Studying the Temporal Dynamics of Word Co-Occurrences: An Application to Event Detection

Daniel Preoţiuc-Pietro Srijith P.K., Mark Hepple,Trevor Cohn

LREC 2016

Positive Psychology Center
University of Pennsylvania

Penn | World Well-Being Project

27 May 2016

Word Co-occurrence

Discover events based on temporal text variation

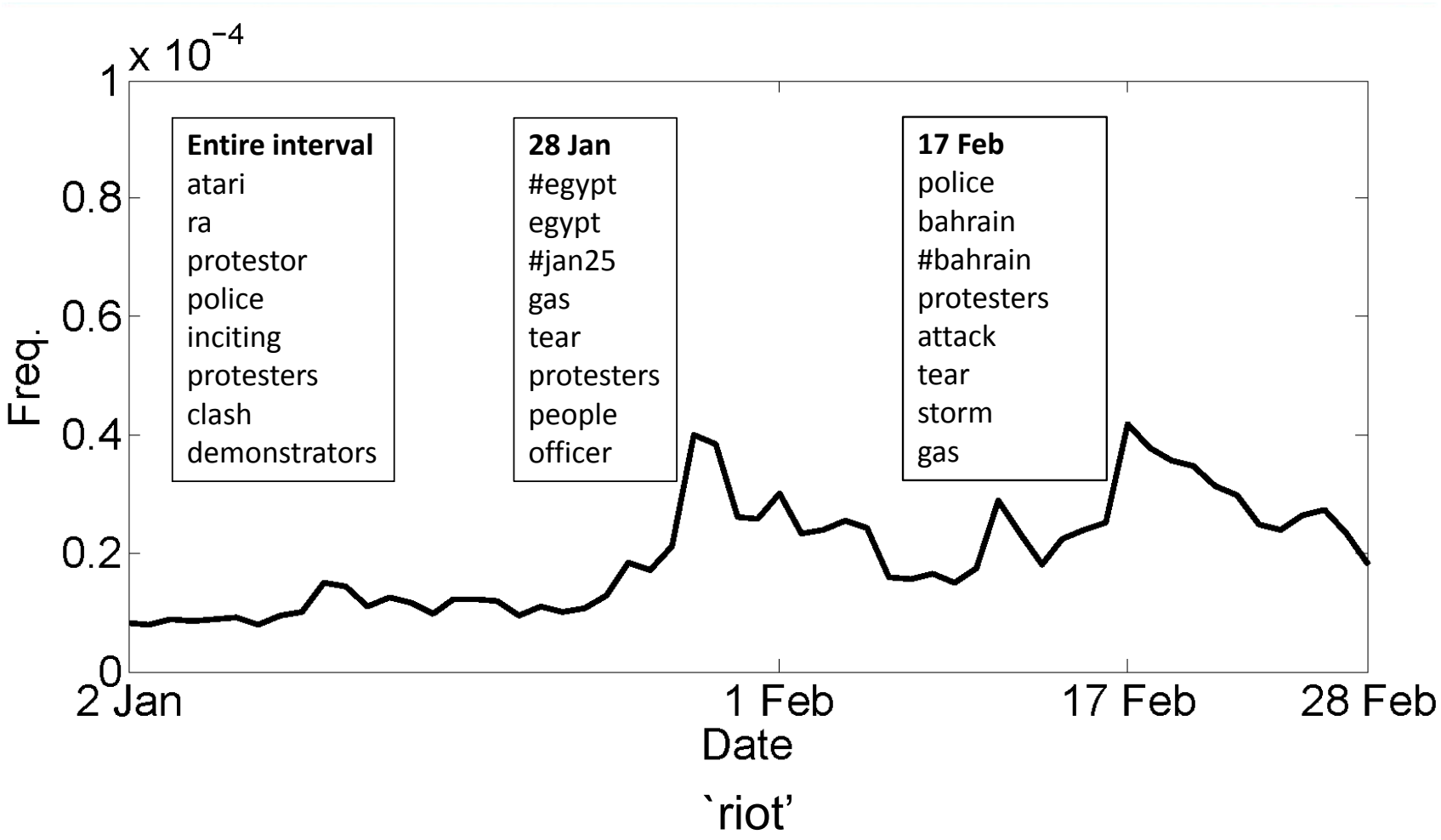
Word co-occurrence computed over large, static corpora (Wikipedia):

- similar concepts
- collocations

Computed over data from social media that reflects timely events (e.g. Twitter):

- current events & news

Example


x 10^−4
1
0.8
0.6
0.4
0.2
0
Freq.
Entire interval
atari
ra
protestor
police
inciting
protesters
clash
demonstrators
28 Jan
#egypt
egypt
#jan25
gas
tear
protesters
people
officer
17 Feb
police
bahrain
#bahrain
protesters
attack
tear
storm
gas
2 Jan
1 Feb
17 Feb
28 Feb
Date
`riot'

Event Detection

Two approaches:

1. Cluster messages
2. Cluster words, then retrieve messages

Event Detection

1. Cluster messages
   - Similarity based
   - classic Topic Detection and Tracking *(Allan 2002)*
   - e.g. cosine similarity, locality sensitive hashing *(Petrovic et al. 2011)*
2. Cluster words, then retrieve messages

Event Detection

1. Cluster messages
2. Cluster words, then retrieve messages
    - Based on time series analysis
    - e.g.: spike detection, dynamic time warping

Method

Cluster words (not messages) in a time interval

Fix time frame and compute word co-occurrence metrics:

$$\mathrm{NPMI}(X, Y) = -\log P(x, y) \cdot \log \frac{P(x, y)}{P(x) \cdot P(y)},$$

NPMI is bounded in the $[-1, 1]$ and has a easy interpretation:

- the maximum value (1) implies that both words appear exclusively together in the same tweet
- a value of 0 implies words appear independently of each other

Normalized PMI

Word1	Word2	NPMI	Type
arrests	yemen	0.699	news
publish	trailers	0.678	news
bestfriends	forming	0.678	news
g-slate	spotted	0.675	news
activist	arrests	0.674	news
china's	stealth	0.674	news
blake	griffin	0.672	proper name
magazines	merchandise	0.669	news
activist	yemen	0.667	news
actors	showcase	0.667	news
cameras	g-slate	0.664	news
angeles	los	0.662	proper name

Table: Top NPMI values for 23 Jan 2011, 9-10am. Word1, Word2 are in alphabetical order.

Clustering

We cluster words in a time frame to uncover terms that uniquely characterise an event.

We use spectral clustering *(Shi & Malik, 2000)* with NPMI as similarity measure.

Spectral clustering partitions a graph where edges are weighted using word co-occurrences.

We have to specify k – number of clusters.

Sample Events

Query: Kubica crash

Label: Formula 1 driver Robert Kubica injured in rally crash

Coherence: 0.47, Magnitude: 140

Date: 06 Feb 2011, 12-1pm

Sample Events

BBC | Sign in | News | Sport | Weather | Capital | TV | Radio | Mo

SPORT FORMULA 1

Sport Homepage
Formula 1
Results
Standings
Race Calendar
Drivers & Teams
Gossip
BBC F1 team
Circuit Guide

A-Z of Sports

Related BBC sites
News
Weather
Sport Relief

Page last updated at 00:00 GMT, Monday, 7 February 2011

E-mail this to a friend Printable version

Robert Kubica's F1 career at risk after rally crash

Kubica's car shows severity of crash

Robert Kubica's Formula 1 career is under threat after his right hand was partially severed in a high-speed rally crash in Italy on Sunday.

SEE AL
- Can K 07 Fe
- Rober 19 No
- Kubic 03 Fe
- F1 tes 03 Fe
- Brawn 03 Fe
- Thurs 03 Fe
- Alons 02 Fe
- Team 31 Ja
- Renau 14 No

Sample Events

Query: Moscow airport bombing

Label: Suicide bomber kills 35 at Moscow airport

Coherence: 0.37, Magnitude: 214

Date: 24 Jan 2011, 5-6pm

Sample Events

BBC | Sign in | News | Sport | Weather | Shop | Earth | Travel | Mo

NEWS

Home | Video | World | US & Canada | UK | Business | Tech | Science | Magazine | Ente

World | Africa | Asia | Australia | Europe | Latin America | Middle East

Moscow bombing: Carnage at Russia's Domodedovo airport

24 January 2011 | Europe

Sample Events

Query: Oprah Winfrey half-sister

Label: Oprah Winfrey has a half-sister

Coherence: 0.29, Magnitude: 43

Date: 24 Jan 2011, 9-10pm

Sample Events

BBC | Sign in | News | Sport | Weather | Shop | Earth | Travel

NEWS

Home | Video | World | US & Canada | UK | Business | Tech | Science | Magazine

Entertainment & Arts

Oprah Winfrey reveals half-sister on TV show

25 January 2011 | Entertainment & Arts

Oprah Winfrey has revealed she has a half-sister she never knew she had.

Winfrey introduced her sibling to viewers on her show, who was given up for adoption by Winfrey's mother nearly 50 years ago.

Patricia, 48, had been trying to trace her birth mother for years and eventually discovered her via a Wisconsin adoption

Winfrey's half-sister was only introduced as Patricia

Quantitative Analysis

First Story Detection Dataset:

2,228 annotated tweets – 27 events during June 2011 to September 2011 *(Petrovic et al. 2012)*

85,000 tweets from the same interval as background

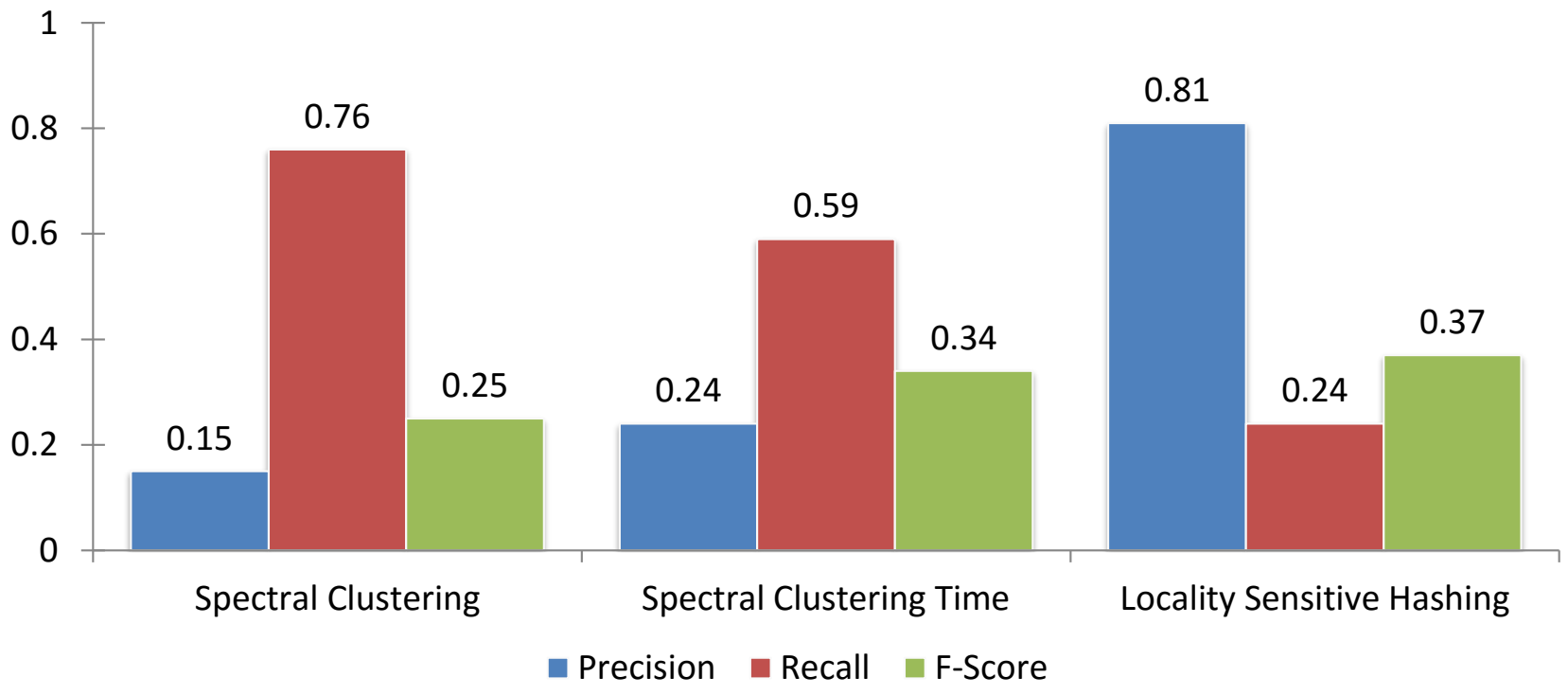
London Riots PHEME Dataset:

10,000 annotated tweets – 7 events during 6–16 August 2011

2,5M tweets as background

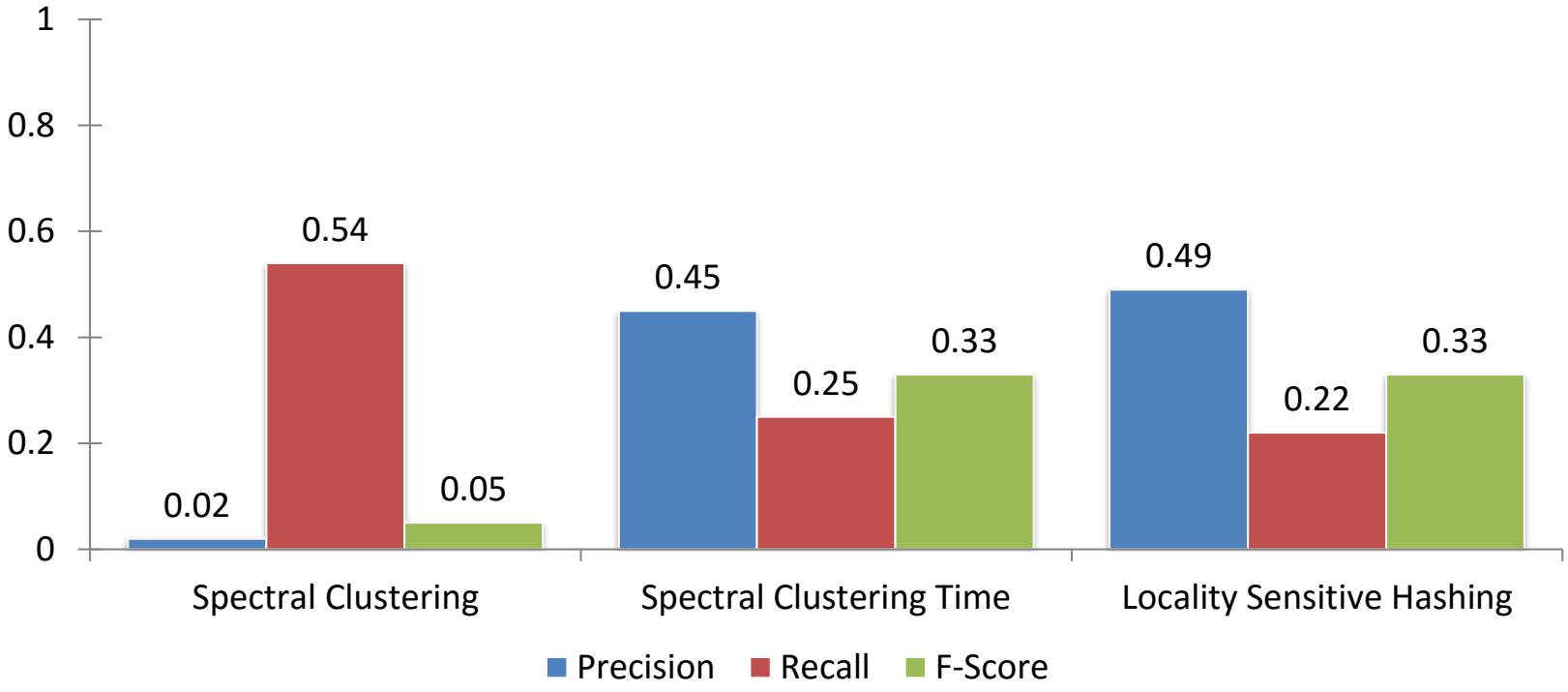
Events all about the London riots – more challenging as the events are related

Quantitative Analysis



First Story Detection Dataset

Quantitative Analysis



London Riots PHEME Dataset

Thank You!

Thank you!

Questions?

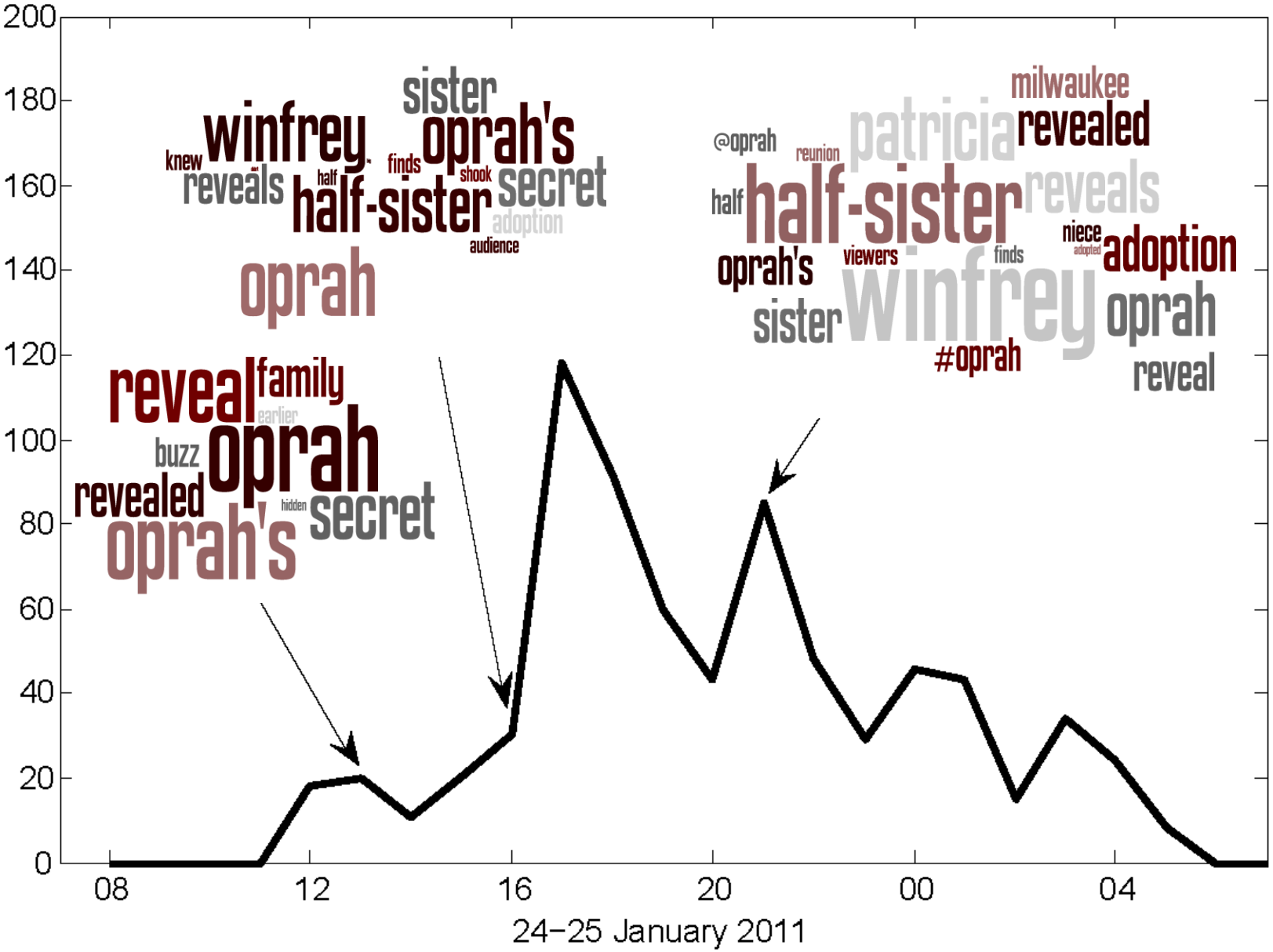
Longitudinal Analysis

Evolutionary spectral clustering methods

Create consistent clusters across consecutive time windows

Discover event evolution and content drift over time

Longitudinal Analysis



Take Aways

Social Media data is highly time dependent.

Words have different proprieties conditioned on time.

By modelling time, we gain a better understanding of real world effects.

Social Media can be used to uncover real world events.