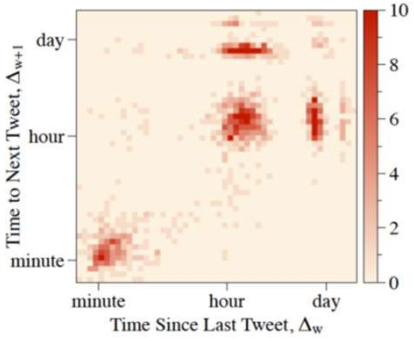
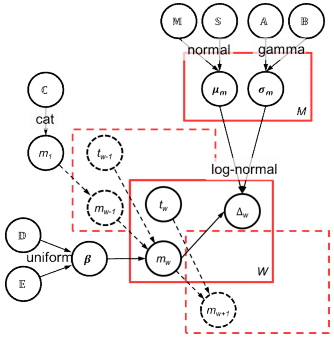
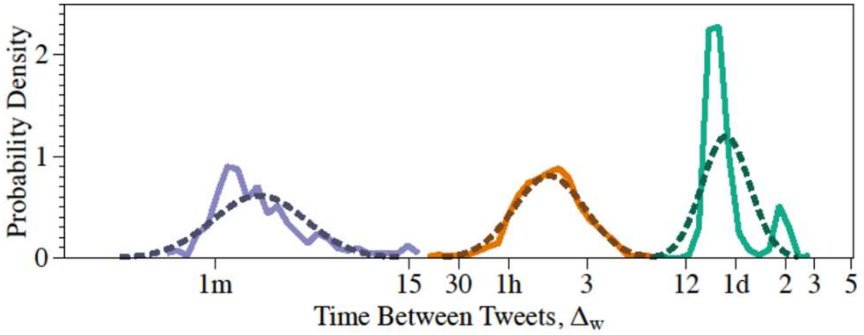
SAND2017-8003C

Exceptional service in the national interest

Sandia National Laboratories







Estimating users' mode transition functions and activity levels from social media

Hamilton Link, Jeremy D. Wendt, Richard V. Field, Jr., and Jocelyn Marthe

Sandia National Laboratories

Albuquerque, New Mexico, U.S.A.

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Social Media Modeling

Sandia National Laboratories

We consider social media research as analogous to language modeling research: build a model to explain the data.

- Language models are likelihood estimators that support many tasks…
  - Translation
  - Automatic summary
  - Bot detection
- Social media models should support diverse tasks as well.
  - Electric power load prediction
  - Network traffic analysis
  - Disease spread modeling & forecasting
  - Market segmentation
  - Bot detection
  - Clustering

Temporal Analysis of Social Media

Sandia National Laboratories

- This work focused on modeling Twitter post time patterns
  - Post times & post intervals
  - Changing patterns throughout the day
  - Diversity among users
- Temporal model presented is in the context of broader social media analysis
  - Temporal properties
  - Text analysis
  - Social network structure

Developing a New Model

Sandia National Laboratories

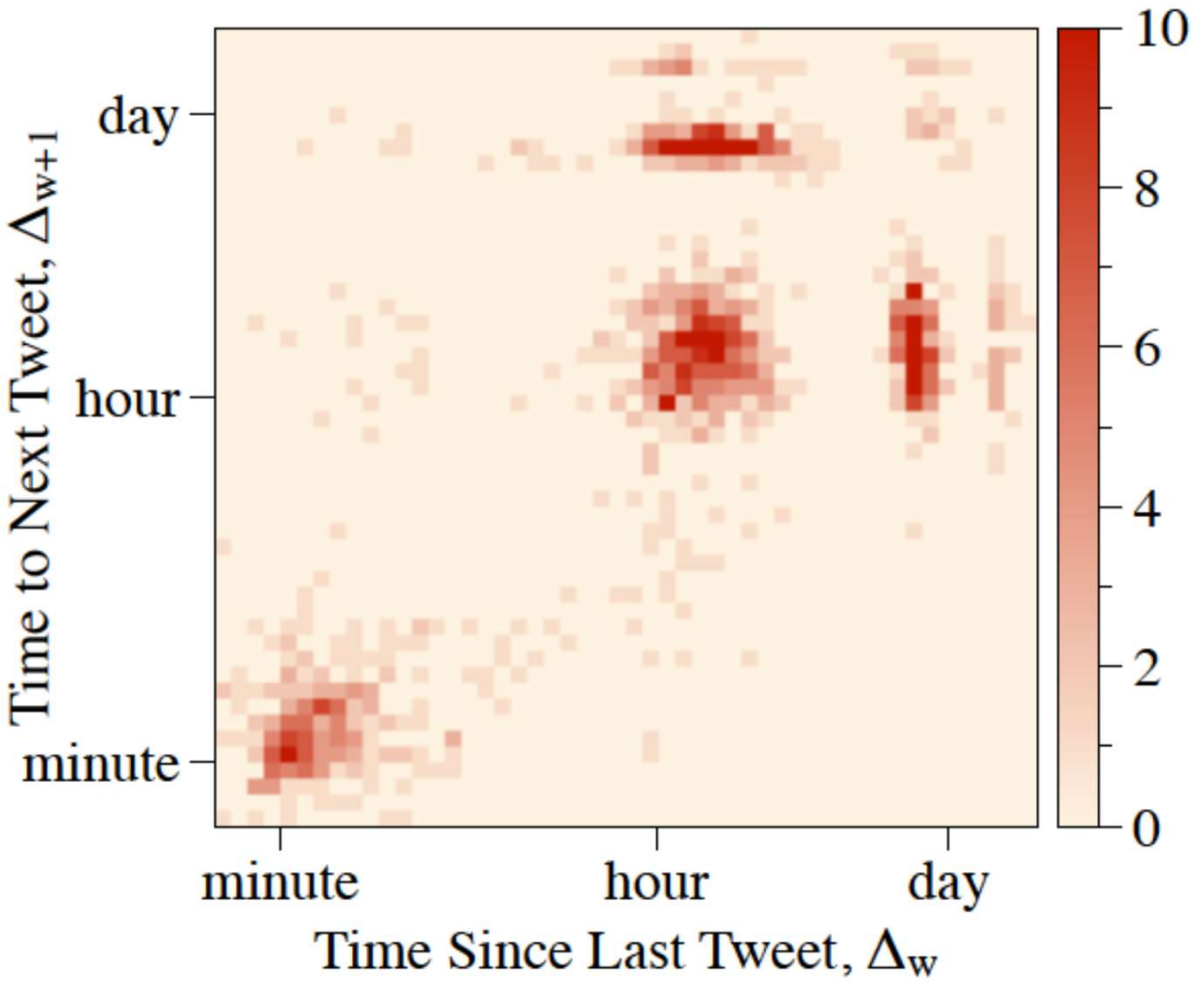
- Observed phenomenon of “modal” behavior & formed hypothesis
- LDA suggested discrete variable chosen in a user-specific mixture, influencing likelihood of observations.
  - Topics -> Activities (Modes)
  - Topic Mixture -> Mode transition function
  - Words -> Intervals
- Final model structure isn’t recognizable as LDA and isn’t trained the same way. Other point processes also could be used.

Observation

Sandia National Laboratories

Many users appear to have discrete “modes” of behavior.

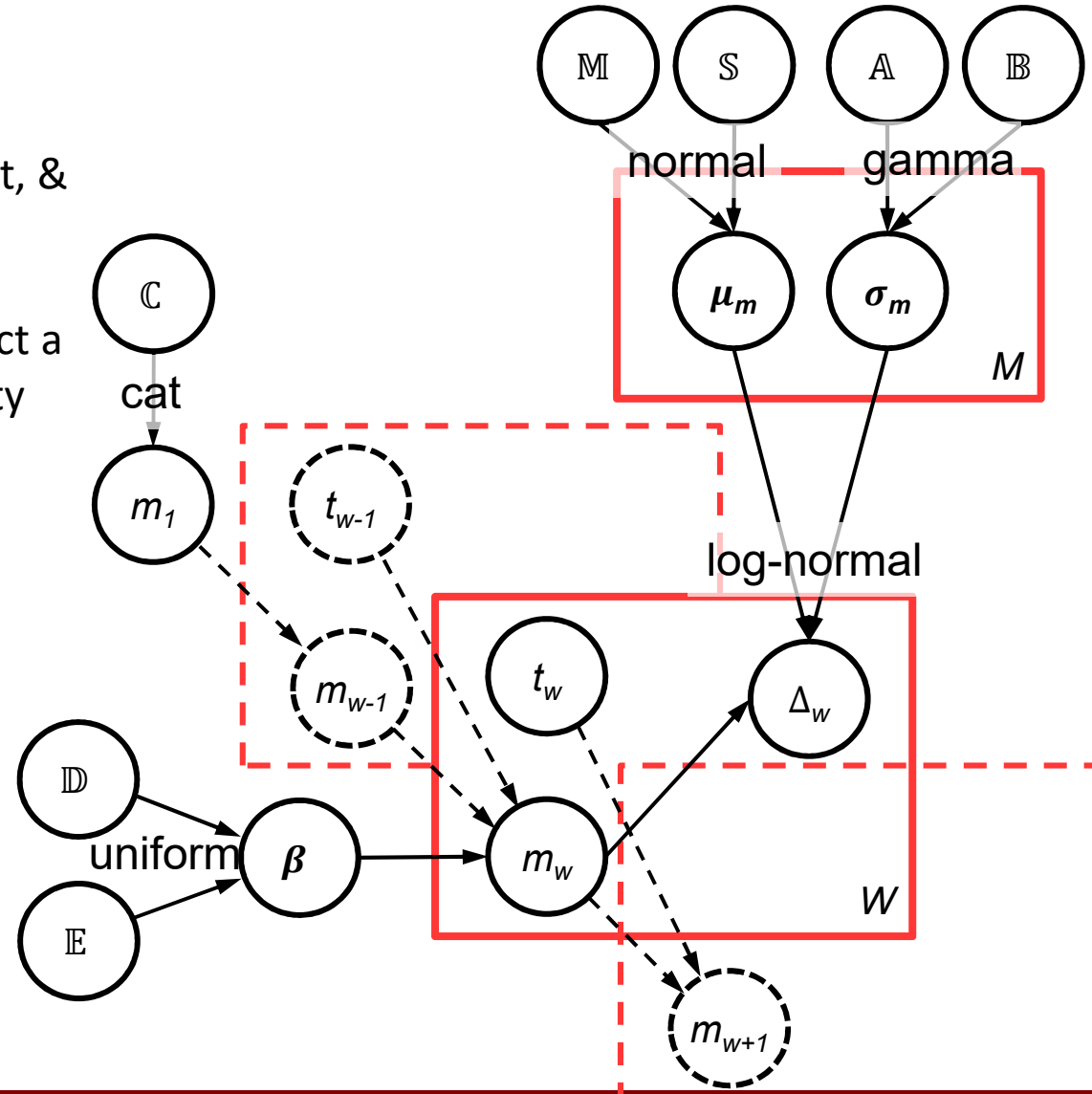
When they are excited, they tend to tweet rapidly for a while. When they are sedate, they tend to tweet at a lower frequency for several messages.



Hypothesis

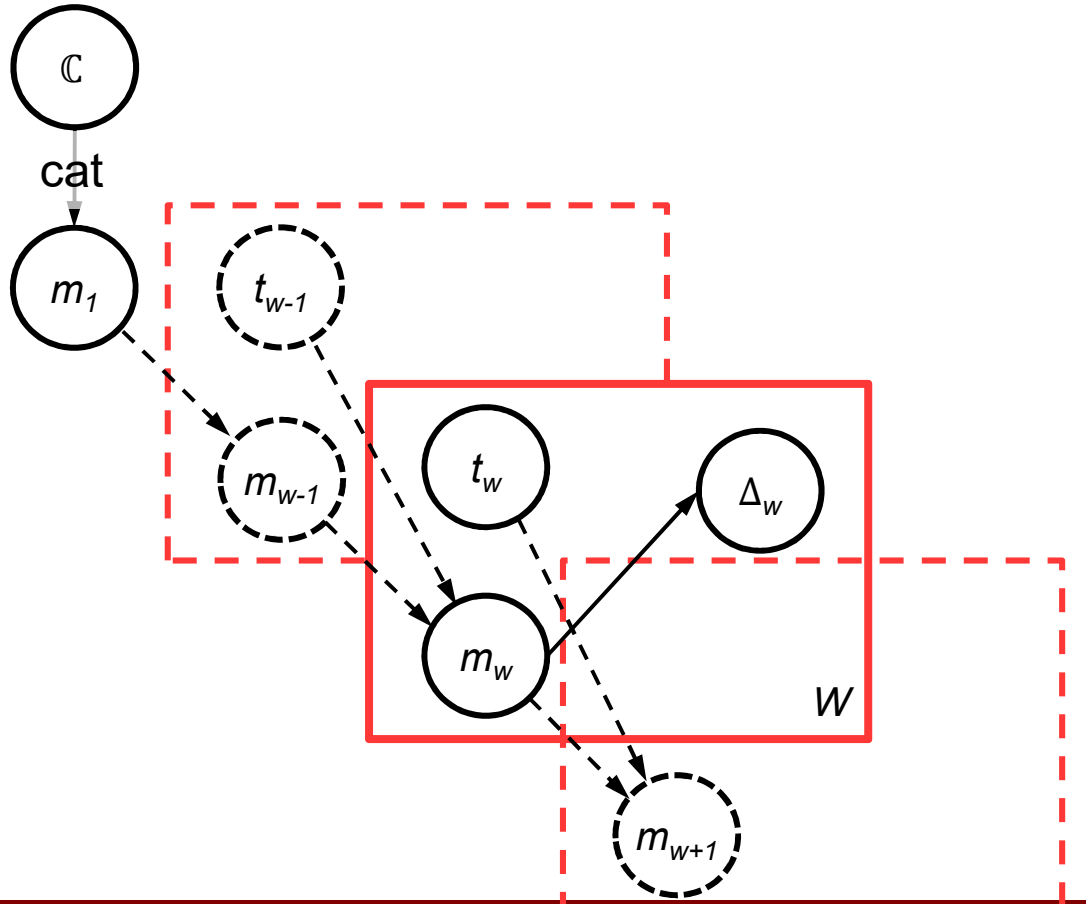
We'd like to
- formalize this pattern,
- quantify our confidence in it, &
- apply this insight.

To do this we chose to construct a Bayesian conditional probability model relating unknowns to observables.



Constructing the Model

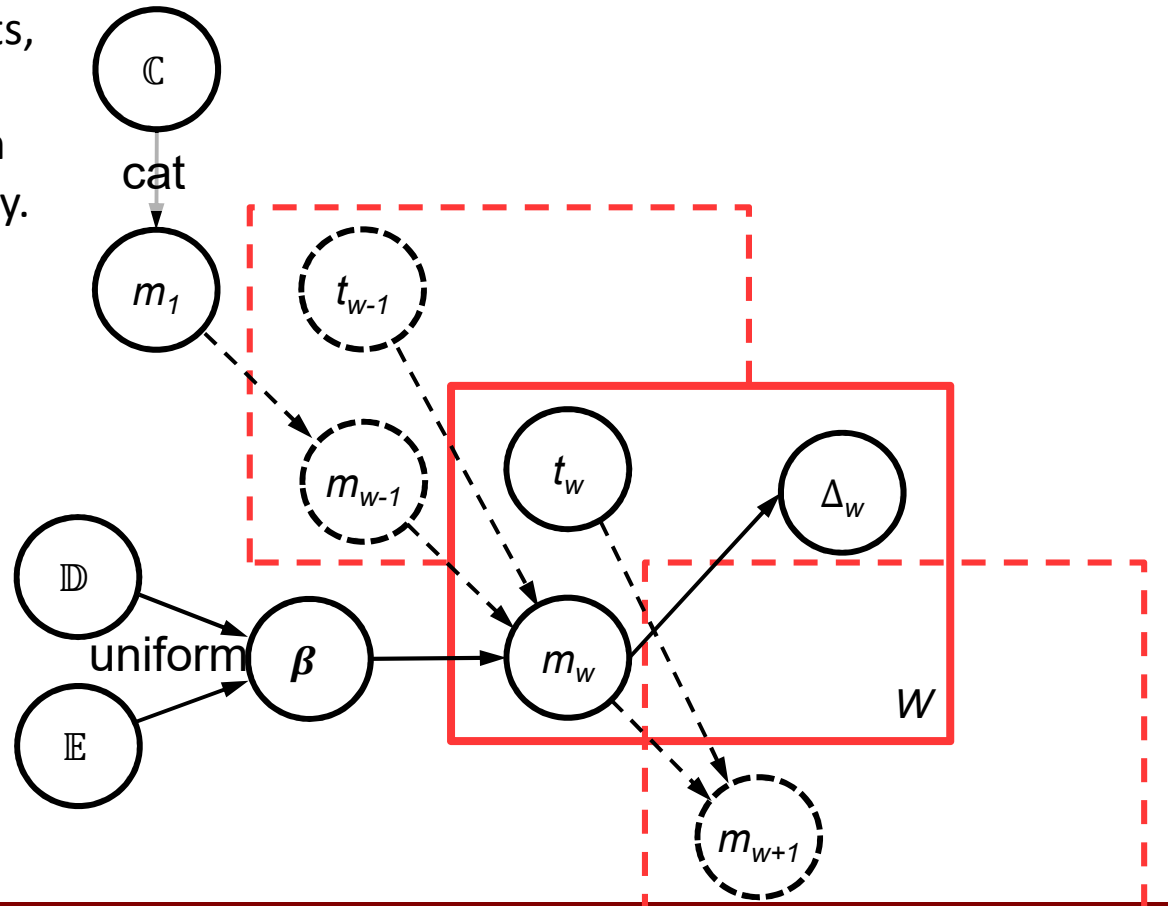
Consider a hidden Markov model, with a user that proceeds through different states of excitation.



Constructing the Model

We do not expect the current state to be sufficient.

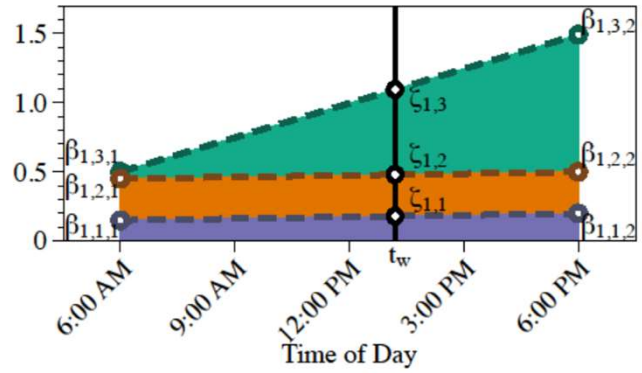
Many people have daily habits, and we extend the model to account for the time of day in the state transition probability.

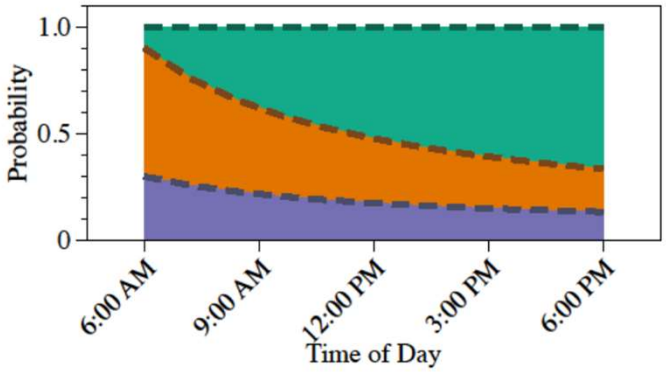


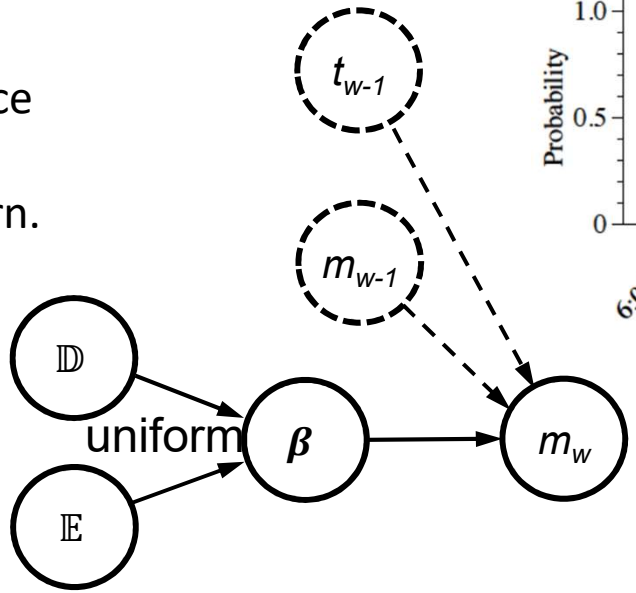
Testing

The state transition probability is a function of the current mode, and the relative proportion of linear functions of time of day.

More complex functions could be used. This on provides for some nonlinear behavior (once normalized) and has a small number of parameters to learn.



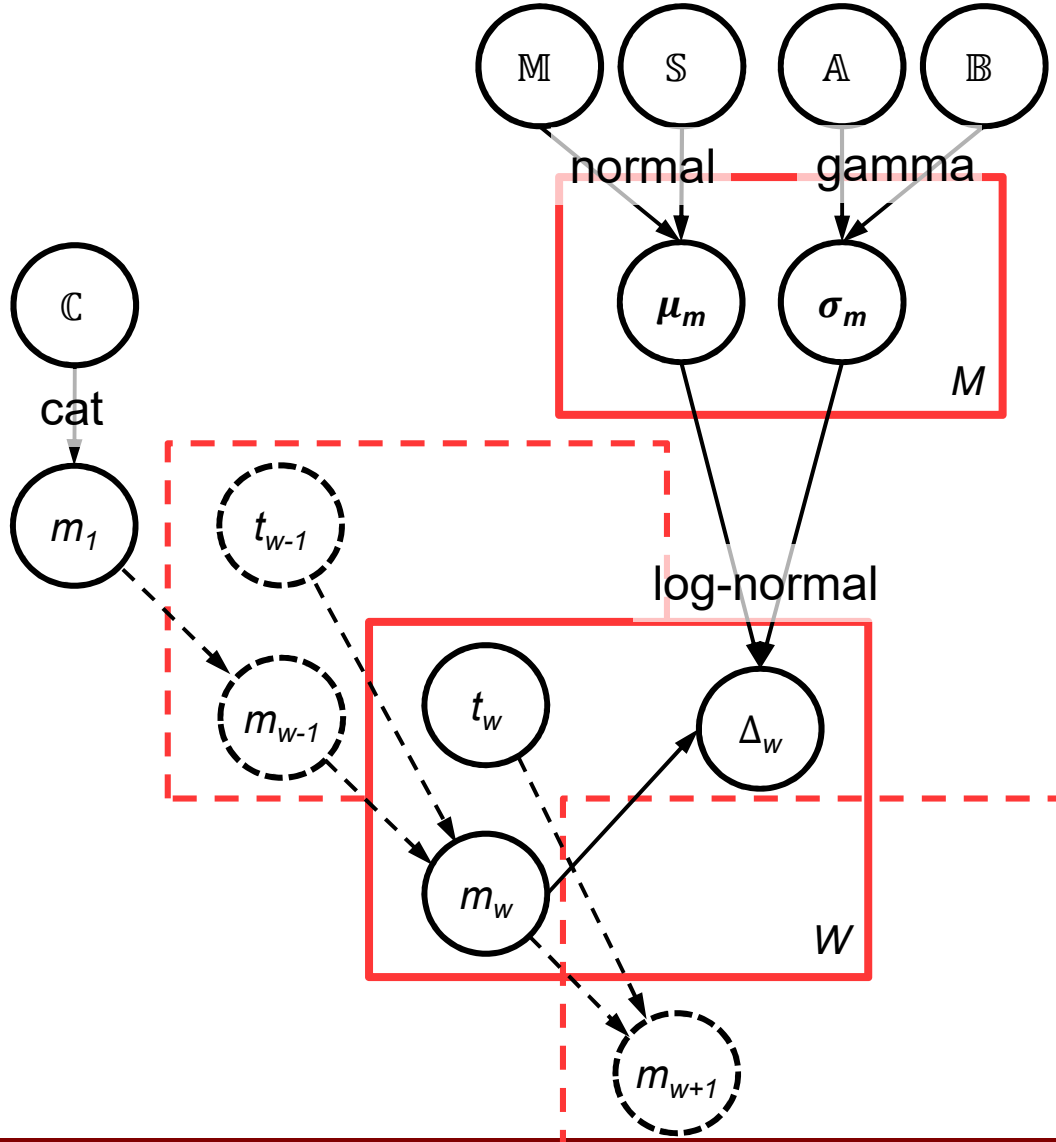




Constructing the Model

The likelihood of interval observations is a parameterized function of the mode.

Log-normal distributions are used with k unknown means & standard deviations.



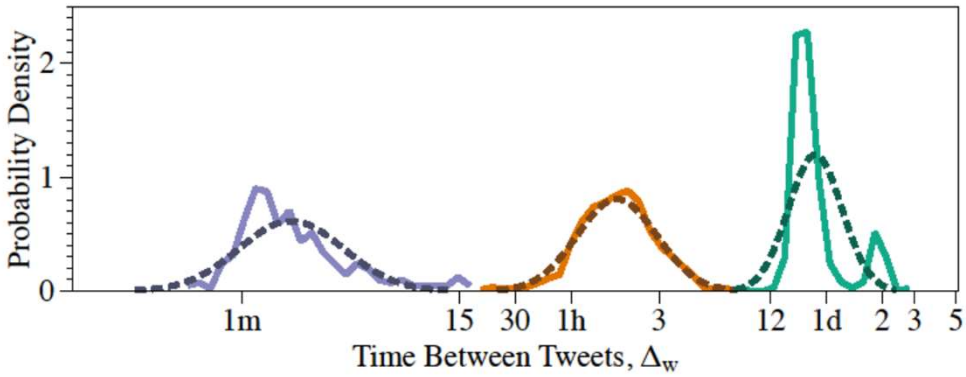
Testing

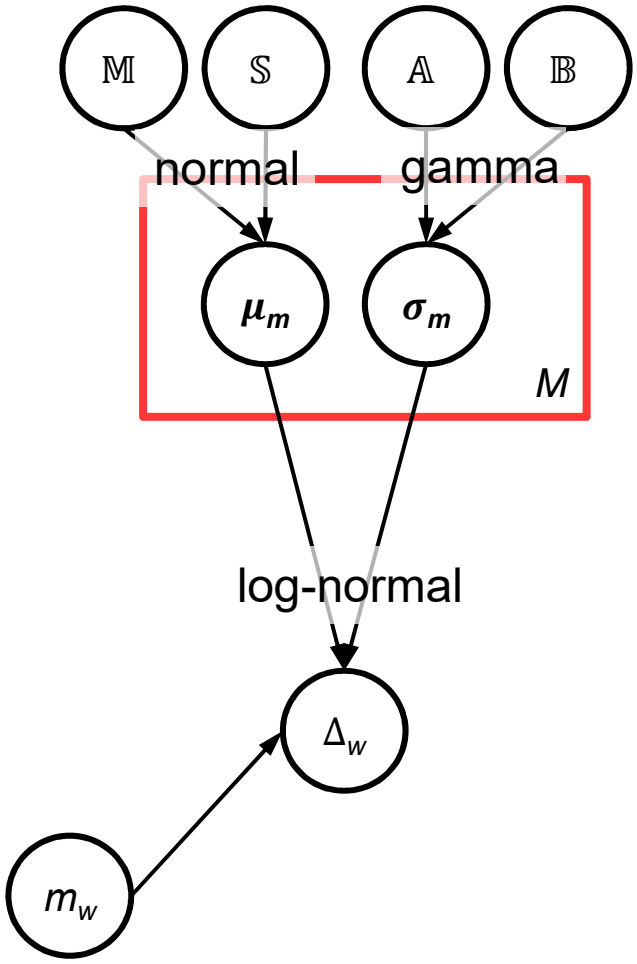
Log-normal was chosen after observation of interval distributions in several domains of human activity.

- Instant messenger traffic, Twitter
- Facebook
- Video attention

Mixture model fits data except…

- Very low energy modes are underrepresented
- Mode-to-mode sequence is not arbitrary

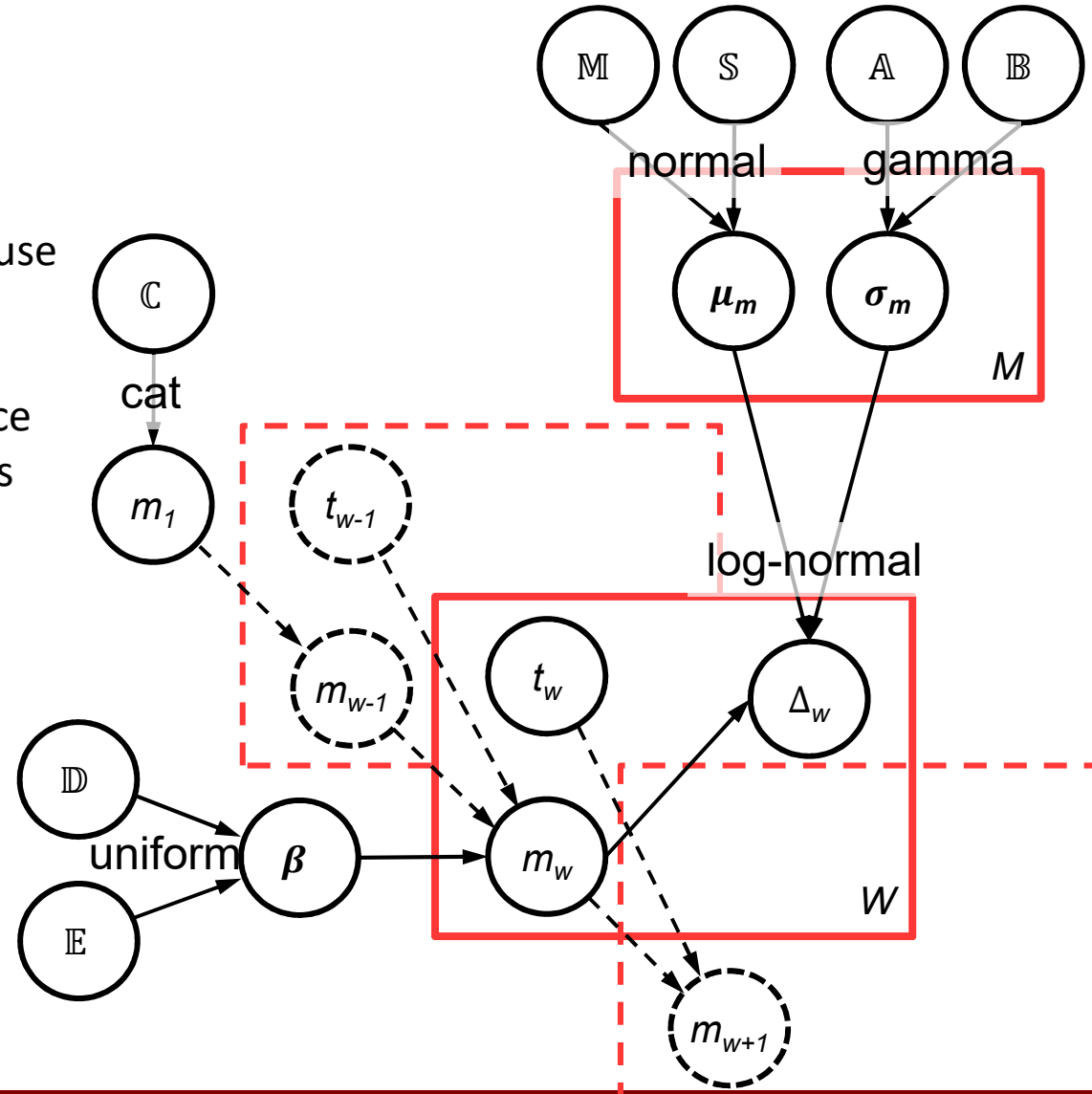




Constructing the Model

JAGS worked well for this problem.

- Discrete mode precluded use of HMC (via Stan)
- Built model in pieces
- Extended model to produce posterior predictive checks

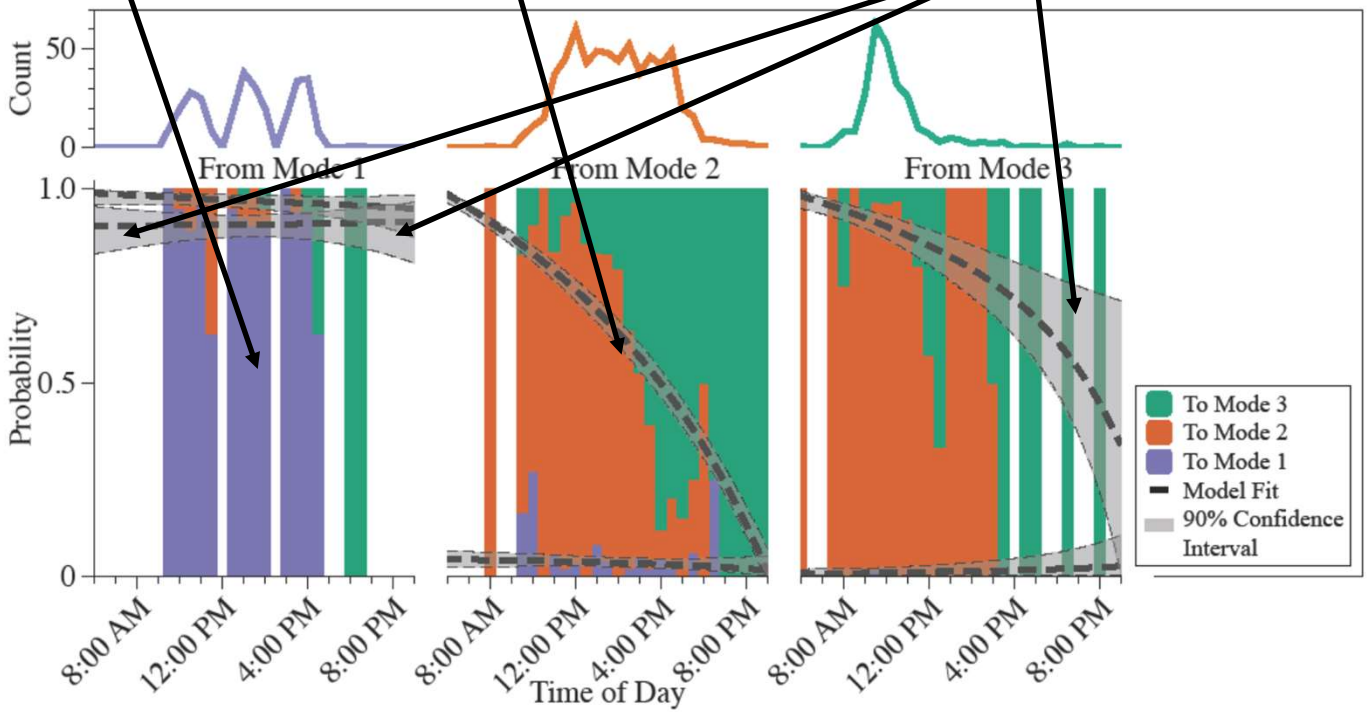


Posterior Predictive Checks

High-energy modes were more likely to be sustained.

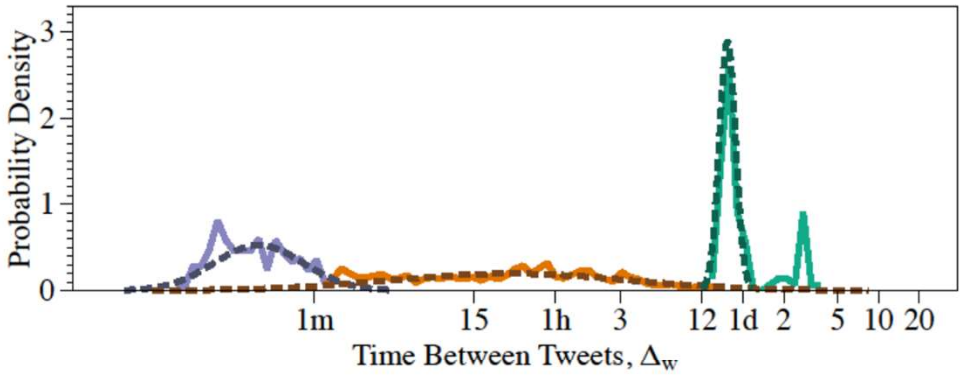
Lower-energy mode transition probabilities shifted throughout the day.

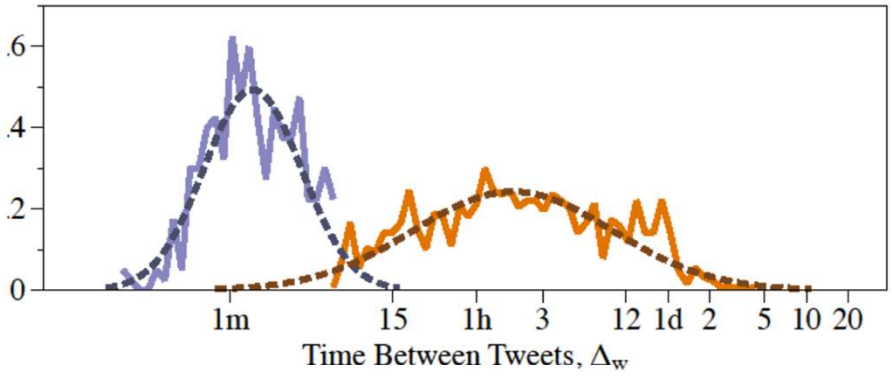
Posterior uncertainty reflected modes & times of day that were underrepresented in the data.

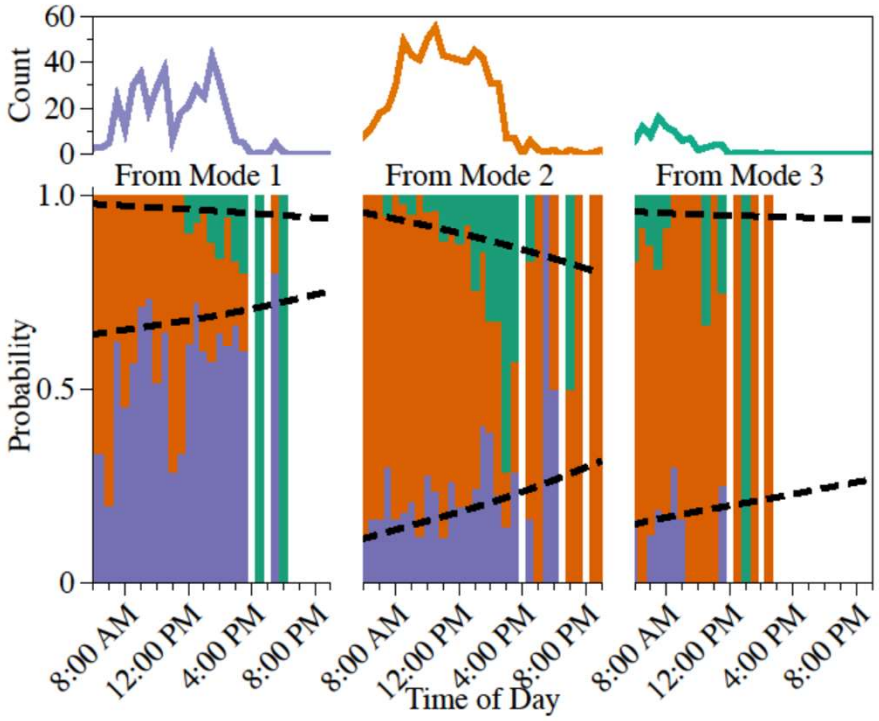


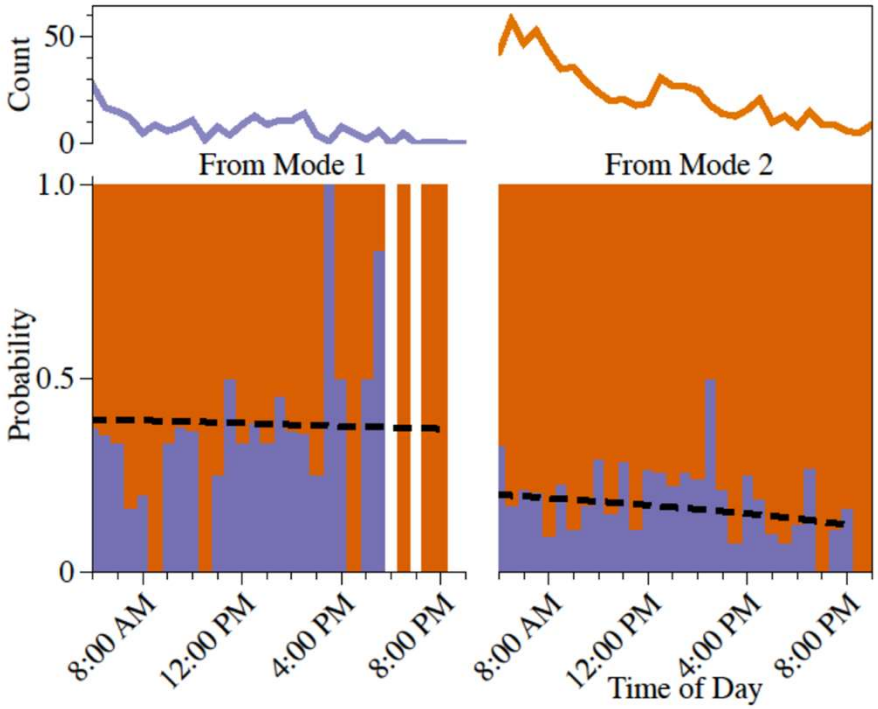
Posterior Predictive Checks

Sandia National Laboratories







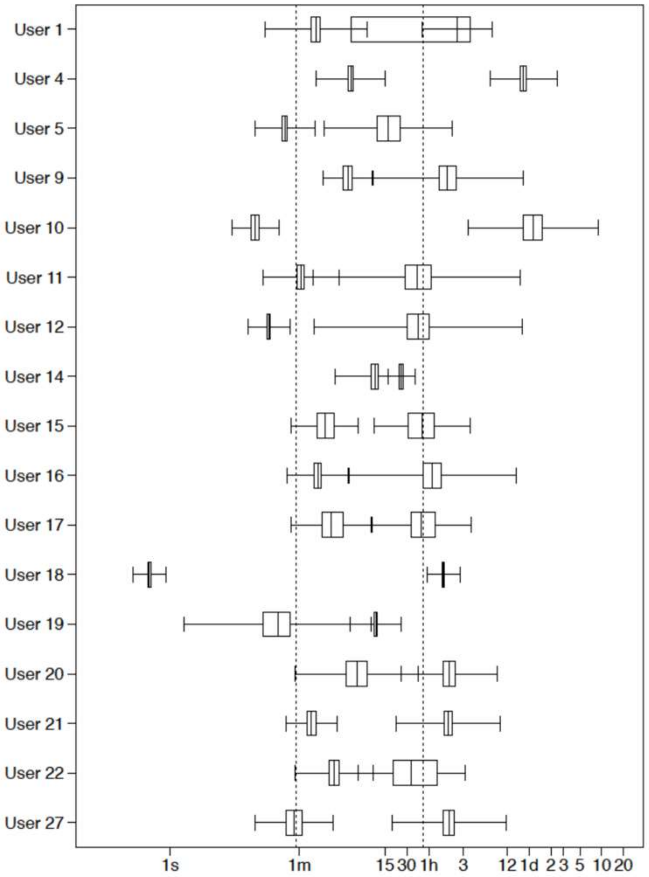


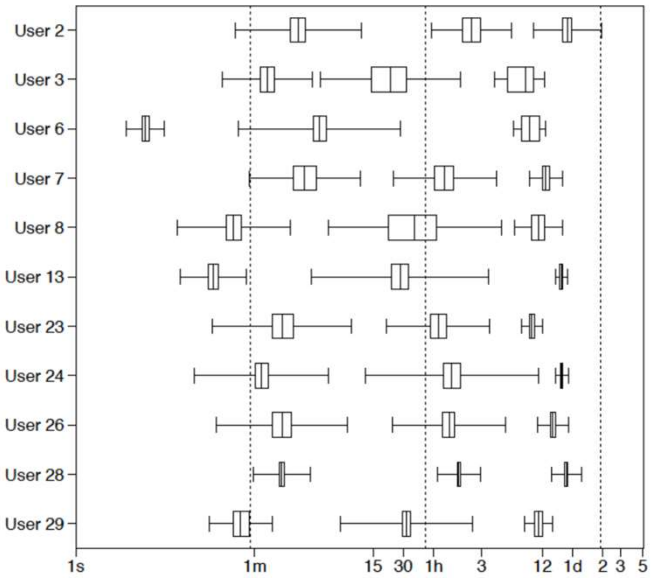
Sensitivity to Priors

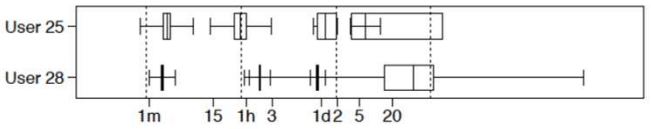
Sandia National Laboratories

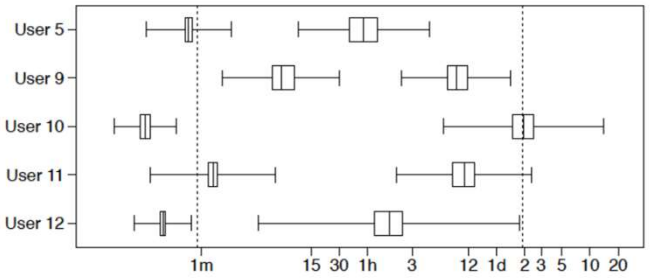
Checking a model with different priors is important.

We found our model was somewhat sensitive to prior mean/stddev (but not too badly).









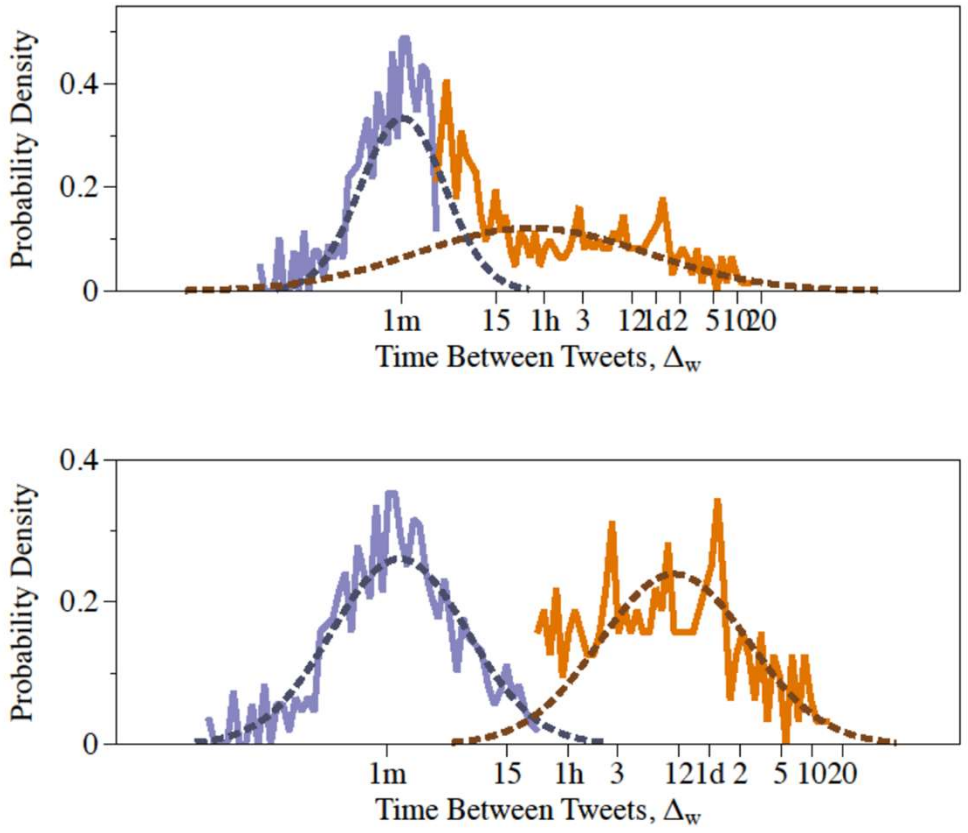
Sensitivity to Priors

Sandia National Laboratories

In our system, sensitivity to priors had the effect of biasing the posterior mode assignments.

In this example, the equal probability point of assigning mode 1 or 2 to a message shifts from about 10 minutes to about 30 minutes.

More data would overcome the choice of prior, or the prior uncertainty in the means and stddevs could be increased.



Conclusions

Sandia National Laboratories

- Model converges well
- Model uncertainty appears in reasonable places
- Model demonstrates flexibility when fit to diverse users
- Priors are stronger than we'd like
- Current model works well enough for some applications
  - Multi-agent simulation for disease spread modeling
  - Market segmentation
  - Bot detection

Next Steps

Sandia National Laboratories

- Model extensions
  - Better transition function representation
    - Periodic instead of a fixed window
  - Time zone inference
  - Accounting for topic & other evidence conditioned on mode
  - Estimating ‘k’
  - Population parameter distributions
  - Perplexity/data likelihood comparisons vs. competing models
- Application extensions
  - Market segmentation (distinguishing bots, people, and organizations)
  - Load forecasting (relating social media activity to power demand)
  - Network security (repurposing model to handle network traffic)

Questions?

Sandia National Laboratories

- Thank you for your time.
- Those interested in this work may reach me at helink@sandia.gov