CHAPTER 5

Quantifying diffusion in social networks: a Bayesian approach

Glenna Nightingale, Neeltje J. Boogert, Kevin N. Laland, and Will Hoppitt

Introduction to social transmission in groups of animals

Social learning is often broadly defined as 'learning that is influenced by observation of, or interaction with, another animal (typically a conspecific) or its products' (Heyes 1994). However, the phrase 'influenced by' is unacceptably vague, and we prefer to characterize social learning as learning that is *facilitated by* observation of, or interaction with, another individual, or its products (Hoppitt and Laland 2013). In addition, we prefer the more specific term 'social transmission' to refer to the process by which behavioural traits spread through groups. We define social transmission as occurring when 'the prior acquisition of a behavioural trait *T* by one individual, *A*, when expressed either directly in the performance of *T* or in some other behaviour associated with *T*, exerts a lasting positive causal influence on the rate at which another individual, *B*, acquires and/or performs the behavioural trait' (Hoppitt and Laland 2013).

The study of social learning was initially motivated by an interest in the cognitive or psychological mechanisms underpinning social learning (e.g. Galef 1988; Heyes 1994), leading to research conducted in controlled laboratory conditions (e.g. Zentall et al. 1996). In recent years, the focus has shifted to animal traditions and culture, driven by the discovery of group-specific behaviour in a number of taxa, including primates (e.g. Whiten et al 1999; S. Perry et al. 2003), cetaceans (e.g. Rendell and Whitehead 2001) and birds (e.g. Madden 2008). Such group-specific behaviour patterns often appear to be the result of different behavioural innovations spreading through groups by social transmission. Many researchers are now studying the conditions under which novel behavioural traits spread and form traditions in the field or in a captive group context, in which the subjects are free to interact with one another (e.g. meerkats (*Suricata suricatta*) (Thornton and Malapert 2009); vervet monkeys (*Chlorocebus pygerythrus*) (Van de Waal et al. 2010); and humpback whales (*Megaptera novaeangliae*) (Allen et al. 2013)). Such research has motivated the development of novel methods for studying social learning in freely interacting groups (Laland and Galef 2009; Kendal, Galef et al. 2010; Hoppitt and Laland 2013). Laboratory experiments provide valuable insights into learning mechanisms, and other aspects of social interaction, but it remains challenging in a natural or field context to ascertain whether social learning has occurred and quantify its impact without the use of sophisticated statistical methods.

In our use of the terminology, 'social transmission' can be distinguished from 'diffusion', as the latter term refers to the observed spread of a trait through a group, irrespective of the cause of the spread. Therefore, a trait might be said to have diffused through a group without any evidence that this occurred by social transmission. For instance, the diffusion may result from independent asocial learning by each individual, or there may be an unlearned social influence on behaviour, as occurs, for instance, when animals influence each other's movements, as is reported in sticklebacks (Atton et al. 2012).

Animal Social Networks. Edited by Jens Krause, Richard James, Daniel W. Franks and Darren P. Croft.

A promising alternative approach to highly structured laboratory experimentation for inferring and quantifying social transmission in naturalistic group contexts is network-based diffusion analysis (also known as NBDA) (Franz and Nunn 2009). Network-based diffusion analysis infers social transmission when the time (or order (Hoppitt, Boogert, et al. 2010)) of acquisition of the trait by individuals in animal groups follows a social network. Similar models have been used in the social sciences (Valente 2005); their relationship with network-based diffusion analysis is discussed in Hoppitt and Laland (2013). In this chapter, we will first briefly review the uses of network-based diffusion analysis and then explain why a Bayesian formulation is required. We will then present our Bayesian formulation of network-based diffusion analysis and test it using simulated data.

Network-based diffusion analysis

Network-based diffusion analysis infers and quantifies the magnitude of social transmission in a set of diffusion data from the extent to which the pattern of spread follows a social network. There are two versions: time of acquisition diffusion analysis (also known as TADA) (Franz and Nunn 2009), which takes as data the times at which individuals acquired the target behavioural trait, and order of acquisition diffusion analysis (also known as OADA) (Hoppitt, Boogert, et al. 2010), which is sensitive only to the order in which they do so. The former is more powerful but makes stronger assumptions. The greater power of time of acquisition diffusion analysis stems from the fact that order of acquisition diffusion analysis is only sensitive to social transmission if it results in a difference in the relative rate of acquisition by individuals, whereas time of acquisition diffusion analysis is also sensitive to absolute changes in the rate of acquisition and thereby has more data with which to detect social transmission (Hoppitt and Laland 2011).

Despite being a recently developed method, network-based diffusion analysis has already been used a number of times to analyse diffusion data from wild and captive animal populations, usually using an association metric to obtain the social network. For example, Aplin et al. (2012) found strong evidence that the time of and probability of discovering of novel food patches followed an association network in a wild population of great tits (*Parus major*), blue tits (*Cyanistes caeruleus*), and marsh tits (*Poecile palustris*). Likewise, Allen et al. (2013) found strong evidence that the acquisition of lobtailing, a foraging innovation, followed an association network in a wild population of humpback whales. Kendal, Custance, et al. (2010) also applied the method to analyse the diffusion of a novel foraging behaviour in lemurs (*Lemur catta*), although there was no evidence that social transmission followed the network in that case.

In other cases, network-based diffusion analysis has been applied to diffusion data arising when captive groups of animals are presented with a novel foraging task. For example, Boogert et al. (2008) constructed a nearest neighbour association network for three groups of five starlings (*Sturnus vulgaris*) each and then separately presented each group with six tasks. Hoppitt, Boogert, et al. (2010) applied a continuous time of acquisition diffusion analysis to these data and found strong evidence of social transmission, although further analysis (Hoppitt, unpublished data) found that there was little evidence that it followed the social network. Specifically, a model with social transmission following the association network had more support than a model without social transmission, but it had similar support to a model that had homogeneous connections between individuals in each group. Under such circumstances, the evidence for social transmission comes from the observation that individuals were more likely and/or faster to solve the task once other individuals in the group had solved it (Hoppitt and Laland 2013). One drawback of many such captive studies is that animals may have little opportunity to avoid each other, as a result of which network-based diffusions may be less like to be detected, either because all network associations are strong or because network connections become superfluous in confined spaces, where learning between poorly connected individuals is feasible.

Network-based diffusion analysis has also been used to analyse diffusion experiments on groups of captive sticklebacks (Atton et al. 2012; M. Webster et al. 2013). Atton et al. (2012) expanded network-based diffusion analysis methodology by recognizing

that individuals can move between multiple states. For example, rather than just moving from 'naïve' (not solved the task) to 'informed' (solved the task) states, individuals move from being 'naïve', to having 'discovered' the task, to 'solving' the task. In principle, social influences might operate on both the discovery and the solving transition, and Atton et al. (2012) found that the social network affected each transition in a different way. Atton et al. (2012) also expanded network-based diffusion analysis to allow for multiple options available to solve the task (cf. Kendal et al. 2010).

While network-based diffusion analysis has, thus far, been used to assess whether the pattern of diffusion follows a measured association network, it could instead be used to compare the support for different hypothesized pathways of diffusion (Franz and Nunn 2009; Hoppitt and Laland 2011). For instance, one could test networks corresponding to different theories of 'directed social learning' (Coussi-Korbel and Fragaszy 1995). For example, the hypothesis that all social transmission is vertical would correspond to an asymmetrical binary network in which all connections lead from parents to offspring (Hoppitt and Laland 2013).

Why do we need Bayesian network-based diffusion analysis?

Network-based diffusion analysis can be expanded such that it quantifies the evidence for social transmission across a number of diffusions (Hoppitt et al. 2010), although care must be taken in the interpretation of such models (Hoppitt and Laland 2011, 2013). This expansion provides a valuable way of combining information arising from diffusions across different groups of animals (e.g. Webster et al. 2013). The expansion of network-based diffusion analysis to multiple diffusions could also be valuable where researchers have repeated diffusions across the same group, or groups, of animals (e.g. Boogert et al. 2008), especially when they only have a limited number of animals, allowing them to obtain good statistical power.

However, a statistical problem arises if they fail to account for the fact that the same individuals are involved in multiple diffusions. To illustrate this, imagine a group of experimental subjects presented with a foraging task and who vary in their asocial learning ability in a way that seems to imply social transmission occurs. For example, the best asocial learner may be well connected to the second-best asocial learner, making it appear that the latter is learning from the former, when in fact they are solving the task independently (i.e. through asocial learning processes). For a single diffusion, this chance possibility is automatically accounted for when assessing the evidence for social transmission.

If multiple foraging tasks are presented to different groups of individuals, and a pattern in the resulting diffusion data consistent with social transmission arises each time, then this adds to the evidence that social transmission is occurring; it is unlikely that a chance pattern of asocial learning abilities, consistent with social transmission, occurs in all cases (although confounding variables are possible) (see Hoppitt, Boogert, et al. 2010 for discussion). If *multiple* diffusions are run on the same individuals, and a similar pattern arises, this too will be taken by the model as being strong evidence of social transmission. However, in this case if a chance pattern of asocial learning ability consistent with social transmission happens to arise over the single group of individuals, it is repeated over multiple diffusions. If a model without random effects is fitted, each diffusion will be unrealistically taken as an independent set of data supporting the hypothesis that social transmission is occurring, thus potentially resulting in a spurious result. In a similar way, a chance pattern in asocial learning abilities counteracting the effects of social transmission would lead us to underestimate the effects of social learning, with an overinflated level of certainty.

By including an individual random effect on the asocial rate of learning, the model accounts for the fact that the same individuals have the same (or similar) asocial learning ability in each diffusion. However, random effects can be difficult to implement using maximum likelihood methods (used to fit network-based diffusion analysis models thus far), especially when the random effects structure is complex, because one has to integrate the likelihood function across all the possible values the random effects could take. It is easier to include random effects in a Bayesian model, using Markov chain Monte Carlo methods (Gelman et al. 2004).

Bayesian methods take a joint prior distribution for the model parameters, quantifying researchers' knowledge about the plausible values those parameters could take before receiving the data, and update this in light of the data to yield a joint posterior distribution. The joint posterior distribution thus quantifies the state of knowledge arising from the data, showing which combinations of parameter values are plausible. The marginal posterior distribution for a parameter (often shortened to 'posterior distribution') is the joint posterior distribution integrated over all the possible values of the other parameters in the model (including all the possible values for each level of each random effect). For relatively simple models, a mathematical expression can be obtained for the exact posterior distribution for each parameter of interest. However, for more complex models, such as those containing random effects, this is not possible, and Markov chain Monte Carlo is used.

Markov chain Monte Carlo is a procedure that simulates drawing values for the model parameters from the joint posterior distribution for all parameters in the model. By drawing a large number of values for one parameter from the joint posterior distribution, one is automatically accounting for uncertainty in the other parameters in the model. Consequently, when using Markov chain Monte Carlo, instead of integrating the likelihood over the random effects numerically, inferences on a given random effect parameter can be made by simulating draws from the marginal posterior distribution of this parameter. For a comprehensive explanation of the use of Markov chain Monte Carlo for random effects models, we refer the reader to Hoff's (2009) book on Bayesian statistics. Here, we develop a Bayesian version of time of acquisition diffusion analysis and then test whether it solves the problems outlined above, using simulated data. To aid explanation of the model formulation, we describe the simulated data first.

Simulated diffusion data

The data were simulated from the time of acquisition diffusion analysis model using the Gillespie algorithm (D. Wilkinson 2012). The social network used to simulate the data consisted of interactions of equal magnitude, set at 1. This allows us to show that the extracted parameter estimates closely match the real ones in a simple case. However, we stress that the approach works effectively for more realistic social networks. Bayesian network-based diffusion analyses for networks with more complex structures can be found in the tutorial posted on the Laland lab website (<http://lalandlab.st-andrews.ac.uk/>).

We simulate data corresponding to ten different tasks performed by the same group of ten individuals. We use a fixed value for the social effect (set at 0.6) and a fixed value for the baseline rate of asocial learning (0.3). The output from each data simulation includes both latency-to-solve times and solving order for the group of ten individuals. Random effects for each individual i were also incorporated in the simulation by assigning a number R_i from a set of numbers with variance 9, and multiplying its rate of asocial learning by a factor of exp (R_i). The incorporation of the random effects at the individual level thus models the heterogeneity arising from variation in asocial learning ability across the ten individuals. For ease of reference, each individual was given a unique name (Ned, Ted, Ron, Wim, Jim, Sue, Fay, Lou, Joe, and May).

Figure 5.1a shows the solve times for the ten diffusions, and Figure 5.1b shows the solve times for the first four tasks, with each task represented by a unique plotting symbol. Finally, Figure 5.1c shows how performance across the tasks varies among individuals. From Figure 5.1c, it is evident that the variation in solve times is highest for Ned, Ted, Jim, Ron, and Wim. For Fay, Lou, Joe, and especially May, the latencies to solve the tasks are much shorter than those of the other individuals, and the variation in times is low (the points on the plot have merged into one).

Previous formulation of time of acquisition diffusion analysis

The time of acquisition diffusion analysis model is based on standard survival models using an exponential distribution. We therefore use survival analysis terminology, referring to the 'hazard function' as giving the instantaneous rate at which an individual acquires the target trait, which in this

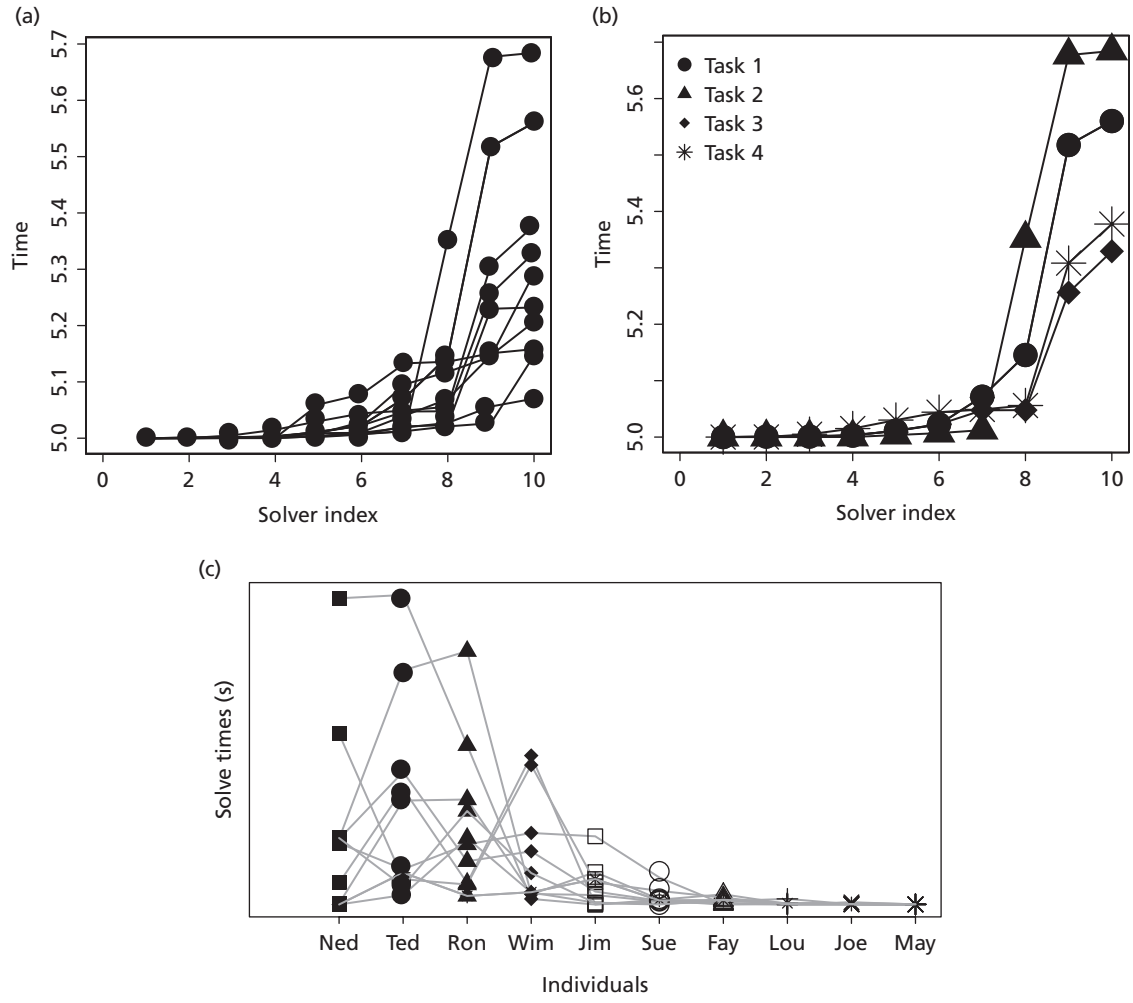

Figure 5.1 (a) Plot showing simulated diffusion times for each of the ten tasks. The points for each diffusion are joined in a separate line and represent a unique task. (b) Plot showing simulated diffusion times for Tasks 1 to 4. Each diffusion represents a unique task and is represented by a unique plotting symbol. (c) Simulated solve times per individual. The points per individual represent the time at which this individual solved the given tasks. Points for each diffusion (task) are joined in a separate line. In cases where the variation in solve times for a given individual is small, some of the points on the graph have merged together.

case is the task solution. There are two parameters of interest in the basic time of acquisition diffusion analysis model: the rate of social transmission between individuals per unit of network connection, s, and the baseline rate of trait performance in the absence of social transmission, λ_0. Throughout this chapter, we refer to the s parameter as the social transmission parameter, and to λ_0 as the baseline parameter.

The hazard function for the model is expressed as:

$$\lambda_i(t) = \lambda_0(t)\big(1 - z_i(t)\big)R_i(t) \tag{1}$$

such that

$$R_i(t) = \left(s\sum_{j=1}^{N} a_{ij}z_j(t) + 1\right) \tag{2}$$

where $\lambda_i(t)$ is the rate at which individual i acquires the task solution at time t, $\lambda_0(t)$ is a baseline acquisition function determining the distribution of latencies to acquisition in the absence of social transmission (that is, through asocial learning), and $z_i(t)$ gives the status (1 = informed, 0 = naïve) of individual i at time t. The $(1-z_i(t))$ and $z_i(t)$ terms ensure that the task solution is only transmitted from informed to uninformed individuals (Hoppitt, Boogert, et al. 2010). Previous versions of time of acquisition diffusion analysis allow for an increasing or decreasing baseline rate $\lambda_0(t)$ (Hoppitt, Kandler, et al. 2010). Here, we restrict ourselves to expanding the version for a constant baseline rate (i.e. $\lambda_0(t) = \boldsymbol{\lambda}_0$) (Hoppitt, Boogert, et al. 2010), although the version for a non-constant baseline rate can be expanded in the same way.

The model assumes that the rate of social transmission between individuals is proportional to the connection between them which is given by a_{ij} (see Equation 2). The model is used to generate a likelihood function, allowing it to be fitted by maximum likelihood or analysed using Bayesian methods. Social transmission is inferred if a model including $\boldsymbol{s}$ is better than a model with $\boldsymbol{s}=0$, using, for example Akaike's information criterion, if maximum likelihood fitting is used, or Bayes factor if Bayesian methods are used (see 'Bayesian formulation of time of acquisition diffusion analysis'). For simplicity, here we assume that a social effect (i.e. $\boldsymbol{s} > 0$) is always indicative of social transmission, although in reality this need not be the case (Atton et al. 2012).

Network-based diffusion analysis can be adapted to include other variables influencing the rate of social transmission or asocial learning that vary across individuals and/or time, by expanding the model for V continuous (or indicator) variables as follows:

$$R_i(t) = \left(s \exp\left(\Gamma_i(t)\right) \sum_{j=1}^{N} a_{ij} z_j(t) + \exp\left(B_i(t)\right) \right) \quad (3)$$

where

$$\Gamma_i(t) = \sum_{k=1}^{V} \gamma_k x_{k,i}(t) \quad (4)$$

and

$$\mathrm{B}_i(t) = \sum_{\mathrm{k}=1}^{V} \beta_k x_{k,i}(t) \quad (5)$$

Here $\Gamma_i(t)$ and $B_i(t)$ are linear predictors similar to those used in a generalized linear model.[1] The term $x_{k,i}(t)$ is the value of the kth variable for individual i at time t, β_k is the coefficient giving the effect of variable k on asocial learning, giving the natural logarithm of the multiplicative effect per unit of $x_{k,i}(t)$. Similarly, γ_k is the coefficient giving the effect of variable k on the rate of social transmission (Hoppitt and Laland 2013).

Bayesian formulation of time of acquisition diffusion analysis

In principle, the formulation of the model can remain the same for a Bayesian approach as for a model fitted by maximum likelihood. However, here we wish to include random effects, and reparameterize the model in a way that makes it easier to use in a Bayesian context. Thus, we apply a Bayesian time of acquisition diffusion analysis to the simulated dataset described in 'Previous formulation of time of acquisition diffusion analysis' to assess its performance under different circumstances. To illustrate the importance of both random effects and social transmission, four models were considered based on their inclusion/exclusion. Two of the models (Models 1 and 2) do not include random effects, while Models 3 and 4 do. Likewise two of the models (Models 1 and 3) do not include an $\boldsymbol{s}$ parameter, while Models 2 and 4 do. Please see Table 5.1 for details.

The linear predictors are easily adapted to include random effects. For Models 3 and 4 for example, random effects $\boldsymbol{\varepsilon}$ at the individual level were considered such that $\boldsymbol{\varepsilon} = \{\varepsilon_1, \ldots, \varepsilon_{10}\}$ and the total number of individuals is ten. The term $R_i(t)$ in Equation 2 is therefore expanded to

$$R_i(t) = \left(s \sum_{j=1}^{N} a_{ij} z_j(t) + \exp\left(\varepsilon_k\right) \right) \quad (6)$$

where $k \in \{1, \ldots, 5\}$ and depends on which task is involved. The rate of trait performance $\lambda_i(t)$ for individual i, at time t therefore becomes

[1] This general formulation allows the effects of individual-level variables on asocial learning and social transmission to differ. Hoppitt, Boogert, et al. (2010) suggested a constrained additive model constraining for all k, and the multiplicative model by constraining for all k.

Table 5.1 Models considered.

Model	Parameters
1	λ_0
2	λ_0, s′
3	λ_0, ε_i (i = 1:10), σ_ε^2 i denotes individual effects
4	λ_0, **s′** ε_i (i = 1:10), σ_ε^2 i denotes individual effects

$$\lambda_i(t) = \lambda_0(t)\left(1 - z_i(t)\right)\left(s\sum_{j=1}^{N} a_{ij} z_j(t) + \exp(\varepsilon_k)\right) \quad (7)$$

To allow us to more easily set a prior distribution reflecting our state of knowledge (see below), Equation 7 is then re-parameterized to obtain

$$\lambda_i(t) = \left(1 - z_i(t)\right)\left(\lambda_0 s\sum_{j=1}^{N} a_{ij} z_j(t) + \lambda_0 \exp(\varepsilon_k)\right)$$

giving

$$\lambda_i(t) = \left(1 - z_i(t)\right)\left(s'\sum_{j=1}^{N} a_{ij} z_j(t) + \lambda_0 \exp(\varepsilon_k)\right)$$

where $s' = \lambda_0 s$. The effect of social interactions on the rate of learning s' and the baseline rate of learning $\boldsymbol{\lambda}_0$ are the two parameters of interest. We refer to the re-parameterized s' as the unscaled social transmission parameter, since it is not scaled such that it is quantified relative to the rate of asocial learning, as s is. The full parameter vector $\boldsymbol{\theta}$ is defined as $\boldsymbol{\theta} = \{s', \lambda_0, \boldsymbol{\varepsilon}, \sigma_\varepsilon^2\}$, where $\boldsymbol{\varepsilon}$ refers to random effects at the task level. The variance term σ_ε^2 denotes the variance for the distribution of the task-level random effects.

Likelihood function for time of acquisition diffusion analysis

Given the observed data, ω, the likelihood that the nth individual learns the behaviour at time t_n (where t_n is the observed time of acquisition for individual i) is expressed as

$$\boldsymbol{L}\left(\lambda_i(t_{n-1}) \mid \boldsymbol{\omega}\right) = \lambda_i(t_{n-1}) \exp\left(-\lambda_i(t_{n-1})\right)[t_n - t_{n-1}] \prod_{j \neq i}^{J} \exp(-\lambda_j(t_{n-1}))[t_n - t_{n-1}] \quad (8)$$

where $n > 1$, and J denotes the number of individuals in the group during the time period $t_n - t_{n-1}$. This expression represents the product of the probability density of individual i solving the task for the first time and the probability density of the naive individuals ($j \neq i$) that did not solve the task for the first time in the time period under consideration. Note that this expression is used when all the individuals in the study have solved the task within the observation period. Individuals which have solved the task are classified as uncensored individuals. In contrast, an individual that does not solve the task during the observation period is classified as a censored individual. Censored individuals are taken into account in the analysis, using the modification below.

The combined likelihood for all the events $n \in \{1:N\}$ in the observation period is expressed as

$$\prod_{n=1}^{N} \boldsymbol{L}\left(\lambda_i(t_{n-1}) \mid \boldsymbol{\omega}\right)\boldsymbol{\varphi} \quad (9)$$

where there are N performance events (i.e. the given task was solved N times), and where $\boldsymbol{\varphi}$ denotes the probability density of the naive individuals which did not perform the trait during the observation period $[t_1 - t_N]$. In particular $\boldsymbol{\varphi}$ is expressed as

$$\prod_{j \neq i} \exp\left(-\lambda_j(t_N)\left[t_Q - t_N\right]\right)$$

The terms t_N and t_Q denote the last time of performance and the time of the end of the observation period, respectively. Equation 9 thus represents the likelihood arising from the data for all individuals, both censored and uncensored.

Prior specification

A fundamental requirement for inferring parameter values using a Bayesian approach is that suitable priors are specified for each parameter. The prior represents the researcher's prior knowledge of the distribution of the parameter's under consideration. The prior specified here for the social transmission parameter s' was a uniform prior such that $\log(s') \sim U[-1,1]$. A similar prior is specified for the baseline parameter $\boldsymbol{\lambda}_0$ such that $\log(\lambda_{0)} \sim U[-10,10]$.

These priors are very wide, which would indicate a lack of prior knowledge about the plausible values

these parameters could take. Furthermore, uniform priors specified on a log scale express a prior belief that the parameters are more likely to be near zero. These are chosen fairly arbitrarily for the purposes of this simulation. In this section, we discuss how a researcher might set these priors so they represent the prior state of knowledge and discuss the circumstances under which this is important. A hierarchical prior (Gelman 2006; Gustafson et al. 2006) is specified for the random effects $\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right)$. A gamma distribution is used as a prior for the variance term σ_ε^2.

Generating posteriors using updating methods

The Bayesian approach (R. King et al. 2010; Lee 1989) usually involves the use of an Markov chain Monte Carlo algorithm, which is deployed to generate a sequence of values which converge to the joint posterior distribution of the parameters (see 'Why do we need Bayesian network-based diffusion analysis?') given the data observed. Note that after the simulations are conducted, the properties of the resulting posterior sample (after removing the output from the initial 10% of simulations, called the 'burn in') will reflect the properties of the posterior distribution of the parameters under consideration. There are various methods of parameter updating, two of which are Metropolis Hastings and Gibbs sampling. The parameters in this analysis were updated using a random walk Metropolis Hastings (Gamerman 1997; Gamerman and Lopes 2006; McCarthy 2007) update method.

For the individual model simulations, each simulation consisted of 10,000 iterations and a conservative burn in (initial 10% of the iterations) was removed before obtaining the posterior sample. In addition, we demonstrate the application of Bayesian methods to diffusion analysis to achieve model discrimination. For this analysis, we employ a reversible jump Markov chain Monte Carlo algorithm to discriminate between the four models considered in this analysis.

Model discrimination

For a given dataset, there are typically a number of plausible candidate models, such as shown in Table 5.1, which may vary in the number of parameters they possess. With a number of plausible models for a given dataset, it is desirable that model discrimination is performed so as to determine which model (or, in some cases, which group of models) has (have) more support for the data observed. To achieve this, the reversible jump Markov chain Monte Carlo algorithm allows for posterior model probabilities (the probability that model is the true model, given that one of them is) to be obtained for each model, which is particularly useful when there is a large number of competing models. A common summary statistic related to the posterior model probabilities is the Bayes factor (Lee 1989), which is simply the ratio of the model probabilities for two specified models (so long as the prior model probabilities are equal). This is usually a preferred measure of relative evidence, since it does not implicitly assume that one of the models in the set considered is true. The Bayes factor is used specifically to compare two competing models (hypotheses) for a given dataset.

Mathematically, for model discrimination, we extend the previous Bayesian approach to treat the model itself as a parameter and then form the joint posterior distribution over parameter and model spaces. However, the posterior distribution is no longer of fixed dimensions, since different models have a different number of parameters. Thus, to explore the posterior distribution and to obtain posterior summary statistics, we use a reversible jump Markov chain Monte Carlo approach (Lesaffre and Lawson 2012).

Note that this approach comprises a two-step algorithm which involves the Metropolis Hastings algorithm and a reversible jump step. The first step involves updating the parameters given the model state, and the second step involves updating the model itself. This results in a sequence of model states at the end of the simulation which represent the exploration of model space during the iterations within the simulation. Figure 5.2 illustrates this concept graphically. Model discrimination in the Bayesian context is discussed in the tutorial provided on Bayesian network-based diffusion analysis on the Laland lab website. In addition, a data example and sample R code are provided to illustrate the concept.

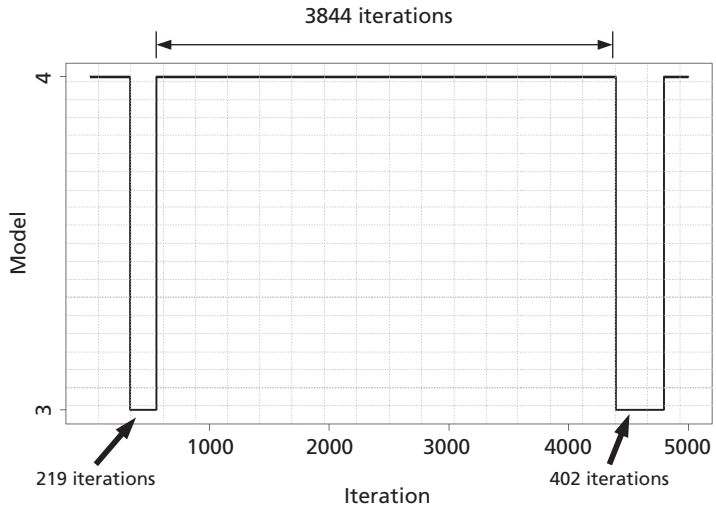

Figure 5.2 Plot illustrating model updating during a theoretical simulation involving two models: Model 3 and Model 4. Note that the model state does not necessarily change at each iteration.

Results

Posterior parameter estimates

The posterior parameter estimates for the parameters in each model are provided in Table 5.2. As noted in 'Simulated diffusion data', we used a fixed value of 0.6 for the social effect and a fixed value of 0.3 for the baseline rate of asocial learning. Below we consider the relative merits of the four models in accounting for the data. However, for the moment we merely draw attention to the fact that the posterior parameter estimates for the model parameters differ between the relevant models. Most strikingly, the parameter likely to be considered of greatest interest to users of network-based diffusion analysis, the social transmission parameter s' differs strongly between Models 2 and 4, which do not, or do, control for random effects, respectively. The posterior estimates for the baseline rate parameter λ_0 showed relatively less variation between models.

From the results, the median posterior estimate of the unscaled social transmission parameter was found to be higher in Model 4 than in Model 2 (the other model which contained this parameter). The log of the estimate in this model (−0.50) is closest to the log of the value (−0.51) used to simulate the data ($s' = e^{-0.50} = 0.61$), whereas the estimate generated by Model 2, which fails to control for random effects, is poor ($s' = e^{-6.62} = 0.001$). Hence, the analysis shows that controlling for random effects is critical to generating accurate estimates of the magnitude of social transmission and that a failure to do so could lead researchers to seriously underestimate or overestimate the social influence.

Figure 5.3 shows the trace plot for the unscaled social transmission parameter s' for the two models. A trace plot is a time series plot which provides a rough indication of how well the Monte Carlo chain has mixed and has explored the posterior distribution. The x-axis represents the iteration number within the simulation, and the simulated values of the parameter are represented on the y-axis. From the trace plots, it is clear that there is more variation in the posterior values in Model 2 than in Model 4, again indicating that Model 2 is poorer.

In both Models 2 and 4, the correlation between the social transmission and baseline rate parameters was negative but small. However, the random effect parameters were found to be correlated with each other and also with the baseline rate (and, to a lesser degree, with the social effect parameter). These correlations explain why the posterior estimates for the baseline rate of solving are so poor, a point to which we return in the final section. For illustration, Figure 5.4 shows the correlation between the baseline rate parameter and the random effect parameters for individuals 9 and 10. The density plots for each random effect parameter are shown in Figure 5.5.

From Figure 5.4, it is clear that there is a negative correlation between the baseline rate parameter

Table 5.2 Posterior parameter estimates are provided in natural logarithms (except the random effects and variance parameter) and are accompanied by symmetric credible intervals for the models considered. For the social transmission and baseline rate parameters, the posterior median and credible intervals are provided. For the random effects and variance parameter, the posterior mean and credible intervals are provided. The median is the preferred summary statistic when the distribution of the parameter of interest is skewed.

Parameter	Model 1	Model 2	Model 3	Model 4
log(s')		−6.62 (−9.81, −2.53)		−0.50 (−0.95, −0.07)
$\log(\lambda_0)$	2.48 (2.27, 2.68)	2.47 (2.25, 2.67)	3.57 (2.38, 5.61)	2.88 (0.78, 4.58)
ε_1 (Ned)			−2.07 (−4.14, −0.72)	−3.65 (−8.59, −0.53)
ε_2 (Ted)			−2.14 (−4.18, −0.83)	−3.94 (−8.12, −0.82)
ε_3 (Ron)			−1.94 (−3.98, −0.62)	−3.29 (−7.62, −0.29)
ε_4 (Wim)			−1.44 (−3.46, −0.16)	−1.82 (−5.48, 0.750)
ε_5 (Jim)			−0.63 (−2.67, 0.79)	−0.06 (−2.14, 2.13)
ε_6 (Sue)			0.48 (−1.61, 1.82)	1.29 (−0.66, 3.42)
ε_7 (Fay)			1.12 (−0.78, 2.46)	1.99 (0.03, 4.11)
ε_8 (Lou)			2.09 (0.07, 3.40)	2.97 (0.99, 5.09)
ε_9 (Joe)			2.89 (0.83, 4.23)	3.76 (1.85, 5.87)
ε_{10} (May)			4.35 (2.32, 5.73)	5.24 (3.23, 7.37)
σ_ε^2			7.29 (3.86, 12.62)	9.79 (5.05, 16.19)

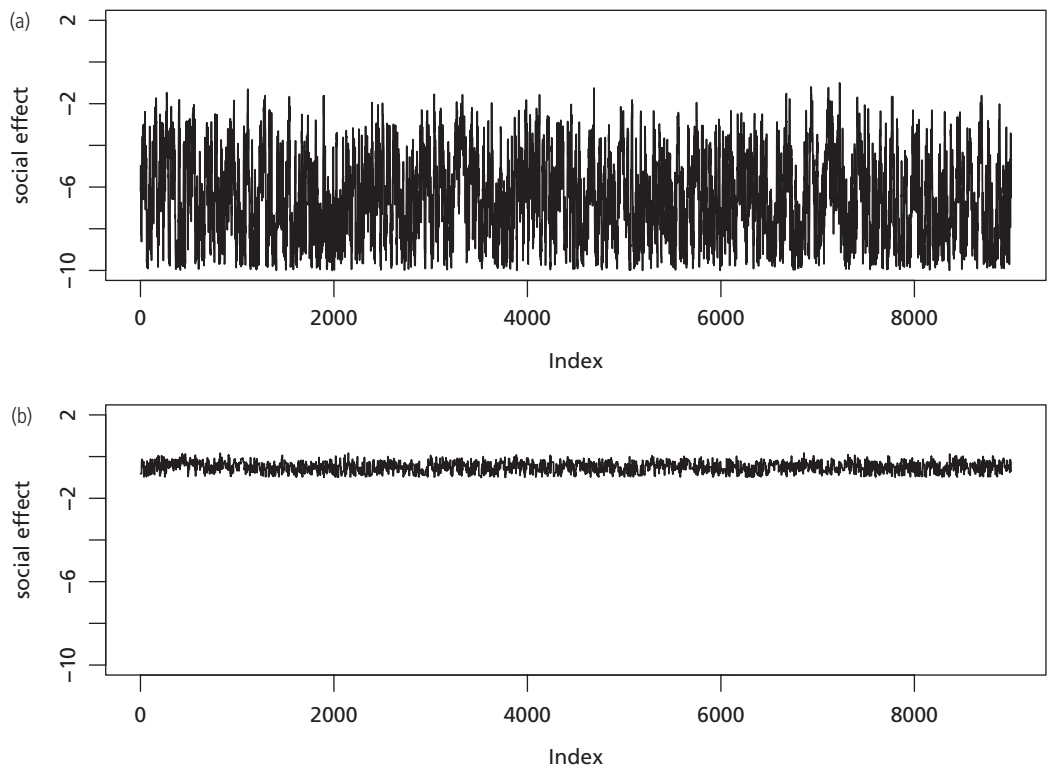

Figure 5.3 Trace plots for the social effect parameters for (a) Model 2, and (b) Model 4. The *x*-axis shows the iteration number, and the *y*-axis shows the log of the simulated value for the parameter. The first 1000 iterations were treated as 'burn in' and were removed before plotting. Identical *y*-axes are used in the plots to highlight the difference in variance of the simulated values obtained under the two models' hypotheses.

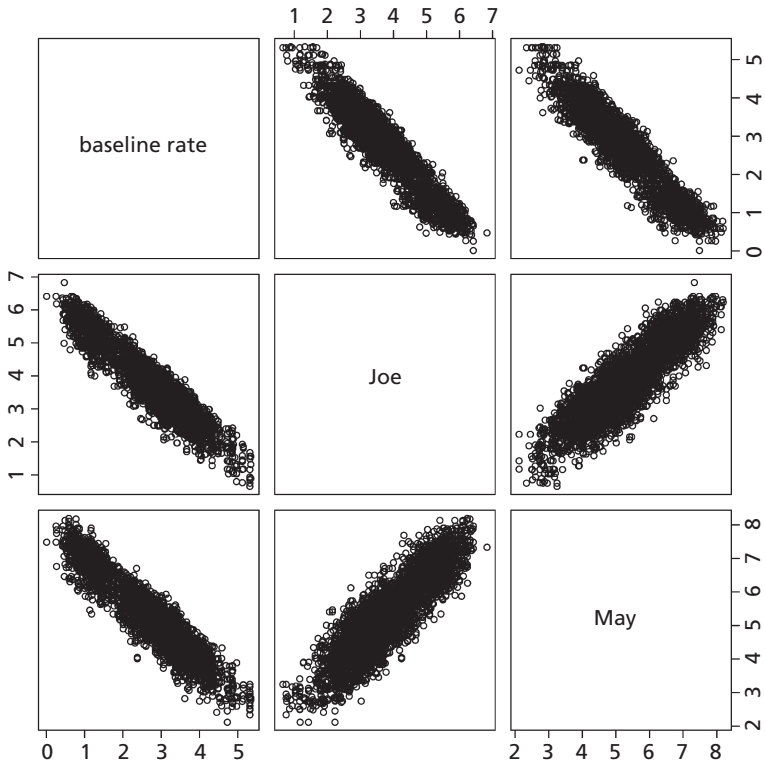

Figure 5.4 Scatter plot showing the correlation in the joint posterior distribution between the baseline rate and the random effect parameters for individual 9 (Joe) and individual 10 (May).

and the random effect parameter representing Joe. A similar effect is observed for May. The interpretation of these posterior correlations would be that, as the posterior value of the random effect parameter for Joe increases, that of the baseline rate parameter decreases. A positive correlation was observed between the random effect parameters for Joe and May. The interpretation in this case would be that, as the posterior value of the random effect parameter representing either of these individuals increases, then that of the other individual would also increase.

The density plots for each random effect parameter are shown in Figure 5a; the density plot for a given parameter can be thought of as a normal approximation to the distribution of this parameter. Each random effect parameter represented in Figure

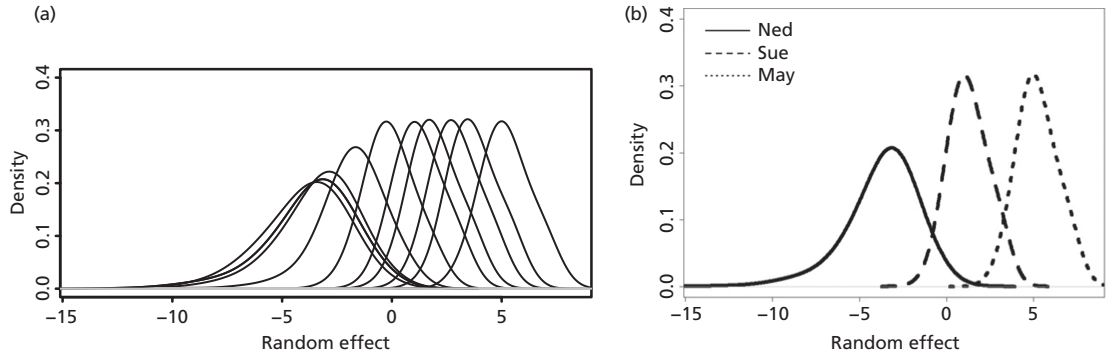

Figure 5.5 Density plots per random effect parameter: (a) ten random effects; (b) three random effects.

5 is associated with a bell-shaped plot, and the width of this 'bell' gives an indication of the spread of the posterior values for this parameter. The x coordinate, corresponding to the apex of the bell, represents the posterior parameter value, which has the highest density and represents the mean (or average) for the distribution. The density plots therefore provide a visual summary of the spread and centre of the distribution of the posterior values for the parameters considered. As expected, for the ten individuals considered, these vary in their effect on the rate of solving, reflecting individual differences in the ability to solve the tasks. This is illustrated more clearly in Figure 5b, where it is evident that May had a positive effect, Ned a negative impact, and Sue had little impact on the rate of solving.

Markov chain Monte Carlo replication

The Markov chain Monte Carlo simulations for Models 2 and 4 were repeated 100 times so as to allow enough time for the respective credible intervals to obtain the values used to simulate the data. From the results of these replications, we note that the credible intervals for the social transmission parameter for Model 4 were observed to be narrower than those for Model 2. In addition, while the credible intervals for the social transmission parameter for Model 4 always contained the value used to generate the data, the credible intervals for Model 2 were not found to contain the parameter value that was used to generate the data. The sole difference between Models 2 and 4 is the inclusion of random effects in Model 4, which thus underlies Model 4's superior performance.

Model discrimination

Model discrimination was performed for three different model comparisons. The posterior model probabilities obtained are shown in Table 5.3. When all four models are considered simultaneously (Comparison 3), there is far more support for Model 4 (the correct model, used to simulate the data) than for any of the other models (i.e. Model 4 was the best model in 94% of the simulations). In addition, when only Models 1 and 2 are considered (Comparison 1; i.e. no random effects) or only Models 3 and 4 are considered (Comparison 2; i.e. random effects), in each case the model that contains a social transmission parameter is by far the best supported model.

Table 5.3 Posterior model probabilities obtained from the application of the reversible jump Markov Chain Monte Carlo algorithm.

Models/ comparison	Comparison 1(Models 1 and 2)	Comparison 2(Models 3 and 4)	Comparison 3(Models 1, 2, 3, and 4)
1	0.11	–	0.0005
2	0.89	–	0.0032
3	–	0.05	0.0550
4	–	0.95	0.9420

These conclusions are reinforced by consideration of the Bayes factor associated with model comparisons, which can be derived by dividing the posterior model probability for the better supported model by the posterior model probability for the alternative. From the analysis using Comparison 3, we note that the Bayes factor in favour of random effects (i.e. (Model 3 + Model 4)/(Model 1 + Model 2)] is greater than 269.5, indicating decisive posterior support for these parameters and their importance in modelling the data observed. The Bayes factor in favour of Model 4, which contains the social transmission parameter, against Model 3, which does not, is 17.2, which suggests that there is very strong evidence for Model 4 (against Model 3). The Bayes factor in favour of the social transmission parameter (i.e. (Model 2 + Model 4)/(Model 1 + Model 3)] is 17.1, which suggests very strong support for the inclusion of this parameter in the model. In this simulated dataset, strong support for s implies a general influence of other individuals on the rate of solving, since here the network is comprised of homogenous patterns of association. However, more typically, support for a model containing s will be indicative of social transmission along pathways of association.

A Bayesian approach to quantifying diffusion on social networks: conclusions and future directions

We have developed a Bayesian version of network-based diffusion analysis as a means to control for

random effects that can be generated by individual differences in ability among datasets that repeatedly test the same group or groups. The application of this approach to a simulated dataset clearly illustrates its merits, which we discuss in this section. The incorporation of random effects to account for heterogeneity in the baseline rate of asocial learning in Model 4 yielded a more realistic estimate (0.61) for the social effect on learning (recall that the value of the unscaled social transmission parameter used to generate the data was 0.6). Conversely, when random effect parameters were left out of the model, the social effect was so seriously underestimated that it would have been falsely regarded as negligible. Importantly, we note that the posterior mean (and standard deviation) for the variance of the random effects is 9.79 (2.96), reflecting substantial variation between the rates at which the individuals learn asocially. This is illustrated clearly in Figure, where individuals Joe and May have relatively high baseline rates while Ned and Ted have much lower rates.

The model discrimination exercise indicated that there was decisive posterior support for the random effect parameters, since Model 4 received the highest posterior support, and the Bayes factor in favour of random effects is greater than 100. Of course, to a large extent this is an artefact of the dataset deployed, and different datasets would give greater or lesser support for the models with random effects. However, the result illustrates that at least in some cases it will be necessary to control for random effects and that the Bayesian network-based diffusion analysis is capable of doing this effectively. The exercise also illustrates how a failure to control for random effects can lead to inaccurate estimates for other parameters of interest—most obviously, the magnitude of the social effect.

We note that the median posterior parameter estimate obtained for the social transmission and baseline rate parameters in the model of choice, Model 4, are not precisely equal to the point values used to simulate the dataset. For the social transmission parameter, the 95% credible interval does contain the value used for simulation, and the median is extremely close. However, this was not the case for the baseline rate parameter. The 95% credible intervals for the baseline rate parameter did not contain the value used to simulate the data. The differences between the point values can be attributed to correlations between model parameters in the joint posterior distribution.

This is apparent with the baseline rate parameter, since it appears in the model as a product with the random effect parameter (i.e. $\lambda_0 \exp(\epsilon_i)$). This means that a range of different combinations of λ_0 and $\exp(\epsilon_i)$, for any given i, can explain the data approximately equally well. For instance, a relatively low value of λ_0 and a relatively high value of $\exp(\epsilon_i)$ would explain the data roughly as well as a relatively high value of λ_0 and a relatively low value of $\exp(\epsilon_i)$. This is what a correlation between two parameters in the posterior distribution is telling us.

An alternative formulation for which the correlations between the baseline rate and the random effect parameters are avoided is to estimate individual baseline rates of asocial learning, λ_{0i}, as random effects, which might be appropriate to researchers who wish to ascertain the asocial performance of particular animals. Whether such alternative formulations are warranted depends on the goals of the researcher. In principle, researchers could make this judgement given the nature of the data. However, we suspect that for most applications of network-based diffusion analysis, the primary objective is accurate estimation of s', with λ_0 treated as a nuisance parameter, and hence the formulation presented here suffices. We have found that alternative formulations of random effects for asocial parameters do not generally affect estimation of s'.

Researchers unfamiliar with use of Bayesian methods might be put off using a Bayesian network-based diffusion analysis by the need to specify a prior distribution, quantifying our state of knowledge about the parameters before receiving the data. Often, they will consider themselves to have no solid basis on which to make judgements about the rate of asocial learning, and rate of social transmission prior to collecting data. Here we will discuss the circumstance under which such choices matter, and, where they do, how a prior distribution might be derived.

In some analyses, parameter estimation might be the key focus—probably focussing on estimating s' with 95% credible intervals—with no need for

model discrimination. In such cases, one can specify a vague prior for parameters, with a large variance reflecting little prior knowledge, without worrying about exactly how large the variance has to be, or exactly what form the prior distribution should take. So long as the prior is fairly flat in the area that the 95% credible intervals fall, our results will not be greatly affected (Jaynes 2003). A pragmatic approach might be to choose a uniform distribution for s' from 0 to a large value far higher than s' could plausible take (see below), and likewise for λ_0.

In contrast, if model discrimination is the aim (e.g. when trying to decide which of a number of social networks best explains a diffusion), then the choice of priors is important, and the prior should reflect our state of knowledge. This is because the evidence for a given model depends not only on the likelihood of the data but also on how concentrated the priors are in the area in which the model parameters are plausibly located. Consequently, the addition of a parameter for which we have a little prior knowledge will penalize a model more than the addition of a parameter for which we have a lot of prior knowledge (see Jaynes 2003 for an explanation of why this is). For this reason, it is important that the prior distribution reflects our state of knowledge, for model discrimination.

A social learning researcher might protest that we do not know how strong social transmission might be (or indeed how rapid asocial learning might be) before conducting the diffusion experiment. However, we argue that researchers do have prior knowledge about such things, and this can be appropriately incorporated into the analysis. To illustrate this, imagine the researchers are conducting a diffusion experiment on swans: no diffusion experiments have ever been conducted on swans, so, on the face of it, they do not know anything about how fast swans might solve the task, whether it is by asocial learning or social transmission. However, there are possibilities that researchers would consider to be, a priori, impossible. Imagine they got the results of an network-based diffusion analysis on swans which estimated $s' = 1000$ per unit connection per second; this would mean that individuals with a single unit of connection to informed swans would, on average, solve the task in 0.001 seconds. Unless the network was quantified on a very small scale, no researcher would believe this result—they would probably assume something had gone wrong in the analysis (e.g. the time units were days, not seconds).

Such reasoning suggests that researchers do have prior knowledge about how fast social transmission could occur if and when it does occur. If they input a prior that allows a large range of values that are a priori implausible, they are penalizing too heavily against models including social transmission, and the Bayes Factors obtained will not reflect the state of our knowledge. Our suggestion for deriving a suitable prior for s' is to start by setting an upper limit, S_{max}, on how fast social transmission could plausibly occur, when all an individual's associates are informed (this can be estimated as $S_{max} = 1/T_{min}$, where T_{min} is the minimum plausible average time it would take individuals to solve a task under such circumstances). Since s' is the rate of social transmission per unit connection, take the average total network connection over all individuals, k. We then set the upper limit as $s'_{max} = S_{max}/k$. One could then specify a uniform prior $s' \sim U(0, s'_{max})$, which would state that we consider all values of s' within this range to be equally plausible.

This is likely to be a conservative approach, since values at the top end of the range are likely to be, in reality, less plausible than lower values, meaning models with s' included will be penalized more heavily than truly reflects our state of knowledge. Consequently, we suggest that researchers also check the sensitivity of their findings to different priors—perhaps altering the value of S_{max}, and also considering vague priors of different forms. For this purpose, we would suggest that one would need to go through a similar exercise to set a lower limit on s' to obtain s'_{min} where $s_{min} = \frac{S_{min}}{k}$, $s_{min} = \frac{1}{T_{max}}$, and T_{max} represents the maximum possible time it is believed to take to solve the task at hand. This would result in a prior specification of $\log(s') \sim U(\log(s'_{min}), \log(s'_{max}))$. Priors can be chosen in a similar way for λ_0, by quantifying λ_{0Max} as the highest plausible rate of asocial learning and setting $\lambda_0 \sim U(0, \lambda_{0Max})$. However, this specification is less important than the specification of the prior for s', if the emphasis is on whether social transmission is occurring, since the same prior can be used for λ_0 in models with and without social transmission.

R code is available on the Laland lab website to demonstrate the Bayesian analysis of diffusion data in a simple form, including just the social effect and baseline rate parameters, and also when random effects are incorporated in the modelling process. The code provided is for demonstrative purposes. Development of a more comprehensive package, which can handle more complicated datasets, is currently underway.

Acknowledgements

Research was supported in part by a Biotechnology and Biological Sciences Research Council grant to K. N. L. and W. H. (BB/D015812/1), a European Research Council Advanced Grant to K. N. L. (EVOCULTURE, Ref. 232823), and a Netherlands Organisation for Scientific Research Rubicon Grant to N. J. B.