

Online News Comments: Exploring, Modeling, and Predicting



Manos Tsagkias, Wouter Weerkamp, Maarten de Rijke
ISLA, University of Amsterdam



Task

Predict the number of comments a news article will receive, having seen an initial number.

RQ #1 What are the dynamics of news comments?

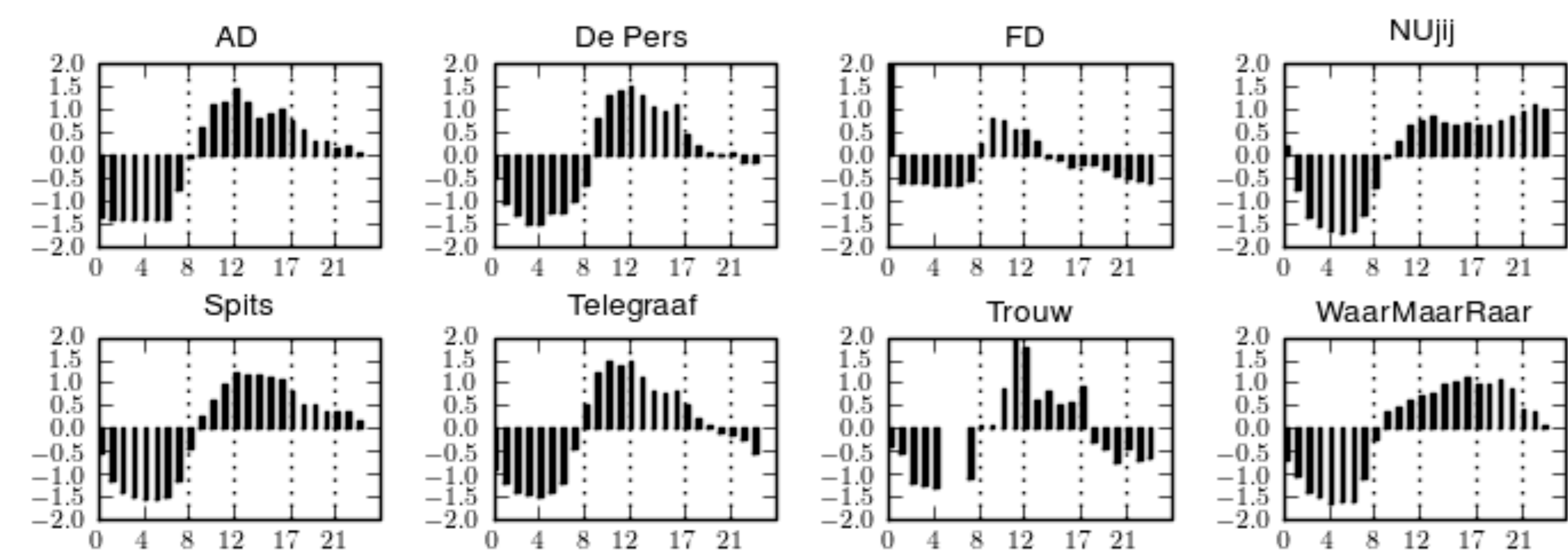
RQ #2 What is a distribution model for news comments?

RQ #3 Can we exploit the correlation between number of comments at early and later time to predict number of comments?

Dataset

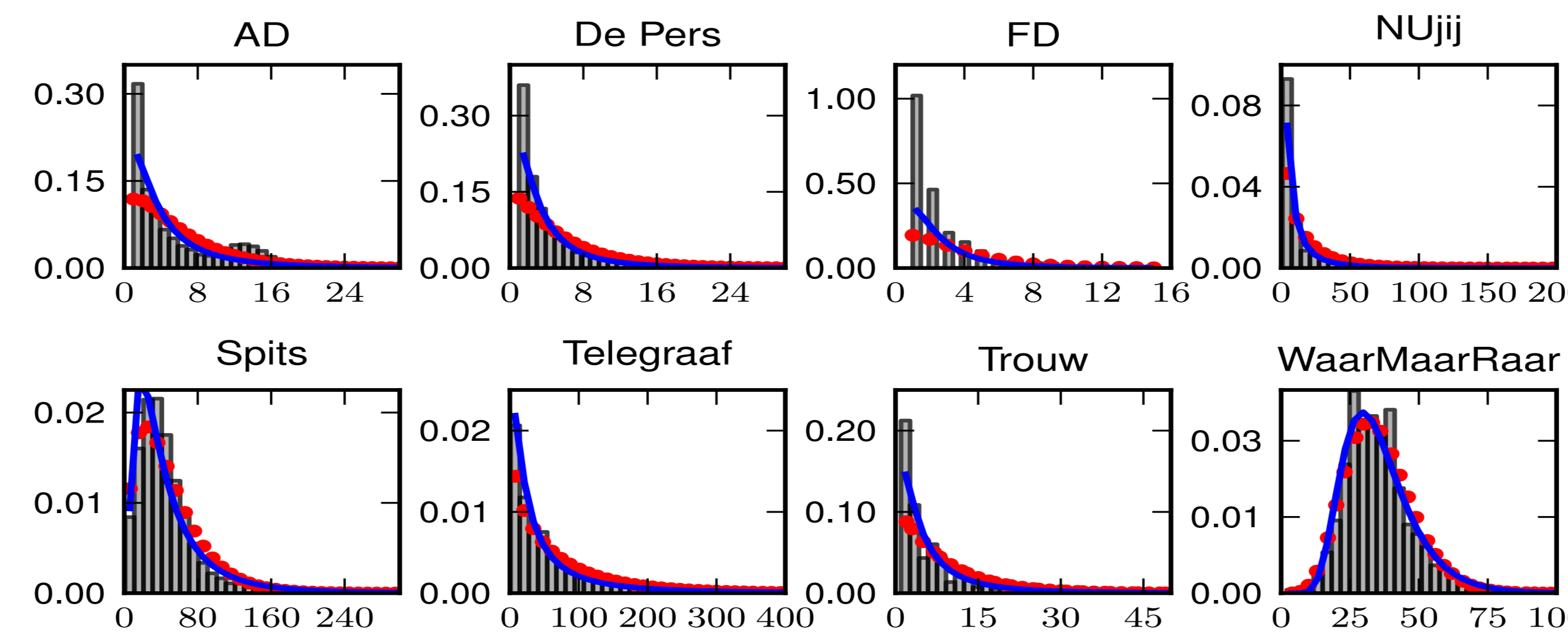
7 Dutch online news agents, 1 collaborative news platform. Date range: November 2008-April 2009. ~300K news articles, ~2M news comments.

Exploring News Comments



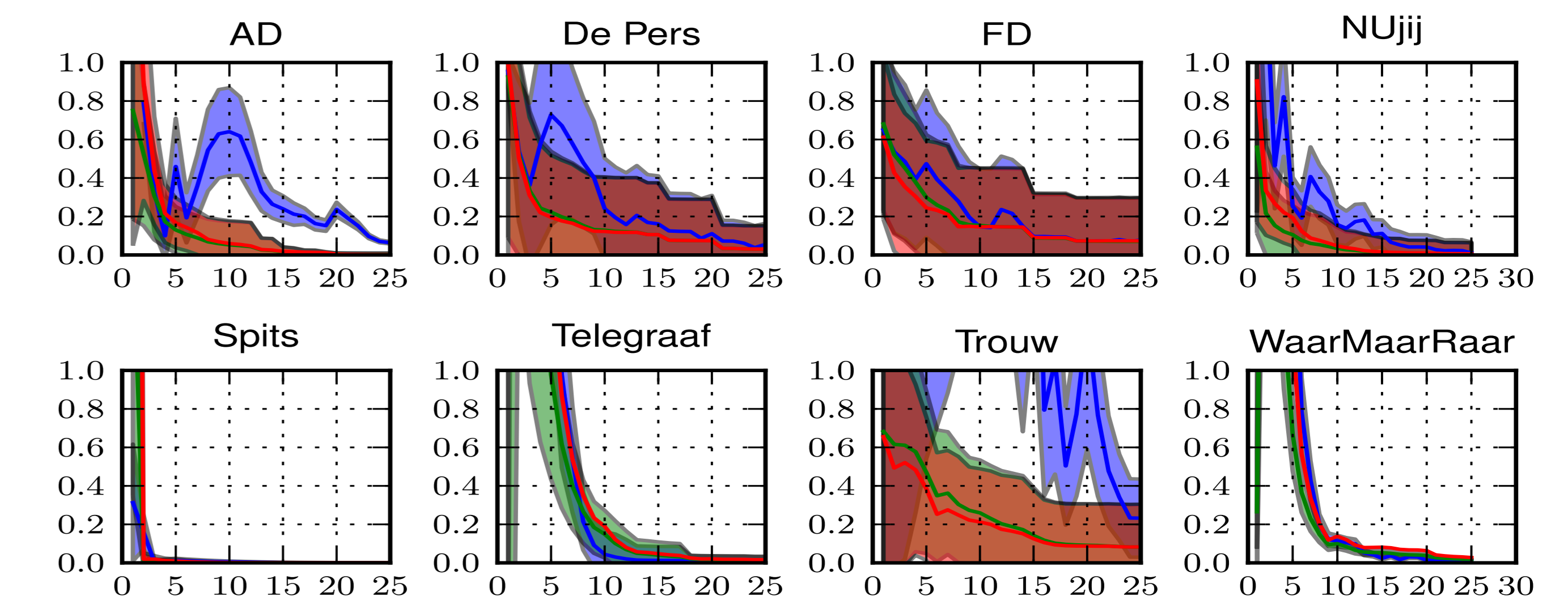
News comment per hour and per source. x-axis shows the time of day, and y-axis represents news comments as z-scores.

Modeling News Comments



PDF of a news article (y-axis) to receive X comments (x-axis). Gray bars show the observed data, the red dots show negative binomial fit, and blue line shows the log-normal fit.

Predicting News Comments



Relative squared error (y-axis) curves for different t_i (x-axis). M1 is trained on only early high commented stories, M2 uses all stories and has fixed slope at 1, M3 uses all stories and the slope is free parameter.

News Comments vs Blog Post Comments

- more news articles are commented (25% vs. 15%).
- but reaction time is slower (6hrs vs. 2hrs).

Temporal Patterns

Monthly: March highest, November low.

Day of the week: weekdays more active.

Hourly: correlation with sleeping/awake time and working, lunch and dinner time.

Source Time Transformation

1 source hour equals the average number of comments entering a news agent system per hour.

- Helps to define "volume" across sources.
- Model the probability of a news article to receive X many comments.
- 2 distributions: **log-normal** and **negative binomial**.
- χ^2 goodness of fit test accepts both distributions.

Linear model on logarithmic scale:

$$\dot{N}_s(t_i, t_r) = \exp\left[\ln(\alpha_0 N_s(t_i)) + \beta_0(t_i) + \frac{\sigma^2}{2}\right]$$

- Parameter estimation on train set, evaluation on test set.
- Report on *related squared error* averaged over all stories.
- M2 shows QRE < 0.2 for less than 10 shrs of observation.

Outlook and Future Work

- Commenting temporal cycles and commenting behavior show to be similar across news agents.
- Both log-normal and negative binomial can model news comments.
- Long-term prediction with small error rate after 10 shrs.
- Confirm findings in other geographic regions and languages.
- How the model can help designers to make more attractive news pages.