Figures and Tables

J.P. Meagher

2023-09-26

Background

Exploratory analysis

This section of the manuscript includes three figures exploring various aspects of the `r/ireland` dataset.

Figure 1

This figure highlights the branching structure associated with a typical sequence of discussions and the circadian rhythm associated with overall activity on the subreddit.

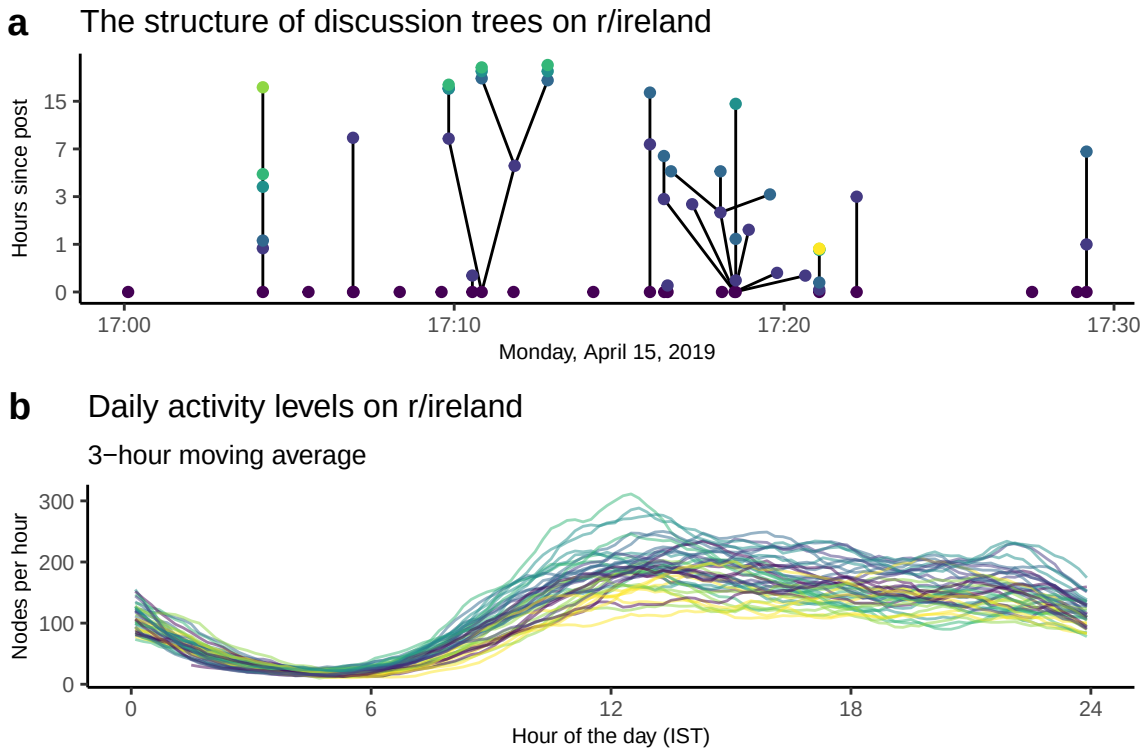


Figure 2

This figure highlights that the mean number of replies to each point within the subreddit seems to follow a circadian rhythm and that the distribution of the number of replies to each point is overdispersed relative to the Poisson distribution.

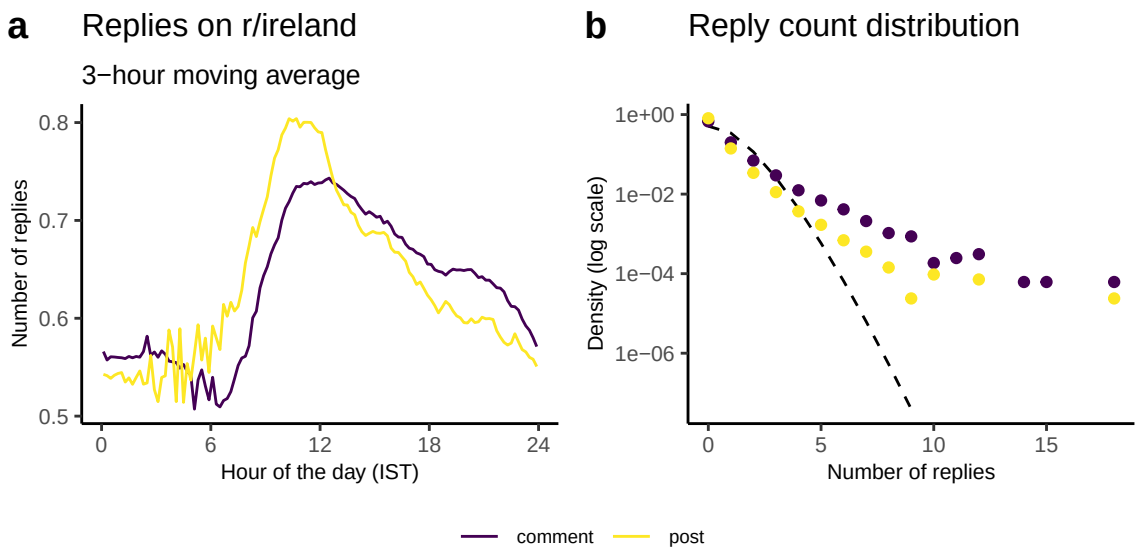
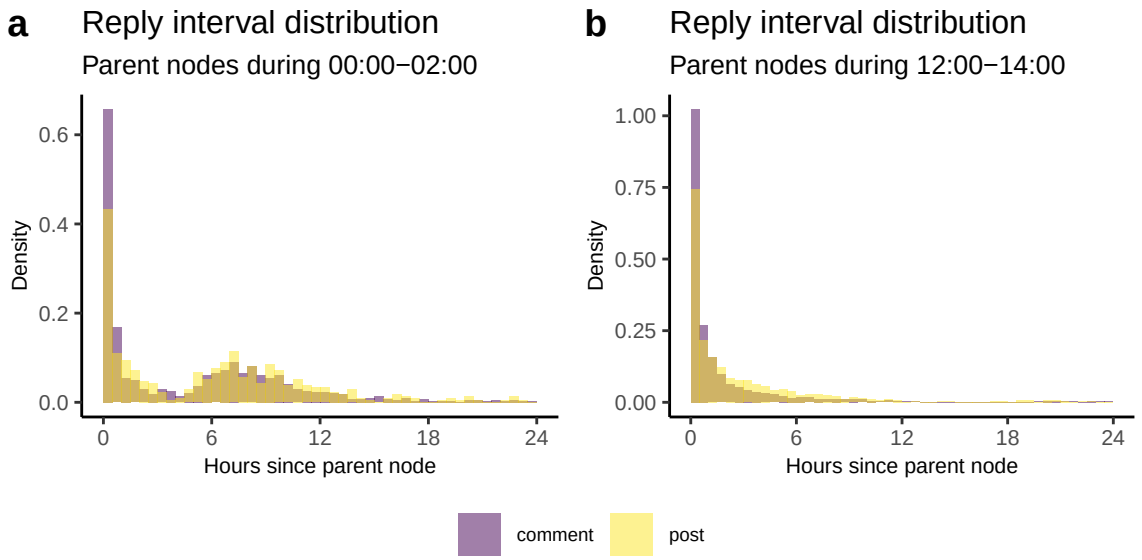


Figure 3

The final figure in this section illustrates that the distribution of generation intervals depends on the time of day that a point arrives.



Methods

This section of the manuscript presents all the methods required to conduct the reported analysis.

The generative model

Figure 4

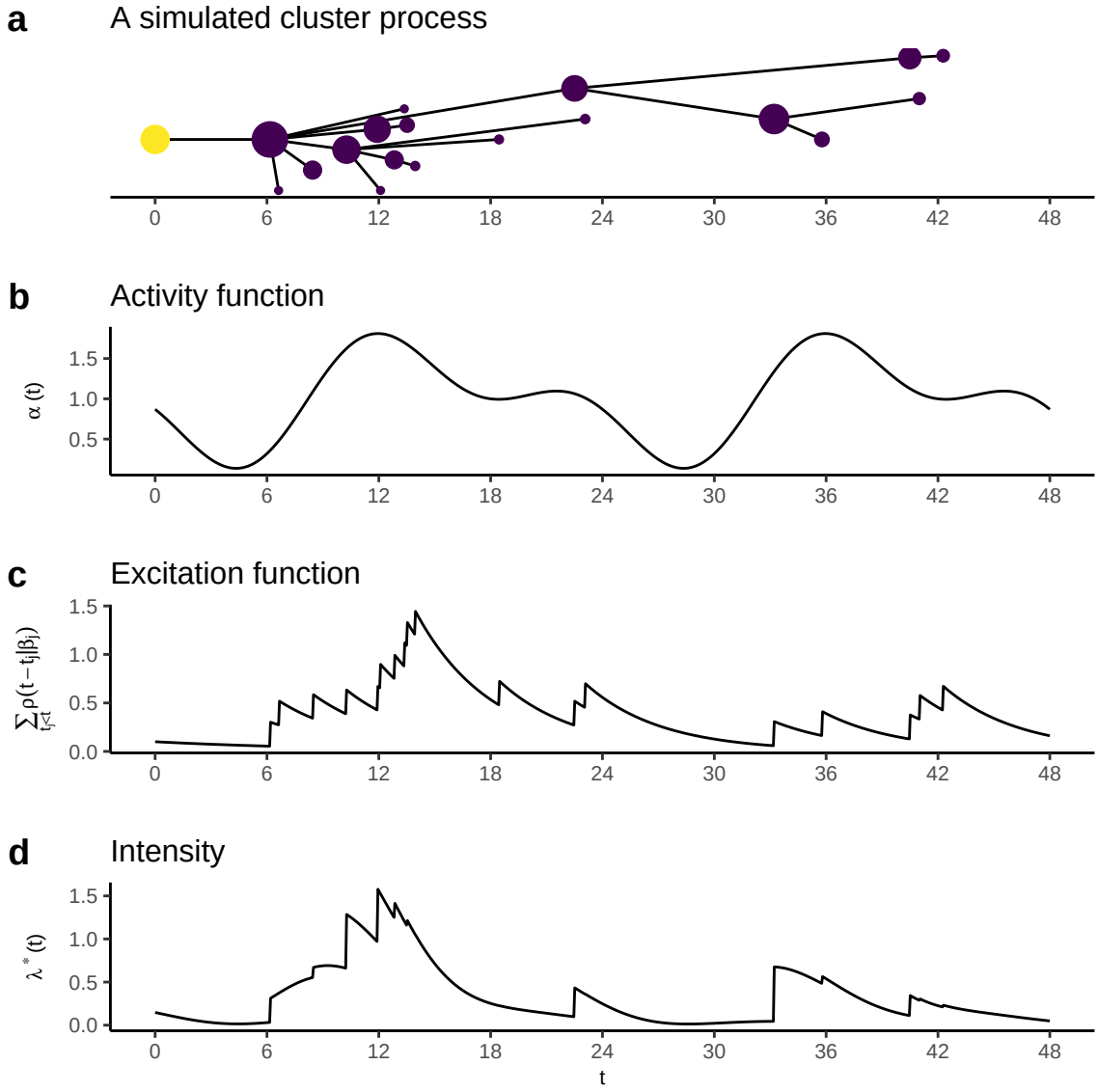
This figure presents a synthetic cluster, highlighting relevant features of the model. The figure consists of 4 elements.

First we present the simulated cluster illustrating individual reproduction numbers and the branching structure.

The second element is the circadian rhythm modulating the overall activity intensity.

The third element is the excitation induces by each new point given it's individual reproduction number and the exponential excitation function.

The fourth and final element is the conditional intensity function associated with the cluster.



Results

This section of the manuscript presents the results of our analysis as a series of tables and figures.

Inference

Table 2

This table summarises posterior distributions for each parameter of the model relating to offspring processes.

Model	μ_1	μ_2	η_1	η_2	ψ_1	ψ_2
$\mathcal{M}_1$	0.66 (0.01)	-	0.33 (0.01)	-	-	-
$\mathcal{M}_2$	0.65 (0.02)	0.67 (0.01)	0.27 (0.01)	0.38 (0.01)	-	-
$\mathcal{M}_3$	0.64 (0.02)	0.64 (0.01)	0.25 (0.01)	0.34 (0.01)	-	-
$\mathcal{M}_4$	0.65 (0.02)	0.65 (0.01)	0.25 (0.01)	0.34 (0.01)	1.15 (0.12)	6.99 (1.63)
$\mathcal{M}_5$	0.65 (0.02)	0.64 (0.01)	0.25 (0.01)	0.34 (0.01)	1.15 (0.12)	-

Table 1: Posterior mean (standard deviation) for the parameters $\boldsymbol{\eta}$, $\boldsymbol{\mu}$, and $\boldsymbol{\psi}$ within each of our candidate models. Considering each of the parameters in turn, the broad agreement on $\boldsymbol{\mu}$ across all the candidate models suggests that the expected number of offspring does not differ between immigrants and offspring. Differences in the offspring distributions for immigrants and offspring are manifest in the memory decay rate, which indicates that the expected generation interval for immigrants is longer than that for offspring. Finally, the values for $\boldsymbol{\psi}$ inferred by $\mathcal{M}_4$ suggest that immigrant points have a moderately heterogeneous offspring process while the offspring process for offspring is relatively homogeneous. As such, we include $\mathcal{M}_5$ in our analysis, which assumes heterogeneous immigrant and homogeneous offspring processes, respectively.

We include calculations to support assertions on the posterior distributions for each parameter included in the text.

```
##          mu[1] mu[2] eta[1] eta[2] psi[1] psi[2] inv_eta[1] inv_eta[2] tq[1]
## ci_lower  0.61  0.62   0.24   0.33   0.92   4.39       3.79       2.82  0.47
## ci_upper  0.70  0.68   0.26   0.35   1.40  10.15       4.21       3.05  0.53
##          tq[2] prop_zero[1] prop_zero[2]
## ci_lower  0.29         0.58         0.52
## ci_upper  0.34         0.62         0.55
```

Figure 5

This table presents out posterior inference for the activity function $\alpha(t)$.

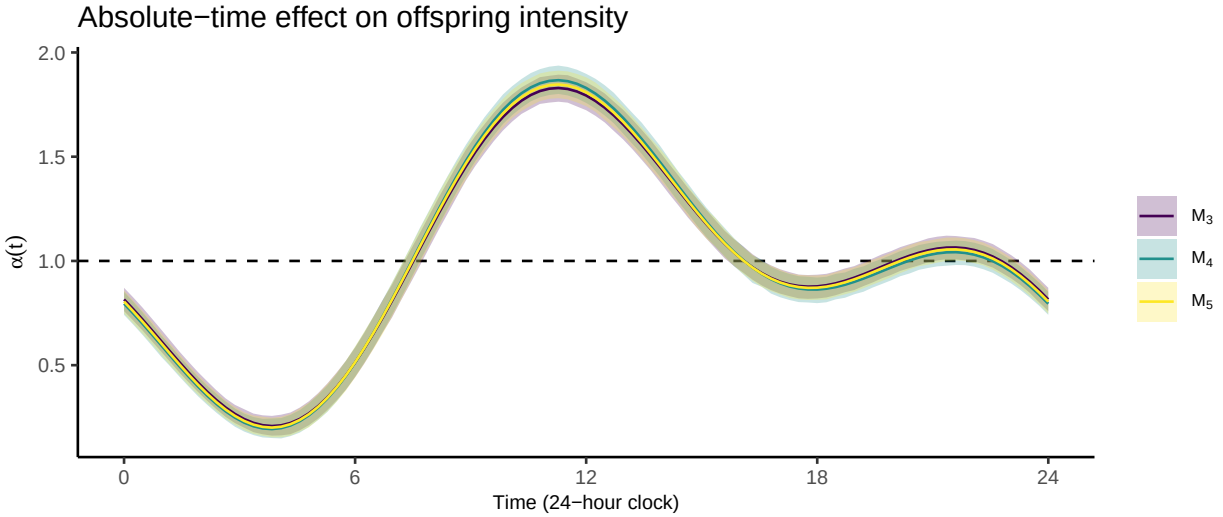


Table 3

This table presents Bayes factors for each model, allowing us to assess the evidence for each model.

We also check the Coefficient of variation associated with each model evidence.

```
## [1] 0.0008276592
```

```
## [1] 0.001166156
```

	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$
$\ln \mathcal{BF}_{l4}$	-494.75	-448.37	-117.35	0.00	-10.65

Table 2: Estimated log Bayes factor for each candidate model relative to $\mathcal{M}_4$, that is, $\ln \mathcal{BF}_{l4}$ for $l = 1, \ldots, 5$. Note that the evidence supporting each candidate model is estimated from the sampled posterior distribution $p(\theta \mid \boldsymbol{y}_{\text{train}}, \mathcal{M}_l)$ via bridge sampling. In each case, the `bridgesampling` algorithm reports a coefficient of variation for the evidence estimate of < 0.005, indicating that we have a precise estimate for each model evidence and, as a result, the corresponding Bayes factors. We find decisive evidence to support our inclusion of a circadian rhythm in the offspring intensity. Furthermore, we find decisive support for the inclusion of heterogeneous immigrant and offspring reproduction numbers.

```
## [1] 0.002134034
## [1] 0.003607418
## [1] 0.002648298
```

Assessing Predictive Performance

Here we assess the precictive performance of each model in terms of expected predictive density and continuous ranked probability score.

Table 4

This table presents the expected log predictive density for each model, allowing us to assess the predictive performance for each model.

	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$
$\widehat{\Delta\text{lpd}}_{l4}$	-1004.6 (78.7)	-892.4 (71.8)	-177.5 (37.5)	0.0 (0.0)	17.2 (11.7)

Table 3: The difference (standard error) in log cluster-wise predictive density on $\boldsymbol{y}_{\text{test}}$ between each model and $\mathcal{M}_4$. We find that $\mathcal{M}_4$ and $\mathcal{M}_5$ offer the best out-of-sample predictive performance in terms of lpd, with $\mathcal{M}_5$ outperforming $\mathcal{M}_4$ slightly. This provides decisive support for the inclusion of a circadian rhythm and heterogeneous immigrant reproduction numbers in our model for online discussion on the `r/ireland` subreddit.

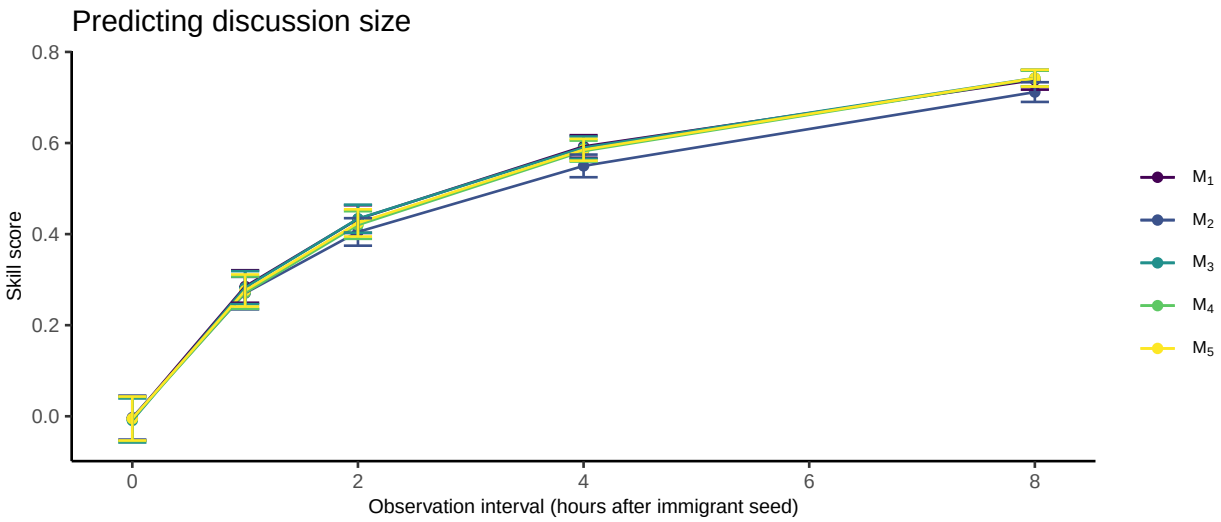
Figure 6

This figure presents our analysis of the CRPS skill score comparing the predictions for the cluster size after 48 hours to the empirical distribution.

We start by computing the mean CRPS for the training set of clusters using the empirical distribution.

Create the data frame required to present the CRPS skill scores for each learning interval.

We then construct the plot illustrating skill scores.



Assessing goodness-of-fit

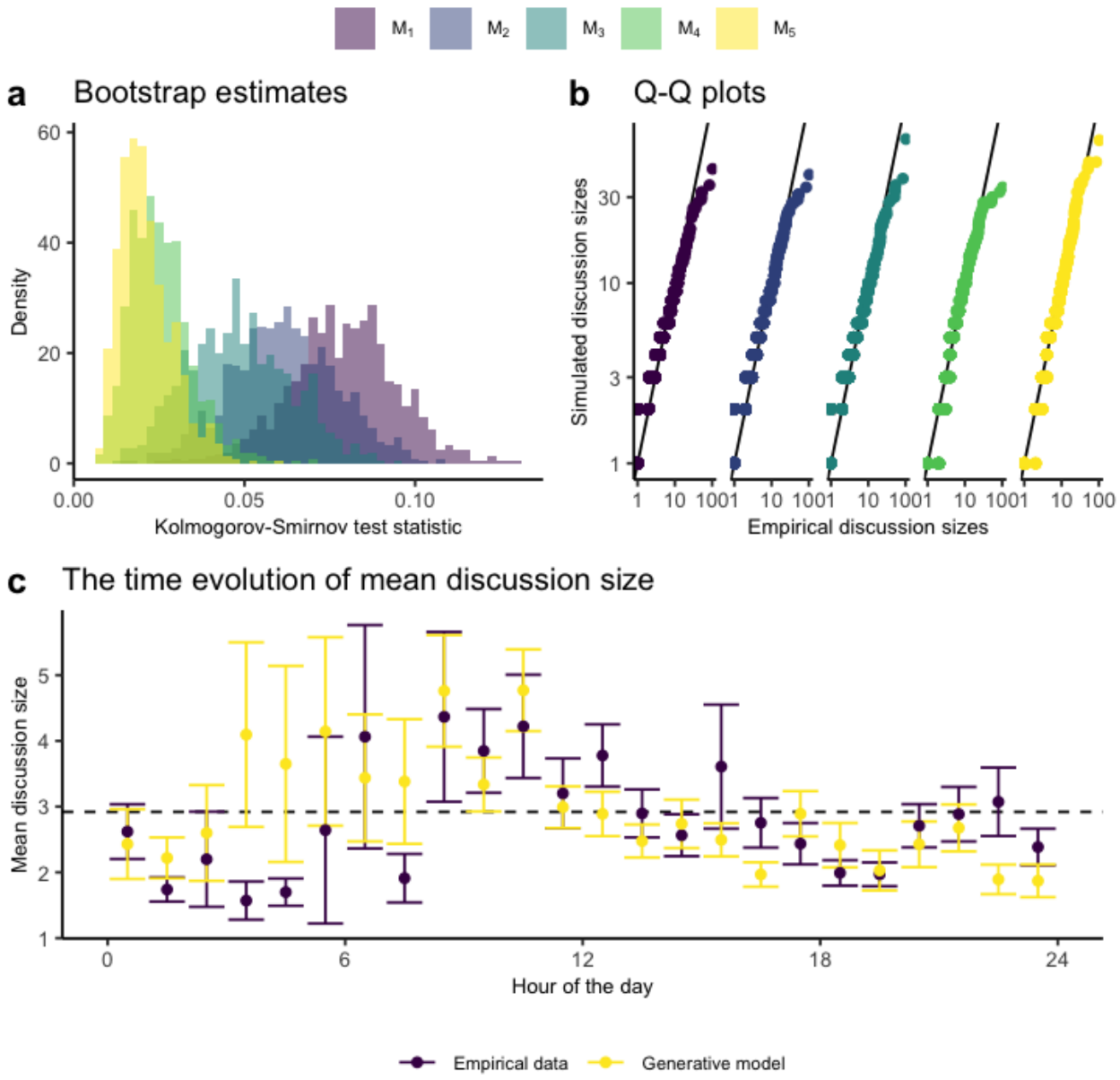
Figure 7

This figure present an analysis of the goodness of fit for each model to the set of all discussions seeded during the training interval.

We first compute bootstrapped estimates of the ks test statistic comparing predicted cluster sizes from each model to the empirical data.

We then compare the mean discussion size for each hour of the day using only our full generative model and the empirical data.

Finally, we produce quantile-quantile plots comparing the distriution of discussion sizes for each model to the empirical distribution



Appendices

Circadian Rhythm

Figure 8

Perform the spectral analysis

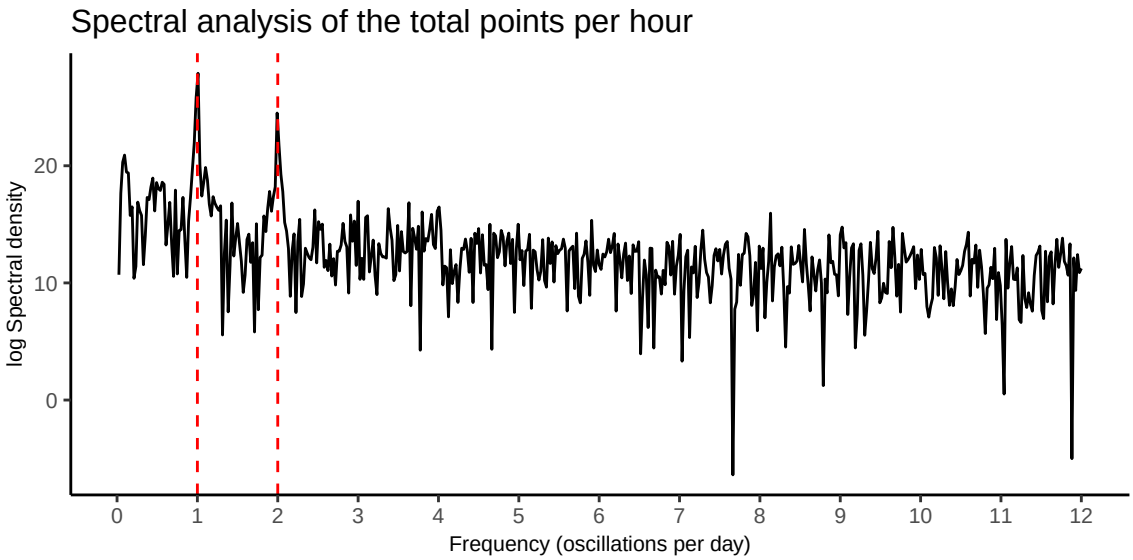


Table 5

Assess evidence for each basis function

	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
$\ln \mathcal{BF}_{l2}$	-170.23	0.00	5.01	16.12

Table 4: Estimated log Bayes factor for each candidate model relative to $\mathcal{M}_2$, that is, $\ln \mathcal{BF}_{l2}$ for $l = 1, \ldots, 4$. The evidence supporting each candidate model is estimated from the sampled posterior distribution $p(\theta \mid \boldsymbol{y}_{\text{train}}, \mathcal{M}_l)$ via bridge sampling. In each case, the `bridgesampling` algorithm reports a coefficient of variation for the evidence estimate of < 0.005, indicating that we have a precise estimate for each model evidence. This analysis presents overwhelming evidence to support a choice of $K > 1$.

Figure 9

Plot the activity function under each model