Mining Volatile Data

On the Spatiotemporal Burstiness of Terms

2012
Theodoros Lappas, Marcos R. Vieira, Dimitros Gunopulos, Vassilis J. Tsotras
(University of California, University of Athens)

14.11.2012 | Beatrix Vad

Outline

1) Terms and Definition
2) Spatiotemporal Burstiness of Terms
3) Applications
4) Experiments and Results
5) Summary

Terms and Definitions

Document Web

The New York Times ON THE WEB

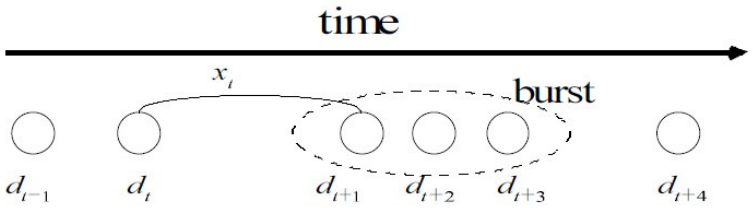
Google news

SPIEGEL ONLINE

Burst Identification

Search for term t in a stream of documents:

- **temporal bursts**
  - timeframe with unusually high frequency
    OR score to indicate strength of burst



- **spatial bursts**
  - geographical regions with unusually high frequency for a given timeframe

snowstorm

Spatiotemporal Term Burstiness

- Multiple document streams

What needs to be known...

- Document Stream
  - + timestamp
  - + geostamp

  - ▶ granularity for identifying set of streams is important

    e.g. set of streams ≙ city

Spatiotemporal Bursts

Given a term t and a set of document streams from different locations:

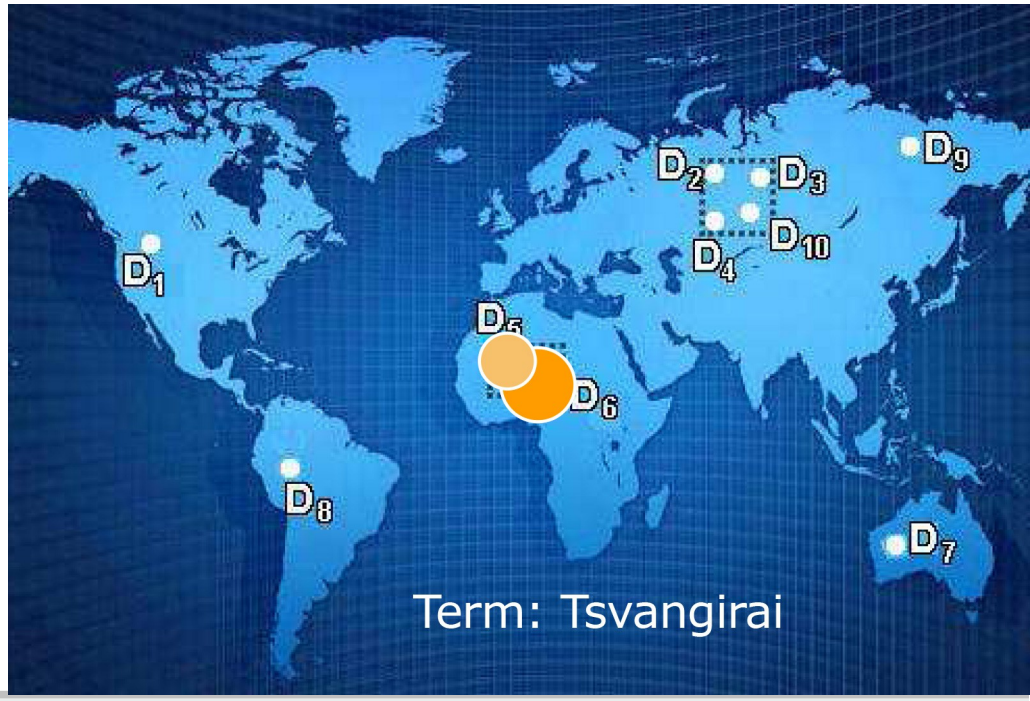
- **Find unusually high frequency in time and space**

Spatiotemporal Burstiness of Terms

Spatiotemporal Burstiness Patterns

- **Combinatorial Patterns**
  global effect
  + temporal interval
  + streams from diff. locations

- **Regional Patterns**
  local impact
  + temporal interval
  + region of streams



Combinatorial Patterns - STComb

- Overlapping segments represent a spatiotemporal pattern
  - burstiness score [0,1]
  - timeframe – span of segment
  - locations – geostamps of streams

Combinatorial Patterns - HSS

- **Finding the Highest-Scoring Subset (HSS):**

  U set of subsets with overlapping segments

  $B_T(I)$ burstiness score of interval

  ▶ Find $I^* \in U$: $I^* = argmax \sum B_T(I)$

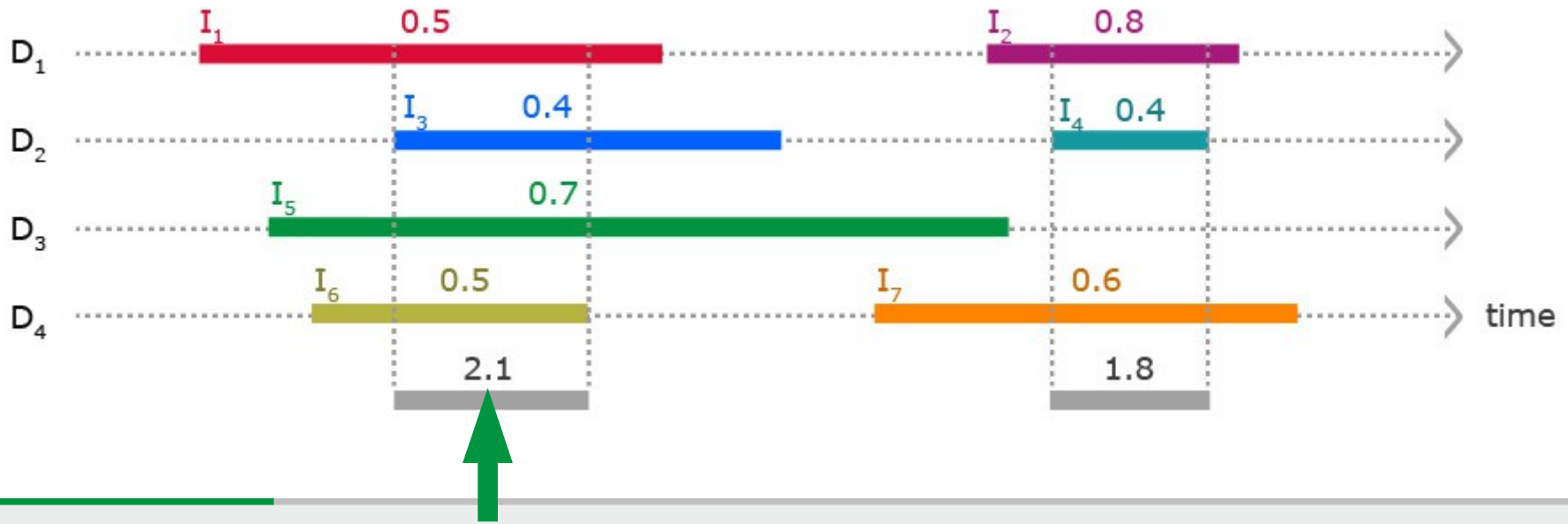
Combinatorial Patterns - HSS

- **Finding the Highest-Scoring Subset (HSS)**:

  U set of subsets with overlapping segments

  $B_T(I)$ burstiness score of interval

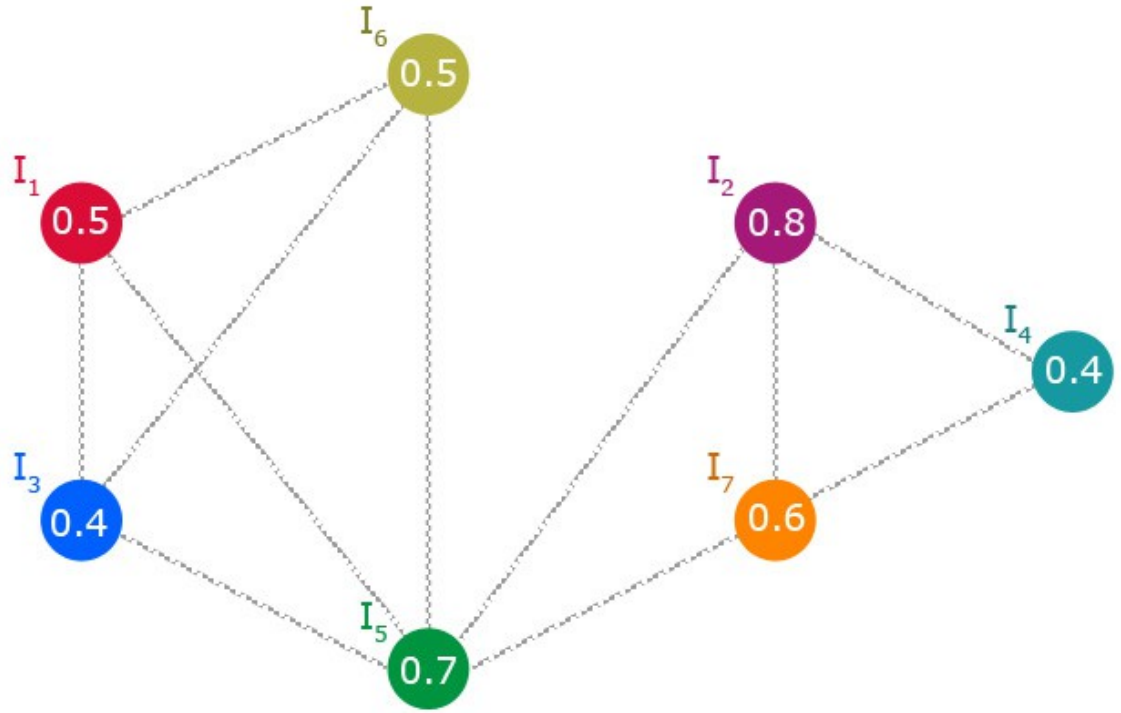
  ▶ Find $I^* \in U$: $I^* = argmax \sum B_T(I)$

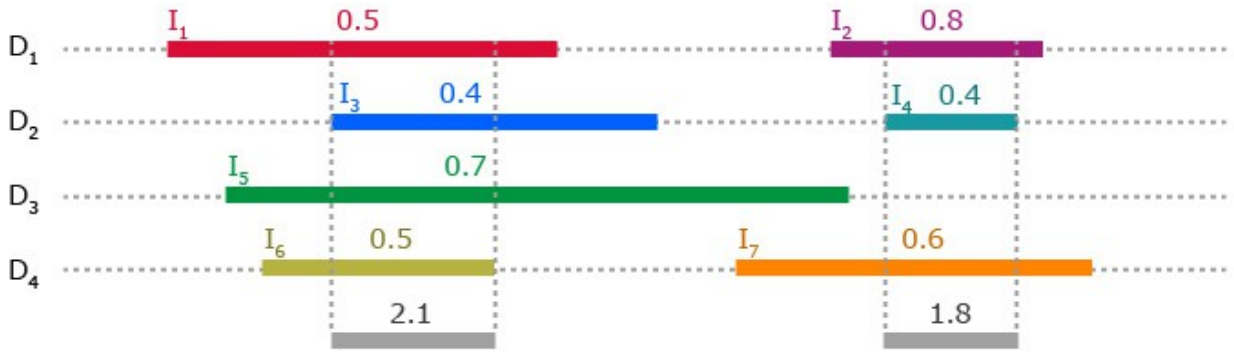


Combinatorial Patterns - MWCI

- **Maximum-Weight Clique Problem**

- Find highest-scoring clique / pattern in O(n logn)





i For multiple patterns apply MWCI iteratively

Regional Patterns - STLocal

Single Data Stream

▼

Snapshot of Entire Collection

▼

Streaming Data

Regional Patterns – Single Stream

- **Dicrepancy Concept**
  - observed frequency
  - expected frequency

$$B(t, D_x[i]) = D_x[i][t] - E_x[i][t]$$

i

$D_x[i][t]$: total frequency of term t in a set of documents at timestamp i

- average observed freq.
- recent measurements
- previous timeframes

Regional Patterns – Snapshot

- Find regions of streams
  - Axis-oriented rectangles → polynomial
  - Bursty rectangles:

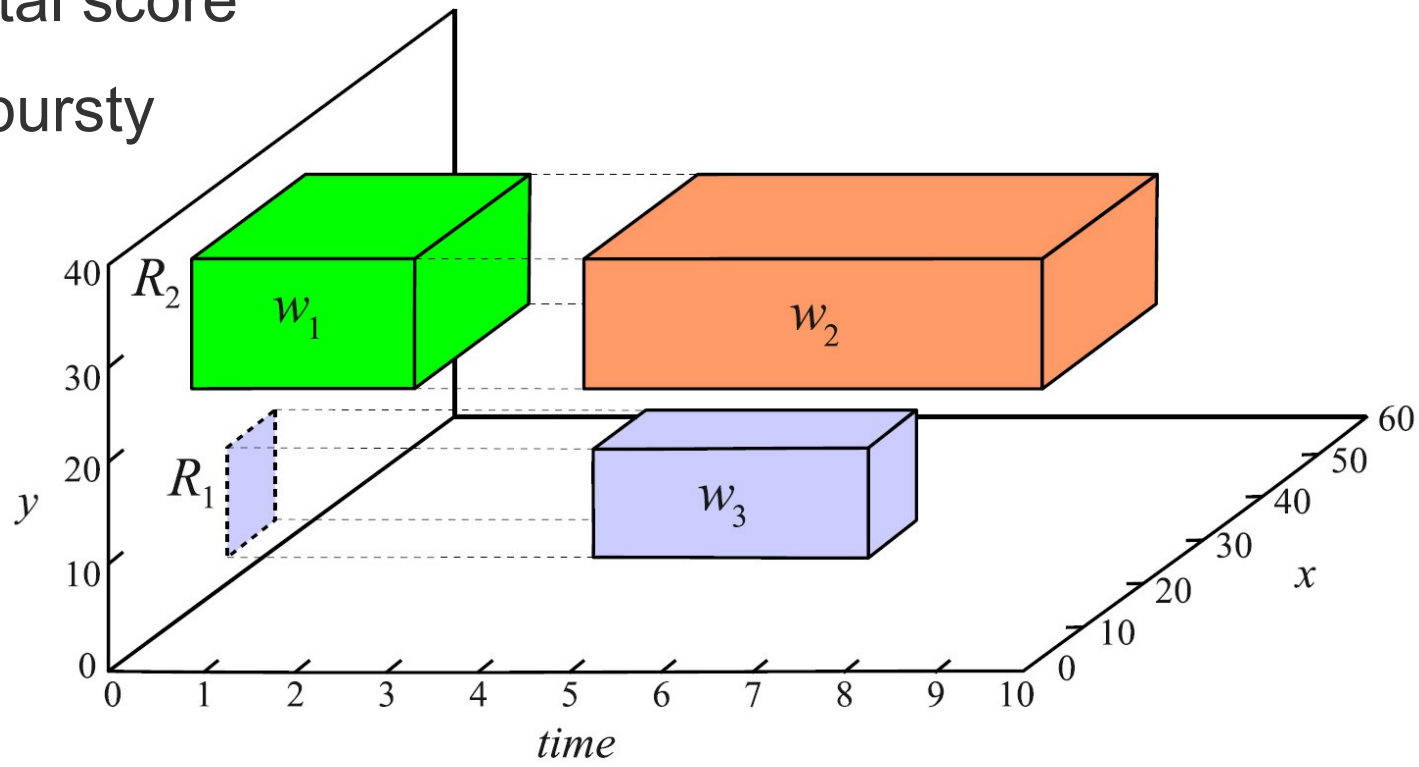
$$\sum_{D_x \in R} B(t, D_x[i]) > 0$$

positive if overall frequency higher than expected

  - Find non-overlapping rectangles and maximize burstiness

Regional Patterns – Streaming Data

- **Maximal Spatiotemporal Window**
  - window is maximal if no super-window with higher total score
  - remove non-bursty streams



Applications

Document Search Engine

- query of terms

- ranked list of documents for influential events with spatiotemporal impact

(influential events affecting multiple places for a long time)

Document Selection

- tracking bursts of t across space & time

- present users with articles according to location & timeframe

Trend Identification

- a set of terms to describe item
- identify when and where it was popular

Experiments and Results

Dataset

- **305,641 articles** form Topix.com
  - 181 countries
  - September 2008 – July 2009
- **Major Events List** from Wikipedia.com
  - query from Major Events List

Precision Evaluation

	#	Query	TB	STLocal	STComb
global	1	Obama	1.0	1.0	1.0
	2	financial crisis	1.0	1.0	1.0
	3	terrorist	1.0	1.0	1.0
	4	Jackson	0.9	1.0	1.0
	5	swine	1.0	1.0	1.0
	6	earthquake	1.0	1.0	1.0
major	7	gaza	1.0	1.0	1.0
	8	ceasefire	1.0	1.0	1.0
	9	Yemenia	1.0	1.0	1.0
	10	piracy	1.0	1.0	1.0
	11	Air France	1.0	1.0	1.0
	12	bush fires	1.0	1.0	1.0
local	13	Nkunda	0.7	1.0	0.8
	14	Vieira	0.8	1.0	1.0
	15	Tsvangirai	0.9	1.0	1.0
	16	Rajoelina	0.7	1.0	1.0
	17	Fujimori	0.8	1.0	1.0
	18	Zelaya	1.0	1.0	1.0

- **TB**: Temporal Search
- **Top-10** documents

- Human annotators
  - relevant
  - not relevant

Similarity of top-k sets:

STComb	TB	0.61
STComb	STLocal	0.58
STLocal	TB	0.67

Pattern Evaluation – Top Bursty

	#	Query	# countries in STLocal	# countries in STComb	# countries in MBR
global	1	Obama	176	136	181
	2	financial crisis	113	159	181
	3	Jackson	132	151	181
	4	terrorists	98	126	167
	5	swine	174	157	181
	6	earthquake	17	81	171
major	7	gaza	174	116	179
	8	ceasefire	36	52	156
	9	Yemenia	19	21	125
	10	piracy	24	39	174
	11	Air France	50	67	179
	12	bush fires	3	30	168
local	13	Nkunda	30	2	118
	14	Vieira	15	22	114
	15	Tsvangirai	4	24	123
	16	Rajoelina	4	30	154
	17	Fujimori	5	19	158
	18	Zelaya	26	55	171

i

top-scoring burstiness patterns

Pattern Evaluation – Top Bursty

	#	Query	# countries in STLocal	# countries in STComb	# countries in MBR
global	1	Obama	176	136	181
	2	financial crisis	113	159	181
	3	Jackson	132	151	181
	4	terrorists	98	126	167
	5	swine	174	157	181
	6	earthquake	17	81	171
major	7	gaza	174	116	179
	8	ceasefire	36	52	156
	9	Yemenia	19	21	125
	10	piracy	24	39	174
	11	Air France	50	67	179
	12	bush fires	3	30	168
local	13	Nkunda	30	2	118
	14	Vieira	15	22	114
	15	Tsvangirai	4	24	123
	16	Rajoelina	4	30	154
	17	Fujimori	5	19	158
	18	Zelaya	26	55	171

Pattern Evaluation – Top Bursty

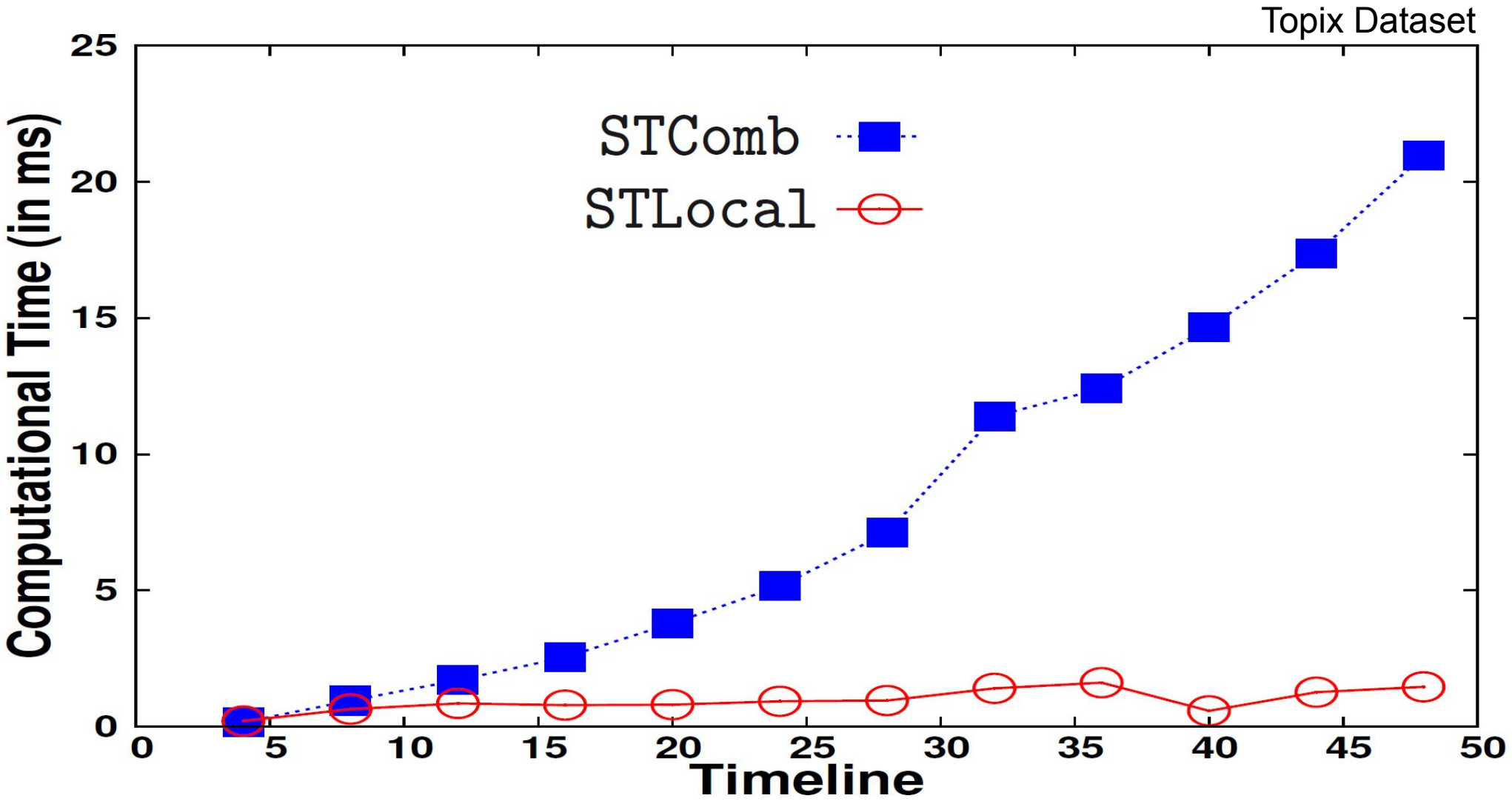
#	Query	# countries in STLocal	# countries in STComb	# countries in MBR
1	Obama	176	136	181
2	financial crisis	113	159	181
3	Jackson	132	151	181
4	terrorists	98	126	167
5	swine	174	157	181
6	earthquake	17	81	171
7	gaza	174	116	179
8	ceasefire	36	52	156
9	Yemenia	19	21	125
10	piracy	24	39	174
11	Air France	50	67	179
12	bush fires	3	30	168
13	Nkunda	30	2	118
14	Vieira	15	22	114
15	Tsvangirai	4	24	123
16	Rajoelina	4	30	154
17	Fujimori	5	19	158
18	Zelaya	26	55	171

global

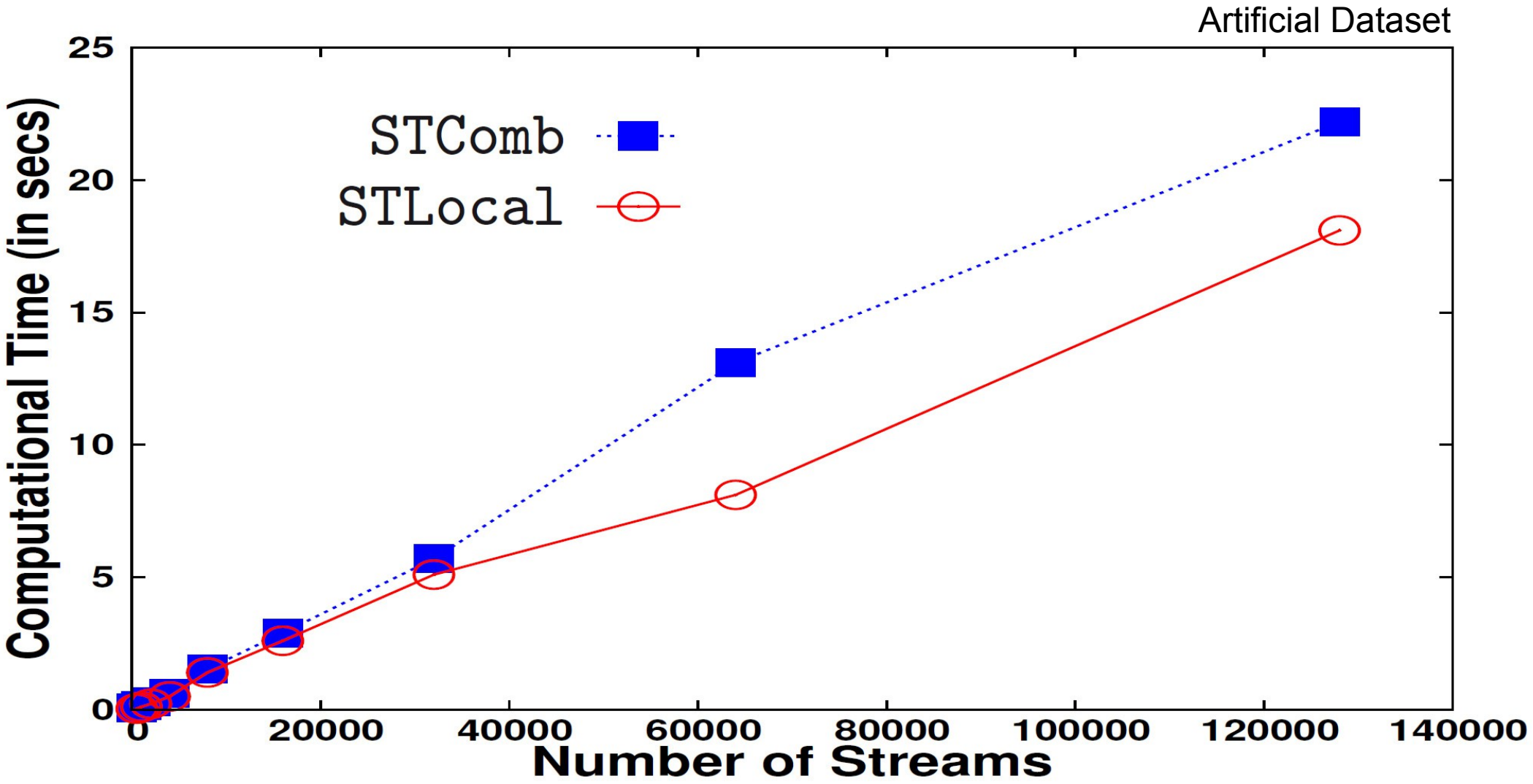
major

local

Performance Evaluation - Speed



Performance Evaluation - Scalability



Summary

- **Spatiotemporal Burstiness of Terms**
- **Patterns**

  Combinatorial | Regional Patterns
- **Applications**

  Search Engine
- **Experiments**

Conclusion and Future Work

- **STComb** → track all locations affected by events with major spatiotemporal impact
  → **Search Engine**
- **STLocal** → bounded by spatial proximity
  → enhance for **Online Streaming Data**
  → extend to support arbitrary regions