GeoBurst: Real-time Local Event Detection in Geo-Tagged Tweet Streams

Chao Zhang[1], Guangyu Zhou[1], Quan Yuan[1], Honglei Zhuang[1],
Yu Zheng[2], Lance Kaplan[3], Shaowen Wang[1], Jiawei Han[1]

[1]UIUC
[2]Microsoft Research
[3]U.S. Army Research Lab

What is a Local Event?

- A local events is an ***unusual activity*** bursted within a ***local area*** and ***specific duration*** while engaging a considerable number of participants.
  - E.g., parade, riot, sport game, concert, accident, disaster.

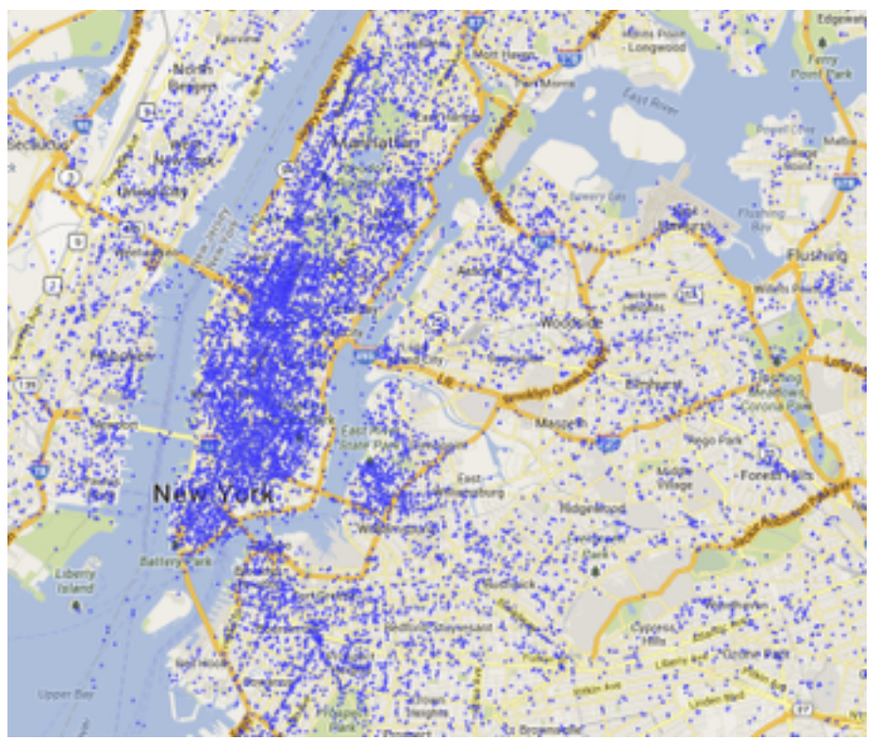
Local Event Detection

- Real-time local event detection is important for various applications
  - disaster monitoring
  - crime alarming
  - activity recommendation
  - …

Why Geo-Tagged Tweet Stream?

- Real-time local event detection is nearly impossible years ago due to the lack of timely and reliable data sources.

- The geo-tagged tweet stream brings new opportunities to this problem because of its (1) sheer size; (2) multi-dimensional information; and (3) real-time nature.



Our Goal

- Given the geo-tagged tweet stream, we aim to
  - detect all local events in any query time window (**batch mode**);
  - update the result list in real time as the query window shifts continuously (**online mode**).

Challenges

1.Integrate multiple types of data.

- Location, time and text have totally different representations.

2.Extracting interpretable events from massive noise.

- Raw tweets are extremely noisy and short.

3. On-line and real-time detection.

- To allow for timely actions, local events should be detected in real time.

Previous Studies

- Most existing event detection methods are designed for detecting ***global events***
  - They can successfully detect events that are bursty in the entire stream;
  - But local events are “bursty” in a small region and involve a limited number of tweets.

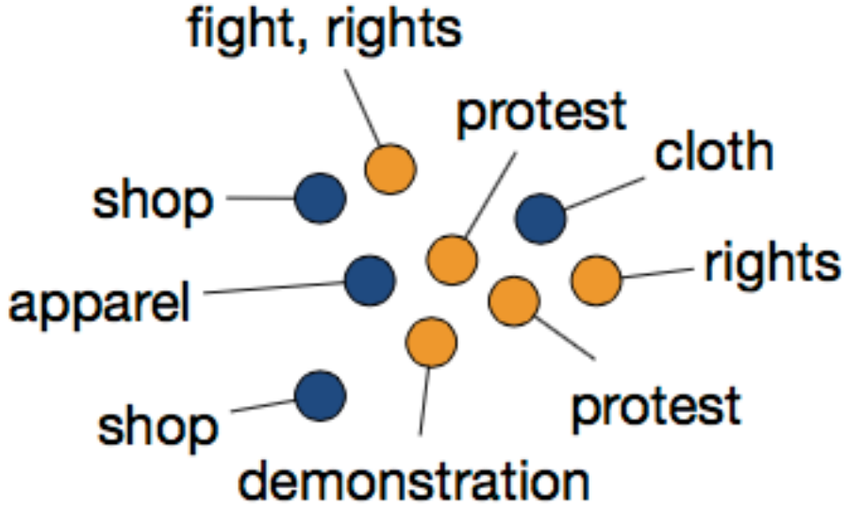
- A few methods for local event detection have been proposed
  - They either do not model the correlations between keywords; or are incapable of detecting local events in real time.

2011 ICWSM. Event detection in twitter.
2012 CIKM. Twevent: segment-based event detection from tweets.
2009 CIKM. Event detection from Flickr data through wavelet-based spatial analysis.
2013 PVLDB. EventTweet: Online localized event detection in the twitter stream.

Our Insight

- A local event usually leads to many related tweets around the location (**a geo-topic cluster**).

- But **a geo-topic cluster is not necessarily a local event**:
  - It may be a routine activity in that region (e.g., shopping).
  - It may be a global event rather than a local one (e.g., TV show).



We define a local event as a geo-topic cluster that shows clear spatiotemporal burstiness.

Overview of GeoBurst

- We propose GeoBurst, a reference-based method for local event detection. It consists of three key components:
  - **a candidate generator** that finds geo-topic clusters in the query time frame, and regard them as candidate events;
  - **a ranking module** that summarizes the routine activities in different regions to filter non-event candidates.
  - **an updater** that updates local events in real time as the query window shifts.

Candidate Event Generation

- The candidate generator finds geo-topic clusters in the query time frame as candidate events.

- Geo-topic cluster: a group of tweets that are geographically close and semantically relevant.

- Challenges for finding geo-topic clusters:
  - How to combine geographical and semantic similarities?
  - How to capture the correlations between different keywords?
  - How to cluster without knowing the number of clusters in advance?

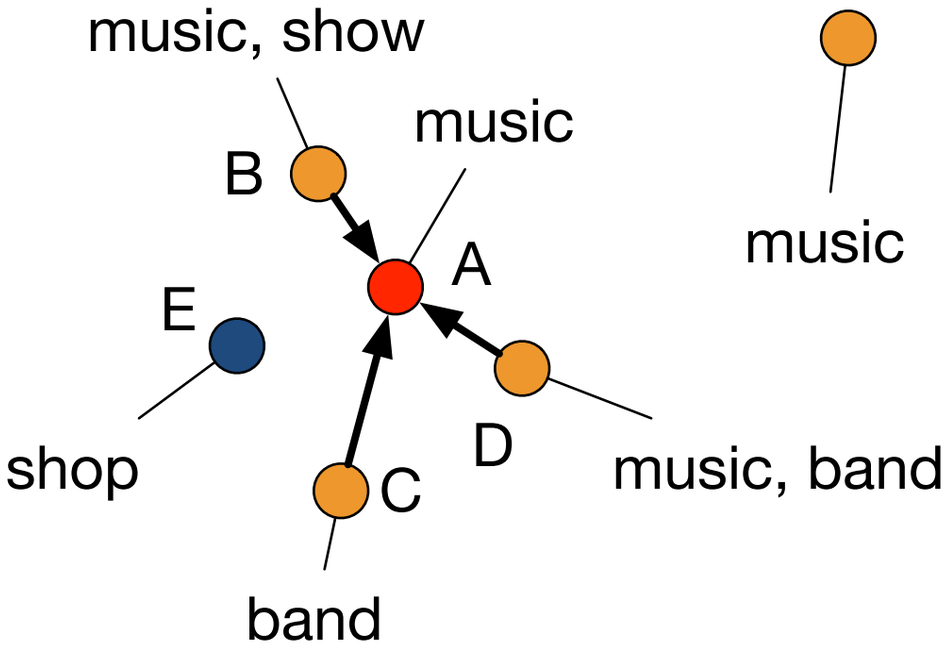
Candidate Event Generation

- Intuition: the spot where the event occurs is acting as a ***pivot*** that produces relevant tweets around it.

- Our clustering algorithm is based on:
  - a geo-topic authority score for each tweet
  - an authority ascent process to find authority maxima as pivots

Geo-topic Authority

- A tweet gets **an authority score** from neighbor tweets where
  - the geographical impact is captured by kernel function;
  - the semantic impact is captured by random walk on the keyword co-occurrence graph.



$$A(d) = \sum_{d' \in N(d)} G(d' \to d) \cdot S(d' \to d)$$

authority

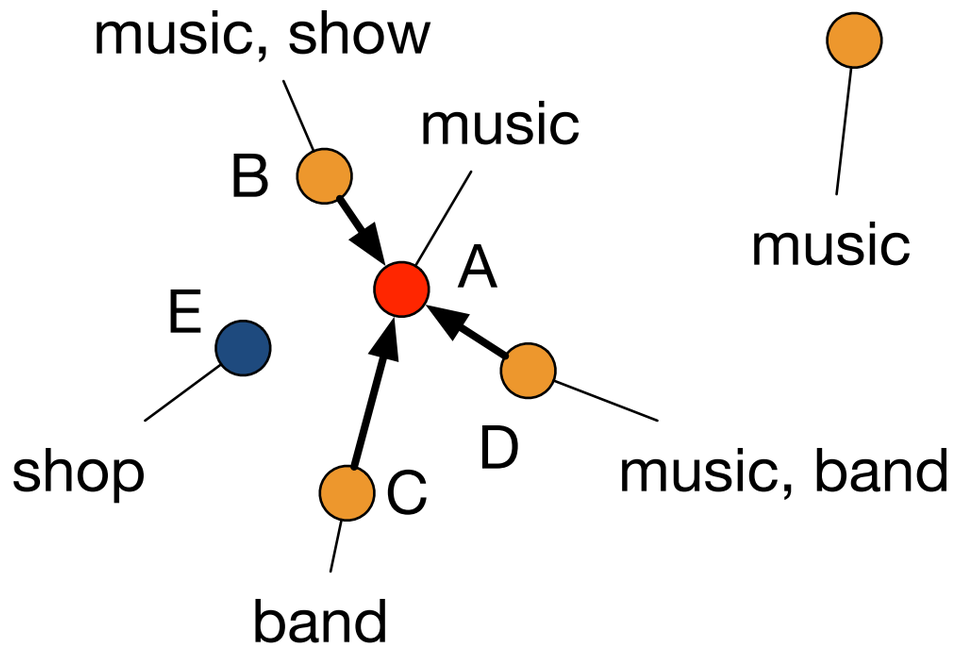
geo-impact

semantic impact

Authority can be interpreted as the total amount of energy received from the neighbors.

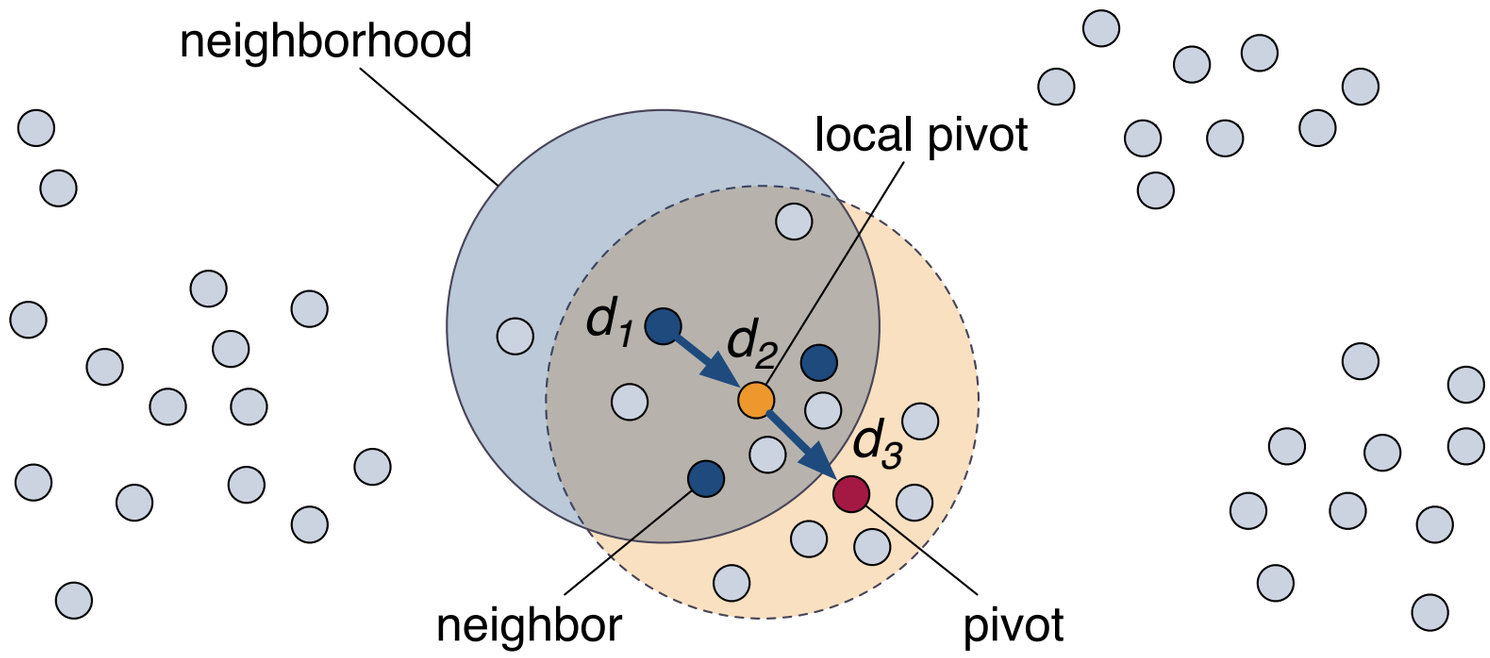
Pivot

- A **pivot** is an authority maximum: a prominent tweet that is surrounded by many relevant tweets.



Authority Ascent

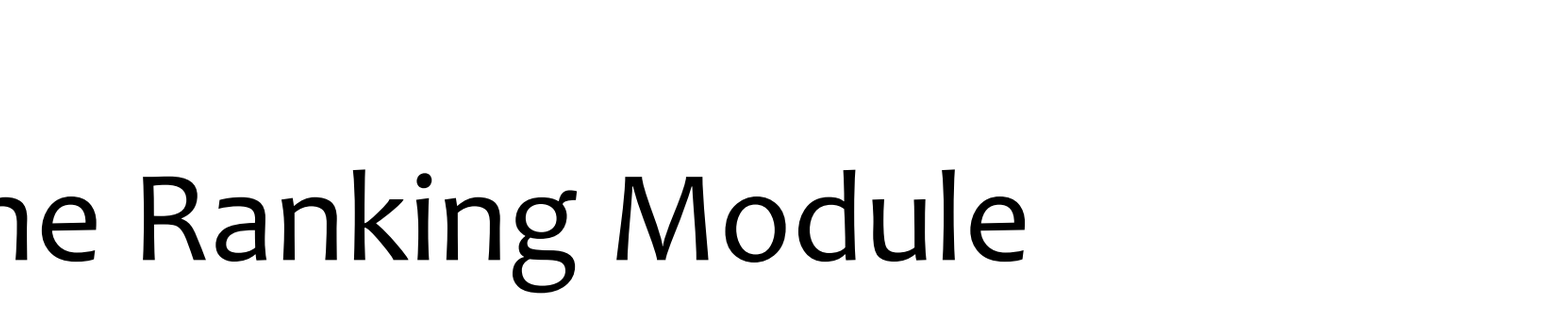
- Now the task is to find all the pivots in the geo-topic space.
- We design an **authority ascent** process to find all pivots.
- **A pivot attracts similar tweets** to form geo-topic clusters.



The Ranking Module

- We design the **activity timeline structure** to summarize the activities in different spatial regions and time periods.

- The summaries in the activity timeline serve as background knowledge to quantify the spatiotemporal burstiness of candidates.



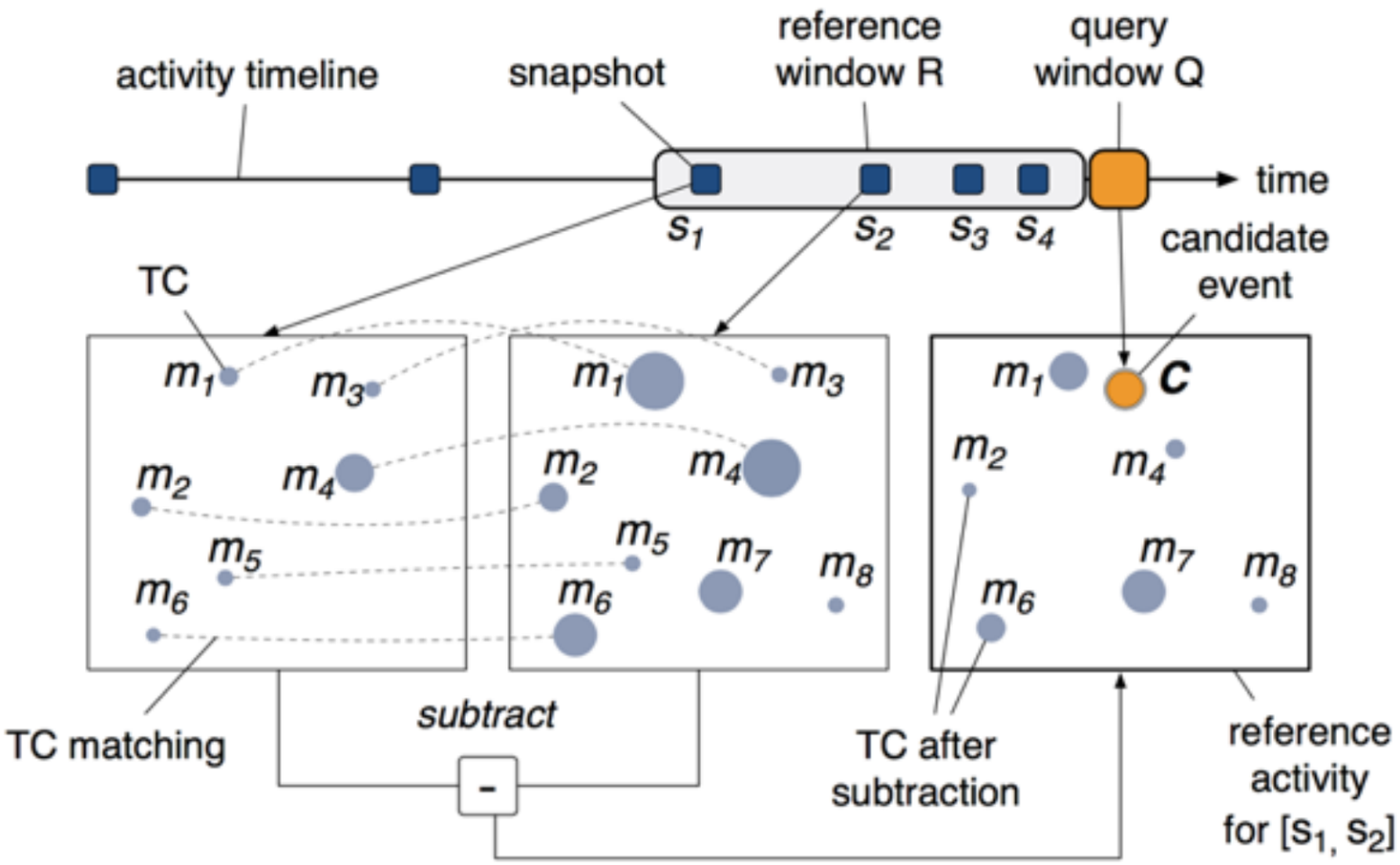
Each snapshot is a set of micro-clusters.

Each cluster is an activity summary for a region.

The Ranking Module

- Retrieve the snapshots in a reference window as background knowledge.

- Compute z-score for each candidate as its ranking score.

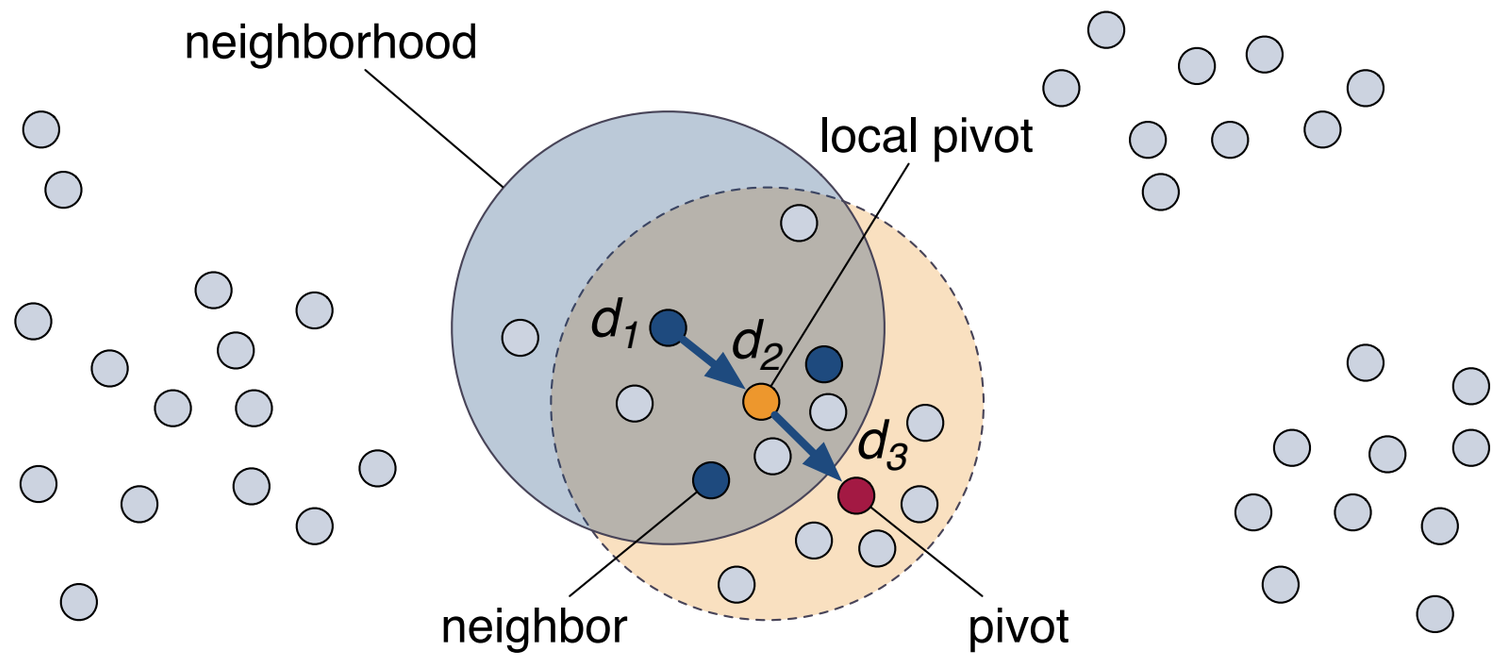


The Update Module

- In the entire process of GeoBurst, the most time-consuming step is pivot finding.

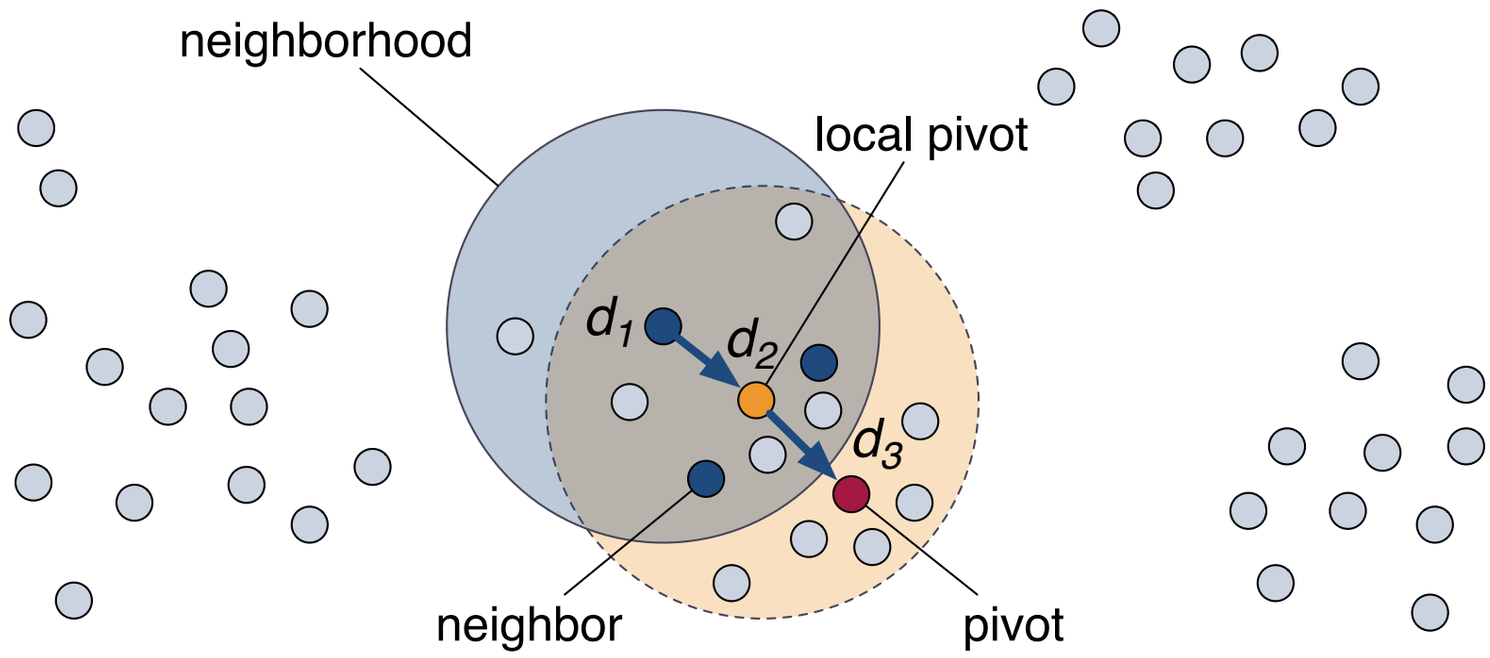
- How to avoid finding pivots from scratch as the query window shifts?

  - The key is to maintain the local pivot for each tweet.



The Update Module

- We design an updating strategy based on the additive property of authority score:
  - subtracting the contributions of outdated tweets
  - emphasizing the contributions of new tweets.



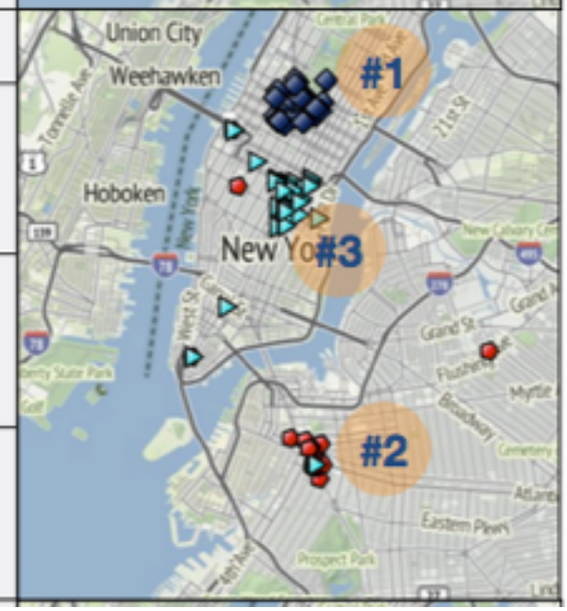
Experimental Settings

- Data:
  - NY: 9M geo-tagged tweets in New York during 3 months.
  - LA: 8M geo-tagged tweets in Los Angeles during 3 months.
- Task: 80 queries with different durations (3h, 4h, 5h, 6h), find top-5 local events in each query window.
- Compared Method: EvenTweet (PVLDB'13), Wavelet (CIKM'09)
- Evaluation: The crowdsourcing platform CrowdFlower
  - Ask the workers to judge whether the result is a local event or not.

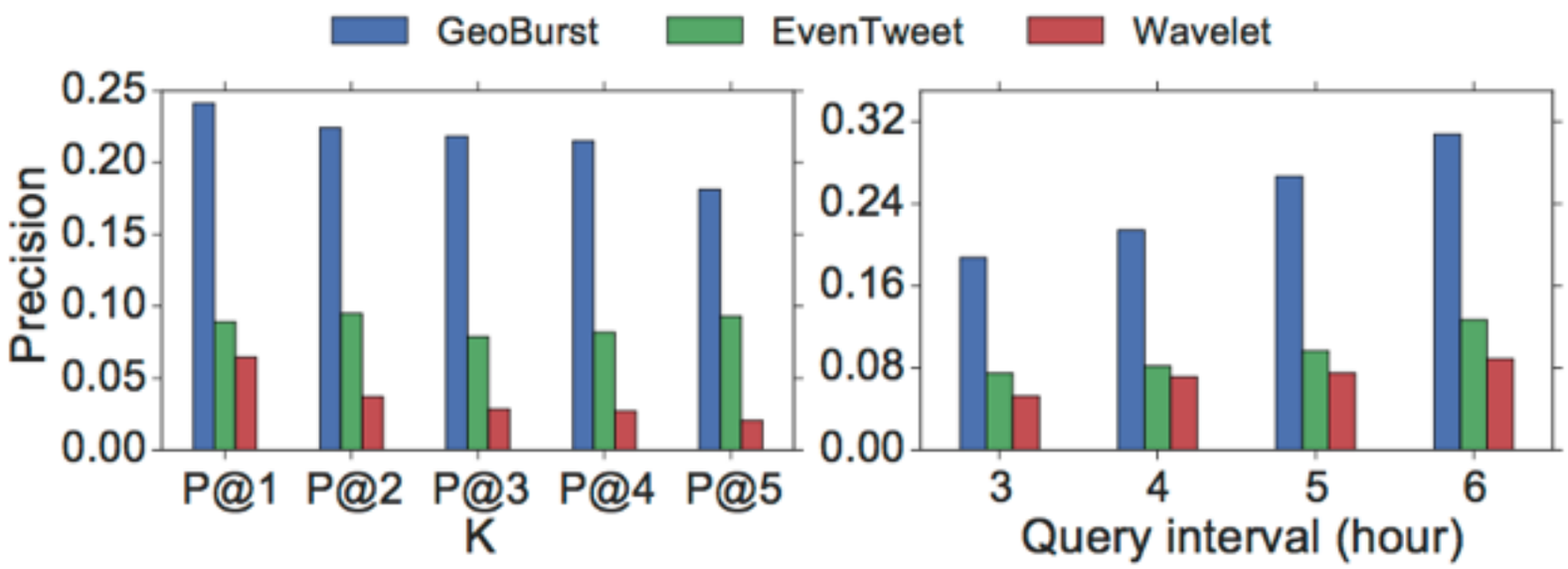
Illustrative Cases

	GeoBurst	Is event?
# 1	1. **Festival of Light**! **#nyfol** (@ **The Archway** under the **Manhattan Bridge** in **Brooklyn, NY**) 2. **#Lasers and beats** under the **Manhattan Bridge**! **#NewYorkFestivalofLight #NYFOL @ DUMBO** 3. **New York Festival of Lights #nyfol #dumbo @ DUMBO, Brooklyn**	**Yes**
# 2	1. **Knicks** vs. **Nets** at **Barclays Center**. @ Barclays Center http://t.co/PILk1xK3Tn 2. **Brooklyn** go hard @ **Barclays Center** http://t.co/iVUsJJ5TNG 3. Let's go **Knicks**! #NETS1107 (@ **Barclays Center** - @brooklynnets for @nyknicks vs @BrooklynNets)	**Yes**
# 3	1. **#Thai Restaurant** #spicythaifood (@ 104 **2nd Avenue** in **New York, NY**) 2. The **ASIAN DISHES** here are always my favorite. @ **Ugly Kitchen** 3. **Dinner time** with my family. Suuuuper Nice **Indian RESTAURANT**! @ Malai Marke **Indian Cuisine**.	No

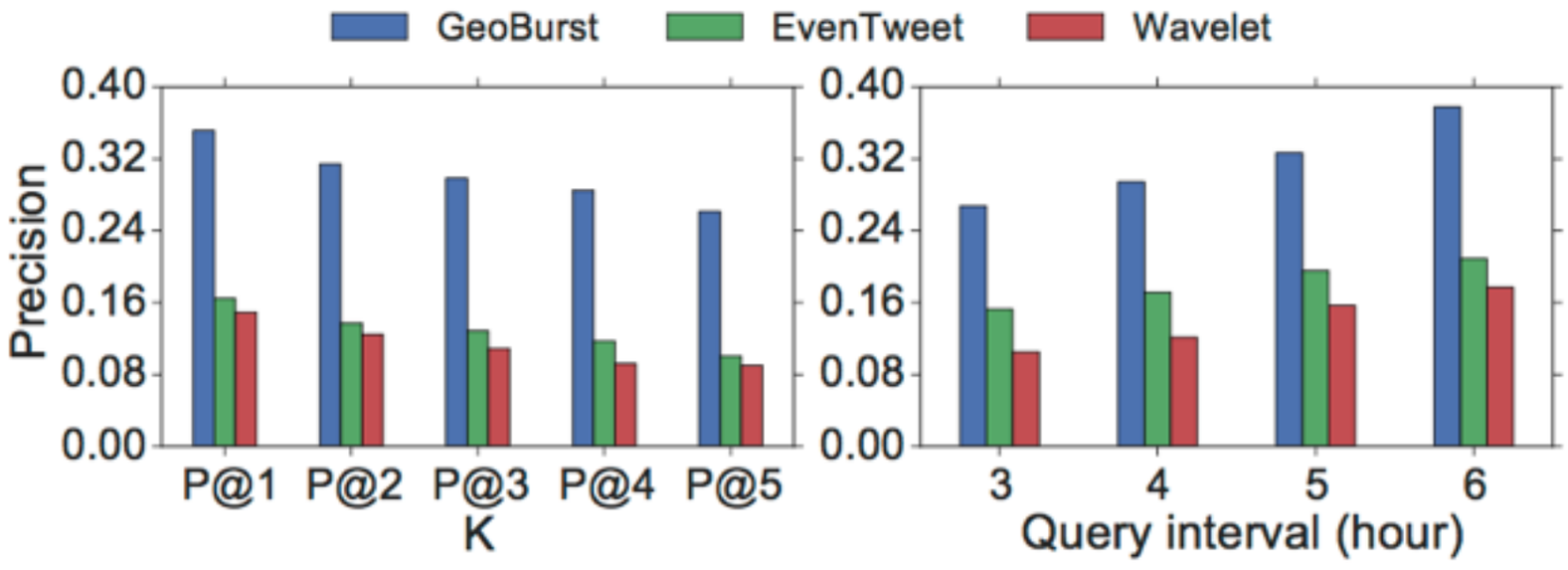
	EvenTweet	Is event?
# 1	1. I practiced... Almost time for **Amy Schumer**. **Jennifer** (@ **Carnegie Hall**) https://t.co/HfqfTLmK2y 2. **2014 Gold Glove Awards Ceremony** with **Hall of Famers, All-Stars Jay Leno @ The Plaza Hotel** 3. My best attempt at a **selfie** with **Hugh Jackman** after **The River at CITS @ The River on Broadway**	No
# 2	1. **Knicks** vs. **Nets** at **Barclays Center**. @ **Barclays Center** http://t.co/PILk1xK3Tn 2. **Budweiser** brings everyone together #family #nonewfriends @ **Alchemy Tavern, Brooklyn** 3. **#Knicks** vs **#nets** with my best gal. @ **Barclays Center Brooklyn** http://t.co/eXXMUKxpIs	**Yes**
# 3	1. **#katespade** @ **Kate Spade / Jack Spade HQ** http://t.co/g6jiFwyc4M 2. Inspiring **keynote** by **Twitter CEO, Dick Costolo** @GirlsWhoCode Gala. http://t.co/yEGh803CuT 3. I wonder if **Jake** from **Statefarm** covers **Jumanji**?	No



Precision



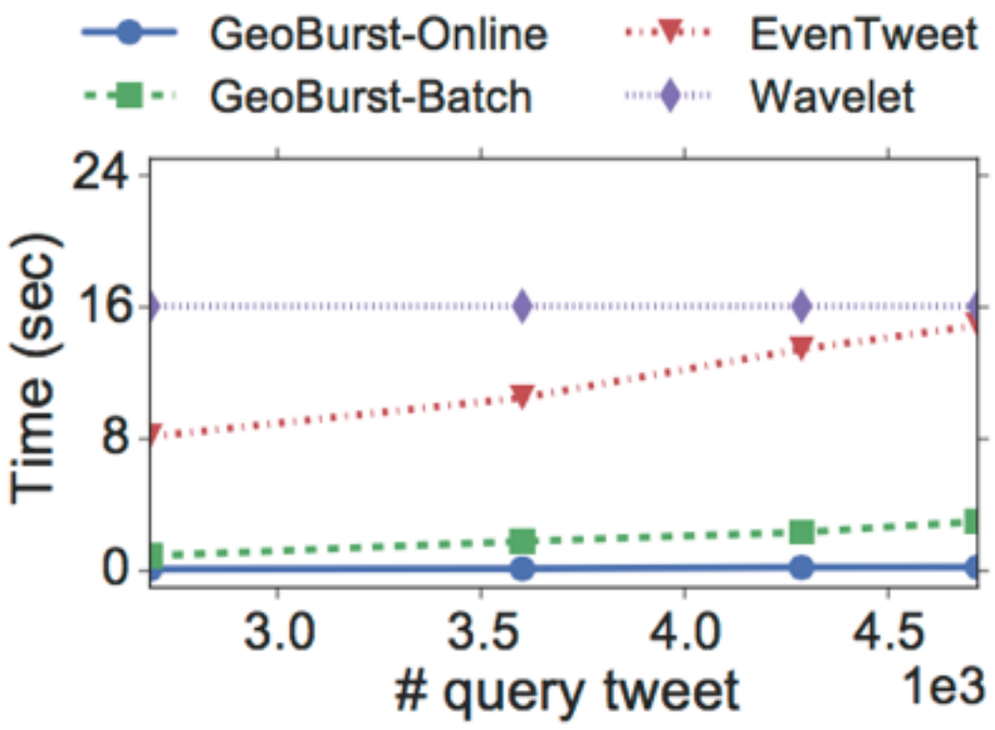
(a) Precision comparison (NY).



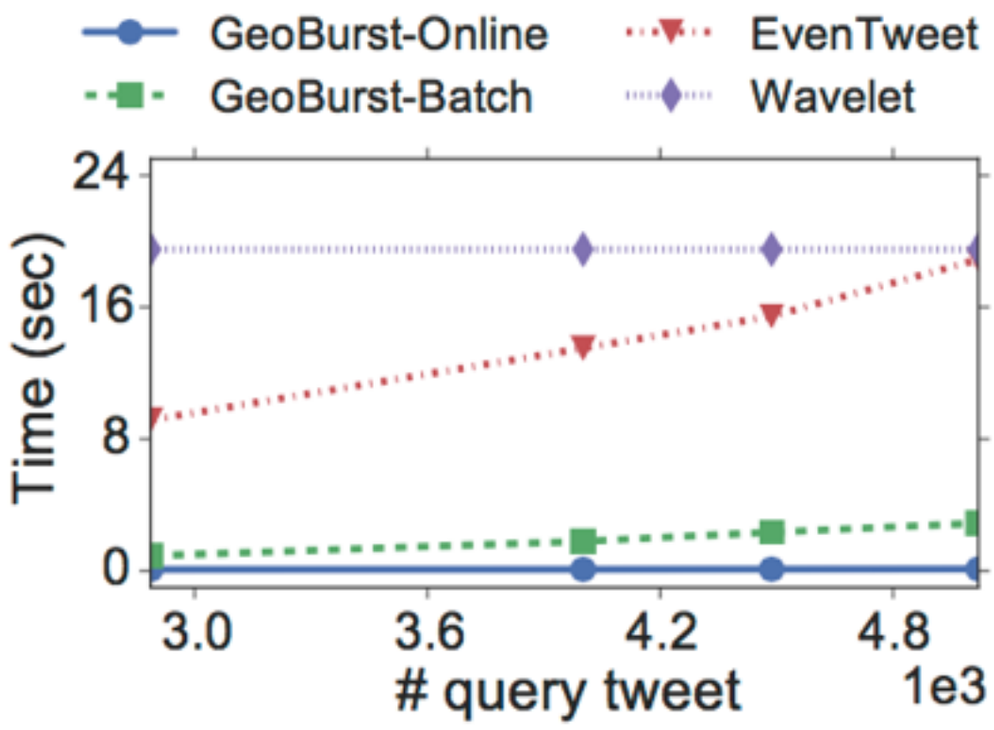
(b) Precision comparison (LA).

Running Time

1. GeoBurst is more efficient than the compared methods even when in batch mode.
2. The online mode of GeoBurst is more efficient.



(a) Running time (NY).



(b) Running time (LA).

Summary

- We study the problem of detecting local events from the geo-tagged tweet stream.
- We proposed the GeoBurst method.
  - It first detects candidate events based on authority ascent, and then ranks the candidates based on background knowledge.
  - It also features an updating module to continuously monitor the stream.
- Experiments demonstrate the effectiveness and efficiency of GeoBurst.
- For future work, we plan to extend GeoBurst to handle the tweets that mention geo-location names but do not have GPS information.

Thanks!