Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora

Jianwen Zhang[†], Yangqiu Song[‡], Changshui Zhang[†], Shixia Liu[‡]

[†]Tsinghua University, Beijing, China, [‡]IBM Research - China

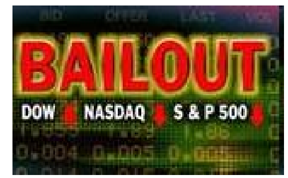
Outline

- Motivation
- Preliminaries
- Evolutionary HDP
- Experiments
  - Synthetic data
  - Finance related online document collections
- Conclusions

Motivation

- Mining cluster evolution patterns in multiple correlated time-varying corpora





News: World braces for insurer AIG's crash | Markets end lower | Congress wrestles with Wall Street bailout package | Obama, McCain debate economic, foreign policy...

Blogs: Financial *crash*: A system in chaos | Preparing for a Financial *Crisis* | Bailout people, not banks

Message boards: Do I have to tell my landlord I *lost* my *job*? | Is *Obama* a US citizen | Canceling my shopping

An example of online textual data from three corpora: news, blogs, and message boards.

Motivation (Cont’)

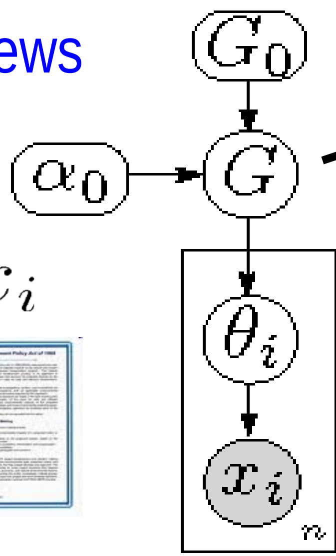
- Patterns to discovery
  - Clusters within each corpus at each epoch
  - Shared clusters among different corpora
  - Evolving of clusters within a corpus and across corpora overtime
- Challenges
  - Single integrated model
  - Commonality & diversity
  - Time dependencies
  - Cluster numbers
- Previous works
  - Multiple corpora
  - Time-varying corpus, evolutionary clustering

Preliminaries

- Dirichlet process mixture models (DPM)
  - DPM: (prior) infinite mixture model which can automatically determine the component number by placing a Dirichlet process (DP) prior for a mixture model

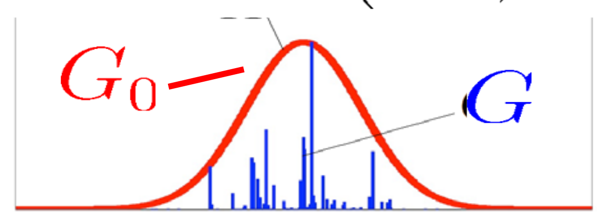
A text corpus, e.g., news

x_i



The DP prior

$$G \sim \mathrm{DP}(\alpha_0, G_0)$$

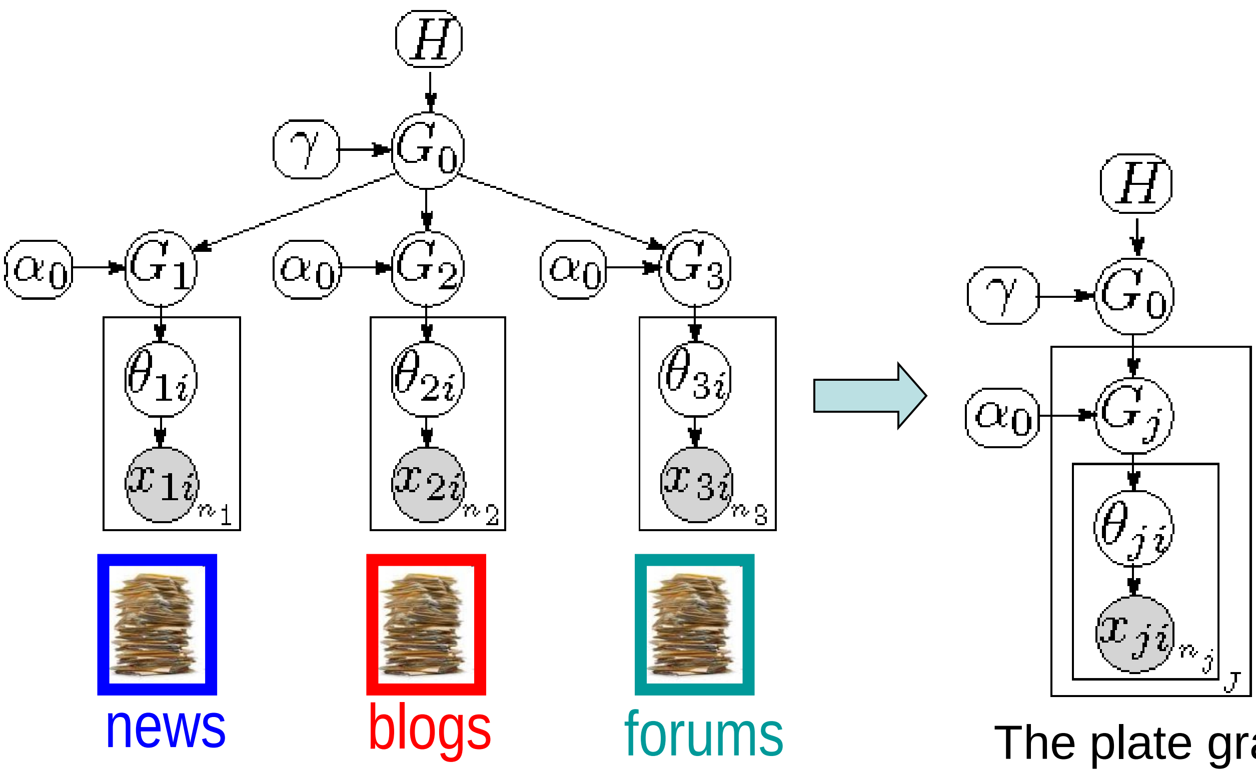


$\theta_i \in$ {politics, market, companies, …}

$$\theta_i \sim G, x_i \sim F(x|\theta_i)$$

Preliminaries (Cont')

- HDP mixture model: to model multiple corpora to enable them sharing components
  - Multiple DPMs sharing a same DP prior



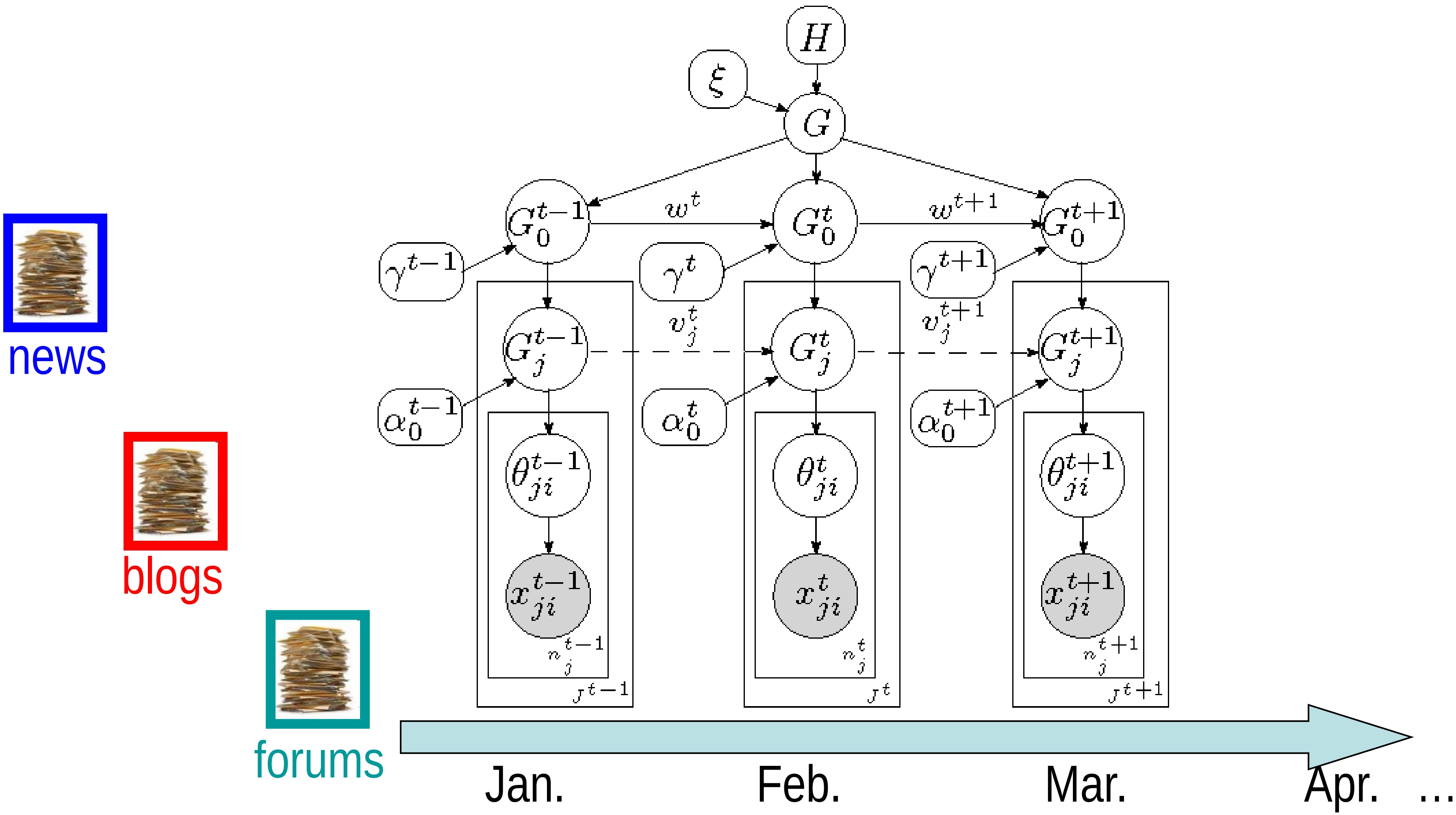
The plate graph

HDP: Multiple DPMs sharing a same DP prior with its parameter drawn from another DP

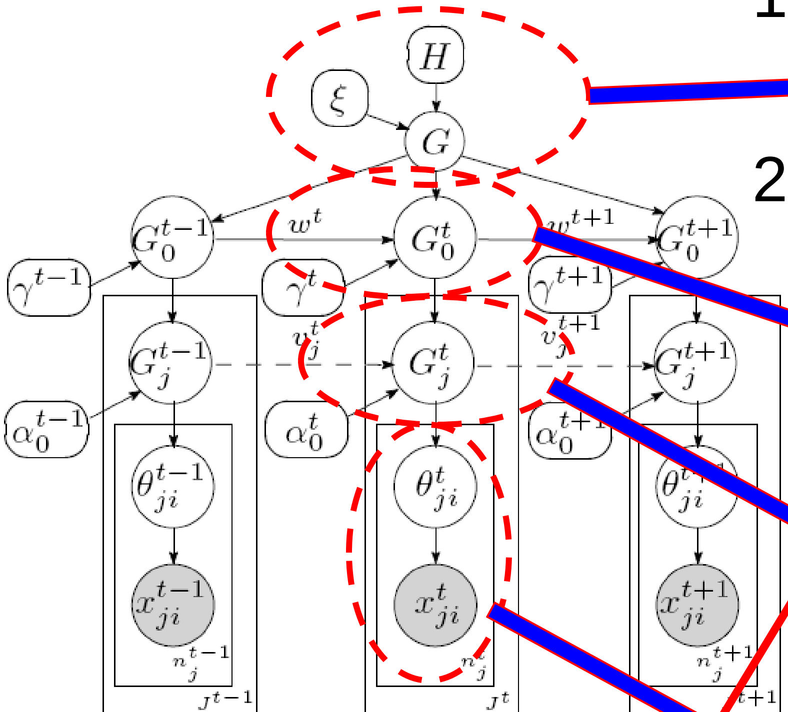
HDP mixture model, Teh *et al. JASA*'06

Evolutionary HDP

- Modeling multiple time varying corpora



Evolutionary HDP (Cont')



The time dependency model

1. Draw an ***overall measure***

$$G \sim \mathrm{DP}(\gamma, H)$$

2. For each epoch t

① Draw the ***snapshot global measure***

$$\boxed{G_0^t \sim \mathrm{DP}\left(\gamma^t, w^t G_0^{t-1} + (1 - w^t) G\right)}$$

② Draw the ***snapshot local measures***

$$\boxed{G_j^t \sim \mathrm{DP}\left(\alpha_0^t, v_j^t G_j^{t-1} + (1 - v_j^t) G_0^t\right)}$$

③ Draw observations

$$\theta_{ji}^t \overset{i.i.d.}{\sim} G_j^t, \quad x_{ji}^t \sim F(x|\theta_{ji}^t)$$

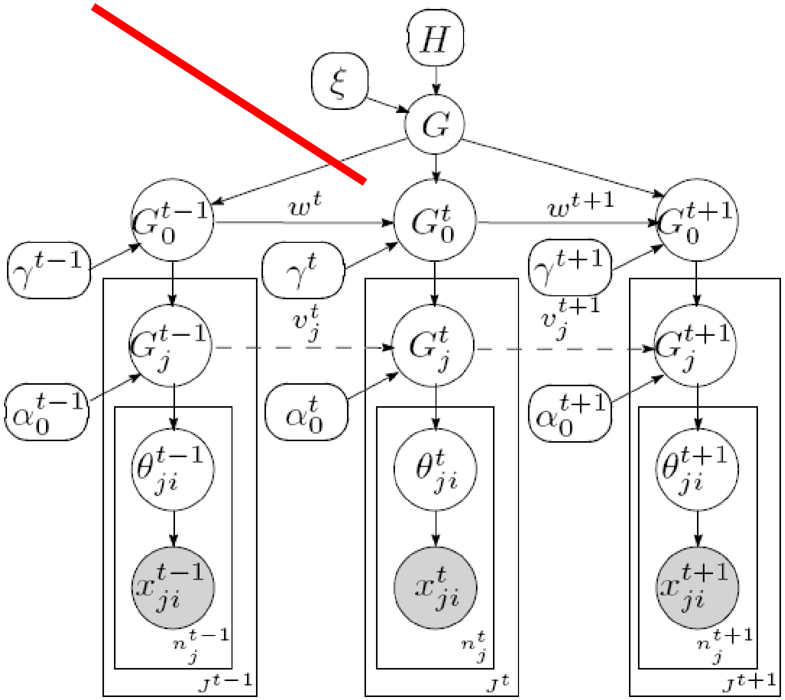
The time dependency model

G **: plays as a bookkeeper of all the components (the common taste)**

$$G_0^t \sim \mathrm{DP}\left(\gamma^t, w^t G_0^{t-1} + (1-w^t)G\right)$$

A part of the atoms of G_0^t are drawn from the previous one G_0^{t-1} while others are drawn from G

Some are inherited from the previous, and some are newly from the common taste.



$$G_j^t \sim \mathrm{DP}\left(\alpha_0^t, v_j^t G_j^{t-1} + (1-v_j^t)G_0^t\right)$$

Similarly…

More…

- Different perspectives to the model (necessary to lead to the sampling scheme)
- Gibbs sampling to infer the model

(Detailing and boring. If you are interested in, we're appreciated if you would like to read the paper instead)

Experiments

- Synthetic data
- Real financial related web text collections

Experiments on synthetic data

$$p_j^t(x) = \sum_{\tau=1}^{3} \frac{1}{3} \text{Multinomial}\left(x; \phi_{k_{j\tau}^t}\right)$$

2-dimenional multinomial

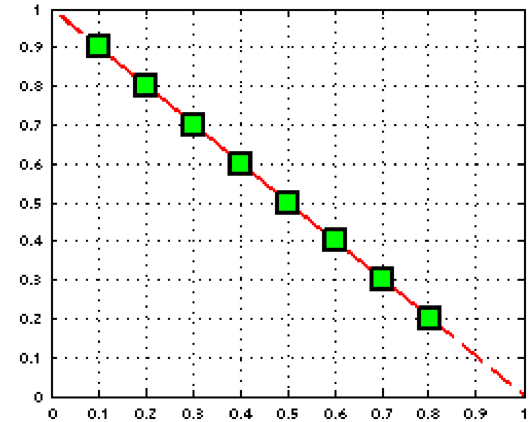
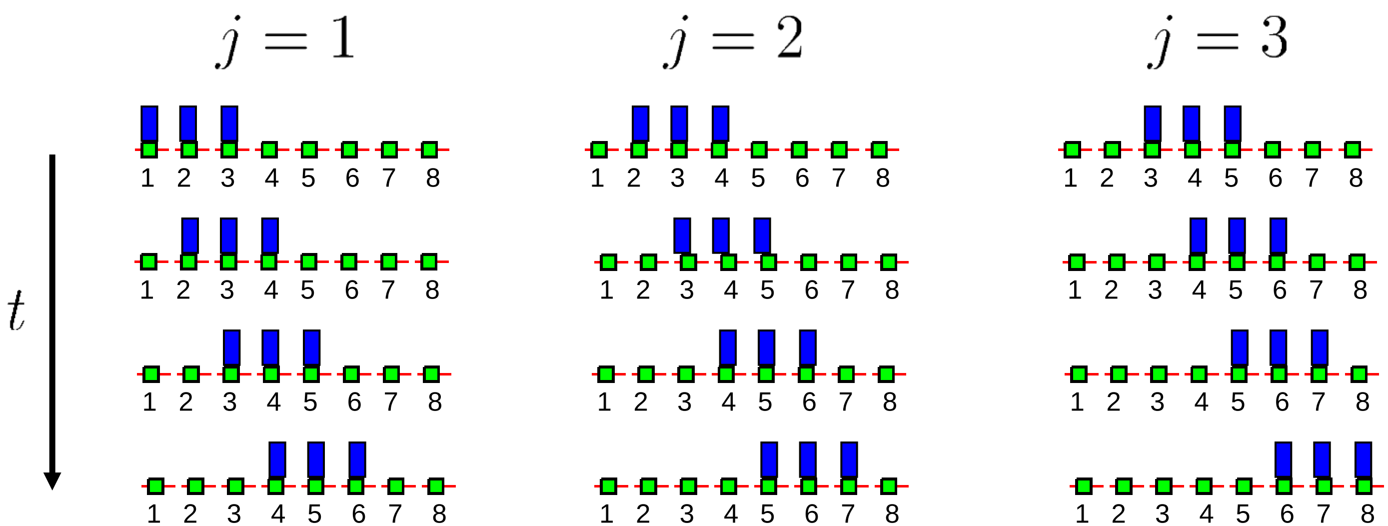


Table 1: Synthetic data set.

Global components (dishes)								
k	1	2	3	4	5	6	7	8
$\phi_{k,1}$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8

Local components (tables) and corpora sizes						
	Tables $(k_{j1}^t, k_{j2}^t, k_{j3}^t)$			Corpora sizes n_j^t		
	$j = 1$	$j = 2$	$j = 3$	$j = 1$	$j = 2$	$j = 3$
$t = 1$	1, 2, 3	2, 3, 4	3, 4, 5	500	300	400
$t = 2$	2, 3, 4	3, 4, 5	4, 5, 6	510	320	430
$t = 3$	3, 4, 5	4, 5, 6	5, 6, 7	520	320	430
$t = 4$	4, 5, 6	5, 6, 7	6, 7, 8	530	340	450



Three corpora, four time epochs

Evaluation criteria

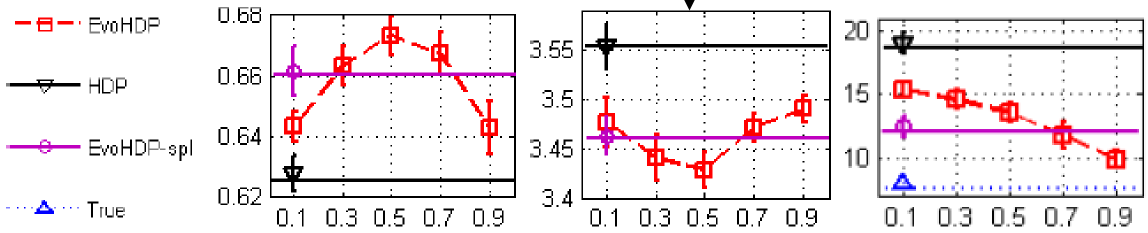
- Static criteria
  - NMI
  - $\log(perword\text{-}perplexity)$ *LogPerp*

$$-\frac{1}{n_{test}}\sum_{t,j,i}\log p\left(x^t_{ji,test}\middle|Model, X_{train}\right)$$

- Temporal criteria
  - Temporal correlations / divergences overtime
- Compared to HDP without considering time dependencies

Better predict ability

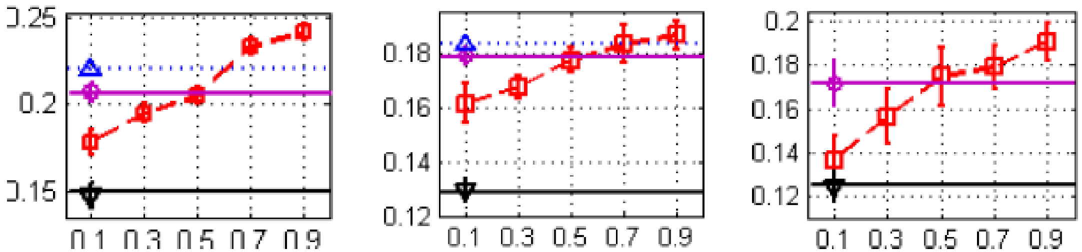
Better clustering performance



(a) NMI (b) LogPerp (c) K

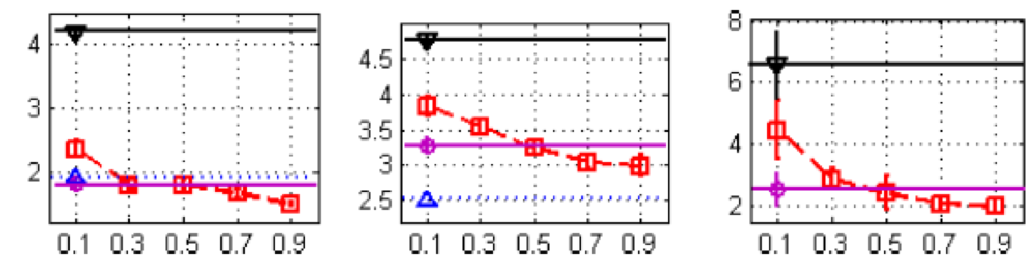
Figure 6: Results on the synthetic data set: static performances, averaged on 10-fold cross validation.

Stronger correlation overtime



(a) IntraCorr (b) InterCorr (c) GCorr

Figure 7: Results on the synthetic data set: temporal correlations, averaged on 10-fold cross validation.

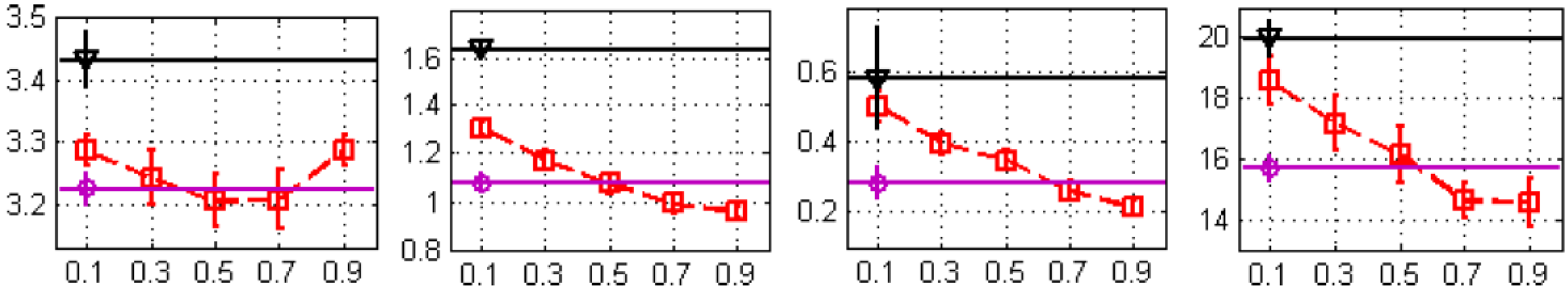


(a) IntraKL (b) InterKL (c) GKL

Figure 8: Results on the synthetic data set: temporal divergences, averaged on 10-fold cross validation.

Experiments on Real Data

- 103,986 text articles queried from a search engine *Boardreader*.
- Financial related. Queries: 20 financial companies' names, e.g., "AIG insurance", "Bank of America", etc.
- Three types. News, blogs, message boards.
- 6 months, Jul. 2008 – Dec. 2008
- Dictionary size $W = 77,999$

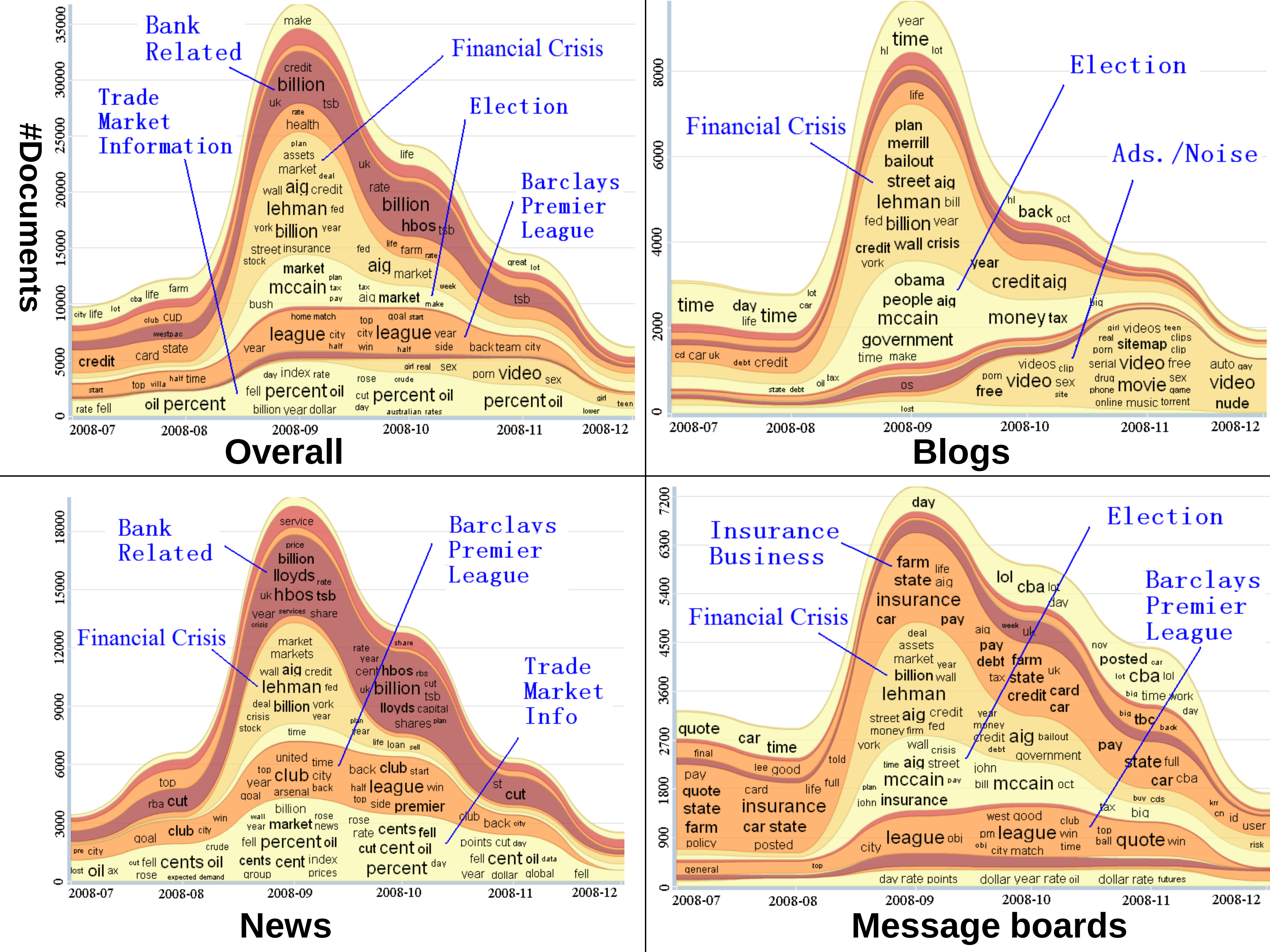


(a) LogPerp (b) InterKL (c) GKL (d) K

Better predict ability

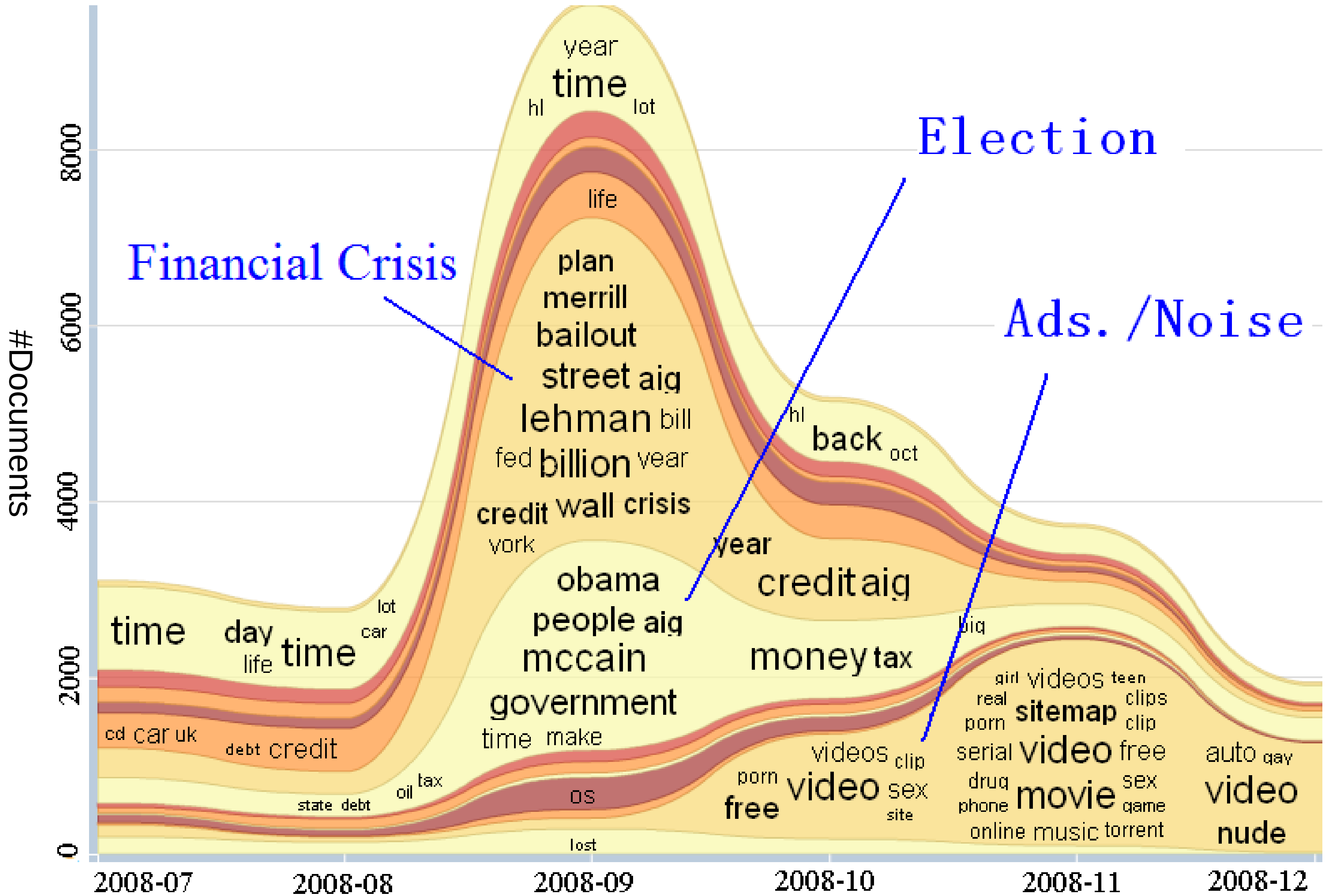
Stronger correlation overtime

Visualization of clusters utilizing the time-based topic visualization tool TIARA (Liu et al. CIKM'09)



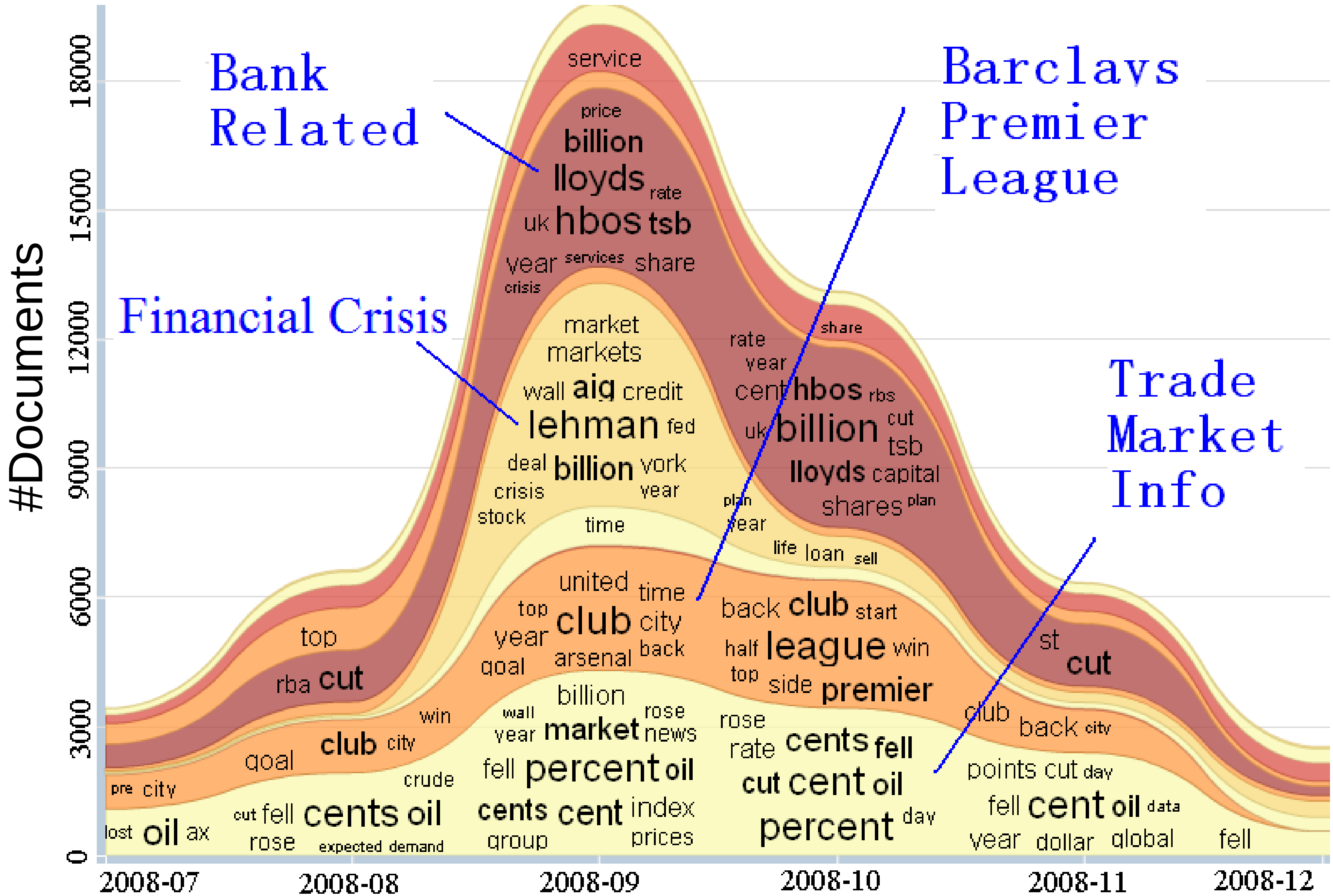
#Documents
Bank Related
Trade Market Information
Financial Crisis
Election
Barclays Premier League
2008-07
2008-08
2008-09
2008-10
2008-11
2008-12
Overall
Financial Crisis
Election
Ads. /Noise
Blogs
Bank Related
Financial Crisis
Barclays Premier League
Trade Market Info
News
Insurance Business
Financial Crisis
Election
Barclays Premier League
Message boards

Blogs

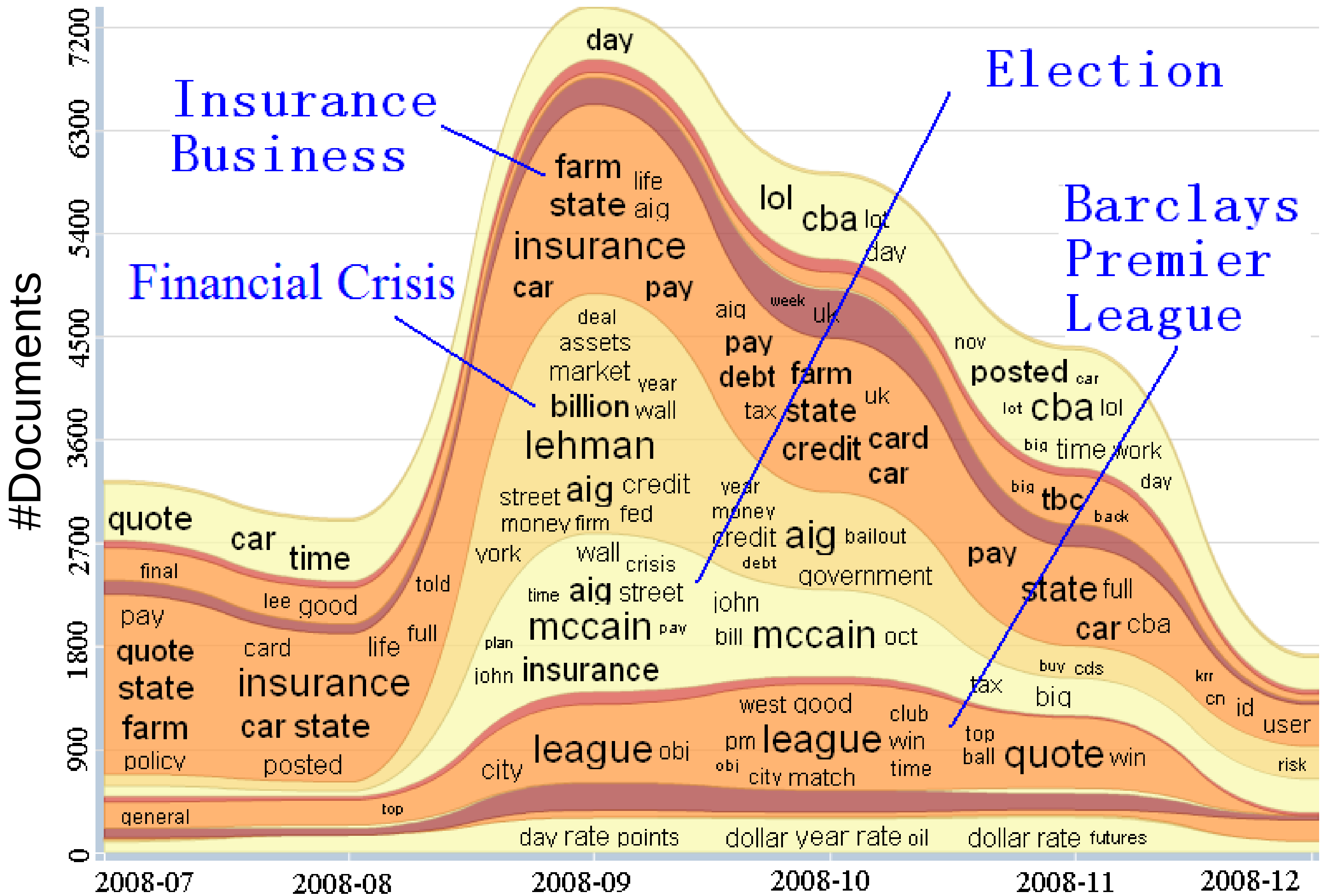


Election
Financial Crisis
Ads./Noise
#Documents
0
2000
4000
6000
8000
2008-07
2008-08
2008-09
2008-10
2008-11
2008-12
year
time
hl
lot
life
plan
merrill
bailout
street aig
lehman bill
fed billion year
credit wall crisis
york
obama
people aig
mccain
government
time make
os
lost
time
day
life
time
lot
car
cd car uk
debt credit
state debt
oil tax
hl
back oct
year
credit aig
money tax
big
porn
free
videos clip
video sex
site
girl videos teen
real
sitemap
clips
porn
clip
serial video free
drug
sex
phone
movie
game
online music torrent
auto gay
video
nude

News

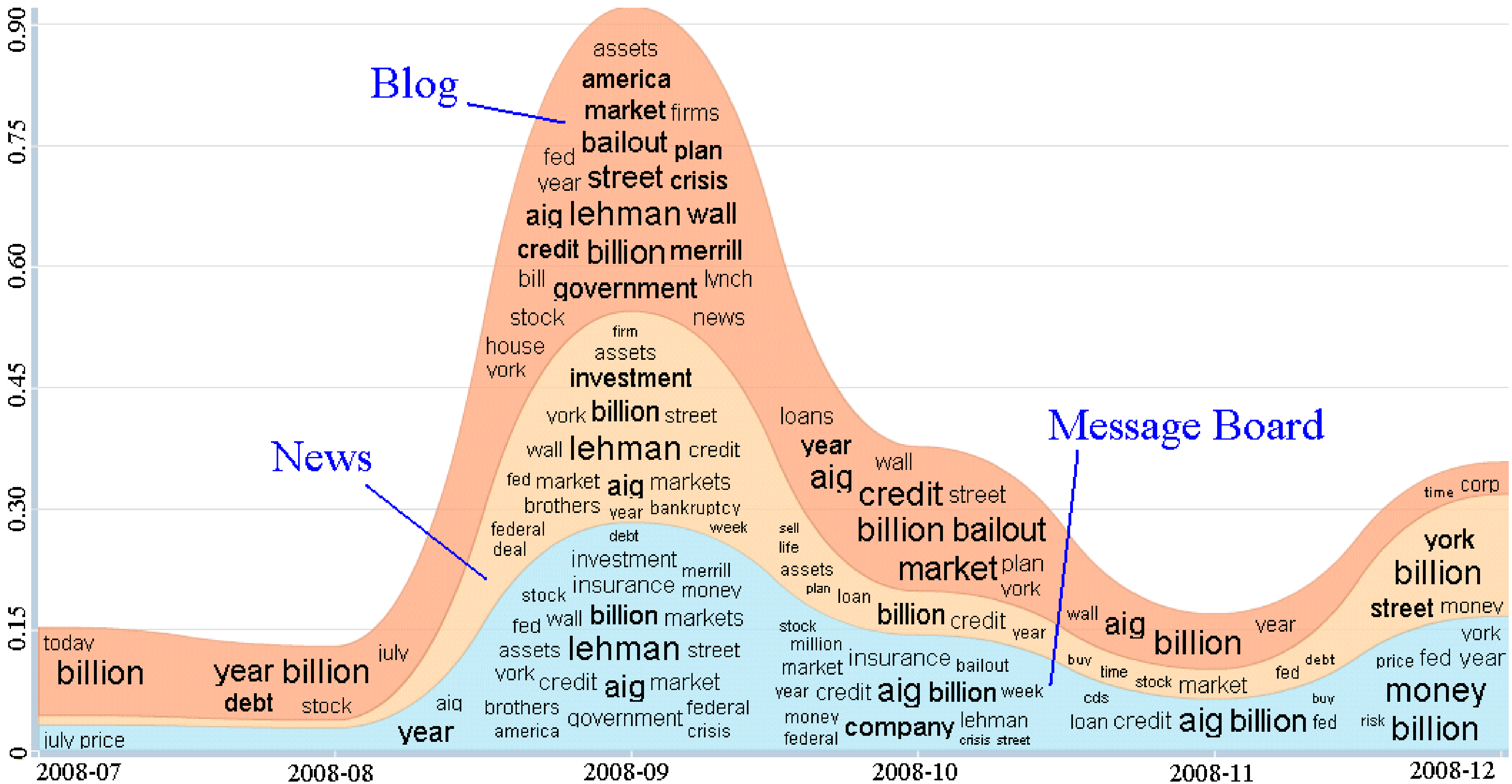

#Documents
18000
15000
12000
9000
6000
3000
0
2008-07
2008-08
2008-09
2008-10
2008-11
2008-12
Bank Related
Financial Crisis
Barclays Premier League
Trade Market Info
service
price
billion
lloyds
rate
uk hbos tsb
year services share
crisis
market
markets
wall aig credit
lehman fed
deal billion york
crisis year
stock
time
united time
top club city
year back
goal arsenal
billion
wall rose
year market news
fell percent oil
cents cent index
group prices
share
rate
year
cent hbos rbs
uk billion cut
tsb
lloyds capital
shares plan
plan
year
life loan sell
back club start
half league win
top side premier
rose
rate cents fell
cut cent oil
percent day
st
cut
club back city
points cut day
fell cent oil data
year dollar global
fell
top
rba cut
goal club city
win
crude
pre city
lost oil ax
cut fell cents oil
rose expected demand

Message Boards

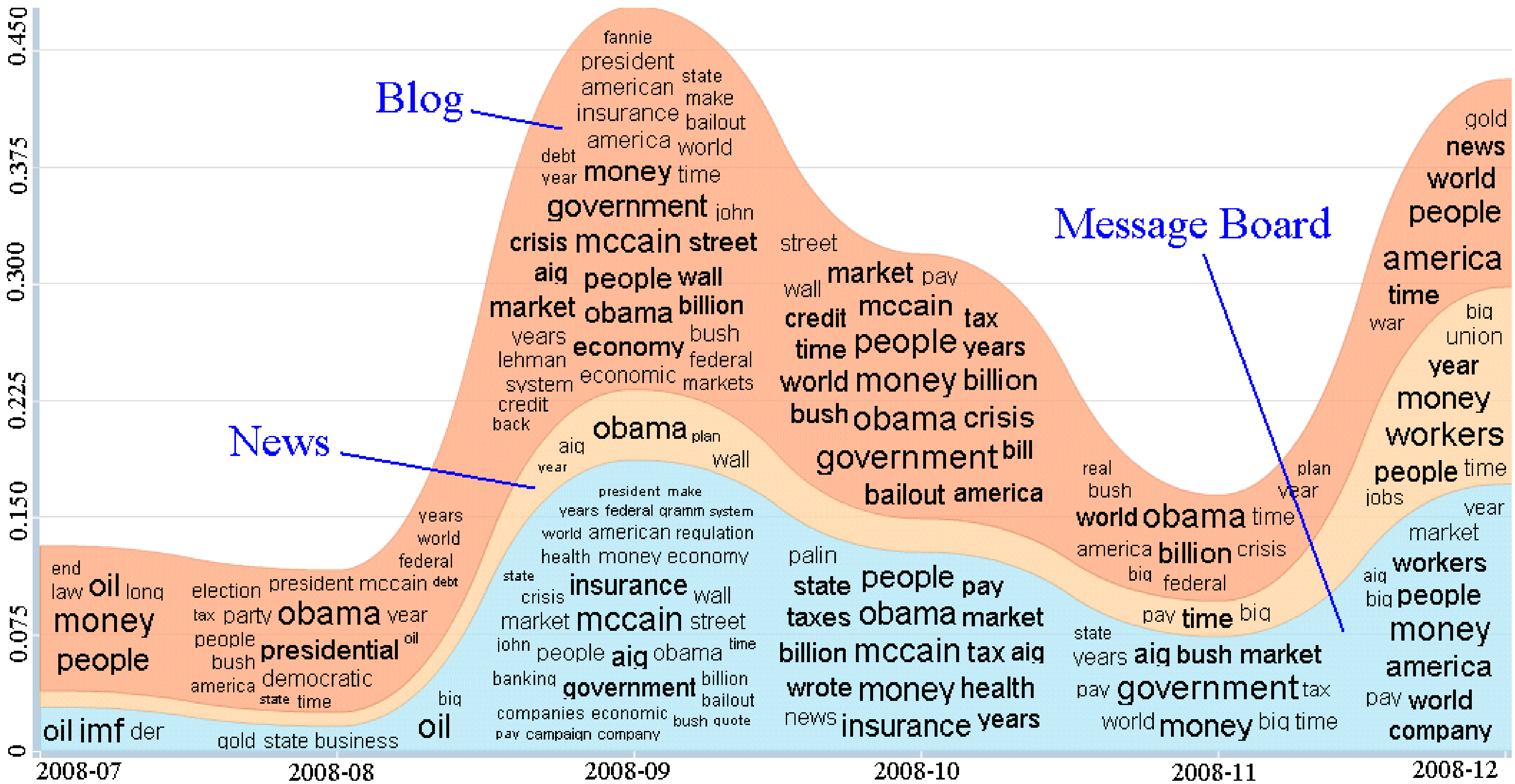


Clusters

Financial crisis $\pi^{t}_{j,k}$



Election $\pi^{t}_{j,k}$



Conclusions

- An EvoHDP model to mine cluster evolution patterns from multiple correlated time-varying corpora
- Extension of the original HDP
- Gibbs sampling
- Better predicting ability and stronger correlations across corpora overtime
- Cluster evolution patterns in real financial related web data


Thank You!!!