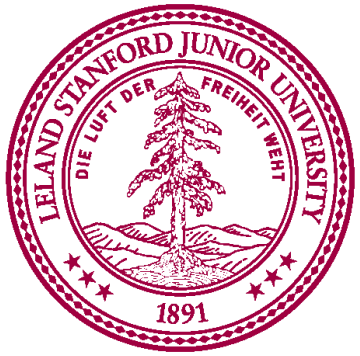
Patterns of Temporal Variation in Online Media

Jaewon Yang and Jure Leskovec
Computer Science
Stanford University



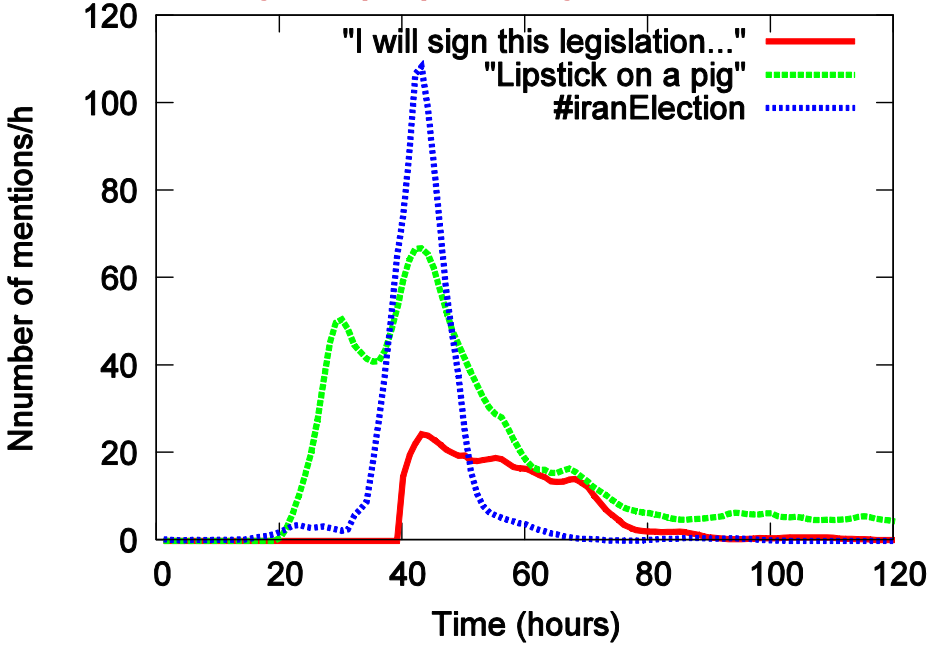
Motivation

- Web content is increasingly dynamic:
  - Social media and user-generated content further intensify this

- How does popularity of web content grow and fade over time?
  - Q1: What are typical patterns of the popularity of a Web content over time?
  - Q2: How do these patterns arise?

Patterns of Popularity

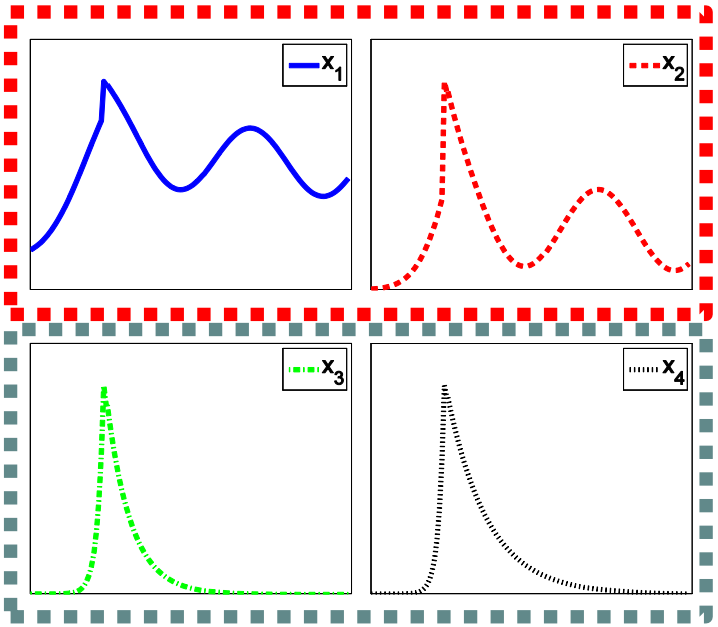
- Rich variability in popularity online content



  - Item i: Piece of content (*e.g.*, phrase, URL, hashtag)
  - Popularity/Volume $x_i(t)$: # of times item *i* mentioned
  - Shape of $x_i(t)$: How volume of item *i* varies over time
- Q: What are typical classes of shapes of $x_i(t)$?

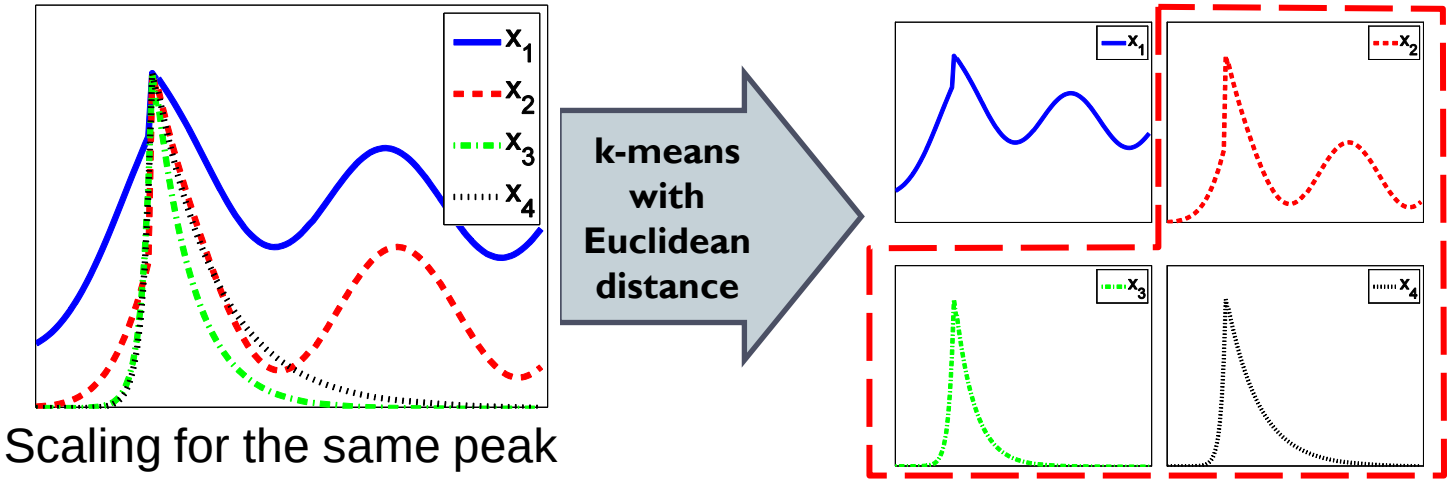
Problem Definition

- Given the volume time series $x_i(t)$
- Goal: Discover types of shapes of $x_i(t)$, i.e., cluster time series $x_i(t)$ by shape



First try: Euclidean + k-means

- Normalizing time series and using standard time series measures (Euclidean Distance or Dynamic Time Warping) can yield wrong results



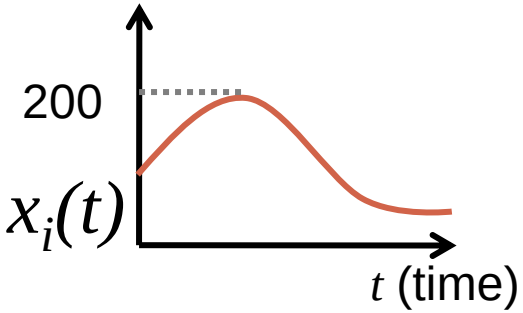
Scaling for the same peak

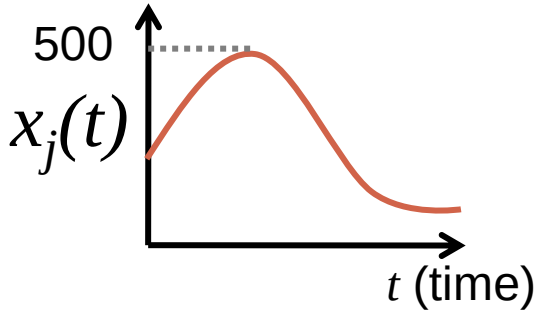
Outline of the Talk

- Goal:
  - Cluster volume time series by shape
- Approach:
  - Distance measure for the shape of time series
  - K-SC: a k-means-like algorithm to cluster time series by their shape
- Experiments
  - Popularity of short phrases and twitter hashtags
  - 6 patterns of temporal variation
  - Predict the temporal pattern of an item based on who mentioned it

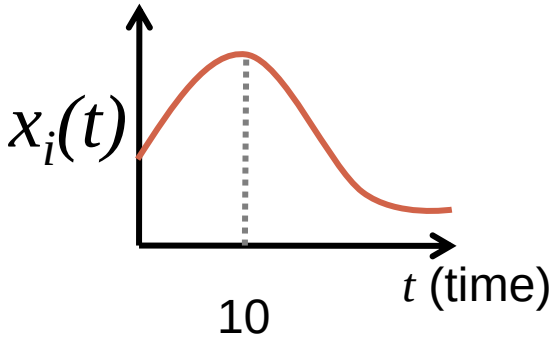
Time Series Distance Measure

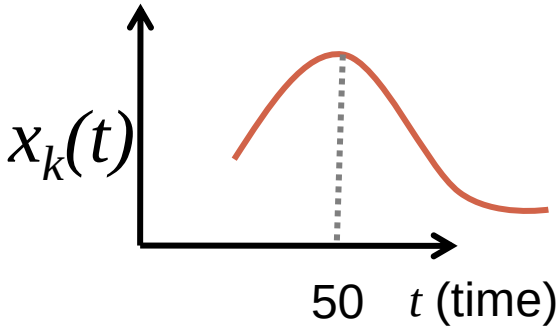
- How to compare volume time series of different items?
  - Invariance to scaling





  - Invariance to translation



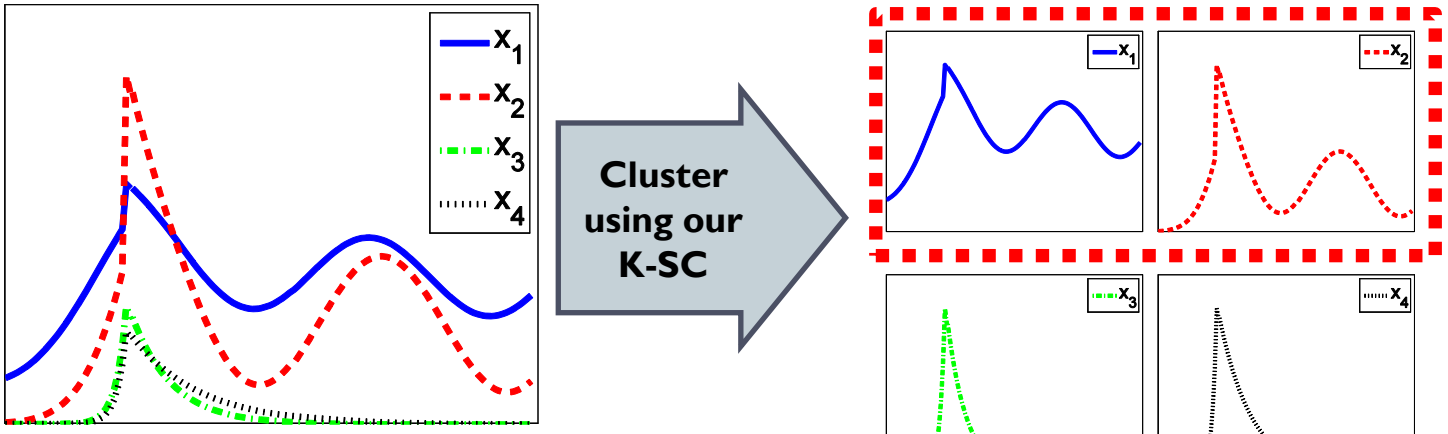


SSI Distance Measure

- **I**nvariant to **S**caling and **S**hifting:

$$\hat{d}^2(x_i(t), x_j(t)) = \min_{\alpha,q} \frac{\sum_t (x_i(t) - \alpha x_j(t-q))^2}{\sum_t x_i(t)^2}$$

  - Find best scaling α and shifting q of time-series x_i and x_j



Optimal scaling and shifting of (x_1, x_2) and (x_3, x_4)

K-Spectral Centroid (K-SC)

- Goal:
  - Cluster time series using SSI distance measure
  - Find typical shapes (i.e., **centroids**) of time series
- k-means is not applicable for SSI distance
  - Centroid is a point that is at minimum distance from all points in the cluster
  - K-means computes the centroid as the average of the points in the cluster
    - minimizes the sum of **Euclidean (not SSI) distances**
- K-Spectral Centroid (K-SC) algorithm
  - k-means-like iterative algorithm
  - Finds centroids under the SSI distance measure

K-SC: Similar to k-Means

- K-SC has same input and output as k-Means
  - Input: Time series x_i, and the number of clusters k
  - Output: Clusters $C_1 \ldots C_k$, Centroids $\mu_{1,\ldots,} \mu_k$
    - C_k : set of time series that belong to cluster k
    - μ_k : typical shape -- time series closest to all the time series in cluster k
- K-SC iteratively finds cluster centroids μ_k and cluster memberships C_k
  - Step 1: Given μ_k , assign points to clusters C_k
  - Step 2: Given clusters C_k , update centroids μ_k

K-SC: Finds better shapes

- Finding cluster memberships C_k:
  - Assign each x_i to the closest μ_κ
- Finding centroid μ_κ:
  - K-SC finds the cluster centroid μ_κ that is the closest to all the time series in C_k **under the SSI distance measure**
  - μ_κ is the typical shape of the time series in cluster C_k

K-SC: Finding μ_k

- SSI distance:

$$\hat{d}(x_i(t), \mu(t))^2 = \min_{\alpha,q} \frac{\sum_t (\mu(t) - \alpha x_i(t-q))^2}{\sum_t \mu(t)^2} = \min_{\alpha,q} \frac{||\mu - \alpha x_{i(q)}||}{||\mu||}$$

- Centroid: $\mu_k^* = \arg\min_{\mu} \sum_{x_i \in C_k} \hat{d}(x_i, \mu)^2$
- For each cluster we compute $M_k = \sum_{x_i \in C_k} (I - \frac{x_i x_i^T}{||x_i||^2})$
- Then:

$$\mu_k^* = \arg\min_{\mu} \frac{\mu^T M_k \mu}{||\mu||^2}$$

  - has a closed form solution!

    μ_k is the eigenvector of M_k *corresponding to the smallest eigenvalue*

Experiments: Datasets

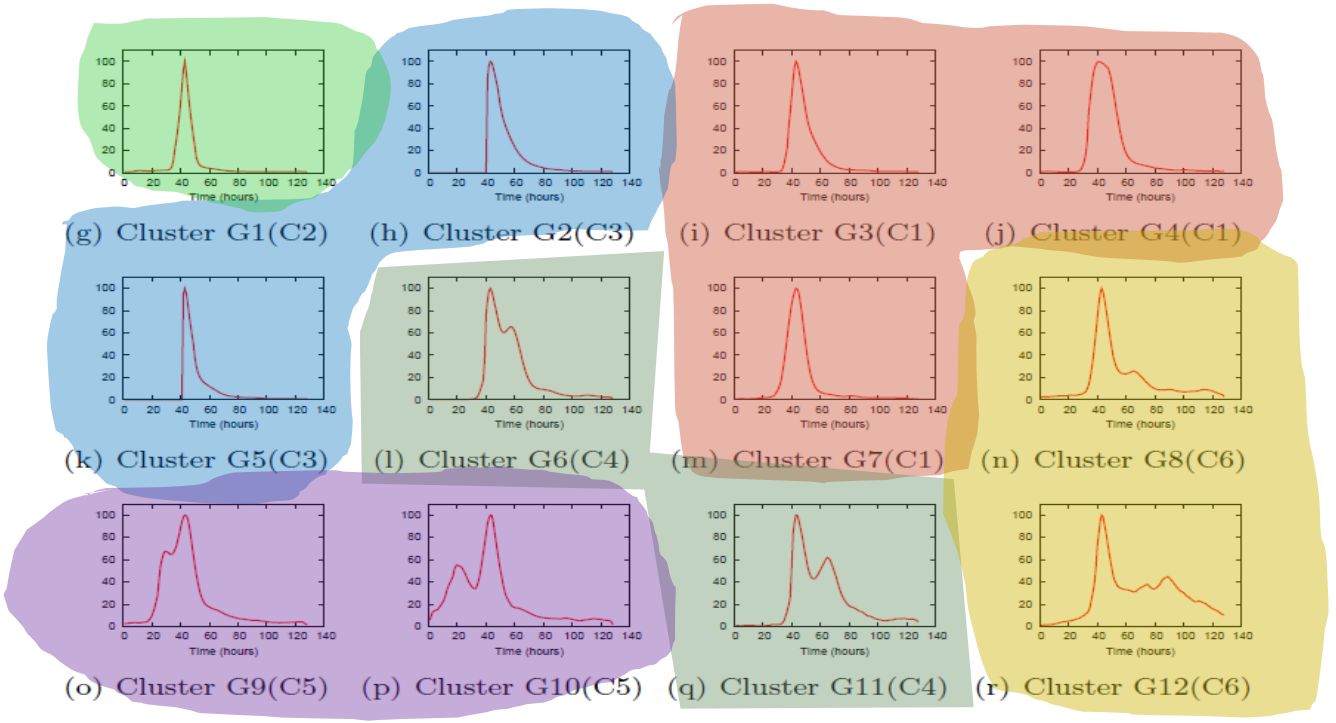
- Memetracker [Leskovec et.al. '09]:
  - Diffusion of short textual phrases on the Web
  - 172M news articles and blog posts
  - Items: 343M short textual phrases
- Twitter:
  - Adoption of Twitter hashtags
  - 580M Twitter posts
  - Items: 6M hashtags

Data Preparation

- Time Unit: 1 hour
- Measure volume of an item during 5 days around its peak volume
  - Most time series are spiky
  - More variation happens after the peak than before the peak
- Choose top 1,000 time series by total volume during the 5 day period

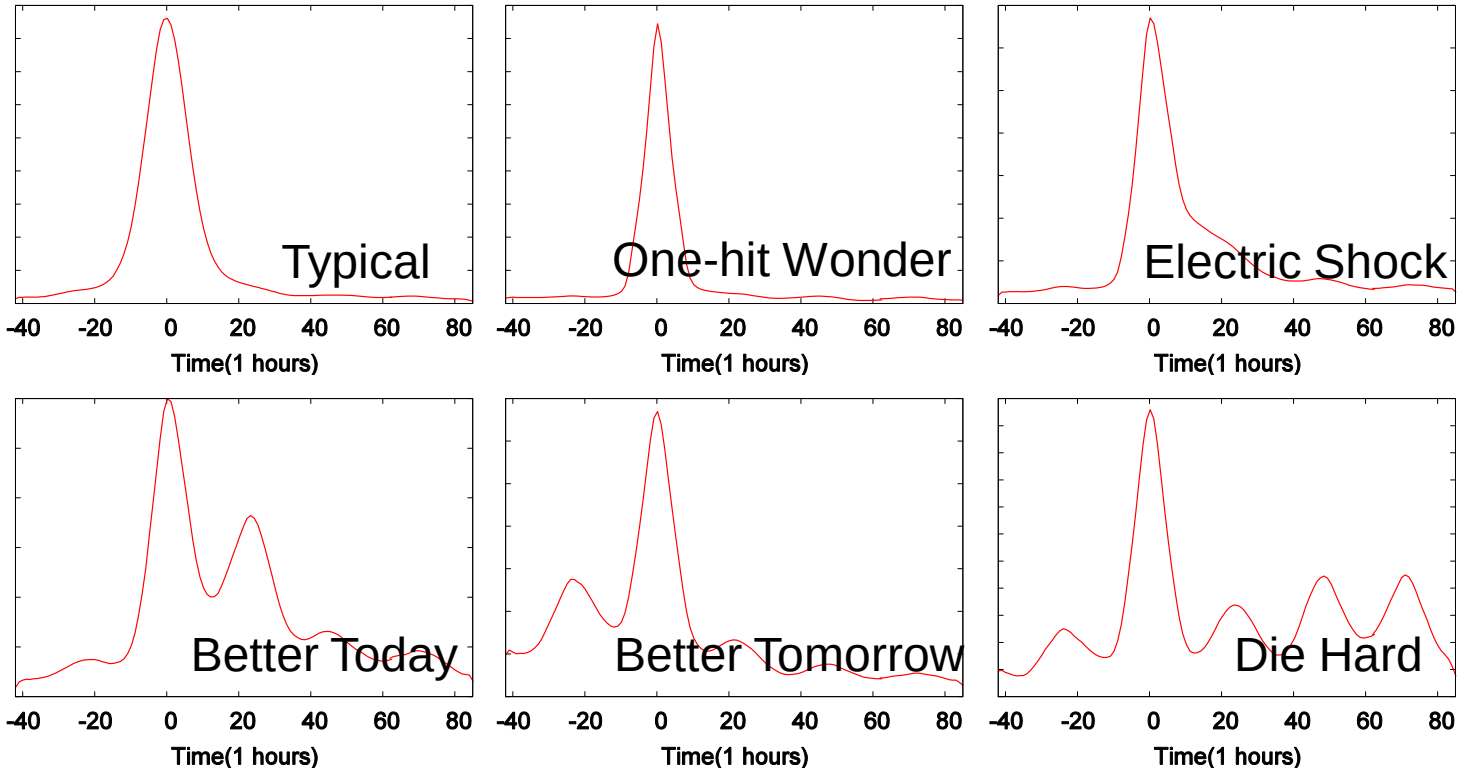
How Many Clusters? **6!**

- K-SC requires the number of clusters (K)
  - Hartigan's Index and Average Silhouette: K=6
  - When K>6, we find replicas of the clusters of K=6



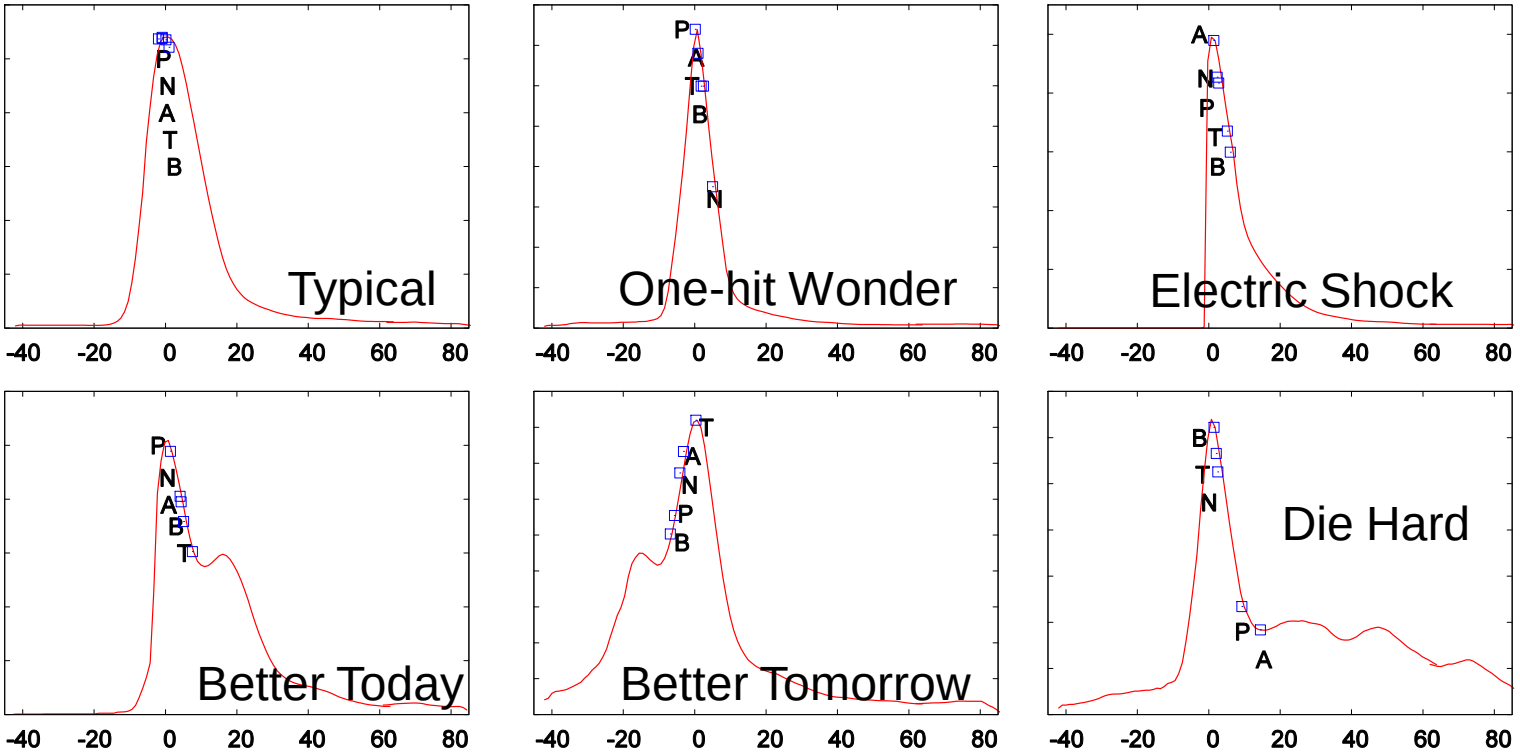
6 Patterns in Twitter Hashtags

- Twitter hashtag volume centroids for K=6



6 Patterns in Memetracker

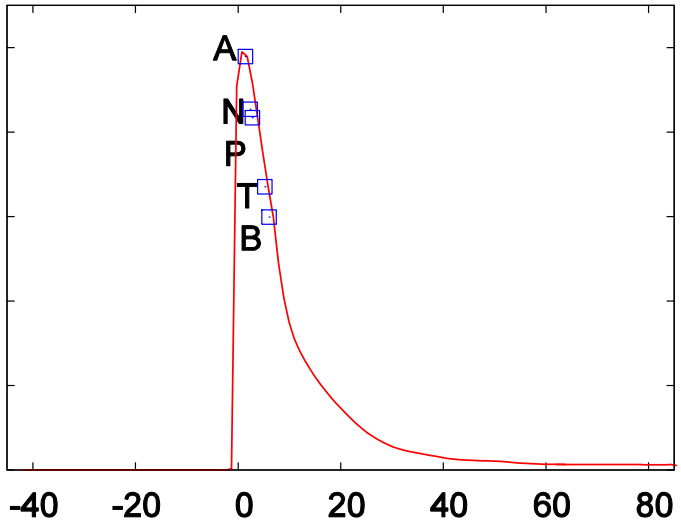
- Memetracker centroids for K=6

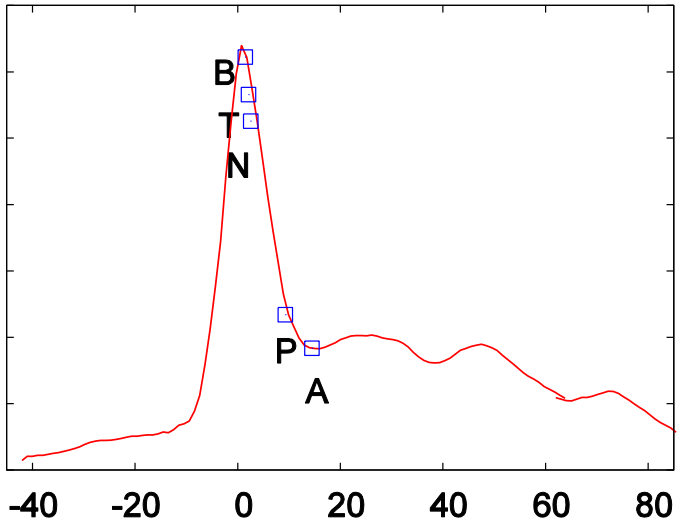


5 media types: **N**: Newspapers, **P**: Professional Blogs, **A**: News Agencies, **T**: Television, **B**: Personal Blogs

□ : When a media type typically mentions phrases in that cluster

Analysis of Clusters





Electric Shock	Die Hard
•Spike by News **A**gencies (AP, Reuters) •Slow and small blog response • Blogs mention 1.3 hours after News media • Blog volume = 29.1% of total volume	• Only cluster that is dominated by **B**loggers both in time and volume • Blogs mention 20 min ***before*** New media • Blog volume = 53.1% of total volume

Predicting Patterns

- Q1: Can we predict the popularity pattern of an item by which media mention it?
  - Classification task of whether a phrase belongs to a cluster or not
- Q2: If so, what kind of information should we look at? (the time of mention, the number of mentions, …)
  - Use different kind of features, and compare the classification accuracy

Predicting Patterns

- Goal: Classify the popularity pattern of an item based on who mentioned it
- Setup:
  - Pick *W* (|*W*|=50) highest volume websites
  - For each item record when sites W mention it
  - Design 3 different feature vectors:
    - TF-IDF: Think of Websites=words, phrases=documents
    - Volume: How many times a website mentions the phrase
    - Temporal: When the website first mentions the phrase
  - Train a logistic regression model

Classification Accuracy

#Websites	50	100	200
TF-IDF	70.1%	77.1%	87.0%
Volume	70.7%	77.1%	86.6%
Temporal	**76.6%**	**81.2%**	**88.7%**

- TF-IDF: Website (word) – phrase (document)
- Volume: How many times the website mentions the phrase
- Temporal: When the website first mentions the phrase (BEST)

- Based on when and which of the 50 websites mentions the phrase, we can predict the shape of volume time series of a phrase!
  - We can also predict the future volume of a phrase [ICDM 10']

Conclusion

- We proposed K-SC
  - K-SC can cluster time series by their shape
  - K-SC finds common shapes of the clusters.
- We found 6 common patterns of temporal variation of popularity of online content:
  - We found the same patterns in 2 different data sets
  - Different types of media give rise to different popularity patterns
  - We reliably classified patterns using the information of 50 websites

Thank You!

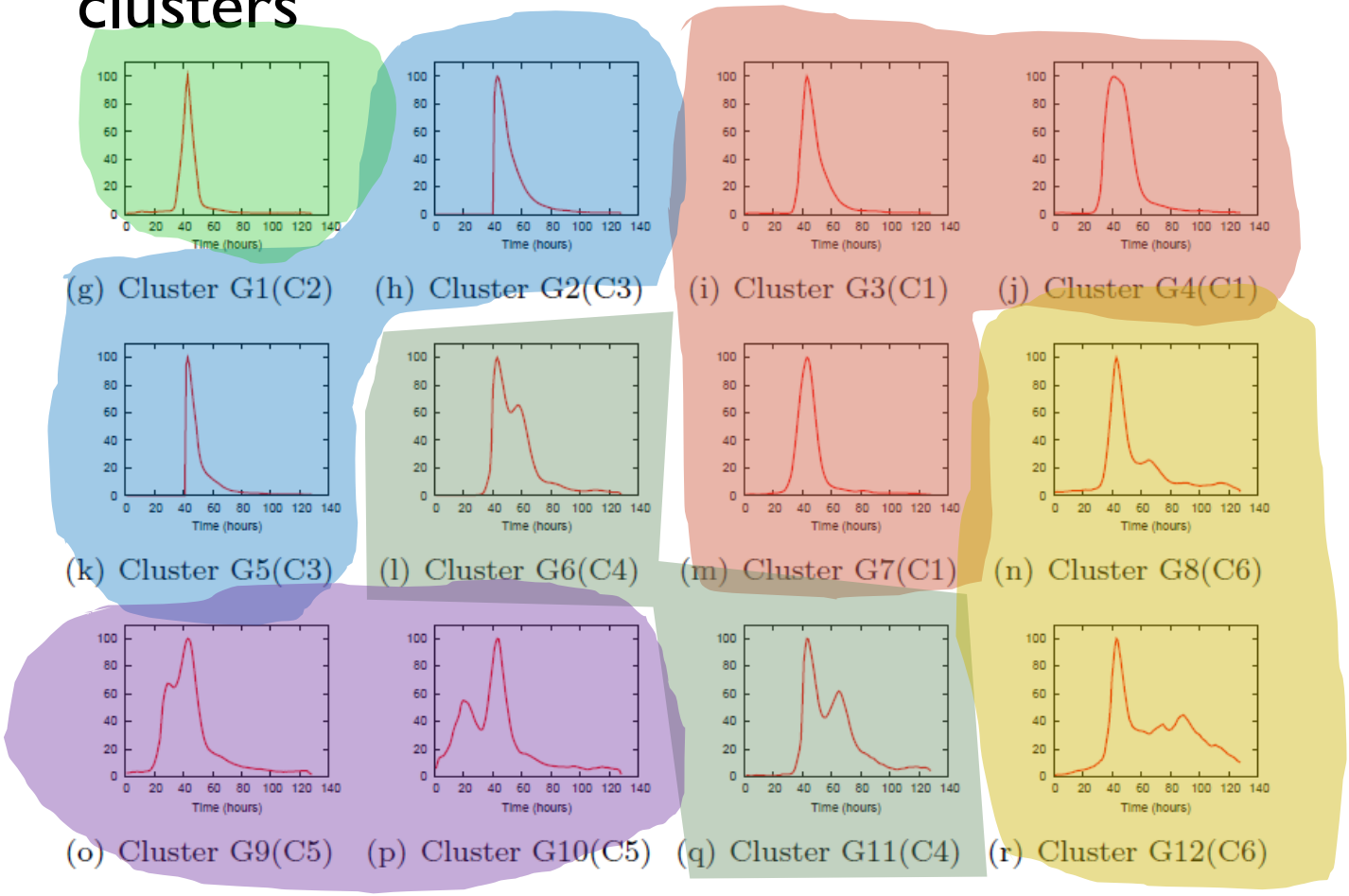
More in the paper:

- Analysis on the patterns in the adoption of Twitter hashtags
- Incremental K-SC using Haar Wavelets to deal with large scale data

Code and data: http://snap.stanford.edu

What if K>6?

- K-SC with K=12 found the replicas of the 6 clusters



K-SC centroid is robust to outliers

- We find the centroid of the cluster of single-peak time series with one outliar
  - K-Means (KM): We take the average of the time series
  - K-SC: We find μ_k^* by minimzing $\sum_{x_i \in C_k} \hat{d}(x_i, \mu)^2$
- K-SC Centroid finds the common shape