BURSTY EVENT DETECTION THROUGHOUT HISTORIES

Debjyoti Paul, Yanqing Peng, Feifei Li

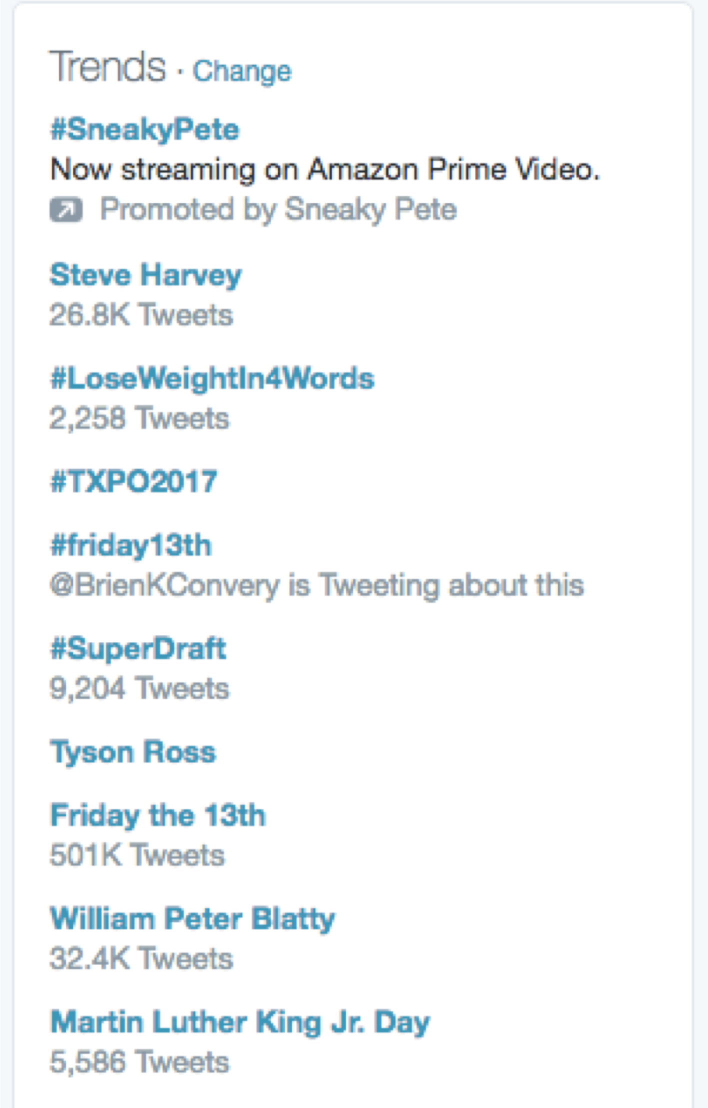
35TH IEEE INTERNATIONAL CONFERENCE ON DATA ENGINEERING
ICDE 2019 RESEARCH TRACK

OVERVIEW

- Twitter trends

- Real-time trending (bursty) event detection
  - Tells people what's happening
  - Help people react to important uprising events in their early stages while they are still developing
  - Well studied problem

- Historical Bursty Events:
  - Not a well studied problem but relevant for data scientists.



Trends · Change

#SneakyPete
Now streaming on Amazon Prime Video.
Promoted by Sneaky Pete

Steve Harvey
26.8K Tweets

#LoseWeightIn4Words
2,258 Tweets

#TXPO2017

#friday13th
@BrienKConvery is Tweeting about this

#SuperDraft
9,204 Tweets

Tyson Ross

Friday the 13th
501K Tweets

William Peter Blatty
32.4K Tweets

Martin Luther King Jr. Day
5,586 Tweets

BURSTINESS

Intuition: Examples of bursty and non-bursty events

- Earthquake: discussed frequently in a time range
- Weather: discussed frequently all the time

Insight: *Bursty* = Surge in incoming rate

Definition: The burstiness of event e at time t is

$$B_e(t) = bf_e(t) - bf_e(t - \tau)$$

where $bf_e(t)$ is the incoming rate of event e within time range $[t - \tau, t)$

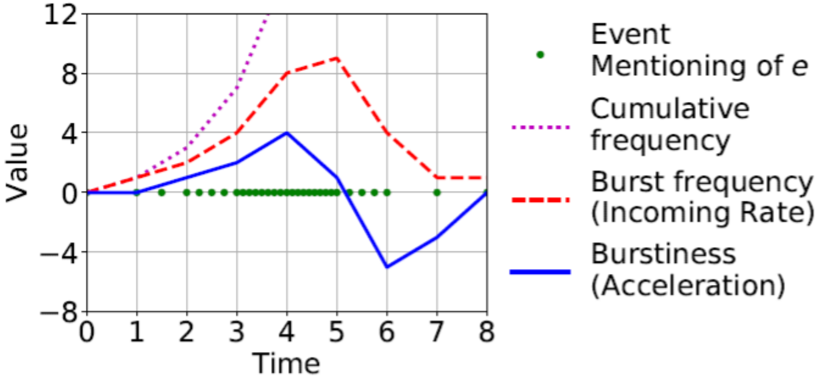
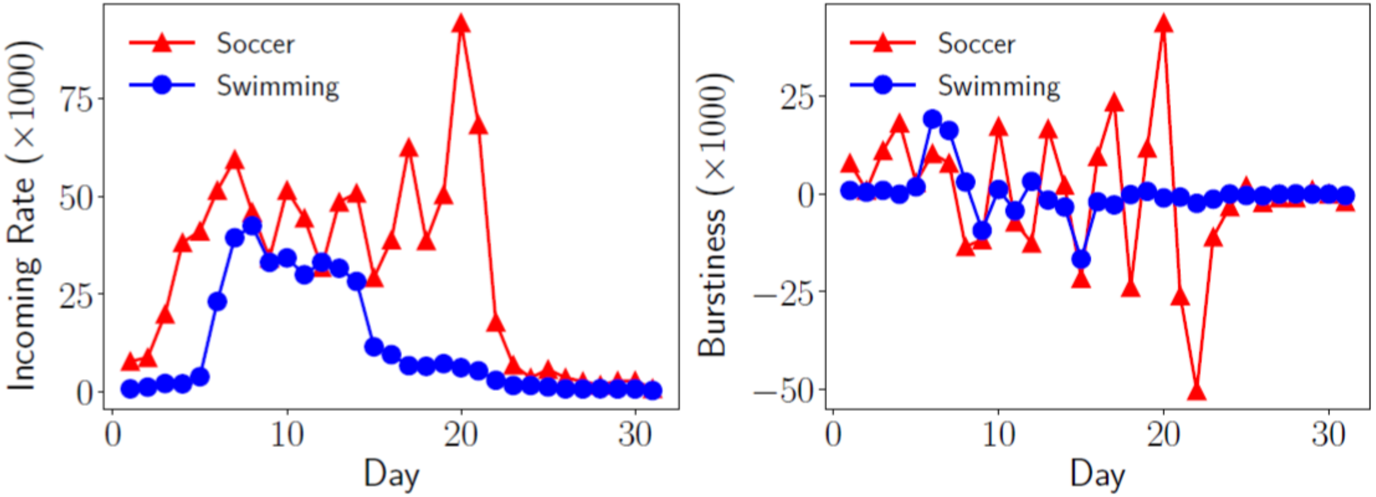


Figure 1: An example of burst where $\tau = 1$.

INCOMING RATE VS BURSTINESS



(a) Incoming rate. (b) Burstiness.

HISTORICAL BURSTY EVENTS

- Interesting problem:
  How to query and analyze bursty events from past efficiently?

- Query Examples:
  1. What are the bursty events in the first week of October in 2016?
  2. Is "Anthem Protest" a bursty event in second week of September in 2017?

- Understand and analyze bursty events by going back and forth in time.

BASELINE SOLUTION

✓ Store timeline curves of all events in the history.

$ Cost: #events * #timestamps

✗ Infeasible!!!

PROBLEM AND DESIGN GOALS

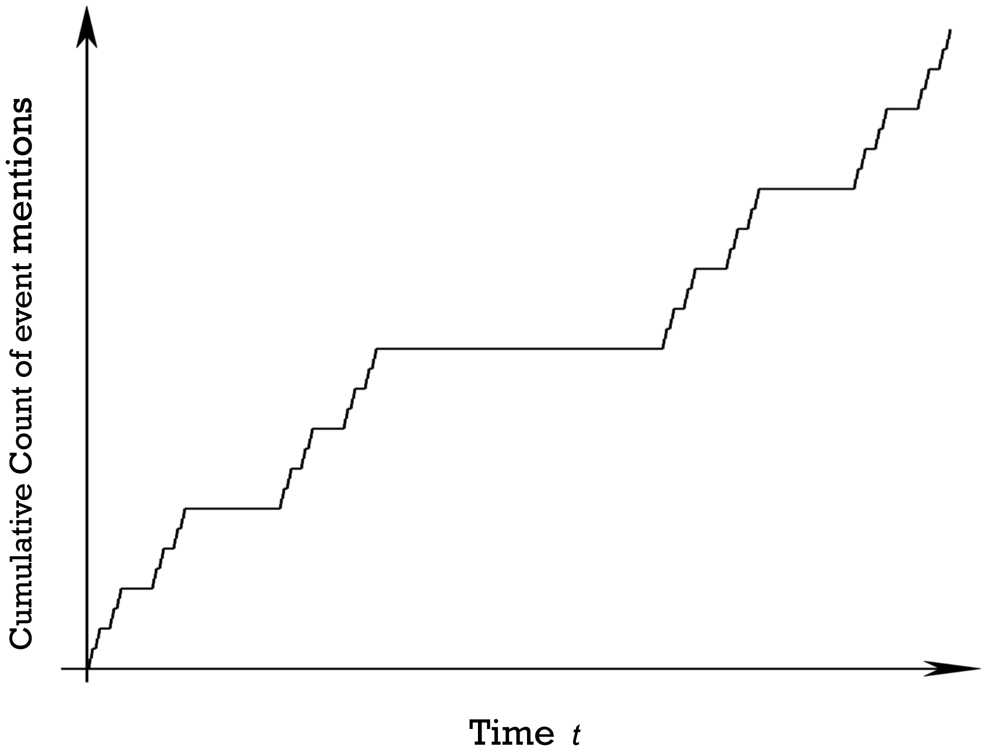
Given a temporal stream of events, design an approach to store the stream with compact space, and answer the following queries with theoretical bounded error:

1. Bursty Point Query: How bursty is this event at this time?
   - Query the burstiness value for event e at time t
2. Bursty Time Range Query: In which time does this event become bursty?
   - Query the timestamps that the burstiness value of event e is above threshold θ
3. Bursty Event Query: What events are bursty at this time?
   - Query the events that has burstiness value above threashold θ at time t

Focus on Bursty Point Queries, then extend to other queries.

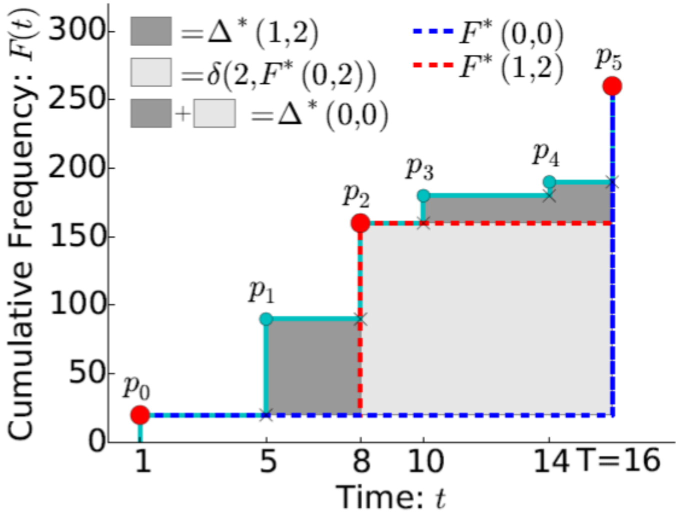
STAIRCASE CURVE

A single event stream represented as a staircase curve.

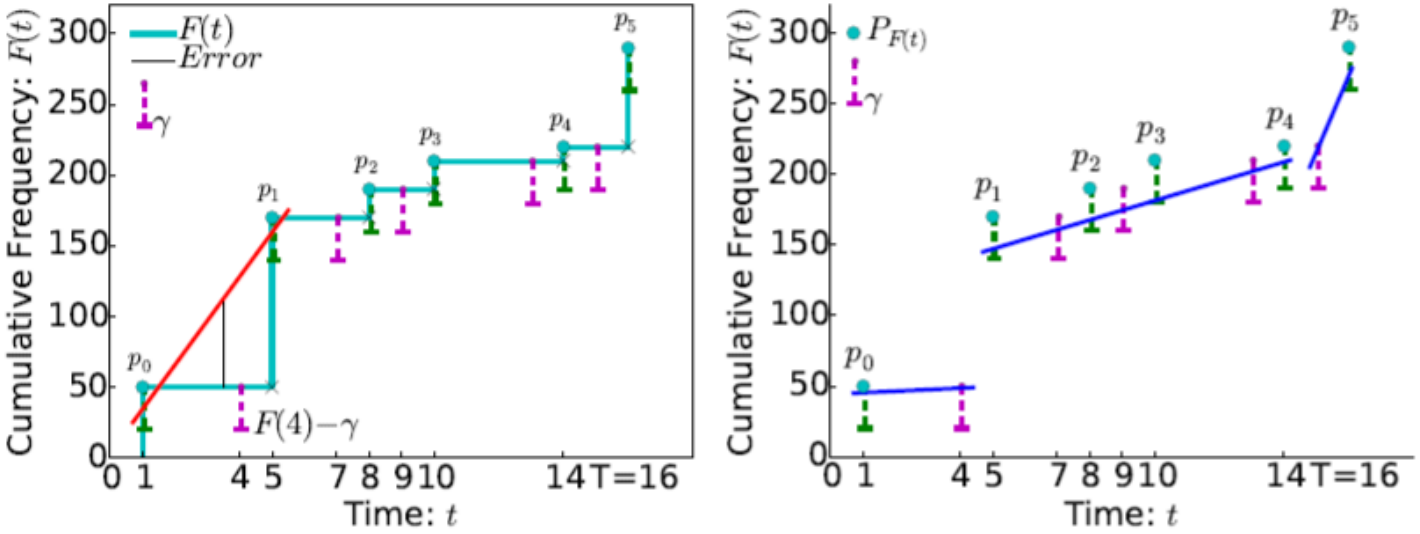


PBE-1 APPROXIMATION: BUFFERED SOLUTION

- Original data F(t): frequency staircase curve
- Compress data F*(t): a staircase curve that under the original staircase
  - "Distance" between F^(t) to F(t) is defined by the area of F-F^(t)
  - Lemma: The corners of the optimal staircase must contain only the corners of F(t)
- Select a subset of staircase corner points to form a sub-staircase
  - Dynamic Programming



PBE-2 APPROXIMATION: ONLINE SOLUTION

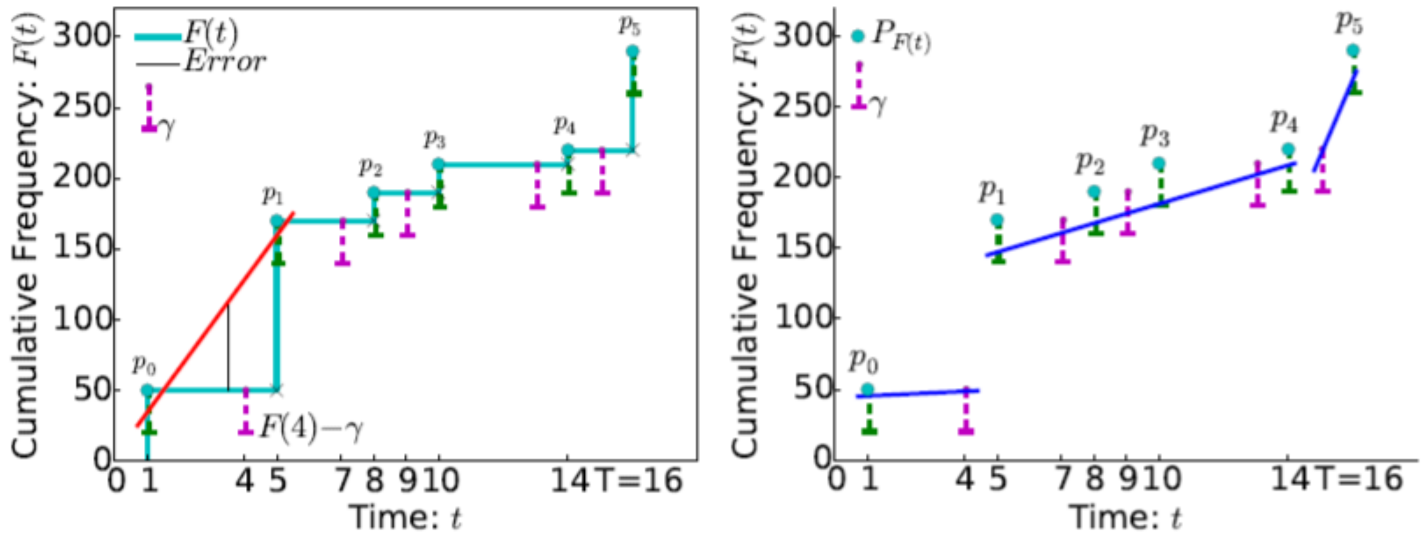


(a) Timestamped frequency ranges A. (b) A PLA L for A.

Figure 3: An example of PBE-2.

- Piecewise Linear Approximation
- Use multiple segments to represent the original staircase

PBE-2 APPROXIMATION: ONLINE SOLUTION



(a) Timestamped frequency ranges A.

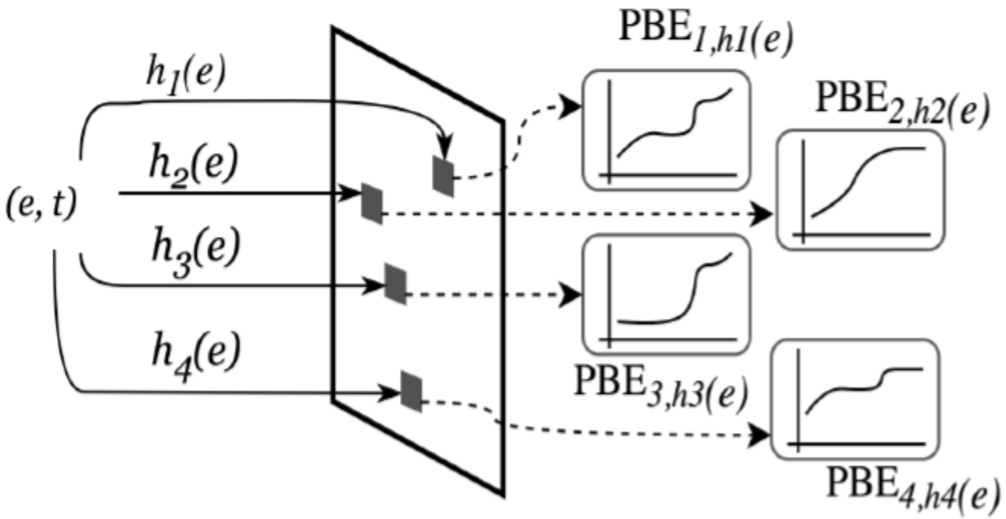
(b) A PLA L for A.

Figure 3: An example of PBE-2.

- Piecewise Linear Approximation
- Use multiple segments to represent the original staircase

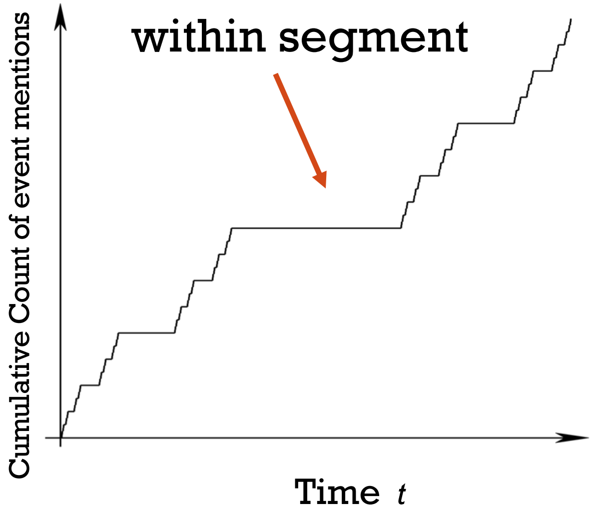
MULTIPLE EVENT STREAM

- Count-Min (CM) Sketch
  - The count–min sketch (CM sketch) is a probabilistic data structure that serves as a frequency table of events in a stream of data
- Combining CM with PBEs



OTHER TYPES OF QUERIES

- Bursty time range query
  - Check only the corner points



0 acceleration within segment

- Bursty event query
  - Log N number of CM-PBE where N is number of events.

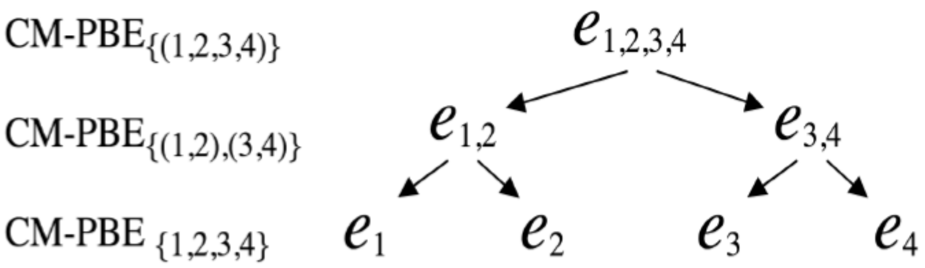
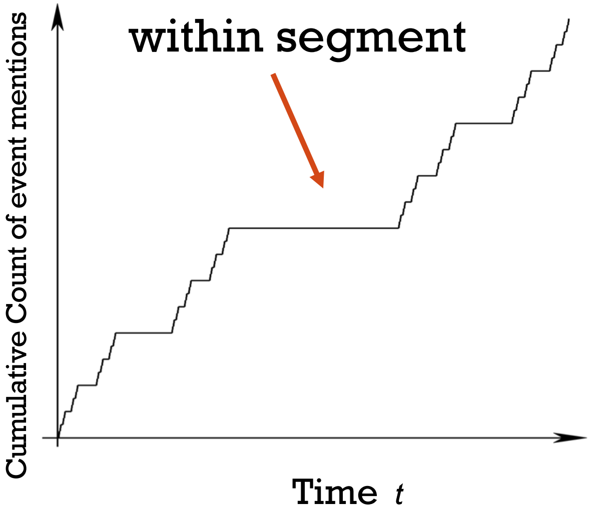


Figure 6: Binary decomposition of the event id space.

OTHER TYPES OF QUERIES

- Bursty time range query
  - Check only the corner points



0 acceleration within segment

- Bursty event query
  - Log N number of CM-PBE where N is number of events.

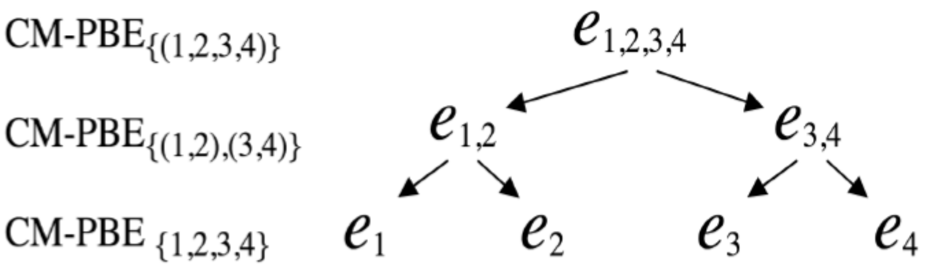
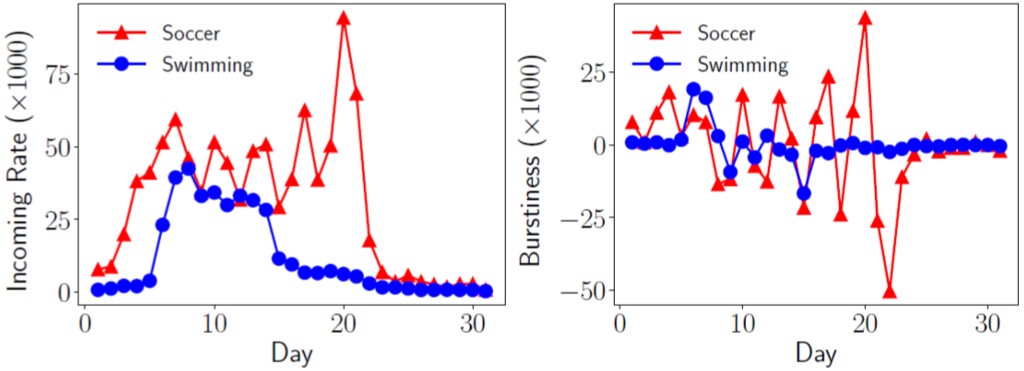


Figure 6: Binary decomposition of the event id space.

EXPERIMENT DATASETS

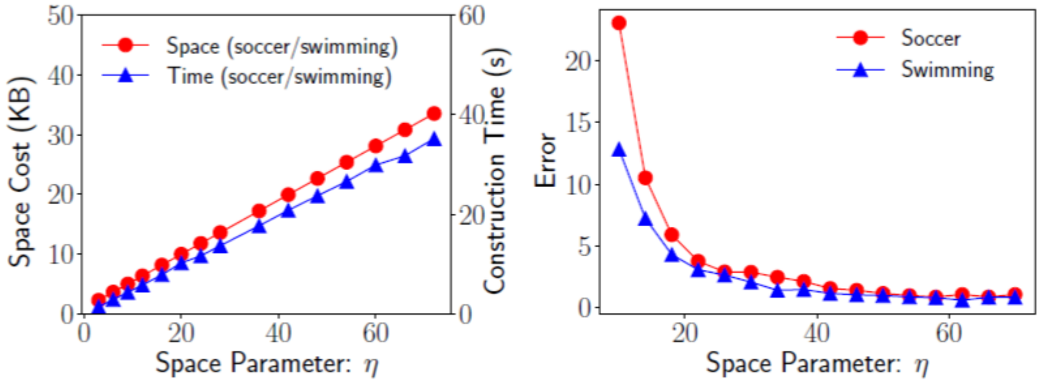


(a) Incoming rate. (b) Burstiness.

Figure 7: Two events in olympicrio. $\tau = 86,400$ seconds (1 day).

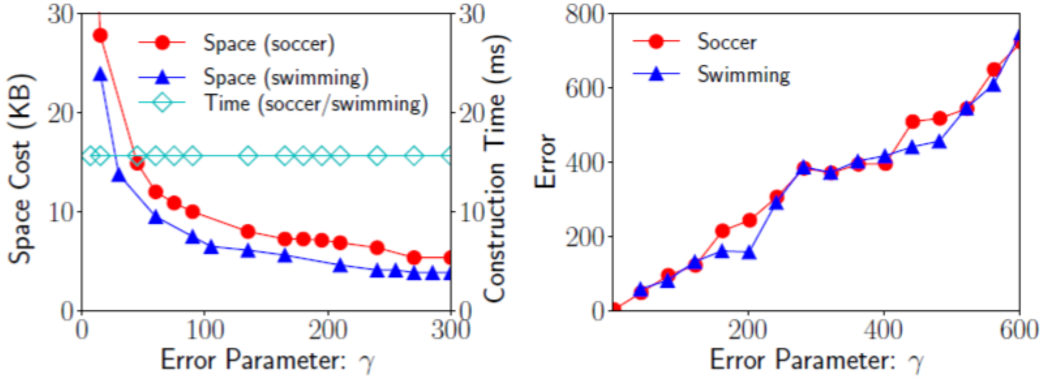
- OlympicRio: 50M tweets in August 2016 about Olympic Games Rio with 864 events.
  - Swimming and Soccer

- USPolitics: 286M tweets from June 2016 to November 2016 on US politics with 1689 events. Randomly sampled to make it as large as OlympicRio.



(a) Space and construction costs. (b) Query accuracy.

Figure 8: PBE-1 parameter study.



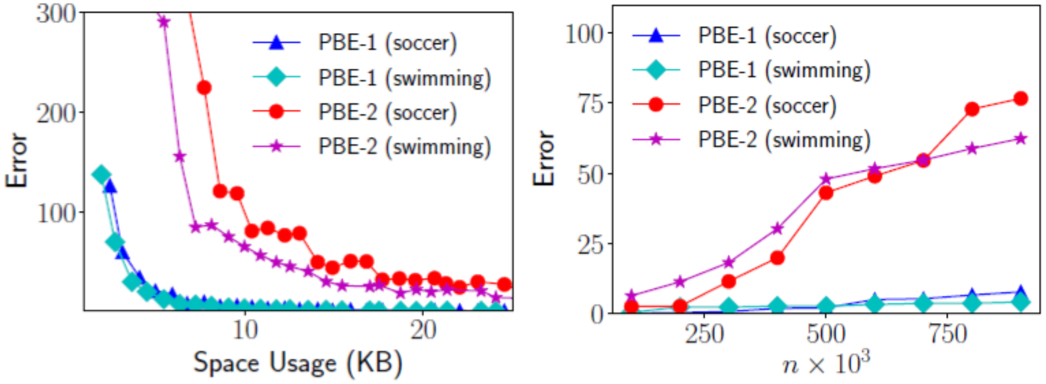
(a) Space and construction costs. (b) Query accuracy.

Figure 9: PBE-2 parameter study.

PARAMETER STUDY

- PBE-1 (offline):
  - Tradeoff: Error vs Space + Time
  - Long construction time (~1min)
  - Small space cost
  - Low error
- PBE-2 (online):
  - Tradeoff: Error vs Space
  - Short construction time (~10ms)
  - Small space cost
  - Relatively high error when compared with PBE-1

SINGLE EVENT STREAM

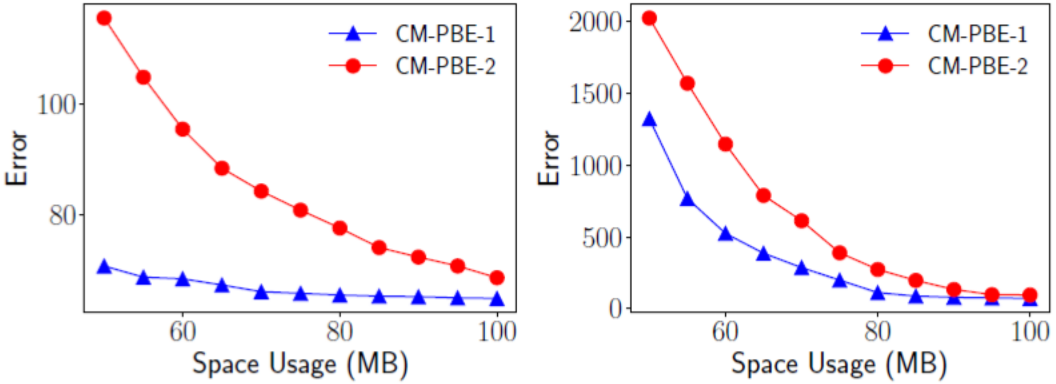


(a) space vs accuracy. (b) n vs accuracy, $|\mathsf{PBE}| = 10$KB.

Figure 10: PBE: single event stream.

- 300x Space save compared with baseline
- Low error for both approaches, PBE-1 (offline) performs better.

MULTIPLE EVENTS STREAM



(a) olympicrio dataset. (b) uspolitics dataset.

Figure 11: CM-PBE: Space vs accuracy.

- 100x Space save compared with baseline
  - 12 GB raw data to 80 MB meta data.
- Low error for both approaches, PBE-1 (offline) performs better.

CONCLUSION

- We have unleashed the potential of Bursty Event Detection for past events.
- Existing work focus on Real-time bursty detection, doesn't discuss on efficient storage for retrieval.
- We propose a framework to answer historical bursty event queries with small space.
  - Single event stream
    - Offline Dynamic Programming: Optimal but requires buffering
    - Online Piecewise Linear Approximation: Fast and no-buffering, but with higher error.
  - Multiple events stream: A variant of Count-Min Sketch
- Supported queries
  - Point query
  - Bursty time range query
  - Bursty event query

REFERENCES

REFERENCES

[1] X. Zhou and L. Chen, "Event detection over twitter social media streams," The VLDB journal, vol. 23, no. 3, pp. 381–400, 2014.
[2] C. C. Aggarwal and K. Subbian, "Event detection in social streams," in SDM, 2012.
[3] C. Li, A. Sun, and A. Datta, "Twevent: segment-based event detection from tweets," in CIKM, 2012, pp. 155–164.
[4] W. Feng, C. Zhang, W. Zhang, J. Han, J. Wang, C. Aggarwal, and J. Huang, "Streamcube: hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream," in ICDE, 2015.
[5] C. Xing, Y. Wang, J. Liu, Y. Huang, and W.-Y. Ma, "Hashtag-based sub-event discovery using mutually generative lda in twitter." in AAAI, 2016.
[6] W. Xie, F. Zhu, J. Jiang, E.-P. Lim, and K. Wang, "Topicsketch: Realtime bursty topic detection from twitter," in ICDE, 2013, pp. 837–846.
[7] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Peaks and persistence: Modeling the shape of microblog conversations," in CSCW, 2011.
[8] C. Zhang, L. Liu, D. Lei, Q. Yuan, H. Zhuang, T. Hanratty, and J. Han, "Triovecevent: Embedding-based online local event detection in geotagged tweet streams," in SIGKDD, 2017, pp. 595–604.
[9] C. Zhang, G. Zhou, Q. Yuan, H. Zhuang, Y. Zheng, L. Kaplan, S. Wang, and J. Han, "Geoburst: Real-time local event detection in geo-tagged tweet streams," in SIGIR, 2016.
[10] D. Paul, F. Li, M. K. Teja, X. Yu, and R. Frost, "Compass: Spatio temporal sentiment analysis of US election what twitter says!" in KDD. ACM, 2017, pp. 1585–1594.
[11] G. Cormode, M. N. Garofalakis, P. J. Haas, and C. Jermaine, "Synopses for massive data: Samples, histograms, wavelets, sketches," Foundations and Trends in Databases, vol. 4, no. 1-3, pp. 1–294, 2012.
[12] G. Cormode and S. Muthukrishnan, "An improved data stream summary: the count-min sketch and its applications," Journal of Algorithms, vol. 55, no. 1, pp. 58–75, 2005.
[13] N. Alon, Y. Matias, and M. Szegedy, "The space complexity of approximating the frequency moments," Journal of Computer and system sciences, vol. 58, no. 1, pp. 137–147, 1999.
[14] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," Communications of the ACM, vol. 13, no. 7, pp. 422–426, 1970.
[15] L. AlSumait, D. Barbara, and C. Domeniconi, "On-line lda: Adaptive ´ topic models for mining text streams with applications to topic detection and tracking," in ICDE, 2008, pp. 3–12.
[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.
[17] Z. Wei, G. Luo, K. Yi, X. Du, and J.-R. Wen, "Persistent data sketching," in SIGMOD, 2015, pp. 795–810.

REFERENCES

[18] J. Kleinberg, "Bursty and hierarchical structure in streams," DMKD, vol. 7, no. 4, 2003.
[19] Y. Zhu and D. Shasha, "Efficient elastic burst detection in data streams," in KDD, 2003. [20] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu, "Parameter free bursty events detection in text streams," in VLDB, 2005, pp. 181–192. [21] Q. He, K. Chang, E.-P. Lim, and J. Zhang, "Bursty feature representation for clustering text streams," in SDM, 2007, pp. 491–496.
[22] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: a survey," SIGMOD, 2013.
[23] R. Lu and Q. Yang, "Trend analysis of news topics on twitter," IJMLC, vol. 2, no. 3, 2012.
[24] E. Schubert, M. Weiler, and H. Kriegel, "Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds," in KDD, 2014.
[25] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on twitter based on temporal and social terms evaluation," in MDM, 2010.
[26] M. A. Cameron, R. Power, B. Robinson, and J. Yin, "Emergency situation awareness from twitter for crisis management," in WWW. ACM, 2012, pp. 695–698. [27] Y. Peng, J. Guo, F. Li, W. Qian, and A. Zhou, "Persistent bloom filter: Membership testing for the entire history," in SIGMOD, 2018.
[28] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han, "Scalable topical phrase mining from text corpora," PVLDB, vol. 8, no. 3, pp. 305–316, 2014.
[29] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, "Topical word embeddings." in AAAI, 2015.
[30] X. Fu, T. Wang, J. Li, C. Yu, and W. Liu, "Improving distributed word representation and topic model by word-topic mixture model," in ACML, 2016.
[31] Q. Li, S. Shah, X. Liu, A. Nourbakhsh, and R. Fang, "Tweetsift: Tweet topic classification based on entity knowledge base an

QUESTIONS

?

AUXILLARY SLIDES

BURSTINESS ILLUSTRATION

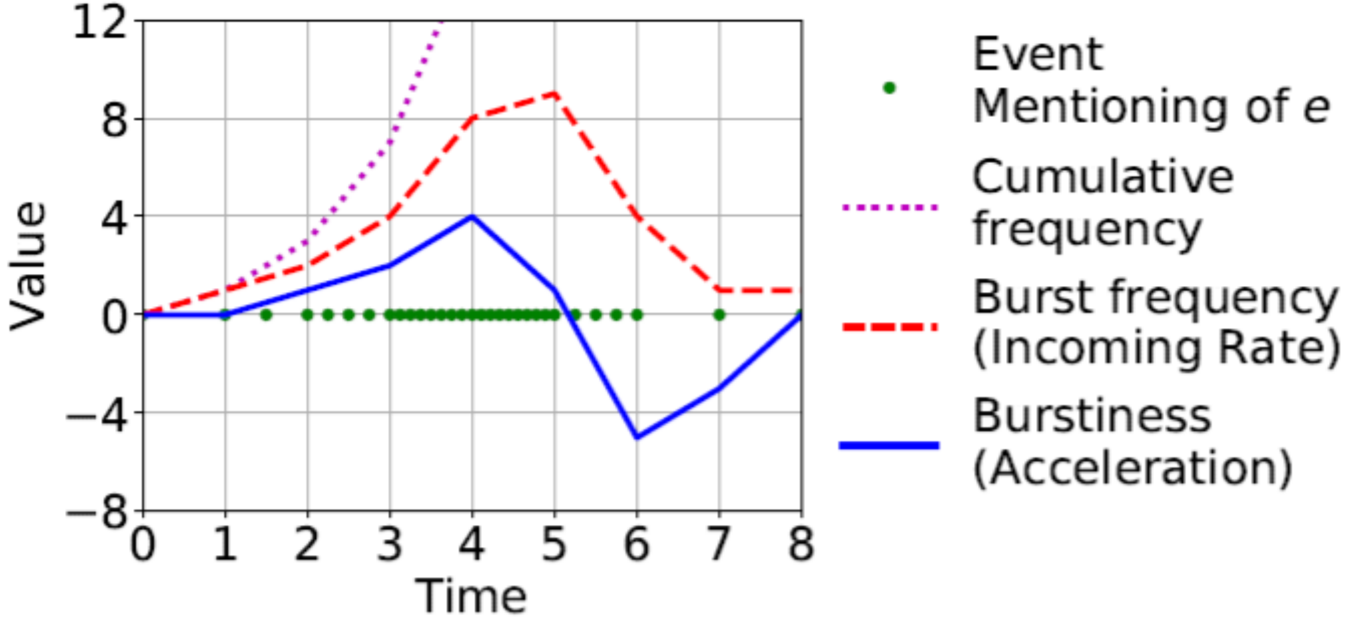


Figure 1: An example of burst where $\tau = 1$.

PBE1: OFFLINE OPTIMAL SOLUTION

- Input: P, The set of corner points in the original staircase
- Input: eta, the number of points in the output
- Output: P*, a subset of the input points with size eta
- Use Dynamic Programming to calculate optimal P*.
- $\Delta^*(i, j)$: The optimal solution when chooseing i points from the first j points in P

$$\Delta^*(i,j) = \min \begin{cases} \min_{x \in [i-1, j-1]} \Delta^*(i-1, x) - \delta(j, F^*(i-1, x)); & \text{Choose the j-th point} \\ \min_{x \in [i, j-1]} \Delta^*(i, x). & \text{Not choose the j-th point} \end{cases}$$

- Buffering in online case
  - Buffer η points, run DP, concatenate optimal staircases