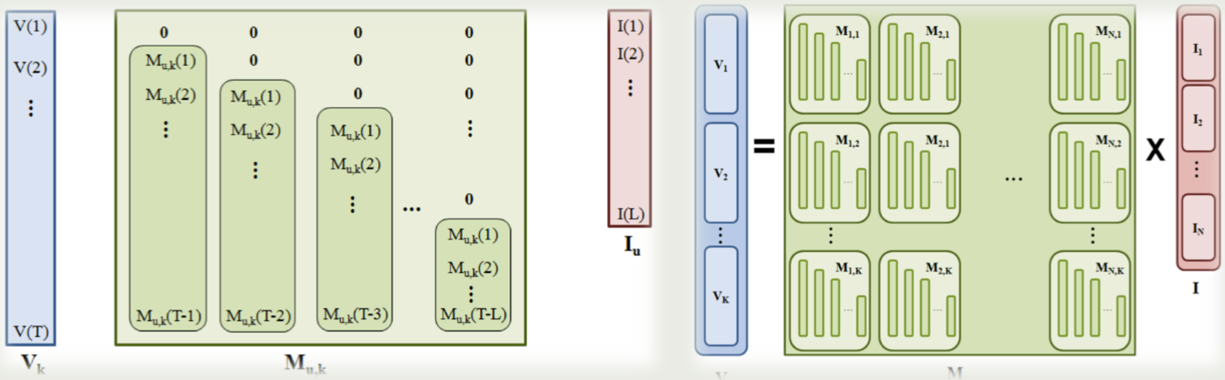
Modeling Information Diffusion in Implicit Networks.

Jaewon Yang， Jure Leskovec

IEEE International Conference On Data Mining (ICDM), 2010

Presenter: SHI, Conglei(clshi@cse.ust.hk)



PROBLEM

- There are some limitations for parameter estimation:
  - Need complete network data:

    FACT: Commonly , we only observe nodes got “infected”.
  - Contagion can only spread over the edges:

    FACT: The diffusion is not just depend on the social network.

METHODS

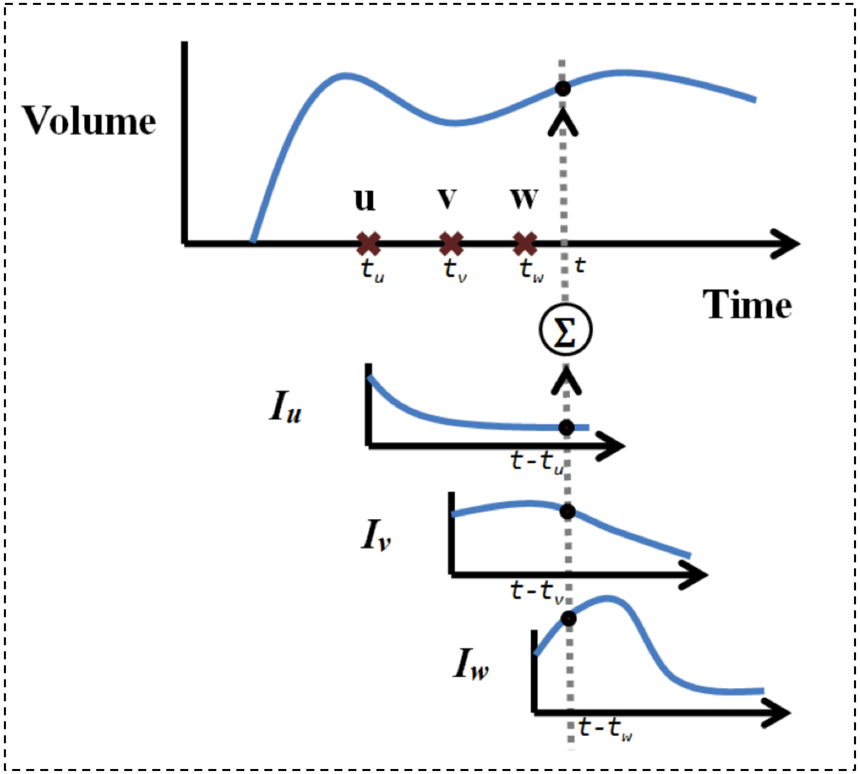
- Focusing on modeling the global influence a node has on the rate of diffusion through the implicit network.
  - Ignore the knowledge of the network
  - Also model how the diffusion unfold over time.
- Proposed Linear Influence Model(LIM)
  - Base Assumption: number of newly infected nodes depends on which other nodes got infected in the past.

LINEAR INFLUENCE MODEL

- $V(t)$: The number of nodes that mention the info at *t*
- $I_u(l)$: The Influence of the node *u* at time *t*

$$V(t+1) = \sum_{u \in A(t)} I_u(t - t_u)$$

- How to model $I_u(l)$?



MODELING INFLUENCE FUNCTION

- Parametric approach:

$$I_u(l) = c_u e^{-\lambda_u l} \qquad I_u(l) = c_u l^{-\alpha_u}$$

  - Too simplistic, assuming all the nodes follow the same form

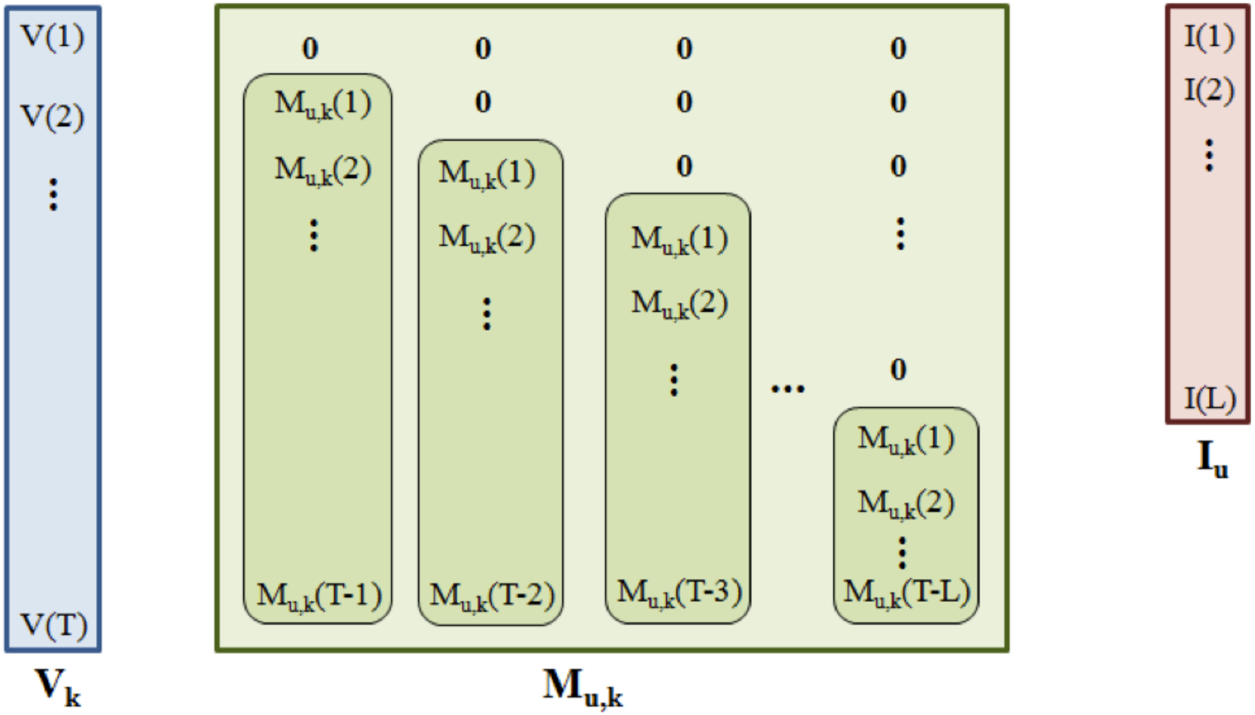
- Non-parametric approach:
  - Do not make any assumption about the shape of function
  - Represent the function as a non-negative vector of length *L*
  - Can study how the function varies for different types.

ESTIMATING FUNCTIONS

- Consider a set of *N* nodes, *K* contagions.
- Design an indicator function $M_{u,k}(t)$. If node u got infected by contagion *k* at time *t*, $M_{u,k}(t) = 1$.
- $V_k(t)$: The number of nodes that got infected by *k* at time *t*.

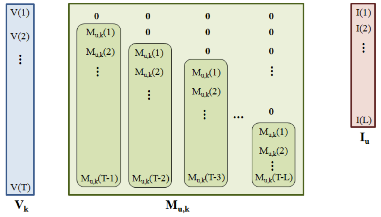
$$V_k(t+1) = \sum_{u=1}^{u=N} \sum_{l=0}^{l=L-1} M_{u,k}(t-l) I_u(l+1)$$

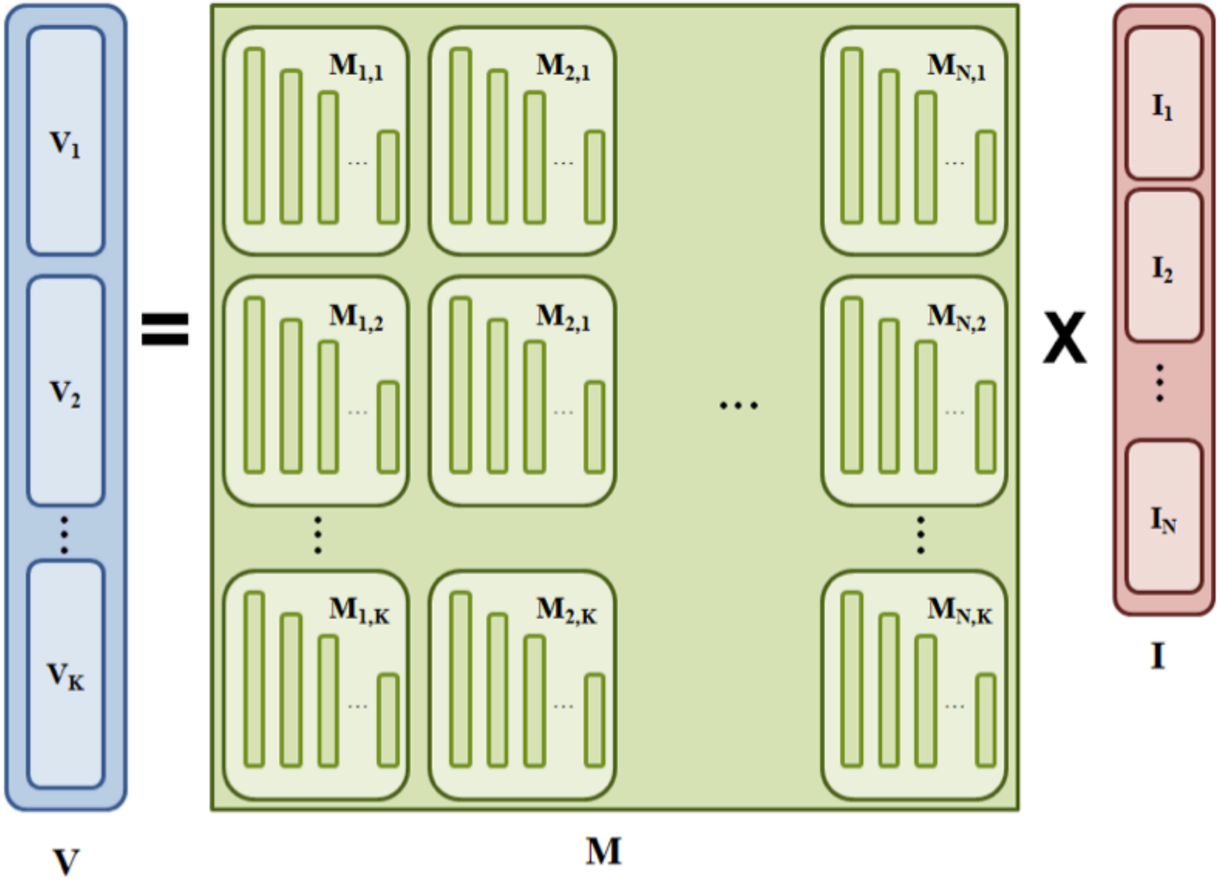
ESTIMATING FUNCTIONS



$$V_k(t+1) = \sum_{u=1}^{u=N} \sum_{l=0}^{l=L-1} M_{u,k}(t-l) I_u(l+1)$$

ESTIMATING FUNCTIONS





ESTIMATING FUNCTIONS

- Minimize

$$||\mathbf{V} - \mathbf{M} \cdot \mathbf{I}||_2^2$$

- Subject to

$$\mathbf{I} \geq 0$$

- This problem is called Non-negative Least Squares(NNLS) problem
- The Matrix *M* is sparse in nature
- Using Reflective Newton Method is very effective.
- Tikhonov regularization is also applied to smooth the estimates.

EXTENSIONS

- Accounting for novelty:
  - One node’s influence is related to the time it appears.
  - Introduce a multiplicative factor $\alpha(t)$.

$$V_k(t+1) = \alpha(t) \sum_{u=1}^{u=N} \sum_{l=0}^{l=L-1} M_{u,k}(t-l) I_u(l+1)$$

  - The equation is convex both $I_u(l)$ and $\alpha(t)$, which means we can use a coordinate descent procedure.

EXTENSIONS

- Accounting for imitation
  - Some information diffusion is the effect of imitation.
  - Introduce $b(t)$ to model the latent volume.

$$V_k(t+1) = b(t) + \sum_{u=1}^{u=N} \sum_{l=0}^{l=L-1} M_{u,k}(t-l) I_u(l+1)$$

  - Also linear.

EXPERIMENTS

- First datasets
  - Memetracker data: Extracting 343 million short textual phrases from 172 million news article and blog post.
  - Time period: Sep.1 2008 to Aug. 31 2009
  - Choosing 1000 phrases with highest volume in a 5 day window around their peak volume

EXPERIMENTS

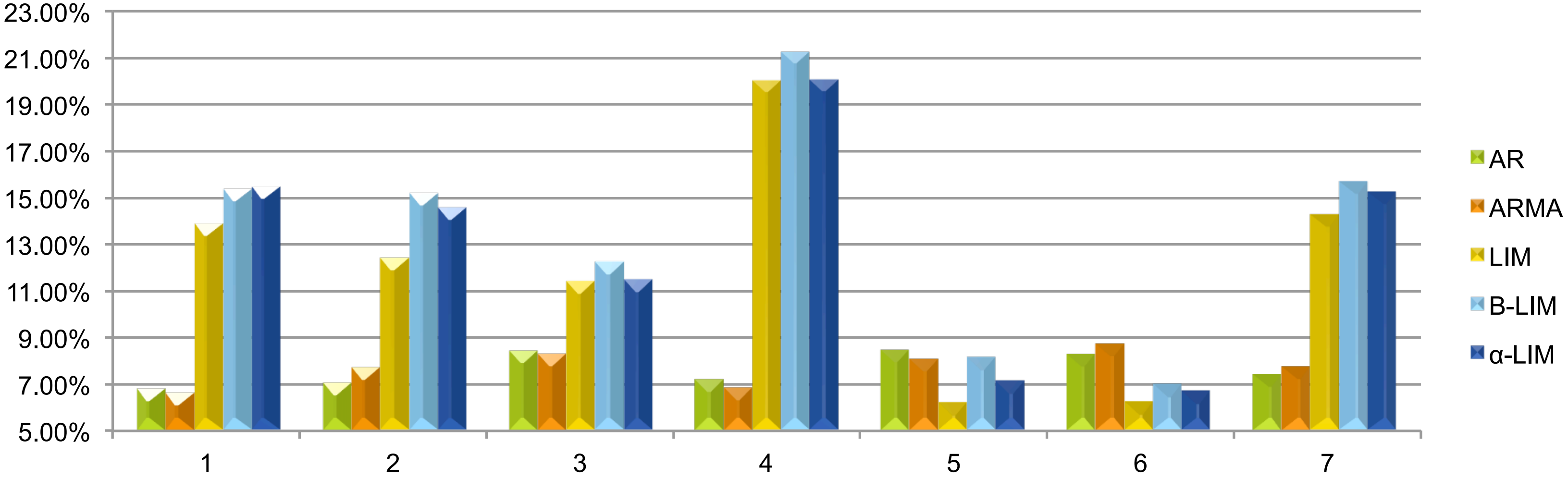
- Second datasets
  - Twitter data: Identifying 6 million different hashtags from a stream of 580 million Twitter posts.
  - Time period: Jun. 2009 to Feb. 2010
  - Choosing 1000 hashtags with highest volume in a 5 day window around their peak volume
  - Grouping users into groups of 100 users.

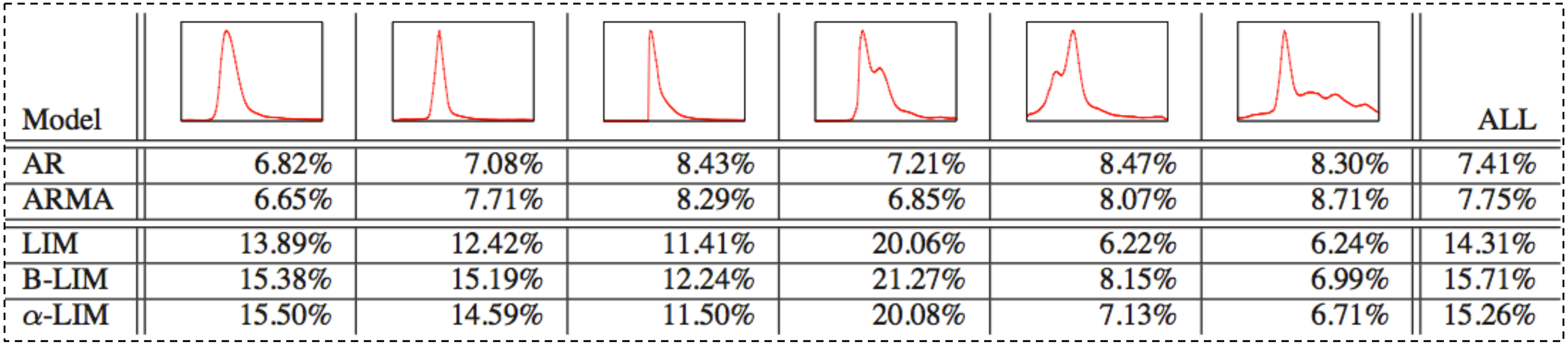
EXPERIMENTS

- Evaluate LIM model on a time series prediction task.
- Employ 10-fold cross validation.
- Calculate $E_k(t+1) = V_k(t+1) - \hat{V}_k(t+1)$
- Relative error is what we want.

$$\sqrt{\sum_{k,t} E_k(t)^2} / \sqrt{\sum_{k,t} V_k(t)^2}$$

RESULT

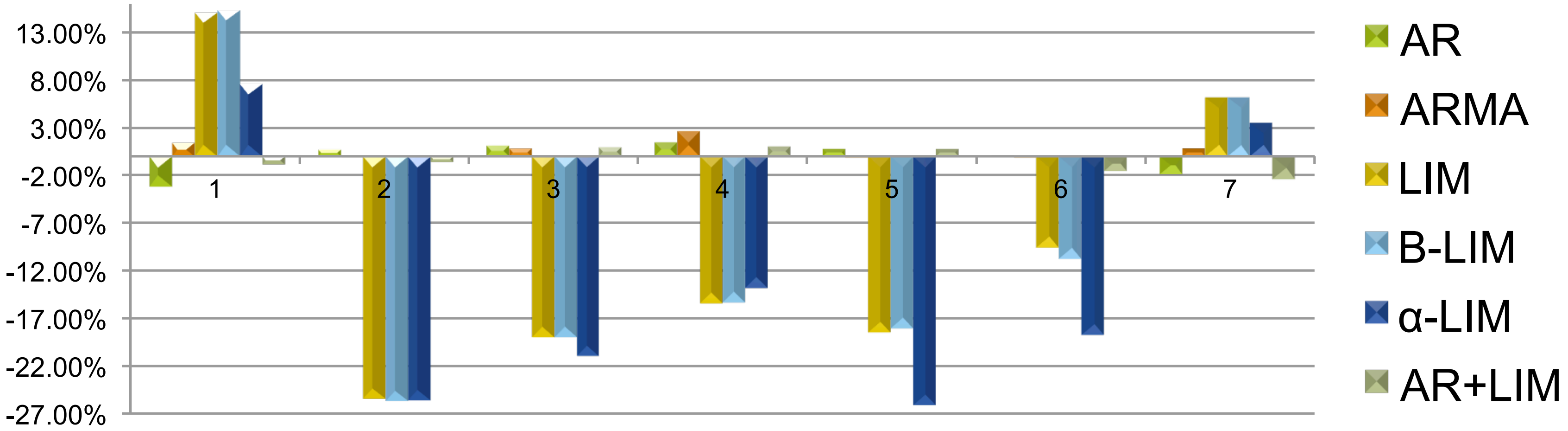


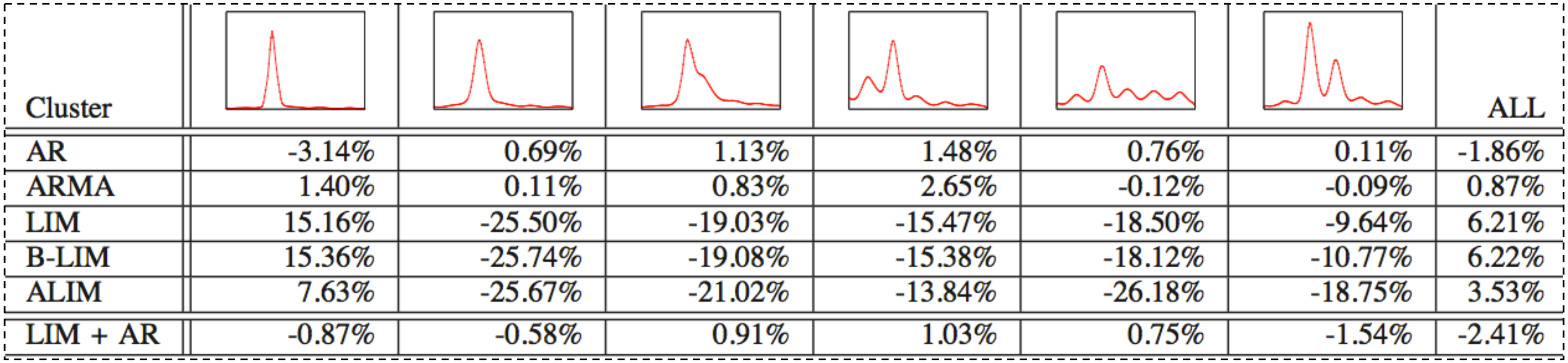


Model							ALL
AR	6.82%	7.08%	8.43%	7.21%	8.47%	8.30%	7.41%
ARMA	6.65%	7.71%	8.29%	6.85%	8.07%	8.71%	7.75%
LIM	13.89%	12.42%	11.41%	20.06%	6.22%	6.24%	14.31%
B-LIM	15.38%	15.19%	12.24%	21.27%	8.15%	6.99%	15.71%
α-LIM	15.50%	14.59%	11.50%	20.08%	7.13%	6.71%	15.26%

Yang, J., & Leskovec, J. Patterns of temporal variation in online media. (WSDM '11)

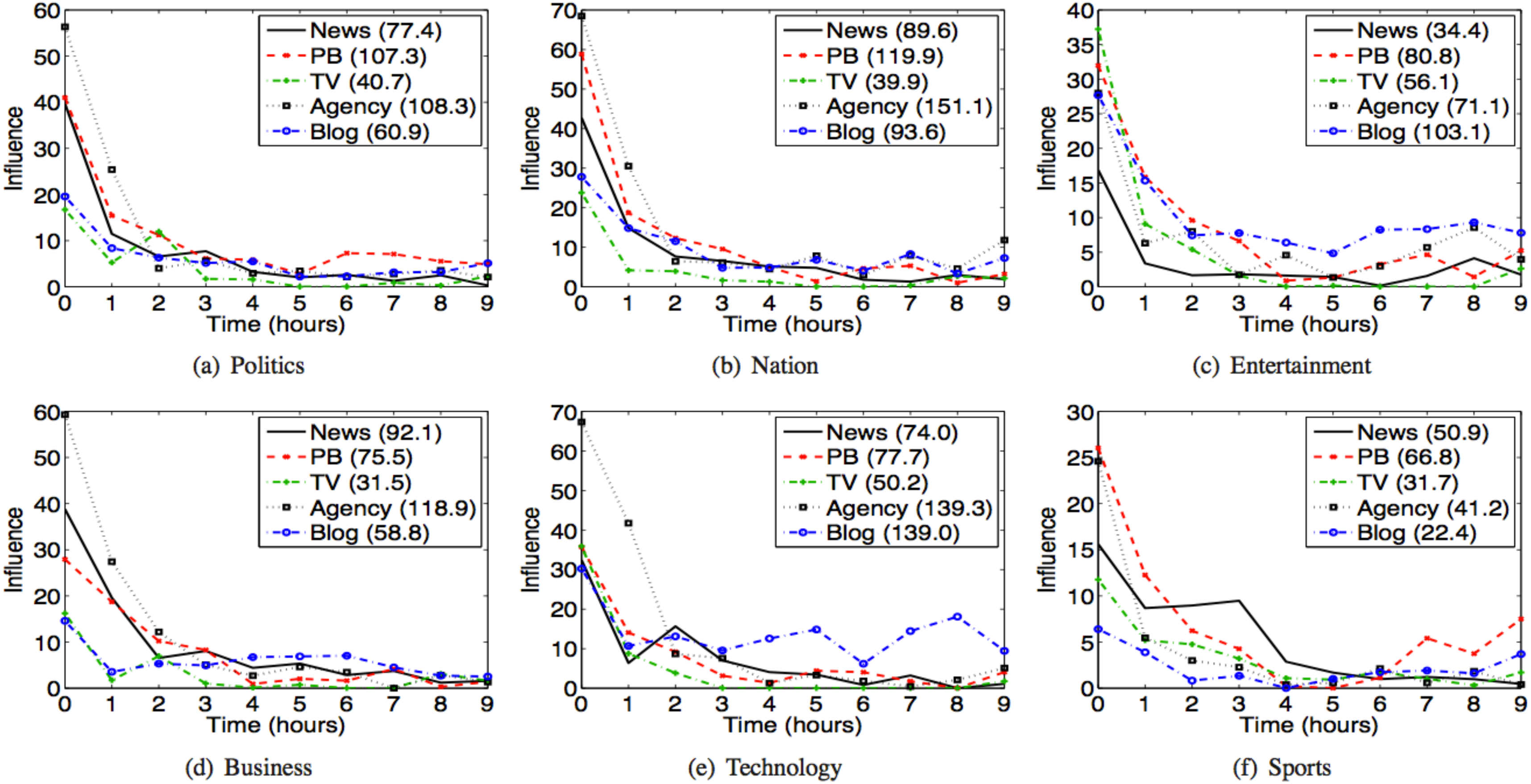
RESULT





Cluster							ALL
AR	-3.14%	0.69%	1.13%	1.48%	0.76%	0.11%	-1.86%
ARMA	1.40%	0.11%	0.83%	2.65%	-0.12%	-0.09%	0.87%
LIM	15.16%	-25.50%	-19.03%	-15.47%	-18.50%	-9.64%	6.21%
B-LIM	15.36%	-25.74%	-19.08%	-15.38%	-18.12%	-10.77%	6.22%
ALIM	7.63%	-25.67%	-21.02%	-13.84%	-26.18%	-18.75%	3.53%
LIM + AR	-0.87%	-0.58%	0.91%	1.03%	0.75%	-1.54%	-2.41%

RESULT



(a) Politics

(b) Nation

(c) Entertainment

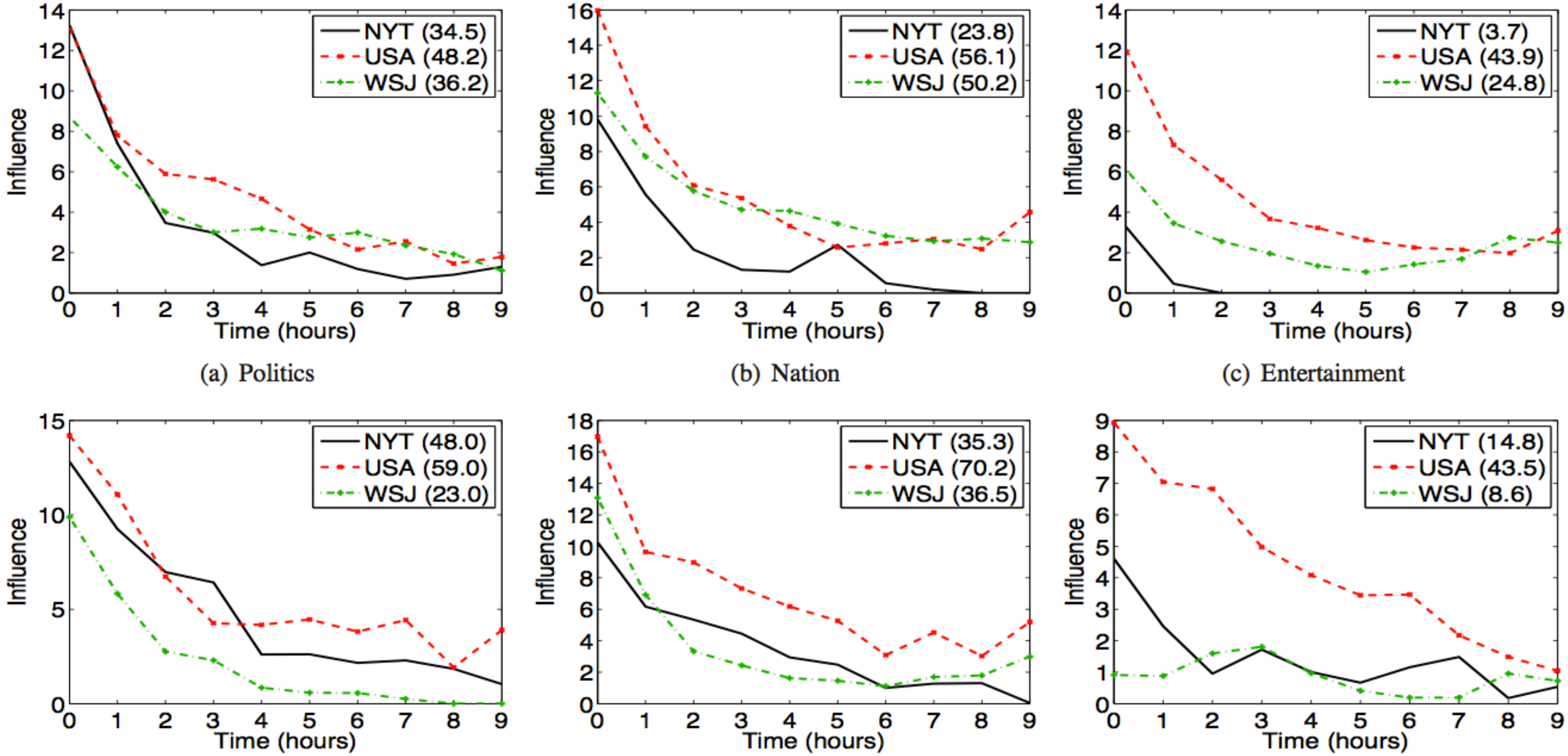
(d) Business

(e) Technology

(f) Sports

Figure 3. Average influence functions of five types of websites: Newspapers (News), Professional Blogs (PB), Television (TV), News Agencies (Agency), and Personal Blogs (Blogs). The number in brackets denotes the total influence of a media type.

RESULT



(a) Politics

(b) Nation

(c) Entertainment

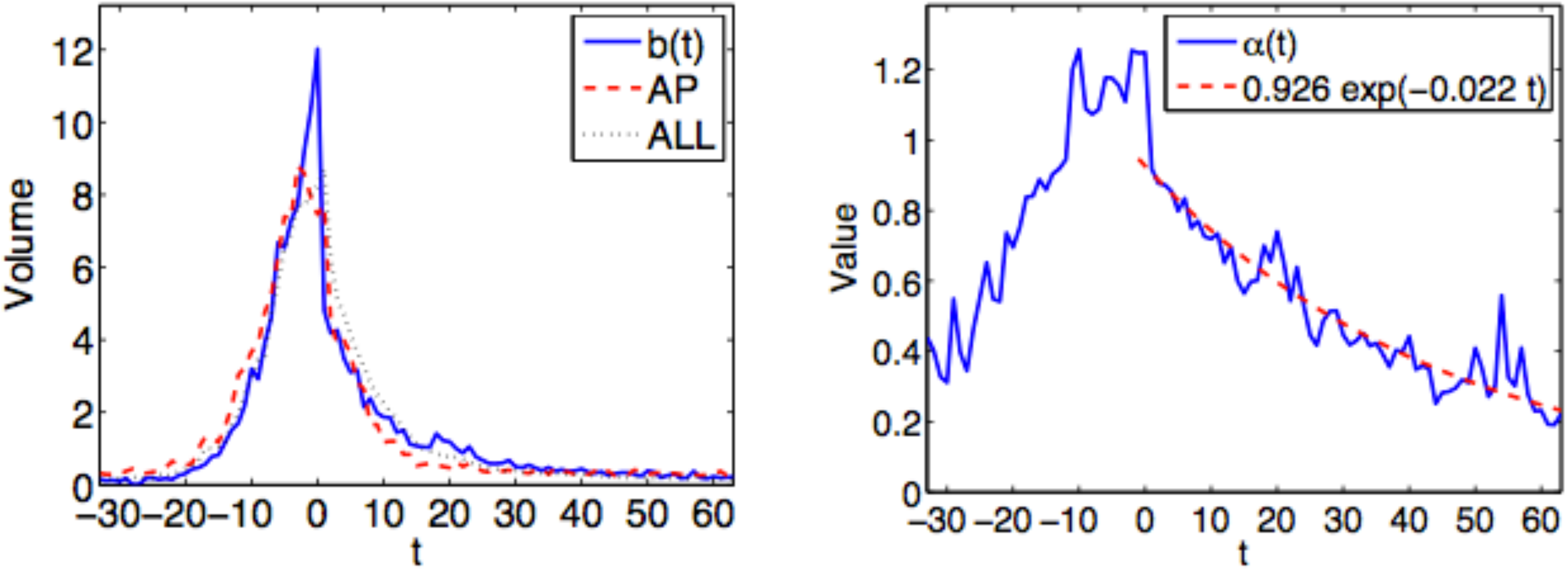
(d) Business

(e) Technology

(f) Sports

Figure 4. Influence functions of the New York Times (NYT), the Wall Street Journal (WSJ), and USA Today (USA).

RESULT

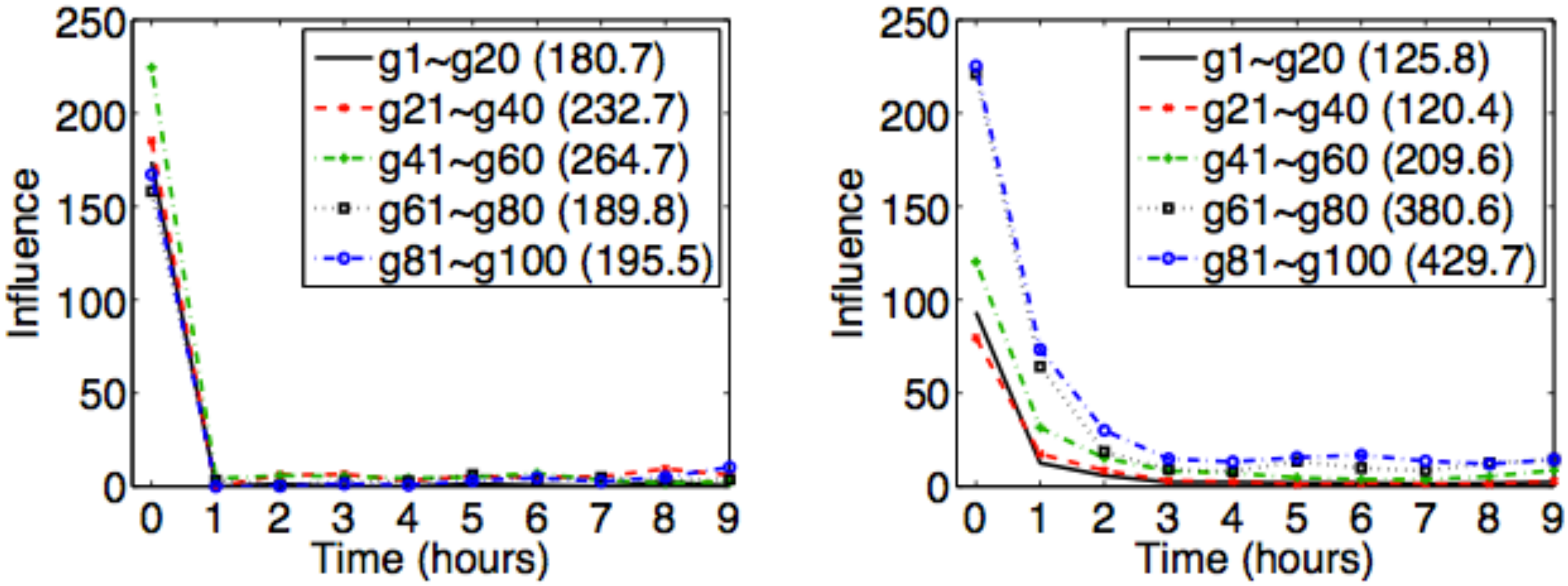


(a) Latent Volume in B-LIM

(b) Recency factor in α-LIM

Figure 5. Latent volume $b(t)$ and the recency factor $\alpha(t)$.

RESULT



(a) Average influence functions of each group based on volumes

(b) Average influence functions of each group based on follower counts

Figure 6. Influence functions of groups of Twitter users.

CONCLUSION

- Proposed the Linear Influence Model.
- Considered some other factors to enhance the model.
- Used large scale of data to justify the effectiveness of the model.
- Opened up a new framework for the analysis of diffusion.
  - Future work: extend the linear model to non-linear model.

THANKS FOR YOUR ATTENTION!