RESEARCH ARTICLE

Invasion biology and the success of social collaboration networks, with application to the Wikipedia

M. Mangel[a,b,1], W.H. Satterthwaite[a], P. Pirolli[c], B. Suh[c,d], and Y. Zhang[c,e]

[a]*Department of Applied Mathematics and Statistics, University of California, Santa Cruz,CA, USA* [b]*Department of Biology, University of Bergen, POB 7803, 5020 Bergen, Norway,* [c]*PARC(Palo Alto Research Center), 3333 Coyote Hill Road, Palo Alto, CA 94304. USA,* [d]*Current address: Adobe Research, 345 Park Avenue San Jose, CA 95110, USA;* [e]*Current address: Google Inc., 2025 Garcia Avenue, Mountain View, CA, USA*

(*Received XX YY 2013; accepted ZZ QQ 2013)*

[1] Corresponding author: msmangel@ucsc.edu

Abstract.

We adapt methods from the stochastic theory of invasions -- for which a key question is whether a propagule will grow to an established population or fail -- to show how monitoring early participation in a social collaboration network allows prediction of success. Social collaboration networks have become ubiquitous and can now be found in widely diverse situations. However, there are currently no methods to predict whether a social collaboration network will succeed or not, where success is defined as growing to a specified number of active participants before falling to zero active participants. We illustrate a suitable methodology with Wikipedias. In general, Wikis are web-based software that allows collaborative efforts in which all viewers of a page can edit its contents online, thus encouraging cooperative efforts on text and hypertext. The English language Wikipedia is one of the most spectacular successes, but not all wikis succeed and there have been some major failures. Using these new methods, we derive detailed predictions for the English language Wikipedia and in summary for more than 250 other language Wikipedias. We thus show how ideas from population biology can inform aspects of technology in new and insightful ways.

Keywords: Invasion biology, stochastic population theory, social collaboration networks, Wikipedia

Introduction

It is difficult to overestimate the effect that Charles Elton's book *The Ecology of Invasions by Animals and Plants* (Elton 1958) had on population biology. Fifty years after its publication, it was still the most highly cited source in the subject (Richardson & Pysek 2008). Although invasion biology covers a wide intellectual swath, one of the fundamental questions concerning invasions is whether a founding population of a certain size will establish itself or fail – i.e. will the descendants of the initial propagule(s) reach a level considered to be self-maintaining or fall back to 0? The answer to this question uses stochastic population theory (MacArthur & Wilson 1967; Mangel 2006).

Social collaboration networks have become ubiquitous (Newman 2010, Easley & Klineberg 2011) and when such a network is started one can envision it as a propagule of users 'invading' the worldwide web. Understanding whether a social collaboration network will succeed or fail is a widely interesting problem, with applications that range from investment in social media, to monitoring terrorist networks, to science education (e.g. Micklos et al. 2011) but we currently lack methods to make predictions of success.

In this paper, we use methods from stochastic population theory (Mangel & Ludwig 1977, Mangel 1994, Mangel 2006) to show how such a prediction can be done. Population biologists will find our approach familiar, but with a new application and connection to the world of technology. Technologists interested in social networks will find novel features that include i) characterizing the stochastic growth of a social collaboration network, ii) a clear definition of success in terms of reaching a specified number of active users before falling to 0 [or any other low number] of active users; iii) derivation and solution of the differential equation characterizing the probability of success; iv) application of Maximum Likelihood Estimation (MLE) and the Akaike Information Criterion (AIC) to determine parameters and then weight possible models of success given data on participation; and v) application of Markov Chain Monte Carlo (MCMC) methods to determine the distribution of parameters for each potential model. We mainly focus on the English language Wikipedia, but consider the remaining 250+ languages in the Supplementary Information (SI). In summary, we show a novel and important interdisciplinary approach that is broadly applicable.

About Wikis

A wiki is a website in which multiple, dispersed participants create and edit interlinked web pages (Mader 2008, Eberscbach et al. 2008). One of the most famous wikis is the Wikipedia, an online encyclopedia in more than 250 different languages, which had remarkable success. Indeed, when viewed backwards across the time scale of a decade, the success of the English Wikipedia (henceforth Wikipedia), measured by the number of individuals actively creating and editing pages (henceforth editors), seems inevitable (Fig. 1a). However, early on the growth of the

Wikipedia shows fluctuations (Fig. 1b) and the monthly change in the number of editors was sometimes very negative (Fig. 1c).

Not all wikis succeed, and some of them – including those created by the *Los Angeles Times,* Penguin publishing, and Amazon.com - were very short-lived (Mader 2008). Because wikis and other social collaboration networks are important in many fields, it is vital to ask how we can predict the success of a social collaboration network given early data, such as that shown in Figs 1b,c. Although wikis can be of many sorts, we use the Wikipedia as a case study.

In light of the fluctuations in the number of editors, the predicted success of a wiki must be a probabilistic quantity. Thus, our goal becomes computing the probability $u(w)$ that a wiki currently with w editors (the w indicating that individuals are writing and editing web pages) reaches a specified upper threshold w_u in which it is considered to be established before it falls to 0 editors.

Methods

In order to keep the technical details to a minimum, we have put most of them in the Appendix, but there are some details that cannot be skimmed over. To capture the dynamics shown in Figure 1 requires a model that is both nonlinear (because of the sigmoidal shape) and stochastic (because of the fluctuations). We let $W(t)$ denote number of editors at time t and $dW = W(t+dt) - W(t)$. The entire trajectory in Figure 1a could then be described by a stochastic differential equation of the form (Schuss 1980)

$$dW = r(W)dt + \sigma(W)dB \quad (1)$$

where $r(W)$ and $\sigma(W)$ are functions of the number of editors, which have to be determined, and dB is an increment in Brownian motion: it is normally distributed with mean 0 and variance $dt + o(dt)$, where $o(dt)$ denotes terms that are higher powers of dt.

For the analogy with a biological invasion, we assume that the number of editors is small, so that 'density dependent' effects that lead to the sigmoidal growth can be ignored. We then Taylor expand the functions on the right hand side of Equation (1), keeping only the linear terms to obtain

$$dW = (r_0 + r_1 W)dt + (\sigma_0 + \sigma_1 W)dB \quad (2)$$

where r_0, r_1, σ_0 and σ_1 are constants to be determined. As these constants vary, we have many possible models, depending upon which of these parameters is set equal to 0. In the absence of fluctuations (i.e. the second term on the right hand side is set equal to 0), when $r_1 > 0$, Equation (2) corresponds to $W=0$ being an unstable steady state and will lead to exponential-like growth in the number of editors. By setting some of the parameters in Equation (2) equal to 0, we generate various forms of the model. For example, a model with constant deterministic and stochastic components is obtained by setting $r_1 = \sigma_1 = 0$.

The statistical interpretation of Equation (2) is that given $W(t) = w$, dW is normally distributed with mean $(r_0 + r_1 w)dt + o(dt)$ and variance $(\sigma_0 + \sigma_1 w)^2 dt + o(dt)$, where $o(dt)$ denotes higher powers of dt.

We used techniques of Maximum Likelihood Estimation (MLE) to determine the most likely values of the parameters using the data shown in Figure 1c and Equation (2). We then used the Akaike Information Criterion (AIC, Burnham & Anderson 2002) to select the most parsimonious models among the variants of in which one or more of the parameters equal to zero and to weight the support that the data give to the various models.

In the Appendix we show that the probability of success $u(w)$ satisfies (also see Mangel & Ludwig 1977, Keizer 1987, Stratonovich 1992, Mangel 1994)

$$\frac{(\sigma_0 + \sigma_1 w)^2}{2}\frac{d^2u}{dw^2} + (r_0 + r_1 w)\frac{du}{dw} = 0 \qquad (3)$$

Since $u(w)$ is the probability of reaching w_u editors before falling back to 0 given that the current number of editors is w, the associated boundary conditions are $u(w_u) = 1$ and $u(w_u) = 1$. For the Wikipedia with data shown in Figure 1c, the solution of Equation (3) is a prediction of the probability success given the data up to month 18.

To reflect model uncertainty, at each value of the number of editors, we computed the mean weighted by AIC weight (Burnham & Anderson 2002)and variance (across the models) of the probability of success. In order to account for parameter uncertainty within each model, rather than between models, we used Markov Chain Monte Carlo (MCMC) methods to obtain distributions for the parameters in each of the models. We used a Bayesian approach to describe the joint posteriors of the model parameters and generate median, central 68%, and central

95% credible intervals on $u(w)$. These details are described in the Appendix.

Based only on a single language Wikipedia, we cannot assess model performance in terms of how well predicted success probability matches the frequency of success of replicate Wikis. To assess the ability of the AIC-determined best model to predict the relative risk faced by multiple Wikipedias, we downloaded data on monthly counts of “active users” (at least 5 edits per month) in 268 different language Wikipedias.

For each language, we identified the first month with a nonzero editor count and then extracted the first 18 monthly counts, allowing the estimation of 18 values of the number of editors and 17 values of the monthly change in the number of editors. We then fit the best supported and most parsimonious model according to our original analysis for the English Wikipedia to each dataset, assessed uncertainty the parameters for each language, and estimated $u(w)$ and its associated uncertainty for a range of values of w, including the actual editor count on the 18th month of that language Wikipedia’s existence. Since some languages have an inherently small user base and would be unlikely to sustain more than 1,000 editors even if successful, we used both 50 editors and 1,000 editors as the upper value w_u.

Results

In Table 1, we show the best models identified by the AIC analysis and in Table 2 the parameters for the 5 best models, which have 97% of the AIC weight.

In Figure 2a, we show the weighted mean probability of success with error bars determined by the square root of the variance. From this figure, we see that based on the weighted mean, once the Wikipedia crossed 250 editors there was about a 70% chance of growing to 1000 editors before returning to 0 editors. The relatively wide error bars in Figure 2a are due to the considerable variation in the individual solutions, which we show in Figure 2b.

In Figure 3, we show two-dimensional plots of the joint posteriors of all pairs of parameters for the English language Wikipedia, but in the case of more complicated models these are still integrated over the uncertainty in the remaining, unseen parameters.

In Figure 4, we bin together languages with similar success probabilities predicted using Model 1 (individual results are in the online Supplementary Information) and count the actual fraction of non-English Wikipedias succeeding or failing. In general, languages with a higher predicted probability of success were more often successful, but there was not a close correspondence between the predicted success probability and the fraction of languages succeeding. In particular, most of the data fall below the 1:1 line if we use the MLE success probability (Figure 4a) but are on either side of the 1:1 line if we use the 68% credible interval (Harte 2001; Figure 4b). The failure of Model 1 to more accurately predict the probability of success across all the Wikipedias may indicate that each language needs its own best model; in the case of languages with small numbers of readers, nonlinear terms in Equation

(1) may become important long before 50 editors is reached.

Discussion

Simberloff (2009) recently reviewed the role of propagule pressure in biological invasions and our results have a clear analogue through the mean rate of growth and its dependence on the current number of editors. The success of a social collaboration network can also be viewed as an "epidemic", characterized (Gladwell 2000) by contagion, the deterministic component Equation (1), which measures the mean rate of growth; and stickiness, the stochastic component of Equation (1), which measures the volatility of editor numbers. Individuals interested in increasing the success probability of a social collaboration network (e.g. Lewis et al. 2012) could use our methods to ask if the probability of success is improved more by increasing the deterministic component or decreasing the volatility.

The probability that we compute does not characterize the sustainability (large-scale adoption in the sense of Mader (2008)) of a wiki or other social collaboration network, but does identify those that are more likely to succeed than not. Indeed, a wiki could fall to 0 editors and then rebound; the analogue of a local extinction and then re-invasion.

Similarly, we have not discussed a mechanism by which the social collaboration network grows in its nonlinear phase. Mechanism becomes much more important if one wishes to understand the full trajectory, including the point of inflection, saturation, and fluctuations around the level of saturation. Although that is a subject for a different paper, we can give a few comments here. First, since editors may join a wiki because of other editors (i.e., my friend is doing this, so I will too) or because of articles that inspire one to become an editor, a minimal model for mechanism should likely include articles and editors as states. In addition, we may also need to consider the behavioral ecology of editors and the 'fitness' metric involvement in a social network. Second, although the paper is about social networks, we assumed no influence of structure on wiki success. That is, we used the simplifying assumption that all editors are the same (have the same influence in generating new editors). But structure is a key characteristic of networks, and social networks often have hubs, where few individuals connect a large proportion of other less-connected individuals. As with invasion biology, in which all invaders are not equal, there are likely key highly connected editors that have more impact than others. We hypothesize that if editors are connected to their articles in a small-world network, then the predicted probability of success of the network would in general be less than the observed probability of success because most editors have little influence on the number of articles, which could in part explain the results in Figure 4.

Our computation of the probability of success for different Wikipedias has additional analogies to the field of Population Viability Analysis (PVA) in conservation biology, which has the goal of predicting extinction probabilities for endangered populations (Gilpin & Soule 1986, Beissinger & Westphal 1998). While PVAs have proven useful in ranking relative risks and evaluating alternative management strategies (Brook et al. 2000), confidence intervals on extinction probability ranging nearly from 0 to 1 are common (Ludwig 1990, Satterthwaite et al. 2002) and results are better interpreted as rankings of relative risks rather than absolute predictions of outcome probabilities (Beissinger & Westphal 1998).

The first major step in making a network tool for collaboration in an organization is to conduct a pilot with both a time frame and a long-term vision (Mader 2008) – this is like seeding with a propagule. The methods we have described allow quantitative evaluation of the probability that the collaborative social network will succeed on a larger scale, given the data of the pilot. For that reason, our approach is broadly applicable and is an example of how ideas from population biology inform understanding of technology (Crawford 2001, Wischmann et al. 2012).

Acknowledgements

This work was sponsored by the Defense Advanced Research Projects Agency Defense Sciences Office (DSO) Program: Thermodynamically Evolving Robust and Adaptive Physical Intelligence (THERA-PI) ARPA Order No. Z191/00, Program Code: 0D10 Issued by DARPA/CMO under Contract No. HR0011-10-C-052; the final stages were supported by NSF grant EF- 0924195 to M.M. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the Defense Advance Research Projects Agency or the U.S. Government. We appreciate the comments of Colin Clark, Christopher Friedrichs, and Donald Ludwig, Robin Snyder, and Justin Yeakel on previous versions of this work.

Literature Cited

Abramowitz M., Stegun IA. 1964. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* Superintendent of Documents, Washington(DC)

Beissinger S. , Westphal MI 1998. On the use of demographic models of population viability in endangered species management *J. Wildlife Manag*. 62: 821–841

Brook BW, O'Grady J, Chapman AP , Burgman MA , Akc¸akaya, HR, Frankham, R. . 2000. Predictive accuracy of population viability analysis in conservation biology. Nature 404:385–387.

Burnham KP, Anderson DR. 2002. *Model Selection and Multimodel Inference: A Practical-Theoretic Approach*, Springer Verlag, New York (NY)

Crawford M. 2011. Biologists using social-networking sites to boost collaboration *BioScience.* 61:736-736

Easley D, Klineberg J. 2011. *Networks, Crowds and Markets. Reasoning About a Highly Connected World.* Cambridge University Press, Cambridge (UK)

Eberscbach A, Glaser M, Heigl, R., Warta A. 2008. *Wiki Web Collaboration,* Springer Verlag, Heidelberg (Germany)

Elton C. 1958. *The Ecology of Invasions by Animals and Plants.* Methuen, London (UK)

Gelman A, Carlin JB, Stern HS, Rubin, DR. 2004. *Bayesian Data Analysis*, Chapman & Hall/CRC Texts in Statistical Science, Boca Raton (FL)

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bernardo JM, et al. editor, *Bayesian Statistics*, Clarendon Press, Oxford (UK), p. 169-193.

Gilpin M, Soule, ME. 1986. Minimum viable populations: processes of species extinction. In Soule ME, editor. *Conservation Biology: The Science of Scarcity and Diversity,* Sinauer Associates, Sunderland, MA (USA), p. 19–34.

Gladwell M 2000. *The Tipping Point. How Little Things Can Make a Big Difference,* Black Bay Books, Little, Brown and Company, New York (USA)

Harte D. 2011. "Non-asymptotic binomial confidence intervals" (Tech. Note, Statistics Research Associates, Wellington, NZ,; www.statsresearch.co.nz/pdf/confint.pdf).

Keizer JE 1987. *Statistical Thermodynamics of Nonequilibrium Processes* Springer Verlag, New York (USA)

Lewis K, Gonzalez M, Kaufman, J. 2012. Social selection and peer influence in an online social network. *Proc. Nat. Acad US* 109: 68-72.

Ludwig D. 1990. Is it meaningful to estimate a probability of extinction? *Ecology* 80: 298-310

MacArthur RH, Wilson, EO. 1967. *The Theory of Island Biogeography.* Princeton University Press, Princeton, NJ (USA)

Mader S. 2008. *Wikipatterns: a practical guide to improving productivity and collaboration in your organization.* Wiley, Indianapolis, IN (USA_

Mangel M, Ludwig D. 1977. Probability of extinction in a stochastic competition. *SIAM J. Appl. Math.* 33:256-266 .

Mangel M. 1994. Barrier transitions driven by fluctuations, with applications to ecology and evolution. *Theor. Pop. Biol* . 45:16-40

Mangel M. 2006. *The Theoretical Biologist's Toolbox. Quantitative Methods for Ecology and Evolutionary Biology.* Cambridge University Press, Cambridge (UK)

Micklos D, Lauter S, Nisselle A. 2011. Lessons from a science education portal *Science* 334:1657-1568

Newman, M. 2010. *Networks An Introduction.* Oxford University Press, New York (USA)

Plummer M, Best N, Cowles K, and Vines K. 2010. coda: Output analysis and diagnostics for MCMC. R package version 0.13-5. http://CRAN.R-project.org/package=coda

R Development Core Team 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Raftery AE, Lewis SM. 1995. The number of iterations, convergence diagnostics and generic Metropolis algorithms. In WR Gilks et al. editors. *Practical Markov Chain Monte Carlo,* Chapman and Hall, London (UK) p. 115-130.

Richardson DM, Pysek P. 2008. Fifty years of invasion ecology – the legacy of Charles Elton. Diversit Distribs. 14:161–168

Satterthwaite WH, Menges ES, Quintana-Ascencio P. 2002. Assessing scrub buckwheat population viability in relation to fire using multiple modeling techniques. *Ecol Applics* 12:1672–1687

Schuss Z. 1980. *Theory and Applications of Stochastic Differential Equations.* John Wiley and Sons, New York (USA)

Simberloff D. 2009. The role of propagule pressure in biological invasions. *Annu. Rev. Ecol. Evol. Syst.* 40:81-102

Stratonovich RL 1992. *Nonlinear Nonequilibrium Thermodynamics I. Linear and Nonlinear Fluctuation-Dissipation Theorems*, Springer-Verlag, New York (USA)

Wischmann S, Floreano D., and Keller L. 2012. Historical contingency affects signaling strategies and competitive abilities in evolving populations of simulated robots. *Proc. Nation. Acad. USA* 109: 864-868

Appendix 1. Further Details on the Methods

AIC Analysis for the Parameters

According to Equation (2), given $W(t) = w$,

$$dW \sim N((r_0 + r_1 w)dt + o(dt), (\sigma_0 + \sigma_1 w)^2 dt + o(dt)) \qquad (4)$$

where $N(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2, and *o(dt)* denotes terms that are higher powers of *dt*.

We used maximum likelihood techniques, assuming the normal distribution for *dW* specified in Equation (4), to determine the most likely values of the parameters in Equation (2) using the data shown in Figure 1c. We report these in Table 1. We assessed support for each of the different models using AIC analysis.

Derivation and Solution of the Equation for the Probability of Success

From the definition of *u(w),* the law of total probability (Mangel 2006) leads to

$$u(w) = E_{dW}\{u(w + dW)\} \qquad (5)$$

Taylor expanding the right hand side of Equation (5) to second order gives

$$u(w) = E_{dW}\{u(w) + \frac{du}{dw}dW + \frac{1}{2}\frac{d^2u}{dw^2}dW^2 + o(dW^2)\} \qquad (6)$$

In light of the properties of the increment of Brownian motion

$$\begin{aligned} E\{dW \mid W(t) = w\} &= (r_0 + r_1 w)dt + o(dt) \\ E\{dW^2 \mid W(t) = w\} &= (\sigma_0 + \sigma_1 w)^2 dt + o(dt) \end{aligned} \qquad (7)$$

so that Equation (6) becomes

$$u(w) = u(w) + (r_0 + r_1 w)\frac{du}{dw}dt + \frac{(\sigma_0 + \sigma_1 w)^2}{2}\frac{d^2u}{dw^2}dt + o(dt) \qquad (8)$$

We divide by dt and let $dt \to 0$ to obtain

$$\frac{(\sigma_0 + \sigma_1 w)^2}{2}\frac{d^2u}{dw^2} + (r_0 + r_1 w)\frac{du}{dw} = 0 \qquad (9)$$

In light of the definition of $u(w)$, the boundary conditions associated with Equation (9) are

$$u(0) = 0; u(w_u) = 1 \qquad (10)$$

The general solution of Equation (9) is

$$u(w) = \frac{\int_0^w \exp\left[-2\int_0^y \frac{r_0 + r_1 x}{(\sigma_0 + \sigma_1 x)^2}dx\right]dy}{\int_0^{w_u} \exp\left[-2\int_0^y \frac{r_0 + r_1 x}{(\sigma_0 + \sigma_1 x)^2}dx\right]dy} \qquad (11)$$

We evaluate this integral exactly for Model 1

$$u(w) = \frac{1 - e^{-\frac{2r_0}{\sigma_0^2}w}}{1 - e^{-\frac{2r_0}{\sigma_0^2}w_u}} \qquad (12)$$

For cases in which the MLE $r_0 = 0$, Equation (12) must be evaluated using l'Hospital's rule (or equivalently by Taylor expanding the exponentials as $r_0 \to 0$).

We evaluated the solution for Model 2

$$u(w)=\frac{\int_0^{w}\exp\left[\frac{1}{\sigma_1}\bullet\frac{2r_0}{\sigma_0+\sigma_1 y}\right]dy}{\int_0^{w_u}\exp\left[\frac{1}{\sigma_1}\bullet\frac{2r_0}{\sigma_0+\sigma_1 y}\right]dy} \tag{13}$$

using numerical integration.

The solution for Model 3

$$u(w)=\frac{\int_0^{w}\exp\left[-\frac{r_1}{\sigma_0^2}y^2\right]dy}{\int_0^{w_u}\exp\left[-\frac{r_1}{\sigma_0^2}y^2\right]dy}=\frac{\int_0^{w\frac{\sqrt{2r_1}}{\sigma_0}}\exp\left[-\frac{y^2}{2}\right]dy}{\int_0^{w_u\frac{\sqrt{2r_1}}{\sigma_0}}\exp\left[-\frac{y^2}{2}\right]dy} \tag{14}$$

can be evaluated quickly by using the numerical approximation for normal distributions (Abramowitz & Stegun 1964).

We also evaluated the solution for Model 4

$$u(w)=\frac{\int_0^{w}\exp\left[-\frac{2}{\sigma_0^2}(r_0 y+\frac{1}{2}r_1 y^2)\right]dy}{\int_0^{w_u}\exp\left[-\frac{2}{\sigma_0^2}(r_0 y+\frac{1}{2}r_1 y^2)\right]dy} \tag{15}$$

numerically.

Finally, for Model 5, no simplifications are possible for the integrands in the exponentials , so we used Eqns 3.3.16-3.1.19 from Abramowitz & Stegun (1964) for evaluation of integrals of the form $\int \frac{mx+n}{ax^2+bx+c}dx$.

Construction of Fig. 2a

To construct Figure 2a, we let subscript *i* denote the result obtained from Equation (11) when model *i* in Table 1 is used (with weights rescaled to sum to 1.0 for the top five models) and define

$$\begin{aligned} \overline{u}(w) &= \sum_{i=1}^{5} AIC_w(i) u_i(w) \\ \overline{u}^2(w) &= \sum_{i=1}^{5} AIC_w(i) u_i^{2V}(w) \\ Var(u(w)) &= \overline{u}^2(w) - (\overline{u}(w))^2 \end{aligned} \qquad (16)$$

to represent the AIC weighted solutions for a mean of Equation (11) and a measure of variability around that mean. For Figure 2a, we show the weighted mean with error bars determined by $\sqrt{Var(u(w))}$.

Bayesian Methods

Our goal is to estimate the joint posterior for the model parameters, which we denote by a vector *v* whose form will vary based on the model under consideration. The posterior predictive distribution for the success probability *u(w)* can then be determined by applying the formula for *u(w)* to draws of model parameters from their joint posterior distribution.

The maximum likelihood fits described earlier allow us to calculate the probability distribution of dW given the parameters and the current number of editors, which we denote by $p(dW|\nu,W)$. To generate the posterior for the model parameters we compute the probability distribution of the parameters, given the current number of editors and the observed change in editors, which we denote by $p(\nu \mid W, dW)$.

Using Bayes's theorem and assuming independent priors on W, dW, and ν, the posterior distribution for the parameters $p_1(\nu \mid W, dW)$ given the current number of editors and the change in the number of editors is

$$p_1(\nu \mid W, dW) = \frac{p_0(\nu) p(dW \mid W, \nu)}{\int p_0(\nu) p(dW \mid W, \nu) d\nu} \propto p_0(\nu) p(dW \mid W, \nu) \qquad (17)$$

In this expression, $p_0(\nu)$ is the prior probability for the parameters (we assume limited knowledge and thus uniform, unbounded priors except that $\sigma_0 > 0$ since the standard deviation in growth rate must be positive even as W becomes small) and $p(dW, W \mid \nu)$ is of the probability of drawing each observed value of dW from a normal distribution with mean and standard deviation calculated according to the current model formulation, given W and the proposed values of ν. The denominator in the middle expression in Equation (17) is a normalization constant, so that the posterior density integrates to 1. The final expression follows from this.

We approximated the posterior distribution for *v* by using Metropolis algorithm Markov-Chain Monte Carlo sampling of the equation for the posterior density given above (Gelman et al. 2004). We used custom R code (R Development Core Team) with symmetric, normal proposal distributions manually tuned to give acceptance probabilities of ~ 0.4. Within each iteration of the chain, we updated each element of *v* one parameter at a time, conditioned on the current value of the other parameters. After (potentially) updating each element of *v*, we calculated *u(w)* for a series of values of *w* ranging from 2.5 to 50 in increments of 2.5 and from 50 to 500 in increments of 25 (and then up to 1000 in increments of 100 for English). We recorded each of these estimates of *u(w)* to generate chains of draws from the posterior predictive distribution.

The priors we used are improper (do not have finite integrals), but our data had a bounded range of support (and sufficient power for likelihoods to dominate priors), resulting in proper posteriors. We also ran the models with a Jeffreys prior (Gelman et al. 2004) for σ_0, $p(\sigma_0) \sim 1/\sigma_0$ for all $\sigma_0 > 0$, with essentially identical results. Since the posterior distributions of individual variables integrate over the full uncertainty in the remaining parameters, their modes do not always coincide exactly with the maximum likelihood estimates.

After manual tuning of the proposal distributions, for each model formulation (i.e., Models 1-5) or dataset (we extended this analysis to multiple languages, see below) we ran an initial burn-in of 20,000 iterations, starting from the MLE estimates of *v*. We then used the diagnostic method of Geweke (1992) as implemented in "gewke.diag" in R library "coda" (Plummer et al. 2010) to compare the mean of the first 10% of the retained chain to the last 50% of the chain. If |z|-scores for all elements of *v* and all estimates of *u(w)* were < 2, we considered the chains to have converged and burn-in adequate. If not, we ran an additional 20,000 iterations until the last 20,000 iterations yielded passing diagnostics. Finally, we used the method of Raftery & Lewis (1995) approach implemented via the command "raftery.diag" in library "coda" to calculate the recommended post burn-in length of the posterior chain for each parameter or predictive output required to estimate the .025 quantile with accuracy of 0.005 with 95% probability. We rounded the required chain length up to the nearest 20,000 (if ≤100,000 iterations required) or 250,000.

Captions for Figures

Figure 1. a) When viewed backwards across about a decade, the growth of the English language Wikipedia seems inevitable, rising exponentially then saturating. b) Early on, however, the number of active editors shows growth but considerable fluctuations. c) When those data are converted to monthly change as a function of the current number of editors, large fluctuations can occur even when the number of editors is considerable.

Figure 2. a) The weighted mean probability of the English Wikipedia hitting 1000 editors before returning to 0 editors given the current number of editors, for the 5 top models in Table 1, with AIC weights rescaled to sum to 1. The error bars are determined by the square root of the AIC weighted variance (**Methods**). b) The individual solutions, along with their AIC weights.

Figure 3. Posterior distributions of parameters for the top five models for early dynamics of editors. For the plots of $u(w)$, the solid line is the posterior median, the heavily dashed lines are the 68% credible interval, and the lightly dashed lines are the 95% credible interval. Panels a-e) correspond to models 1-5.

Figure 4. Cross-language success in predicting the probability of different language Wikipedias succeeding. We bin languages a) by MLE predicted success, and plot the proportion of models with predicted success in each bin that actually succeeded (error bars are 68% binomial confidence intervals (*9*)) or b) based on the lower bound of the 68% credible interval on success probability.

Figure 1a

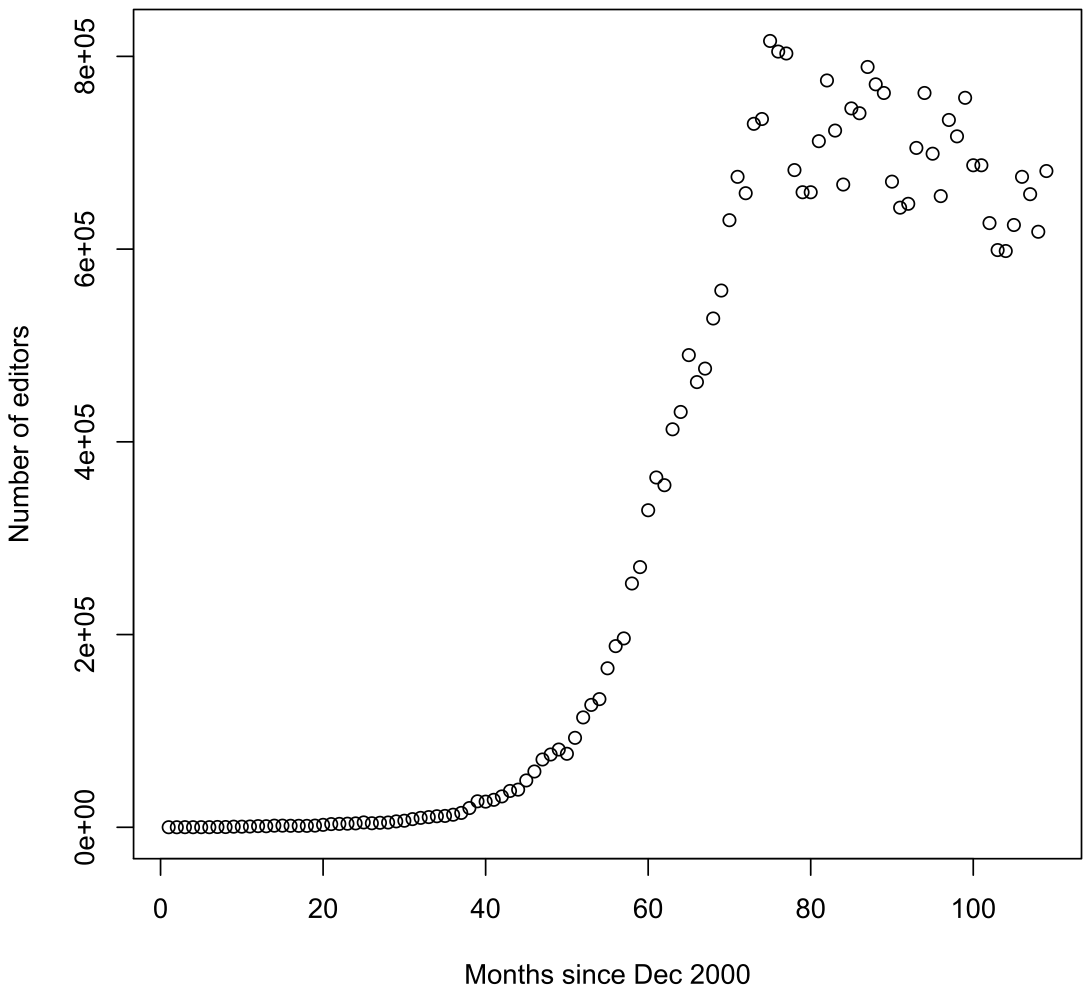
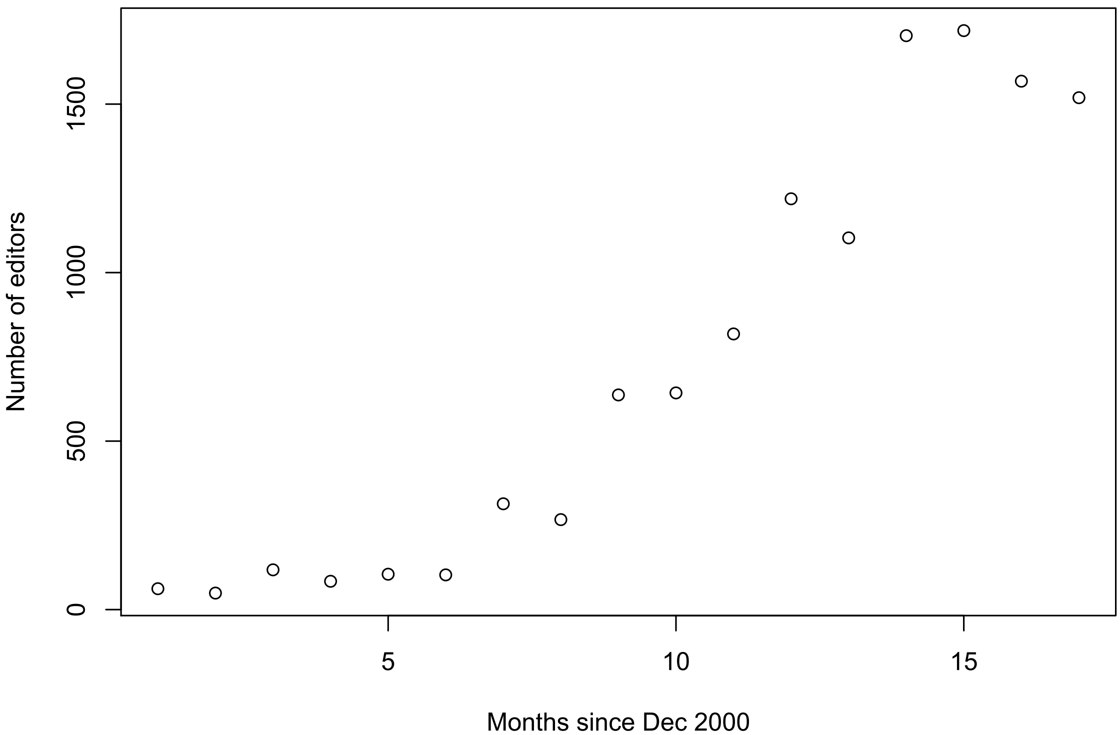

Number of editors
0e+00
2e+05
4e+05
6e+05
8e+05
0
20
40
60
80
100
Months since Dec 2000

Figure 1b



0
500
1000
1500
Number of editors
5
10
15
Months since Dec 2000

Figure 1c

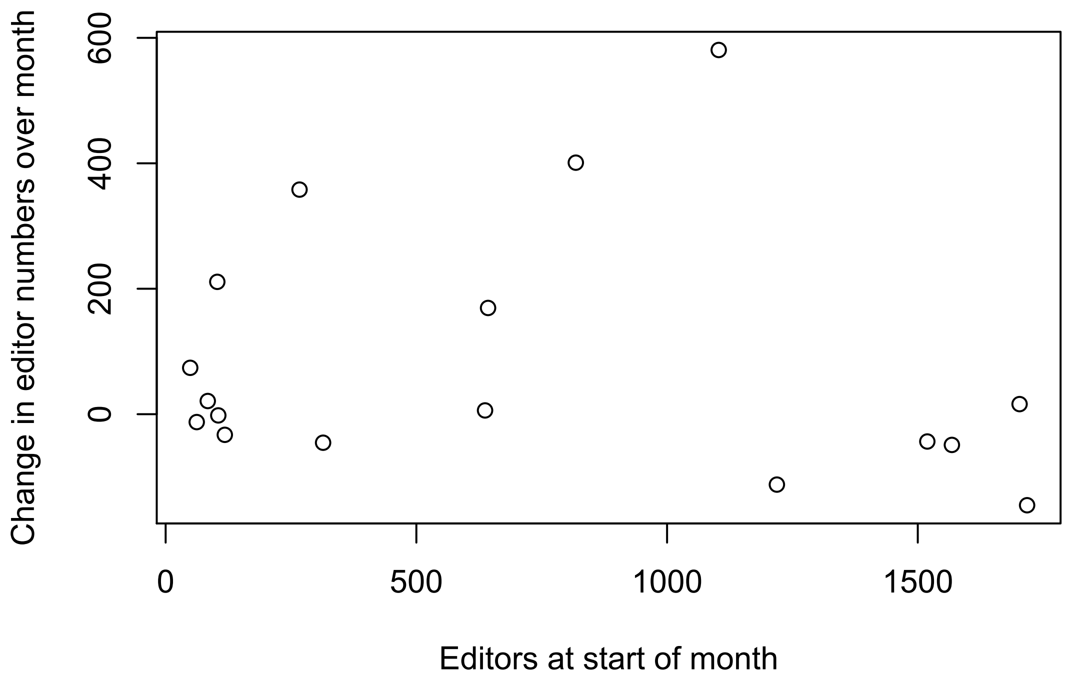

Change in editor numbers over month
600
400
200
0
0
500
1000
1500
Editors at start of month

Figure 2a

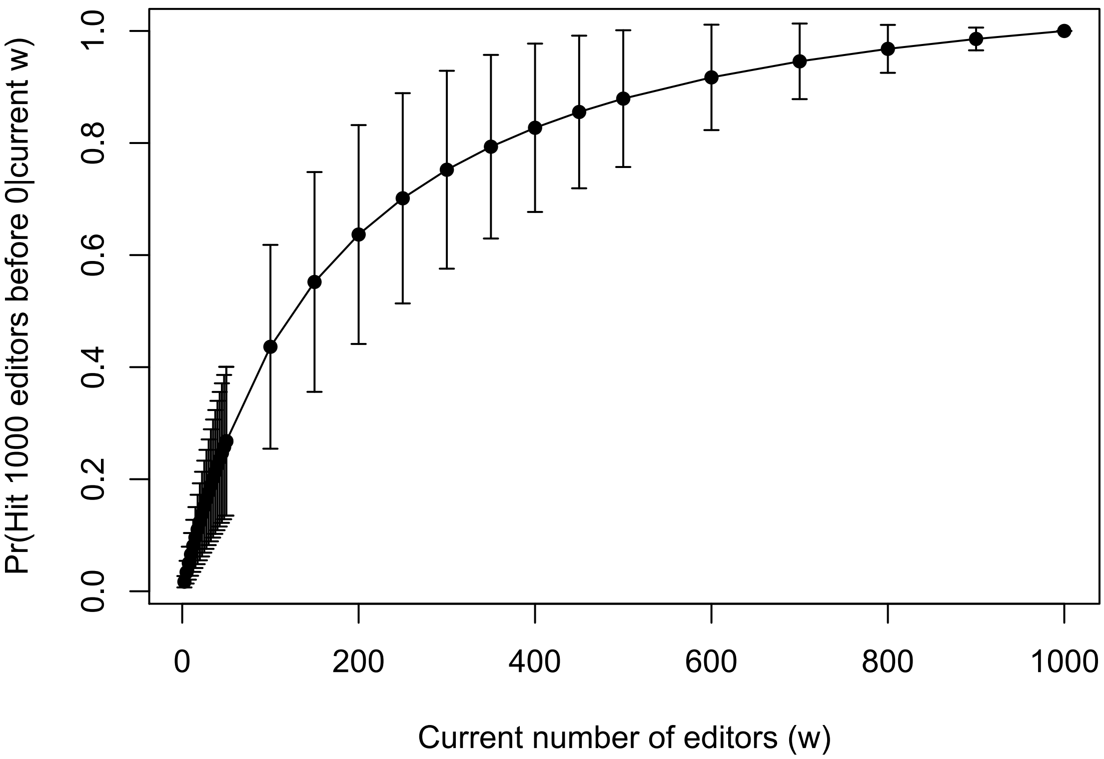

Pr(Hit 1000 editors before 0|current w)
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
Current number of editors (w)

Figure 2b

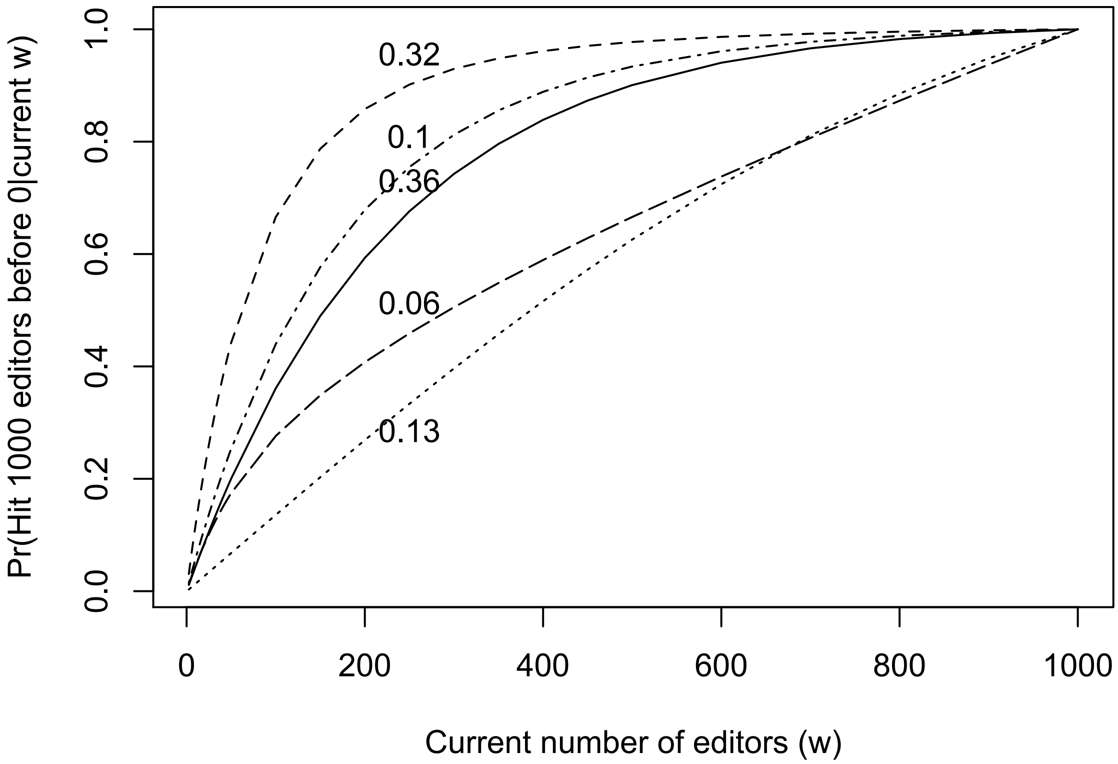

0.32
0.1
0.36
0.06
0.13
Pr(Hit 1000 editors before 0|current w)
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
Current number of editors (w)

Figure 3a

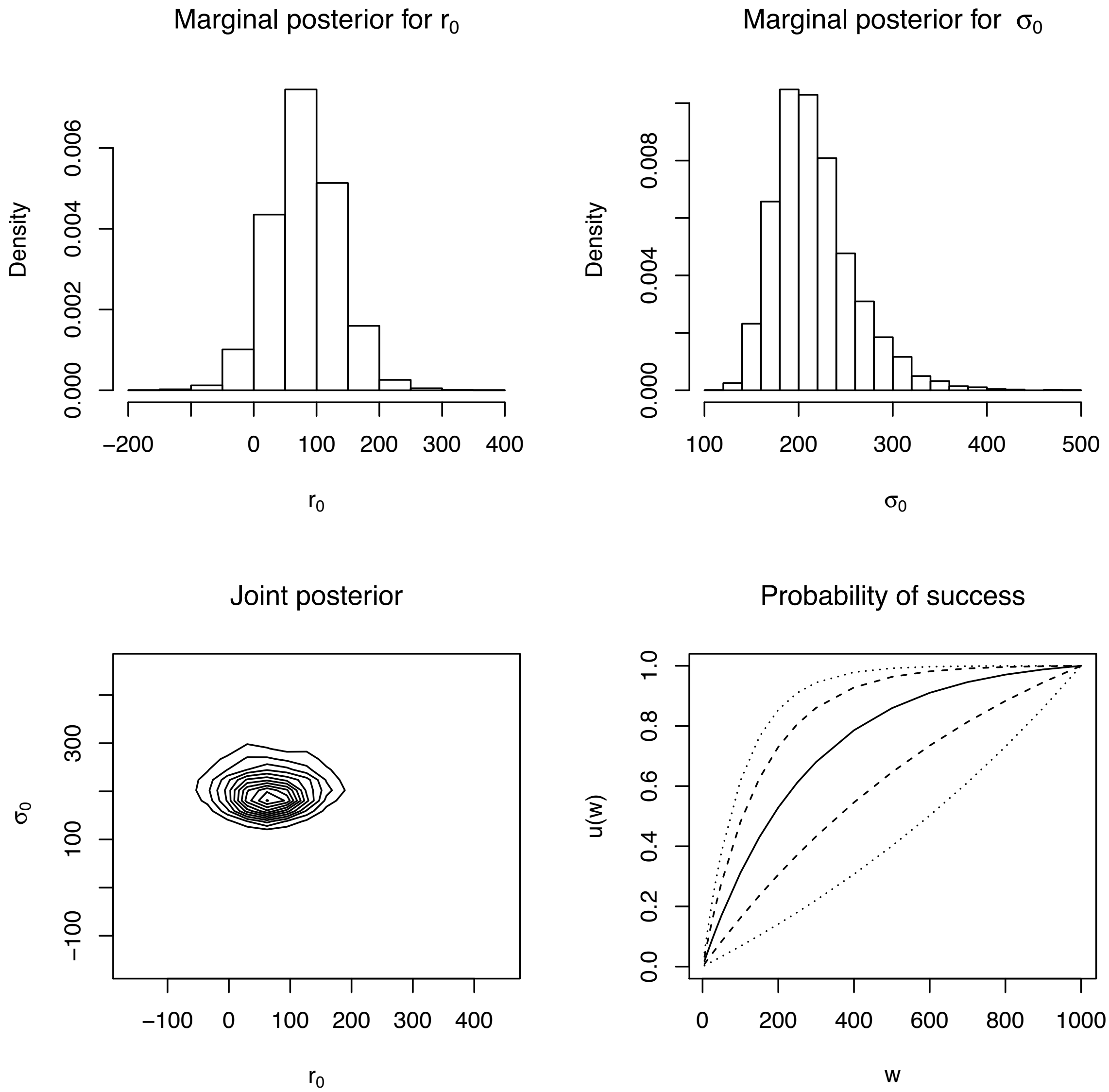
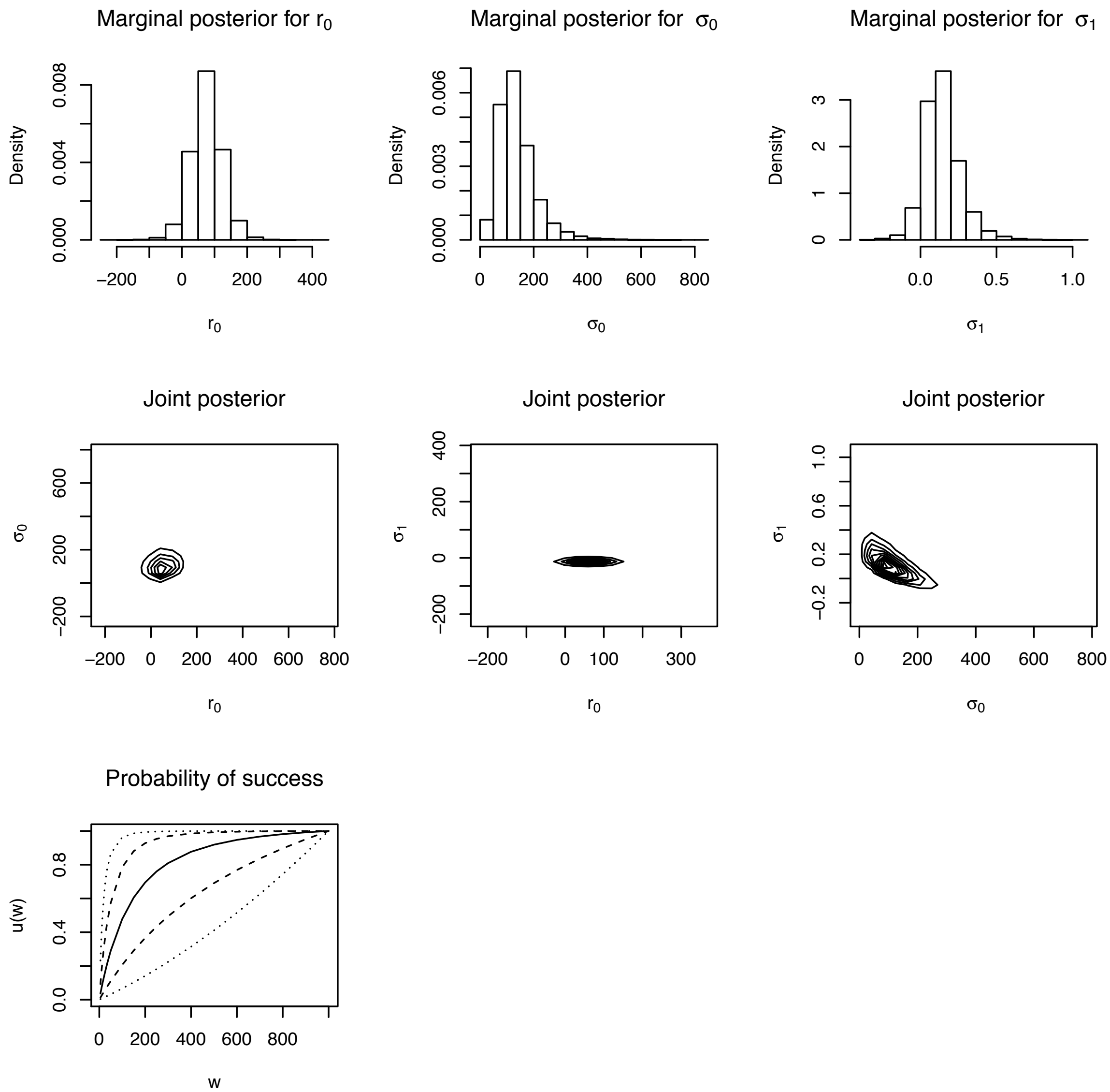

Marginal posterior for r_0
Density
0.000
0.002
0.004
0.006
−200
0
100
200
300
400
r_0
Marginal posterior for σ_0
Density
0.000
0.004
0.008
100
200
300
400
500
σ_0
Joint posterior
σ_0
−100
100
300
−100
0
100
200
300
400
r_0
Probability of success
u(w)
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
w

Figure 3b



Marginal posterior for r_0
Marginal posterior for σ_0
Marginal posterior for σ_1
Density
r_0
σ_0
σ_1
Joint posterior
Joint posterior
Joint posterior
Probability of success
u(w)
w

Figure 3c

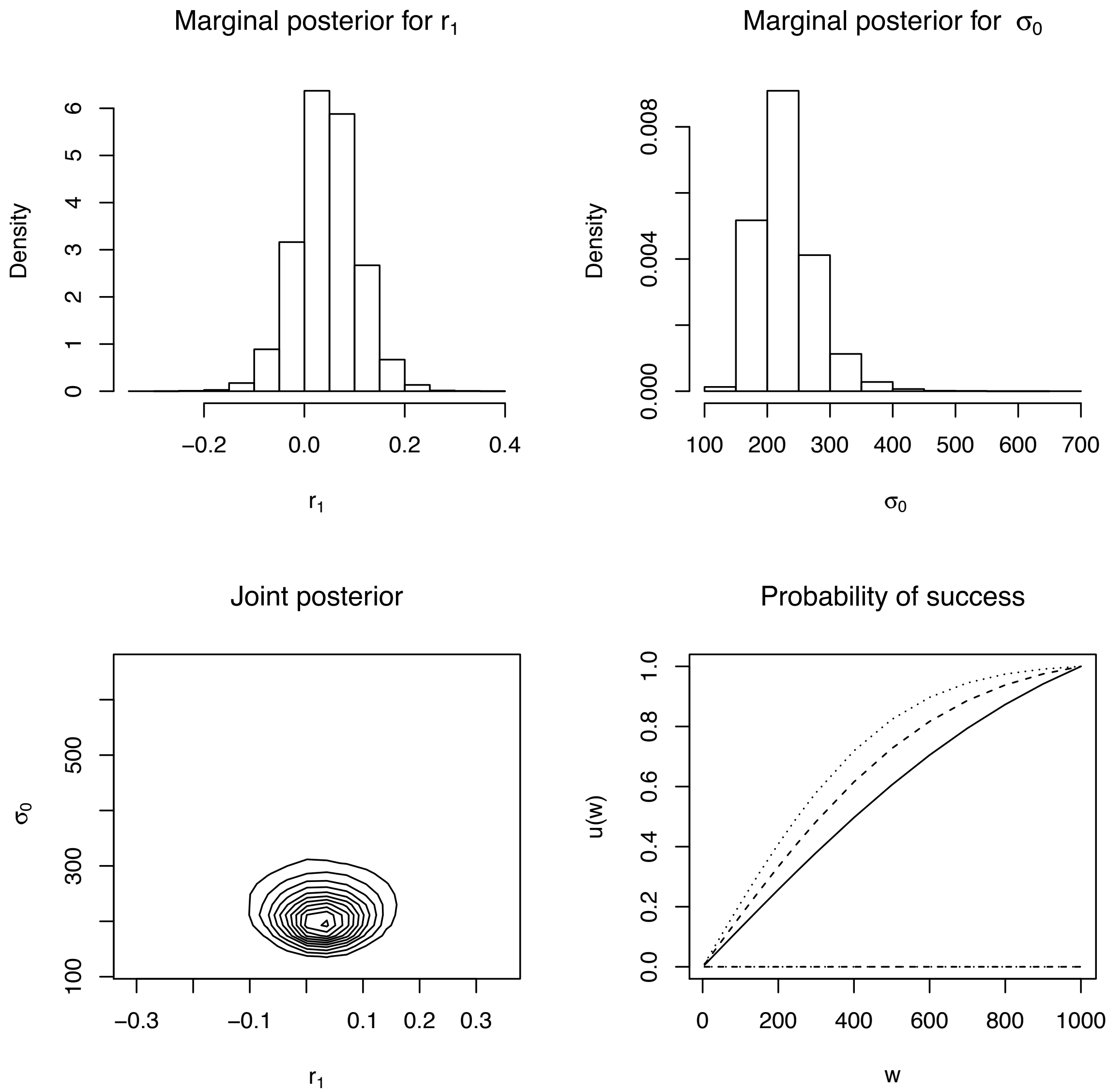

Marginal posterior for r1
Density
r1
Marginal posterior for σ0
Density
σ0
Joint posterior
σ0
r1
Probability of success
u(w)
w

Figure 3d

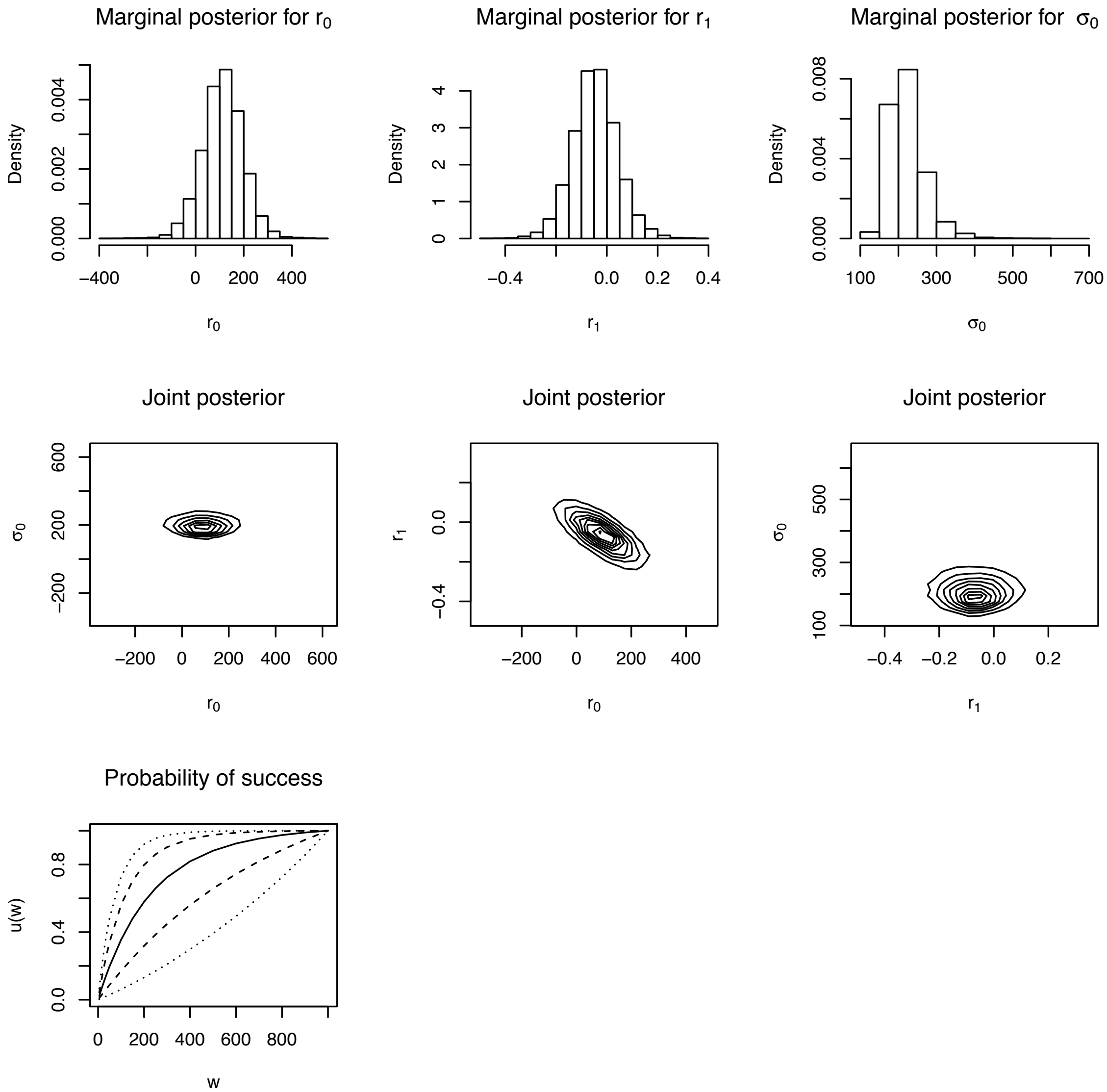

Marginal posterior for r_0
Density
r_0
Marginal posterior for r_1
Density
r_1
Marginal posterior for σ_0
Density
σ_0
Joint posterior
σ_0
r_0
Joint posterior
r_1
r_0
Joint posterior
σ_0
r_1
Probability of success
u(w)
w

Figure 3e

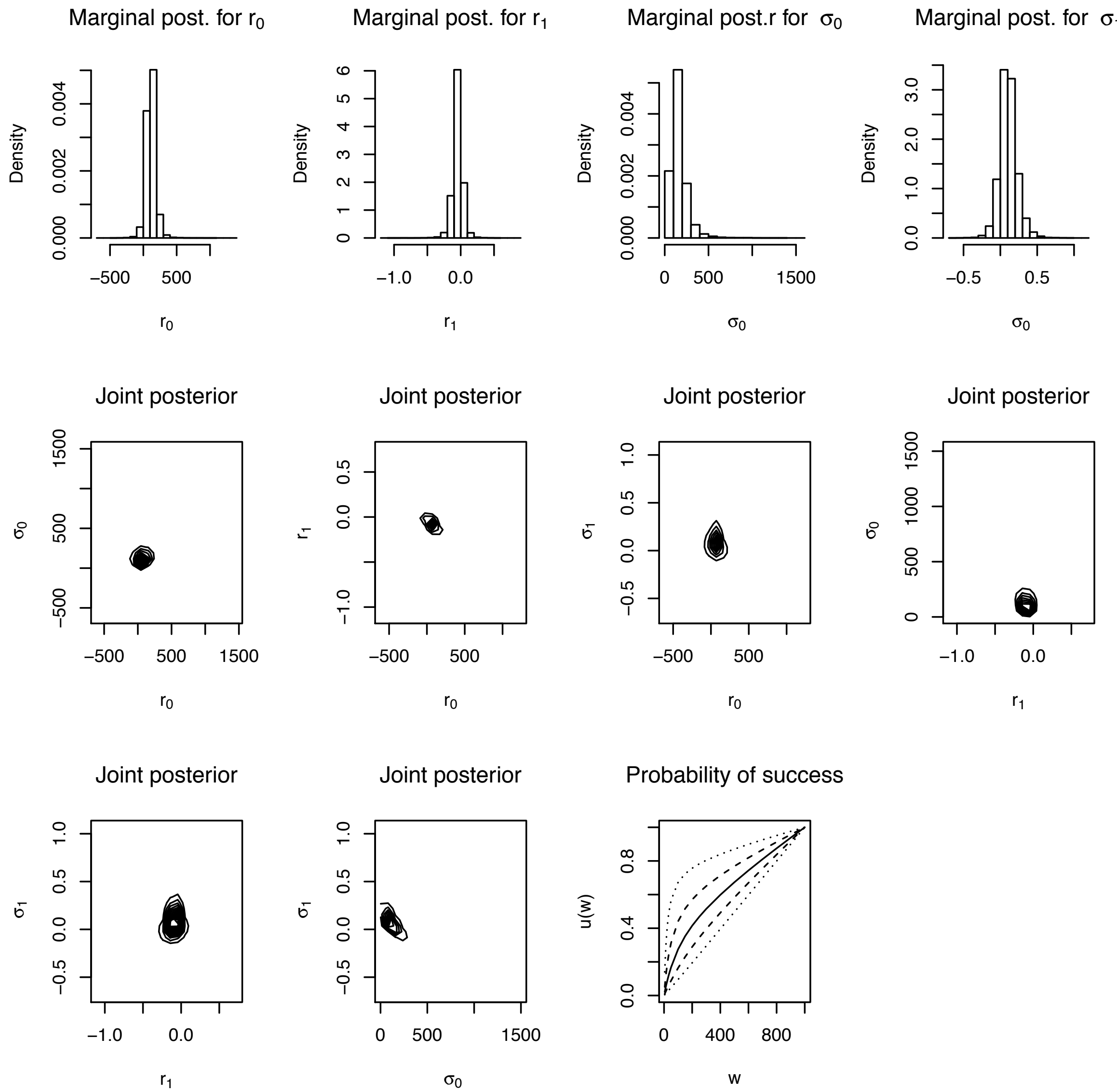

Marginal post. for r_0
Marginal post. for r_1
Marginal post.r for σ_0
Marginal post. for σ
Density
r_0
r_1
σ_0
Joint posterior
σ_0
r_1
σ_1
Probability of success
u(w)
w

Figure 4a

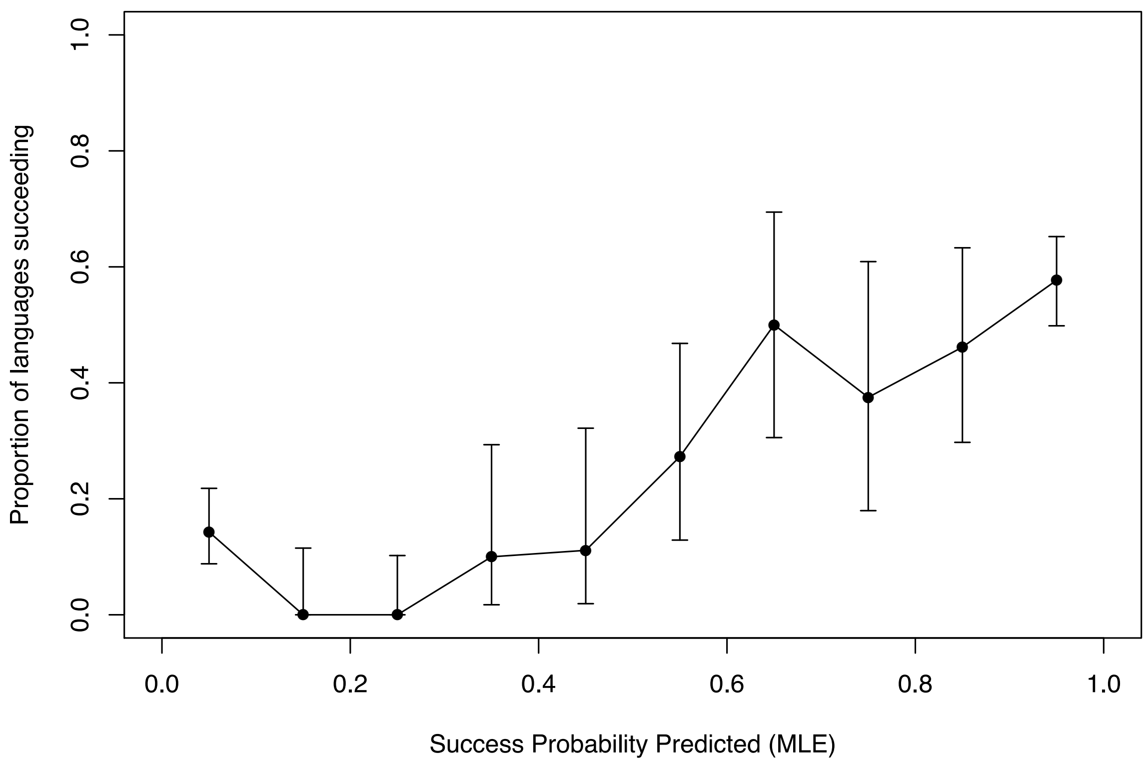

Proportion of languages succeeding
0.0
0.2
0.4
0.6
0.8
1.0
Success Probability Predicted (MLE)

Figure 4b

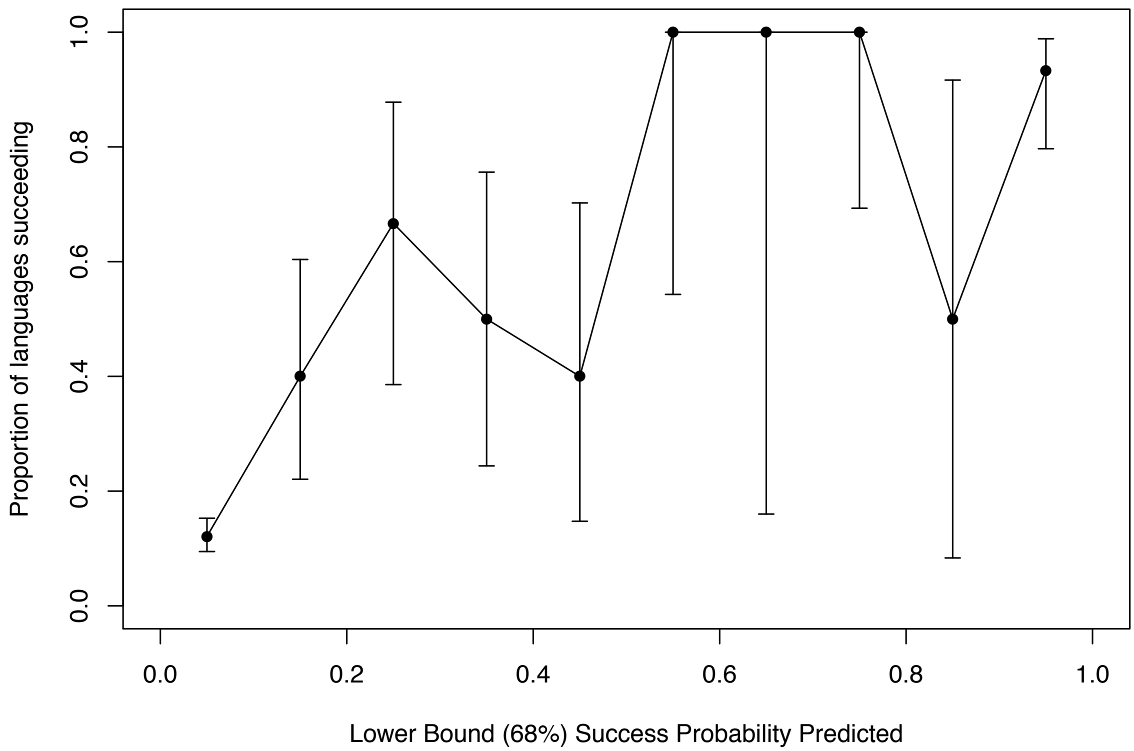

Proportion of languages succeeding
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Lower Bound (68%) Success Probability Predicted

Table 1. Results of the AIC Analysis for the data shown in Figure 1c under different assumptions about the stochastic dynamics* (k=number of parameters)

Model	k	AICc	AIC Weight ($AIC_w(i)$)
1) $r=r_0$, $\sigma=\sigma_0$	2	232.01	0.36
2) $r=r_0$, $\sigma=\sigma_0+\sigma_1 w$	3	232.22	0.32
3) $r=r_1 w$, $\sigma= \sigma_0$	2	234.12	0.12
4) $r= r_0+ r_1 w$, $\sigma= \sigma_0$	3	234.62	0.098
5) $r= r_0+ r_1 w$, $\sigma= \sigma_0+\sigma_1 w$	4	235.57	0.061
6) $r= r_1 w$, $\sigma= \sigma_0 + \sigma_1 w$	3	238.73	0.012
7) $r= r_0$, $\sigma= \sigma_1 w$	2	239.00	0.011
8) $r= r_1 w$, $\sigma= \sigma_1 w$	2	240.95	0.004
9) $r= r_0+ r_1 w$, $\sigma= \sigma_1 w$	3	241.64	0.0029

*Given $W(t) = w$ and setting $dW = W(t+dt) - W(t)$

$$dW \sim N((r_0 + r_1 w)dt + o(dt), (\sigma_0 + \sigma_1 w)^2 dt + o(dt))$$

where $N(\mu,\sigma^2)$ is a normal distribution with mean μ and variance

Table 2 Non-zero MLE values of parameters for the five best-supported models from Table 1

Model	Parameters
1	r_0=81.99, σ_0=192.86
2	r_0=78.79, σ_0=101.30, σ_1=0.13
3	r_1=0.0453, σ_0=205.20
4	r_0=115.04, r_1=-0.0467, σ_0=190.74
5	r_0=66.70, r_1=0.0361, σ_0=95.23, σ_1=0.14