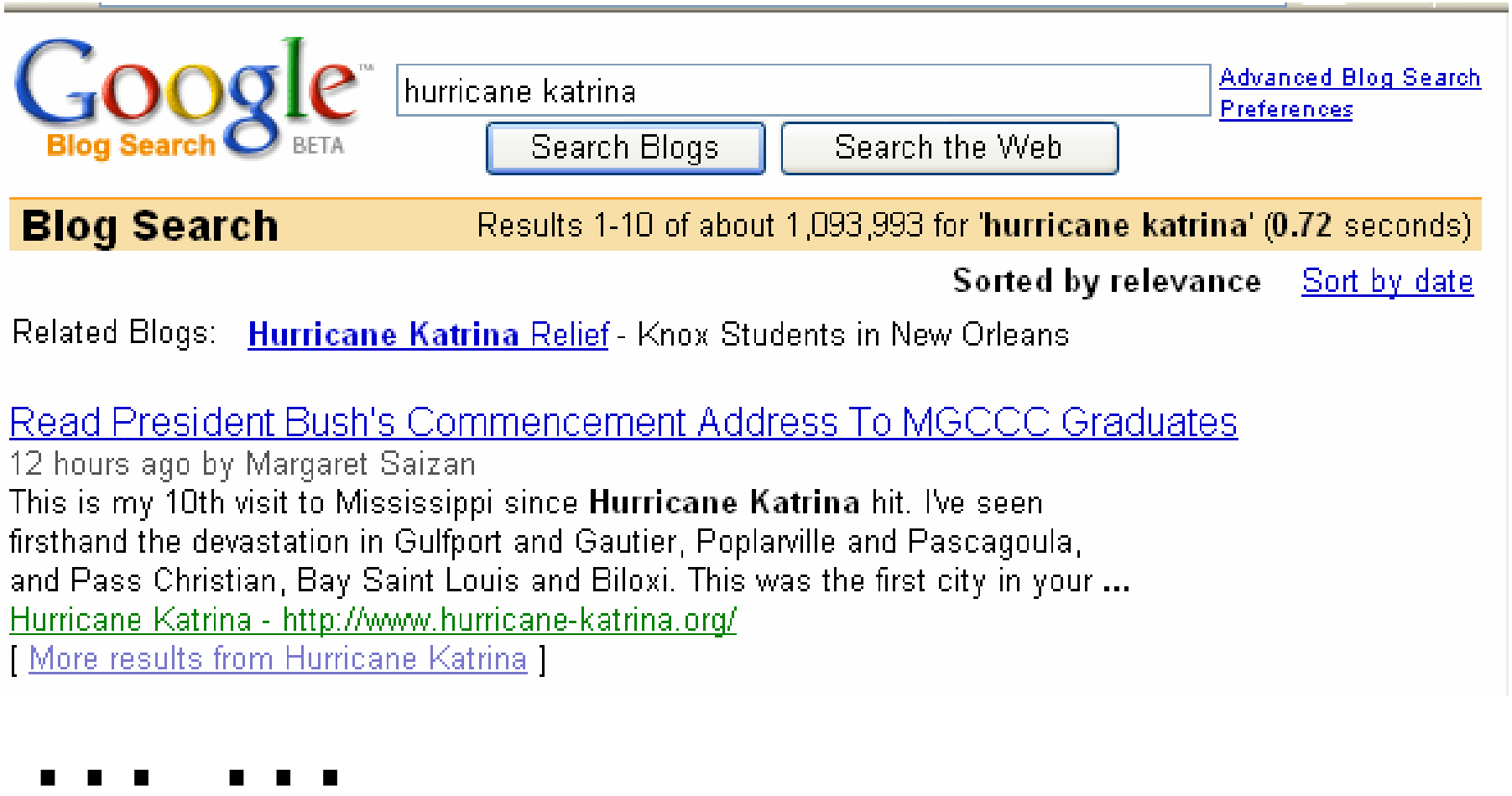
A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs

Qiaozhu Mei[†], Chao Liu[†], Hang Su[‡], and ChengXiang Zhai[†]

[†]: University of Illinois at Urbana-Champaign

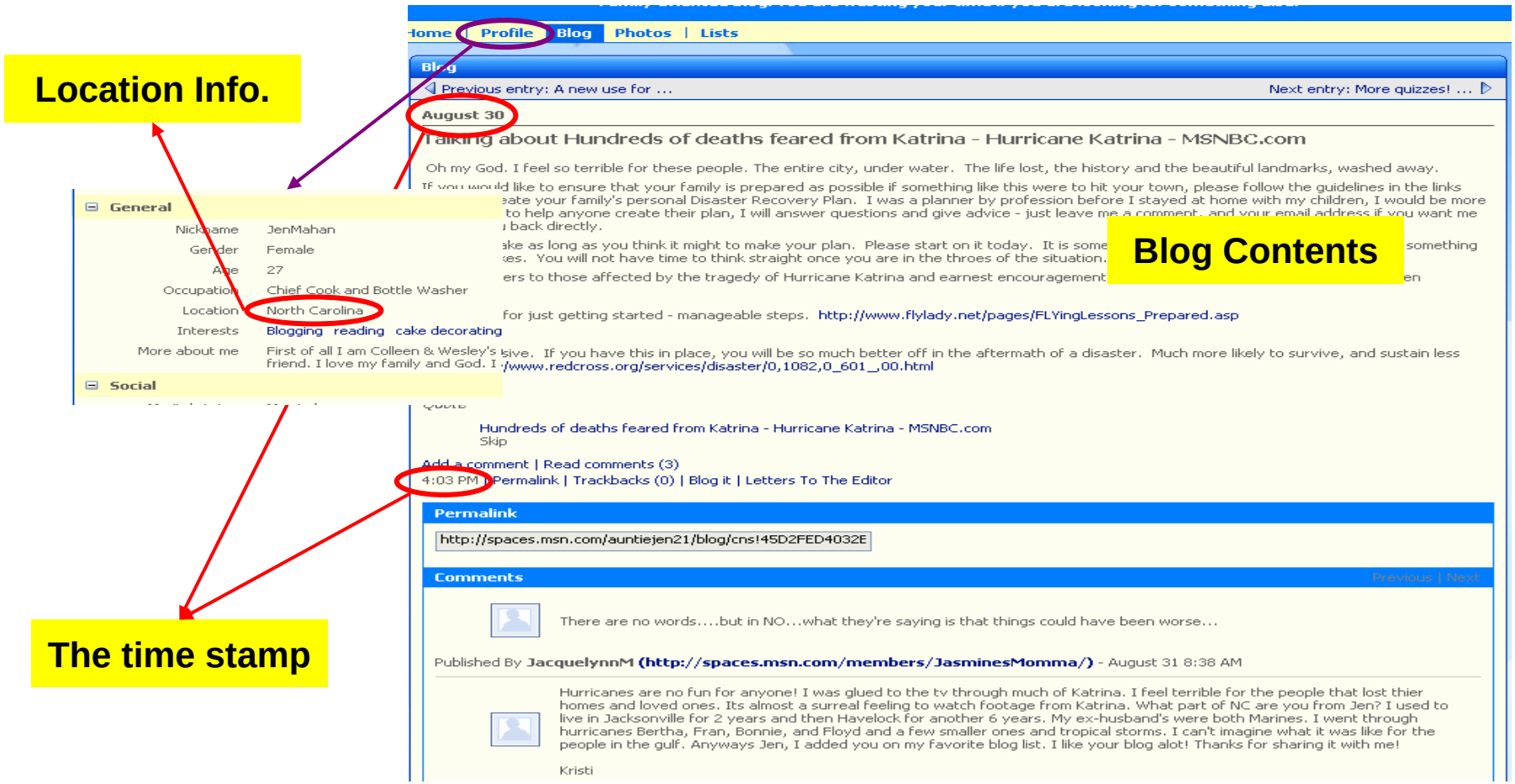
[‡]: Vanderbilt University

Weblog as an emerging new data…



■ ■ ■ ■ ■ ■

An Example of Weblog Article



Characteristics of Weblogs

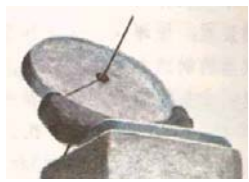
Highly personal
With opinions

Interlinking &

Forming communities

With mixed topics

Time



Location

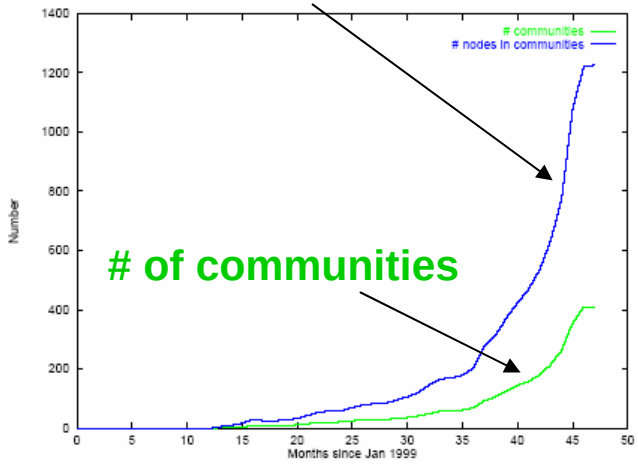
Associated with time & location

Immediate response to events

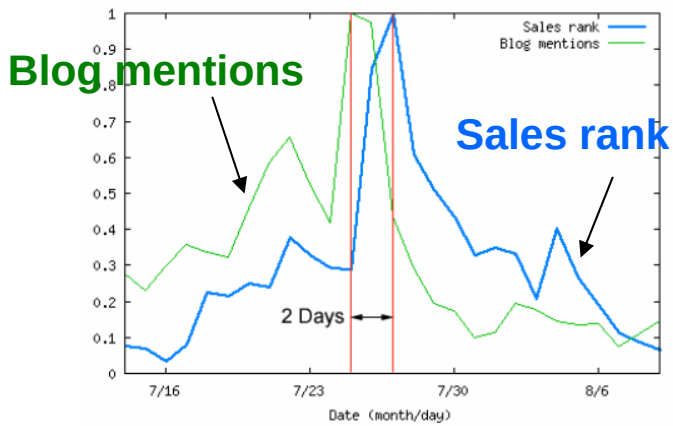
Existing Work on Weblog Analysis

- Interlinking and Community Analysis
  - Identifying communities
  - Monitoring the evolution and bursting of communities
  - E.g., [Kumar et al. 2003]

# of nodes in communities



- Content Analysis
  - Blog level topic analysis
  - Information diffusion through blogspace
  - Use topic bursting to predict sales spikes
  - E.g., [Gruhl et al. 2005]



Query: Lance Armstrong OR Tour de France

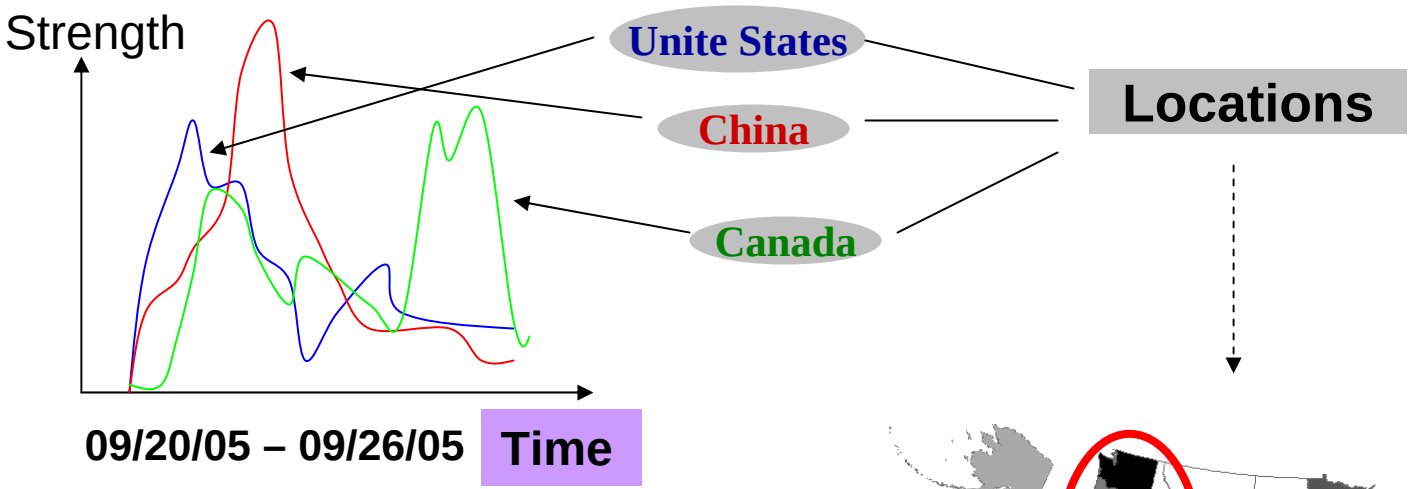
How to Perform Spatiotemporal Theme Mining?

- Given a collection of Weblog articles about a topic with time and location information
  - Discover multiple themes (i.e., subtopics) being discussed in these articles
  - For a given location, discover how each theme evolves over time (generate a theme life cycle)
  - For a given time, reveal how each theme spreads over locations (generate a theme snapshot)
  - Compare theme life cycles in different locations
  - Compare theme snapshots in different time periods
  - …

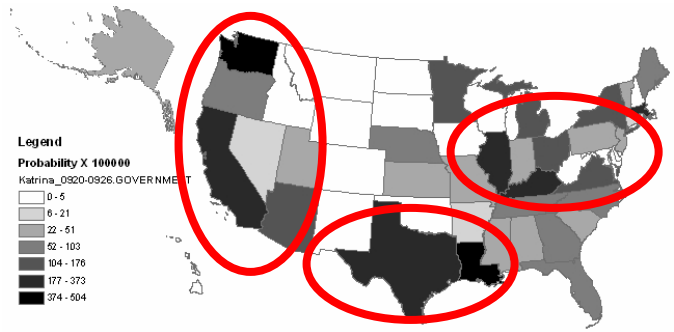
Spatiotemporal Theme Patterns

Theme life cycles

Discussion about "Release of iPod Nano" in articles about "iPod Nano"



Discussion about "Government Response" in articles about Hurricane Katrina



A theme snapshot

Applications of Spatiotemporal Theme Mining

- Help answer questions like
  - Which country responded first to the release of iPod Nano? China, UK, or Canada?
  - Do people in different states (e.g., Illinois vs. Texas) respond differently/similarly to the increase of gas price during Hurricane Katrina?
- Potentially useful for
  - Summarizing search results
  - Monitoring public opinions
  - Business Intelligence
  - …

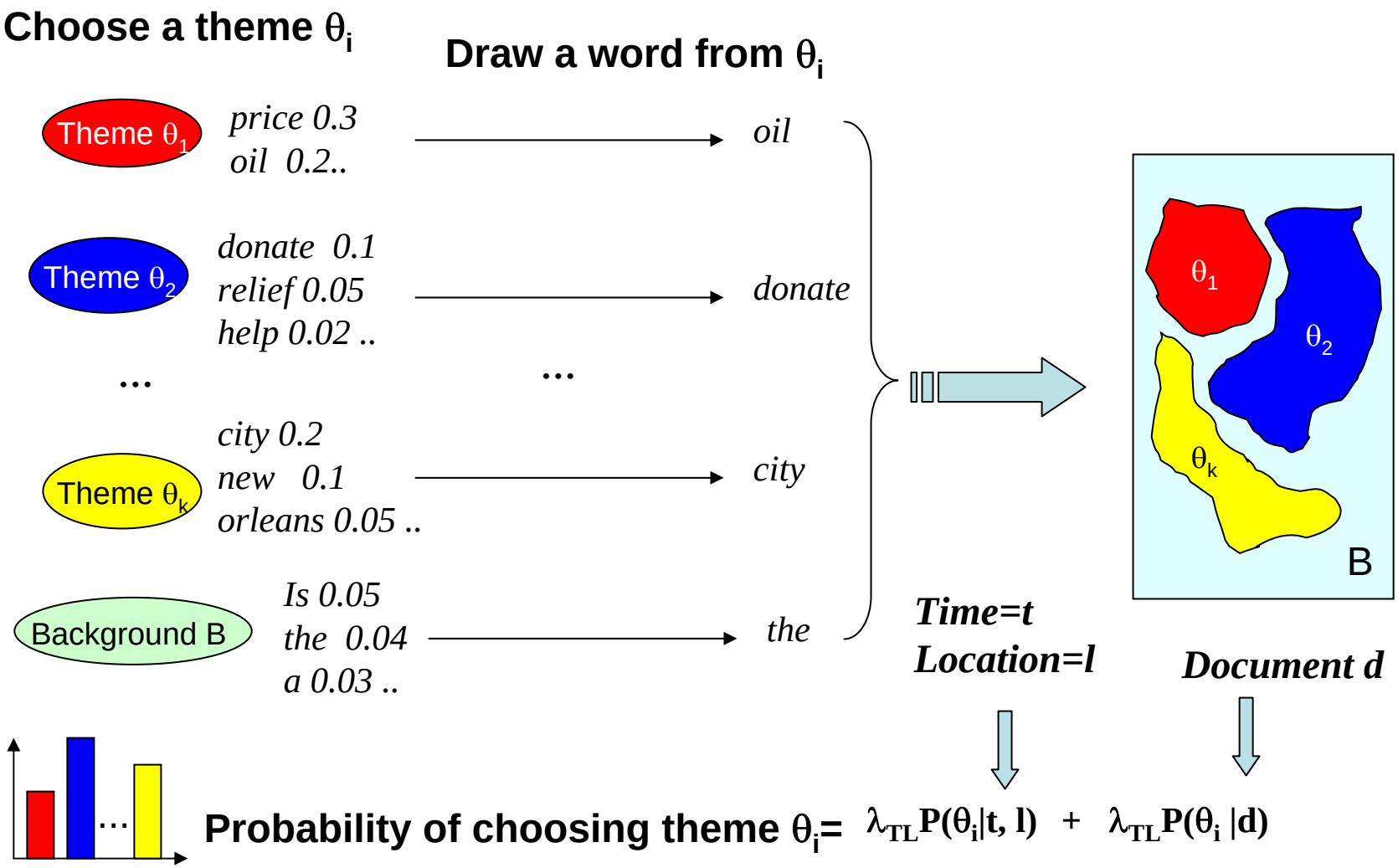
Challenges in Spatiotemporal Theme Mining

- How to represent a theme?
- How to model the themes in a collection?
- How to model their dependency on time and location?
- How to compute the theme life cycles and theme snapshots?
- All these must be done in an unsupervised way…

Our Solution: Use a Probabilistic Spatiotemporal Theme Model

- Each theme is represented as a multinomial distribution over the vocabulary (language model)
- Consider the collection as a sample from a mixture of these theme models
- Fit the model to the data and estimate the parameters
- Spatiotemporal theme patterns can then be computed from the estimated model parameters

Probabilistic Spatiotemporal Theme Model



The "Generation" Process

- A document d of location l and time t is generated, word by word, as follows
  - First, decide whether to use the background theme θ_B
    - With probability λ_B , we'll use the background theme and draw a word w from $p(w|\theta_B)$
  - If the background theme is not to be used, we'll decide how to choose a topic theme
    - With probability λ_{TL}, we'll sample a theme using the "shared spatiotemporal distribution" $p(\theta|t,l)$
    - With probability 1- λ_{TL}, we'll sample a theme using $p(\theta|d)$
  - Draw a word w from the selected theme distribution $p(w|\theta_i)$
- Parameters
  - $\{p(w|\theta_B), p(w|\theta_i), p(\theta|t,l), p(\theta|d)\}$ (will be estimated)
  - λ_B =*Background noise;* λ_{TL}=*Weight on spatiotemporal modeling* (will be manually set)

The Likelihood Function

$$\log p(C) = \sum_{d \in C} \sum_{w \in V} c(w,d) \times \log \left[\lambda_B P(w \mid B) + (1-\lambda_B) \sum_{j=1}^{k} p(w \mid \theta_j)((1-\lambda_{TL})\, p(\theta_j \mid d) + \lambda_{TL}\, p(\theta_j \mid t_d, l_d)) \right]$$

Count of word w in document d

Generating w using the background theme

Generating w using a topic theme

Choosing a topic theme according to the document

Choosing a topic theme according to the spatiotemporal context

Parameter Estimation

- Use the maximum likelihood estimator
- Use the Expectation-Maximization (EM) algorithm
- $p(w|\theta_B)$ is set to the collection word probability

E Step

$$p(z_{d,w}=j)=\frac{(1-\lambda_B)p^{(m)}(w|\theta_j)[(1-\lambda_{TL})p^{(m)}(\theta_j|d)+\lambda_{TL}p^{(m)}(\theta_j|t_d,l_d)]}{\lambda_B p(w|B)+(1-\lambda_B)\sum_{j'=1}^{k}p^{(m)}(w|\theta_{j'})[(1-\lambda_{TL})p^{(m)}(\theta_{j'}|d)+\lambda_{TL}p^{(m)}(\theta_{j'}|t_d,l_d)]}$$

$$p(y_{d,w,j}=1)=\frac{\lambda_{TL}p^{(m)}(\theta_j|t_d,l_d)}{(1-\lambda_{TL})p^{(m)}(\theta_j|d)+\lambda_{TL}p^{(m)}(\theta_j|t_d,l_d)}$$

M Step

$$p^{(m+1)}(\theta_j|d)=\frac{\sum_{w\in V}c(w,d)p(z_{d,w}=j)(1-p(y_{d,w,j}=1))}{\sum_{j'=1}^{k}\sum_{w\in V}c(w,d)p(z_{d,w}=j')(1-p(y_{d,w,j'}=1))}$$

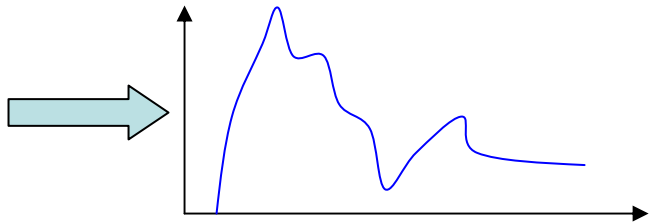
$$p^{(m+1)}(\theta_j|t,l)=\frac{\sum_{d:t_d=t,l_d=l}\sum_{w\in V}c(w,d)p(z_{d,w}=j)p(y_{d,w,j}=1)}{\sum_{d:t_d=t,l_d=l}\sum_{j'=1}^{k}\sum_{w\in V}c(w,d)p(z_{d,w}=j')p(y_{d,w,j'}=1)}$$

$$p^{(m+1)}(w|\theta_j)=\frac{\sum_{d\in C}c(w,d)p(z_{d,w}=j)}{\sum_{w'\in V}\sum_{d\in C}c(w',d)p(z_{d,w'}=j)}$$

Probabilistic Analysis of Spatiotemporal Themes

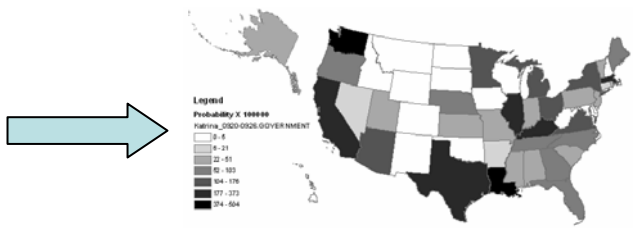
- Once the parameters are estimated, we can easily perform probabilistic analysis of spatiotemporal themes
  - Computing theme life cycles given location

$$p(t \mid \theta_j, \tilde{l}) = \frac{p(\theta_j \mid t, \tilde{l})\, p(t, \tilde{l})}{\sum_{\tilde{t} \in T} p(\theta_j \mid \tilde{t}, \tilde{l})\, p(\tilde{t}, \tilde{l})}$$



  - Computing theme snapshots given time

$$p(\theta_j, l \mid \tilde{t}) = \frac{p(\theta_j \mid \tilde{t}, l)\, p(\tilde{t}, l)}{\sum_{\tilde{l} \in L} \sum_{j'=1}^{k} p(\theta_{j'} \mid \tilde{t}, \tilde{l})\, p(\tilde{t}, \tilde{l})}$$



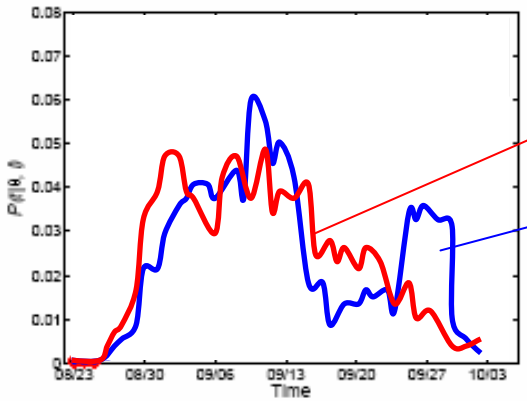
Experiments and Results

- Three time-stamped data sets of weblogs, each about one event (broad topic):

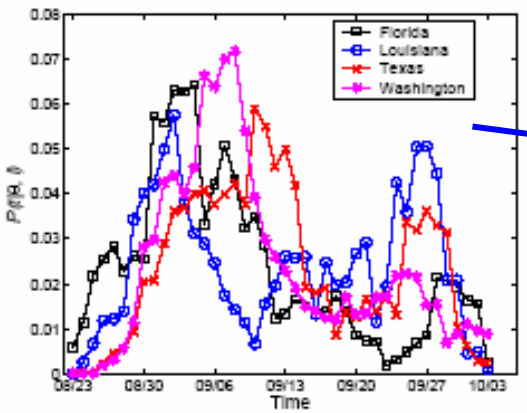
Data Set	# docs	Time Span(2005)	Query
Katrina	9377	08/16 -10/04	Hurricane Katrina
Rita	1754	08/16 - 10/04	Hurricane Rita
iPod Nano	1720	09/02 - 10/26	iPod Nano

- Extract location information from author profiles
- On each data set, we extract a set of salient themes and their life cycles / theme snapshots

Theme Life Cycles for Hurricane Katrina



(a) Theme life cycles in Texas (Hurricane Katrina)



(b) Theme "New Orleans" over states (Hurricane Katrina)

Oil Price

New Orleans

price 0.0772
oil 0.0643
gas 0.0454
increase 0.0210
product 0.0203
fuel 0.0188
company 0.0182
…

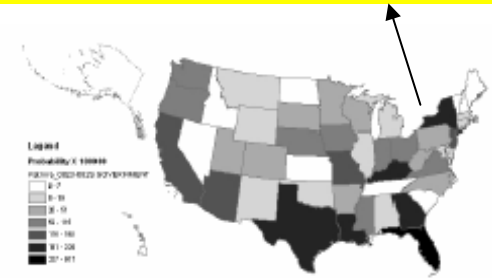
city 0.0634
orleans 0.0541
new 0.0342
louisiana 0.0235
flood 0.0227
evacuate 0.0211
storm 0.0177
…

Theme Snapshots for Hurricane Katrina

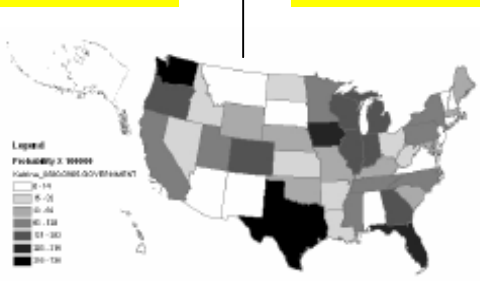
Week2: The discussion moves towards the north and west

Week1: The theme is the strongest along the Gulf of Mexico

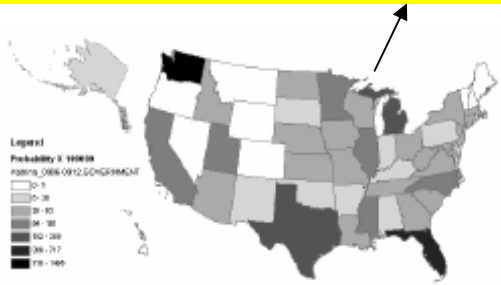
Week3: The theme distributes more uniformly over the states



(a) Week1: 08/23-08/29



(b) Week Two: 08/30-09/05



(c) Week Three:09/06-09/12

Theme 1
Government Response
bush 0.0716374
president 0.0610942
federal 0.0514114
govern 0.0476977
fema 0.0474692
administrate 0.0233903
response 0.0208351
brown 0.0199573
blame 0.0170033
governor 0.0142153

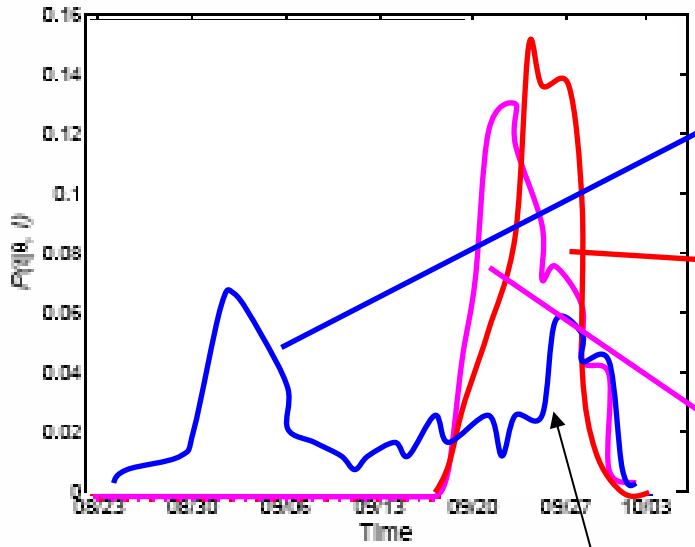
(d) Week Four: 09/13-09/19

(e) Week Five: 09/20-09/26

Week4: The theme is again strong along the east coast and the Gulf of Mexico

Week5: The theme fades out in most states

Theme life cycles for Hurricane Rita



(d)Theme life cycles in Louisiana (Hurricane Rita)

Hurricane Katrina: Government Response

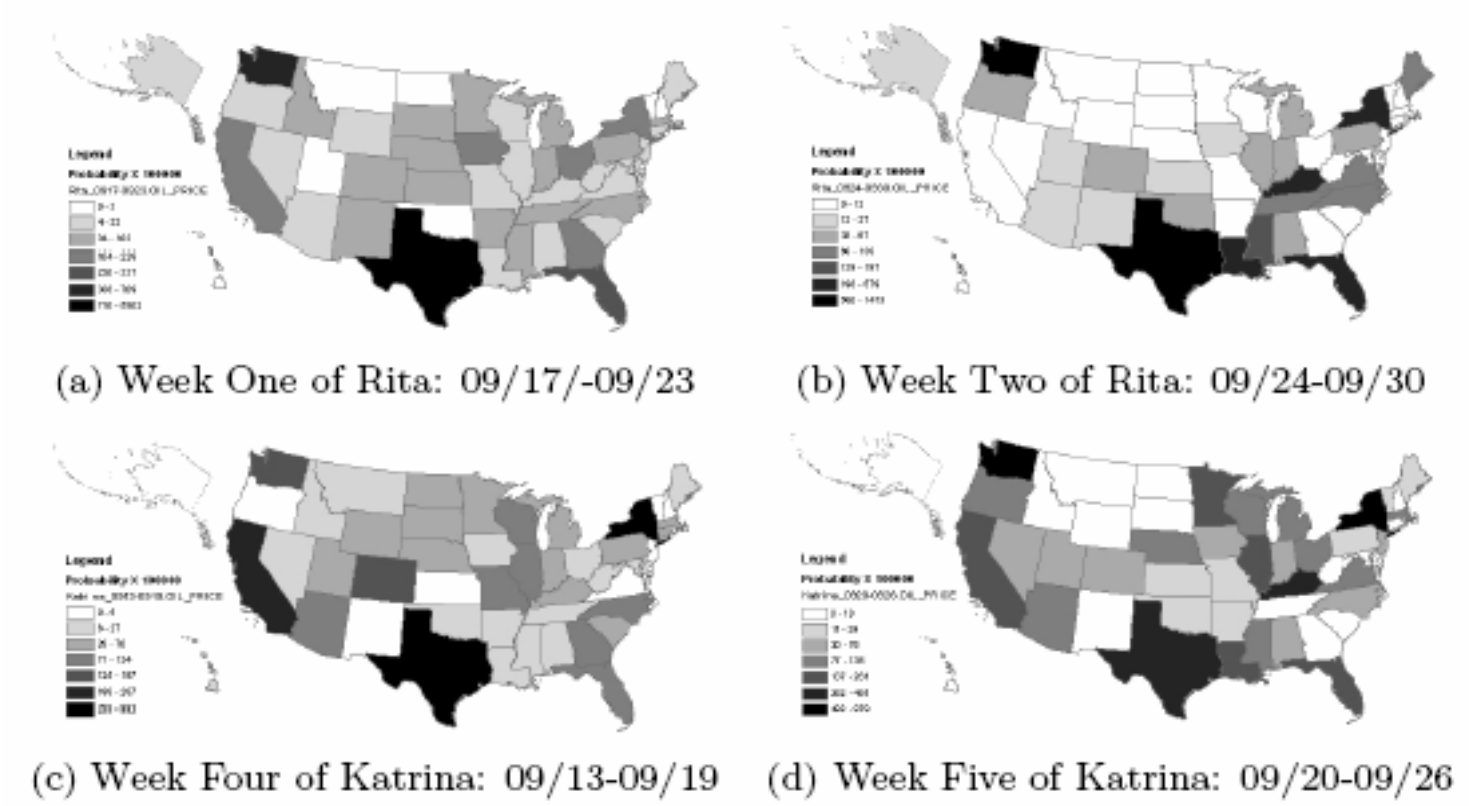
Hurricane Rita: Government Response

Hurricane Rita: Storms

A theme in Hurricane Katrina is inspired again by Hurricane Rita

Theme Snapshots for Hurricane Rita

Both Hurricane Katrina and Hurricane Rita have the theme "Oil Price"



(a) Week One of Rita: 09/17/-09/23

(b) Week Two of Rita: 09/24-09/30

(c) Week Four of Katrina: 09/13-09/19

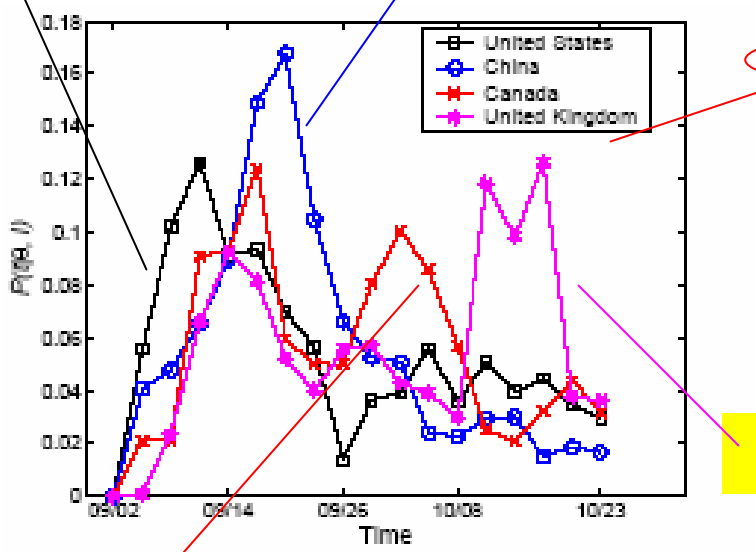
(d) Week Five of Katrina: 09/20-09/26

The spatiotemporal patterns of this theme at the same time period are similar

Theme Life Cycles for iPod Nano

United States

China



Theme "Release of Nano" over states (iPod Nano)

Release of Nano

Canada

United Kingdom

ipod 0.2875
nano 0.1646
apple 0.0813
september 0.0510
mini 0.0442
screen 0.0242
new 0.0200
…

Contributions and Future Work

- Contributions
  - Defined a new problem -- spatiotemporal text mining
  - Proposed a general mixture model for the mining task
  - Proposed methods for computing two spatiotemporal patterns -- theme life cycles and theme snapshots
  - Applied it to Weblog mining with interesting results
- Future work:
  - Capture content dependency between adjacent time stamps and locations
  - Study granularity selection in spatiotemporal text mining

Thank You!