Supplementary Material for "The Bayesian Echo Chamber"

Fangjian Guo
Duke University
Durham, NC, USA
guo@cs.duke.edu

Charles Blundell
Gatsby Unit, UCL
London, UK
c.blundell@gatsby.ucl.ac.uk

Hanna Wallach
Microsoft Research
New York, NY, USA
wallach@microsoft.com

Katherine Heller
Duke University
Durham, NC, USA
kheller@stat.duke.edu

1 INFLUENCE VIA TURN-TAKING

In this section, we provide appropriate priors and details of an inference algorithm for the variant of Blundell et al.'s model [2012] described in section 2 of the paper. For real-world group discussions, the utterance start times $\mathcal{T} = \{\mathcal{T}^{(p)}\}_{p=1}^{P}$ and durations $\mathcal{D} = \{\{\Delta t_n^{(p)}\}_{n=1}^{N^{(p)}(T)}\}_{p=1}^{P}$ are observed, while parameters $\Theta = \{\lambda_0^{(p)}, \{\nu^{(qp)}\}_{q\neq p}, \tau_T^{(p)}\}_{p=1}^{P}$ are unobserved; however, information about the values of these parameters can be quantified via their posterior distribution given $\mathcal{T}$ and $\mathcal{D}$, obtained via Bayes' theorem, i.e.,

$$P(\Theta \,|\, \mathcal{T}, \mathcal{D}) \propto P(\mathcal{T} \,|\, \Theta, \mathcal{D})\, P(\Theta). \tag{1}$$

The likelihood term has the form

$$P(\mathcal{T} \,|\, \Theta, \mathcal{D}) = \prod_{p=1}^{P} \left(\exp\left(-\Lambda^{(p)}(T)\right) \prod_{n=1}^{N^{(p)}(T)} \lambda^{(p)}(t_n^{(p)}) \right), \tag{2}$$

where $\Lambda^{(p)}(T) = \int_0^T \lambda^{(p)}(t)\,\mathrm{d}t$ is the expected total number of utterances made over the entire observation interval from 0 to T [Daley and Vere-Jones, 1988].

Like Blundell et al., we place an improper prior over $\lambda_0^{(p)} > 0$. We also use priors to ensure that the multivariate Hawkes process is stationary. Specifically, we employ the stationarity condition of Bremaud and Massouli [1996]. If $\boldsymbol{M}$ is a $P \times P$ matrix given by

$$M^{(qp)} = \int_u^\infty \left| g^{(qp)}(t,u) \right| \mathrm{d}t = \nu^{(qp)} \tau_T^{(p)}, \tag{3}$$

then this condition requires the spectral radius of $\boldsymbol{M}$ to be strictly less than one. This condition is not straightforward to enforce with tractable constraints; however, since the spectral radius of $\boldsymbol{M}$ is upper-bounded by any matrix norm, the condition may be enforced by requiring that $\|\boldsymbol{M}\| < 1$ for any norm $\|\cdot\|$. We use the maximum absolute column sum norm:

$$\|\boldsymbol{M}\|_{1\to 1} = \max_{\|x\|_1 = 1} \|\boldsymbol{M}x\|_1 \tag{4}$$

$$= \max_{p=1,\cdots,P} \tau_T^{(p)} \sum_{q\neq p} \nu^{(qp)}. \tag{5}$$

Rewriting this expression implies an improper joint prior over $\{\tau_T^{(p)}\}_{p=1}^{P}$ and $\{\{\nu^{(qp)}\}_{q\neq p}\}_{p=1}^{P}$ in which

$$0 < \tau_T^{(p)} < \frac{1}{\sum_{q\neq p} \nu^{(qp)}} \text{ and} \tag{6}$$

$$0 < \nu^{(qp)} < \frac{1}{\tau_T^{(p)} - \sum_{r\neq q, r\neq p} \nu^{(rp)}}. \tag{7}$$

Although the resultant posterior distribution $P(\Theta \,|\, \mathcal{T}, \mathcal{D})$ is analytically intractable, posterior samples can be drawn using either the conditional intensity function approach or the cluster process approach described by Rasmussen [2013]. Like Blundell et al., we take the former approach and use a slice-within-Gibbs algorithm [Neal, 2003] that sequentially samples each parameter from its conditional posterior.

This slice-within-Gibbs algorithm requires frequent evaluation of the likelihood in equation 2; however, the computational cost can be reduced by noting that the product over rate functions can be efficiently computed using the following recurrence relation:

$$\lambda^{(p)}(t_n^{(p)}) = \lambda_0^{(p)} + \left(\lambda^{(p)}(t_{n-1}^{(p)}) - \lambda_0^{(p)}\right) \exp\left(-\frac{t_n^{(p)} - t_{n-1}^{(p)}}{\tau_T^{(p)}}\right) + \sum_{q\neq p} \sum_{m: t_{n-1}^{(p)} \leq t'^{(q)}_m < t_n^{(p)}} \nu^{(qp)} \exp\left(-\frac{t_n^{(p)} - t'^{(q)}_m}{\tau_T^{(p)}}\right)$$

for $n = 2, 3, \ldots, N^{(p)}(T)$. The initial term is

$$\lambda^{(p)}(t_1^{(p)}) = \lambda_0^{(p)} + \sum_{q\neq p} \sum_{m: t'^{(q)}_m < t_1^{(p)}} \nu^{(qp)} \exp\left(-\frac{t_1^{(p)} - t'^{(q)}_m}{\tau_T^{(p)}}\right).$$

2 INFLUENCE VIA LINGUISTIC ACCOMMODATION

In this section, we provide a directed graphical model, appropriate priors, and details of an inference algorithm for our model, the Bayesian Echo Chamber.

The likelihood term implied by our model is

$$P(\mathcal{W} \mid \Theta, \mathcal{T}, \mathcal{D}) = \prod_{p=1}^{P} \prod_{n=1}^{N^{(p)}(T)} P(\boldsymbol{w}_n^{(p)} \mid \{\{\boldsymbol{w}_m^{(q)}\}_{m:t'^{(q)}_m < t_n^{(p)}}\}_{q \neq p}, \Theta).$$

A directed graphical model depicting the structure of $P(\boldsymbol{w}_n^{(p)} \mid \{\{\boldsymbol{w}_m^{(q)}\}_{m:t'^{(q)}_m < t_n^{(p)}}\}_{q \neq p}, \Theta)$ is in figure 1.

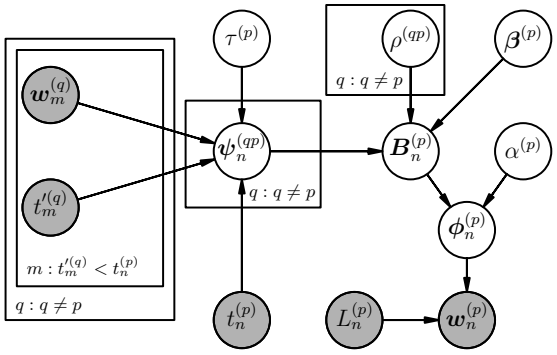


Figure 1: Directed Graphical Model Depicting the Structure of $P(\boldsymbol{w}_n^{(p)} \mid \{\{\boldsymbol{w}_m^{(q)}\}_{m:t'^{(q)}_m < t_n^{(p)}}\}_{q \neq p}, \Theta)$.

We place a gamma prior over $\rho^{(qp)}$, with a shape parameter chosen to encourage shrinkage towards zero. Due to the additive nature of $\boldsymbol{B}_n^{(p)}$, the value of $\beta_v^{(p)}$ should be comparable in magnitude to $\sum_{q \neq p} \rho^{(qp)} \psi_{v,n}^{(qp)}$. We therefore place a gamma prior over each $\beta_v^{(p)}$, with shape and scale parameters chosen to yield this property for real-world data sets. We also place broad gamma priors over $\alpha^{(p)}$ and $\tau_L^{(p)}$. In practice, inference is insensitive to the specific values of the shape and scale parameters of these priors, provided they are broad. For our experiments, we used $\alpha^{(p)} \sim \text{Gamma}\,(10, 10)$, $\beta_v^{(p)} \sim \text{Gamma}\,(10, 20)$, $\rho^{(qp)} \sim \text{Gamma}\,(1, 2)$, and $\tau^{(p)} \sim \text{Gamma}\,(10, 10)$.

Although the resultant posterior distribution $P(\Theta \mid \mathcal{W}, \mathcal{T}, \mathcal{D})$ is intractable, posterior samples of $\{\alpha^{(p)}, \boldsymbol{\beta}^{(p)}, \{\rho^{(qp)}\}_{q \neq p}, \tau_L^{(p)}\}_{p=1}^{P}$ can be drawn using a collapsed[1] slice-with-Gibbs algorithm that sequentially samples each parameter from its conditional posterior. Pseudocode for this approach is given in algorithm 1. Each parameter is sampled in a univariate fashion, except for $\boldsymbol{\beta}^{(p)}$, which is drawn using multivariate slice sampling with the hyperractangle method [Neal, 2003]. To improve mixing, we drew ten samples of $\boldsymbol{\beta}^{(p)}$ during each Gibbs sweep. When implemented in Python, we were able to draw 4,000 posterior samples (including 1,000 burn-in samples) of $\{\alpha^{(p)}, \boldsymbol{\beta}^{(p)}, \{\rho^{(qp)}\}_{q \neq p}, \tau_L^{(p)}\}_{p=1}^{P}$ in at most a couple of hours for all data sets used in our experiments.

[1]Probability vectors $\{\{\boldsymbol{\phi}_n^{(p)}\}_{n=1}^{N^{(p)}(T)}\}_{p=1}^{P}$ can be integrated out using Dirichlet–multinomial conjugacy.

Algorithm 1 Inference Algorithm

```
for i = 1, 2, ··· , I do
    for p = 1, 2, ··· , P do
        Slice sample α^(p)
        Slice sample τ_L^(p)
        for q ≠ p do
            Slice sample ρ^(qp)
        end for
        for j = 1, 2, ··· , 10 do
            Slice sample β^(p) (multivariate)
        end for
    end for
end for
```

3 EXPERIMENTS

The salient characteristics of all data sets used in our experiments are provided in table 1. For each data set obtained from TalkBank [MacWhinney, 2007], the "TalkBank" column contains the data set identifier within the "Meetings" section of the TalkBank database. The "No. Tokens" column indicates the total number of tokens in each data set after restricting the vocabulary to the $V = 600$ most frequent stemmed types. The "Tokens Removed" column contains the percentage of tokens that were discarded via this step.

Table 2 contains predictive log probabilities for several additional data sets. The "Family Discussion" and "University Lecture" data sets are conversation transcripts from the Santa Barbara Corpus of Spoken American English [MacWhinney, 2007]. These data sets capture the back-and-forth of real-world conversations. The "January 29, 2008 FOMC Meeting" data set is one of the FOMC meeting transcripts used our exploratory analysis. The salient characteristics of these data sets are given in table 1. For all but one of these additional data sets, the Bayesian Echo Chamber out-performed a unigram language model and Blei and Lafferty's dynamic topic model [2006] by predicting higher probabilities of held-out data for both a 90%–10% and an 80%–20% training–testing split.

Posterior means and standard deviations of the influence parameters $\{\{\rho^{(qp)}\}_{q \neq p}\}_{p=1}^{P}$ inferred from the DC

Table 1: Salient Characteristics of Data Sets.

Data Set	TalkBank	No. People	No. Utterances	No. Tokens	Tokens Removed
Synthetic	–	3	300	15,070	0.00%
University Lecture	SB/12	5	138	3,482	4.42%
Birthday Party	SB/49	8	454	4,229	5.88%
DC v. Heller	SCOTUS/07-290	10	365	15,104	7.21%
L&G v. Texas	SCOTUS/02-102	6	200	8,573	5.47%
Citizens United v. FEC	SCOTUS/08-205b	10	345	12,700	7.41%
12 Angry Men	–	12	312	6,350	5.25%
January 29, 2008 FOMC Meeting	–	4	101	13,505	13.74%

Table 2: Additional Predictive Log Probabilities of Held-Out Data.

	10% Test Set			20% Test Set		
Data Set	Our Model	Unigram	DTM	Our Model	Unigram	DTM
University Lecture	-528.23±0.06	-541.23±0.05	**-520.74**	**-1972.67**±0.13	-2009.62±0.12	-2110.66
Birthday Party	**-1883.45**±0.11	-1961.4±0.11	-1900.68	**-4384.42**±0.16	-4625.57±0.20	-4498.467
January 29, 2008 FOMC Meeting	**-3187.73**±0.04	-3338.59±0.10	-3211.09	**-17342.43**±0.21	-17779.01±0.24	-17726.64

v. Heller Supreme Court case using our model are given in tables 3 and 4, respectively. These values were obtained using 3,000 samples from the posterior distribution. To further illustrate posterior uncertainty, influence networks drawn using 25%, 50% (i.e., median), and 75% posterior quantiles are shown in figure 2. These networks look very similar to each other.

Posterior means and standard deviations of the influence parameters $\{\{\rho^{(qp)}\}_{q\neq p}\}_{p=1}^{P}$ inferred from "12 Angry Men" using our model are provided in tables 5 and 6, respectively. These values were obtained using 3,000 samples from the posterior distribution. To further illustrate posterior uncertainty, influence networks drawn using 25%, 50%, and 75% posterior quantiles are provided in figure 3. As with the DC v. Heller case, these networks look very similar to one another.

Log probabilities, obtained using a 90%–10% training–testing split and a vocabulary of $V = 300$ types, are provided for the tied and untied combined models in table 7. The tied model, whose likelihood is the product of the Bayesian Echo Chamber's likelihood and that of Blundell et al.'s model but with shared influence parameters, assigned lower probabilities to held-out data than the fully factorized (i.e., untied) model.

Data Set	Tied	Untied
L&G v. Texas	-5507.11±0.15	**-5502.87**±0.15
DC v. Heller	-6321.30±0.16	**-6303.55**±0.15
Citizens United v. FEC	-4795.24±0.18	**-4777.96**±0.17
"12 Angry Men"	-4014.56±0.24	**-3987.20**±0.23

Table 7: Log Probabilities of Held-Out Data for the Combined Model with Tied and Untied Parameters.

Acknowledgements

Thanks to Juston Moore and Aaron Schein for their work on early stages of this project, and to Aaron for the "Bayesian Echo Chamber" model name. This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #IIS-1320219, and in part by NSF grant #SBE-0965436. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*.

Blundell, C., Heller, K. A., and Beck, J. (2012). Modelling reciprocating relationships with Hawkes processes. In *Advances In Neural Information Processing Systems*.

Bremaud, P. and Massouli, L. (1996). Stability of nonlinear Hawkes processes. *The Annals of Probability*, pages 1563–1588.

Daley, D. J. and Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes*. Springer.

MacWhinney, B. (2007). The TalkBank project. *Creating and Digitizing Language Corpora: Synchronic Databases*.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, 31(3):705–767.

Rasmussen, J. G. (2013). Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642.

Table 3: Posterior Means of $\{\{\rho^{(qp)}\}_{q\neq p}\}_{p=1}^{P}$ Inferred from the DC v. Heller Case.

	To									
From	DELLI	GURA	ROBE	CLEME	STEV	SCAL	KENN	GINS	SOUT	BREY
DELLI	–	65.85	109.92	109.36	86.78	125.29	143.29	71.21	82.77	72.98
GURA	10.18	–	4.79	2.43	7.75	3.27	2.62	3.17	6.84	5.08
ROBE	161.29	37.99	–	6.44	3.76	5.12	4.77	7.62	3.99	7.67
CLEME	5.12	37.41	11.02	–	16.53	4.84	6.03	15.77	12.53	32.13
STEV	3.93	9.27	3.89	4.37	–	3.10	2.70	2.91	4.42	3.12
SCAL	50.53	15.45	7.77	5.62	4.64	–	3.41	6.04	7.17	6.83
KENN	180.91	2.90	5.86	50.67	13.75	4.93	–	4.50	5.53	5.63
GINS	6.98	9.91	11.29	4.55	3.22	4.08	2.69	–	2.95	3.68
SOUT	3.34	4.34	3.86	5.90	3.55	3.54	2.59	3.22	–	4.50
BREY	8.24	16.48	5.18	2.45	3.71	3.22	2.99	4.31	5.26	–

Table 4: Posterior Standard Deviations of $\{\{\rho^{(qp)}\}_{q\neq p}\}_{p=1}^{P}$ Inferred from the DC v. Heller Case.

	To									
From	DELLI	GURA	ROBE	CLEME	STEV	SCAL	KENN	GINS	SOUT	BREY
DELLI	–	7.36	11.08	8.99	10.30	12.25	12.73	10.12	10.94	9.22
GURA	6.12	–	3.88	2.34	5.90	3.07	2.51	3.12	5.20	4.39
ROBE	14.60	12.93	–	5.76	3.48	4.92	4.75	6.68	3.91	7.48
CLEME	4.69	6.32	7.11	–	9.73	4.45	5.15	9.35	8.49	8.81
STEV	3.65	7.56	3.68	4.16	–	3.12	2.73	2.99	4.40	3.08
SCAL	14.02	9.42	6.84	5.02	4.43	–	3.33	5.67	6.46	6.28
KENN	14.84	2.70	5.46	17.43	10.61	4.68	–	4.14	5.13	5.64
GINS	6.32	8.14	9.74	4.24	3.09	4.09	2.58	–	2.79	3.56
SOUT	3.52	4.11	3.75	5.85	3.72	3.25	2.47	3.16	–	4.86
BREY	7.19	8.40	4.89	2.43	3.58	3.05	2.95	3.91	4.53	–

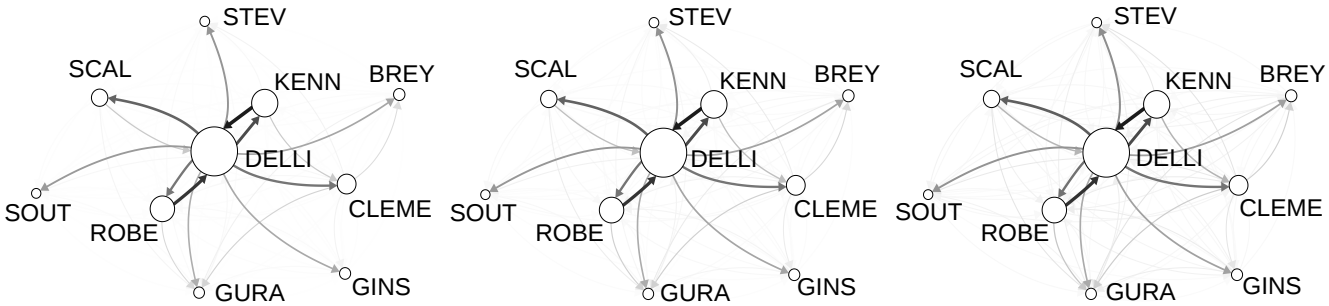


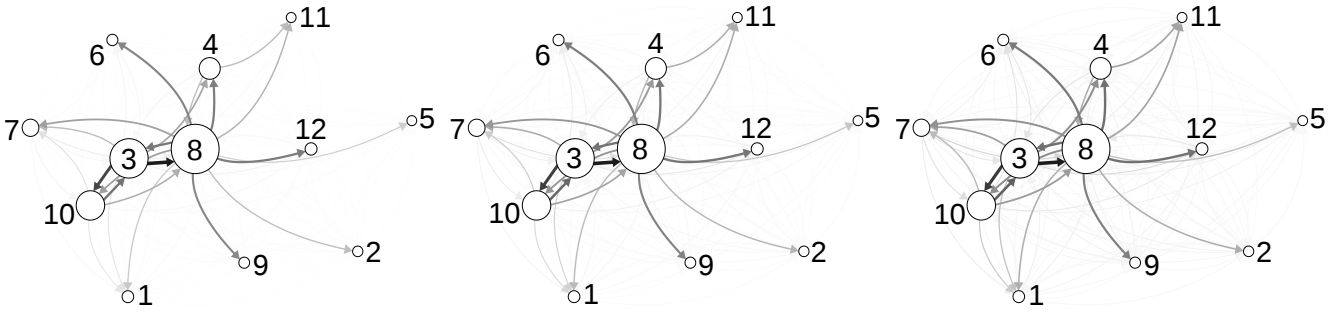
Figure 2: Influence Networks for the DC v. Heller Case Drawn Using (a) 25%, (b) 50%, and (c) 75% Quantiles.

Table 5: Posterior Means of $\{\{\rho^{(qp)}\}_{q\neq p}\}_{p=1}^{P}$ Inferred from "12 Angry Men."

	To											
From	Juror 8	Juror 3	Juror 10	Juror 7	Juror 1	Juror 4	Juror 6	Juror 11	Juror 12	Juror 9	Juror 2	Juror 5
Juror 8	–	82.09	61.82	61.08	39.80	80.38	75.30	48.96	80.12	72.86	43.58	33.98
Juror 3	134.35	–	112.40	47.76	27.56	59.83	10.62	6.46	16.85	5.28	5.51	6.84
Juror 10	53.38	88.01	–	30.65	24.54	4.71	12.72	3.41	9.48	4.08	4.25	5.50
Juror 7	19.56	11.53	13.97	–	8.24	3.69	5.19	3.35	4.46	4.25	4.86	4.08
Foreman	5.12	5.51	3.30	2.99	–	3.02	3.74	2.66	4.82	2.57	3.11	3.18
Juror 4	43.05	11.73	2.88	2.88	3.44	–	2.47	46.62	4.08	6.47	4.55	3.46
Juror 6	5.79	3.23	3.03	3.16	2.76	2.56	–	2.82	3.14	3.11	3.30	3.23
Juror 11	3.39	2.76	2.63	2.17	2.28	2.80	2.32	–	2.61	2.50	2.40	2.44
Juror 12	9.61	3.49	3.00	3.47	2.84	2.91	4.44	3.59	–	2.64	3.62	2.76
Juror 9	4.28	3.44	2.56	2.95	2.68	2.73	2.96	4.44	3.05	–	2.67	2.82
Juror 2	2.85	2.99	2.84	2.67	3.34	2.41	3.34	2.16	3.14	2.49	–	3.05
Juror 5	2.88	2.49	2.59	2.38	2.54	2.23	2.47	2.77	2.53	2.73	2.35	–

Table 6: Posterior Standard Deviations of $\{\{\rho^{(qp)}\}_{q\neq p}\}_{p=1}^{P}$ Inferred from "12 Angry Men."

	To											
From	Juror 8	Juror 3	Juror 10	Juror 7	Juror 1	Juror 4	Juror 6	Juror 11	Juror 12	Juror 9	Juror 2	Juror 5
Juror 8	–	13.10	14.50	14.09	14.12	14.54	13.52	12.53	13.51	11.50	12.35	11.31
Juror 3	15.21	–	18.19	17.49	16.01	16.63	8.71	5.76	12.62	4.86	5.00	6.17
Juror 10	14.47	16.84	–	17.76	14.06	4.36	10.20	3.33	8.20	3.82	4.22	5.15
Juror 7	12.34	10.15	10.36	–	7.30	3.44	4.91	3.09	4.66	4.08	4.57	3.80
Foreman	4.58	5.18	3.42	3.12	–	3.10	3.80	2.63	4.66	2.57	3.16	3.06
Juror 4	12.45	9.17	2.84	2.81	3.32	–	2.44	15.45	4.04	6.41	4.27	3.38
Juror 6	5.51	3.42	3.04	3.14	2.66	2.44	–	2.72	3.06	3.06	3.43	3.30
Juror 11	3.46	2.75	2.55	2.26	2.26	2.77	2.28	–	2.60	2.47	2.33	2.41
Juror 12	8.14	3.38	3.00	3.44	2.76	2.87	4.24	3.51	–	2.63	3.46	2.63
Juror 9	4.35	3.39	2.51	2.97	2.64	2.70	2.90	4.59	3.00	–	2.58	2.76
Juror 2	2.79	2.90	2.68	2.64	3.42	2.23	3.41	2.20	3.07	2.49	–	3.21
Juror 5	2.83	2.53	2.68	2.49	2.46	2.28	2.55	2.76	2.63	2.80	2.35	–



(a) (b) (c)

Figure 3: Influence Networks for "12 Angry Men" Drawn Using (a) 25%, (b) 50%, and (c) 75% Quantiles.