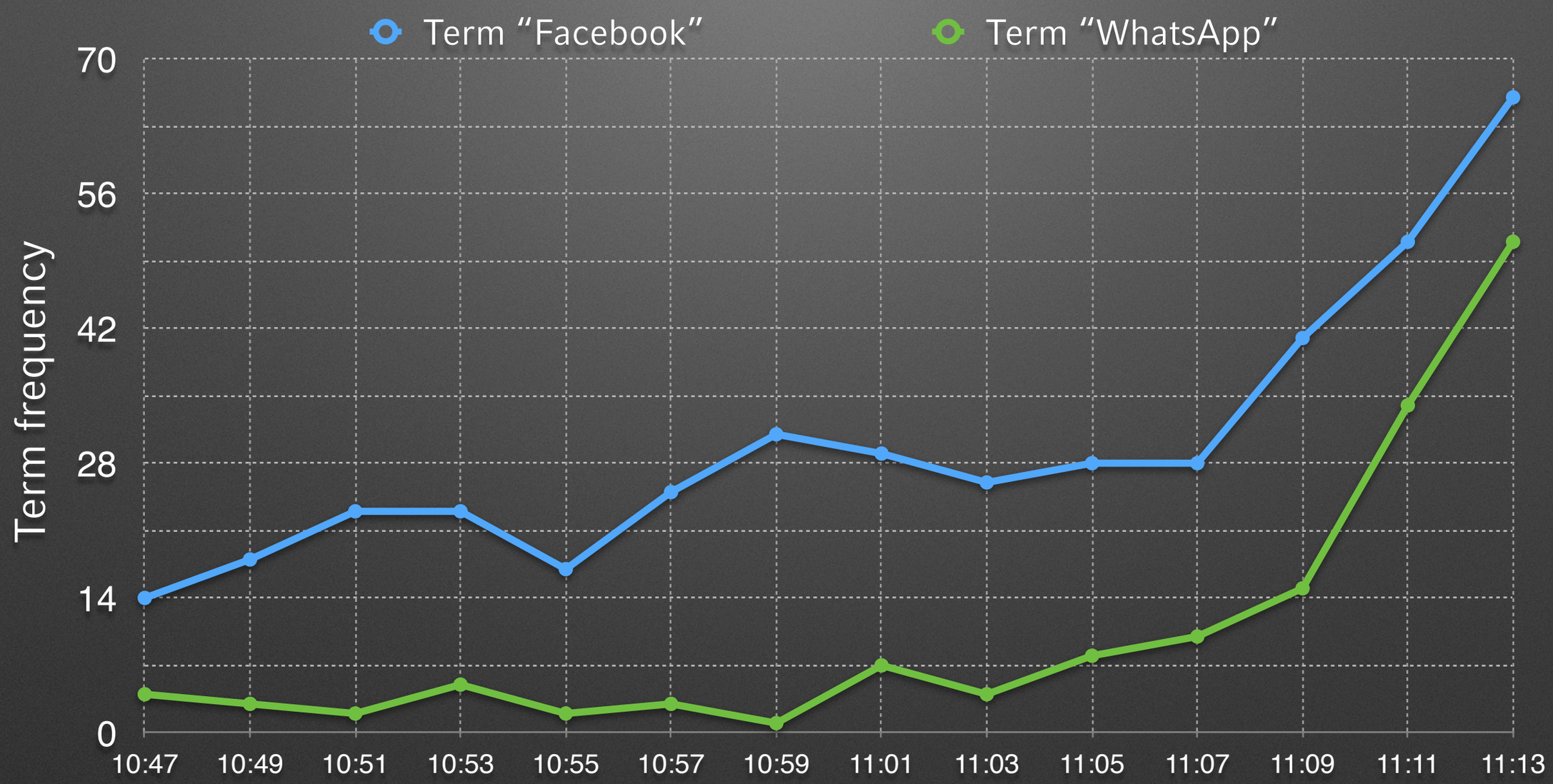
SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds

Erich Schubert, Michael Weiler, Hans-Peter Kriegel

Institute of Informatics
Database Systems Group
Ludwig-Maximilians-Universität München

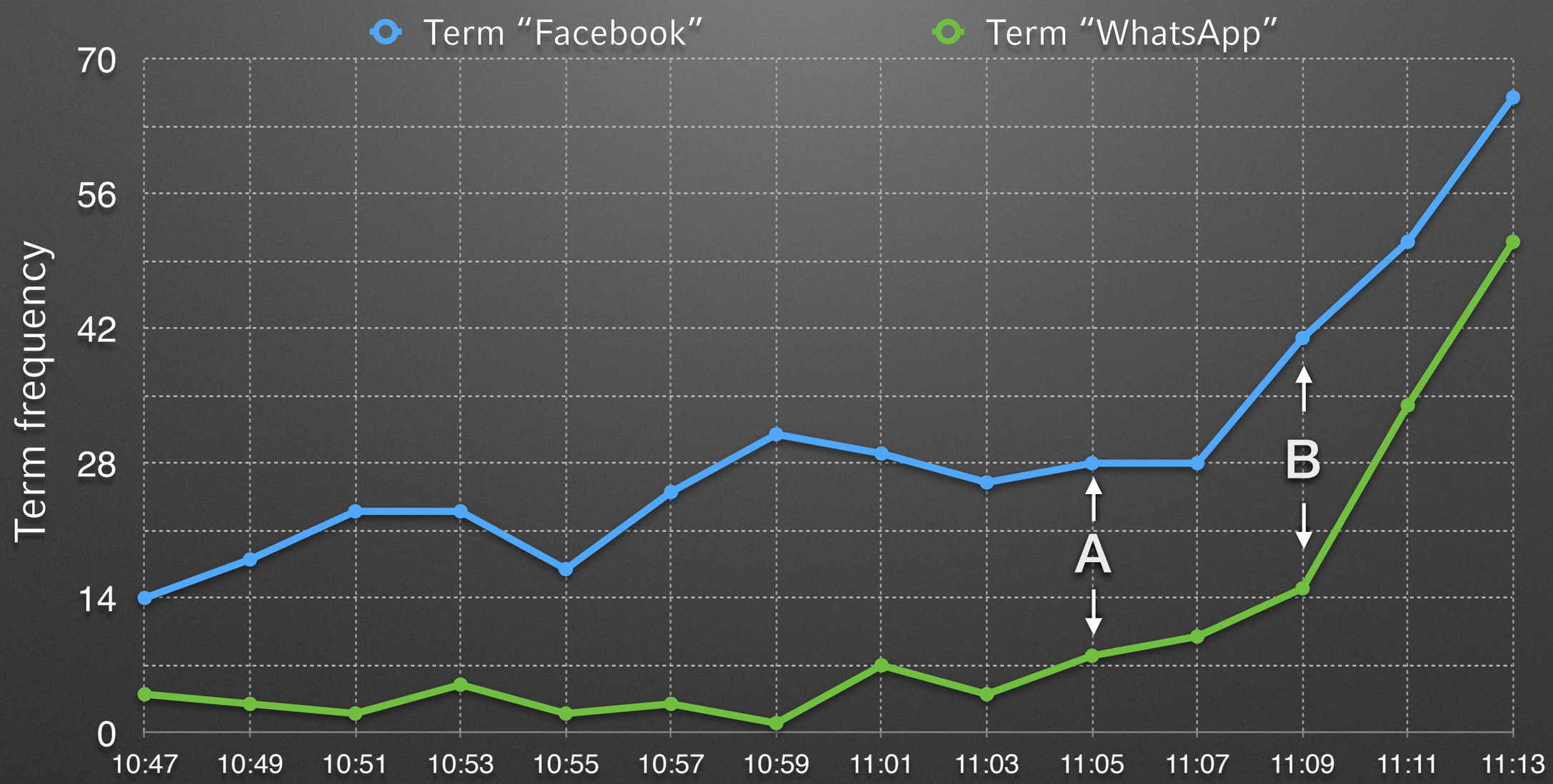
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Trend detection on streams should be early and accurate



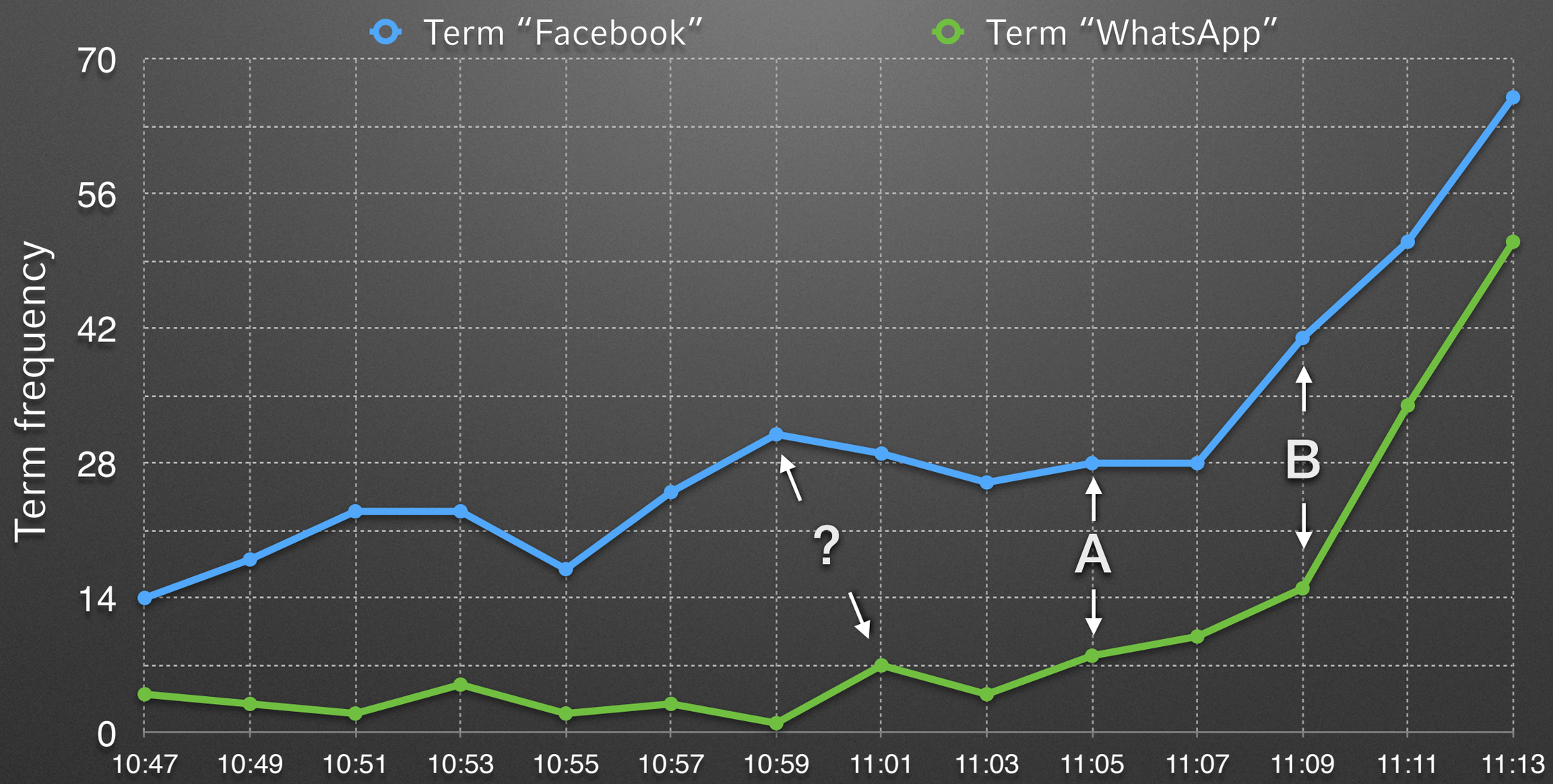
Twitter Streaming API on Feb. 19th 2014

Trend detection on streams should be early and accurate



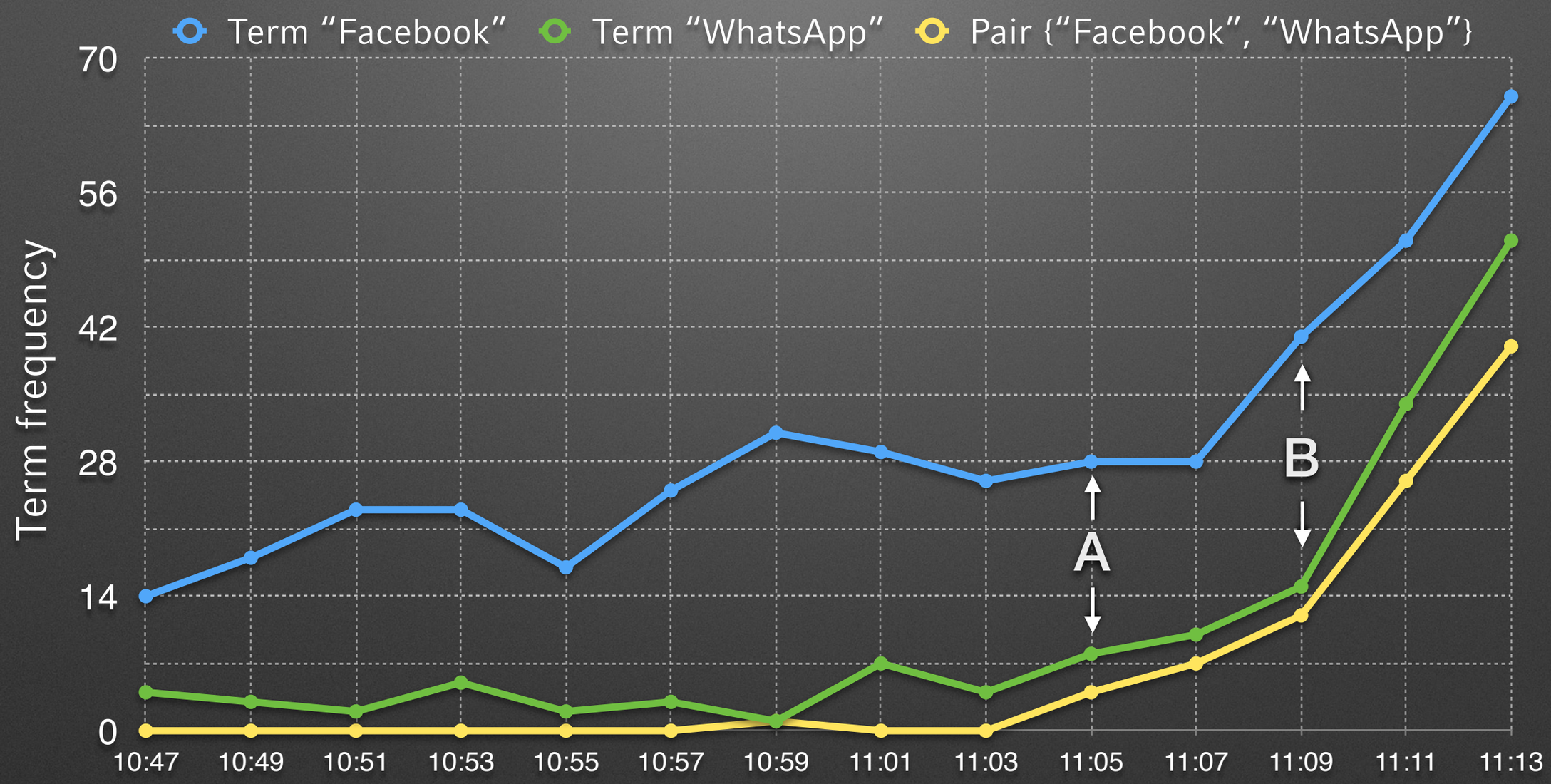
Twitter Streaming API on Feb. 19th 2014

Trend detection on streams should be early and accurate



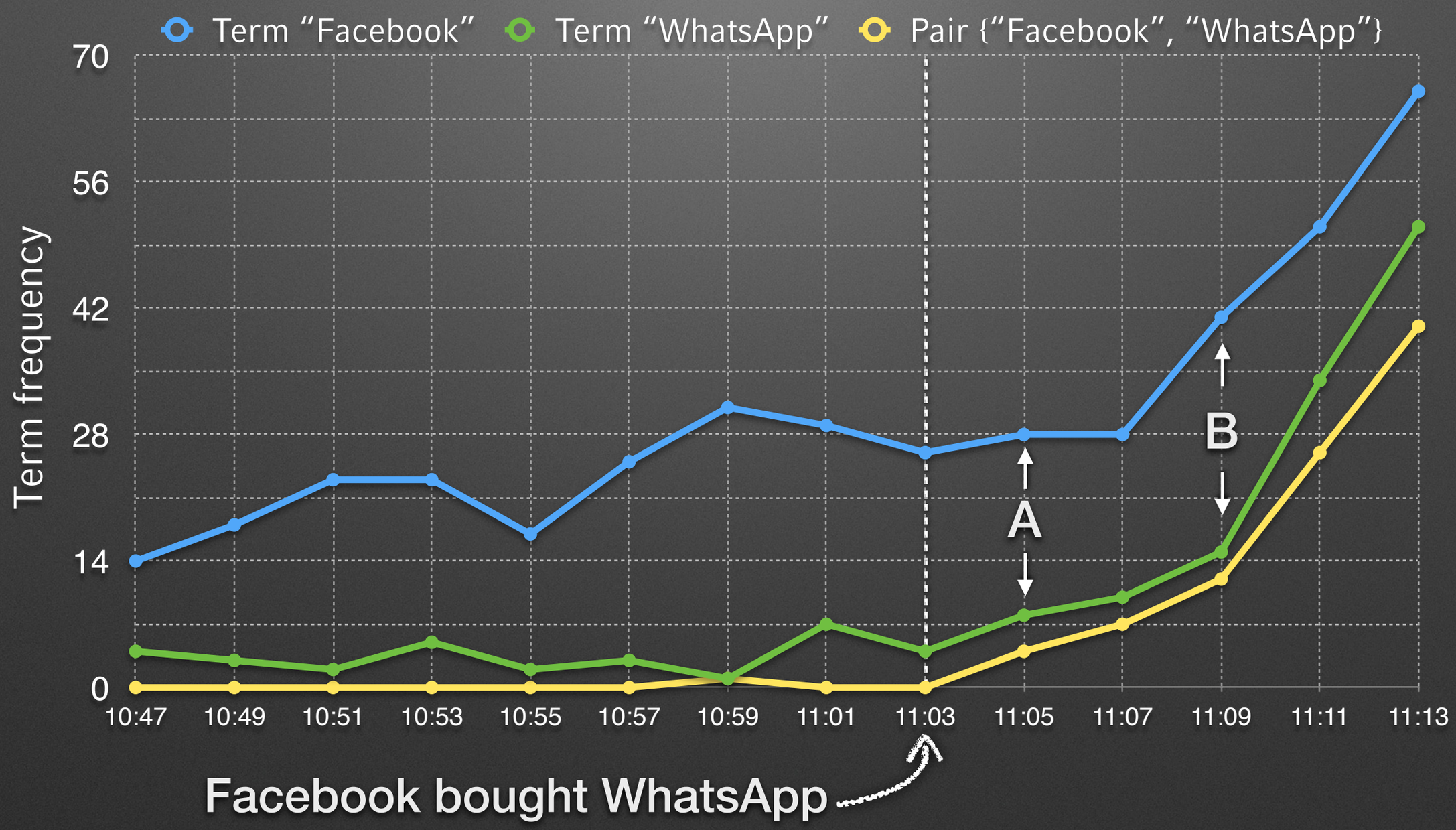
Twitter Streaming API on Feb. 19th 2014

Trend detection on streams should be early and accurate



Twitter Streaming API on Feb. 19th 2014

Trend detection on streams should be early and accurate



Problem description

1. **Statistical significance score**
   Popular topics ≠ trending topics (e.g. Obama)
2. **Track interacting terms**
   - Facebook bought WhatsApp
   - Edward Snowden traveled to Moscow
3. **Scalability**
   Efficient calculation for **all** terms and pairs

SigniTrend on textual streams

tracking both: single terms and pairs

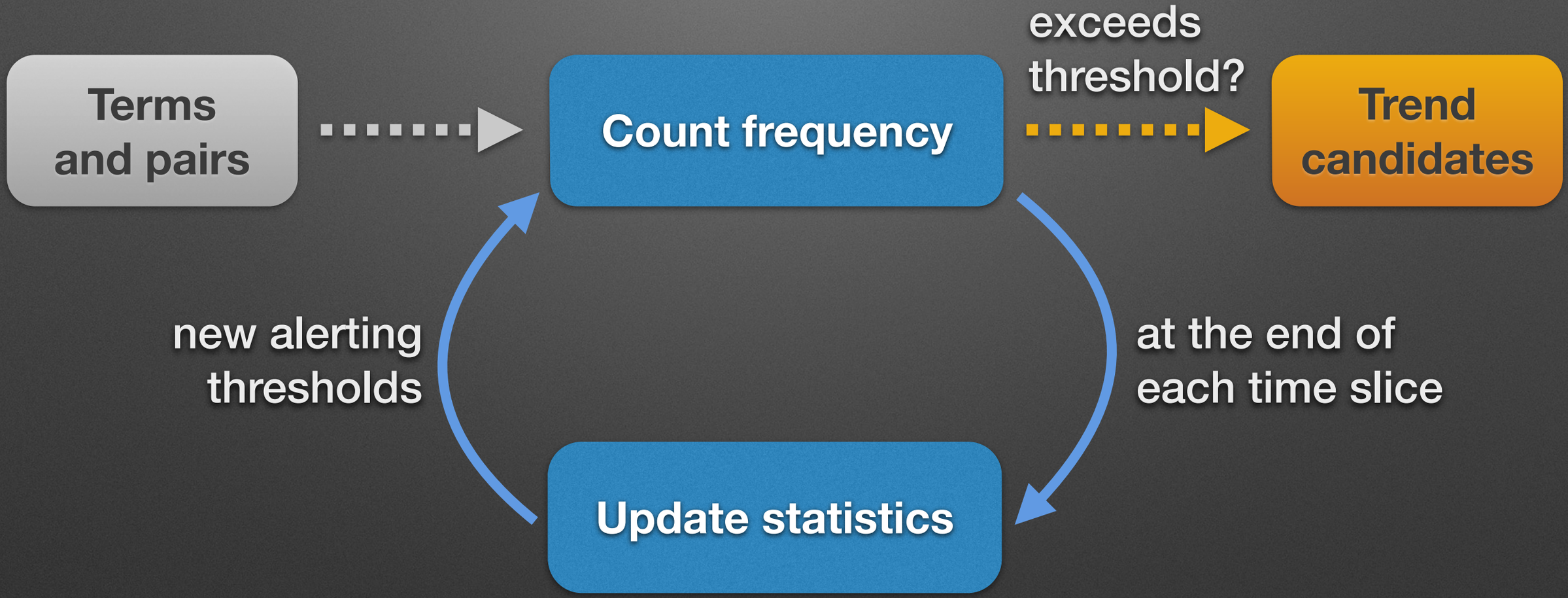
A. Preprocessing (stopwords, stemming, duplicates)

B. Trend detection cycle

- Temporal slicing for statistical aggregation
- Score all terms and pairs based on expectations from past slices

C. Refinement with clustering

Trend detection cycle



Update statistics
for time slice t and term or pair e

- How many standard deviations is the current frequency x higher than its mean

$$z\left(x_{t,e}\right) := \frac{x_{t,e} - \mu_{t-1,e}}{\sigma_{t\text{-}1,e}}$$

Update statistics
for time slice t and term or pair e

- How many standard deviations is the current frequency x higher than its mean

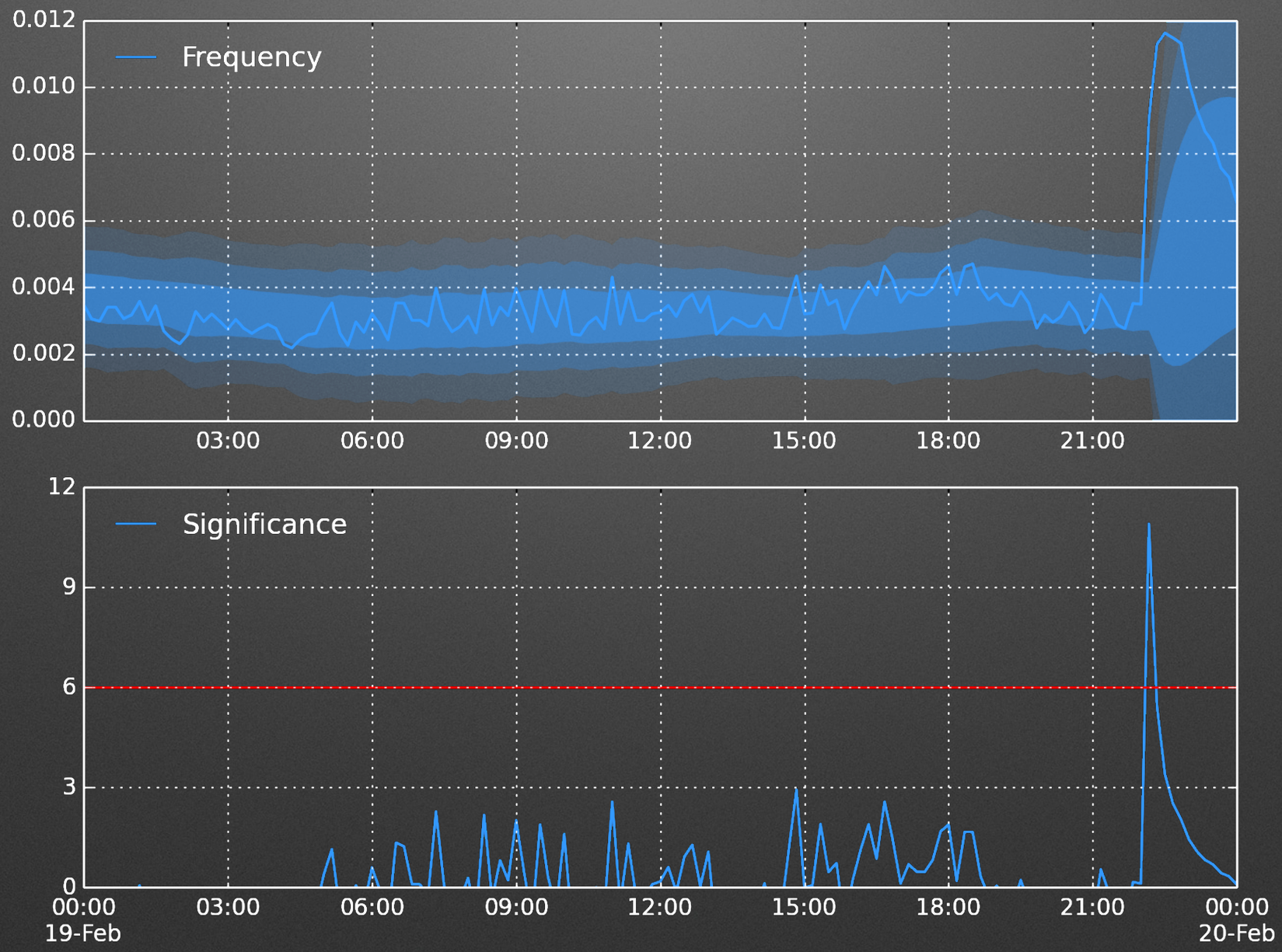
$$z\left(x_{t,e}\right) := \frac{x_{t,e} - EWMA_{t-1,e}}{\sqrt{EWMVar_{t-1,e}}}$$

- Exponential weighted moving average/variance for continuous estimation on a stream [Finch09]

$$\begin{aligned}
\triangle_{t,e} &\leftarrow x_{t,e} - EWMA_{t-1,e} \\
EWMA_{t,e} &\leftarrow EWMA_{t-1,e} + \alpha \cdot \triangle_{t,e} \\
EWMVar_{t,e} &\leftarrow (1-\alpha) \cdot \left(EWMVar_{t-1,e} + \alpha \cdot \triangle_{t,e}^2\right)
\end{aligned}$$

[Finch09] T. Finch. Incremental calculation of weighted mean and variance. Technical report, University of Cambridge, 2009

Significance and frequency for term “Facebook”



How to track statistics of all pairs efficiently?

Problem: Too many terms and pairs to track everything

2013 News Dataset

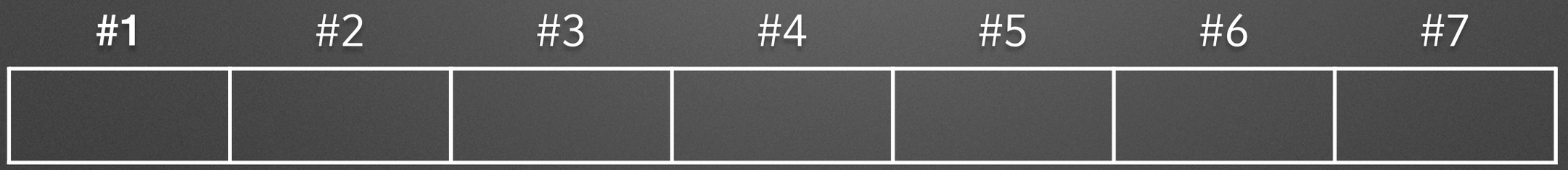
	STEMMED TERMS	OBSERVED PAIRS
TOTAL	56,661,782	660,430,059
UNIQUES	300,141	71,289,359

Therefore, we designed an **efficient hashing scheme**
(based on Bloom Filters and Heavy Hitters)
for probabilistic upper-bound statistics

Hashing scheme for efficient tracking

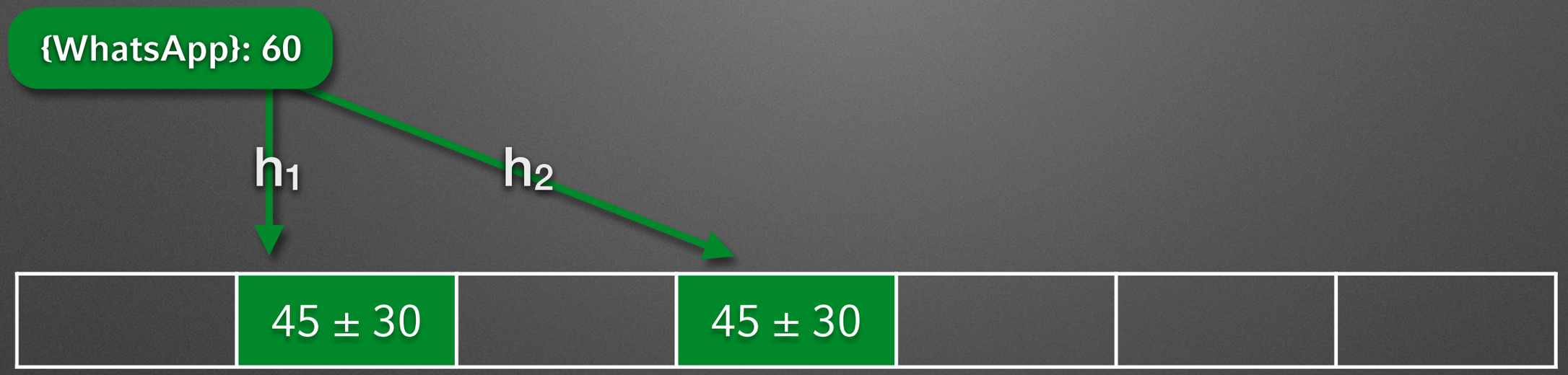
L=7 buckets, K=2 hash functions

{WhatsApp}: 60



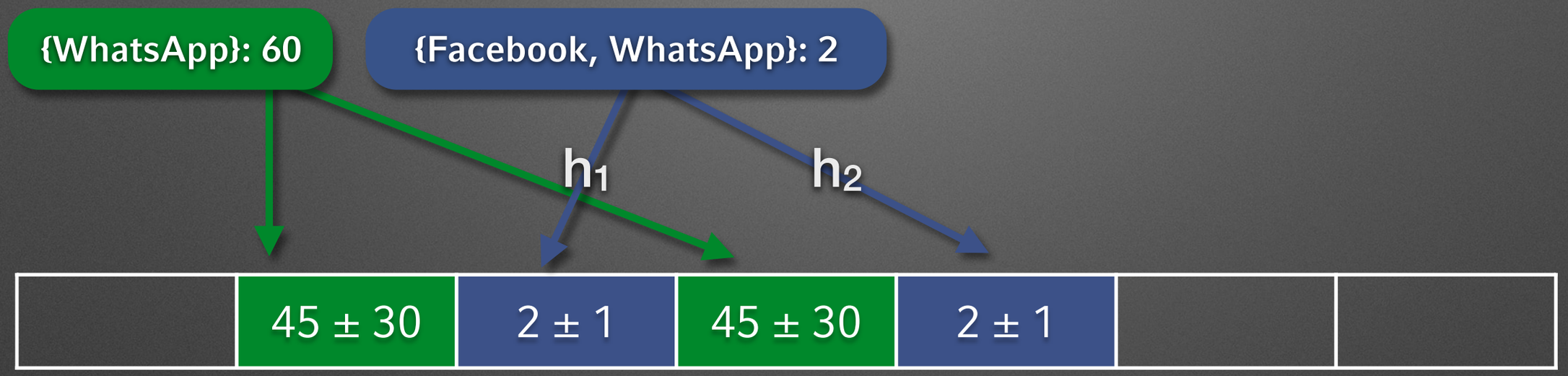
Hashing scheme for efficient tracking

L=7 buckets, K=2 hash functions



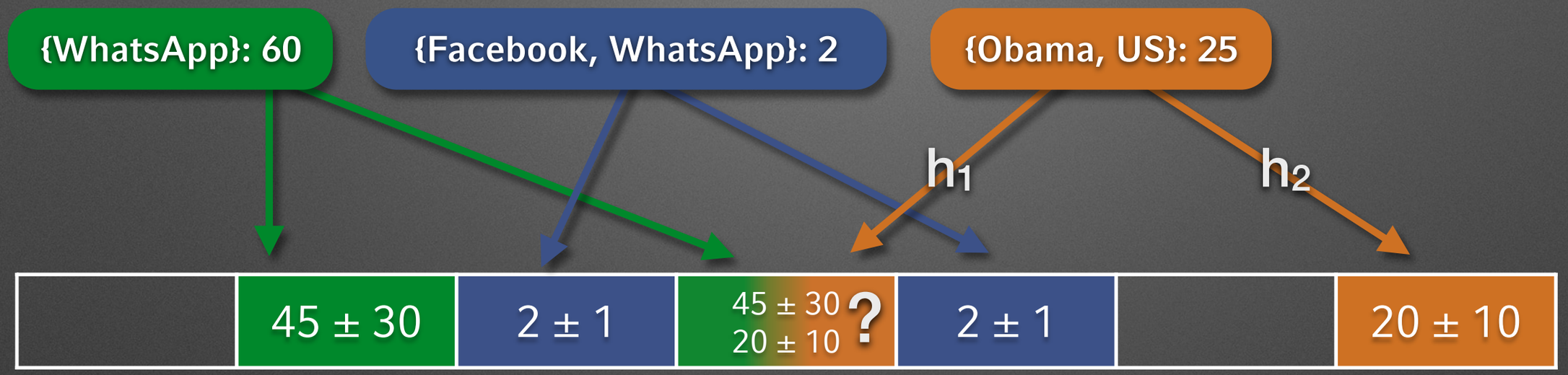
Hashing scheme for efficient tracking

L=7 buckets, K=2 hash functions



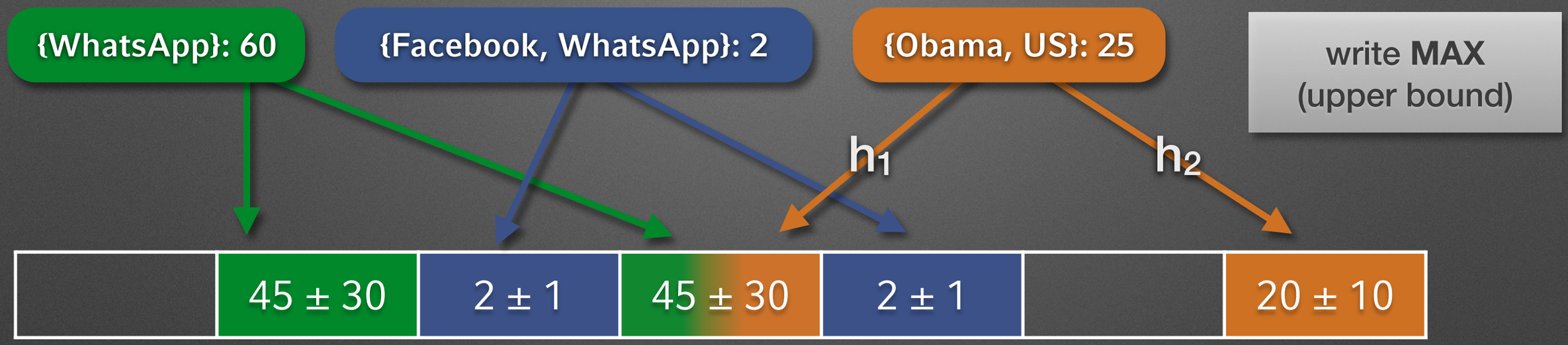
Hashing scheme for efficient tracking

L=7 buckets, K=2 hash functions



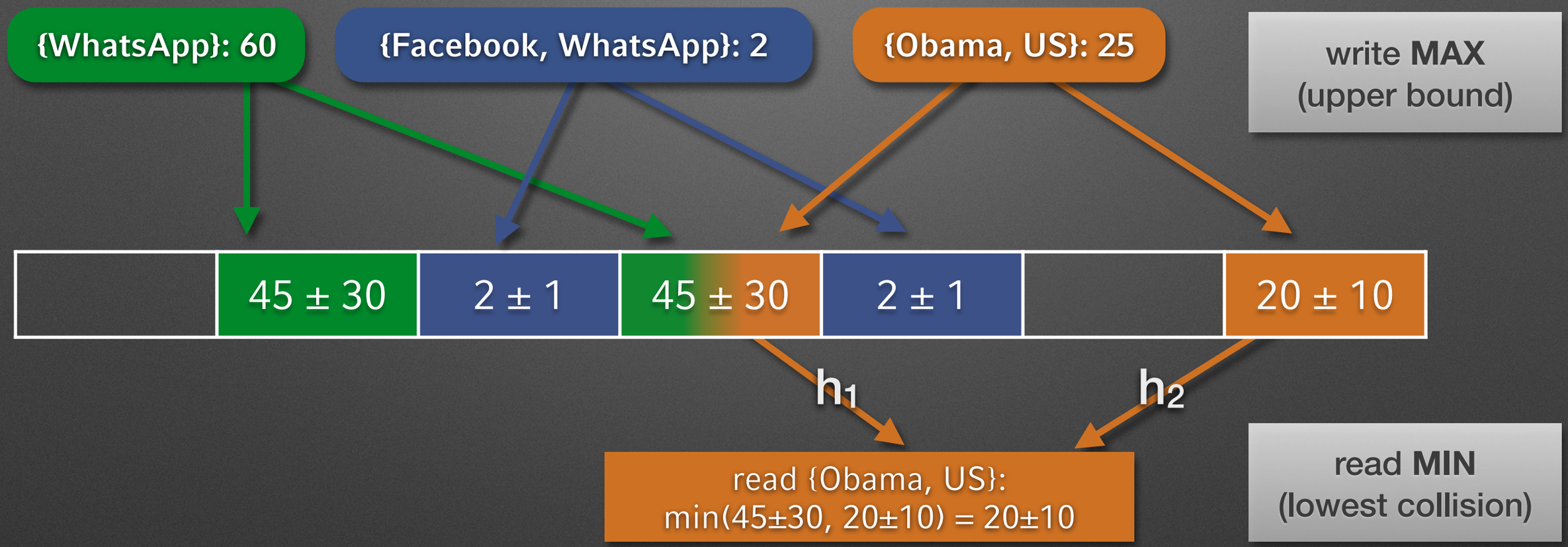
Hashing scheme for efficient tracking

L=7 buckets, K=2 hash functions



Hashing scheme for efficient tracking

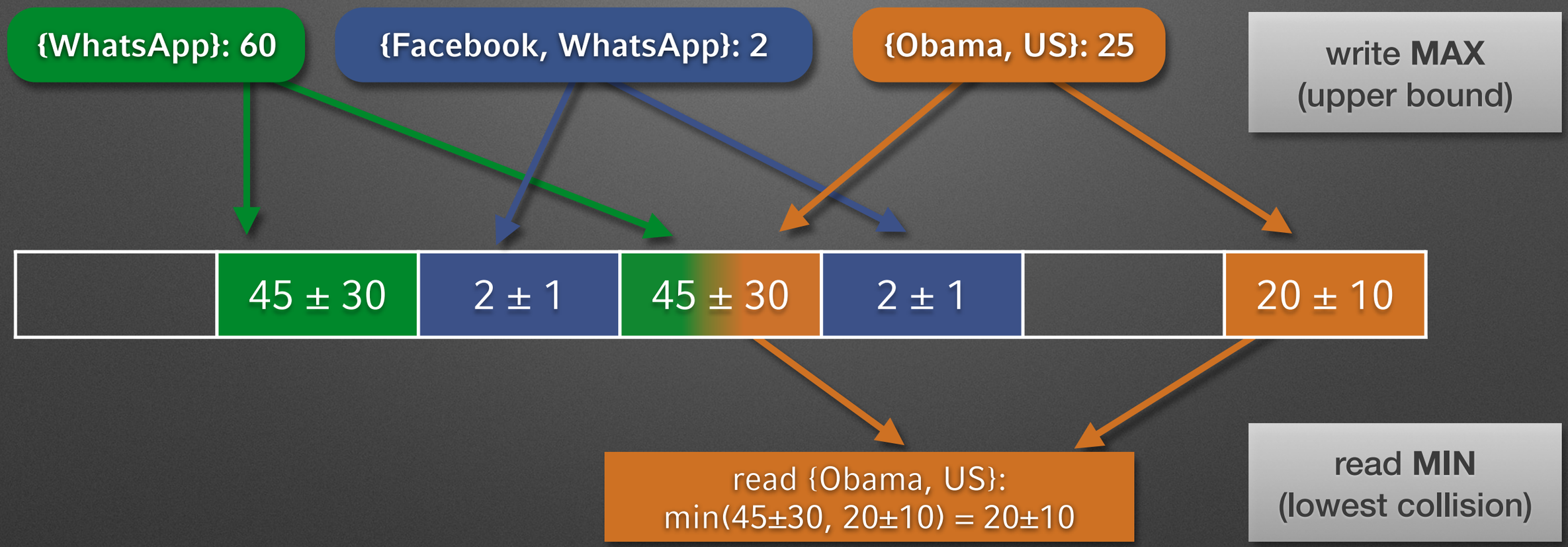
L=7 buckets, K=2 hash functions



Upper-bound estimate for mean and its variance

Hashing scheme for efficient tracking

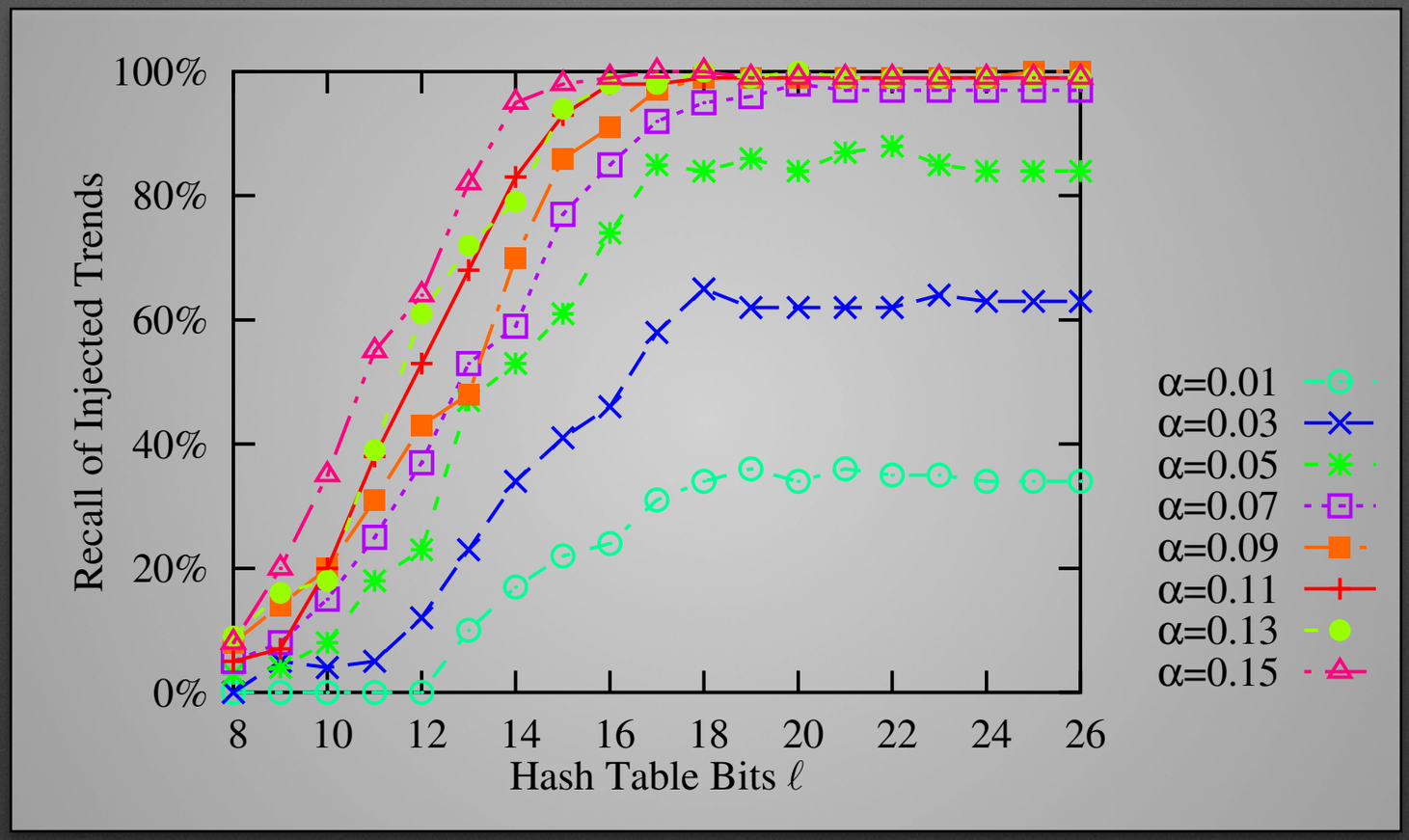
L=7 buckets, K=2 hash functions



Upper-bound estimate for **mean** and its **variance**
Performance on news dataset: **104s/day** with a **Raspberry-Pi**

Artificial trends evaluation

Inject artificial words with frequency α
e.g. “Obama meets <X123> Netanyahu”



Hash table size large enough → recall saturation

Refinement & clustering

- **Inverted index** (Apache Lucene) to verify **trend candidates** and measure **exactly (without hashing)** for precise reporting (false-positives can be eliminated)

- **Single Link** clustering with **Ward** of remaining trends (similarity matrix is built with the exact significance of all pairs)

- **Future work:** include topic modeling techniques (e.g. pLSI, LDA)

Thank You!

Questions?