

Automatic Discovery of Patterns in News Content



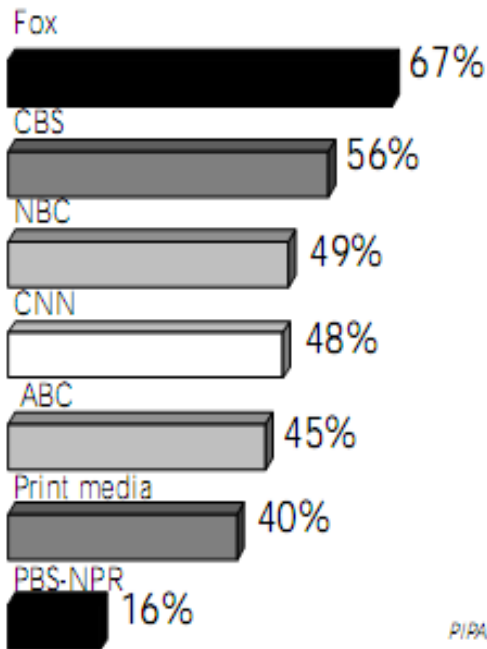
- **Nello Cristianini**
Intelligent Systems Lab
University of Bristol

An Interesting Fact

Evidence of Links Between Iraq and al-Qaeda

Is it your impression that the US has or has not found clear evidence in Iraq that Saddam Hussein was working closely with the al-Qaeda terrorist organization?

US has:

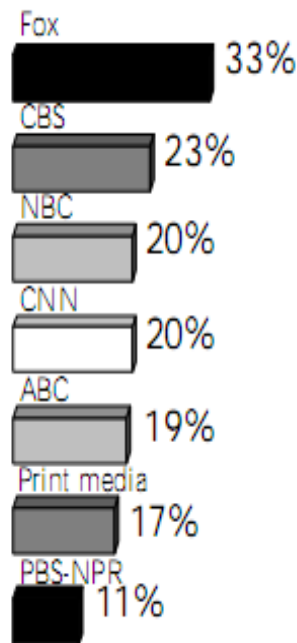


PIPA/KN 10/03

Weapons of Mass Destruction

Since the war with Iraq ended, is it your impression that the US has or has not found Iraqi weapons of mass destruction?

US has:

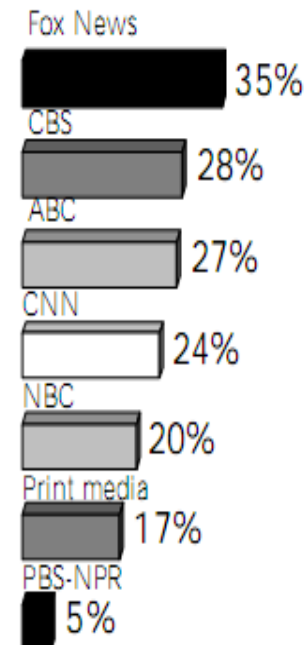


PIPA/KN 10/03

World Public Opinion

Thinking about how all the people in the world feel about the US having gone to war with Iraq, do you think:

The majority of people favor the US having gone to war



PIPA/KN 10/03

Feedback

- **FACT:** beliefs (not just opinions) of readers depend on the news content they choose to consume.
- e.g.: heavy consumers of news have unrealistic expectations about crime rates, usage of drugs, teen sex, (**over**estimated) and prevalence of ethnic minorities, older people, lower social classes (**under**estimate).

Cultivation Theory

- **Cultivation theory:** repeated exposure to a message shapes expectations, beliefs.

→ STUDY BIAS IN NEWS CONTENT

- Social scientists study how news are chosen, presented, narrated... as biases in news content can **affect (as well as reflect) biases** in society.

News Content Analysis

- **Study how news are chosen, presented, narrated... and how this affects (as well as reflects) biases in society / public opinion.**
- This is done “by hand”, on small numbers of news outlets, for short periods, for pre-specified questions (“coding approach”).

News Coding



Appendix 1 Coding scheme: MDHH television broadcasts (translated from Dutch)

1. Date:
2. Channel:
3. Title of the item:
4. Starting time (full four digits):
5. Ending time (full four digits):
6. Length (in minutes):
7. Genre:
 1. News
 2. Documentary
 3. Drama
 4. Comedy
 5. Ceremony
 6. Discussion panel
 7. Artistic performance
 8. Other
8. Producer:
 1. Israeli
 2. Non-Israeli
 3. Co-production
9. Is the item aired only/mostly on Holocaust Memorial Day?
 1. Yes
 2. No—airs on other days
 3. Other
10. Does the item address the Holocaust?
 1. Yes
 2. No
 3. Other



The **MediaPatterns** Project

Involved about 10 people
(+ extra social scientists) over 5 years

Goals:

- to automate the analysis of news content,
- to understand the workings of the media system,
- to understand how **science can be automated**,
- to operate with challenging patterns on large datasets..
- to enjoy creating a large-scale infrastructure

MediaPatterns.enm.bris.ac.uk

The Problem with Large Projects...

Cannot cover all aspects



Getting the Data

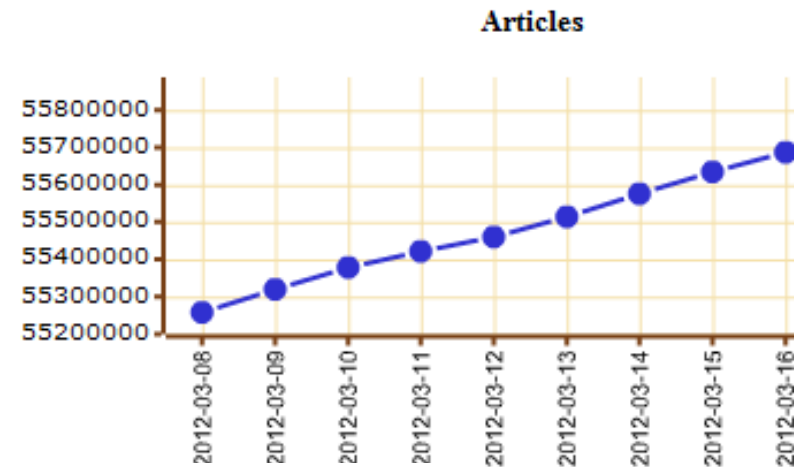
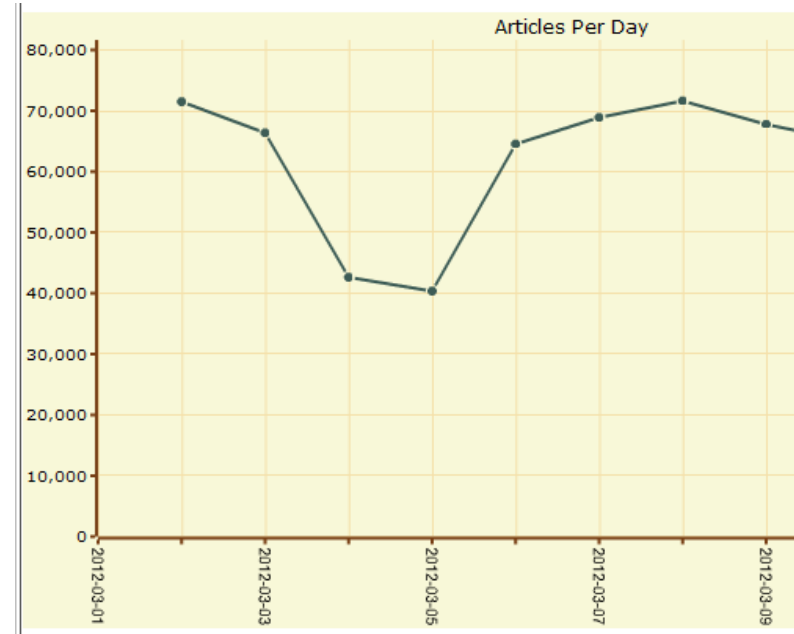


The NOAM infrastructure

- We developed a dedicated infrastructure for **News Outlet Analysis and Monitoring**
- To gather and annotate data about:
 - News outlets
 - News items
 - News stories
 - Named entities

The Data

- We gather about 60K news items per day, from > 1000 outlets , 4400 rss feeds, in 22 languages , from 195 countries (machine translated into English if necessary)
- We have analysed ~55 million news items



Analysis of the Data

- We are interested in macroscopic patterns found in the global news-system contents.
- What kind of stories / people / topics make news?
- What do editors want? What do readers want?
- What patterns in style and narrative can be found?
- Can we measure how people are affected?

Our Questions

1. What is in the news ?
 1. What people ?
 2. Which stories are covered by whom?
2. What do readers want?
3. Any patterns in style?
And narrative?
4. Can we measure public mood?

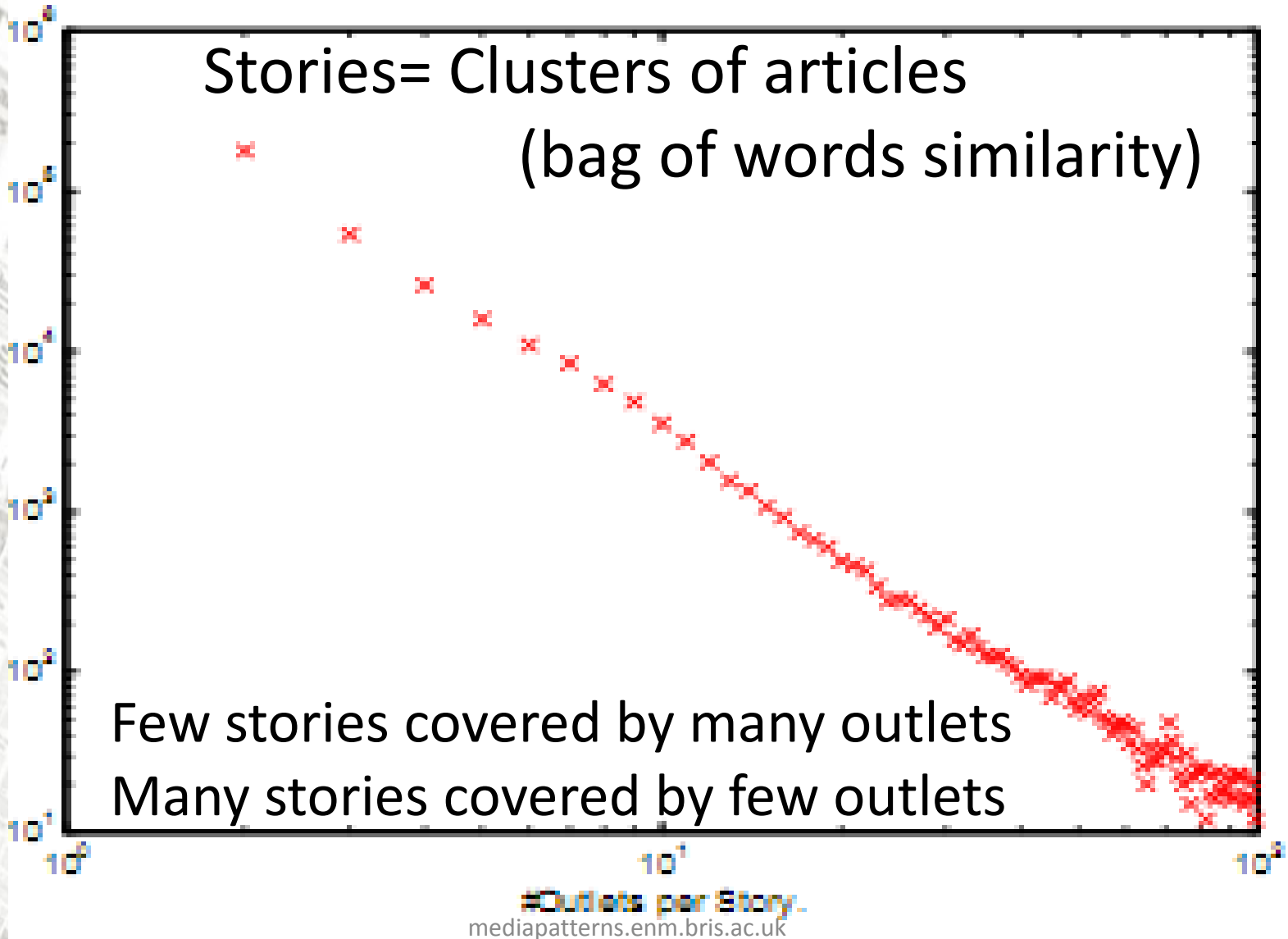


Question 1: what is in the news?

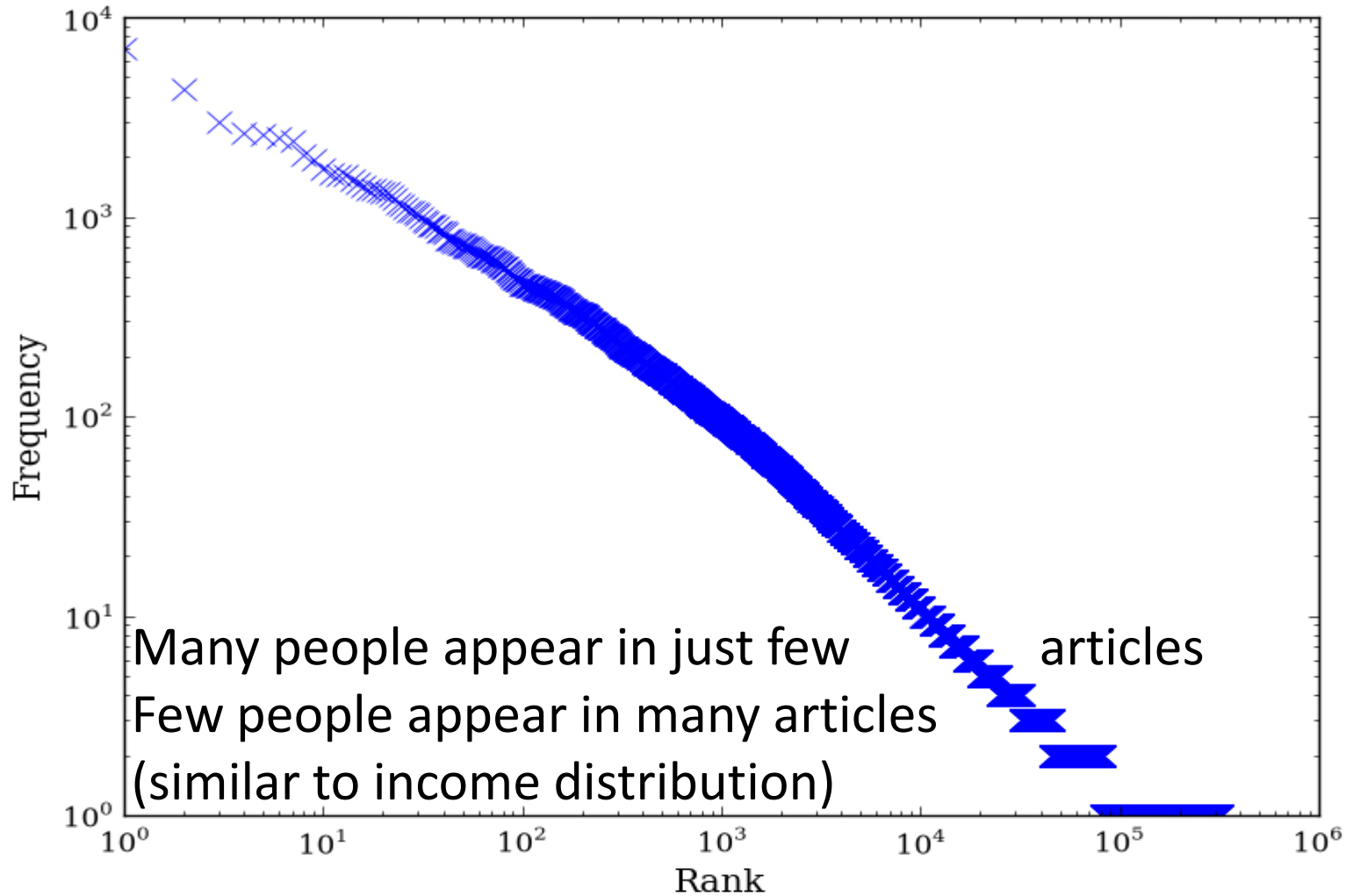
- There are stories about people.
But: which stories and which people make it?
- Short answer:
the same few stories and the same few people occupy the most “real estate”
- A power law...

[skipping method: stories are clusters of articles; entities are extracted, disambiguated, and their properties computed based on large data sets]

From Articles to Stories



People in the News



MediaPatterns

More in detail...

- Which kinds of people are present in which kinds of stories?
- What determines which stories are covered and which ones are neglected in a given outlet ? And in general ?

People in the News



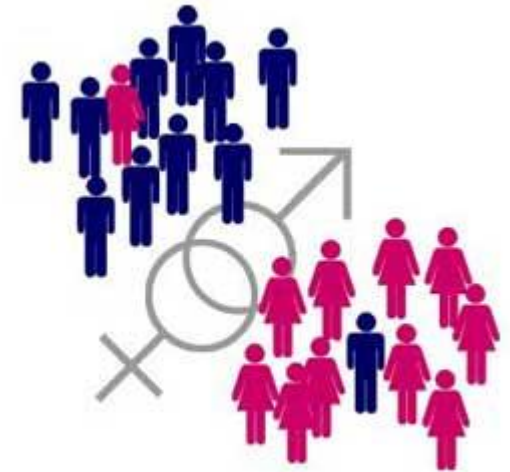
M:F Ratio

Consider the Top-100 richest people in the world.
90 are men, 10 are women.

We call this the M/F ratio.

The M/F ratio varies with domain:

- Of the top 50 richest athletes, all 50 are male.
- Of the top 100 celebrities, 35 are female.
- Of the top -10 fashion models all 10 are female.



Gender Bias in the Media

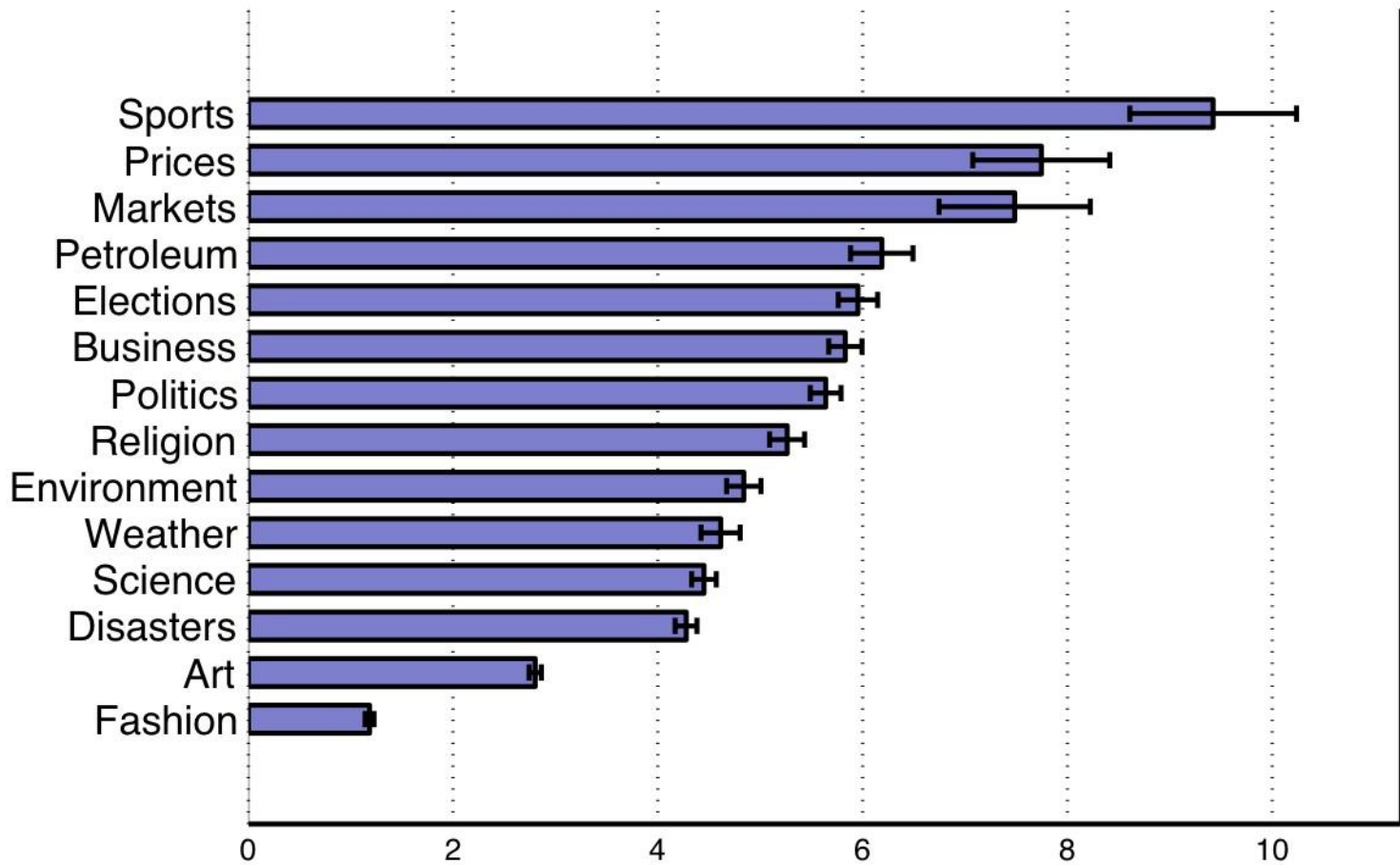
- What about media attention?
- Of the Top-1000 most mentioned people, how many are male?
- How does this change by topic?
- This involved analysing 476,528 articles in English language, and detecting their topic, as well as people and their gender.

Detecting Topics

- Support Vector Machines trained on Reuters and New York Times tags as ground truth
- High precision requested for tags to be applied (so: many articles left untagged)

	Topic
1	SPORTS
2	MARKETS
3	FASHION
4	DISASTERS
5	ART
6	BUSINESS
7	INFLATION-PRICES
8	RELIGION
9	POLITICS
10	SCIENCE
11	WEATHER
12	PETROLEUM
13	ELECTIONS
14	ENVIRONMENT

M:F by Topic



Flaounas, I., Ali, O., Bie, T.D., Mosdell, N., Lewis, J., Cristianini, N.: Massive-scale automated analysis of news-content: Topics, style and gender. Submitted for Publication (2011)

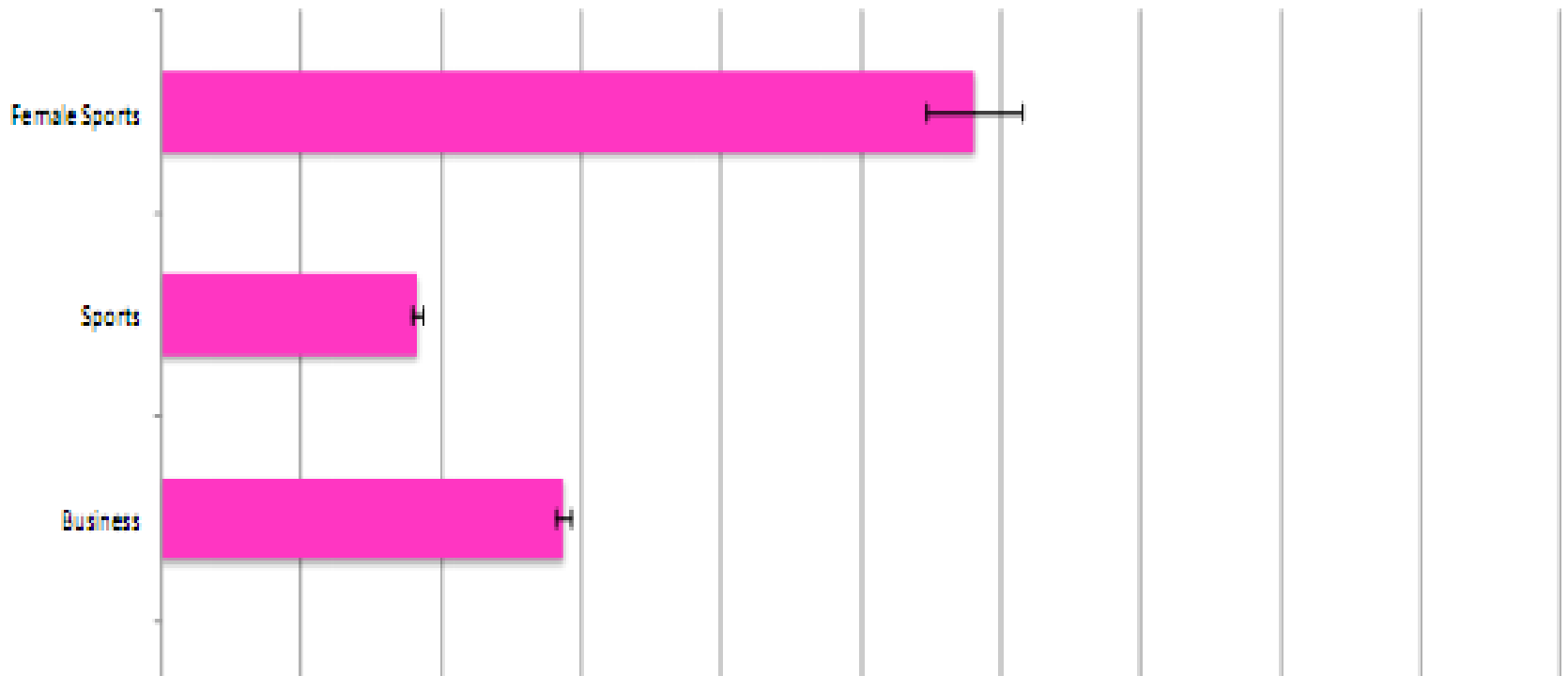
Validation

As validation we added the topic “female sports” where we expected more female names than in general sports.

While this was observed, the bias was still in favour of males.

The general pattern observed for income was also observed for media attention:

MF Sport > MF General > MF Fashion



Observations

- Q: How does this shape beliefs, attitudes and opinions about gender?



Which stories
are covered
by which
outlets?



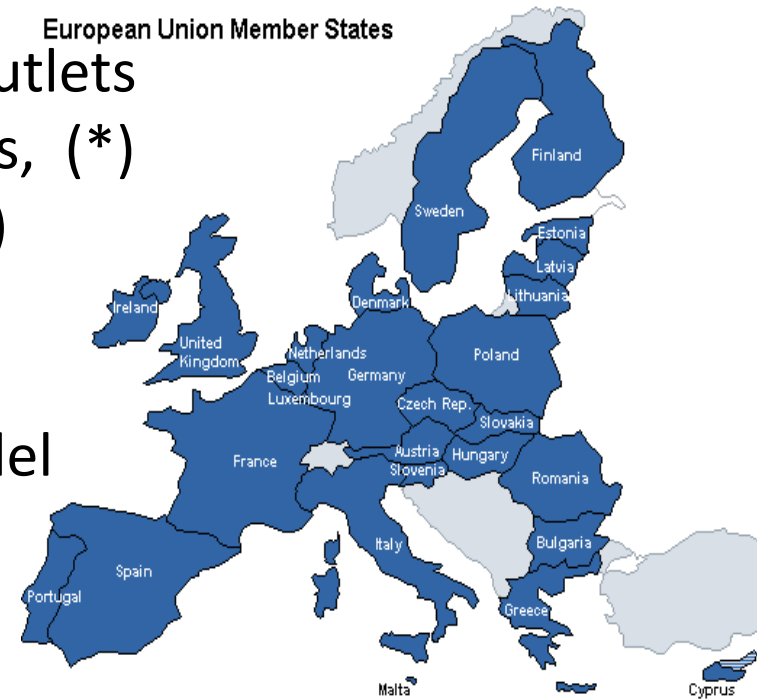
Mapping the EU Mediasphere

We machine translated the top EU outlets into English (Moses + our innovations, (*) trained on europarl + other corpora)

Details skipped but fun:
eg we re-create a new language model every day automatically

Clustered all articles into stories (bag of words)

Question: **which outlets tended to carry the same stories?**



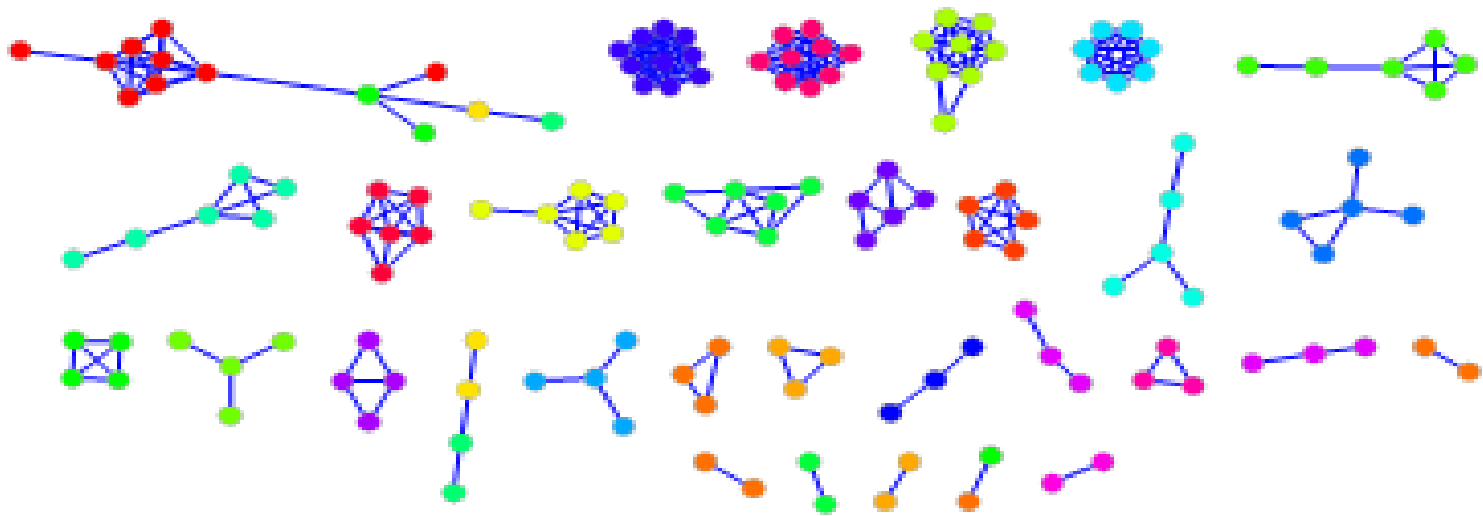
The Data

- Top-10 media outlets per country
- over the 27 EU countries
- in 22 different languages
- for a 6 months period
- A total of 1.3M news items.



Outlets Covering Same Stories

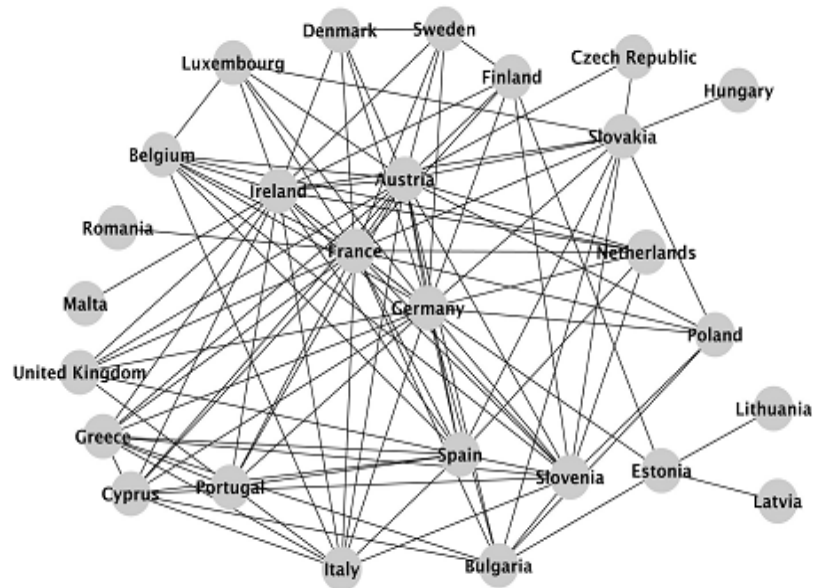
Link outlets if they share more stories than expected by chance (chi-square scores).



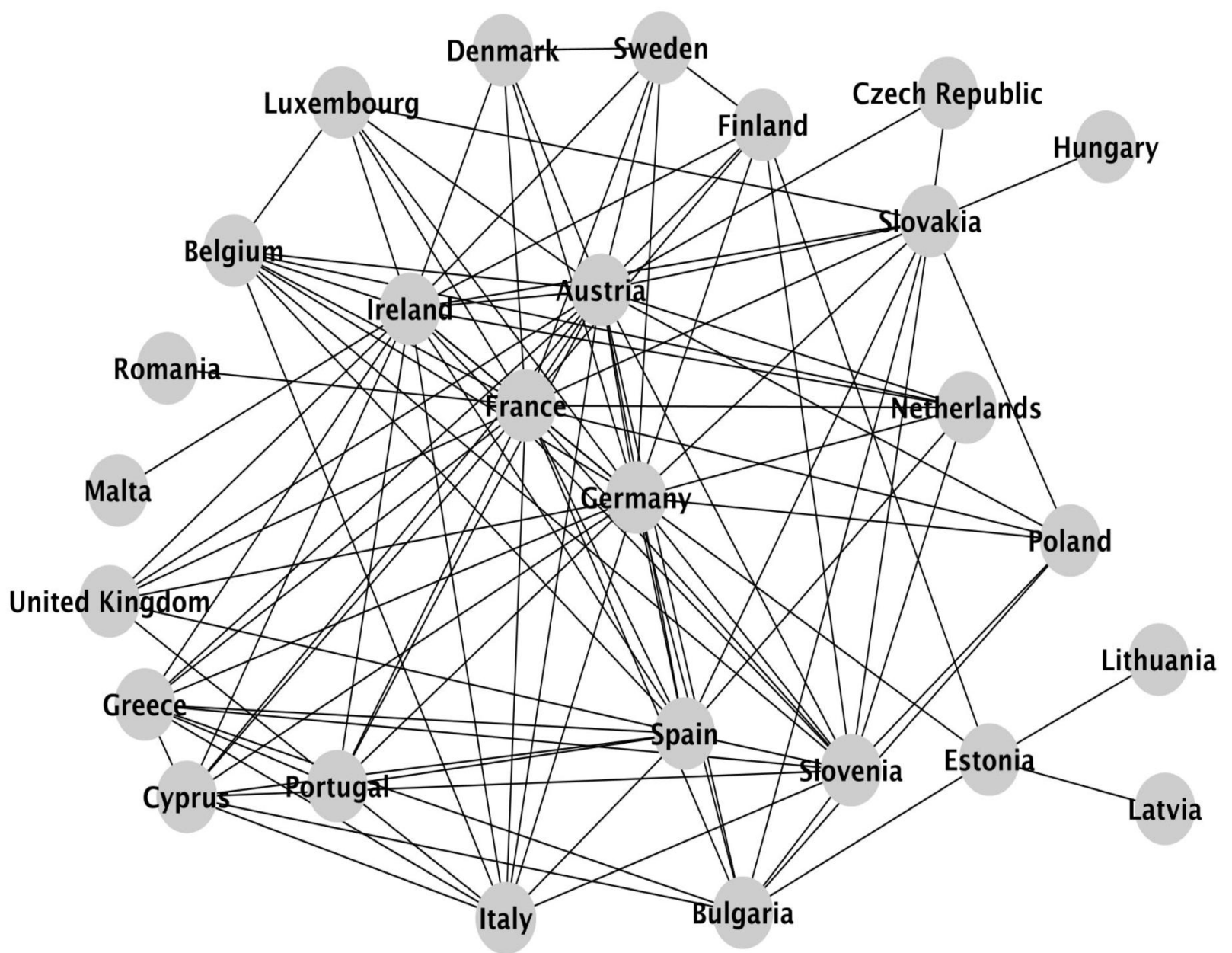
The probability of two non-singleton nodes from the same country to end up in the same connected component is **82.9%** ($p < 0.001$).

Linking Countries

- Since countries essentially match the communities, we generate a network of countries



We go as sparse as possible while keeping the network connected.



Flaounas, I., Turchi, M., Ali, O., Fyson, N., Bie, T.D., Mosdell, N., Lewis, J., Cristianini, N.: The structure of eu mediasphere. PLoS ONE p. e14243 (2010)
mediapatterns.enm.bris.ac.uk

Explaining the Relations

- Thousands of different editors make their choices independently every morning based on their own different goals.... yet ...

We found significant ($p < 0.001$) correlation of countries' *media-content similarity* to their:

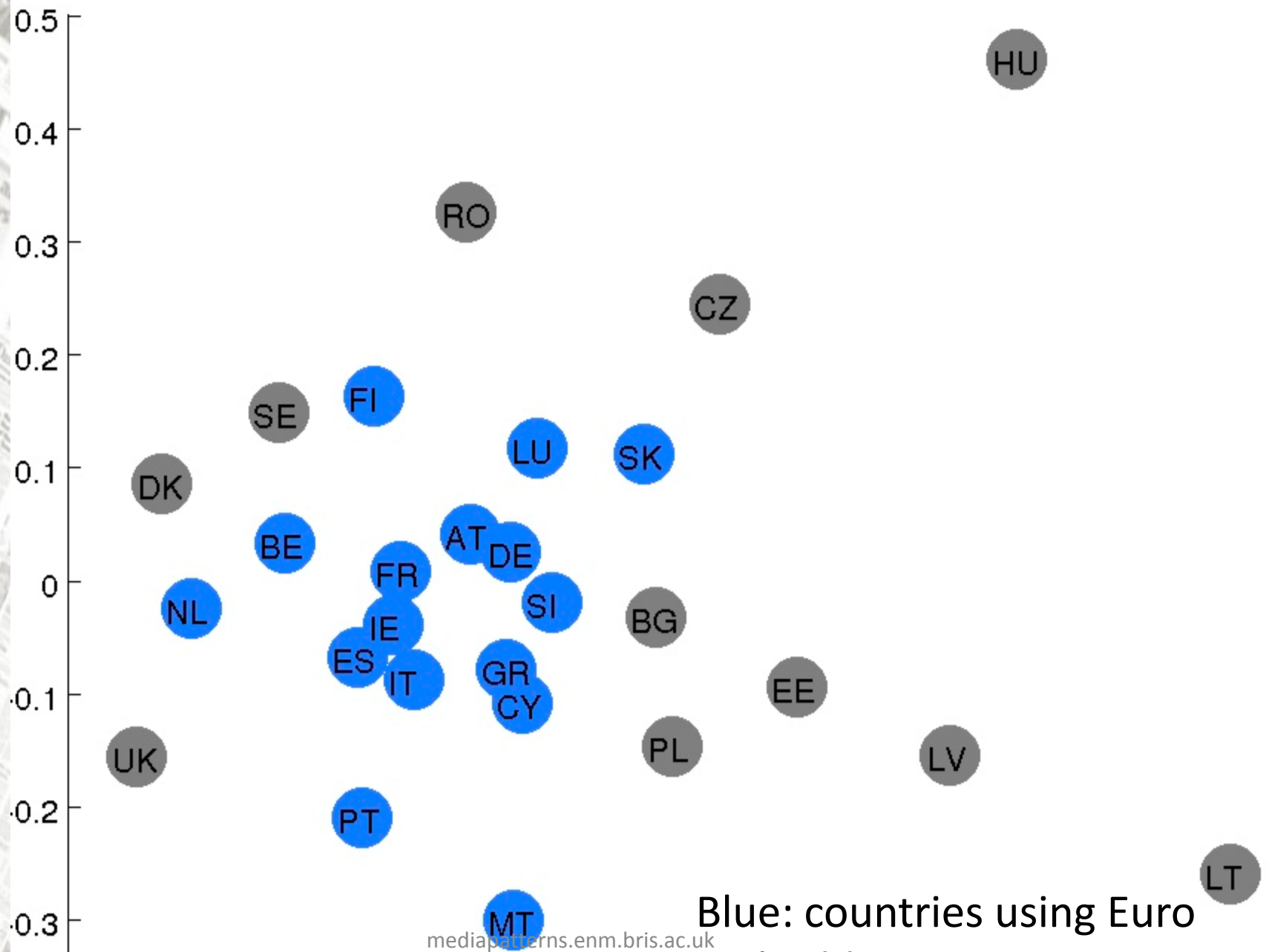
Geographical proximity — based on sharing of borders **33.86%**

Economical proximity — based on trade volume **31.03%**

Cultural proximity — based on song contest voting patterns **32.05%**



MediaPatterns



Question 2

What
Readers
Want



What Readers Want

- Can we predict the preferences of readers?
- Data is available through RSS feeds...
- How to exploit it?

Most Popular

Shared

Read

Video/Audio

Milly's mother collapses in court

1

Sex, religion and gossip fuels superbrands

2

'Exploding' watermelons in China

3

Queen begins first Ireland visit

4

The parental spending craze

5

Players held over drug possession

6

Is it fair to fine fat people for not dieting?

7

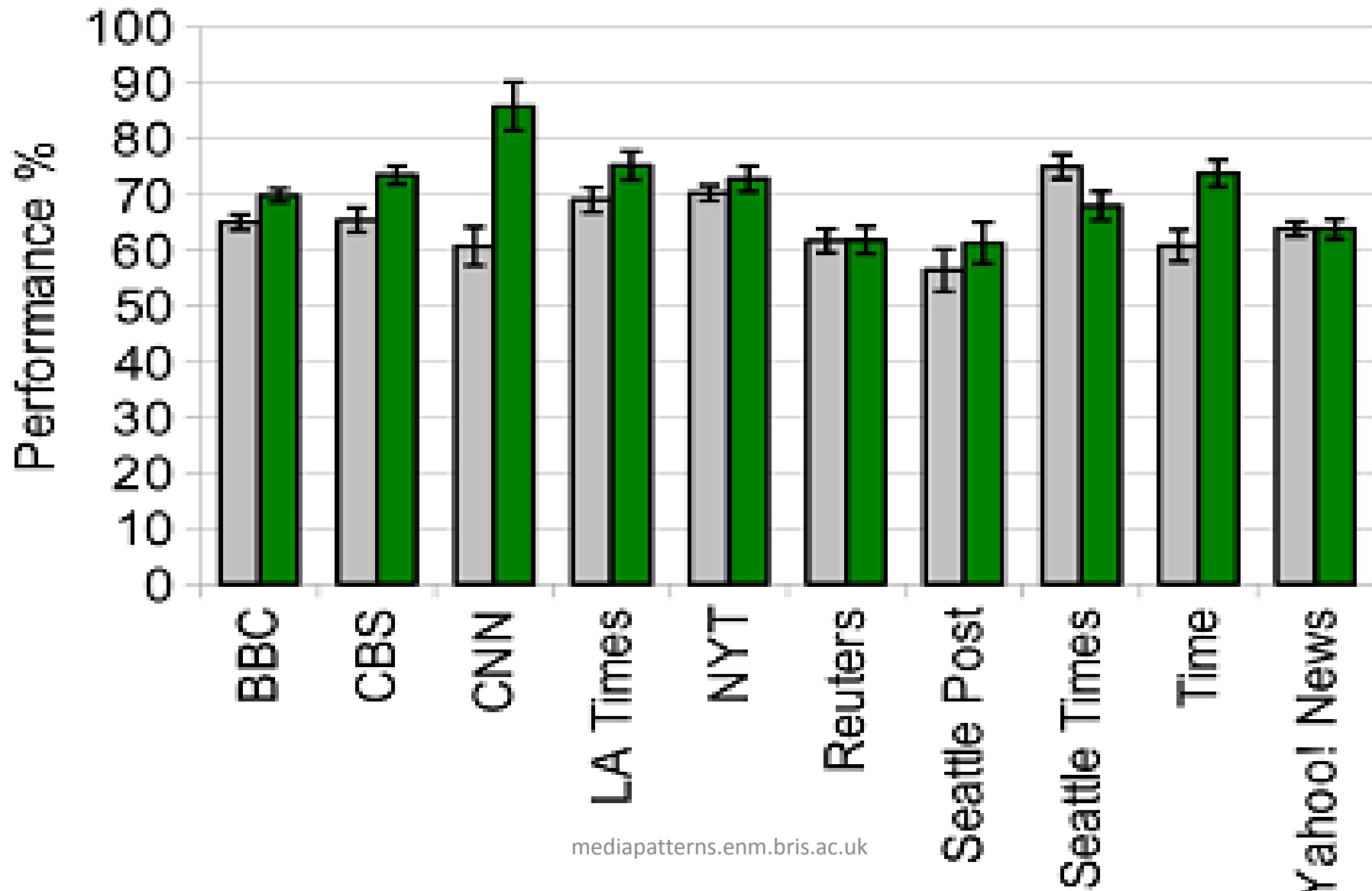
Israeli parents named baby 'Like'

8

What Readers Want

- Even just using very basic data (textual content of title and snippet) it is possible to predict the preferences of readers...
- Not a simple classification task, needs to be framed as a “learning to rank” task (competitive nature of the process)
- Linear scoring function inferred from data, with SVMs, and used to choose which of 2 articles is likely to be preferred, by average reader...

Ranking SVM



Average relative distances between outlets

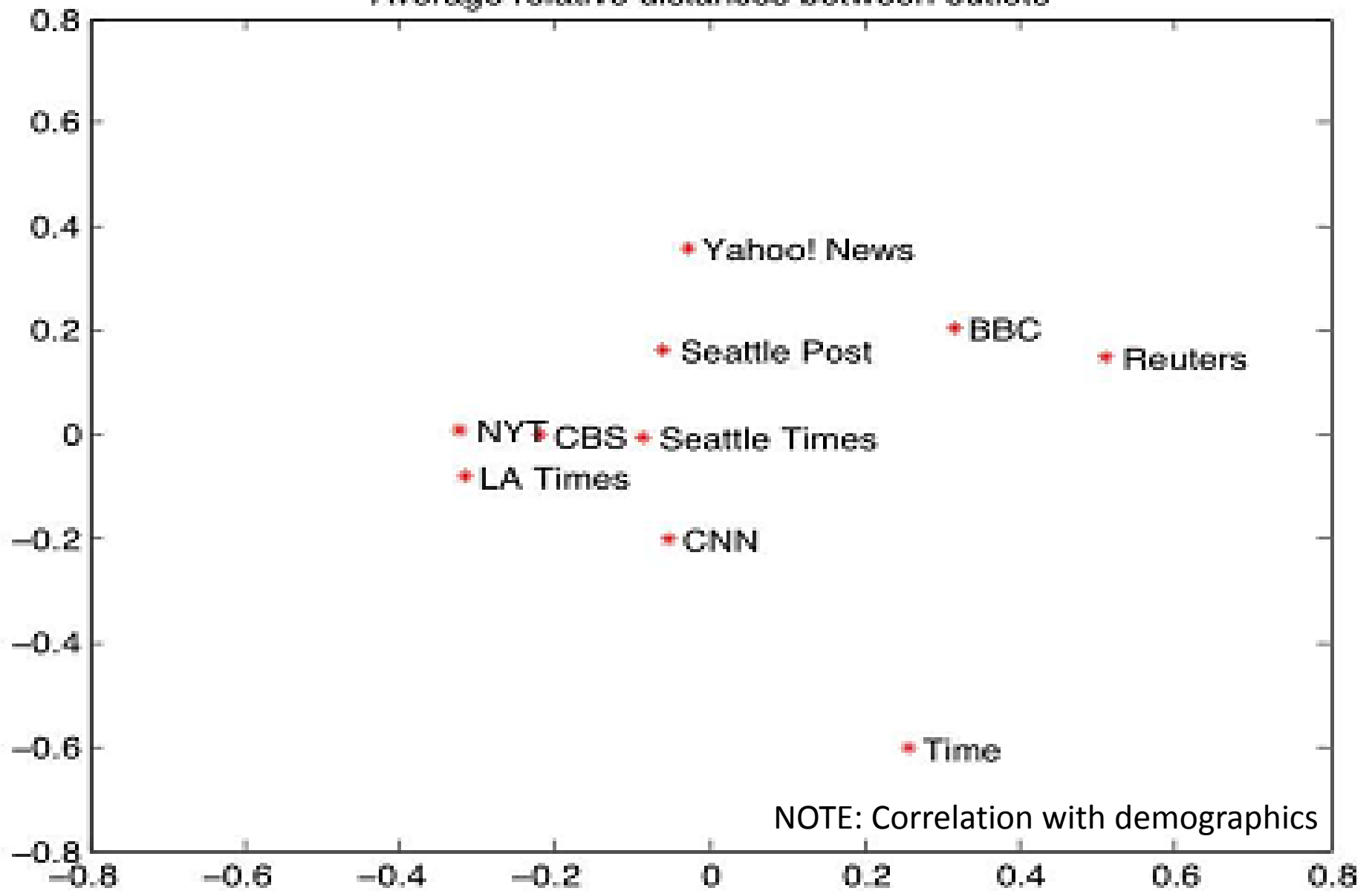
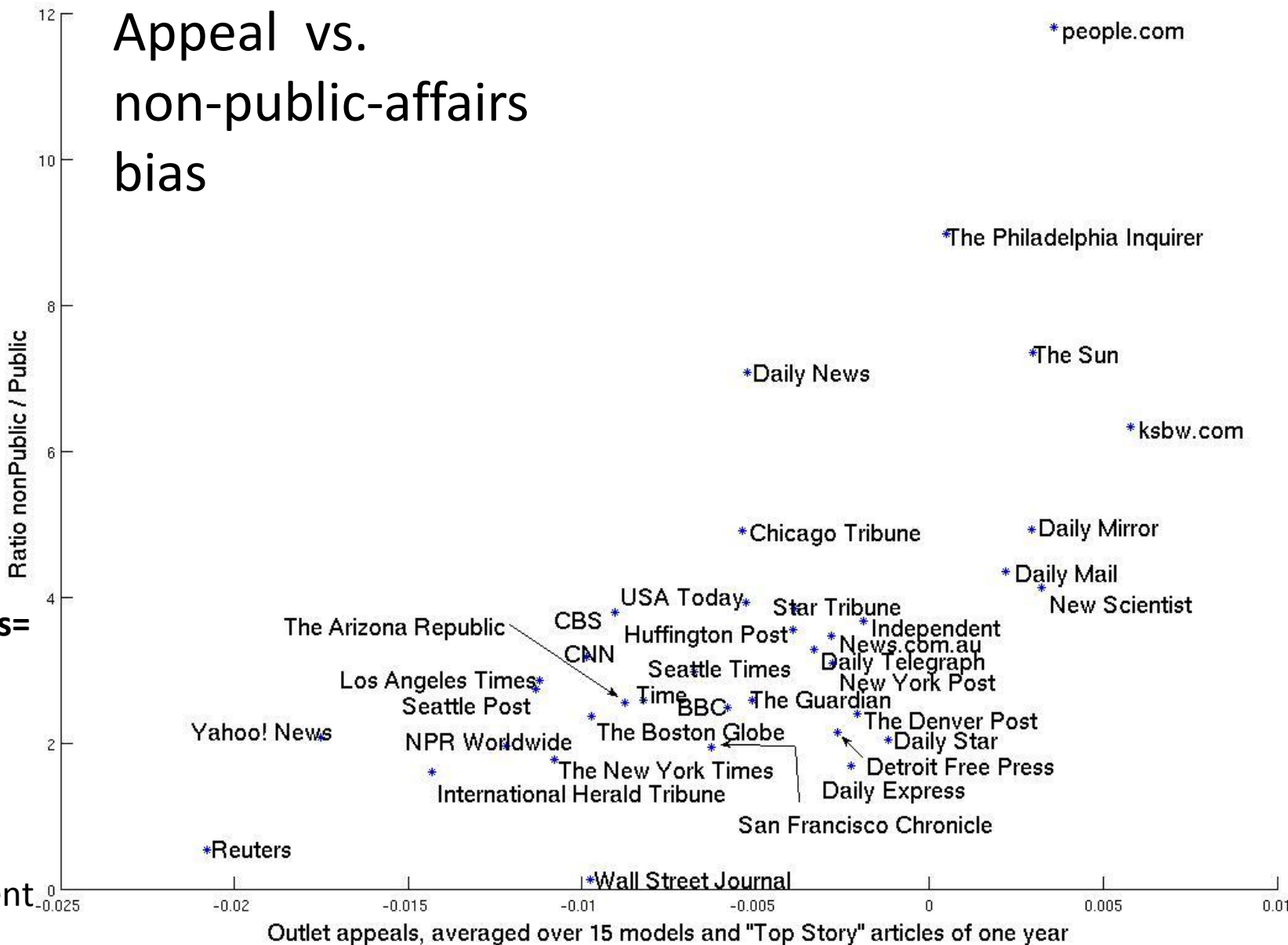


Fig. 3 Relative distances of outlets, plotted in the first two dimensions of multidimensional scaling.

Appeal vs. non-public-affairs bias



Public affairs=
politics
economics
finance...

Non-PA=
entertainment
sport
...

Question 3

Writing Style and Narrative Patterns



Writing Style

Interested in large scale patterns involving writing style:

- Readability

- Language subjectivity

- The first is captured with a standard measure (FRET) that has been shown to correlate well with ease of comprehension.

- The second quantity was captured by detecting the adjectives in the text and measuring their “polarity”.

Readability

Flesch Reading Ease Test

The higher the FRET the easier the text to read.

Scores range from 0-100.

10K random news items per topic

$$\text{FRET (article)} = 206.835 - (1.015 \cdot \text{ASL}) - 84.6 \cdot \text{ASW}$$

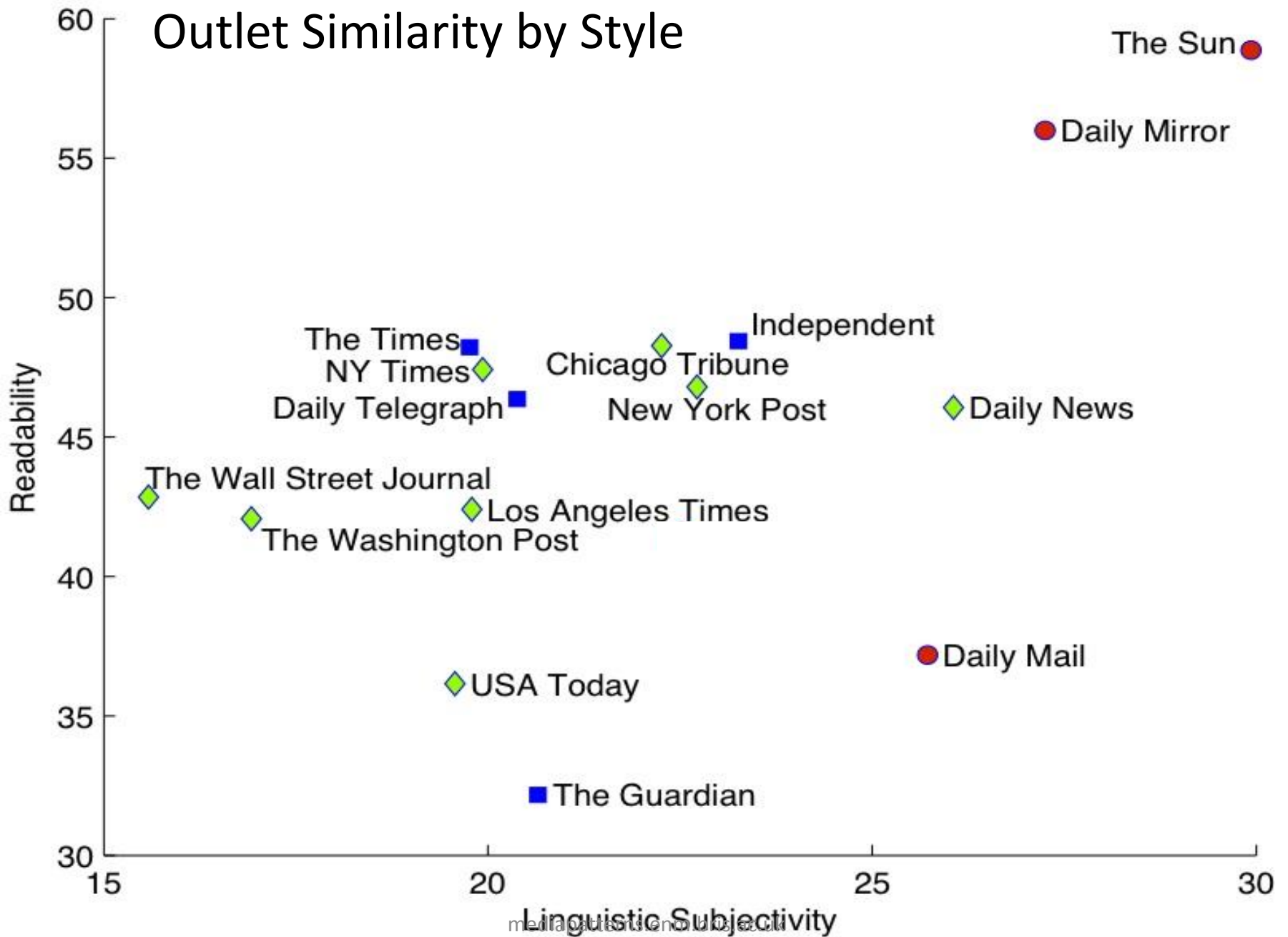
ASL=average sentence length

ASW = average syllables per word

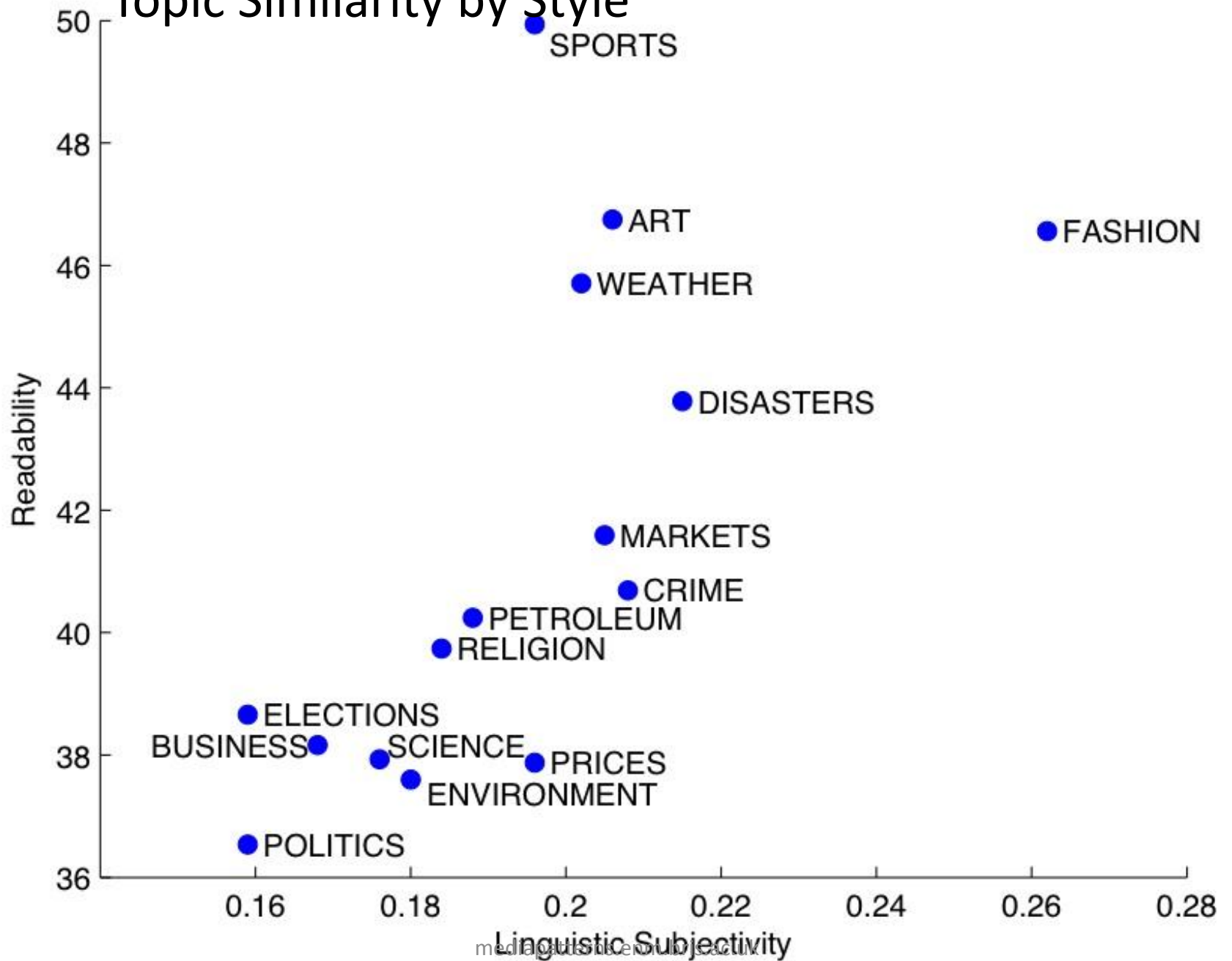
Linguistic Subjectivity

- We measure the percentage of sentimental adjectives over the total number of adjectives.
- Adjectives detection by Stanford POS tagger.
- We check for each adjective the presence of a SentiWordnet sentimental score >0.25 (percentage of adjectives that have either positive or negative sentiment score $>25\%$)

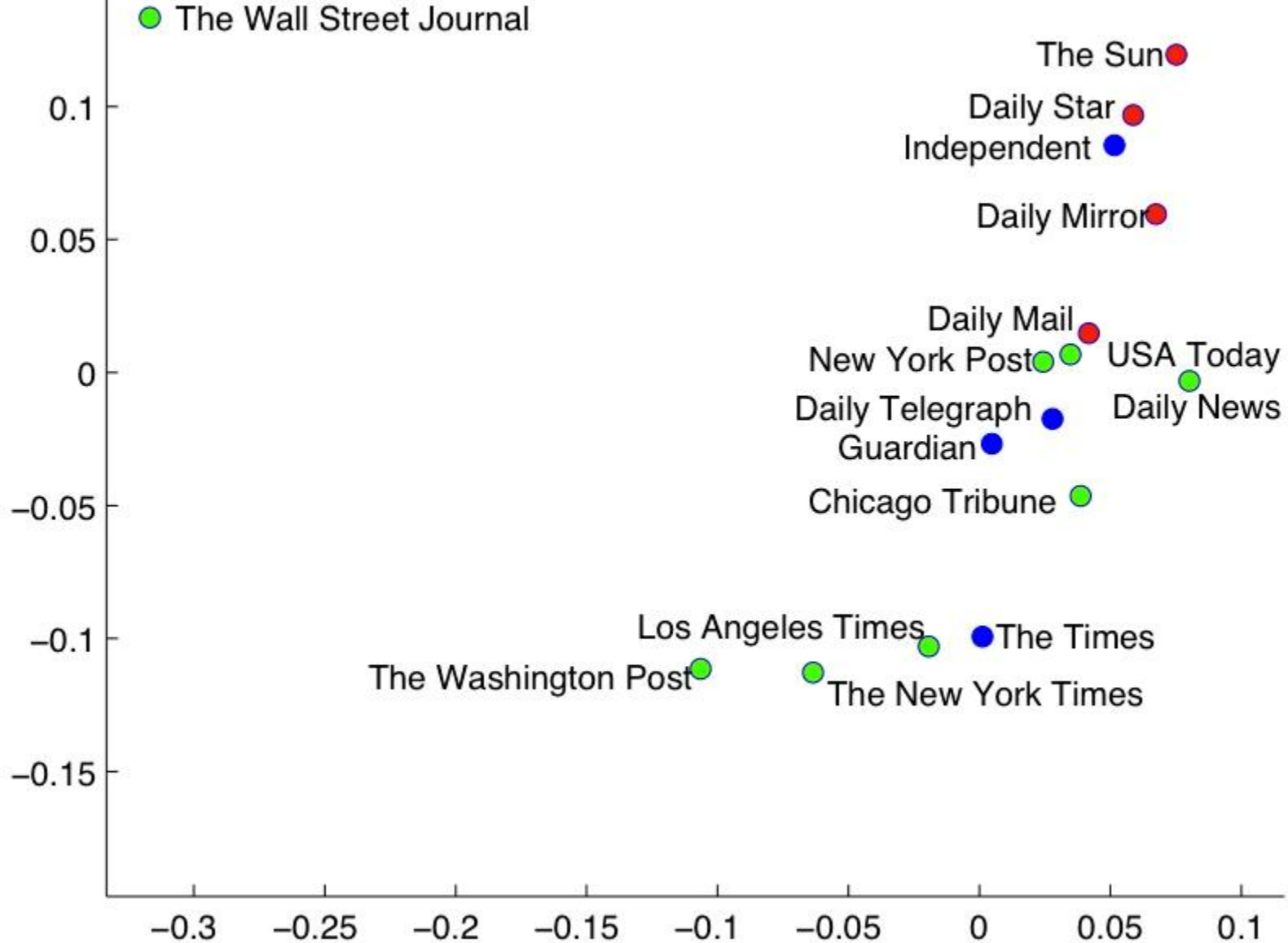
Outlet Similarity by Style



Topic Similarity by Style

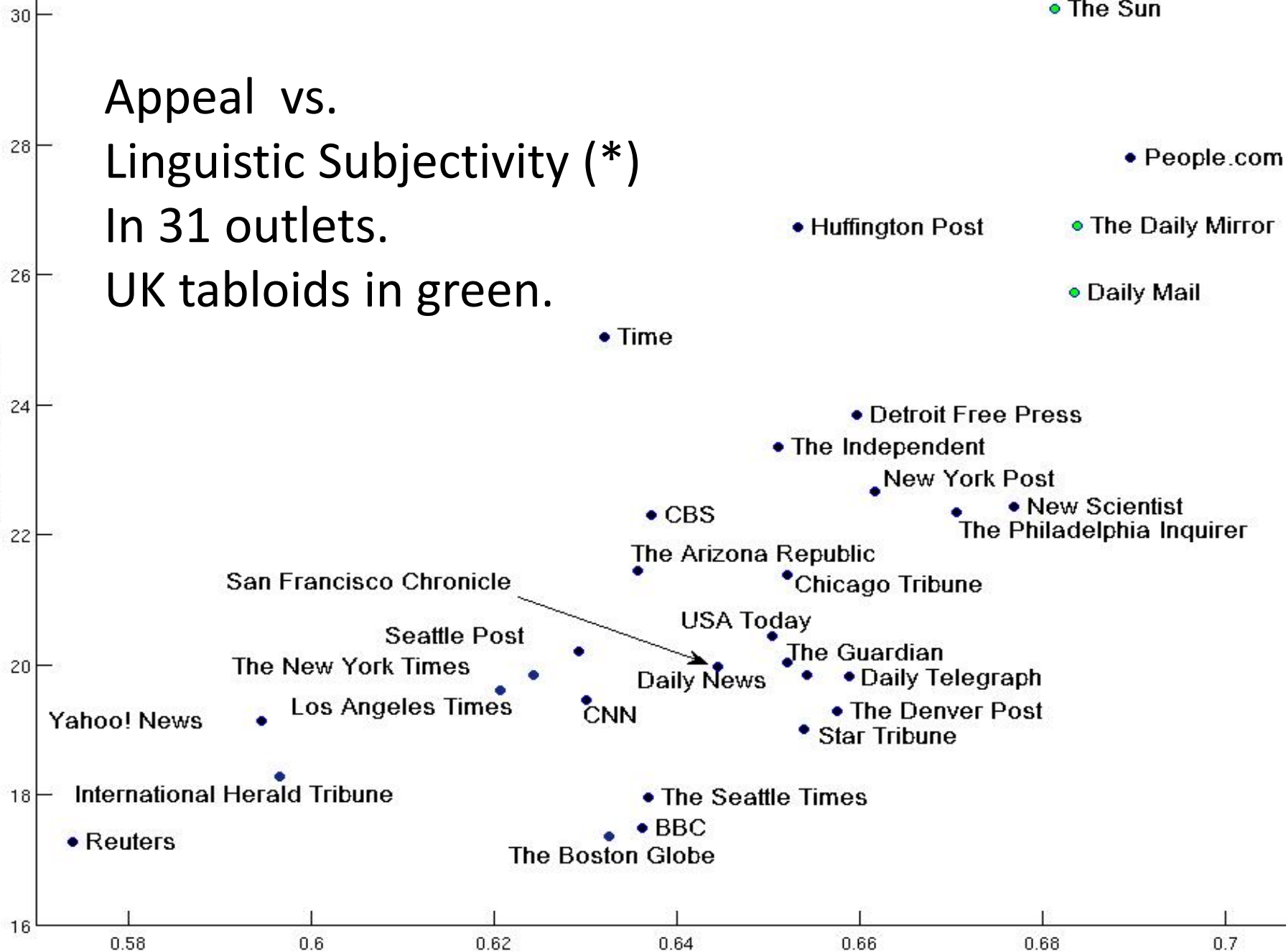


Outlet Similarity by Topic Distribution



Appeal vs.
Linguistic Subjectivity (*)
In 31 outlets.
UK tabloids in green.

Linguistic subjectivity



Validations

- **CBC Newsround is the most easily readable outlet.**
than For validation reasons we added a set of articles from the BBC show CBBC-Newsround, which is a current affairs programme written specifically for children. As expected the CBBC news were found to be the most readable collection of articles with a mean readability score of 62.50 (S.E.M. = 0.27).
- **Op/Ed Articles are more subjective than average.**
We collected 5766 **Op/Ed articles** of that kind in our period of study from 57 different media and we found that their linguistic subjectivity has a mean of 26.15% (S.D.=0.29%) while the mean subjectivity of main articles is 19.45% (S.D. = 0.22%).
- Note: we also found a 72.5% correlation between readability and subjectivity (Spearman correlation, $p=0.003$).

The background of the slide features a collage of newspaper clippings. Visible text includes 'EL PUNTI', 'Vistes es doctors mundial amb 16 an...', 'Trouble the Dia', and 'me'.

Relation: style vs. demographics

- For UK newspapers we obtained reader-demographics: [gender, age group, income group] and we computed similarity between outlets.
- (from: Newspaper Marketing Agency at: www.nmauk.co.uk)
- **Writing Style Correlates with Reader Demographics - 31.43%**
(Kendall correlation of the pairwise euclidean distances between outlets, **p=0.048**)

NARRATIVE ANALYSIS

- Social scientists like to think in terms of social actors, their actions, and the narrative linking them.
- Roberto Franzosi of Emory University developed Quantitative Narrative Analysis (QNA) to identify key actor / action patterns in a set of articles.
- Actors and actions (and sphere of action) are hand annotated...
- We identified ACTORS with NOUN PHRASES, and ACTIONS with VERBS, and SVO triplets with narrative units

NY Times Corpus, Year 2002 – crime stories

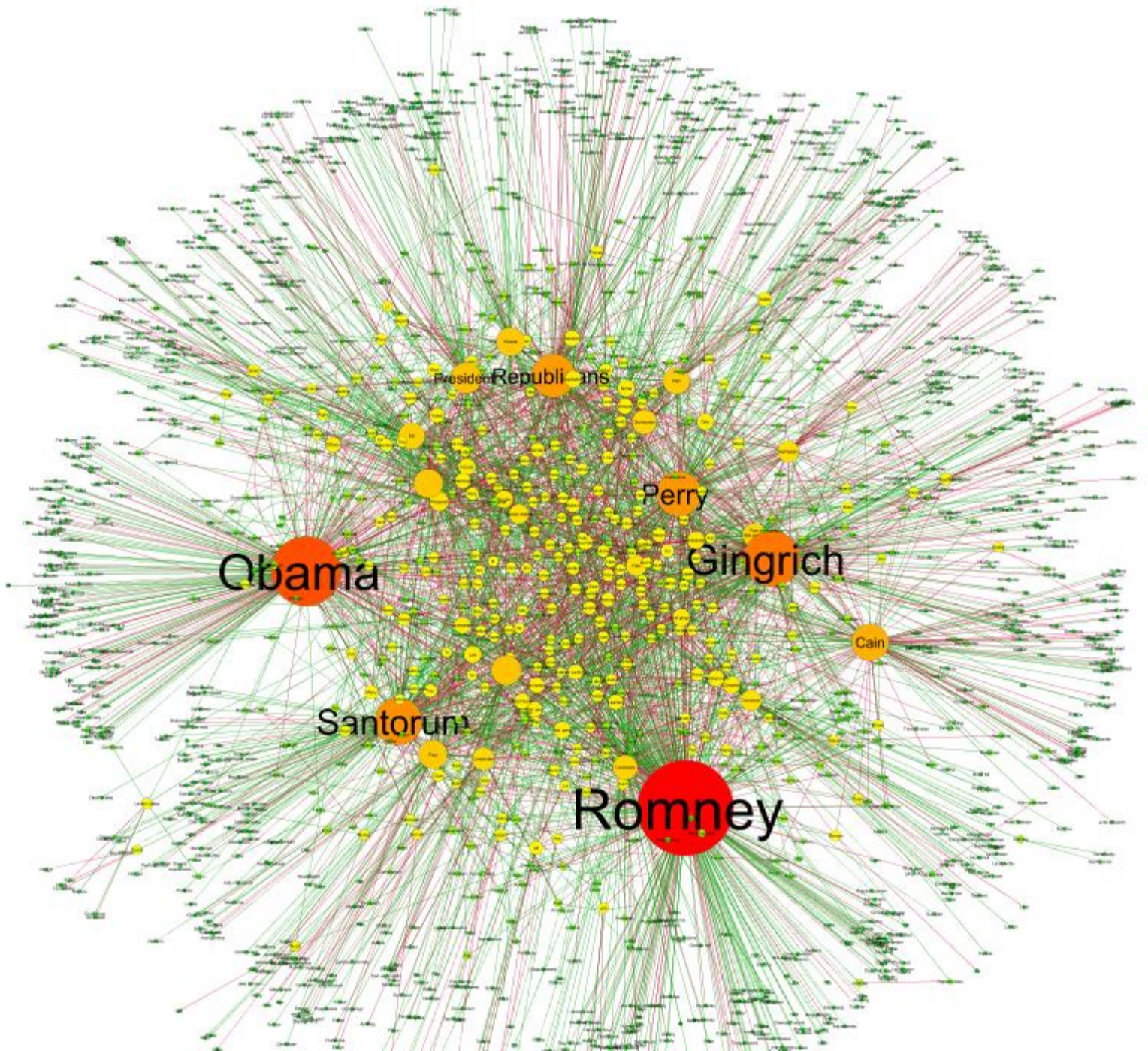
Crime against Person		Crime against Property	
Subjects	Objects	Subjects	Objects
Priest	People	Man	Money
Man	Boy	Police	Bank
Troops	Child	Soldiers	Records
Reyes	Girl	Winona Ryder	Millions
Geoghan	Man	Priest	Weapons
Shanley	Woman	People	Wallet
Forces	Jogger	Jason Bogle	Trade Secret
Police	Victim	Investigators	Steven Seagal
United States	Minors	Employee	Most
Others	Me	Agents	Man

Networks of political support

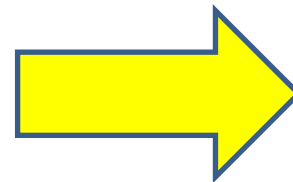
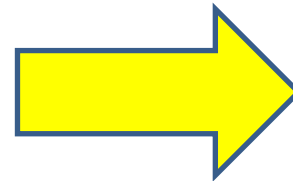
Example: US Presidential Elections

Types of information about actors:

- Party loyalties
- Subject / object bias
- Positive / negative action bias

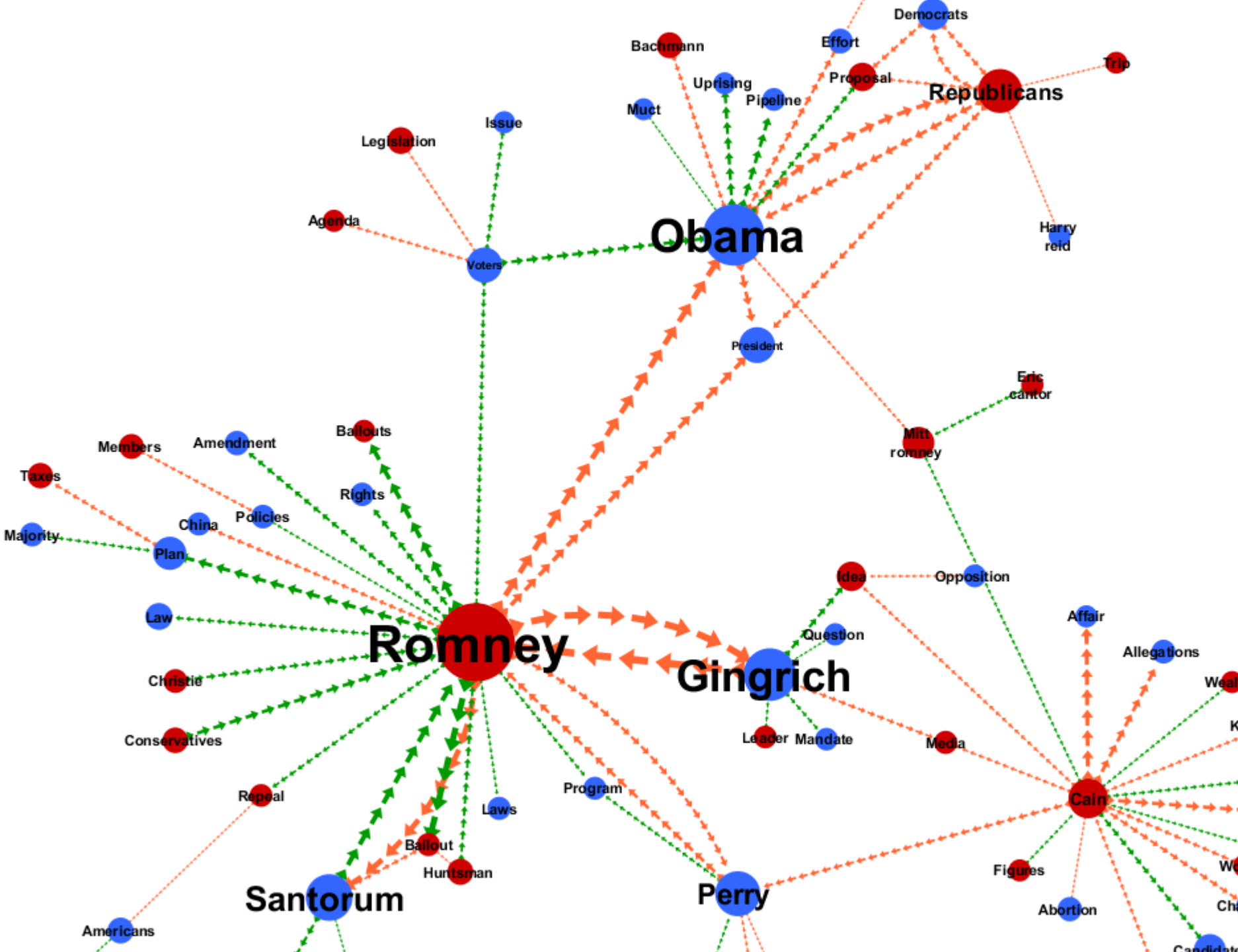


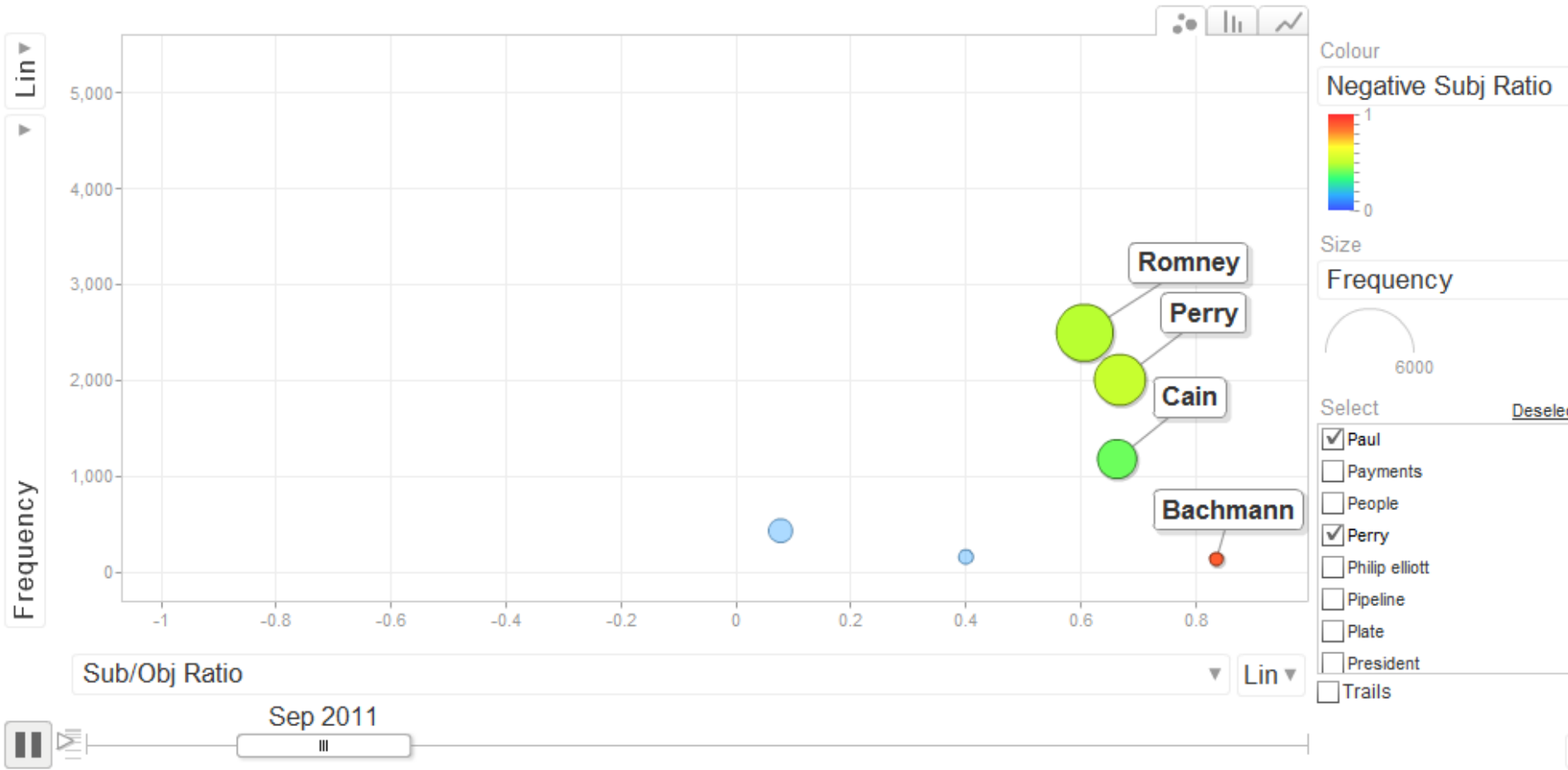
- ★ Gingrich Oppose Romney
- ★ Romney Oppose Gingrich
- ★ Romney Endorse Bailout
- ★ Romney Oppose Obama
- ★ Romney Oppose Santorum
- ★ Santorum Endorse Romney
- ★ Romney Endorse Bailouts
- ★ Romney Oppose President
- ★ Obama Oppose Republicans
- ★ Mitt Romney Oppose President
- ★ Republicans Oppose Obama
- ★ Obama Oppose President
- ★ Conservatives Endorse Romney
- ★ Cain Oppose Allegations
- ★ Cain Oppose Accusations
- ★ Perry Oppose Romney
- ★ Obama Endorse Uprising
- ★ Santorum Endorse Earmarks
- ★ Romney Endorse Rights
- ★ Romney Oppose Perry



- ★ Obama
- ★ Perry
- ★ Gingrich
- ★ President
- ★ Democrats
- ★ China
- ★ Santorum
- ★ Muct
- ★ Uprising
- ★ Harry Reid
- ★ Bailouts
- ★ Bailout
- ★ Christie
- ★ Future
- ★ Repeal
- ★ Mitt Romney
- ★ Cain
- ★ People
- ★ Republicans
- ★ Romney

Note: in primary phase
many people oppose Romney
Creating errors





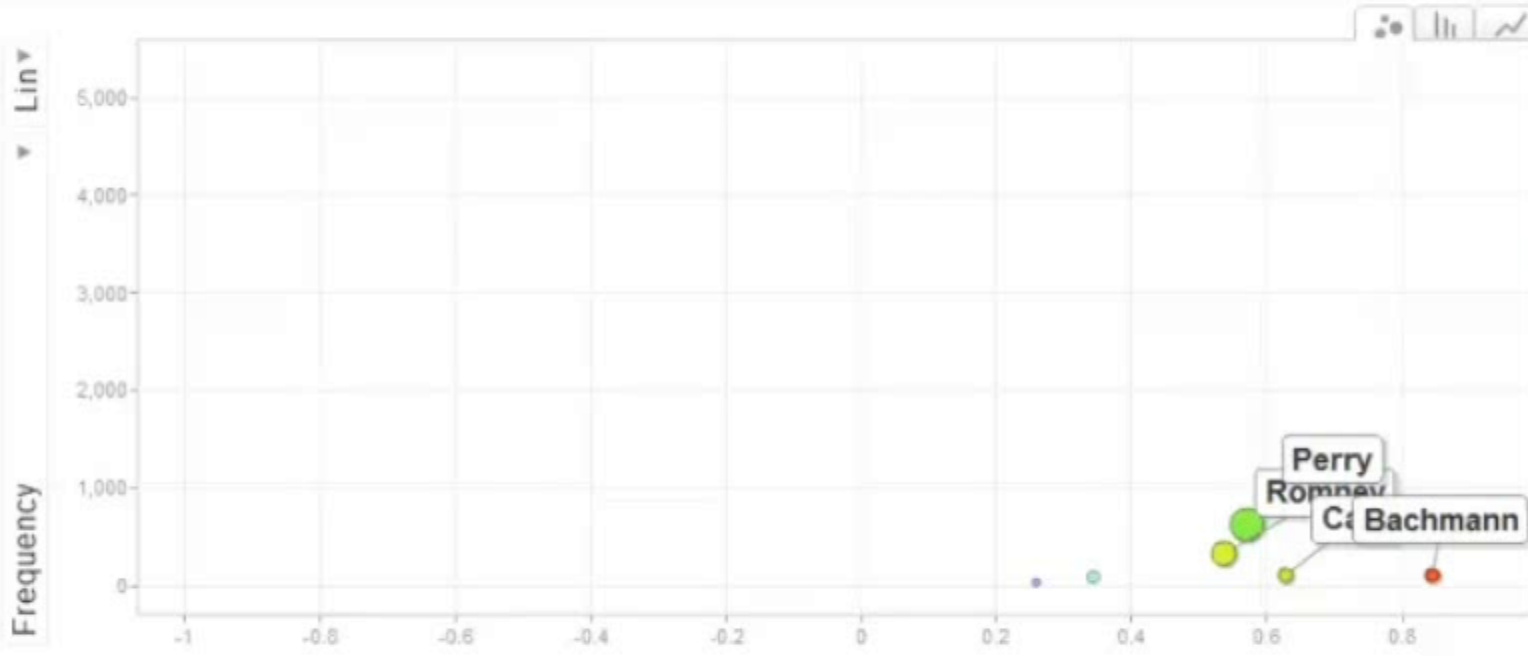
Motion Chart_Updated ☆

Nello Cristianini ▾

Share

File Edit View Insert Format Data Tools Help Last edit was made 22 hours ago by nello.cristianini

Edit Settings Publish Add Gadget to iGoogle... more ▾



Colour
Negative Subj Ratio ▾
1
0

Size
Frequency ▾
6000

Select Deselect all

- Anderson
- Appointment
- Bachmann
- Bailout
- Bailouts
- Bail
- Trails

Sub/Obj Ratio

Sep 2011

Play

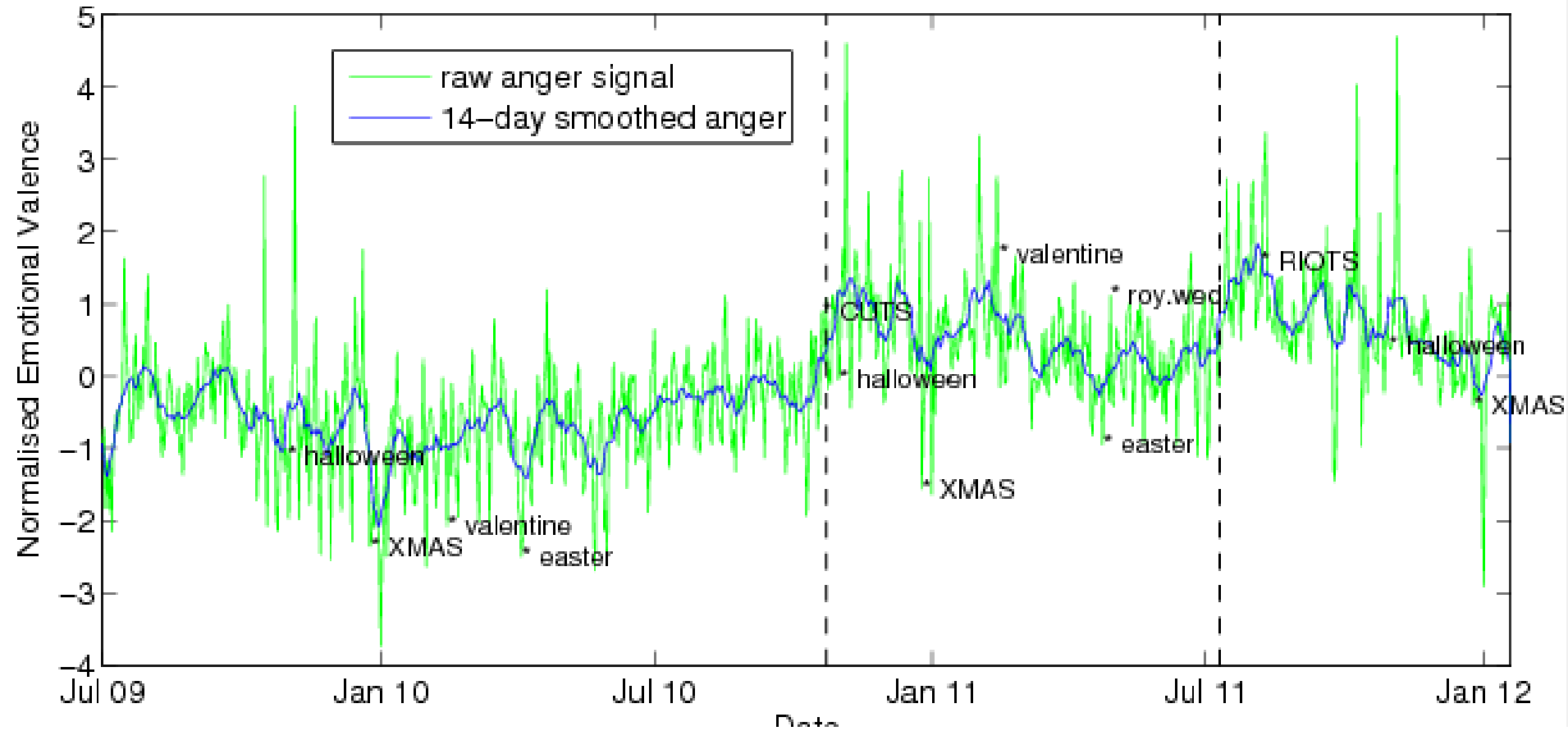
+ Sheet 1 Gadget3 Gadget2 Gadget1 ▾

Question 4

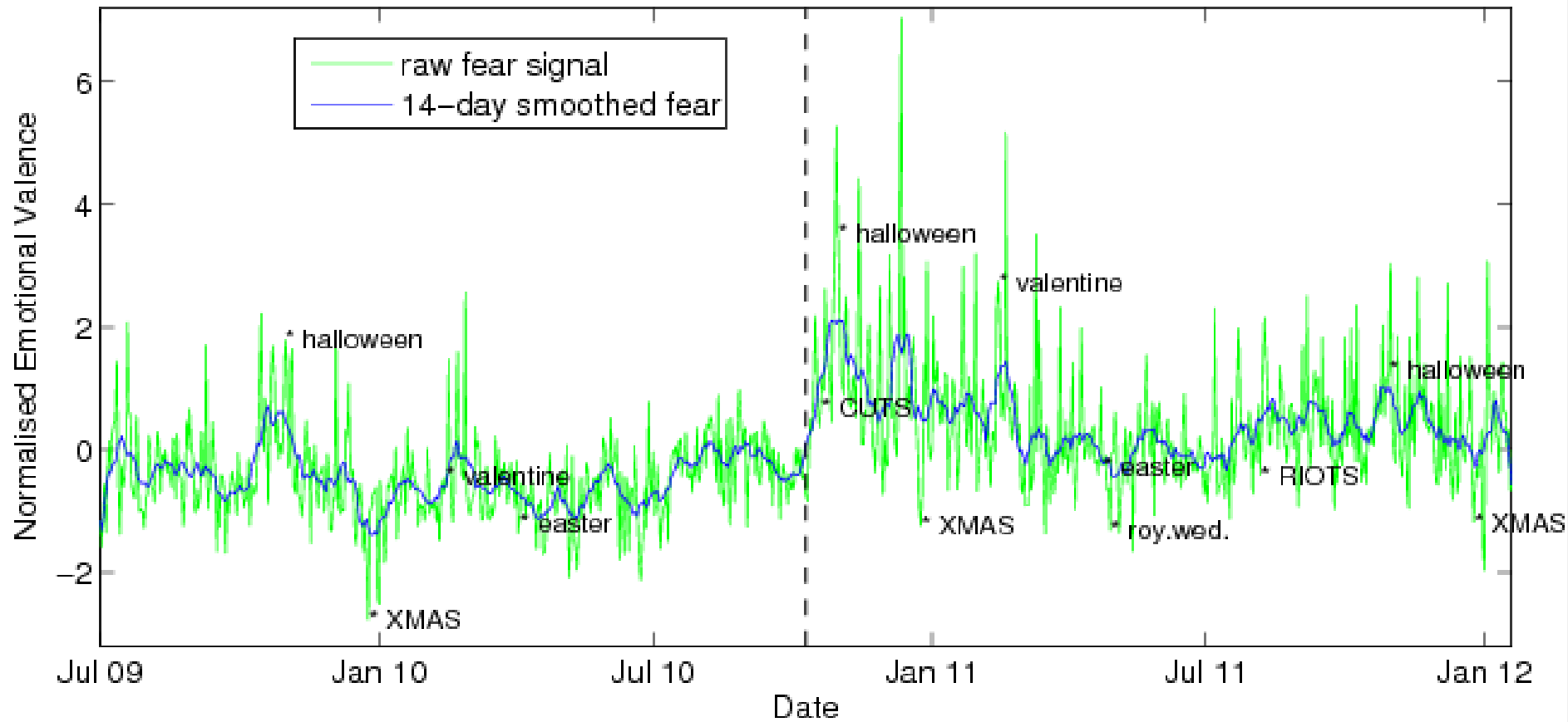
Measuring Public Mood



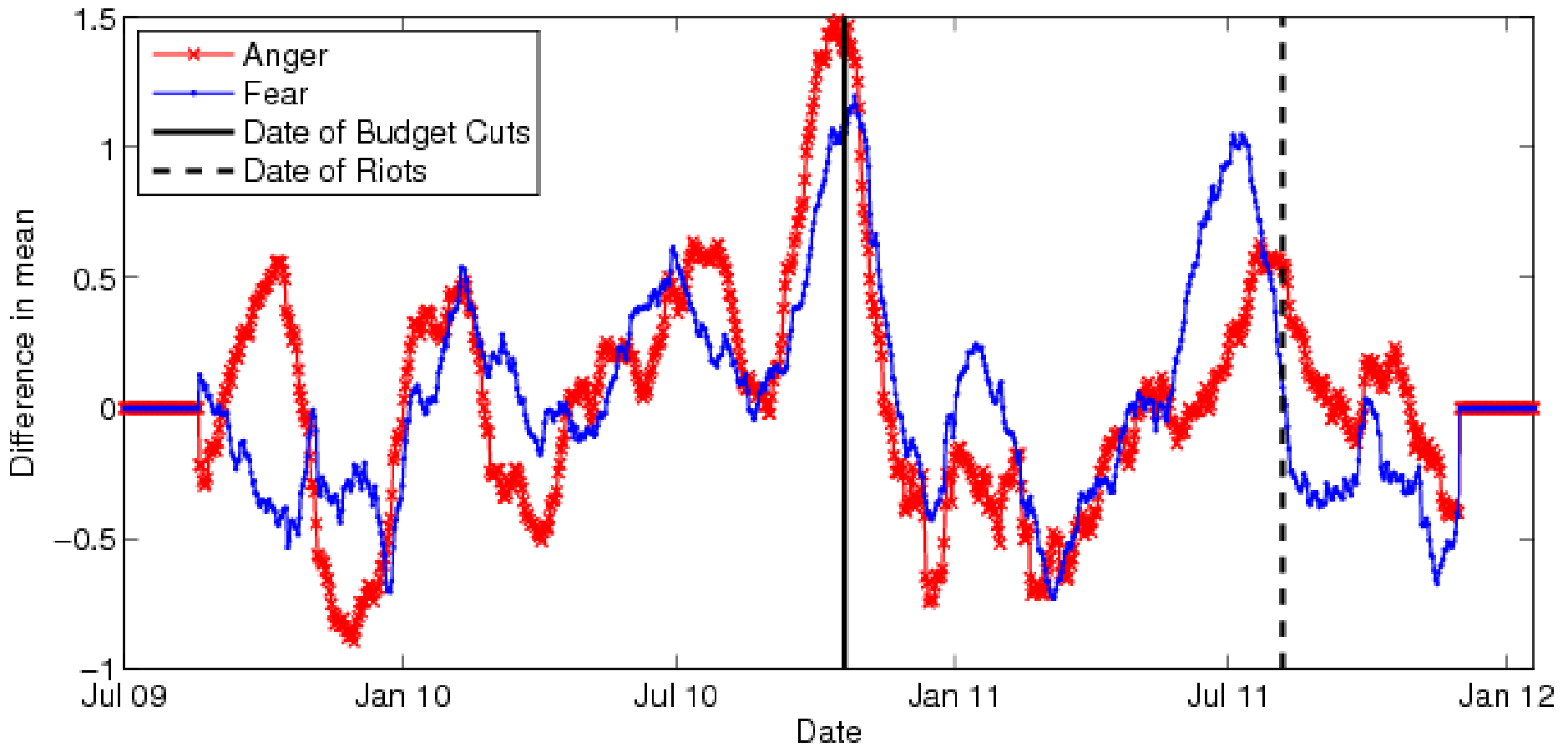
933 Day Time Series for Anger in Twitter Content



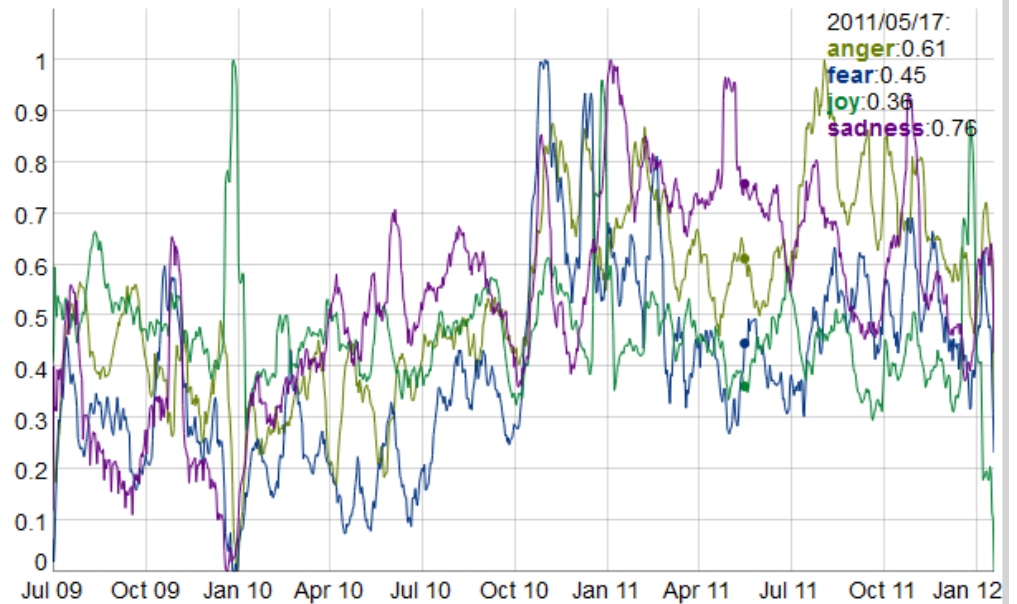
933 Day Time Series for Fear in Twitter Content



Rate of Mood Change by Day using the Difference in 50-day Mean

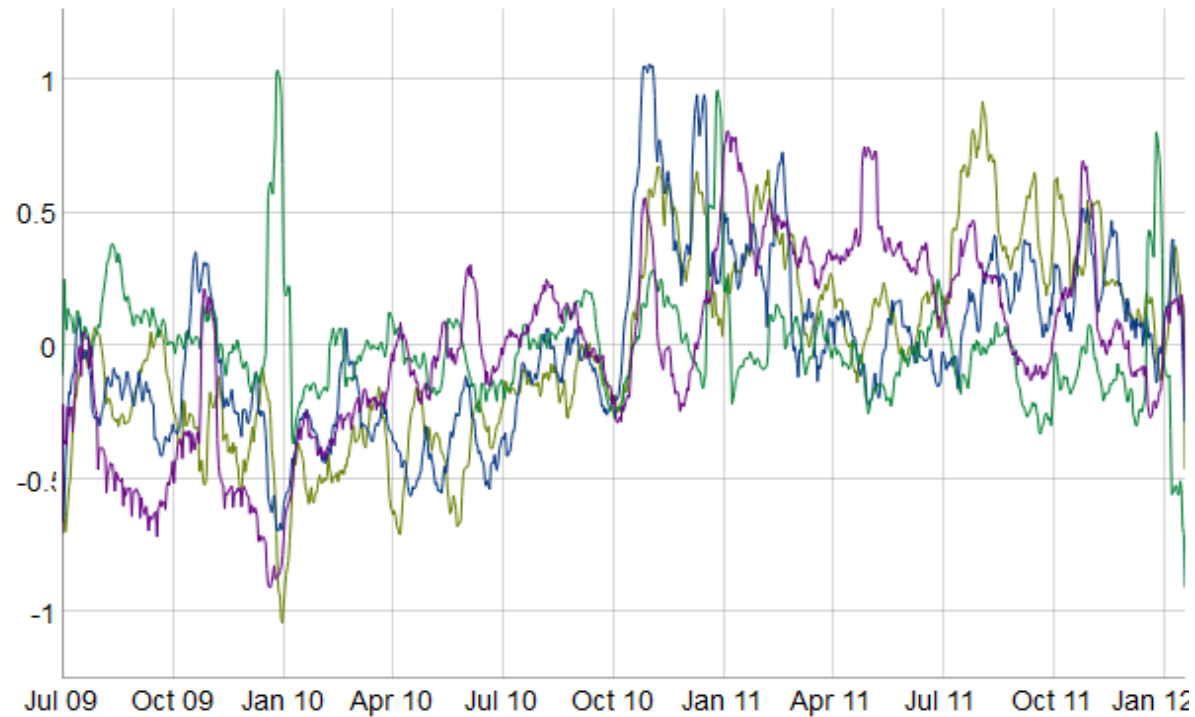


The Face of Britain...



Animation of Mood Changes

Mood Changes in UK Twitter Content 2009-2012



Intelligent Systems Laboratory - University of Bristol
09/11/2011

Face Visualisation Tool: courtesy of <http://grimace-project.net/>
mediapatterns.enm.bris.ac.uk



Demos

celebwatch.enm.bris.ac.uk

foundintranslation.enm.bris.ac.uk

geopatterns.enm.bris.ac.uk/epidemics

electionwatch.enm.bris.ac.uk

mediapatterns.enm.bris.ac.uk

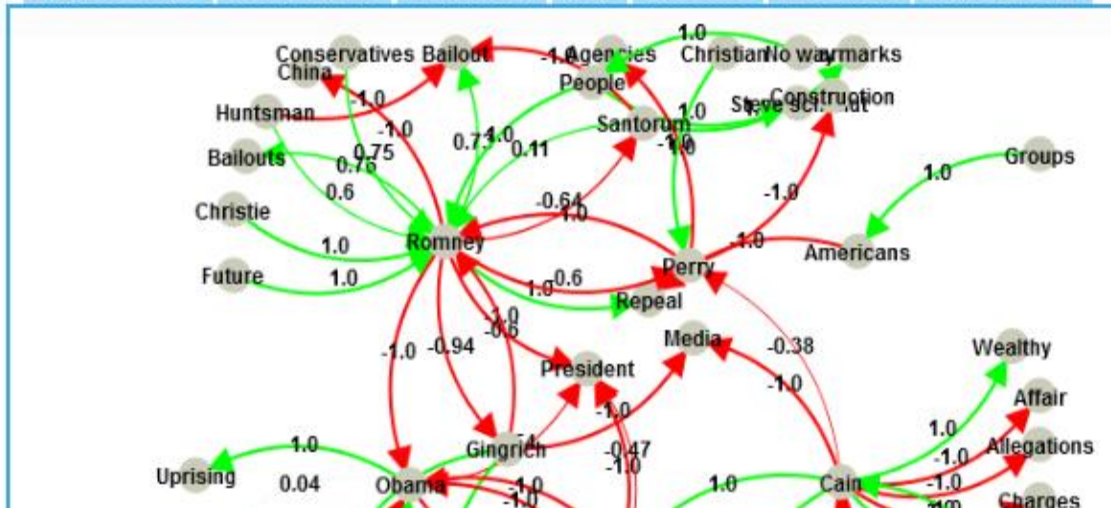


Actor Spectrum



- ★ [Obama](#)
- ★ [Perry](#)
- ★ [Gingrich](#)
- ★ [President](#)
- ★ [Democrats](#)
- ★ [China](#)
- ★ [Santorum](#)
- ★ [Muc](#)
- ★ [Uprising](#)
- ★ [Harry Reid](#)
- ★ [Bailouts](#)
- ★ [Bailout](#)
- ★ [Christie](#)
- ★ [Future](#)
- ★ [Repeal](#)

[Triplet Graph](#) [Actor Space](#) [Actor Bias](#) [Map](#) [Timeline](#) [Bar Chart](#) [Activity Map](#)



Relations



- ★ [Gingrich Oppose Romney](#)
- ★ [Romney Oppose Gingrich](#)
- ★ [Romney Endorse Bailout](#)
- ★ [Romney Oppose Obama](#)
- ★ [Romney Oppose Santorum](#)
- ★ [Santorum Endorse Romney](#)
- ★ [Romney Endorse Bailouts](#)
- ★ [Romney Oppose President](#)
- ★ [Obama Oppose Republicans](#)
- ★ [Mitt Romney Oppose President](#)
- ★ [Republicans Oppose Obama](#)
- ★ [Obama Oppose President](#)
- ★ [Conservatives Endorse Romney](#)
- ★ [Cain Oppose Allegations](#)

Topics

- ★ [Arts](#)
- ★ [Business](#)
- ★ [Environment](#)
- ★ [Politics](#)
- ★ [Religion](#)
- ★ [Crime](#)
- ★ [Disasters](#)
- ★ [Elections](#)
- ★ [Fashion](#)
- ★ [Prices](#)
- ★ [Markets](#)
- ★ [Petroleum](#)
- ★ [Religion](#)
- ★ [Science](#)
- ★ [Sports](#)
- ★ [Travel](#)
- ★ [Weather](#)

Search

[Map](#) [Bar Chart](#) [Text Only](#)



Greece Activity



Countries

- ★ [Austria](#)
- ★ [Belgium](#)
- ★ [Bulgaria](#)
- ★ [Cyprus](#)
- ★ [Czech Republic](#)
- ★ [Denmark](#)
- ★ [Estonia](#)
- ★ [Finland](#)
- ★ [France](#)
- ★ [Germany](#)
- ★ [Greece](#)
- ★ [Hungary](#)
- ★ [Ireland](#)
- ★ [Italy](#)
- ★ [Latvia](#)
- ★ [Lithuania](#)
- ★ [Luxembourg](#)

Markets: Greece

[Demand and a rise in grain prices](#)

Regardless of the other factors that drive the markets, the fear of shortcomings in the market cereal prices as demand for us and China. China is growing demand for kitchen oils and for an [Read more...](#)

Turchi, M., Flaounas, I., Ali, O., Bie, T.D., Snowsill, T., Cristianini, N.: Found in translation. In: Machine Learning and Knowledge Discovery in Databases, European Conference (ECML/PKDD), Lecture Notes in Computer Science, vol. 5782, pp. 746–749. Springer, Bled, Slovenia (2009)

Top memes (2010-08-26)

is Prev Next is

- [the anniversary](#)
- [checkpoint](#)
- [powerful](#)
- [north korean leader kim jong](#)
- [michael enright](#)

← < Aug 2010 > →						
S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	1	2	3	4

In the news (103)

[North Korean leader reported to be in China](#)

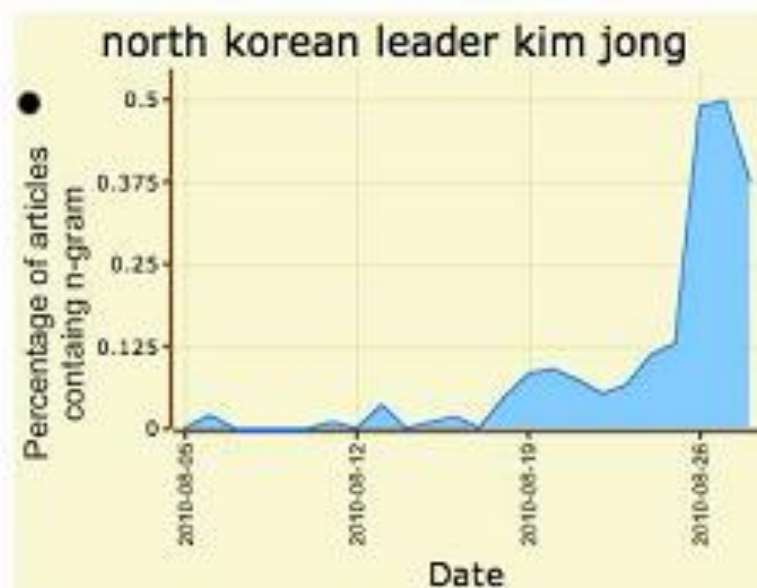
[2010-08-26 06:00:12] North Korean leader Kim Jong Il is visiting China, the South Korean JoongAng--

[Kim Jong-Il may be visiting China: S.Korea official](#)

[2010-08-26 06:00:25] North Korean leader Kim Jong-Il may have left for China Thursday in what would...

[N Korea's Kim Jong-il 'in China'](#)

Meme timelines



Cluster details [1-17]

kim jong | jong | leader kim jong | il | jong - kim jong - leader kim | kim jong - il | jong - il | korean leader | north korean leader | - il | north korean | korean leader kim jong | korean leader kim | north korean leader kim | leader kim jong - il | leader kim jong - to china | succession | kim 's | kim ' | jilin | that kim | korea 's | korea

Snowsill, T., Flaounas, I., Bie, T.D., Cristianini, N.: Detecting events in a million new york times articles. In: Machine Learning and Knowledge Discovery in Databases, European Conference (ECML/PKDD). pp. 615–618. Barcelona, Spain (2009)

Celebrity Watch

Celebrity Watch

... the World's hottest celebrity gossip, tracking 11325 Celebrities in 1000 News Outlets!
[Read on...](#)

Mariah Carey Activity



Map **Timeline** Social Network Text Only



Mariah Carey (27)

[Inside Mariah Carey's baby shower!](#)

9 Mar 2011 17:21:59 GMT

On March 6, Mariah Carey and Nick Cannon gathered an intimate group of friends and family at the rooftop Conservatory Grill at the Montage hotel to celebrate the arrival of their twin boy and girl, due in late April or early May. [Read more on Inside Mariah Carey's baby shower!](#)

[Read more...](#)

[Mariah Carey Dishes On Her Baby Shower!](#)

9 Mar 2011 21:19:57 GMT

MARIAH Carey says her baby shower is the first one she's ever been to. The superstar singer — who's expecting twins with husband Nick Cannon — gathered an intimate group of friends and family at the rooftop Conservatory Grill at the Montage hotel to celebrate

Search

Hot Today

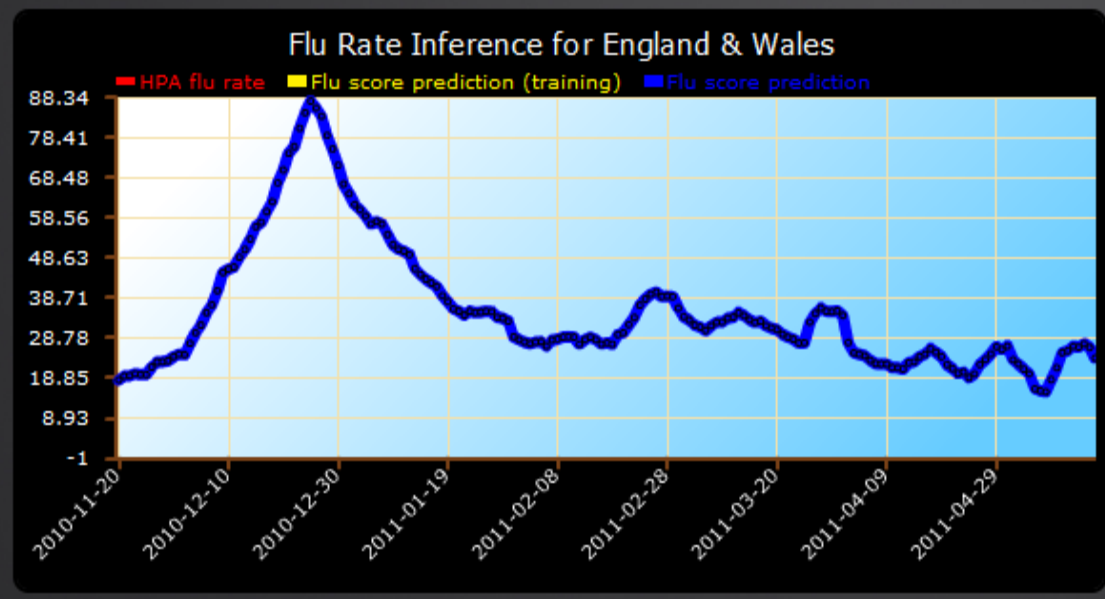
- Mariah Carey (27)
- Nelly Furtado (13)
- Collins (20)
- Mark Cuban (7)
- Carey (14)
- George Michael (12)
- Jon Cryer (55)
- Till (5)
- John Stamos (24)
- Chuck Lorre (72)
- Phil Collins (22)
- Tyson (17)
- Mike Tyson (18)
- Alexander McQueen (8)
- Johnny Depp (11)
- Vanessa Hudgens (3)
- Tom Izzo (1)
- Matt Damon (8)
- Leslie Moonves (3)
- Webster (10)

Hot This Month

- Justin Bieber (33)
- Natalie Portman (14)
- Colin Firth (20)
- James Franco (5)
- Britney Spears (54)
- Kim Kardashian (33)
- Michael Jackson (14)
- Perry (54)
- Anne Hathaway (7)
- Christina Aguilera (20)
- Nicole Kidman (8)
- Jennifer Aniston (18)
- Oprah Winfrey (30)

Here are the inferred flu rates for **England & Wales** in the last 6 months based on geolocated Twitter content

Check Regional Flu Inferences:
for the last 6 months since June 2009



Current Flu Rate
23.76

Flu Detector uses the content of Twitter to nowcast flu rates in several UK regions. Inferences are compared with official ILI rates from **HPA**. Performance evaluation results are available [here](#).

The methodology is described in the following papers:

Tracking the flu pandemic by monitoring the Social Web, Lamos and Cristianini, CIP 2010.

Flu Detector - Tracking Epidemics on Twitter, Lamos, De Bie and Cristianini, ECML PKDD 2010.

Vertical axis **Inferred flu rate** indicating the number of GP consultations per 100,000 citizens where the diagnosis' result was Influenza-like Illness (ILI)
Horizontal axis Date in yyyy-mm-dd format



Conclusions

- Many diverse case studies, trying to make a single point: standard machine learning technologies can help the social sciences to enter its “big data” (or -omics) phase.
- We started from a single question: can we capture macroscopic scale patterns in the contents of the global media system?
(of the type that a single observer cannot see)

Conclusions

- This journey took us to deal with named entity disambiguation, social network analysis, narrative analysis, machine translation, topic detection, sentiment analysis... but also databases, data visualisation, data mining...
- I hope I managed to convey just part of the fun we had (dealing with social scientists, lawyers, psychologists)...



Thanks To

Ilias Flaounas, Omar Ali, Elena Hensinger, Bill Lampos, Marco Turchi, Saatviga Sudhahar,
(Intelligent Systems Laboratory, Bristol)

Justin Lewis, Nick Mosdell
(School of Journalism, Cardiff)

Roberto Franzosi
(Emory University)



MediaPatterns.enm.bristol.ac.uk



mediapatterns.enm.bris.ac.uk