied vehicles to attend the large incident. A simple combinatorial calculation shows that there are 25800 distinct CMCs possible. In fact, the majority of these could be ruled out on operational grounds so that only 230 CMCs turned out in this instance to need serious consideration.

The strict (policy driven) operating requirement for the cover-moves problem is that a solution has to be found within *one minute* of the notification of occurrence of a large incident. Suppose that we would be satisfied with any *one* of the a best alternatives (as measured by lowest expected fatality rate). We define

$$P_a = \Pr\{\text{At least one of the best } a \text{ alternatives is found out of } K, \text{ amongst } n \text{ randomly selected alternatives}\}. \quad (1)$$

We can easily calculate this, at least approximately, as follows. Suppose that there are K alternatives ($K = 230$, in the above example) and that we can examine n out of the K possibilities. Now, if a is small compared with n and K then the probability we do not find *any* of the a in the n is simply the probability they are all in the remaining $(K - n)$ alternatives. This is approximately

$$\left(\frac{K-n}{K}\right)^a.$$

Thus the probability, P_a, that we *do* find at least one of the a in the n examined is approximately:

$$P_a = 1 - \left(\frac{K-n}{K}\right)^a. \quad (2)$$

In our example, if simulations of a year's operation were used, then one could only make 20 runs in the one minute allowed. However if, say, each run simulated only two months' operation, then about 120 runs of the simulation model could be made. Table 1 shows the probability of finding one of the top a for a selection of a values for this latter situation. It will be seen that there is a 99.9% chance of finding one of the top ten CMCs.

Table 1: Value of P_a, as given by Equation (2), for selected a, when $n = 120$, $K = 230$.

a	1	2	3	5	10
P_a	0.522	0.771	0.891	0.975	0.999

The above analysis demonstrates the obvious benefit of being able to compare as many alternatives as possible, provided, as we have tacitly assumed, that the performance measure can be calculated exactly for each alternative so that comparisons can be made without error. However in practice there is a (time) cost attached to the calculation of of an accurate estimate of the performance measure. This has to be allowed for in deciding simulation run length. We discuss this in the next Section.

3 THE SIMULATION RUN-LENGTH PROBLEM

We discuss some related optimization/selection problems first, then consider the specific run-length problem considered in this paper.

3.1 General Search/Selection Problems

The above FSEC example is typical of a class of problems that falls between, on the one hand, the full optimization problems discussed for example by Andradottir (2006) or Olafsson (2006), and on the other hand, the ranking and selection problems, discussed for example by Pichitlamken and Nelson (2001), Boesel et al. (2003), and Kim and Nelsom (2006) or the optimal computer budget allocation (OCBA) problems discussed by Chen (2002). In these latter problems the number of possible choices is sufficiently small for all to be potentially examinable. In our problem we do not have the intractably large set of possible decisions of a full optimization problem. But the set of possible decisions is still too large to be exhaustively considered. However it is possible to examine a meaningful fraction of the set.

The natural approach is the OCBA one, but with the added need to decide on the fraction of the full set of possible decision choices that should be examined.

The process of selecting the decisions to be examined by simulation can be regarded as one of random sampling, involving the convolution of *two* independent random quantities: (i) the error arising from the variability in the simulation and more interestingly (ii) the random variation in the objective function arising from the random sampling. A possibility therefore is to estimate both distributions separately and so arrive at a criterion for optimal choice of run length and consequently the fraction of possible decision choices to be sampled. This is the approach adopted here.

Such an approach could be applied to the full optimization problem as well.

3.2 The Problem

We formulate the simulation run length problem as follows. Let K be the number of alternative decisions under consideration. Let the true performance measure of the system operating under the ith alternative be X_i. We consider the situation where we select alternatives at random. The performance X of a randomly selected alternative is therefore a random variable taking values in the set $\{X_i\}$. We denote the cumulative distribution function of X by $F_X(x)$. Suppose that the total time available for carrying out simulation runs is T and that $n < K$ simulation runs are to be made each of length $t = T/n$.

For the ith simulation run let the *observed* performance measure be

$$Y_i = \mu + X_i + \eta + \varepsilon_i,\ i = 1,2,...,n\,. \tag{3}$$

All the terms on the right: $\eta,\ X_i,\ \varepsilon_i,\ i = 1,2,...,n$ are assumed to be mutually independent random variables.

We assume that the performance measure, X_i, is a rate quantity (averaged over time) so that it is independent of the run length t. Moreover its variability stems purely from the random process of selecting alternatives to investigate, and *not* from simulation variability.

The term μ is a common mean that it enables us to take

$$E(X_i) = 0 \tag{4}$$

and simplify derivation and presentation of numerical results slightly. We write the variance of X_i as

$$Var(X_i) = \sigma^2,\quad i = 1,2,...,n\,. \tag{5}$$

The quantity ε_i is an 'error' term with mean zero and variance

$$Var(\varepsilon_i) = \tau^2/t = n\tau^2/T,\quad i = 1,2,...,n\,. \tag{6}$$

The term η is also a random variable with mean zero and variance

$$Var[\eta(t)] = \theta^2/t = n\theta^2/T\,, \tag{7}$$

and is included to allow the case where common random numbers are used in the n simulation runs. Thus, we assume that the total error is $\eta+\varepsilon_i$ with η the error arising from the use of common numbers that is removeable when the difference between observations is taken.

If the X_i were known then our best choice would be the alternative i_0 corresponding to the smallest X_i. However we do not know the true X_i but instead have the observed performance measures, Y_i, which we can order as:

$$Y_{(1)} < Y_{(2)} < ... < Y_{(n)}. \tag{8}$$

We therefore take the alternative , $\hat{i}_0$, corresponding to the smallest, $Y_{(1)}$, in the expectation that when the errors ε_i are small compared to the X_i , then this will be a good choice. We can study this by calculating the probability, P_a, of (1), only now where we have selected $\hat{i}_0$. In what follows, for simplicity we write $X_{(1)}$ (rather than $X_{\hat{i}_0}$) for the performance measure corresponding to $Y_{(1)}$. Also it is important in what follows, to note the dependence of $X_{(1)}$ on n, which, again for simplicity, we do not display explicitly.

To simplify the discussion, we shall assume that both X and Y are continuously distributed. Though Y will typically be continuous, X may not be. In the FSEC example, X definitely is not, but provided K, the number of alternatives is reasonable large, (say more than 100 or 200) the assumption that it is continuous should not make a great quantitative difference. We write $F_X(\cdot)$, $F_Y(\cdot)$ and $F_\varepsilon(\cdot)$ for the cumulative distribution functions (cdf) of X, Y and ε; and write $f_X(\cdot)$, $f_Y(\cdot)$ and $f_\varepsilon(\cdot)$ for their probability density functions (pdf).

Each Y can be regarded as the sum of X and ε where each has been sampled independently. Thus $F_Y(\cdot) = F_{X+\varepsilon}(\cdot)$ can be obtained by convolution.

We now calculate $F_{X_{(1)}}(\cdot)$. Suppose that $X_{(1)} = X_i$. An elementary conditional argument shows that

$$\Pr(X_{(1)} = X_i \text{ and } X_i > x) = \int_{-\infty}^{\infty} \{\int_{-\infty}^{y-x} f_X(y-u) f_\varepsilon(u)[1-F_Y(y)]^{n-1} du\} dy \,. \tag{9}$$

Hence, as any of the i are equally likely to be selected, we have

$$\Pr(X_{(1)} < x) = F_{X_{(1)}}(x) = 1 - n(\int_{-\infty}^{\infty} \{\int_{-\infty}^{y-x} f_X(y-u) f_\varepsilon(u)[1-F_Y(y)]^{n-1} du\} dy \,, \tag{10}$$

and

$$f_{X_{(1)}}(x) = n(\int_{-\infty}^{\infty} f_X(x) f_\varepsilon(y-x)[1-F_Y(y)]^{n-1} dy \,. \tag{11}$$

It is clear from the definition of $X_{(1)}$ that it does not depend on η or its distribution and this is borne out by (10) and (11).

It should be noted that n enters these formulas not only as a multiplicative factor but also through the distributional dependence of ε on n, which has been suppressed for notational simplicity.

Equation (10) allows us to find the probability, P_a, that the selected alternative, $\hat{i}_0$ (corresponding to $Y_{(1)}$), is actually one of the a alternatives with the smallest performance measure. From (10), we have

$$P_a = F_{X_{(1)}}(x_{a/K}) \tag{12}$$

where $x_{a/K}$ is the quantile

$$x_{a/K} = F_X^{-1}(a/K) \quad (= X_{(a)}) \tag{13}$$

of the distribution of the performance measure X, when this is treated as a random variable.

The best choice of n is simply that value which maximizes P_a, i.e.

$$n_{\text{opt}} = \arg\max\{F_{X_{(1)}}(x_{a/K})\}. \tag{14}$$

An analytical expression for this value seems difficult to obtain because of the complicated way in which n enters into the expression for $F_{X_{(1)}}(x_{a/K})$. A numerical calculation is possible. However before we can do this we need to determine $F_X(\cdot)$ and $F_\varepsilon(\cdot)$, as $F_{X_{(1)}}(x_{a/K})$ depends explicitly on these. We consider this next.

4 DETERMINATION OF THE DISTRIBUTIONS OF X AND ε

The choice of an appropriate form for the distribution of X is an interesting problem in its own right. Assuming that we are interested in performance measure minimization, then we would be particularly interested assuming the correct form for left tail behaviour. It is possible to give examples of problems where a normal distribution would seem appropriate, and we shall only consider this case here.

The assumption of a normal distribution for that of the error components η and ε in equation (3) seems less contentious, and this will be assumed here.

Figure 1 gives the distribution for the observations (i.e. the Y_i) for the fire service cover-moves problem of Section 2.

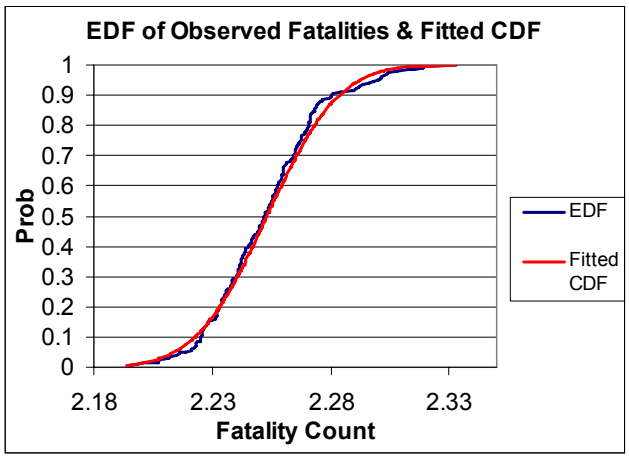

Figure 1: Distribution of fatalities count for 230 cover-move choices.

We shall assume that we can make a set of simulation runs off-line in order to estimate the parameters of the normal distributions of X and ε. Let a set of n runs as given in (3) be called a *trial* and consider a set of m such trials with observations

$$y_{ij} = \mu + x_i + \eta_j + \varepsilon_{ij},\ i = 1,2,...,n,\ j = 1,2,...,m\,. \quad (15)$$

We can estimate the x_i and η_j as fixed effects by minimizing the sum of squares

$$S = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \mu - x_i - \eta_j)^2 \quad (16)$$

subject to

$$\sum_{i=1}^{n} x_i = \sum_{j=1}^{m} \eta_j = 0\,. \quad (17)$$

This gives the estimates

$$\hat{\mu} = (nm)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\,, \quad (18)$$

$$\hat{x}_i = m^{-1}\sum_{j=1}^{m} y_{ij} - \hat{\mu},\ i = 1,2,...,n\,, \quad (19)$$

$$\hat{\eta}_j = n^{-1}\sum_{i=1}^{n} y_{ij} - \hat{\mu},\ j = 1,2,...,m\,. \quad (20)$$

The corresponding estimate for σ^2 is then:

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\hat{x}_i^2\,. \quad (21)$$

From equation (6), the variance of the ε is estimated by

$$n\hat{\tau}^2 = (nm)^{-1}\sum_{j=1}^{m}(y_{ij} - \hat{\mu} - \hat{x}_i - \hat{\eta}_j)^2 \quad (22)$$

where, with no loss of generality, we have taken $T = 1$.

Figure 2 shows the results of $m = 5$ trials comprising $n = 10$ runs each where we have set $n = 120$, the value discussed in Section 2. The estimates obtained from these observations are

$$\hat{\sigma} = 0.0322 \text{ and } \hat{\tau} = 0.000563\,.$$

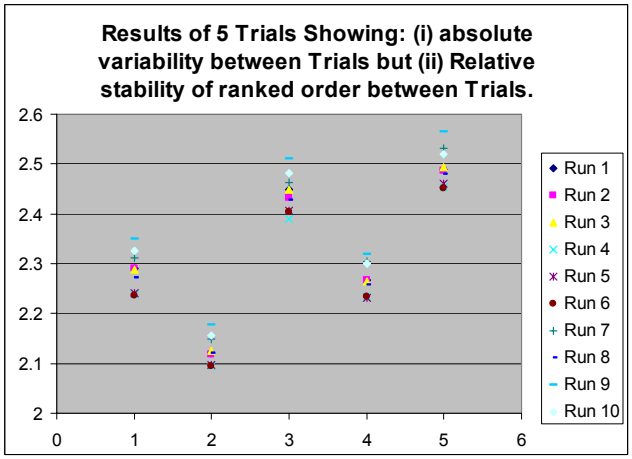

Figure 2: Fatality counts of 5 trials with 10 runs in each.

5 THE DISTRIBUTION OF $X_{(1)}$

Once the distributions of X and ε have been estimated, the distribution of $X_{(1)}$ can be obtained from equations (9) and (10). We have carried out this numerically using simple quadrature by Simpson's Rule with appropriately selected ranges of integration over which the integrand is not negligible. Figure 3 shows the behaviour of the densities for selected n. It will be seen that as n increase, the general location of the densitites initially moves to the left, indicating that lower (i.e. better $X_{(1)}$ values) are obtained. However, as n continues to increase, the direction changes and the densities start to move right. There is clearly an optimum n as far as general location is concerned.

We can make this more precise by considering how the probability P_a of equation (1) changes. We can calculate this from equation (12). Table 2 gives the value of P_a for selected n for the cover-moves example with $a = 10$ (and $K = 230$). In this case we see that our setting of $n = 120$ is actually not optimal and we could have achieved a higher probability with $n = 200$. This shows that in this exmaple, we could actually have made runs sufficiently short to have examined all the 230 alternatives. Note that even when we examine all alternatives, we cannot guarantee to select ono of the best a because we cannot completely eliminate all simulation experimental error.

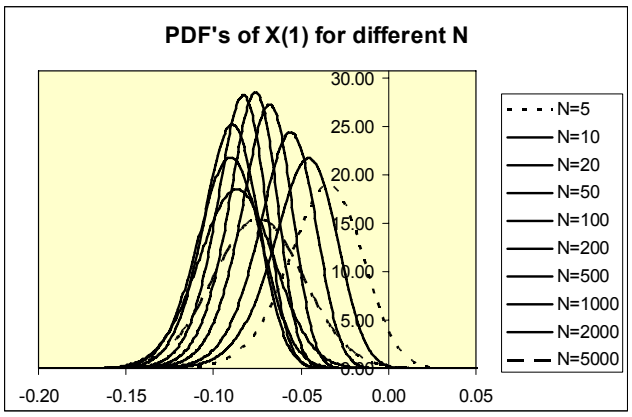

Figure 3: Probability density functions of $X_{(1)}$ for selected n.

Table 2: Value of P_a, calculated from Equation (12), for selected n, when $a = 10$, $K = 230$.

n	100	200	500	1000	2000
P_a	0.969	0.991	0.991	0.978	0.932

6. CONCLUSIONS

We have given a preliminary analysis of how simulation runs should be set up when there is a limit to the time available for simulation runs. This problem can arise when simulation is used in real time decision making. For expositional clarity, the statistical model used in the analysis and discussed in Section 3 has been kept simple. However it does allow for the use of common random numbers in making the simulations.

We have shown how the model might be used in practice in Sections 4 and 5, and illustrated it with an application to a real problem arising in the provision of fire service emergency cover.

An issue of especial interest which we have not discussed, is the form that the distribution of the performance measure might take when it is randomly sampled. We have only considered the normal model for this, but there are theoretical reasons why power law distributions may be more appropriate in certain situations. It is hoped to discuss this issue elsewhere.

REFERENCES

Andradottir, S. 2006. An Overview of Simulation Optimization via Random Search. In Handbooks in Operations research and Management Science, SG Henderson and BL Nelson, eds. Amsterdam: North-Holland

Boesel, J. 2003. Simulating Aircraft Delay Absorption. Proceedings of the 2003 Winter Simulation Conference. S. Chick, P.J. Sanchez, D. Ferris, and D.J. Morrice, eds, IEEE Piscataway. 1663 – 1669.

Boesel, J., Nelson, B.L., and Kim, S-H. 2003. Using ranking and selection to "clean up" after simulation optimization. *Operations Research,* 51, 814-825.

Chen, C.-H. 2002. Very Efficient Simulation for Engineering Design Problems with Uncertainty. In Modeling and Simulation-Based Life Cycle Engineering, K. Chong, S. Saigal and S Thynell, eds) pp 291-302, London: Spon Press.

EUROCONTROL-CFMU 2002. Flow and Capacity Management Unit. <http://www.cfmu.eurocontrol.int/cfmu/opsd/public/standard_page/operational_services_atfcm_management.html>

Law, A.M. and Kelton, W.D. 1991. *Simulation Modeling and Analysis 2nd Ed.* New York: McGraw-Hill.

Olafsson, S. 2006. Chap 21. Metaheuristics. In Simulation, Handbooks in Operations research and Management Science, SG Henderson and BL Nelson, eds. Amsterdam: North-Holland

Kim, S.-H. and Nelson, B.L. 2006. Selecting the Best System. Chap 17. In Simulation, Handbooks in Operations research and Management Science, SG Henderson and BL Nelson, eds. Amsterdam: North-Holland

Wrigley, D. and Taylor, B. 2003. SIMBRIG – Simple Brigade Model, Meeting the Need for Fast operational analysis support at the formation level. <http://www.dcmt.cranfield.ac.uk/ismor/ISMOR/2003/wrigley.ppt>

AUTHOR BIOGRAPHY

RUSSELL C. H. CHENG is Professor of the Operational Research Group at the University of Southampton. He has an M.A. and the Diploma in Mathematical Statistics from Cambridge University, England. He obtained his Ph.D. from Bath University. He is a former Chairman of the U.K. Simulation Society, a Fellow of the Royal Statistical Society and the British Computer Society, Member of the Operational Research Society. His research interests include: variance reduction methods and parametric estimation methods. He was a Joint Editor of the *IMA Journal of Management Mathematics*. His email and web addresses are <R.C.H.Cheng@maths.soton.ac.uk> and <www.maths.soton.ac.uk/staff/Cheng>.