

# SourceSeer: Forecasting Rare Disease Outbreaks

## Using Multiple Data Sources - Supplementary Material

Theodoros Rekatsinas<sup>1</sup>, Saurav Ghosh<sup>2</sup>, Sumiko R. Mekar<sup>3</sup>, Elaine O. Nsoesie<sup>3</sup>  
John S. Brownstein<sup>3</sup>, Lise Getoor<sup>4</sup>, Naren Ramakrishnan<sup>2</sup>

thodrek@cs.umd.edu, sauravcsvt@cs.vt.edu

{sumiko.mekaru, elaine.nsoesie, john.brownstein}@childrens.harvard.edu,

getoor@soe.ucsc.edu, naren@cs.vt.edu

<sup>1</sup>University of Maryland, <sup>2</sup>Virginia Tech <sup>3</sup>Boston Children's Hospital, <sup>4</sup>University of California, Santa Cruz

### 1 Gibbs Sampling for STAT

We provide the detailed derivation of the Gibbs sampling algorithm in Section 3. The joint distribution corresponding to the topic model is:

$$\begin{aligned} \Pr(\mathbf{w}, \mathbf{t}, \mathbf{l}, \mathbf{z}, \phi, \theta, \xi; \alpha, \beta, \gamma) = \\ &= \prod_{z=1}^K \Pr(\phi_z; \beta) \Pr(\xi_z; \gamma) \prod_{l=1}^L \Pr(\theta_l; \alpha) \\ &\cdot \prod_{s=1}^S \prod_{i=1}^{N_s} \Pr(z_{si} | l_{si}, \theta_l) \Pr(w_{si} | \phi_{z_{si}}) \Pr(t_{si} | \xi_{z_{si}}) \end{aligned}$$

Next, we marginalize over all  $\phi$ ,  $\xi$  and  $\theta$ . We have:

$$\begin{aligned} \Pr(\mathbf{w}, \mathbf{t}, \mathbf{l}, \mathbf{z}; \alpha, \beta, \gamma) \\ &= \int_{\phi} \int_{\theta} \int_{\xi} \Pr(\mathbf{w}, \mathbf{t}, \mathbf{l}, \mathbf{z}, \phi, \theta, \xi; \alpha, \beta, \gamma) d\xi d\theta d\phi \\ &= \int_{\phi} \prod_{z=1}^K \Pr(\phi_z; \beta) \prod_{s=1}^S \prod_{i=1}^{N_s} \Pr(w_{si} | \phi_{z_{si}}) d\phi \\ &\cdot \int_{\xi} \prod_{z=1}^K \Pr(\xi_z; \gamma) \prod_{s=1}^S \prod_{i=1}^{N_s} \Pr(t_{si} | \xi_{z_{si}}) d\xi \\ &\cdot \int_{\theta} \prod_{l=1}^L \Pr(\theta_l; \alpha) \prod_{s=1}^S \prod_{i=1}^{N_s} \Pr(z_{si} | l_{si}, \theta_{l_{si}}) d\theta \\ &= \int_{\phi} \prod_{z=1}^K \Pr(\phi_z; \beta) \prod_{s=1}^S \prod_{i=1}^{N_s} \Pr(w_{si} | \phi_{z_{si}}) d\phi \\ &\cdot \int_{\xi} \prod_{z=1}^K \Pr(\xi_z; \gamma) \prod_{s=1}^S \prod_{i=1}^{N_s} \Pr(t_{si} | \xi_{z_{si}}) d\xi \\ &\cdot \int_{\theta} \prod_{l=1}^L \Pr(\theta_l; \alpha) \prod_{s=1}^S \prod_{i=1}^{N_s} \Pr(z_{si} | l_{si}, \theta_{l_{si}}) d\theta \end{aligned}$$

We focus on the different integrals in the expression presented above. We start with the integral over  $\phi$ .

$$\begin{aligned} &\int_{\phi} \prod_{z=1}^K \Pr(\phi_z; \beta) \prod_{s=1}^S \prod_{i=1}^{N_s} \Pr(w_{si} | \phi_{z_{si}}) d\phi \\ &= \prod_{z=1}^K \int_{\phi_z} \Pr(\phi_z; \beta) \prod_{s=1}^S \prod_{i=1}^{N_s} \Pr(w_{si} | \phi_{z_{si}}) d\phi_z \\ &= \prod_{z=1}^K \int_{\phi_z} \frac{\Gamma(\sum_{r=1}^V \beta_r)}{\prod_{r=1}^V \Gamma(\beta_r)} \prod_{r=1}^V \phi_{zr}^{\beta_r-1} \prod_{r=1}^V \phi_{zr}^{n_{zr}^z} d\phi_z \\ &= \prod_{z=1}^K \int_{\phi_z} \frac{\Gamma(\sum_{r=1}^V \beta_r)}{\prod_{r=1}^V \Gamma(\beta_r)} \prod_{r=1}^V \phi_{zr}^{\beta_r+n_{zr}^z-1} d\phi_z \\ &= \prod_{z=1}^K \frac{\Gamma(\sum_{r=1}^V \beta_r)}{\prod_{r=1}^V \Gamma(\beta_r)} \frac{\prod_{r=1}^V \Gamma(n_{zr}^z + \beta_r)}{\Gamma(\sum_{r=1}^V n_{zr}^z + \beta_r)} \end{aligned}$$

where  $n_r^z$  denotes the number of times word  $r$  was associated with topic  $z$  across all sources and entries. Similarly we have for the  $\xi$  part:

$$\begin{aligned} &\int_{\xi} \prod_{z=1}^K \Pr(\xi_z; \gamma) \prod_{s=1}^S \prod_{i=1}^{N_s} \Pr(t_{si} | \xi_{z_{si}}) d\xi \\ &= \prod_{z=1}^K \int_{\xi_z} \Pr(\xi_z; \gamma) \prod_{s=1}^S \prod_{i=1}^{N_s} \Pr(t_{si} | \xi_{z_{si}}) d\xi \\ &= \prod_{z=1}^K \frac{\Gamma(\sum_{t=1}^T \gamma_t)}{\prod_{t=1}^T \Gamma(\gamma_t)} \frac{\prod_{t=1}^T \Gamma(m_t^z + \gamma_t)}{\Gamma(\sum_{t=1}^T m_t^z + \gamma_t)} \end{aligned}$$

where  $m_t^z$  denotes the number of times time-point  $t$  was associated with topic  $z$  across all sources. Finally, we focus on the  $\theta$  integral. We follow a similar analysis and have the following:

$$\begin{aligned}
& \int_{\theta} \prod_{l=1}^L \Pr(\theta_l; \alpha) \prod_{s=1}^S \prod_{i=1}^{N_s} \Pr(z_{si} | l_{si}, \theta_{l_{si}}) d\theta \\
&= \prod_{l=1}^L \int_{\theta_l} \Pr(\theta_l; \alpha) \prod_{s=1}^S \prod_{i=1}^{N_s} \Pr(z_{si} | l_{si}, \theta_{l_{si}}) d\theta_l \\
&= \prod_{l=1}^L \int_{\theta_l} \frac{\Gamma(\sum_{z=1}^K \alpha_z)}{\prod_{z=1}^K \Gamma(\alpha_z)} \prod_{z=1}^K \theta_{lz}^{\alpha_z-1} \prod_{z=1}^K \theta_{lz}^{o_l^z} d\theta_l \\
&= \prod_{l=1}^L \int_{\theta_l} \frac{\Gamma(\sum_{z=1}^K \alpha_z)}{\prod_{z=1}^K \Gamma(\alpha_z)} \prod_{z=1}^K \theta_{lz}^{\alpha_z+o_l^z-1} d\theta_l \\
&= \prod_{l=1}^L \frac{\Gamma(\sum_{z=1}^K \alpha_z)}{\prod_{z=1}^K \Gamma(\alpha_z)} \frac{\prod_{z=1}^K \Gamma(o_l^z + \alpha_z)}{\Gamma(\sum_{z=1}^K o_l^z + \alpha_z)}
\end{aligned}$$

where  $o_l^z$  denotes the number of times location  $l$  was associated with topic  $z$  across all sources and their entries. Eventually we have that the joint distribution is given by:

$$\begin{aligned}
\Pr(\mathbf{w}, \mathbf{t}, \mathbf{l}, \mathbf{z}; \alpha, \beta, \gamma) &= \prod_{z=1}^K \frac{\Gamma(\sum_{r=1}^V \beta_r)}{\prod_{r=1}^V \Gamma(\beta_r)} \frac{\prod_{r=1}^V \Gamma(n_r^z + \beta_r)}{\Gamma(\sum_{r=1}^V n_r^z + \beta_r)} \\
&\cdot \prod_{z=1}^K \frac{\Gamma(\sum_{t=1}^T \gamma_t)}{\prod_{t=1}^T \Gamma(\gamma_t)} \frac{\prod_{t=1}^T \Gamma(m_t^z + \gamma_t)}{\Gamma(\sum_{t=1}^T m_t^z + \gamma_t)} \\
&\cdot \prod_{l=1}^L \frac{\Gamma(\sum_{z=1}^K \alpha_z)}{\prod_{z=1}^K \Gamma(\alpha_z)} \frac{\prod_{z=1}^K \Gamma(o_l^z + \alpha_z)}{\Gamma(\sum_{z=1}^K o_l^z + \alpha_z)}
\end{aligned}$$

The goal of Gibbs sampling is to approximate the conditional distribution  $\Pr(\mathbf{z} | \mathbf{w}, \mathbf{t}, \mathbf{l}; \alpha, \beta, \gamma, \Psi)$ . Using the chain rule we have the following for the conditional probability:

$$\begin{aligned}
& \Pr(z_{si} | \mathbf{w}, \mathbf{t}, \mathbf{l}, \mathbf{z}_{-si}; \alpha, \beta, \gamma) \\
&= \frac{\Pr(z_{si}, w_{si}, t_{si}, l_{si} | \mathbf{w}_{-si}, \mathbf{t}_{-si}, \mathbf{l}_{-si}, \mathbf{z}_{-si}; \alpha, \beta, \gamma)}{\Pr(w_{si}, t_{si}, l_{si} | \mathbf{w}_{-si}, \mathbf{t}_{-si}, \mathbf{l}_{-si}, \mathbf{z}_{-si}; \alpha, \beta, \gamma)} \\
&\propto \frac{n_{w_{si}}^{k, -(s,i)} + \beta_{w_{si}}}{\sum_{r=1}^V n_r^{k, -(s,i)} + \beta_r} \cdot \frac{m_{t_{si}}^{k, -(s,i)} + \gamma_{t_{si}}}{\sum_{t=1}^T m_t^{k, -(s,i)} + \gamma_t} \\
&\cdot \frac{o_{l_{si}}^{k, -(s,i)} + \alpha_{l_{si}}}{\sum_{l=1}^L o_l^{k, -(s,i)} + \alpha_l}
\end{aligned}$$

where  $-si$  in the superscript indicates that the current example has been excluded by the count summations.

## 2 Topic Model Evaluation for Common Diseases

We evaluate the performance of the proposed topic model at discovering common disease topics and their temporal patterns. Table 1 shows three topics related to avian flu, dengue and swine flu. Again, we present their most likely words and their prominence histograms over time. For all

three topics we see that the corresponding disease keywords, i.e., “influenza”, “dengue” and “gripe” (flu) are ranked first. For the avian influenza topic, the proposed topic model is able to discover the correlation among words referring to both the *causes*, i.e., “mosquito”, “larvas”, “zancudos” (mosquitos), and the *symptoms*, i.e., “fiebre” (fever), of the disease. Regarding the dengue topic, our approach is able to identify the main transmission root of dengue which is via the *aedes aegypti* mosquito, as well as, the fact that dengue is more prominent in rural and agricultural areas. Finally, a similar performance is observed for the swine flu topic. Our approach can identify the correlation between the word “bacteria” and swine flu - bacteria co-infections play a key role in swine flu deaths - and the correlation between “paracetamol” and swine flu, one of indicated medication substances for the disease.

## 3 Detailed Evaluation of Different Prediction Approaches

In this section, we present the detailed evaluation results for the performance of BSR, KeyWord, LocSeer, and SourceSeer on predicting Hantavirus incidences from January 2013 to March 2014. Table 2 shows the precision, recall and F1 score of the four approaches aggregated over Chile, Argentina, Brazil and Uruguay. For LocSeer and SourceSeer we report the results for the configuration  $k$  that obtained the best performance.

As shown, SourceSeer obtains the best F1-score and recall for most of the months. We observe that the recall obtained by LocSeer is comparable to that of SourceSeer, while the recall of BSR is significantly lower compared to that of both LocSeer and SourceSeer. The latter is expected as BSR can only predict outbreaks for states where a sufficient number of outbreaks has occurred in the past. In fact, due to its design BSR fails completely to forecast outbreaks for states or countries where no outbreaks have been observed in the past (e.g., the outbreak in Brazil for October 2013 and the outbreak in Uruguay for March 2013). However this mechanism limits the number of false positives significantly, and thus, for many months we observe slightly higher or comparable precision scores for BSR with those of SourceSeer. On the other hand the precision scores of LocSeer are significantly lower compared to SourceSeer. The reason for this behavior is the increased number of false positives returned by LocSeer even after the thresholding mechanism was employed. Regarding KeyWord, we observe that the model performs reasonably well when there is an increase in the number of outbreaks in previous weeks leading to increased keyword counts. However, the model performs poorly for cases when a small number of words was observed in the previous weeks. For example KeyWord failed to forecast Hantavirus outbreaks in August and September 2013 because only one incidence was mentioned in July.

Table 1: Three discovered topics that are related to Influenza (Avian Flu), dengue and Swine Flu. Histograms show the topic prominence over time; The top words with their probability in each topic are shown.

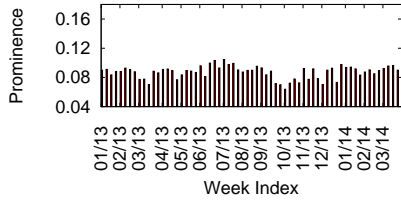
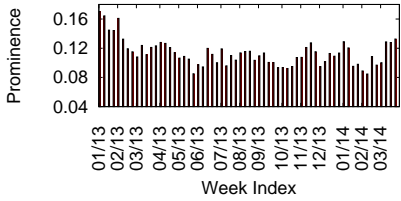
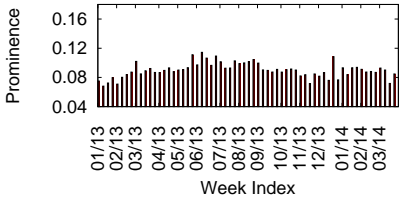
Influenza		Dengue		Swine Flu	
					
influenza	0.0567	dengue	0.2095	gripe	0.0522
mosquito	0.0495	aegypti	0.0166	h1n1	0.0351
pacientes	0.0258	agua	0.0137	infectadas	0.0043
aviar	0.0144	mosquitos	0.0058	flu	0.0024
larvas	0.0096	agricultura	0.0019	bacteria	0.0021
fiebre	0.0088	respiratoria	0.0018	enfermo	0.0008
surto	0.0061	rurales	0.0006	vacinas	0.0008
zancudos	0.0008	agropecuario	0.0006	nasal	0.0008
avian	0.0006	hemorragias	0.0005	paracetamol	0.0007
h5n1	0.0003	suero	0.0004	swine	0.0005

Table 2: BSR, KeyWord, LocSeer and SourceSeer on predicting hantavirus outbreaks. Notation ( $k\%$ ) denotes the best performing configuration for LocSeer and SourceSeer.

Month	BSR			KeyWord			LocSeer (5%)			SourceSeer (5%)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
01/13	0.5	0.17	0.25	<b>0.67</b>	0.33	0.44	0.13	<b>0.67</b>	0.22	0.44	<b>0.67</b>	<b>0.53</b>
02/13	0.52	0.78	0.62	<b>0.67</b>	<b>1.0</b>	<b>0.80</b>	0.12	<b>1.0</b>	0.21	0.5	<b>1.0</b>	0.67
03/13	<b>0.7</b>	0.35	0.46	0.6	<b>0.75</b>	<b>0.67</b>	0.29	0.5	0.37	0.5	0.5	0.5
04/13	<b>0.78</b>	0.59	0.67	0.33	0.25	0.28	0.6	0.75	0.67	0.57	<b>1.0</b>	<b>0.73</b>
05/13	<b>0.51</b>	0.48	<b>0.54</b>	0.29	0.4	0.34	0.14	0.2	0.16	0.38	<b>0.6</b>	0.47
06/13	<b>0.22</b>	0.68	<b>0.33</b>	0	0	0	0.14	<b>1.0</b>	0.25	0.14	<b>1.0</b>	0.25
07/13	<b>0.22</b>	0.68	<b>0.33</b>	0	0	0	0.14	<b>1.0</b>	0.25	0.2	<b>1.0</b>	<b>0.33</b>
08/13	0.4	0.6	0.47	0	0	0	0.2	<b>1.0</b>	0.33	<b>0.67</b>	<b>1.0</b>	<b>0.80</b>
09/13	0.5	0.33	0.39	0	0	0	0.23	<b>1.0</b>	0.37	<b>0.67</b>	<b>0.67</b>	<b>0.67</b>
10/13	<b>0.62</b>	0.24	0.35	0.5	0.4	0.44	0.31	<b>0.8</b>	0.45	0.38	0.6	<b>0.47</b>
11/13	<b>0.89</b>	0.44	0.59	0.75	0.5	<b>0.6</b>	0.21	<b>0.83</b>	0.34	0.45	<b>0.83</b>	0.58
12/13	0.9	0.32	0.47	<b>0.75</b>	0.27	0.40	<b>0.75</b>	<b>0.55</b>	<b>0.63</b>	0.67	<b>0.55</b>	0.60
01/14	0.65	0.49	0.56	0.43	0.38	0.40	0.19	0.5	0.28	<b>0.71</b>	<b>0.63</b>	<b>0.67</b>
02/14	0.56	<b>0.74</b>	0.64	0.43	0.5	0.46	0.27	0.67	0.38	<b>0.67</b>	0.67	<b>0.67</b>
03/14	0.55	<b>0.88</b>	<b>0.68</b>	<b>0.57</b>	0.8	0.66	0.29	0.8	0.42	0.5	0.8	0.62