LABOR
LABORATORIO RICCARDO REVELLI
CENTRE FOR EMPLOYMENT STUDIES

Working Paper no. **142**

We ran one billion agents.
Scaling in simulation models.

Ross Richardson
Institute for New Economic Thinking at the Oxford Martin School;
Mathematical Institute, University of Oxford, UK

Matteo Richiardi
Institute for New Economic Thinking and Nuffield College, Oxford, UK
Collegio Carlo Alberto, Moncalieri, Italy

Michael Wolfson
University of Ottawa, Canada

June, 2015

We ran one billion agents.

Scaling in simulation models.

Ross Richardson

Institute for New Economic Thinking at the Oxford Martin School;
Mathematical Institute, University of Oxford, UK

Matteo Richiardi

Institute for New Economic Thinking and Nuffield College, Oxford, UK;
Collegio Carlo Alberto, Moncalieri, Italy

Michael Wolfson

University of Ottawa, Canada

Abstract

We provide a clarification of scaling issues in simulation models, distinguishing between sample size determination, discovery of emergent properties involving a qualitative change in the behaviour of the system at an aggregate level, and 'true' scaling, the dependence of the quantitative behaviour of the system at any given level of aggregation, to its size. Scaling issues arise because we want to understand what happens when we run one billion agents, without actually having to run one billion agents. We discuss how we can use the Buckingham Pi theorem, a key tool in dimensional analysis, to provide guidance on the nature and structure of scaling relationships in agent-based models.

Acknowledgements

We benefited from insightful discussions with Dan Tang. Errors and/or omissions remain our sole responsibility.

1. Introduction

The ability to scale up is increasingly considered an essential feature of agent-based (AB) models. Lisenko and D'Souza (2008), for instance, simulate Epstein and Axtell's Sugarscape model (Epstein and Axtell, 1996) with one million agents, while Robert Axtell himself, in a TED talk, explicitly aims at modelling the economy with 150 million agents, anticipating a one-to-one map with the US economy (Axtell, 2011). This would in principle allow all the projected individual agents' trajectories to differ, avoiding the distortions involved in having multiple copies of the same agent, as entailed by drawing the simulated population from a representative sample of the real population, with weights attached to each individual in the sample. At the same time, improvements in computer hardware and software architectures, in particular the advent of parallel computing[1], make such goals increasingly within reach, while the excitement brought about by the 'big data revolution'[2] further raises expectations by making available the information necessary to calibrate such models. In addition, many economic variables stretch over orders magnitudes, with a small (potentially very small) number of entries having a disproportional importance in the economy. Examples are power law distributions of income and wealth, the size of cities and firms, stock market returns, trading volume, international trade, executive pay, etc. (see Gabaix, 2009 for a review of the literature). Power laws are relations of the type $Y = kX^{\alpha}$ and are often represented as linear relationships on a log-log scale ($\log Y = c + \alpha \log X$) which remain stable over several orders of magnitude (hence they are referred to as scale-invariant): zooming in or zooming out does not alter the picture.[3] Modelling fat tails requires that rare categories are covered. This in turn requires either to scale up the model in order to allow for such rare categories, or to purportedly distort the underlying distributions to permit a small number of extreme cases in small populations.[4]

Indeed, there are a number of motivations for running AB models on scales much smaller than a one-to-one map of the real world, and they are largely associated with the computational demands entailed with executing such full scale models. At best, the time required to run an AB model scales linearly with the number of agents involved. This can be seen by just considering the extra iterations across the group of agents that is necessary to cover the whole population. However, whenever interactions between agents are involved, for example, when each agent has to communicate with a subset of other agents, or the global information of the model is calculated from agent interactions e.g. during price formation in a market, combinatorial explosion can lead to super-exponential scaling of the execution time. Moreover, computational constraints such as the need to distribute such large amounts of information across computer cores and memory nodes, and the necessary communications across these cores and the memory hierarchy, can further reduce the simulation speed to a

[1] See for instance Lysenko and D'Souza, 2008; Caron-Lormier et al., 2008; Coakley et al., 2012; Holcombe et al., 2013.

[2] The Economist (2010).

[3] By contrast, Gaussian densities rapidly decay away from the mean. The difference is highlighted by the expression 'fat tail', which applies to power law distributions: rare events are less rare, and extreme events are still possible.

[4] A crucial question in dealing with a model that is suspected to display power law behaviour is how many orders of magnitude need to be analysed to test if α remains constant: a typical case of sample size determination (see section 2.1). "As a rule of thumb, the scaling region should persist for at least three orders of magnitude on both axes for a reliable estimate of the critical exponent [α]" (Christensen and Moloney, 2005).

point where the time required to run full scale models becomes prohibitive. One may also need to consider the financial costs involved of building and maintaining such extensive computer hardware required to facilitate the simulation of full scale models.

If a model has to be used by policy makers (e.g. at the Federal Reserve) or regulators (e.g. the United States Securities and Exchange Commission), a necessary requirement is that it can provide timely insight into the problem at hand and guidance on what to do. Models that take ‘too long’ to run and produce data that is ‘too large’ to analyse in the required time, are of limited interest for such users. But speed of execution is important also for academic research. *Understanding of model behaviour* often requires sophisticated and computationally expensive tools of sensitivity analysis (Saltelli et al., 2000, 2008). *Robustness analysis* – checking whether relaxation or replacement of some of the assumptions lead to dramatic changes in model outcome– becomes unfeasible if a model takes too long to run. *Estimation* possibly involves millions of runs, at different values of the parameters (Grazzini and Richiardi, 2015; Tsionas et al., 2015).

The above discussion suggests the importance of creating reduced scale models that can run at a fraction of the time of full scale models, much like engineers, architects and city planners build geometrically reduced scale models to test before building the real thing. Indeed, guidance from these disciplines can be brought to bear in producing reduced scale AB models in a rigorous and consistent manner. In particular, we show how we can employ the Buckingham Pi theorem, a key tool in dimensional analysis, to rewrite the model in terms of dimensionless parameters. This implies a relationship between the dimensional variables that, if respected, ensures the model behaviour remains the same no matter the scale (they are deemed *similar* in the engineering domain). As we shall see however, application of the Buckingham Pi theorem is not always possible; moreover, additional scale effects might arise due to the discretisation induced by small N. Whether these discretisation effects are of practical relevance depends on the characteristics of the system (such as how small are the values of discrete variables involved in the scaling relationship).

As a preliminary step however, we wish to clarify some confusion about what scaling really is, and why it matters for modelling purposes. We distinguish scaling from emergence and sample size determination. *Sample size determination* points to increasing population size up to the point where some desired level of statistical significance is guaranteed, when analysing the model results. *Emergence* involves a qualitative change in the behaviour of the system that is a characteristic of the macro level and cannot be inferred by just looking at the constituent parts: an organism is different from the sum of its cells, or, as the Nobel prize-winning physicist Philip Anderson wrote back in 1972, *more is different* (Anderson, 1972). *Scaling* refers to a quantitative change in the behaviour of the system, as measured by some statistics y, which occurs at any given level of aggregation as the size of the system S (including e.g. population size) changes: $y = y(S)$.

The remainder of the paper is organised as follows: section 2 elaborates on the distinction between scaling, emergence, and sample size determination; section 3 shows how to handle scaling issues using the Buckingham Pi theorem; section 4 offers an example of how to control for scaling effects in a simple AB model of job search; section 5 comes back to a queuing problem which displays scaling effects, that we introduce in section

2, and explains why the Buckingham Pi theorem cannot be applied in that setting; section 6 offers our concluding remarks.

2. Needs for increasing population size

2.1 *Rare events.*

A first case when large-scale simulations are needed is when we are concerned with the stochastic variability of the simulation outcomes, particularly when rare events are concerned. Consider for instance the probability *p* of the occurrence of some event (e.g. a radical innovation) or of some individual trait (e.g. exceptional charisma), or a mean *x* (e.g. the mean income in the top 0.01% of the income distribution).[5] Our question, in its basic form, is then how to select the population size so that the confidence interval of the statistics of interest at any given level does not exceed some chosen margin of error. Suppose for instance that we are interested in 95% confidence intervals. As predicted by the Central Limit Theorem, both $\hat{p}$ and $\hat{x}$ are approximately normally distributed in sufficiently large populations.[6] We can thus construct a confidence interval for $\hat{x}$ using $ME = z*SE$, where *ME* is the margin of error, *z* is the *z*-score[7] and *SE* is the standard error of the estimator. In the case of a proportion or probability, we have $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$, and therefore $ME = z\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$. Solving for *N* gives

$$N^* = \left(\frac{z}{ME}\right)^2 \hat{p}(1-\hat{p}). \tag{1}$$

Without prior knowledge of $\hat{p}$, a conservative assumption is to calculate sample size for $\hat{p}$ = 0.5. For instance, if ME is set to 0.005 (a confidence interval of 1 percentage point) and we require a 99% confidence interval, we get N^* = 66,349. Supposing we can safely assume that $\hat{p}$ is smaller than 5%, we can reduce our sample size to N^* = 12,607. Of course, if we are interested in a rare occurrence we should also lower the accepted margin of error which, *ceteris paribus,* calls for an increase in the sample size.[8]

In case of a population mean, we simply plug in the standard error of the mean, $SE = \frac{s}{\sqrt{N}}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$, where *s* is the standard deviation of the underlying variable of interest (e.g. income), from which we obtain[9]

[5] Note that proportions are just particular cases of means, where the underlying variable is an indicator (a variable that can only take a value of 0 or 1).

[6] Here we assume that the underlying distributions do not change with population size, i.e. there are no "real" scaling problems (as defined in section 2.3).

[7] E.g. 1.645 for a 90% confidence interval, 1.96 for a 95% confidence interval, 2.58 for a 99% confidence interval.

[8] One could for instance specify the ME as a fraction of $\hat{p}$: ME = $\alpha\hat{p}$. For instance, if we are interested in a property that is supposed to characterize one person in 10.000 ("the top 0.01%"), and we are willing to tolerate a 95% confidence interval equal to 1/10 of the predicted frequency (α = .05), we need a sample size of over 15 million individuals.

[9] With very small populations (i.e. *N* < 30), the normal approximation fails and the Student's T distribution should be used, with the corresponding *t*-score. This slightly complicates the computation as sample size affects the *t*-score as well as *N*, via the degrees of freedom in the T distribution.

$$N^* = \left(\frac{z}{ME}\right)^2 s. \qquad (2)$$

Here we can't avoid guessing a value for the standard deviation s, as there is no theoretical maximum value, but the guess can then be validated in the artificial data.

This is how we should dimension sample size such that, at any point in time (in the stationary state), our statistics of interest are measured within given bounds from their theoretical level, whatever this is. Note that it might well be the case that the 'true' values of the statistics of interest are dependent on sample size itself (a case of scaling effects, as we will discuss in section 2.3): as long as we are able to put a reasonable upper value on the standard deviation, we will measure it with enough precision. Note also that the 'true' values could display autocorrelation: for instance, the proportion of the super-rich might be dependent on the business cycle. Hence, if we want to describe the behaviour of the system, a single measurement at a given point in time is likely to prove grossly inadequate, and (longitudinal) time averages should then be used (Grazzini and Richiardi, 2015). By extending the length of the observation window and averaging over a longer period of time, it is possible to improve on the level of precision attained. In principle, given enough simulated data, the statistics of interest can be estimated with any level of accuracy, for any sample size, even if displaying autocorrelation, as long as the statistics are stationary. However, as we have noted above, exploiting this property to excessively reduce sample size increases exposure to scaling effects. Similarly, if the model is ergodic (roughly speaking, this means that all simulated runs are alike[10]) we can average over multiple replications to improve on accuracy and/or reduce sample size (but see the caveats before). Finally, averaging over long periods of time or high numbers of replications is the only way to obtain measurements if systemic events are concerned: these are events that characterise the whole system (like a sudden collapse of economic activity, or an outbreak of an epidemic), and thus preclude the exploitation of individual variability at any given point in time for estimating the underlying probability distribution. The formulas above can then be used to have an initial guess of how many simulation periods or how many simulation runs will be necessary to estimate such systemic events at the desired level of accuracy.

2.2 Emergent behaviour

A second case when large-scale simulations are needed to fully understand the properties of a system is when repeated interactions between the agents qualitatively affect their behaviour, to a point where some new regularity emerges at an aggregate level. Emergent behaviour resides in the properties of the ensemble rather than of any individual state, and it arises when the environment interacts with the system to select the allowable states (Bar-Yam, 2004). We referred to the difference between cells and an organism previously, and another example is pedestrian movements. At small densities, pedestrian movements are determined by individual

[10] See Grazzini and Richiardi (2015) for a more in-depth discussion of ergodicity and non-ergodicity issues in AB modeling.

preferences and aims, and they are largely unpredictable. At higher densities however, pedestrian movements become constrained by repulsive interactions, leading to the predictable dynamics of separate lanes of uniform walking direction in crowds of oppositely moving pedestrians, or oscillations of the passing direction at bottlenecks (Helbing et al., 2001). These dynamics are self-organised (there is no top-down coordination) and characterise the behaviour of the system only at an aggregate level: the crowd behaves as if it was a separate entity, with its own laws of motion which bear little resemblance to those of the individuals composing it. The same is true of the flocking of birds (which are not led in any way, even though they may appear to be), ants foraging for food (each follows a set of simple rules, but the colony as a whole acts in a sophisticated way), the growth of tumours (which elaborate sophisticated communication and decision-making), traffic jams on a motorway (even though all of the cars are moving forward, the traffic jam tends to move backwards), and the formation and scattering dynamics of insurgent militia groups fighting guerrilla wars (Bohorquez et al., 2009).

Also, when there are multiple equilibria, achieving the necessary coordination to switch between equilibria might become increasingly difficult as sample size increases, causing an ergodic system to behave as if it was *de facto* non-ergodic (Grazzini and Richiardi, 2015). A case when this might happen is when social norms are involved: once a norm is established (e.g. wearing a tie in the workplace) it might be very difficult to change it, especially if it is shared by many individuals (e.g. many work interactions). At the micro level, each individual can be in one of many states, a property that is lost with aggregation. Another example is the behaviour of attendees at a public performance: with only a few people, everybody comfortably sits down, and those who stand up are kindly requested to return to their previous position and not to obstruct the view of others; as the number of participants increases however, the likelihood of this happening decreases, and in the end everybody has to stand up.

In all these examples, emergence of a collective behaviour depends on *density*. The probability of observing an emergent phenomenon increases with density, generally following an S-shaped curve, with an extreme case being a step function around a deterministic density threshold. Often, this deterministic threshold is obtained asymptotically when the size of the system is infinite. For instance, a classic problem in percolation theory considers a regular lattice where each cell is occupied with probability p. Percolation is obtained whenever a cluster –a group of nearest neighbouring occupied cells– extends across opposite sides of the lattice, a property that can be assessed only at the macro level. In a finite lattice, for any occupancy probability p, there are some random configurations where the system percolates, and some others where the system does not percolate, hence there exists only a probability of percolation q: an S-shaped curve describes how the percolation probability q increases with the occupation probability p. The critical occupation probability (or density) p_c can then be defined as a fraction τ of the possible configurations that percolate, where the remaining fraction 1- τ do not percolate (a standard value for τ is .5). As the size of the lattice increases, the S-shaped curve becomes steeper, converging to a step function for lattices of infinite size: above a critical occupation probability the probability of percolation is 1; below the critical density, the probability is 0 (e.g. the best recent estimate is $p_c = 0.59274621$ for two-dimensional infinite square lattices).

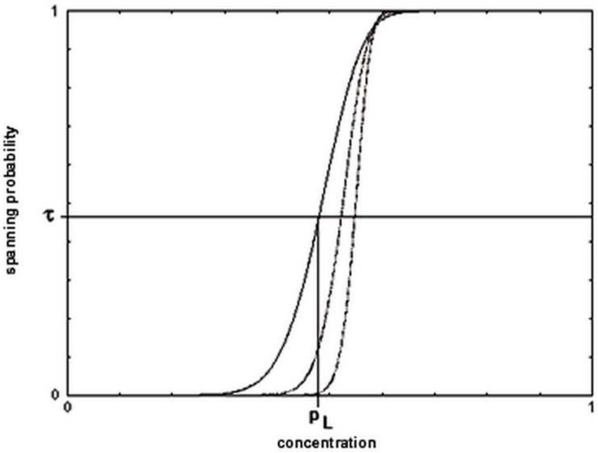

Figure 1: Percolation (spanning) probability against occupation probability (concentration), on finite square lattices of different sizes L. The lattice is viewed as a torus, so that percolation is to be understood as wrapping along either of the lattice dimensions. By fixing a value τ, one can find a sequence of values $p_L(\tau)$ for increasing lattice sizes L, approaching the critical threshold p_c. Source: de Oliveira et al. (2003).

This discussion exemplifies the difference between emergence and scaling: percolation is an emergent property of a system, which changes its qualitative behaviour. When a system is at the critical threshold, many quantities – the cluster number density, the average cluster size, etc. – are insensitive to the underlying lattice details, and depend only on the dimension of the system (1, 2, …, n dimensions, corresponding to a line, a plane or a volume etc.). As such, they are characterised by *universal critical exponents*.[11] On the other hand, the critical occupation probability depends on the size of the lattice (and on other lattice details), and is therefore scale-sensitive. This finally leads us to a discussion of scaling.

2.3 *Scaling issues*

A third case when one might consider running 'large' simulations occurs when the behaviour of the model quantitatively depends on its scale. In order to define scale, let (d_1, d_2, …, d_r) denote a fundamental system of units, such as money, time, number of individuals and number of firms, while (x_1, x_2, …, x_n) are quantities that are measurable with respect to this system, including parameters, dependent variables, and independent variables. The dimension of each x_i, denoted as $[x_i]$, is then given by

$$[x_i] = d_1^{\alpha_{1,i}} d_2^{\alpha_{2,i}} \dots\ d_r^{\alpha_{r,i}} \qquad (3)$$

for suitable exponents ($\alpha_{1,\mathrm{i}}$, $\alpha_{2,\mathrm{i}}$, …, $\alpha_{r,i}$). For instance, the hiring rate is defined as persons hired per period, and thus has dimension NT^{-1}, where N measures individuals and T measures time. A model is said to be invariant under a change in units $d_j \rightarrow \lambda_j d_j$, j = 1, …, r, if

[11] See for instance Christensen and Moloney (2005).

$$x_i \rightarrow \lambda_1^{\alpha_{1,i}}\lambda_2^{\alpha_{2,i}} \ldots \lambda_r^{\alpha_{r,i}}x_i, \quad i = 1, \ldots, n \tag{4}$$

for any (λ_1, λ_2, …, λ_r) > 0. This implies that any relationship of the type $x_1 = f(x_2, \ldots, x_n)$ between measurable quantities in the model satisfies the following *scaling property* (Zohuri, 2015):

$$\lambda_1^{\alpha_{1,1}}\lambda_2^{\alpha_{2,1}} \ldots \lambda_r^{\alpha_{r,1}} f(x_2, \ldots, x_n) = f(\lambda_1^{\alpha_{1,2}}\lambda_2^{\alpha_{2,2}} \ldots \lambda_r^{\alpha_{r,2}}x_2, \ldots, \lambda_1^{\alpha_{1,n}}\lambda_2^{\alpha_{2,n}} \ldots \lambda_r^{\alpha_{r,n}}x_n) \tag{5}$$

For instance, a model of bank failures is scale invariant if the fraction of banks that become bankrupt remains constant when we multiply the number of banks in the system by a factor λ (i.e. if we change the units in which we count banks to $1/\lambda$ units).

However, there are cases where a simple normalisation does not remove scale effects. An example is city dynamics, where many urban properties Y are described by scaling relations of the form $Y = cN^{\beta}$, where c and β are constants. Superlinear scaling ($\beta > 1$) is common with 'social' quantities (such as wages, or inventions), while urban infrastructures are generally subject to sublinear scaling ($\beta < 1$) (Bettencourt, 2013).

Another example is the calculations of service-providing elements, as described by the Erlang loss model, a classical result in queuing theory. The model is commonly used by telephone system designers to estimate the number of lines, telephone circuits, telephone switching equipment or call centre staff (more generally: capacity) required to meet given quality standards, but can also describe the number of copies of a book a library needs to own in order to keep unmet requests under control, for instance.[12] The model assumes that there are N homogenous servers working in parallel and no extra waiting space; customers that find all N servers busy upon arrival are blocked (lost). Under the further assumptions that customers arrive according to a Poisson process with rate ν and that service times are independent and exponentially distributed with mean $1/\mu$, the steady-state probability p that a customer is blocked is given by the Erlang B formula[13]

$$p = \frac{\frac{\lambda^N}{N!}}{\sum_{i=0}^{N} \frac{\lambda^i}{i!}} \tag{6}$$

where N is capacity, i.e. the number of identical parallel resources such as servers, telephone lines, book copies, etc. and $\lambda = \nu/\mu$ is the offered load. Note that λ is a dimensionless unit and it is equal to the mean arrival rate multiplied by the mean holding time.[14] Figure 2 depicts the relationship between the offered load (λ) and capacity (N), for three different levels of the blocking probability p.[15]

[12] We thank Dan Tang for having shared this example with us.

[13] The case when queuing is allowed is described using the Erlang C formula.

[14] In telephony the load unit is referred to as an *erlang*.

[15] See van Leeuwaarden and Temme (2009).

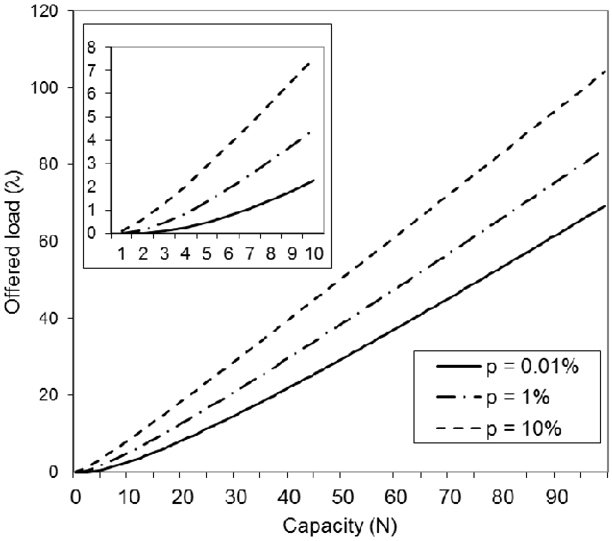

Figure 2. The Erlang B relationship between offered load λ and capacity N (number of servers), for different values of the blocking probability p. The smaller diagram zooms in for smaller values of N.

The relationship is nonlinear: a higher load per server can be carried, ceteris paribus, when there are more servers. Said differently, there are *increasing returns to scale* at a system level. The relationship between λ and N tends asymptotically to linearity. However, the normalised statistics ‘offered load per server’ is still scale-sensitive for a large interval of N (figure 3).

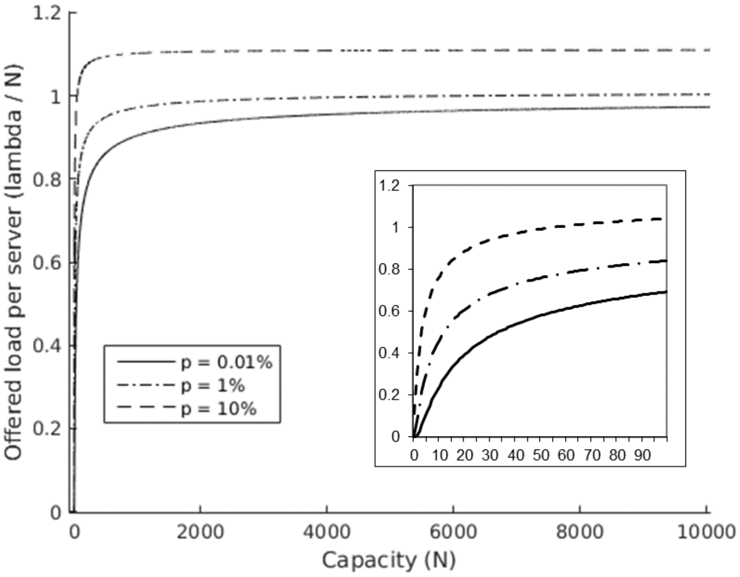

Figure 3. The Erlang B relationship between offered load per server λ/N and capacity N (number of servers), for different values of the blocking probability p. The smaller diagram zooms in for smaller values of N.

Whether scaling is an issue in practical applications where the system can in principle exhibit increasing or decreasing returns to scale, but tends asymptotically to constant returns to scale, depends therefore on the size of the real system. To continue with the queuing example, suppose the real world capacity is measured in tens or hundreds of thousands of servers. Then, a simulation which considers only 1,000 servers would provide a good approximation, in per server terms, if the blocking probability is not too low. However, if the real capacity of the system is measured in hundreds of servers, then further scaling down introduces a significant bias, which has to be taken into consideration.

3. Scaling in a rigorous and consistent way: the Buckingham Pi Theorem

3.1 Similitude and the importance of scaling AB models properly

Engineers, architects and city planners often create reduced-scale models on which to test and perform experiments before building the real thing. Indeed, the topic is known as *similitude*, and the three main criteria under which the similarity of reduced-scale physical models to full-scale models is assessed are: *geometric similarity* (the lengths of components in the model have all been scaled by the same factor in all three spatial dimensions), *kinematic similarity* (the rates of change of components are all scaled by the same factor), and *dynamic similarity* (the ratios of forces acting on components, surfaces and fluids are constant).

Building such reduced-scale models minimise the costs and time involved in constructing well-functioning full-scale systems. Insights gained by performing experiments on reduced-scale models are often invaluable – think of the importance of building structurally safe bridges and buildings, of constructing aeroplanes that are aerodynamically stable, and of designing cities with smoothly flowing traffic. Analogously, it might be desirable to keep an AB model to a small scale: we then need to understand how to consistently scale the model, in order to achieve the benefits of a large scale simulation without having to bear its cost.

A danger that exists in AB modelling is the temptation, when wanting to change the scale of the model, to naively vary only the size of the population of agents and leave all other quantities and parameters in the model the same[16]. This is equivalent to building a reduced-scale model of an aeroplane by scaling down the length of the plane's body whilst keeping the width of the wings at full scale. Clearly any experiments in an aerodynamic wind tunnel on such a scale model will not provide the necessary insight into the maximum load that the wings can support in a properly scaled aeroplane, and applying any information obtained in designing a full size plane would be downright dangerous! In the same regard, if policy makers and regulators applied insights gained from an AB model whose parameters have not been consistently scaled, it could have similar

[16] The variation in the population size of agents is further complicated if the model contains more than one class of agent or aggregate grouping of agents. For example, when scaling a model of health inequality such as in Wolfson et al. (2016), if one varies the number of people, how should one scale the number of neighbourhoods that the people live in?

catastrophic consequences. Returning to the finite lattice percolation model discussed earlier, if we consider the occupation probability as analogous to the number of agents in an AB model, increasing the number of occupied lattice points raises the likelihood that the system is in a percolating phase. However, if we want to build a larger scale model of the system that maintains the same probability of being in a percolating phase, it is necessary to scale other aspects of the model, such as increasing the size of the lattice (the number of lattice points).

Also, when scaling up a population in an AB model by some factor α, it is not generally the case that all other parameters in the model should be multiplied by α as well – such transformations may move the model into a different phase of behaviour. So we are faced with a problem of how to consistently vary the numerous parameters that are often prevalent in AB models, in order to avoid the behaviour of the model changing merely because the inconsistent scaling of parameters has moved it to a completely different phase of the system.

Fortunately, guidance from engineering disciplines can be brought to bear in scaling AB models in a rigorous and consistent manner. As discussed by Sterrett (2009), the general methodology that informs engineers on how to build reduced scale models in a consistent manner has been developed over centuries by many people including Newton. However it was only in 1914 that Buckingham made the formalism explicit, with his famous Pi theorem (Buckingham, 1914). It has since been made more mathematically rigorous by Langhaar (1951) and Palacios (1964), though Buckingham's publication is normally cited in the engineering literature. The first time Buckingham's Pi theorem was advocated for use in scaling AB models appears to be in Osgood (2009), who discusses many of the concepts in section 3 and 4 of this paper, although his focus is on reducing the size of the AB model for reasons of computational feasibility. Moreover, Sterrett (2015) claims that "it remains an open research question whether, and how, the concept of similar systems might be applied to… economics." We provide an example in section 4, which demonstrates its use in an AB model of job search.

3.2 Invariance of the real world to unit-systems and the Buckingham Pi theorem

A key insight of the real world is that its properties are invariant to the systems of units that are used to measure them.[17] For example, the properties of a structure like a suspension bridge should not depend on whether its length is specified in miles or kilometres, or whether the maximum mass it can carry is measured in kilograms or pounds. Mathematical and computational models that attempt to represent real world systems should also capture this property. The Buckingham Pi theorem offers a way of exploiting the dimensionality of quantities in the model to re-specify it in a form that no longer depends on physical (dimensional) units such as time, length, money etc. By expressing the model and it's parameters in a dimensionless form, the model becomes invariant to changes in the scales of dimensional units (i.e. seconds, metres, dollars).

[17] Note that we are not discussing the minimum size of the measurement scale i.e. the granularity or 'resolution' of the measuring apparatus, merely the unit that the measurement is reported in, so Mandelbrot's (1967) famous remark about the length of the coastline of Great Britain depending on the length of the measuring stick is irrelevant.

The Buckingham Pi theorem claims the following. For any physical (dimensionally consistent[18]) relationship that is expressed as a function $x_1 = f(x_2, \dots, x_n)$ where x_1 is the dependent variable and the other x_i are *n*-1 non-zero independent variables specified in *r* independent physical dimensions, the relationship can equivalently be written in terms of (*n*-*r*) dimensionless quantities π_j in the form $\pi_1 = g(\pi_2, \dots, \pi_{n-r})$, where π_1 is the corresponding dependent variable in the dimensionless form of the model. We shall see how to obtain the dimensionless variables π_j in the next subsection.

Note that it is not necessary to know the form of the functions *f* or *g*, and that often the hardest part of the procedure is determining which variables x_i (called the 'governing parameters') are relevant to the problem. The process of specifying the model in terms of dimensionless variables enables all the useful information relating the dependent variable of interest to the independent variables to be presented in the most efficient way, analogous to the method of *Principal Components Analysis* in Econometrics. In particular, the Buckingham Pi theorem allows to reduce the number of variables by the number of fundamental dimensions, from *n* to *n-r*, thus reducing the potential number of parameters to estimate and simplifying the calibration of such a model.

From an engineer's perspective, if we modelled a system at full scale with a relation $x_1 = f(x_2, \dots, x_n)$, whilst the reduced-scale model has parameters and variables described by $x'_1 = f(x'_2, \dots, x'_n)$, then from Buckingham's Pi theorem, we can express these equations as $\pi_1 = g(\pi_2, \dots, \pi_{n-r})$ and $\pi'_1 = g(\pi'_2, \dots, \pi'_{n-r})$, respectively. We are now in a position to define two systems as *similar* whenever they have the same dimensionless variables i.e. $\pi_i = \pi'_i$, and thus behave in the same way. These two systems may have physical quantities such as length or oscillating behaviour with frequencies that differ by several orders of magnitude, however if they have the same π_i variables, they represent the same state of the system. Note that the choice of dimensionless parameters is not unique, with π_i^p also being a valid dimensionless variable for any non-zero rational number *p*.

3.3 Finding the dimensionless π variables

As Barenblatt (2003) argues, dimensional analysis alone is not usually sufficient to prove self-similarity and to find scaling laws. However, he also comments that "the basic difficulty always lies in finding an appropriate model, even a preliminary one. This is a matter of art, and no general recipe can be offered here. But when a researcher arrives at a particular model, and has the intention of working with this model, a certain general system of rules can be recommended," (p. 91). It is our intention to describe this system of rules here.

[18] Dimensional consistency (or homogeneity) requires that all additive terms of an equation must have the same dimensions (if the left-hand side of an equation represents a velocity, the terms on the right-hand side must also have dimensions of [Length] / [Time]) in order to be physically meaningful.

Referring to our relation $x_1 = f(x_2, \dots, x_n)$, if the mathematical form of f is known, then the governing parameters x_i are the variables and parameters that appear in the equation, along with any initial and boundary conditions. If, on the other hand, f is not known, then "the governing parameters must be chosen on the basis of a qualitative model of the phenomenon, to be constructed by each investigator using his/her own experience and intuition as well as an analysis of previous studies," (Barenblatt, p.91).

In order to construct the dimensionless variables, it is necessary to choose r dimensionally distinct 'scaling variables' $\{s_j\}$ (also known as 'repeating variables') from the set $\{x_i\}$. "It is preferable to select those… whose importance to the phenomenon being studied is most firmly established," (Barenblatt, p.92). As the scaling variables s_j are dimensionally distinct (independent), they must span all the r physical dimensions in the model $\{D_i\}$. In addition, the scaling variables must be dimensionally distinct, by which we mean that $[s_j] \neq [s_k]^z$ for all $j \neq k$ and any number z.[19]

Then, for each of the remaining n-r non-scaling variables t_i from the set of governing parameters $\{x_i\}$, a dimensionless variable π_i is constructed by choosing non-zero rational numbers $a_1, a_2, \dots, a_r$ such that the dimensions cancel on the right hand side of the following equation:

$$[\pi_i] = [t_i] \cdot [s_1]^{a_1} \cdot [s_2]^{a_2} \cdot \dots \cdot [s_r]^{a_r} \qquad (7)$$

Note that there is a different dimensionless variable π_i for each non-scaling governing variable t_i.[20] We use the convention that the first n-r governing parameters ($x_1, x_2, \dots, x_{n-r}$) refer to the n-r non-scaling variables ($t_1, t_2, \dots, t_{n-r}$) respectively, while the last r governing parameters ($x_{n-r+1}, x_{n-r+2}, \dots, x_n$) refer to the r scaling variables ($s_1, s_2, \dots, s_r$) respectively.

After finding the expressions for all n-r dimensionless variables π_i in terms of their corresponding non-scaling variable t_i and the set of scaling variables $\{s_i\}$, we estimate the numerical values of $\{\pi_i\}$ by plugging in the estimated values of all the governing parameters $\{x_i\}$, the physical variables of the model. This determines the state of the model in dimensionless form, while the relationship between the dependent dimensionless variable π_1 and the independent dimensionless variables $\{\pi_{j\neq1}\}$ can be discovered through simulation and experiment,

[19] For example, if we represent the dimensions of variable x_i by $[x_i]$, if x_i = 30 km/hour, then by $[x_i]$ we mean Length · Time^{-1}. If the dimensions of a model are in terms of the three dimensions Length, Time and Mass, then each of these dimensions must appear in at least one of the three scaling variables' dimensions $\{[s_j]\}$. In our example, if $[s_1]$ = Length · Time^{-1}, and the dimensions of s_2 also only represent Length and Time, then $[s_2] \neq$ Length^u · Time^{-u} for any number u, although it is possible for $[s_2]$ = Length^u · Time^v for any rational u, v, as long as $u \neq -v$ (so that $[s_2]$ is not merely $[s_1]^z$). To complete this particular example, $[s_3]$ must involve Mass as s_1 and s_2 do not feature this dimension, while $[s_3]$ may include the other dimensions of length and time raised to any rational powers as well.

[20] For example, if the non-scaling variable t_1 represents a force in the model, the unit of force is the Newton, which can be expressed in terms of SI units as kilograms · metres · second^{-2}. In this instance, $[t_1]$ = Mass · Length · Time^{-2}. If, in our example, the scaling variables had the following dimensions $[s_1]$ = Mass, $[s_2]$ = Length^3 (e.g. s_2 represents a volume), and $[s_3]$ = Length · Time^{-1} (s_3 represents a velocity), then by inspection we can see that a = -1, b = -2/3 and c = 2. In this case, we are left with $[\pi_1]$ = Mass^0 · Length^0 · Time^0 = 1, i.e. π_1 is dimensionless. This gives us the functional form for the dimensionless variable, $\pi_1 = t_1 \cdot s_1^{-1} \cdot s_2^{-2/3} \cdot s_3^2$. Repeating this procedure to produce all n-r dimensionless variables π_i allows us to express the model in dimensionless form.

observing how the state of the system changes as we vary the dimensionless variables. This captures the behaviour of the system in its most fundamental form.

3.4 Using the Buckingham Pi methodology to scale up models

Take a model at one scale *M* that is expressed in terms of governing parameters $\{x_i\}$ partitioned into scaling variables $\{s_i\}$ and non-scaling variables $\{t_i\}$, and the same model at another scale *M'* that is expressed in terms of governing parameters $\{x_i'\}$ partitioned into scaling variables $\{s_i'\}$ and non-scaling variables $\{t_i'\}$. They can be compared in terms of their dimensionless variables π_i and π'_i respectively, to see if the two models are in *similar* states (i.e. $\pi_i = \pi_i'$ for all *i*). Moreover, if we know the physical quantities involved in *M* (so we know $\{s_i\}$ and $\{t_i\}$), and we also have the scaling variables $\{s_i'\}$ of model *M'*, we can determine the non-scaling variables $\{t_i'\}$ by equating each dimensionless variable $\pi_i = \pi_i'$ and solving for $\{t'_i\}$:

$$t_i' = t_i \cdot \left(\frac{s_1}{s_1'}\right)^{a_1} \cdot \left(\frac{s_2}{s_2'}\right)^{a_2} \cdot \ldots \cdot \left(\frac{s_r}{s_r'}\right)^{a_r} \qquad (8)$$

This tells us the necessary values of the physical parameters that we must use in order to ensure that the model has been consistently scaled. A general overview to the practical aspects of this procedure can be found in Barenblatt (2003, p.92).

We are also now in a position to see how the dependent variable x_1 on the left-hand side of the original formulation $x_1 = f(x_2, x_3, \ldots, x_n)$ varies with scale. As Osgood (2009) suggests, for reasons of computational feasibility, we may perform a small scale AB model *M* whose output of interest is x_1 (equivalently the non-scaling variable t_1). We can then determine the value x_1' (equivalent to the non-scaling variable t_1') that we would obtain from simulating a larger (or full-scale) model using equation (8).

There is a straightforward procedure involving some linear algebra to calculate the formulas for the dimensionless variables π_i; this will be presented in section 4. We will then demonstrate how to apply this procedure by applying it to a simple AB model.

It is important to note that, whilst these techniques are important in ensuring consistency when changing the scale of a model, there may still be discretisation effects at the smallest scale of AB model, when only a few agents are simulated. These are analogous to the notion of 'incomplete similarity' in the engineering domain, for instance in hydraulic modelling. Whilst it may be possible to scale down the geometric aspects of the model so that all lengths are reduced by a constant factor (geometric similarity), the properties of the fluid (e.g. viscosity and surface tension) and the properties of surface roughness or sediment size (e.g. in river bed modelling) may not be easy to scale down, and so are often not faithfully represented in the reduced-scale model.

Another issue, as Osgood (2009) warns, can occur when modelling heterogeneity within the agent population. If an agent can only exist in one of a finite number of states, the likelihood that there are no agents with a particular state increases as the agent population size is reduced; in this case, the heterogeneity of the model cannot be fully realised. However, on closer inspection, this is nothing more than the problem of sample size determination that we discussed in section 2.1: the sample size is insufficient for guaranteeing that the frequency of some particular state is different from 0 at the desired confidence level.

Finally, note that if the goal is to scale up AB models to larger sizes rather than scaling them down, both the discretisation effect and heterogeneity issue becomes less significant (unless of course we also increase the number of possible discrete states).

4. Applying the Buckingham Pi procedure to scale a simple AB job search model

4.1 The model

We apply the Buckingham Pi procedure to demonstrate consistent scaling of a simple AB job search model. This is one of the sample models included in the java-based JAS-mine AB modelling suite: a description of the model and the source code are available on the JAS-mine website.[21] The model consists of N worker agents, who apply for job vacancies whenever they are unemployed. At each time-step, a number of new vacancies V are created and subsequently remain open for H time-steps. Then, each unemployed worker sends a number of job applications A per time-step. When it is time for the vacancy to close (H time-steps after it opened), an applicant is randomly accepted for the job, and all the other applications that the successful applicant has sent (to other vacancies) are removed from the system. In addition, the other applicants who are still unemployed are informed that they have been unsuccessful with their application for this vacancy. To simplify the model, employment relationships only last for a single time-step, after which the worker becomes unemployed and again proceeds apply to the available job vacancies. As such, each vacancy has a list of applicants, and each unemployed worker has a list of vacancies they have applied to. The model thus involves four independent dimensions: ‘Workers’, ‘Applications’, ‘Vacancies’ and ‘Time’.

At time 0, there are no open vacancies, and all workers start as unemployed. The model then undergoes a transient phase where the number of open vacancies in the system linearly increases (at a rate of V new vacancies per period) until time $\tau = H$, when the first vintage of vacancies closes and is processed. During this initial period, job applications also pile up in the system up until $\tau = H$, when the employment rate jumps to its equilibrium level.

Assume we are interested in modelling job queues Q, defined as the average number of applicants per open vacancy. This is our dependent variable, called x_1 using the conventions in section 3. Q has dimensions

[21] http://www.jas-mine.net/demo/applications.

Applications · Vacancies^{-1}. Up to time $\tau = H$, the average number of applicants per open vacancy is simply equal to $N \cdot A / V$ (a constant number of new vacancies and new applications are added at each time-step, so Q remains constant). After time $\tau = H$, the job queue decreases as successful applicants withdraw their applications to other vacancies. The equilibrium level Q^* is reached at time $\tau = 2H$, when the last vacancy to open during the initial period of total unemployment (and consequently having receiving more than the equilibrium number of applications), is finally closed.

To use the Buckingham Pi's procedure, as there are four independent dimensions, we must therefore choose four 'scaling variables' (s_1, s_2, s_3 and s_4) for the analysis. These scaling variables cannot be dimensionless and must be linearly independent, spanning the space of all the four dimensions (in our case 'Workers', 'Applications', 'Vacancies' and 'Time'). Recall that these scaling variables s_i must be ones that are believed to have the greatest influence on the dependent variable x_1, and should also be the variables that we explicitly wish to specify for each scale of the model. This is an easy choice in our simple model, where we have the current simulation time τ whose dimension is (obviously) Time, the number of new vacancies per time-step V with dimensions Vacancies · Time^{-1}, the number of applications an unemployed worker submits per time-step A with dimensions Applications · Worker^{-1} · Time^{-1}, and the number of workers N with dimension Workers. The only other parameter in the model, the number of time-steps that a job vacancy remains open H, has dimension Time which is dimensionally dependent on the simulation time (both variables have the same dimension). H cannot therefore be a scaling variable s_i, and must be a non-scaling variable, which we label t_2. Lastly, the only other non-scaling variable in this prescription is the dependent variable of interest, the Job Queue x_1, which by convention is labelled as the first non-scaling variable t_1.

To solve for the dimensionless variables π_i, we simply take the logarithm of equation (7), and remember that $[\pi_i] = 1$,

$$0 = \log[t_i] + a_1 \log[s_1] + a_1 \log[s_2] + ... + a_r \log[s_r] \qquad (9)$$

As the dimensions of the scaling variables s_i are linearly independent and the equation must hold for all i, constraints are placed on the values of the exponents a_1, a_2 and a_r. More details on the derivation of the formulae for the π_i are described in the Appendix. Applied to the job search model, we get

$$\pi_1 = Q \cdot N^{-1} \cdot A^{-1} \cdot V$$
$$\pi_2 = H \cdot \tau^{-1} \qquad (10)$$

Each dimensionless variable π_1 and π_2 is associated with one non-scaling variable (Q and H respectively), and expressed as a combination of powers of the non-scaling variable and the set of scaling variables $\{s_i\}$, just as described by equation (7). The formulae for the dimensionless variables π_i are specific to the choice of scaling variables $\{s_i\}$ and the dimensionality of the dependent variable. Choosing a different dependent variable with the same dimensions as Q would produce the same formulae for the dimensionless variables π_i, given the same scaling variable set. Moreover, changing the dependent variable would only affect the formula for π_1. Changing the scaling variables $\{s_i\}$ to ones with different combinations of dimensions, on the other hand, will change the formulae for all dimensionless variables $\{\pi_i\}$.

4.2 Using the dimensionless variables to scale the Applications model

Consider running the model at two scales *M* and *M'*; this entails two sets of (independent) governing parameters $\{N, A, V, \tau, H\}$ and $\{N', A', V', \tau', H'\}$ respectively. We assume that *M* is computationally more feasible to run than *M'*, i.e. *M* requires less computing time and/or memory, for example, because it involves fewer numbers of worker agents ($N < N'$). In order to infer the properties of *M'* from simulations of *M*, the two versions of the model must be *similar* (defined in section 3.2 as meaning $\pi_i = \pi'_i$). If this does not hold, then it is impossible to rule out a difference in behaviour of *M'* compared to *M* as the two versions of the model are in a different phase-state. We therefore construct *M'* so that it is similar to *M* by equating the dimensionless variables of *M'* to those in *M*. This provides two additional constraints on the governing parameters $\{N', A', V', \tau', H'\}$ of model *M'*, as demonstrated in equation (11).

$$\begin{gathered} Q \cdot N^{-1} \cdot A^{-1} \cdot V = \pi_1 = \pi'_1 = Q' \cdot N'^{-1} \cdot A'^{-1} \cdot V' \\ H \cdot \tau^{-1} = \pi_2 = \pi'_2 = H' \cdot \tau'^{-1} \end{gathered} \tag{11}$$

We can use the formulae in equation (11) to solve for the dependent variable in the model *M'*, in our case, the job queue *Q'*, as in equation (12):

$$Q' = Q \cdot (N'/N\,) \cdot (A'/A\,) \cdot (V/V') \tag{12}$$

Thus we can determine how *Q'* is related to the corresponding value *Q*, conditional on knowledge of how the parameters {*N'*, *A'* and *V'*} compare to {*N*, *A* and *V*}. Note that the latter equation (12) seems to intuitively make sense; the job queue *Q'* increases if we increase the number of workers *N'* and the number of applications each worker makes *A'*, compared to parameters of the simulated version of the model (*N* and *A* respectively). Additionally, the job queue shortens if we increase the number of vacancies that open at each time-step *V'*. Note that no knowledge of the underlying processes (the functions *f* and *g* in section 3) has been used to derive these scaling relations, merely the dimensions of the relevant variables.

We show the estimates of the equilibrium values of Q^* for several scaled parameter sets in table 1, along with the average equilibrium values obtained by simulation *Q*. Note that as we are interested in the equilibrium, we do not include values for the simulation time τ, although we illustrate the typical evolution of Q through time by providing simulated time-series of $Q(\tau)$ in figures 4 and 5.

	Independent Variables (the model parameters)				**Dependent Variable**	
Simulation Run ('scale')	Duration of open vacancies, *H*	Number of workers, *N*	Applications per time-step, *A*	New vacancies per time-step, *V*	Estimated equilibrium job queue, *Q**	Realised equilibrium job queue, *Q*
1 (scale *M*)	30	1,000	1	30	-	16.72 ± 0.16
2	30	10,000	1	30	167.2	167.64 ± 0.50
3	10	250,000	2	15,000	16.72	16.67 ± 0.01
4	5	370,000	3	9,000	61.86	61.67 ± 0.05
5	150	1,000	1	30	16.72	16.68 ± 0.07
6	10	1,000	1	30	16.72	16.76 ± 0.30

Table 1. Simulations of the Applications Model. The values of the parameters {*H*, *N*, *A* and *V*} and the estimated *Q** and realised *Q* values of the equilibrium job queue for distinct simulation runs. The estimates *Q** have been calculated using equation (12) and knowledge of the realised value *Q* for Simulation Run 1 (corresponding to scale *M*).

Note that there is no entry for the estimated job queue *Q** in Simulation Run 1, as this is the baseline version of the model (corresponding to scale *M*) to which the other simulation runs are compared. The realised value of the job queue *Q* in this run, along with the other model parameters {*H*, *N*, *A*, *V*} are used in equation (12) in order calculate the predicted ('estimated') value of the job queue *Q**, given the parameter sets {*H'*, *N'*, *A'*, *V'*} – the new 'scales' *M'* – of the other simulation runs.

Simulation Run 2 is an example of inconsistent scaling, in the sense that by only increasing the number of agents, we are not scaling the system up as a whole, merely changing one of the model parameters and thereby effecting the behaviour of the model output – as seen by the fact that the realised equilibrium job queue *Q* has increased by the same factor as the number of agents. Indeed, equation (12) correctly predicted this new job queue value, as can be seen in the column for *Q**.

If we wanted to be able to scale the system up or down, while ruling out any change in behaviour caused by moving the system to a different phase-state (exhibited as a change in the dimensionless parameters π'_i of the system), we can use equation (10) to guide us. We enforce similarity of the scales by equating all the dimensionless parameters between the initial scale and the desired scale. Having done this earlier to derive equation (12), setting *Q'* = *Q* in this equation (12) gives the following relation:

$$1 = (N'/N) \cdot (A'/A) \cdot (V/V') \qquad (13)$$

So knowing the change in scale of any two parameters from {*N*, *A*, *V*} fixes the change in scale of the third. We demonstrate this in Simulation Run 3, where we randomly chose a set of parameters to allow us to consistently scale up the Applications model from 1000 agents to 250,000 agents. As can be seen in table 1, the observed equilibrium job queue Q^* of the simulation was 16.67 applications per vacancy, in close agreement with the predicted Q value of 16.72.

In Simulation Run 4, we demonstrate the ability to estimate the job queue for a much more computationally intensive parameter set, involving 370,000 workers and the creation of 9,000 new vacancies at each time-step. The use of the Buckingham Pi methodology presented here allowed us to estimate the correct equilibrium value of the job queue, without having to endure the far greater computational costs involved in obtaining the value through simulation. We do, however, provide the result of the simulation as a cross-check.

Simulation Runs 5 and 6 illustrate the power of the Buckingham Pi methodology in determining which parameters are not considered to affect the dependent variable. In this case, equation (12) suggests there is no dependence of the job queue Q on parameter H, the number of time-steps that a vacancy remains open. We test this by comparing Simulation Runs 5 and 6 to Simulation Run 1 and find there is indeed no effect on the realised equilibrium job queue value Q.

We illustrate the time-series of the job queue Q in versions of the model that have been scaled consistently, (i.e. in engineering terms, they are *similar*) in figure 4. (Note that we ignore kinematic similarity for the time being, as we discuss the impact of changing parameter H next).

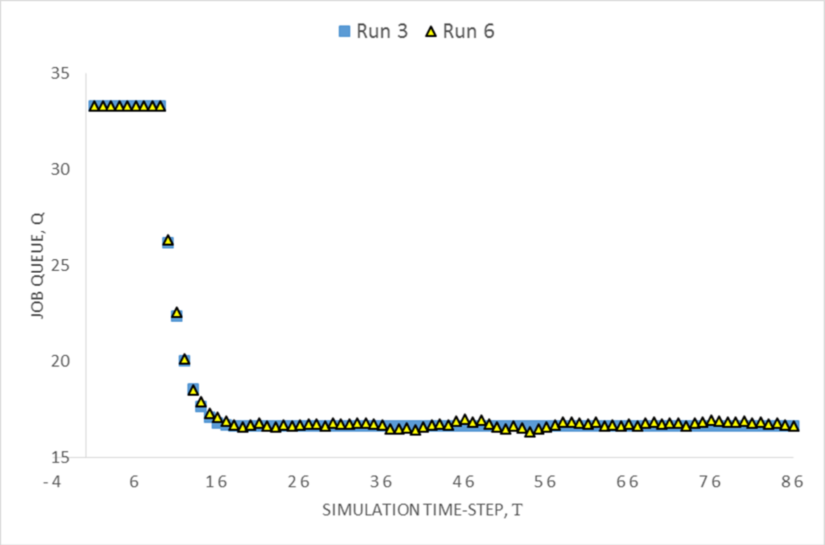

Figure 4. The time-series of the job queue Q for simulations that have been scaled consistently, i.e. they are *similar*.

Finally, we focus on the second dimensionless variable in the model, using equation (11) and the same procedures to derive the following equation for the duration of open vacancies:

$$\tau'^{-1} = \tau^{-1} \cdot (H/H') \tag{14}$$

This suggests that reducing the value of the duration for which vacancies remain open to a third of the initial scale (i.e. $H' = H/3$) will triple the 'speed' of processes in the system (i.e. $1/\tau' = 3/\tau$), as τ is the simulation time so τ^{-1} is the rate of evolution in the system. We illustrate this behaviour by comparing the time series' of the job queue Q in Simulation Runs 1 and 6 in figure 5. We also provide an example of slowing the speed of processes by increasing H by a factor of 5 in Simulation Run 5.

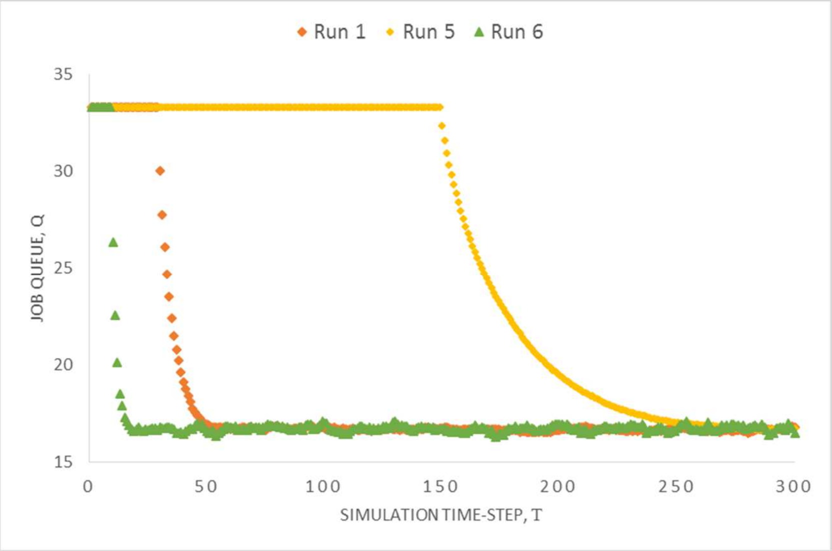

Figure 5. The time-series of the job queue Q for Simulation Runs 1, 5 and 6 are geometrically similar, but not kinematically similar. The parameter controlling the number of time-steps that a job vacancy remains open H influences the speed of the convergence to the equilibrium job queue.

To summarise, without any detailed knowledge of the underlying processes and structure of the model, the procedure we have outlined in sections 3 and 4 has allowed us to correctly predict the dependent variable – the job queue Q – of a simple AB job search model, for a variety of different parameter sets ('scales') based on the simulation results from one simulation using a reference parameter set. We have been able to understand the influence that the independent variables had on the dependent variable, and also the relationships between the independent variables. While Osgood (2009) also applies this procedure on an AB epidemiological model to good effect, we cannot be sure it will always work so successfully. Indeed, Barenblatt (2003) warns that "the situation when everything including scaling laws can be obtained through dimensional analysis alone is in fact very rare," (p. 69).

5. A case when the Buckingham Pi theorem cannot help with scaling: the Erlang model

In section 2.3 we described the Erlang B queuing model, as an example where the behaviour of the system (the blocking probability p) depends on the size of the system (the number of servers, N). One might wonder if the Buckingham Pi theorem can offer some guidance on how the model scales, without knowing the exact scaling relationship, equation (6). After all, it is quite unlikely that a mathematical expression for a relationship between aggregate variables can be derived in an AB model. However, at a closer inspection it is clear that the Buckingham Pi theorem cannot be applied in this setting.

The model, in its compact form, has n = 3 variables: the blocking probability p, the parameter measuring the offered load λ, and the number of servers N, and only one dimension, Servers. From these 3 variables, we need to select a scaling variable, however given that both p and λ are dimensionless parameters, they must be non-scaling variables (t_1 and t_2 using the labelling conventions from section 3). The number of servers N, must therefore be used as the scaling variable (s_1 by convention of section 3). Equation (7) then gives us the formulae

$$[\pi_1] = [p] \cdot [N]^a$$
$$[\pi_2] = [\lambda] \cdot [N]^b \tag{7}$$

However, in order for the right-hand sides of both expressions to be dimensionless, the exponents a and b have to be equal to zero. Thus, no scaling variables show up in the π terms – indeed, p and λ are already dimensionless π terms. There is no room to manoeuvre: the model has no variables that can offset the effect of N on the model output.

Specifying the model in terms of the arrival rate ν (the parameter of the Poisson distribution) and the (inverse of the) mean call duration μ (the parameter of the negative exponential distribution) does not help either. Now the number of variables is equal to n = 4 (p, ν, μ, N) and the number of dimensions has increased to r = 3 (Servers, Calls and Time); hence, three dimensionally independent variables are required as scaling variables. However, we cannot use both ν and μ as scaling variables because they have the same relative dimensions ($\text{Calls}^k \cdot \text{Time}^{-k}$ for $k = \pm 1$) and so are not dimensionally independent. Solving for the equations nonetheless simply restates the definition of the offer load λ: $\pi_2 = \lambda = \nu/\mu$.

6. Conclusions

We defined scale effects as a situation in which some outcome of interest y depends on the size of the system S, and we suggested a way in which this dependency can be modelled even without knowing the explicit functional form $y = y(S)$. This involves the 'nondimensionalisation' of the system, whenever possible, using the Buckingham Pi theorem. The theorem suggests to look at transformations of the outcome variables of interest that are invariant to scaling, and suggests a way to scale the system consistently. Given its simplicity, the Buckingham Pi theorem is a very general and powerful result, and it provides necessary conditions for scaling an AB model from one 'size' (set of parameters) to another in a way that guarantees similarity between the scales. If two versions of a model with differing scales are not deemed *similar* in the engineering sense, then it should not be surprising if their behaviour differs. Moreover, the Buckingham Pi framework provides valuable insights into the nature of the scaling issue even when the theorem cannot be applied, as in the Erlang B example.

However, one might ask why the scaling parameters (e.g. the number of agents, the size of the simulated world, etc.) are not simply treated as any other parameter of the model (such as behavioural parameters). This would entail applying the tools of sensitivity analysis to understand the effects of the scaling parameters on the model outcome, in exactly the same way we do for the other parameters: by simulating the model for different parameters values. The answer is, obviously, that in many cases it is computationally too costly to run the necessary number of simulations with realistic values of the parameters. While we might be able to put limits on the admitted values of behavioural parameters, they remain intrinsically unknown (hence the interest in estimating them). On the other hand, often we know the 'real' value of the scaling parameters already; we do not even need to estimate them and can simply apply them in the model. The problem is that they are often in the order of millions (e.g. the number of individuals, household, or firms in an economy). Therefore, it would be invaluable to be able to understand the behaviour of a full scale model simply by simulating it on vastly reduced scales. When the Buckingham Pi theorem is applicable, that is, when there are enough dimensionally independent scaling variables in the system to span all the scalable dimensions, it is possible to put constraints on how such small scale versions must look. This, however, should not dispense from running some simulations at a much larger scale, in order to see whether additional scaling effects are present in the model, possibly due to the discrete nature of some variable. We leave the investigation of how to deal with these issues open for further research.

References

Anderson PW (1972). More Is Different. *Science* 177(4047): pp. 393-396.

Axtell RL (2011). Modeling the Economy with 150 Million Agents, TEDxUVM "Big Data, Big Stories", University of Vermont, Vermont Complex System Center, 28 October 2011.

Bar-Yam Y (2004). A Mathematical Theory of Strong Emergence Using Multiscale Variety. *Complexity* 9(6): pp. 15-24.

Barenblatt, GI (2003). *Scaling*. Cambridge University Press, Cambridge, UK.

Bettencourt LMA (2013). The Origins of Scaling in Cities. *Science* 340: pp. 1438-1441.

Buckingham E (1914) On Physically Similar Systems: Illustrations of the Use of Dimensional Equations. *Physical Review* 4: pp. 345–376.

Bohorquez JC, Gourley S, Dixon AR, Spagat M, Johnson NF (2009). Common ecology quantifies human insurgency. *Nature* 462: pp. 911-914.

Caron-Lormier G, Humphry RW, Bohan DA, Hawes C, Thorbek P (2008). Asynchronous and synchronous updating in individual-based models. *Ecological Modelling*, 212(3): pp. 522–527.

Coakley S, Gheorghe M, Holcombe M, Chin S, Worth D, Greenough C (2012). Exploitation of High Performance Computing in the FLAME Agent-Based Simulation Framework. Proceedings of the 14th International Conference on High Performance Computing and Communications (HPCC2012), Liverpool, UK, New York: IEEE, 2012, pp. 538–545.

Cockrell RC, Christley S, Chang E, An G (2015). Towards Anatomic Scale Agent-Based Modeling with a Massively Parallel Spatially Explicit General-Purpose Model of Enteric Tissue (SEGMEnT_HPC). PLoS ONE 10(3): e0122192.

Christensen K, Moloney NR (2005). *Complexity and Criticality*. Imperial College Press, London, UK.

De Oliveira PMC, Nóbrega RA, Stauffer D (2003). Corrections to finite size scaling in percolation. Brazilian *Journal of Physics* 33(3): pp. 616-618.

Epstein JM, Axtell RL (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. The MIT Press, Cambridge, MA.

Filatova T, Verburg PH, Parker DC, Stannar CA (2013). Spatial agent-based models for socio-ecological systems: Challenges and prospects. *Environmental Modelling and Software*, 5: pp. 1-7.

Gabaix X (2009). Power Laws in Economics and Finance. *Annual Review of Economics* 1: pp. 255-93.

Grazzini J, Richiardi M (2015). Consistent Estimation of Agent-Based Models by Simulated Minimum Distance. *Journal of Economic Dynamics and Control* 51: pp. 148-165.

Helbing D, Molnar P, Farkas IJ, Bolay K. (2001). Self-organizing pedestrian movement. *Environment and Planning B: Planning and Design* 28: pp. 361-383.

Holcombe M, Coakley S, Kiran M, Chin S, Greenough C, Worth D, Cincotti S, Raberto M, Teglio A, Deissenberg C, van der Hoog S, Dawid H, Gemkow S, Harting P, Neugart M. (2013). Large-Scale Modeling of Economic Systems. *Complex Systems* 22: pp. 175-191.

Langhaar HL (1951). *Dimensional Analysis and Theory of Models*. John Wiley.

Lysenko M, D'Souza R (2008). A Framework for Megascale Agent Based Model Simulations on Graphics Processing Units. *Journal of Artificial Societies and Social Simulation* 11(4): art. 10.

Mandelbrot B (1967). How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension. *Science*, 156 (3775).

Osgood N (2009). Lightening the Performance Burden of Individual-Based Models through Dimensional Analysis and Scale Modeling. *System Dynamics Review* 25(2): pp. 101-134.

Palacios J (1964). *Dimensional Analysis*. (Translated from the Spanish by Lee and Roth) Macmillan.

Richiardi M, Richardson R (2016, forthcoming). Agent-based Computational Demography and Microsimulation using JAS-mine. In: Grow A, van Bavel J. Agent-Based Modeling in Population Studies. Springer Series on Demographic Methods and Population Analysis, Berlin. Available at www.jas-mine.net.

Saltelli A, Chan K, Scott, EM (2000). *Sensitivity analysis*. Jonh Wiley & Sons Ltd., Chichester, West Sussex England.

Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S. (2008). *Global sensitivity analysis. A primer*. John Wiley & Sons, Ltd., Chichester, West Sussex England.

Sterrett, SG (2009). Similarity and Dimensional Analysis. In: Mejers A (ed.). *Philosophy of Technology and Engineering Sciences*. Elsevier/North Holland, Amsterdam: pp. 799-823.

Sterrett SG. (2015) *Physically Similar Systems: a history of the concept.* http://philsci-archive.pitt.edu/11352/ [Accessed on 7th June 2015].

The Economist (2010). Data, data everywhere. 25 February 2010.

Tsionas M, Grazzini J, Richiardi M (2015). Bayesian inference in ergodic AB models. Institute for New Economic Thinking, mimeo.

van Leeuwaarden JSH, Temme NM (2009). Asymptotic Inversion of the Erlang B Formula. *SIAM Journal of Applied Mathematics*, 70(1): pp. 1-23.

Wolfson M, Gribble S, Beall R (2016, forthcoming). Exploring Contingent Inequalities – Building the Theoretical Health Inequality Model. In: Grow A, van Bavel J. Agent-Based Modeling in Population Studies. Springer Series on Demographic Methods and Population Analysis, Berlin.

Zohuri B (2015). *Dimensional Analysis and Self-Similarity Methods for Engineers and Scientists*. Springer.

Appendix

In this Appendix we demonstrate a method for deriving the formulae for the π_i based on solving a matrix equation, which reflects the underlying need to solve a set of simultaneous linear equations that constrain the exponents on the governing parameters. We start by defining a block matrix structure as in table A1.

		Governing Parameters, x_i						
		Non-scaling Variables, t_i				**Scaling Variables, s_i**		
		$x_1 = t_1$	$x_2 = t_2$	…	$x_{n-r} = t_{n-r}$	$x_{\text{n-r+1}} = s_1$	…	$x_{\text{n}} = s_r$
Dimensions	D_1	r by n-r matrix, T				r by r matrix, S		
	…							
	D_r							
Dimensionless Variables, π_i	π_1	n-r by n-r Identity matrix, I				n-r by r matrix, P		
	…							
	π_{n-r}							

Table A1. Block structure of the Dimensional Matrix.

The entries of the sub-matrices S and T in table 1 are the exponents of the dimensions on the governing parameters x_{i}. Note that the dependent variable is represented on the left-most column, x_1 (equivalently, t_1). Also, if a governing parameter is dimensionless, as is possible for non-scaling variables such as unemployment rates or the blocking probability in the Erlang model (i.e. $[x_k] = [t_k] = 1$), the associated column k will only have entries of zeroes in matrix T. In this case, the governing parameter *is* the dimensionless variable, i.e. $\pi_k = x_k$.

Entries in matrices I and P reflect the exponents of the governing parameters x_{i} in the expression for the dimensionless π_i variables; once matrix P is known, it is easy to obtain the formulae for the π_i, as we shall explain. The n-r by n-r identity matrix I occupying the bottom left block associate a dimensionless π_i variable for each non-scaling variable t_i, while the matrix P is calculated by solving the implied set of simultaneous linear equations that constrain the exponents of the governing parameters, as in equation (9). This can be expressed succinctly as solving the following matrix equation:

$$P = - (S^{-1} \cdot T)^T \tag{A1}$$

It is now clear why the scaling variables $\{s_i\}$ have to be (dimensionally) independent and span the number of dimensions (for matrix *S* to be full rank) – this is because the *S* needs to be invertible to solve for *P*.

In our job search model example, the matrix form is represented below (table A2), after solving for matrix *P*.

		Non-scaling Variables		**Scaling Variables**			
		Job queue, *Q*	Duration of open vacancies, *H*	Number of workers, *N*	Applications per time-step, *A*	New vacancies per time-step, *V*	Current simulation time, τ
Dimensions	Workers	0	0	1	-1	0	0
	Applications	1	0	0	1	0	0
	Vacancies	-1	0	0	0	1	0
	Time	0	1	0	-1	-1	1
Dimensionless Variables	π_1	1	0	-1	-1	1	0
	π_2	0	1	0	0	0	-1

Table A2. The Dimensional Matrix for the Applications model

The formulae for the dimensionless variables can now be deduced by reading the entries for the bottom two rows of Table 2. The entries are the exponents of the governing parameters associated with each dimensionless variable, and give equation (10) in the text.