# TopicSketch: Real-time Bursty Topic Detection from Twitter

**Wei Xie**, Feida Zhu, Jing Jiang, Ee-Peng Lim and Ke Wang*
Living Analytics Research Centre
Singapore Management University

* Ke Wang is from Simon Fraser University, and this work was done when the author was visiting Living Analytics Research Centre in Singapore Management University.

# Twitter as News Media

- Twitter works as a huge news media.
- For some topics, especially bursty topics, news first appears in Twitter, rather than traditional news media.
- It is interesting and also useful to detect bursty topics from Twitter.

# Handling Tweet Stream is Challenging

- **Large Volume**
  Number of tweets per day : 340 million

- **Large Velocity**
  Number of tweets per second : 9,000 (average) / 143,000 (peak)

- **Large Variety**
  All kinds of activities and topics appear in Twitter

# Outline

- Motivation
- Related Work
- Proposed Method
  - Intuition
  - Indicator of burst
  - Assumptions
  - Solution
  - Framework
  - Dimension reduction
- Experiment
- Conclusion

# Related Work

# Related Work

- **Topic Modelling**
  —Liangjie Hong, et al. A time-dependent topic model for multiple text streams. KDD 2011
  —Qiming Diao, et al. Finding Bursty Topics from Microblogs. ACL 2012

# Related Work

- **Topic Modelling**
  —Liangjie Hong, et al. A time-dependent topic model for multiple text streams. KDD 2011
  —Qiming Diao, et al. Finding Bursty Topics from Microblogs. ACL 2012

- **Topic Detection & Tacking**
  —Sasa Petrovic, et al. Streaming First Story Detection with application to Twitter. HLT-NAACL 2010
  —Chenliang Li, et al. Twevent: segment-based event detection from tweets. CIKM 2012

# Related Work

- **Topic Modelling**
  —Liangjie Hong, et al. A time-dependent topic model for multiple text streams. KDD 2011
  —Qiming Diao, et al. Finding Bursty Topics from Microblogs. ACL 2012

- **Topic Detection & Tacking**
  —Sasa Petrovic, et al. Streaming First Story Detection with application to Twitter. HLT-NAACL 2010
  —Chenliang Li, et al. Twevent: segment-based event detection from tweets. CIKM 2012

**Both of them face difficulty to handle large tweet stream, as they need to process very huge historical data.**



LARC
LIVING ANALYTICS RESEARCH CENTRE

# Intuition

# Intuition

- **Rather than keep the big historical data, maybe we can take a snapshot of the current data stream.**

# Intuition

- **Rather than keep the big historical data, maybe we can take a snapshot of the current data stream.**
- **At least, it takes much smaller space and hopefully we can efficiently infer topics from it.**

# Intuition

- **Rather than keep the big historical data, maybe we can take a snapshot of the current data stream.**

- **At least, it takes much smaller space and hopefully we can efficiently infer topics from it.**

# But How?

# Acceleration as an Indicator

Adopt the concepts in physics:

# Acceleration as an Indicator

Adopt the concepts in physics:

**Velocity**

# Acceleration as an Indicator

Adopt the concepts in physics:

**Velocity** the rate of change of the volume of tweet stream

# Acceleration as an Indicator

Adopt the concepts in physics:

**Velocity** the rate of change of the volume of tweet stream

$$v = \frac{\Delta x}{\Delta t}$$

# Acceleration as an Indicator

Adopt the concepts in physics:

**Velocity** the rate of change of the volume of tweet stream

$$v = \frac{\Delta x}{\Delta t}$$

**Acceleration**

# Acceleration as an Indicator

Adopt the concepts in physics:

**Velocity** the rate of change of the volume of tweet stream

$$v = \frac{\Delta x}{\Delta t}$$

**Acceleration** the rate of change of the velocity of tweet stream



LARC
LIVING ANALYTICS RESEARCH CENTRE

# Acceleration as an Indicator

Adopt the concepts in physics:

**Velocity** the rate of change of the volume of tweet stream

$$v = \frac{\Delta x}{\Delta t}$$

**Acceleration** the rate of change of the velocity of tweet stream

$$a = \frac{\Delta v}{\Delta t}$$

LARC
Living Analytics Research Centre

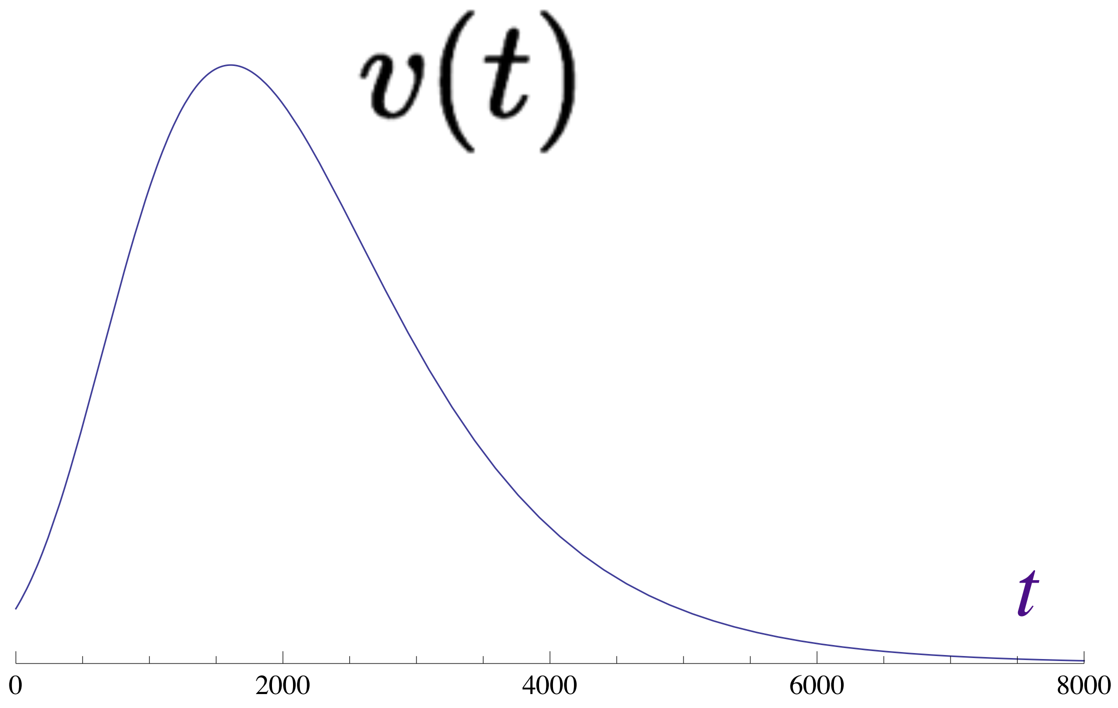
# Acceleration as an Indicator

**Acceleration: a very good early indicator of burst.**

# Acceleration as an Indicator

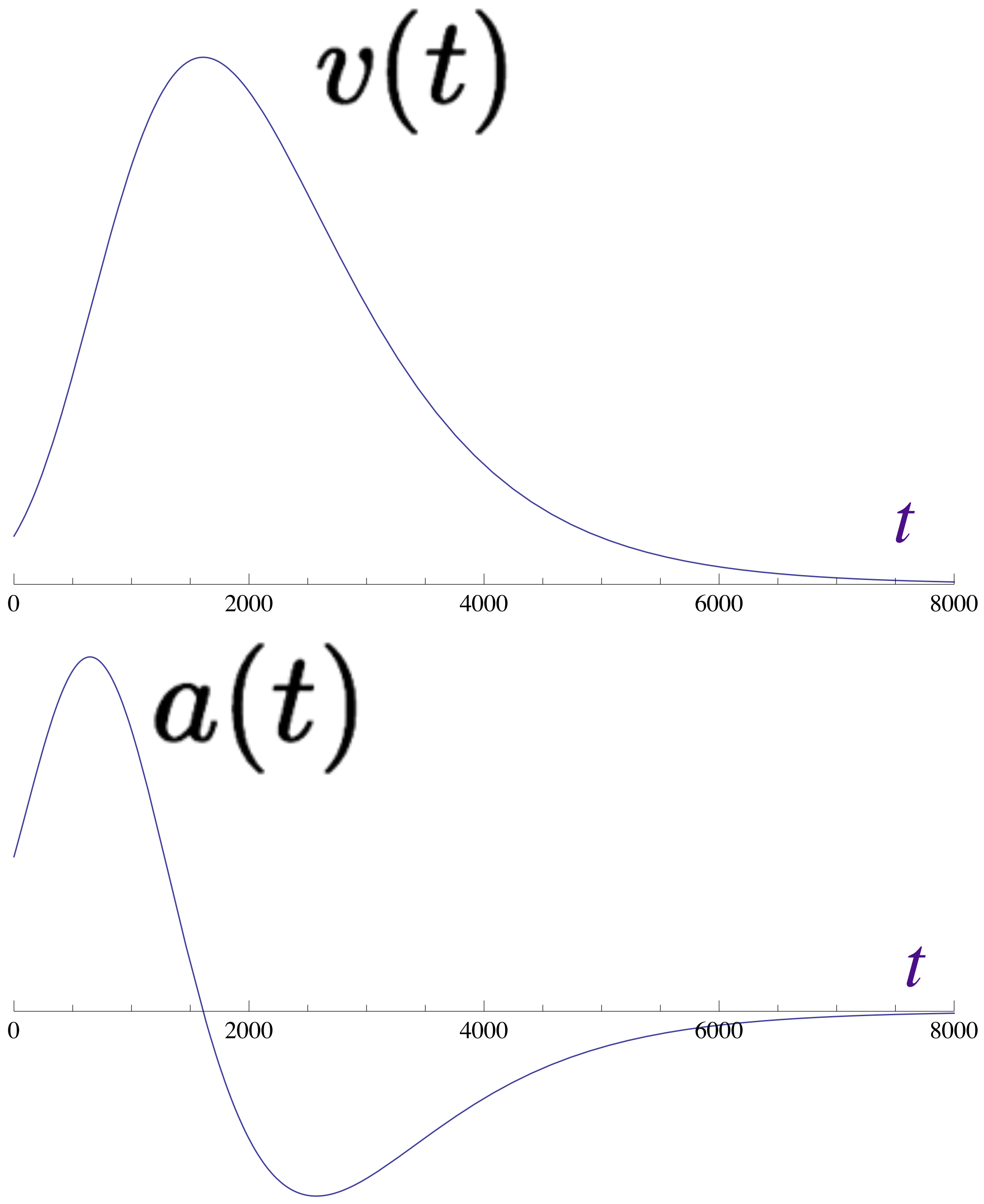
**Usually we can observe the peak of acceleration earlier than the peak of velocity.**

# Acceleration as an Indicator




**Usually we can observe the peak of acceleration earlier than the peak of velocity.**

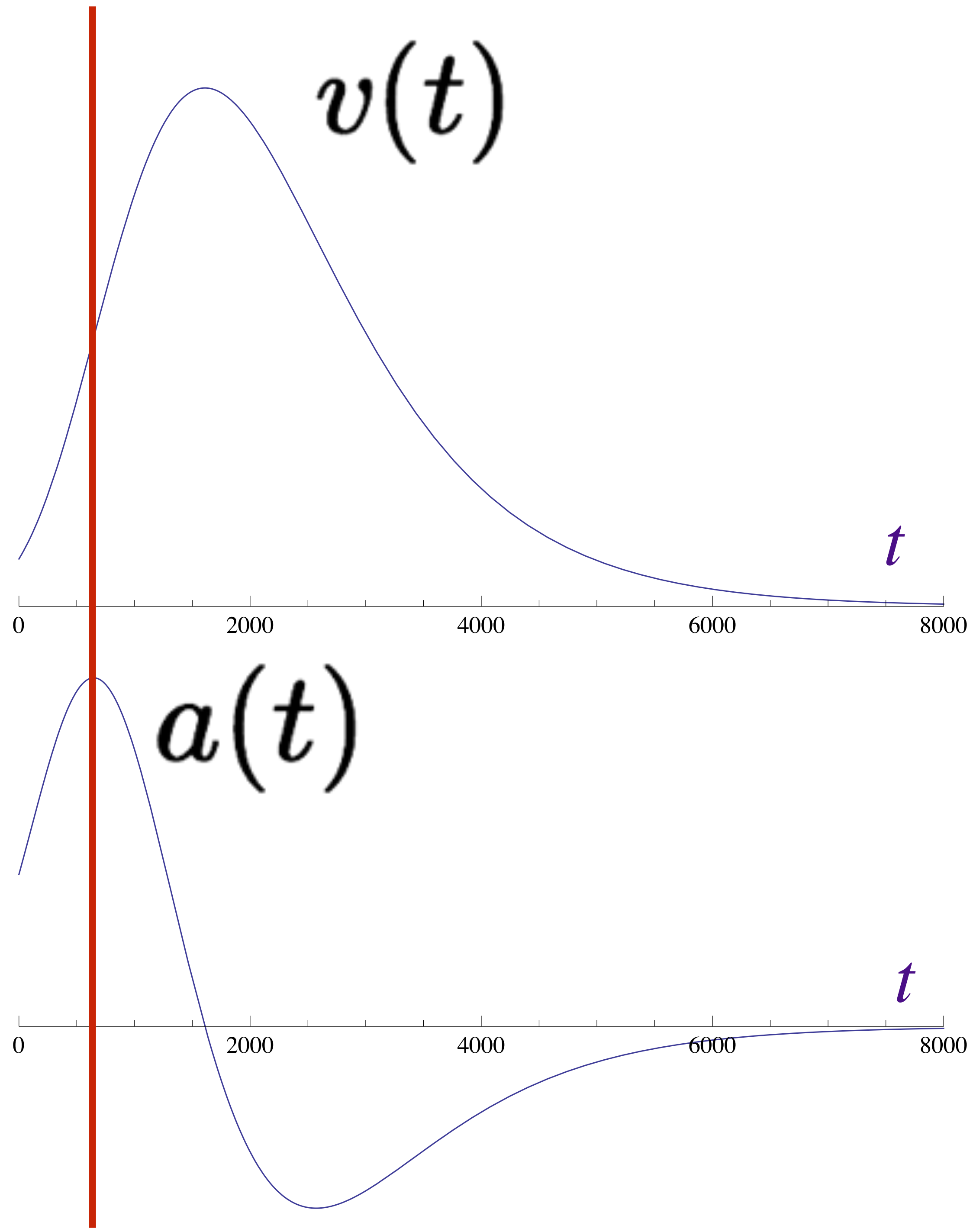
# Acceleration as an Indicator




**Usually we can observe the peak of acceleration earlier than the peak of velocity.**

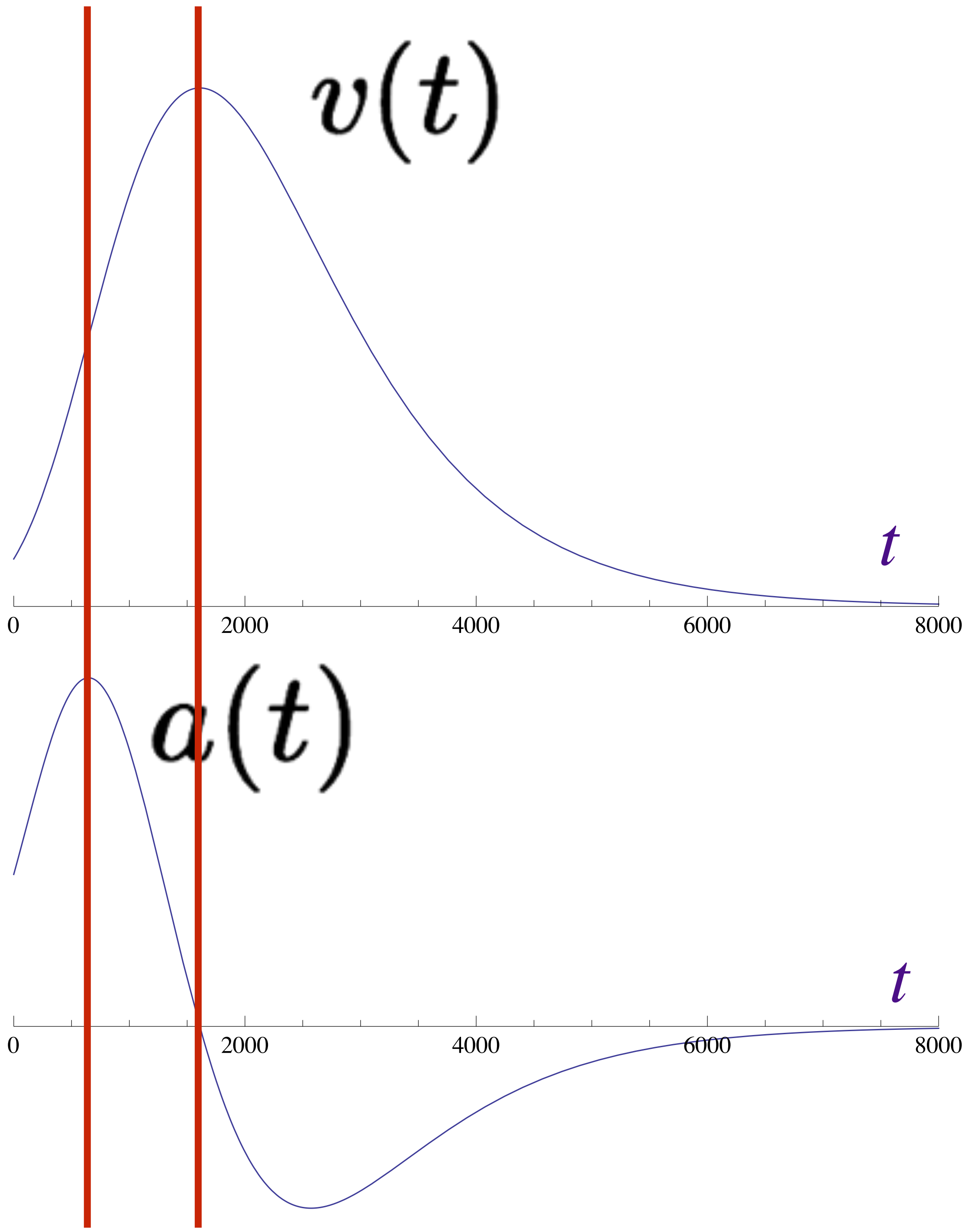
LARC
LIVING ANALYTICS RESEARCH CENTRE

# Acceleration as an Indicator



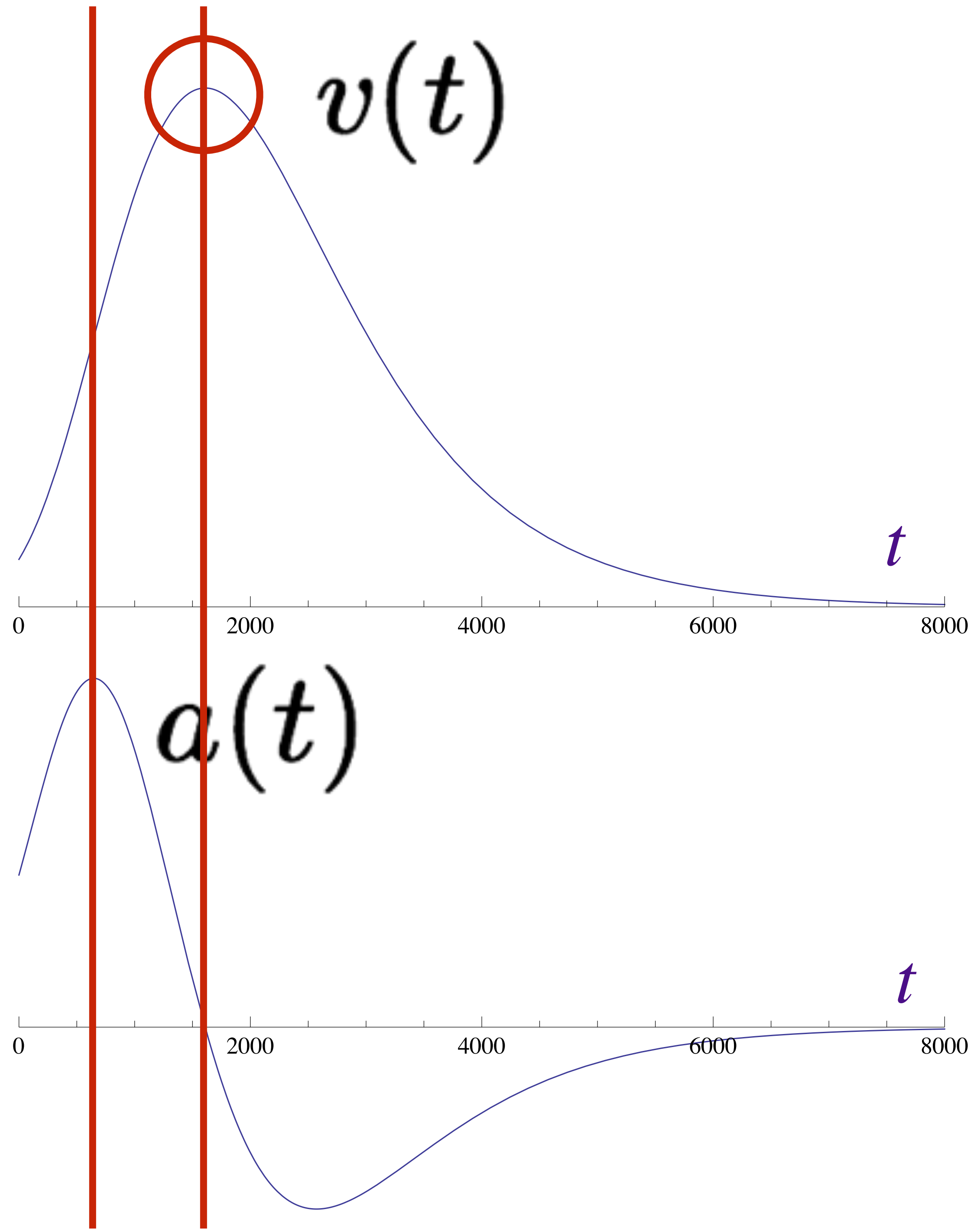

**Usually we can observe the peak of acceleration earlier than the peak of velocity.**

# Acceleration as an Indicator



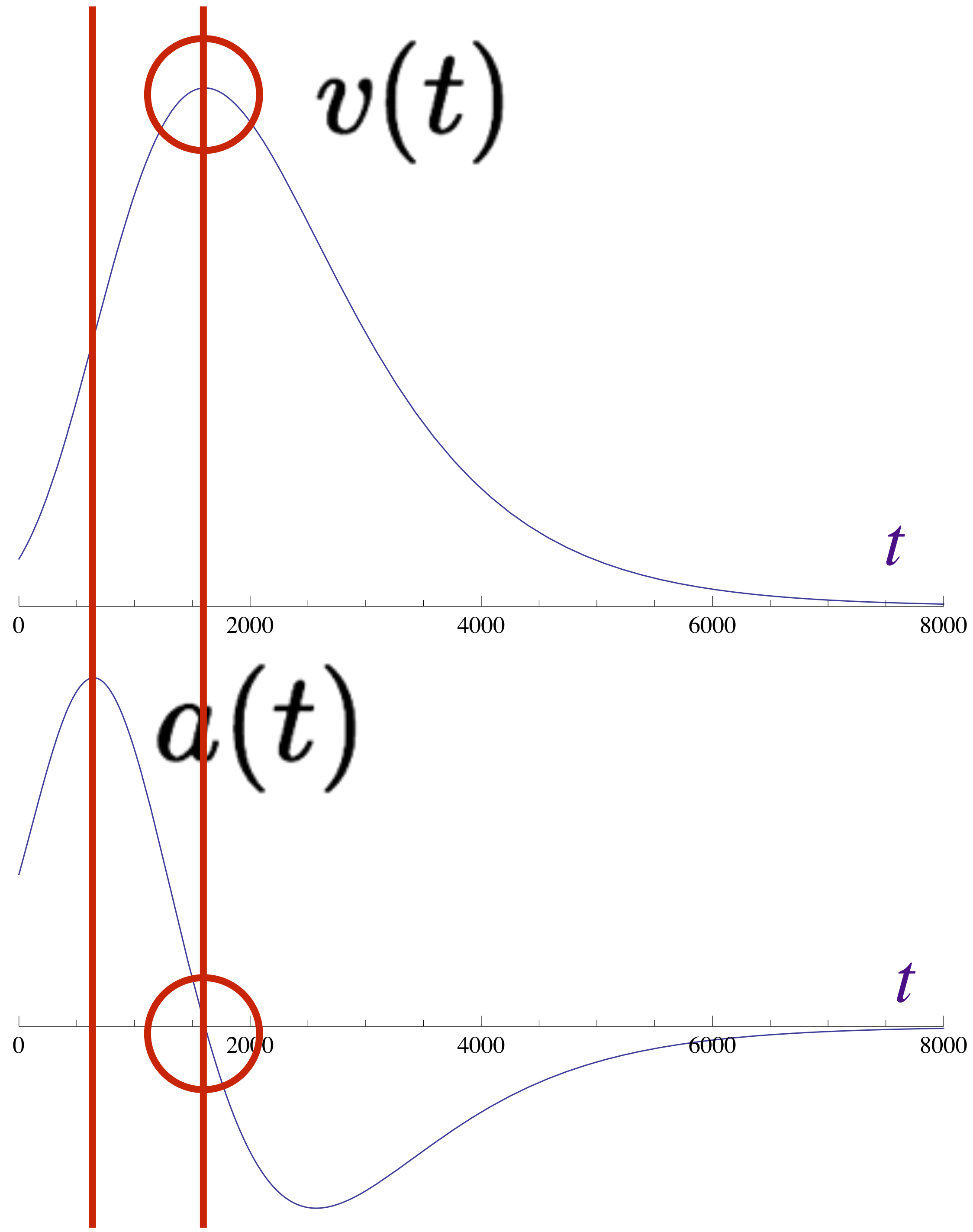

**Usually we can observe the peak of acceleration earlier than the peak of velocity.**

# Acceleration as an Indicator




**Usually we can observe the peak of acceleration earlier than the peak of velocity.**

# Acceleration as an Indicator




**Usually we can observe the peak of acceleration earlier than the peak of velocity.**

# Acceleration as an Indicator

**Acceleration: a very good early indicator of burst.**

# Acceleration as an Indicator

**Acceleration: a very good early indicator of burst.**

1. **Is there any burst at all?**

# Acceleration as an Indicator

**Acceleration: a very good early indicator of burst.**

**1. Is there any burst at all?**

**The acceleration of the whole tweet stream.**

# Acceleration as an Indicator

**Acceleration: a very good early indicator of burst.**

**1. Is there any burst at all?**

**The acceleration of the whole tweet stream.**

**2. Is there any word bursting?**



LARC
LIVING ANALYTICS
RESEARCH CENTRE

# Acceleration as an Indicator

**Acceleration: a very good early indicator of burst.**

1. **Is there any burst at all?**

   **The acceleration of the whole tweet stream.**

2. **Is there any word bursting?**

   **The acceleration of each word in the tweet stream.**

# Acceleration as an Indicator

**Acceleration: a very good early indicator of burst.**

1. **Is there any burst at all?**

   **The acceleration of the whole tweet stream.**

2. **Is there any word bursting?**

   **The acceleration of each word in the tweet stream.**

3. **Is there any topic bursting?**

# Acceleration as an Indicator

**Acceleration: a very good early indicator of burst.**

1. **Is there any burst at all?**

   **The acceleration of the whole tweet stream.**

2. **Is there any word bursting?**

   **The acceleration of each word in the tweet stream.**

3. **Is there any topic bursting?**

   **The acceleration of each pair of words in the tweet stream.**

LARC
LIVING ANALYTICS RESEARCH CENTRE

# Assumptions

# Assumptions

- Each topic is represented as a distribution over words $p_k$.

# Assumptions

- Each topic is represented as a distribution over words $p_k$.
- Tweet stream is modelled as a mixture of multiple latent topic streams. The stream of topic k has velocity $v_k(t)$ and acceleration $a_k(t)$.

# Assumptions

- Each topic is represented as a distribution over words $p_k$.
- Tweet stream is modelled as a mixture of multiple latent topic streams. The stream of topic k has velocity $v_k(t)$ and acceleration $a_k(t)$.
- Each tweet is related to only one topic.

# Assumptions

- Each topic is represented as a distribution over words $p_k$.
- Tweet stream is modelled as a mixture of multiple latent topic streams. The stream of topic k has velocity $v_k(t)$ and acceleration $a_k(t)$.
- Each tweet is related to only one topic.

**The final goal is to discover these unknown $p_k$ and $a_k(t)$ from a snapshot of the tweet stream.**



LARC
Living Analytics Research Centre

# Sketch as Snapshot

(1). $\mathbb{S}''(t)$: The acceleration of the total number of tweets in $D(t)$, $\mathbb{S}(t) = |D(t)|$.

(2). $\mathbb{X}''(t)$: The acceleration of each word, $\mathbb{X}(t)$ is a $N$-dimension vector such that $\mathbb{X}_i(t) = \sum_{d \in D(t)} \frac{d(i)}{|d|}$, $(1 \leq i \leq N)$.

(3). $\mathbb{Y}''(t)$: The acceleration of each pair of words, $\mathbb{Y}(t)$ is a $N \times N$ matrix such that

$$\mathbb{Y}_{i,j}(t) = \begin{cases} \sum_{d \in D(t)} \frac{d(i)^2 - d(i)}{|d|(|d|-1)} & , \quad i = j \\ \sum_{d \in D(t)} \frac{d(i)d(j)}{|d|(|d|-1)} & , \quad i \neq j \end{cases}$$

$(1 \leq i \leq N, 1 \leq j \leq N)$.



LARC Living Analytics Research Centre

# Properties

$$\mathbb{S}''(t) = \sum_{k=1}^{K} a_k(t) \tag{1}$$

$$E[\mathbb{X}''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \tag{2}$$

$$E[\mathbb{Y}''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \cdot p_k^T \tag{3}$$

# Properties

$$\mathbb{S}''(t) = \sum_{k=1}^{K} a_k(t) \quad (1)$$

$$E[\mathbb{X}''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \quad (2)$$

$$E[\mathbb{Y}''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \cdot p_k^T \quad (3)$$



LARC
LIVING ANALYTICS RESEARCH CENTRE

# Properties

$$\mathbb{S}''(t) = \sum_{k=1}^{K} a_k(t) \qquad (1)$$

$$E[\mathbb{X}''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \qquad (2)$$

$$E[\mathbb{Y}''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \cdot p_k^T \qquad (3)$$

**The topics with small accelerations will be filtered out.**

# Properties

$$\mathbb{S}''(t) = \sum_{k=1}^{K} a_k(t) \quad (1)$$

$$E[\mathbb{X}''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \quad (2)$$

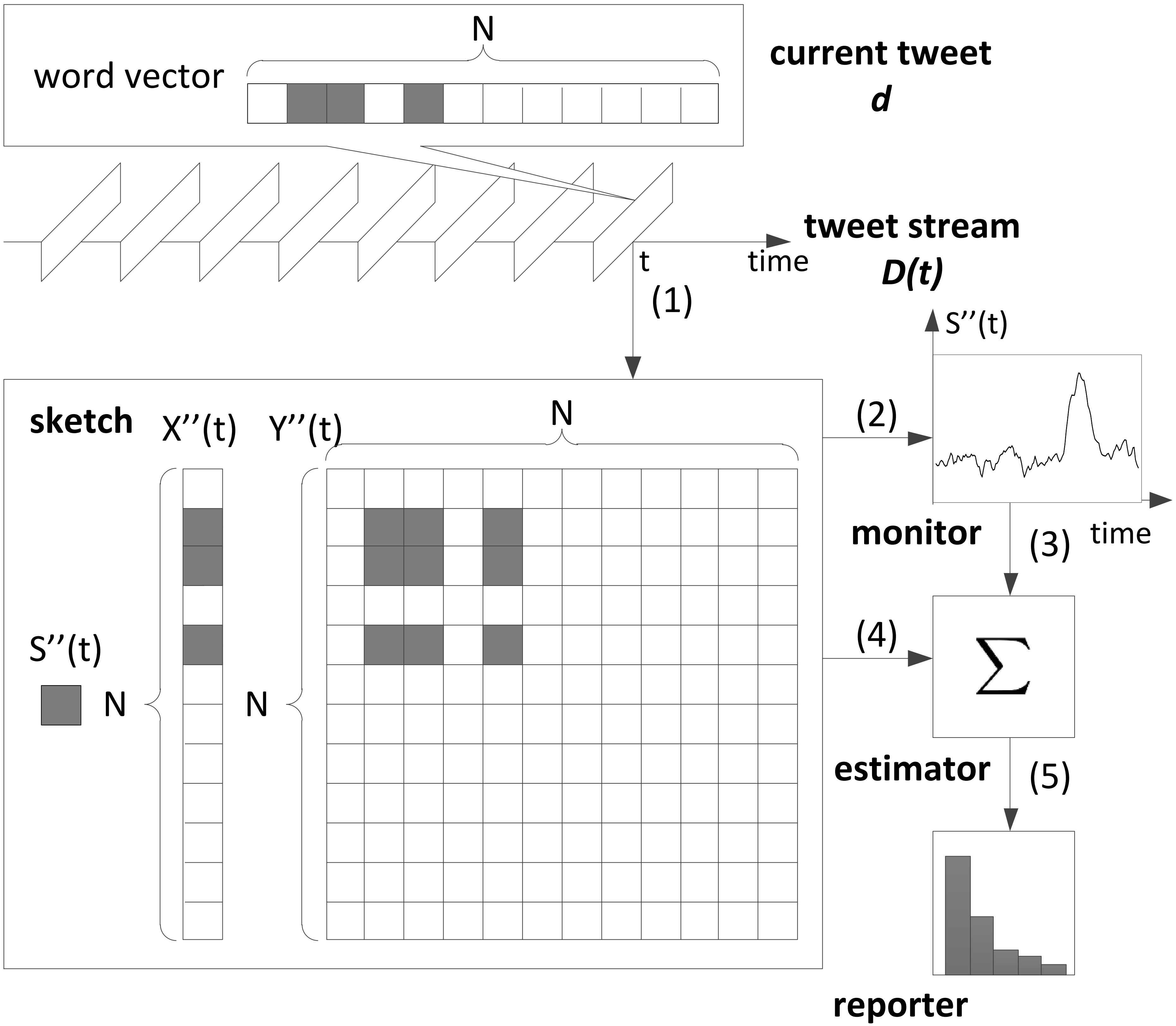
$$E[\mathbb{Y}''(t)] = \sum_{k=1}^{K} a_k(t) \cdot p_k \cdot p_k^T \quad (3)$$

**The topics with small accelerations will be filtered out.**

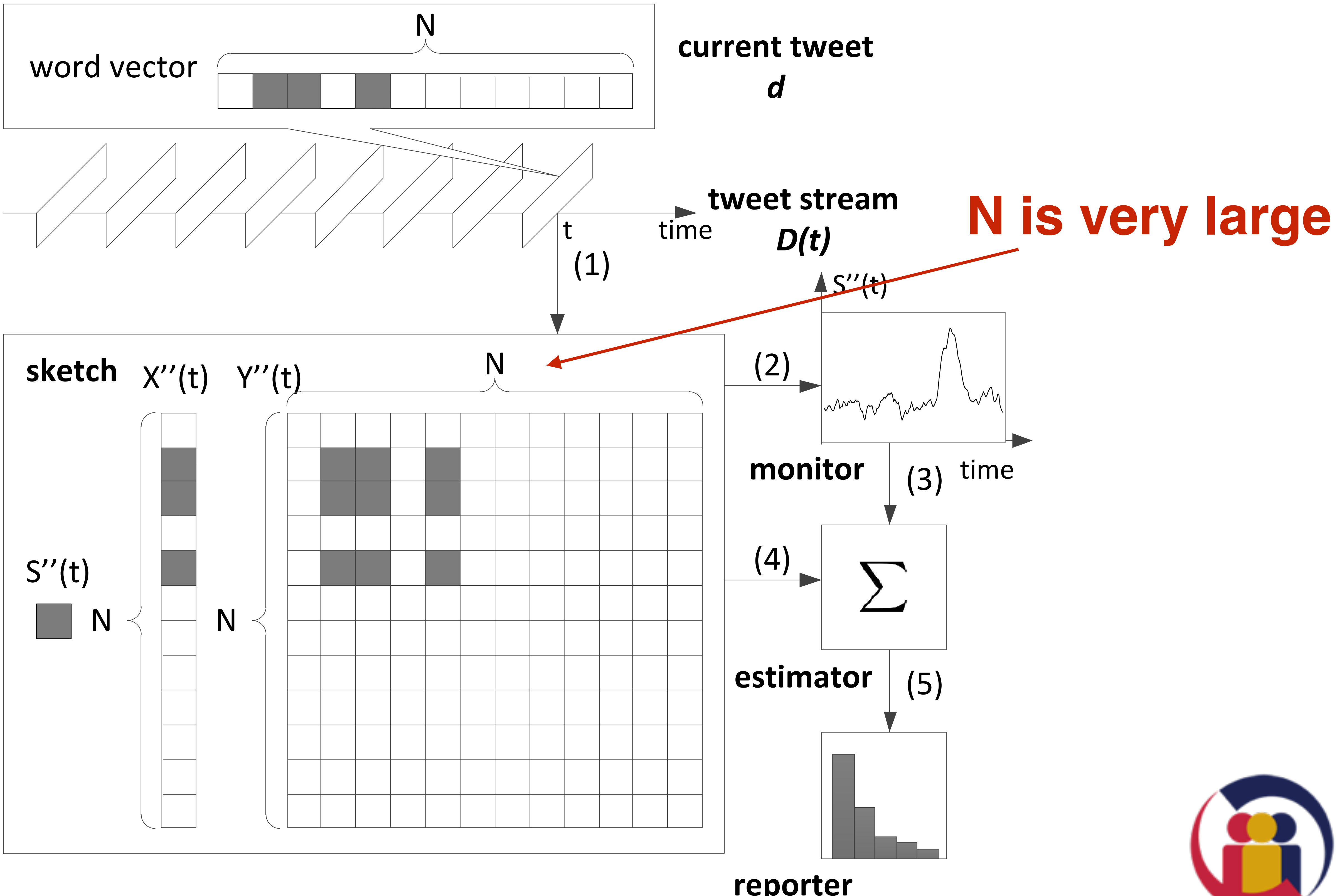
**Minimise the difference between observation and expectation.**

LARC
LIVING ANALYTICS RESEARCH CENTRE

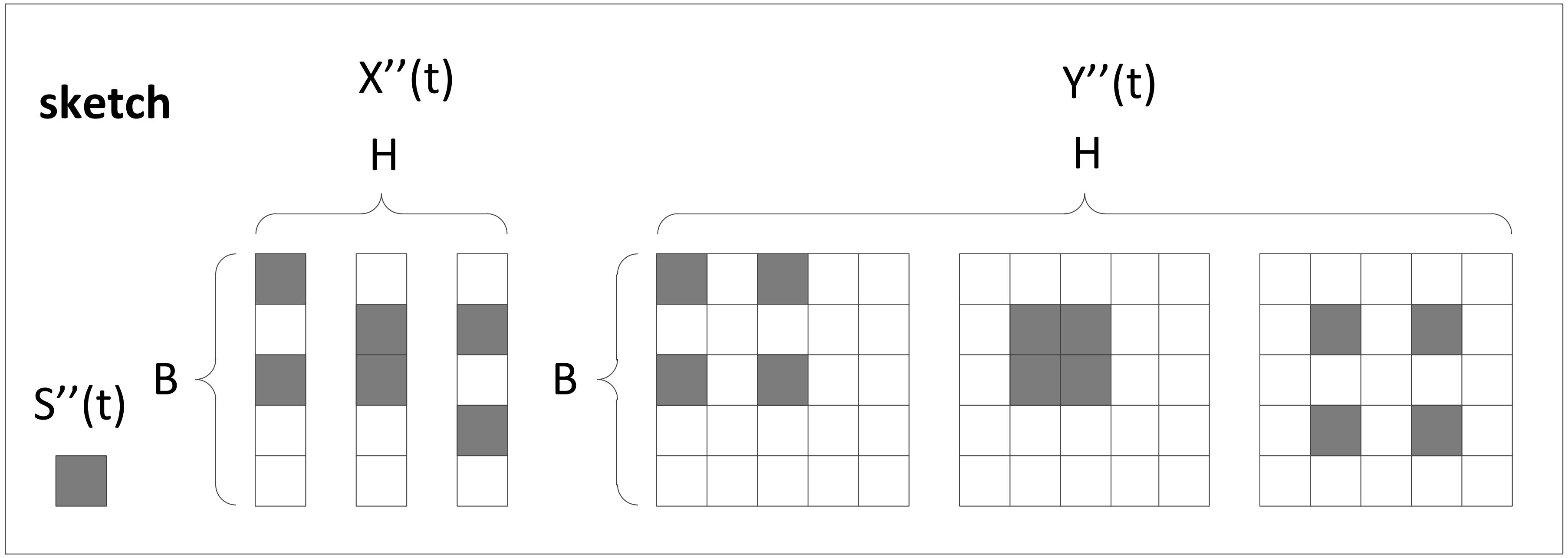
# Real-time Framework



# Real-time Framework



# Dimension Reduction




**From $O(N^2)$ to $O(H*B^2)$, $B \ll N$, $H \ll N$**

G. Cormode and S. Muthukrishnan. **An improved data stream summary: the count-min sketch and its applications.** Journal of Algorithms, 55(1):58–75, 2005.

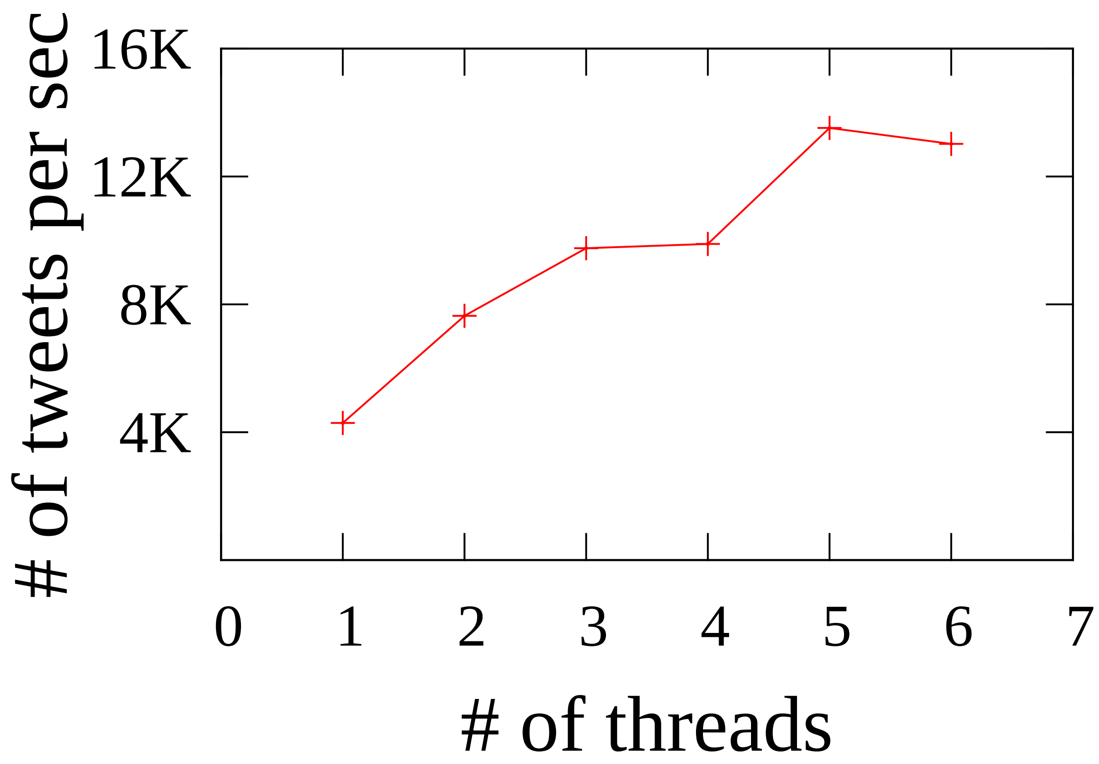
# Efficiency Evaluation

**Dataset** : Singapore based Twitter data, which contains over 30 millions tweets. We use these tweets to simulate a live tweet stream.

# Efficiency Evaluation

**Dataset** : Singapore based Twitter data, which contains over 30 millions tweets. We use these tweets to simulate a live tweet stream.

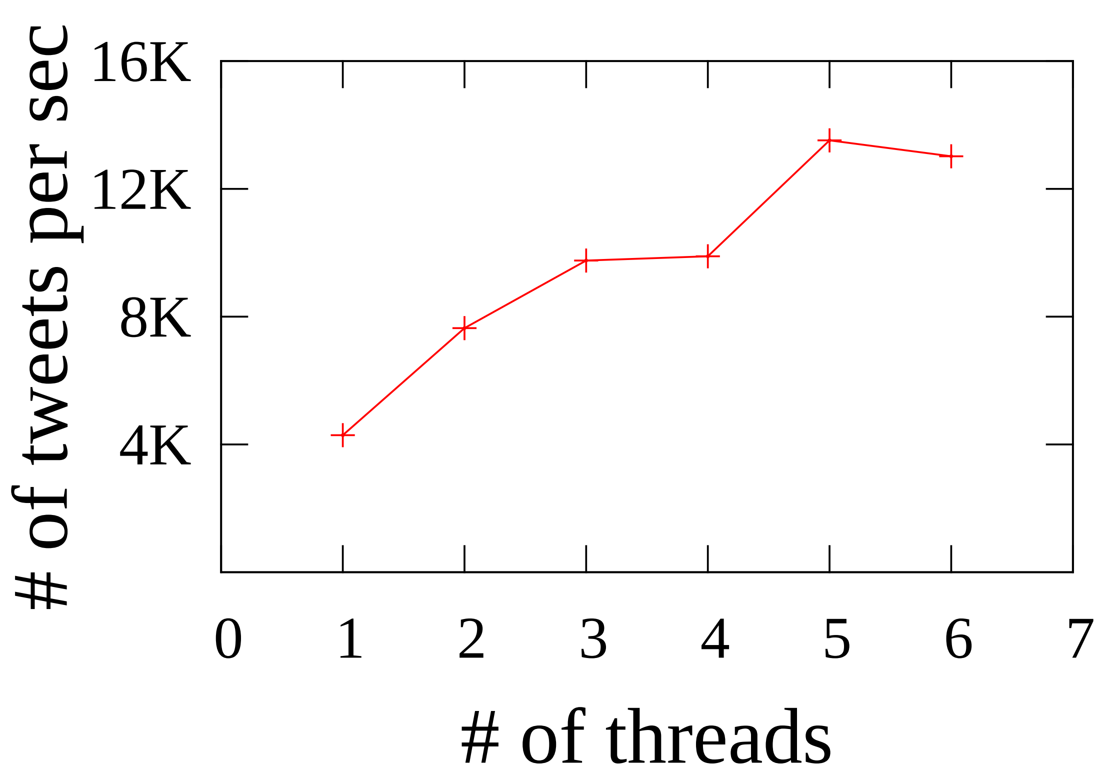
**Throughput**



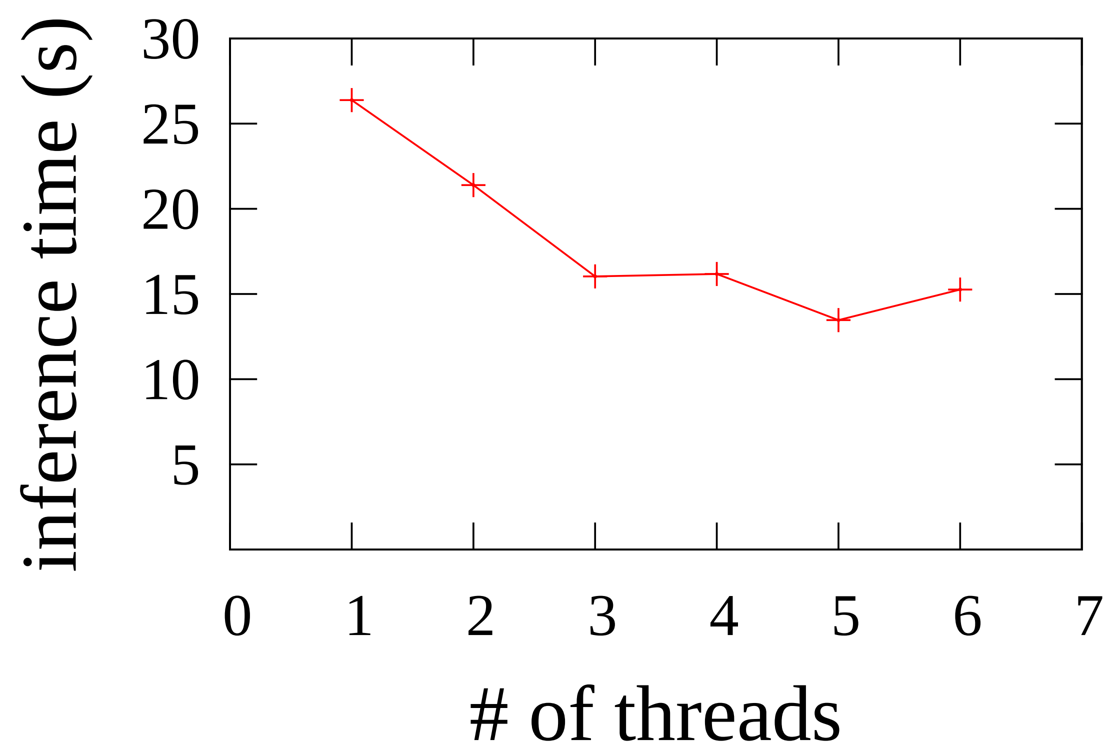
# Efficiency Evaluation

**Dataset** : Singapore based Twitter data, which contains over 30 millions tweets. We use these tweets to simulate a live tweet stream.

**Throughput**




**Inference time**



# Effectiveness Evaluation

- Compare with **Twevent**
- Use the same dataset which contain over 4 million tweets
- List all the events detected by both algorithms between June 7, 2010 to June 12, 2010, in which period several big events happened.

# Effectiveness Evaluation

| Event | Sub-event | TopicSketch | Twevent |
|---|---|---|---|
| Steve Jobs released iPhone 4 during WWDC2010 | Farmville client for iPhone 4 was demonstrated. | #wwdc, iphone, farmville | steve jobs, iMovie, wwdc, iphone, wifi |
| | Retina display of iPhone 4 was introduced. | iphone, 4, #wwdc, display, retina | |
| | iMovie for iPhone 4 was demonstrated. | iphone, 4, imovie, #wwdc | |
| | New iPhone 4 was available in Singapore in July. | iphone, 4, singapore, july | |



LARC
LIVING ANALYTICS RESEARCH CENTRE

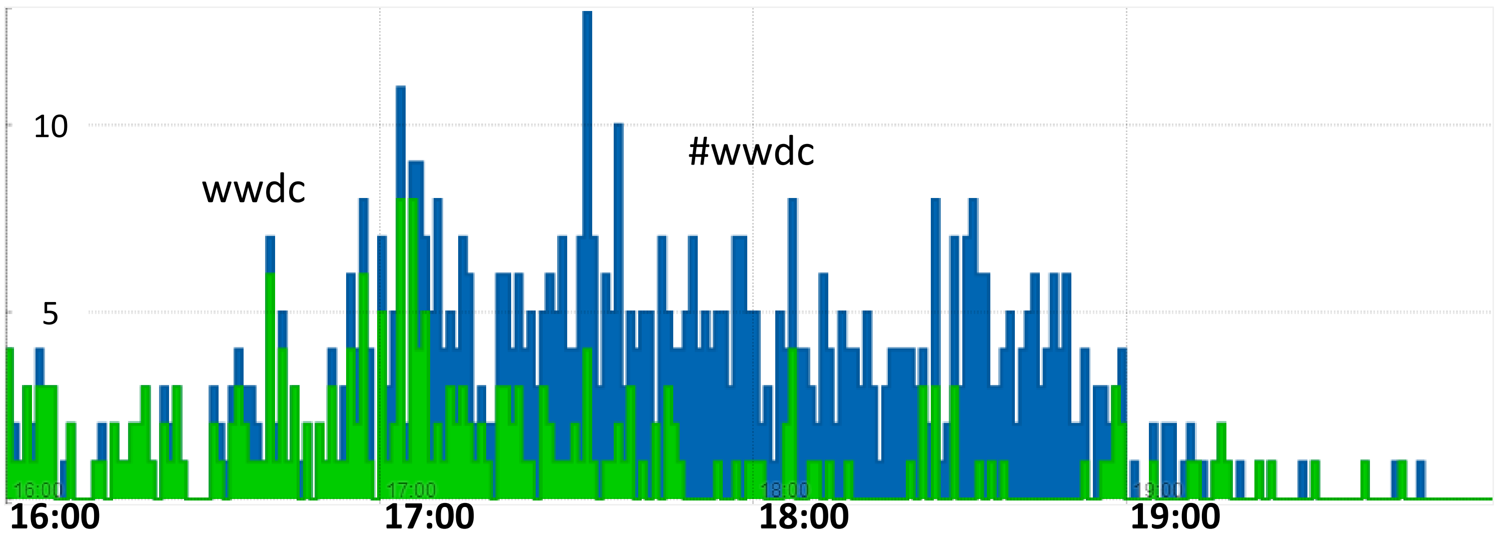
# Effectiveness Evaluation

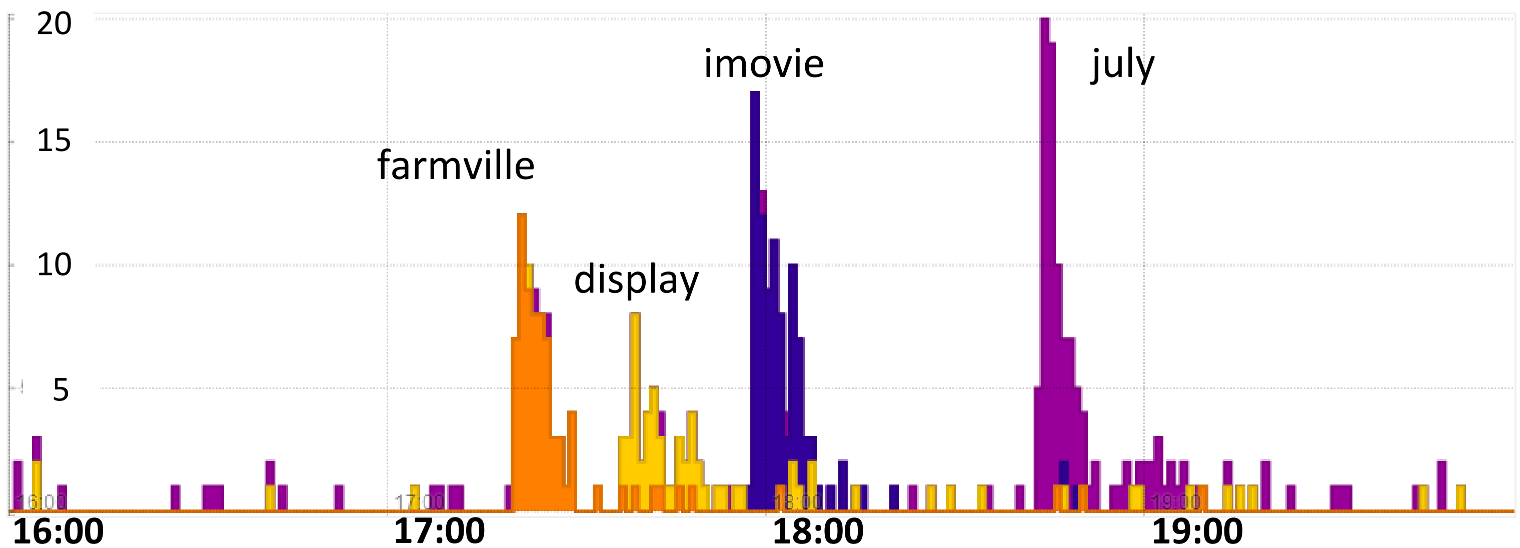
| Event | Sub-event | TopicSketch | Twevent |
|---|---|---|---|
| Steve Jobs released iPhone 4 during WWDC2010 | Farmville client for iPhone 4 was demonstrated. | #wwdc, iphone, farmville | steve jobs, iMovie, wwdc, iphone, wifi |
| | Retina display of iPhone 4 was introduced. | iphone, 4, #wwdc, display, retina | |
| | iMovie for iPhone 4 was demonstrated. | iphone, 4, imovie, #wwdc | |
| | New iPhone 4 was available in Singapore in July. | iphone, 4, singapore, july | |



LARC
LIVING ANALYTICS RESEARCH CENTRE

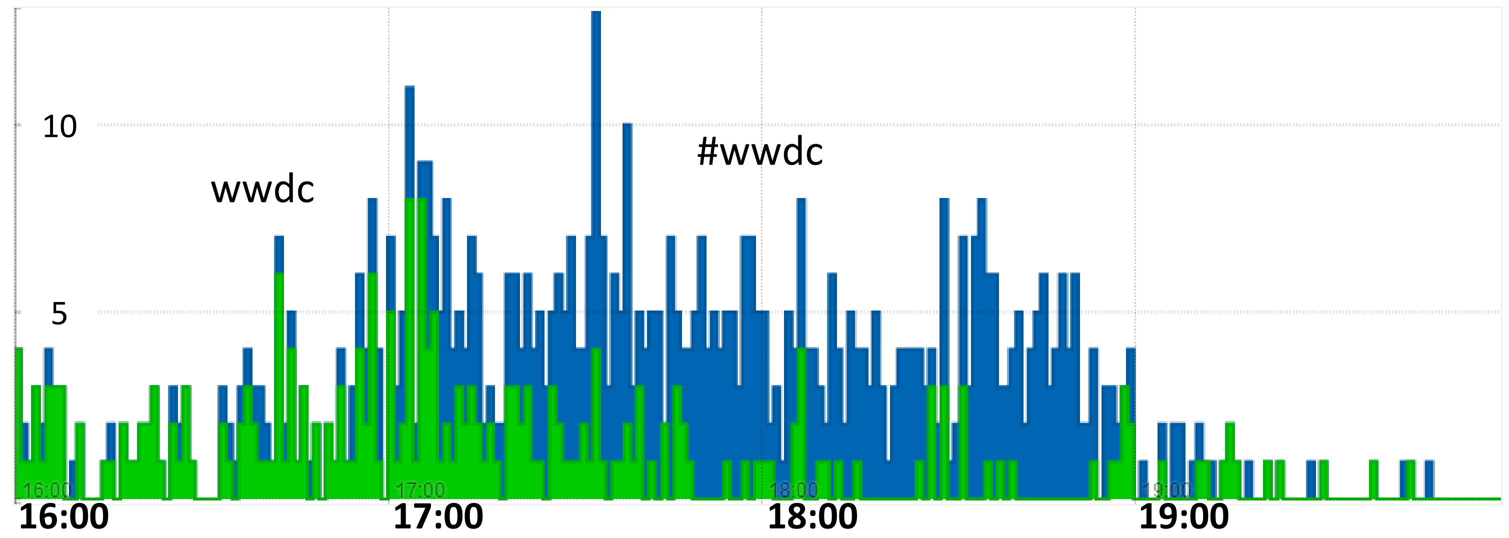
# Effectiveness Evaluation

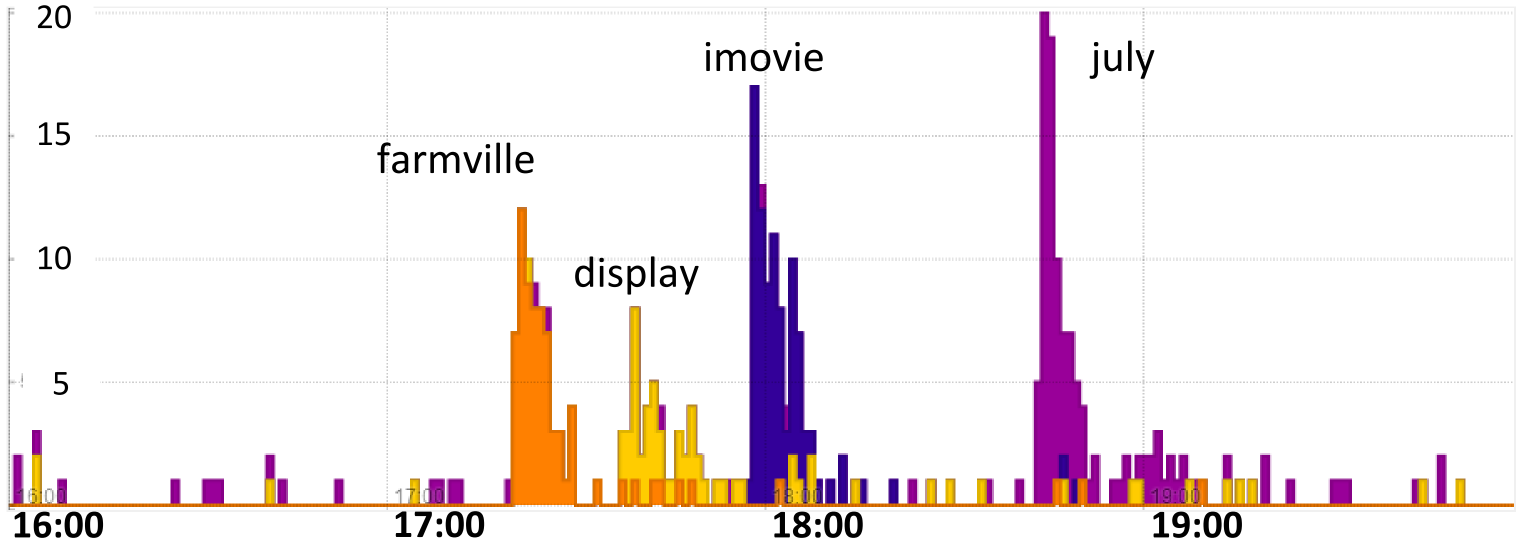



LARC
LIVING ANALYTICS RESEARCH CENTRE

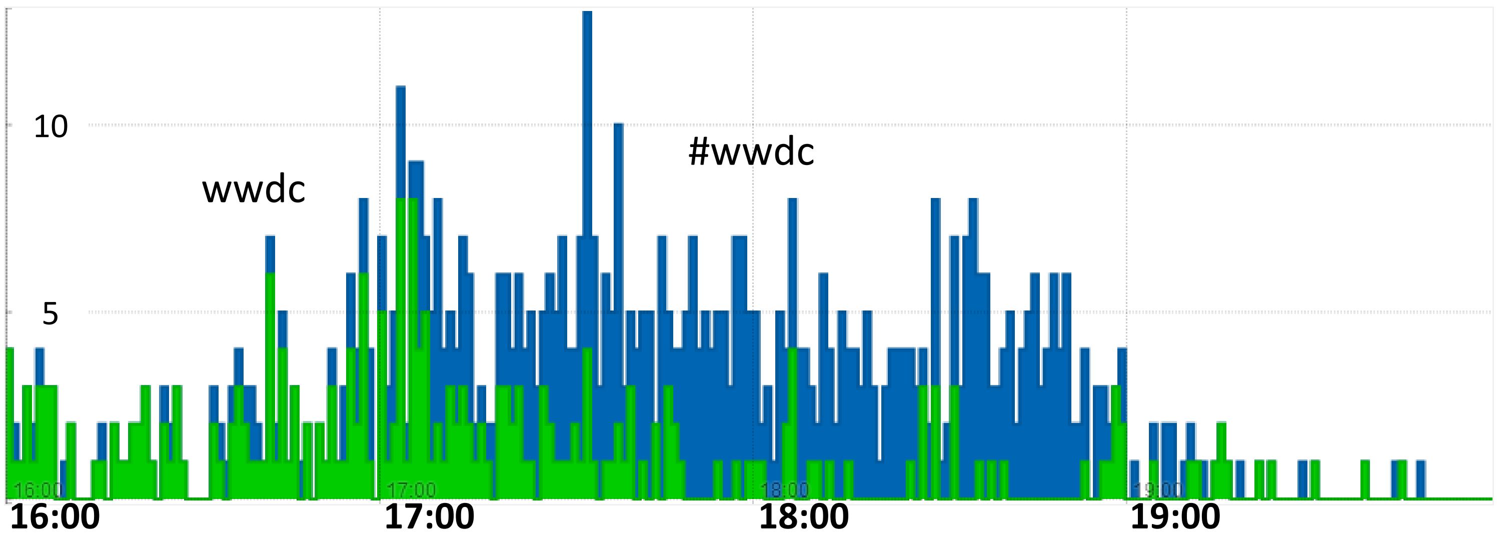
# Effectiveness Evaluation



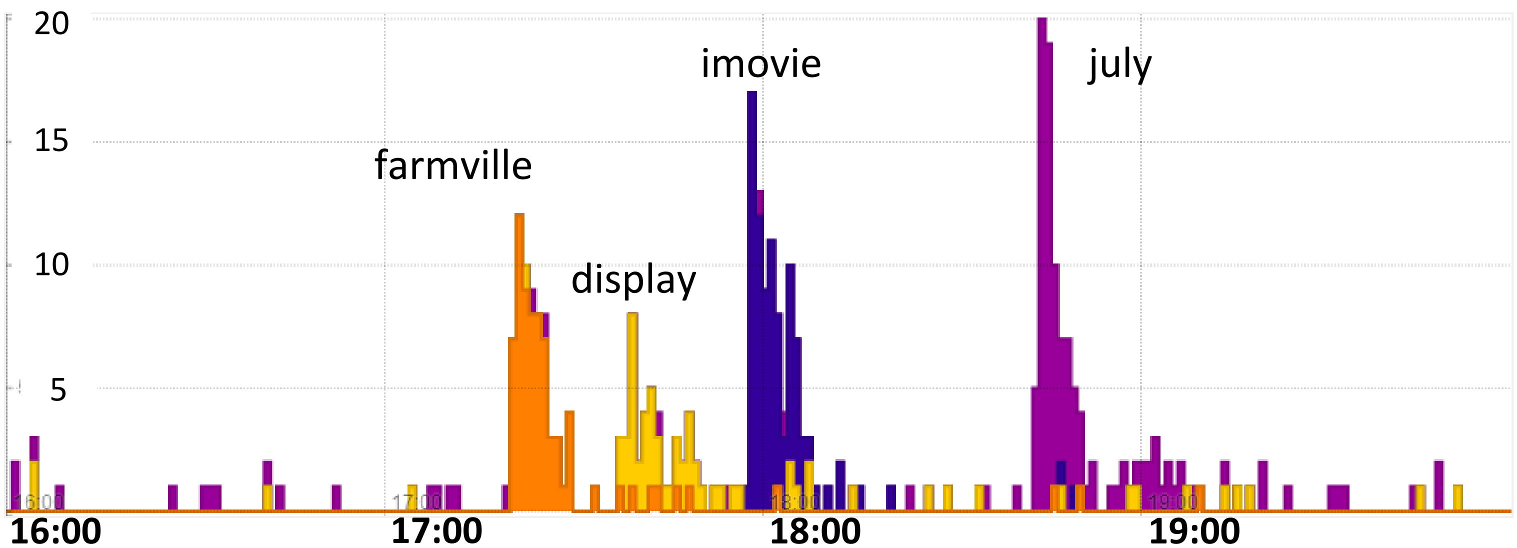




**Event**

# Effectiveness Evaluation



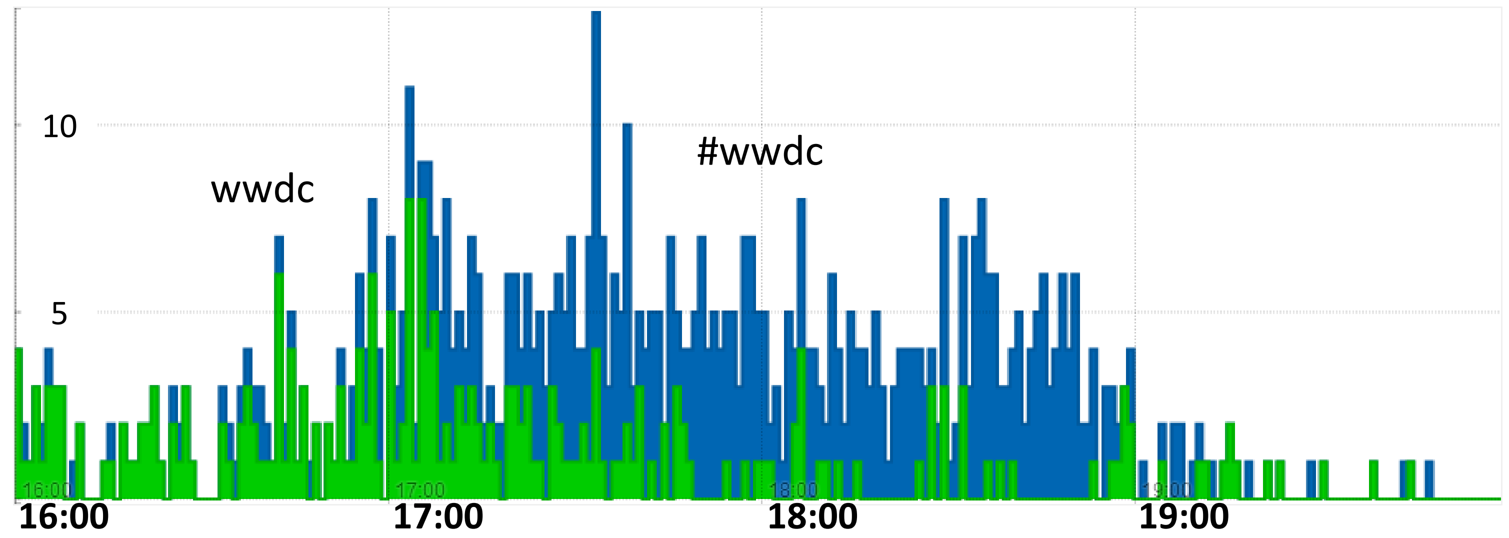

**Event**




**Sub-event**

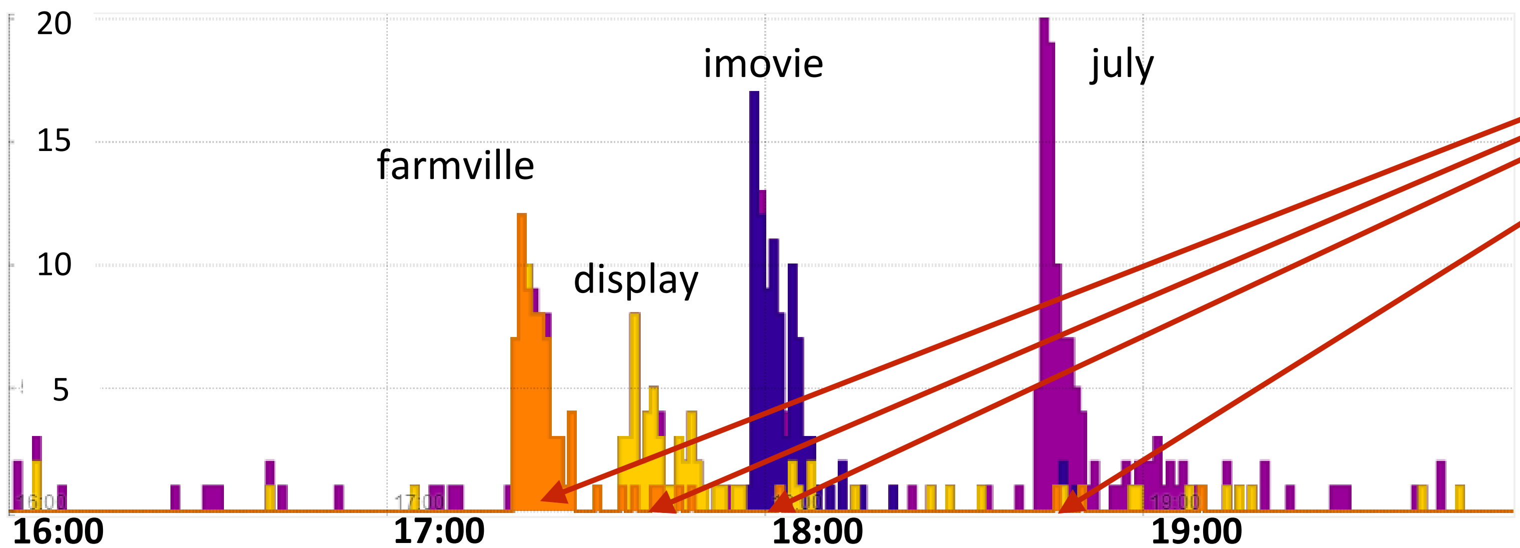
LARC
Living Analytics Research Centre

# Effectiveness Evaluation




**Event**

**Detection**




**Sub-event**

# Conclusion

- We proposed **TopicSketch** a framework for real-time detection of bursty topics from Twitter.

- We developed a concept of “sketch” which provides a “snapshot” of the current tweet stream. It can be updated efficiently. And we can find bursty topics from it efficiently.

- TopicSketch provides a temporally-ordered sub-events to describe the event, which is more informative than the traditional methods.



LARC
LIVING ANALYTICS RESEARCH CENTRE

# Thanks