Dynamical Classes of Collective Attention in Twitter

Janette Lehmann, *Bruno Gonçalves, José J. Ramasco, Ciro Cattuto*
WWW-2012, Lyon

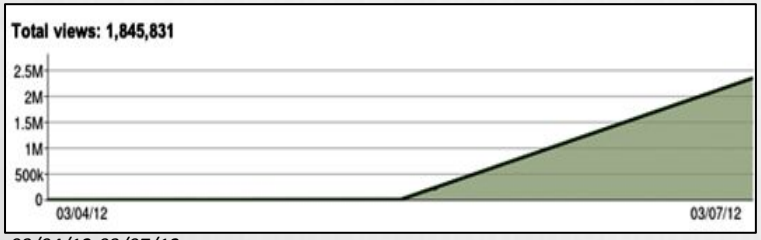
UNIVERSITAT POMPEU FABRA

ISI Foundation

Outline

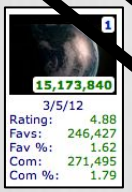
1. **Motivation**
2. **Data and methodology**
3. **Activity peak detection**
4. **Classes of peaks**
5. **Social semantics**
6. **Information spreading**
7. **Summary**

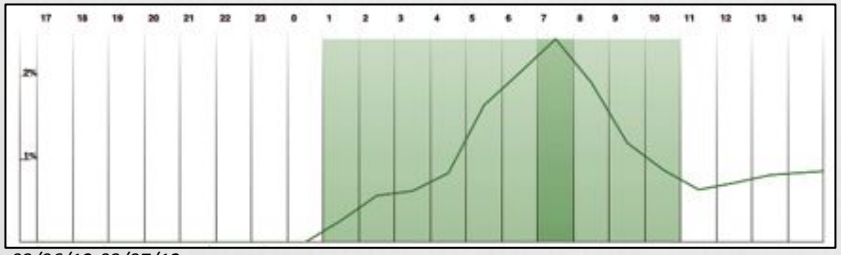
Motivation



03/04/12-03/07/12

YouTube [Mar 5, 2012]
KONY 2012 is a film and campaign by Invisible Children that aims to make Joseph Kony famous, not to celebrate him, but to raise support for his arrest and set a precedent for international justice. (Invisible Children)



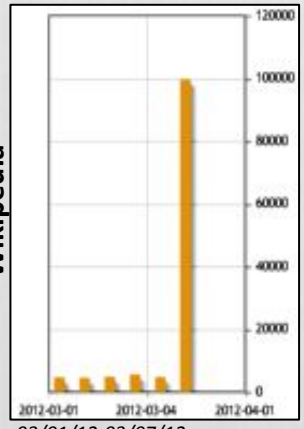


03/06/12-03/07/12

Twitter [Mar 7, 2012]
Let's see if we can keep the world wide web from com-busting: Watch the film. Sign the pledge. #KONY2012 http://bit.ly/Ald4aj

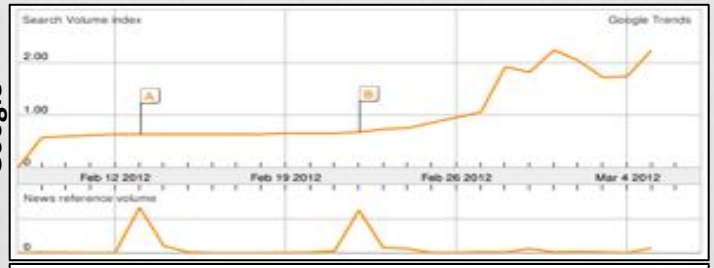
➔ #stopkony

Wikipedia

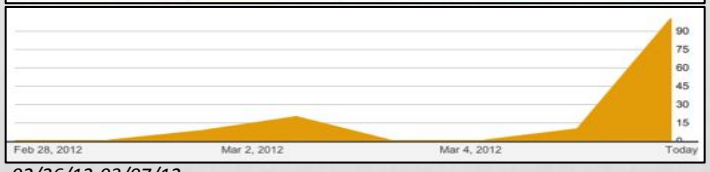


03/01/12-03/07/12

Google



Yahoo!



02/26/12-03/07/12

Motivation

Wikipedia

- Joseph Kony
- Soviet invasion of Afghanistan

Twitter

- #iLoveWhenYou
- #stopkony
- Peyton Manning

YouTube

- Big Cats + Mirrors
- DollarShaveClub.com
- KONY 2012

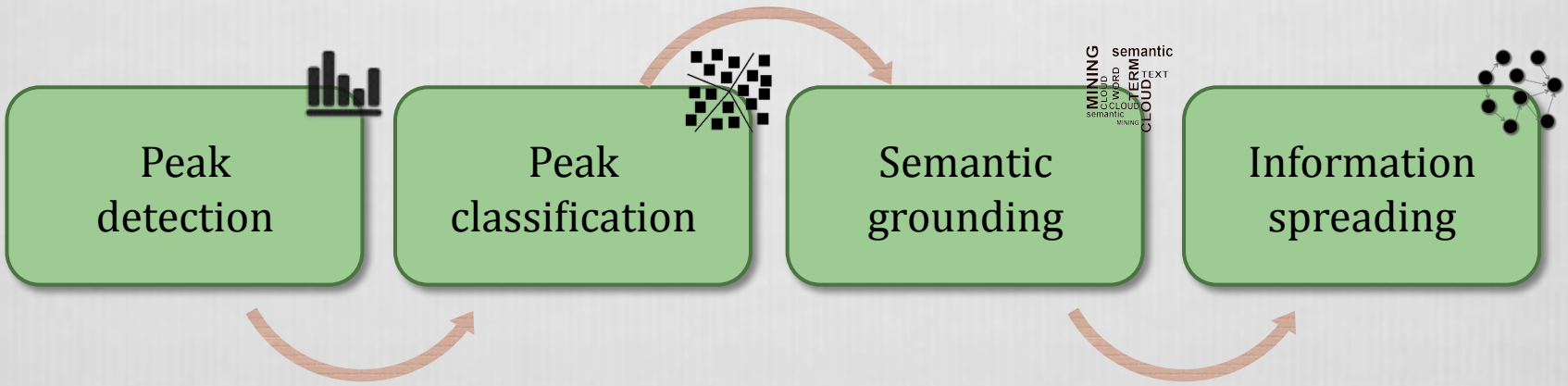
Trends [07/05/2012]

- **Investigation of the characteristics of focused attention in online systems**
  - **Endogenous: internal propagation**
  - **Exogenous: injection from external sources**
- **Twitter: Micro-blogging system, one of the most popular online social network**
- **Analyze and classify peaks of hashtag-usage**

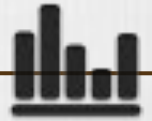
Do variations in activity profiles offer differences in their semantics and the way the information is spread?

Data and Methodology

Twitter data stream and REST API	
Period	11/20/08 – 05/27/09
#Tweets	131.7 Mio
#Tweets with Hashtags	4.3 Mio
#Hashtags	408 K
#Users	6.5 Mio
#Users of the follower network	2.7 Mio



Activity Peak Detection

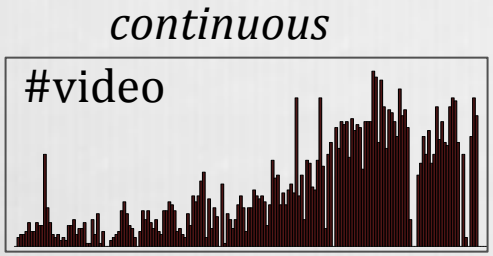


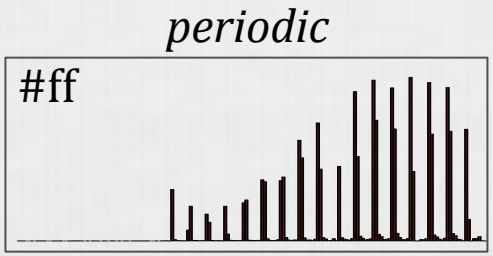
	Method	Data source
M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos [2004]	• Normalized Discrete Fourier Transform • Moving average	MSN query logs
R. Crane and D. Sornette [2008]	• Power law distribution • Poisson process	YouTube videos
J. Yang and J. Leskovec [2011]	• Similarity metric • K-Spectral Centroid clustering algorithm • Predictive model	Twitter News articles Blog posts

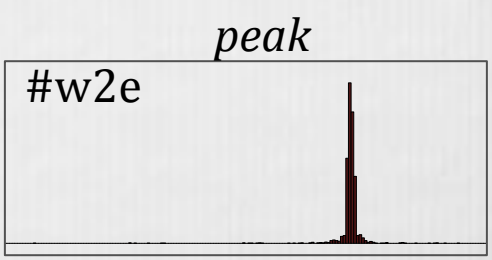
Problems:

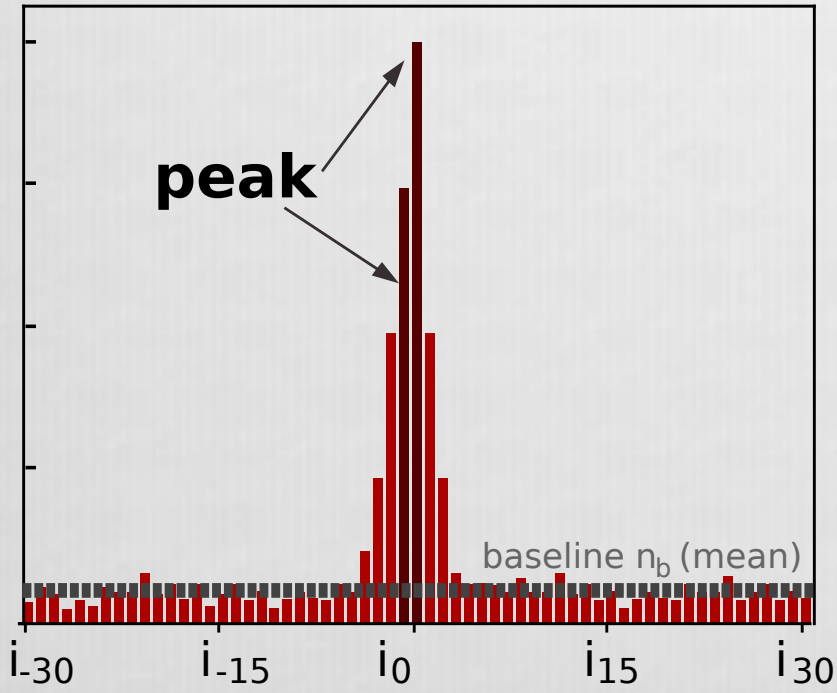
- Computational complexity for large-scale data analysis
- Cannot capture all activity profiles in Twitter
- Twitter reacts in seconds/minutes and includes anticipation

Activity Peak Detection









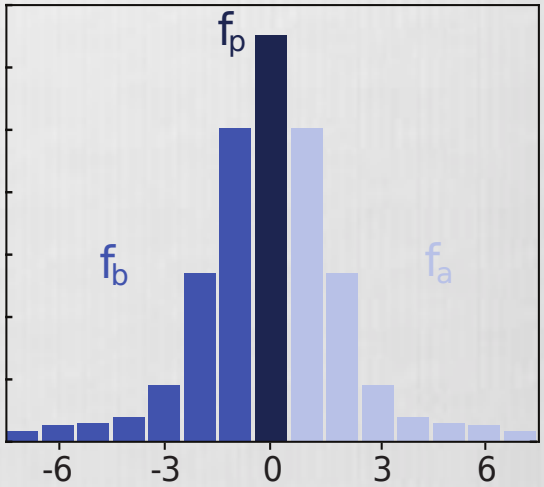
$$p(i_0) = \frac{n(i_0) - n_b}{\max(n_b, 10)} \geq 10$$

- **Peak: relative activity to baseline have to be 10 times larger**
- **Minimal level of activity expected**
- **Selection of isolated popularity bursts (no other peaks one week before/after)**
- **We detected 115 peaks**

Classes of Peaks

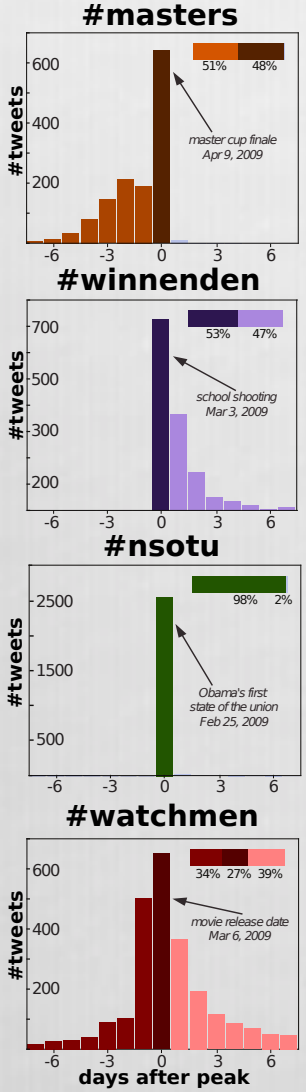
- **Investigation into the shapes of peaks**
  - **Before peak: Endogenous/exogenous cause**
  - **After peak: relaxation shape**
- **Hashtag usage timeline is summarized to a triple (f_b, f_p, f_a) representing fractions of tweets posted before (f_b), during (f_p), and after (f_a)**

$$f_b + f_p + f_a = 1$$



- **Clustering using EM algorithm with an optimal Gaussian mixture model**
- **We found 4 clusters**

Classes of Peaks

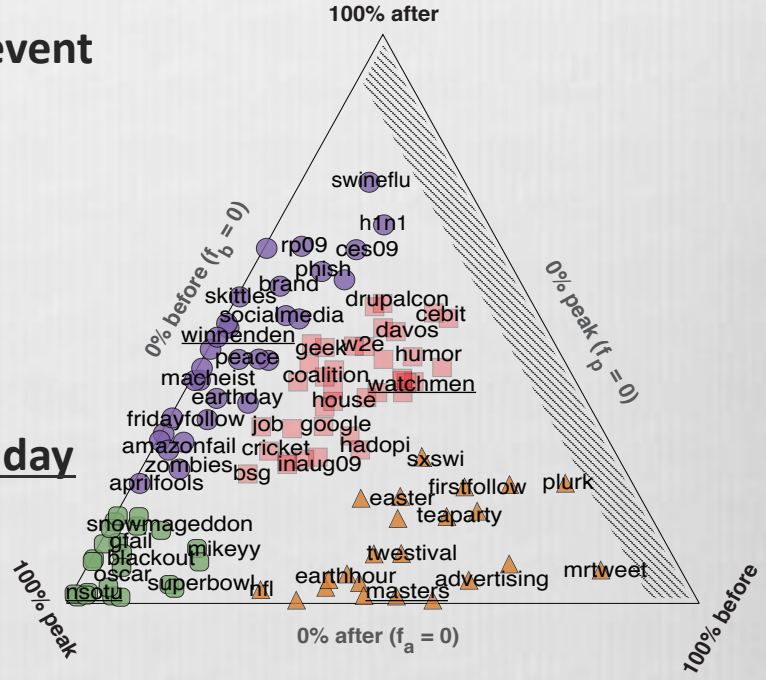


- **Anticipatory behavior**
- **Increasing amount of tweets until the event**
- **Sharp drop of attention after the event**

- **Unexpected events**
- **Driven by exogenous sources**

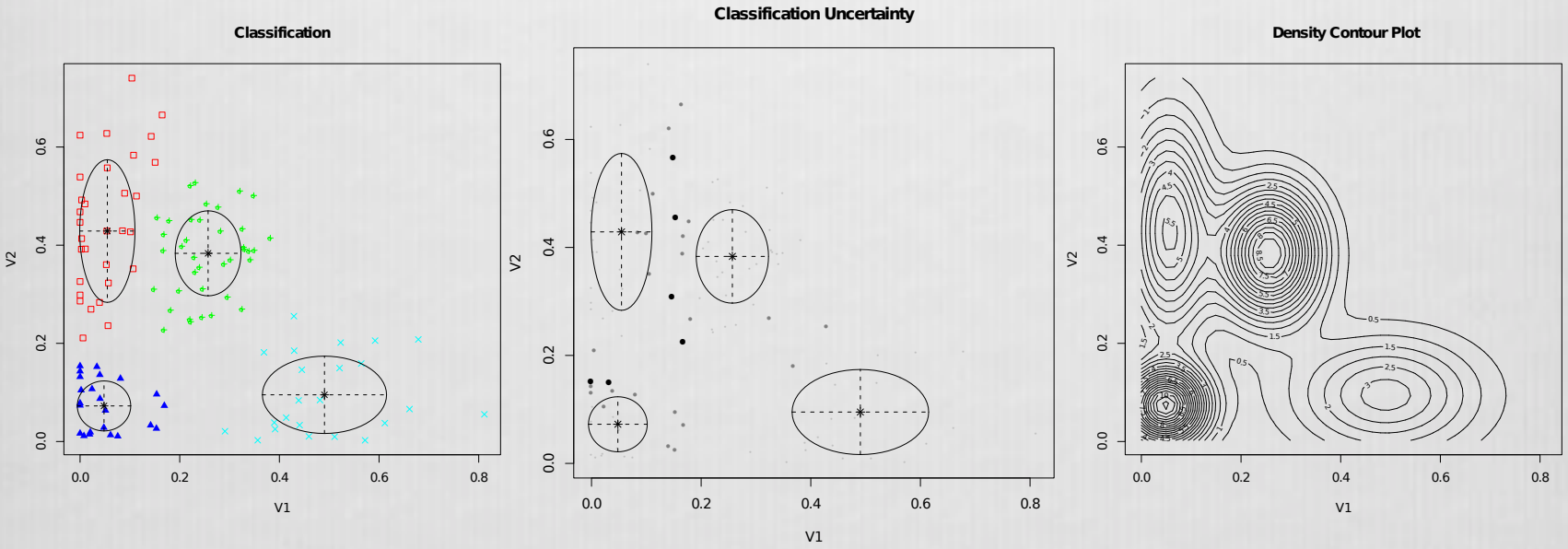
- **Activity concentrated on the peak day**
- **Events that only discussed while they are happen**

- **Collective attention is built up to a peak intensity, then attention shifts away**



Classes of Peaks

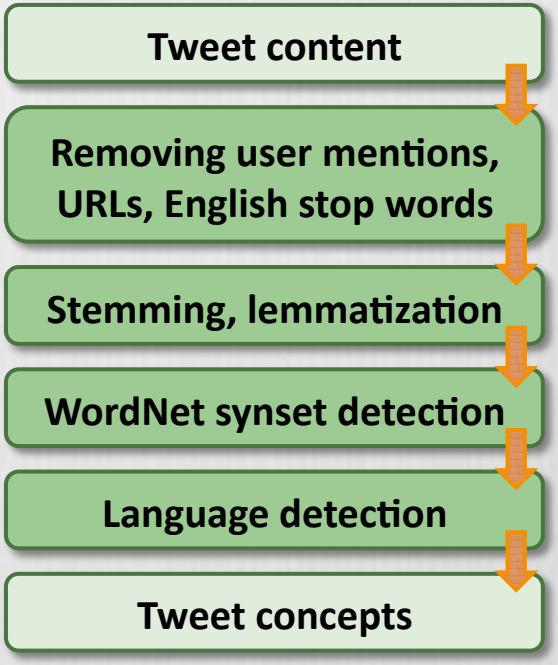
- **Classification robustness**
  - 77% of the hashtags have a classification uncertainty below 5%
  - 6% of them have a classification uncertainty in excess of 20%



Social Semantics

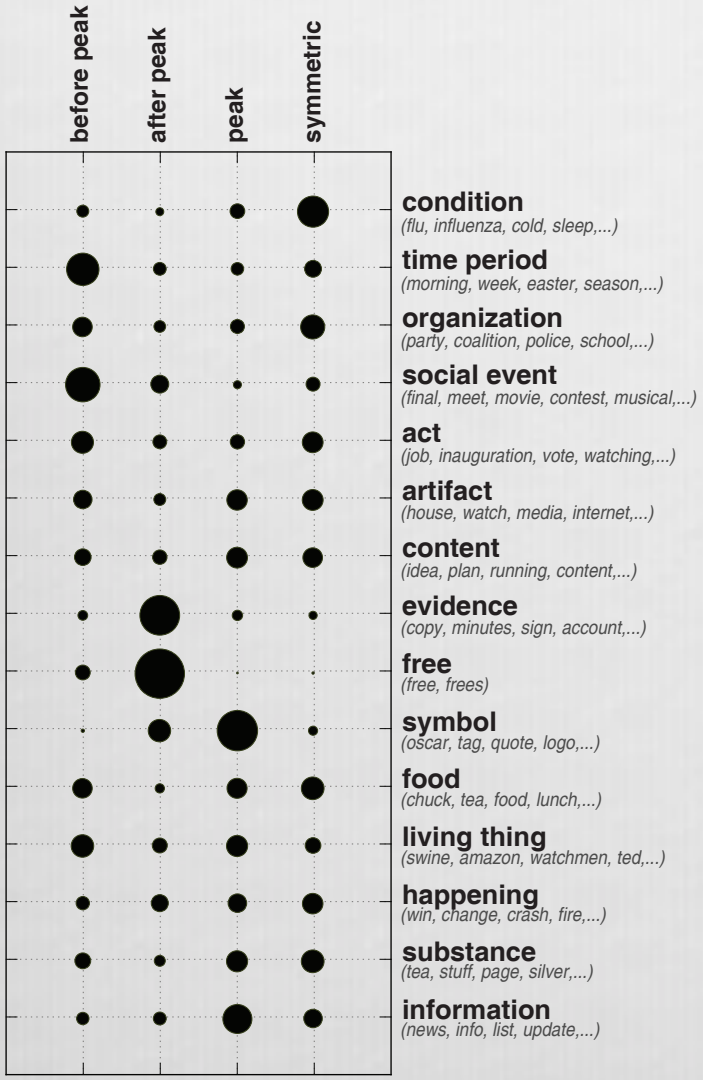
- Analyzing semantic differences of peak classes

- We found 18,000 distinct concepts
- WordNet provides hierarchical structures of concepts
- Mapping the detected concepts to concepts at depth 4



- The applied semantic grounding is simple, but gives a first insight in the semantic differences of the peak classes

Social Semantics



Before peak

- *"social events", "time period"*
- **Social anticipation of a known event**

After peak

- *"free", "evidence"*
- **Marketing events**

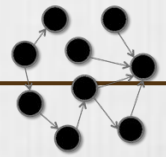
Peak

- *"symbol"*
- **Sport, media events**

Symmetric

- *"condition", "organization"*
- **Conventions, pandemics**

Information Spreading



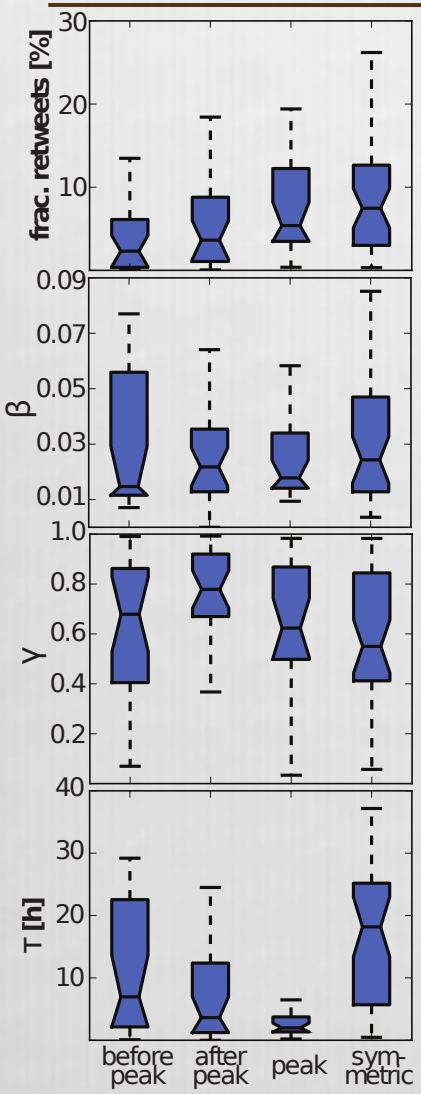
- **Analyzing information spreading of the peak classes**

- **Information spreading is modeled as an epidemic process**
- **Based on follower network of Twitter**

- **Infection = posting a tweet that contains the hashtag**
- **Infection period = time span from the first to the last tweet containing the hashtag**
- **Infection source = internal (follower network) or external**

β – Fraction of seeders γ - Fraction of a user's followers who got infected τ – Infection period

Information Spreading



Before peak

- **No fast information spreading, more discussions (fraction of retweets low)**

After peak

- **Exogenous (Y high)**

Peak

- **Endogenous (fraction of retweets high)**

Symmetric

- **Endogenous (fraction of retweets high)**
- **Users participate during the whole time (τ high)**

β – Fraction of seeders

γ - Fraction of a user's followers who got infected

τ – Infection period

Summary

		Description	Examples
	Before peak	• Anticipatory discussions • Scheduled social events	#easter #earthhour #sxswi
	After peak	• Unexpected events • Marketing, catastrophes (exogenous)	#amazonfail #macheist #free
	Peak	• Unexpected/planned events • Sport, media, games, blackouts (exogenous/endogenous)	#oscar #superbowl #gfail
	Symmetric	• Increasing/decreasing attention • Conventions, pandemics (endogenous)	#cebit #bsg #davos

Summary

- **Identification of 4 classes of peak types offering a different semantic fingerprint and information spreading behavior**
- **Approach to detect peak classes is simple and can be applied in large-scale situations**
- **Improving of the semantic analysis in the future (e.g. sentiment analysis, crowd sourcing)**
- **Epidemic processes play a minor role**

Thank you

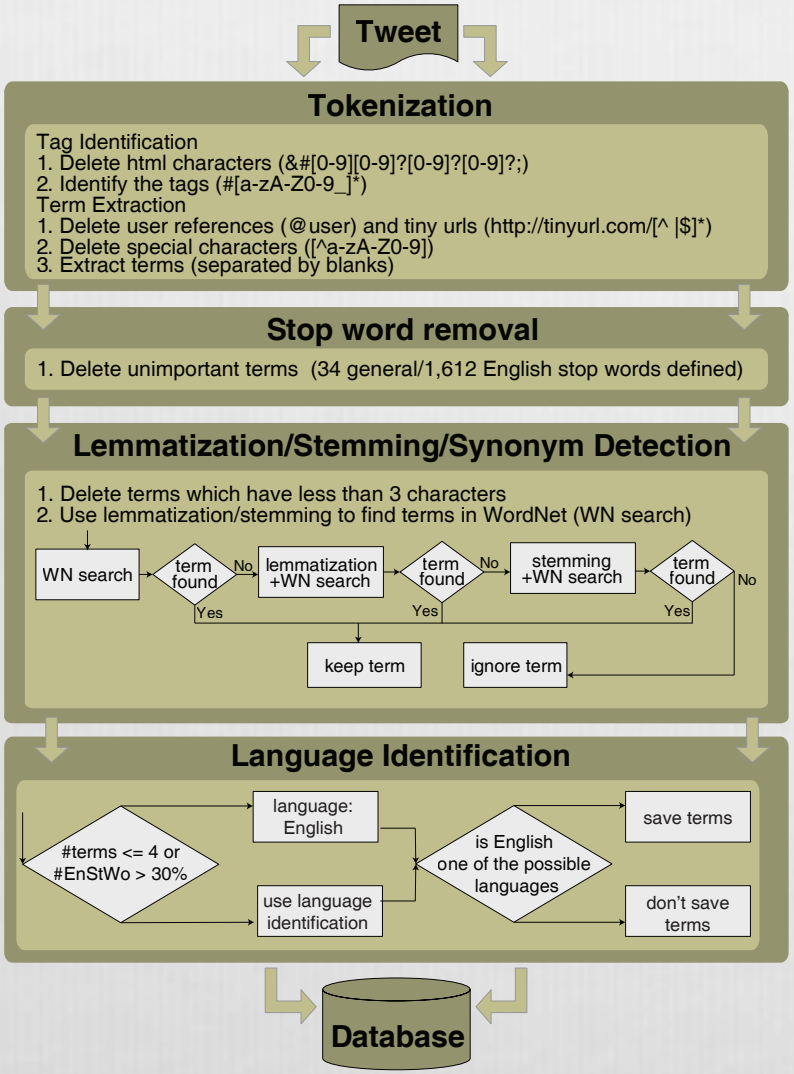
- **Janette Lehmann** **janette.lehman01@estudiant.upf.edu**
- **Bruno Gonçalves** **b.goncalves@neu.edu**
- **José J. Ramasco** **jramasco@ifisc.uib.es**
- **Ciro Cattuto** **ciro.cattuto@isi.it**

References

- R Crane and D Sornette. Robust dynamic classes revealed by measuring the response function of a social system. PNAS, 105:15649, 2008.
- J. Yang and J. Leskovec. Patterns of temporal variation in online media. In Proc. of the fourth ACM international conference on Web search and data mining, pages 177–186, 2011.
- M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In Proc. 2004 ACM SIGMOD international conference on Management of data, pages 131–142, 2004.

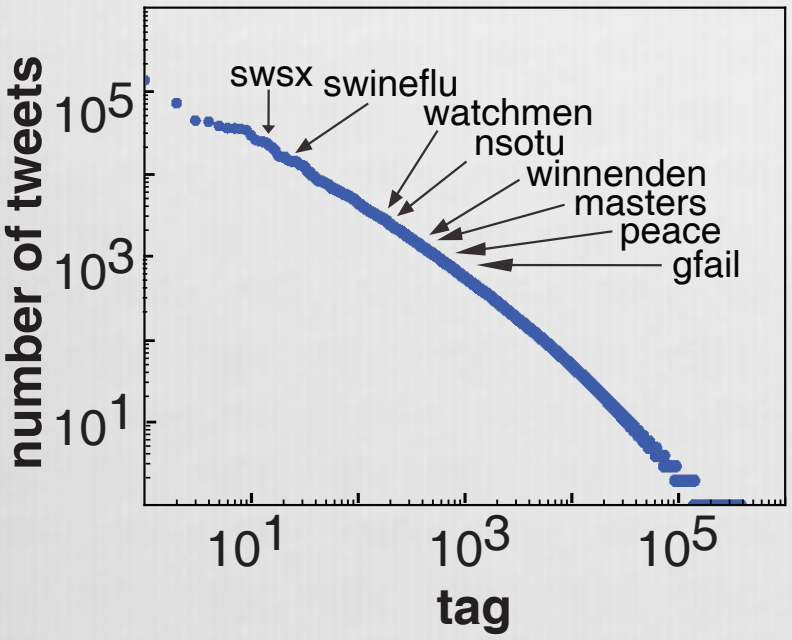
All Hashtags

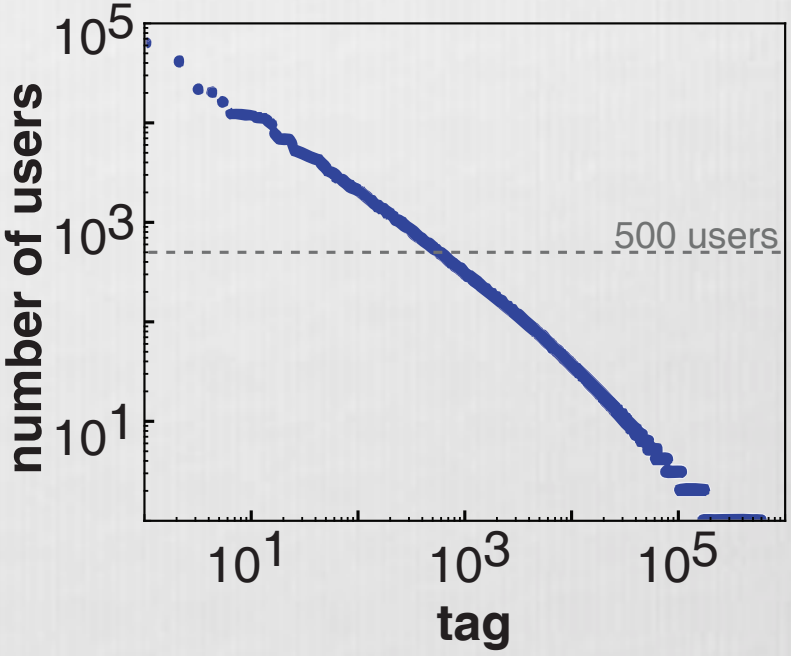
Semantic Grounding



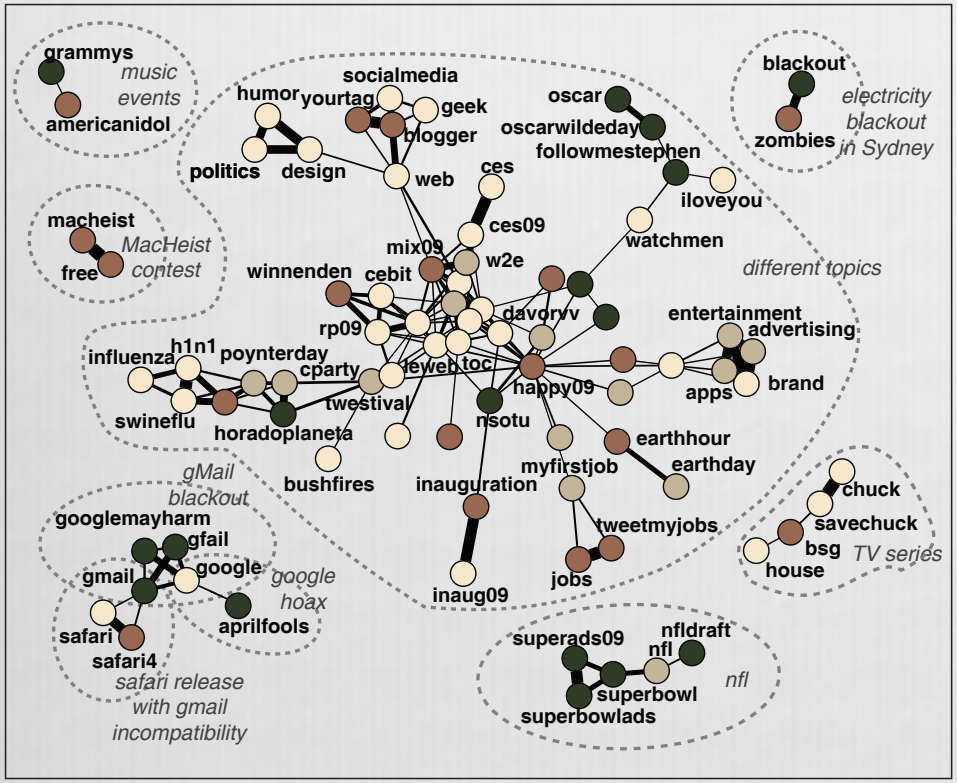
- **Language detection using TextCat**

Hashtag Statistics

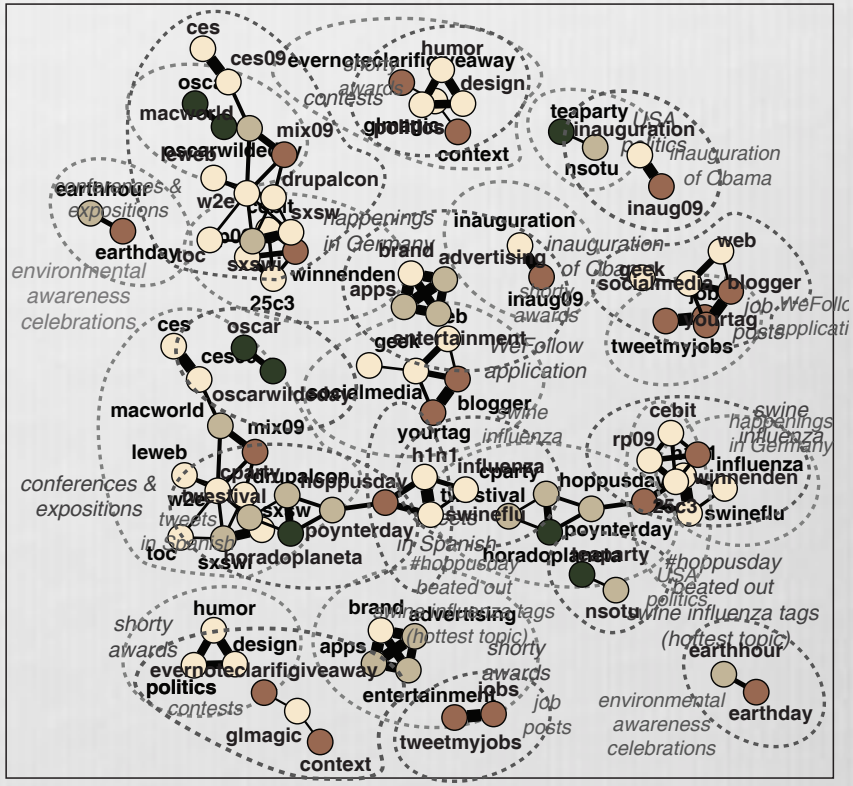




Hashtag Similarities

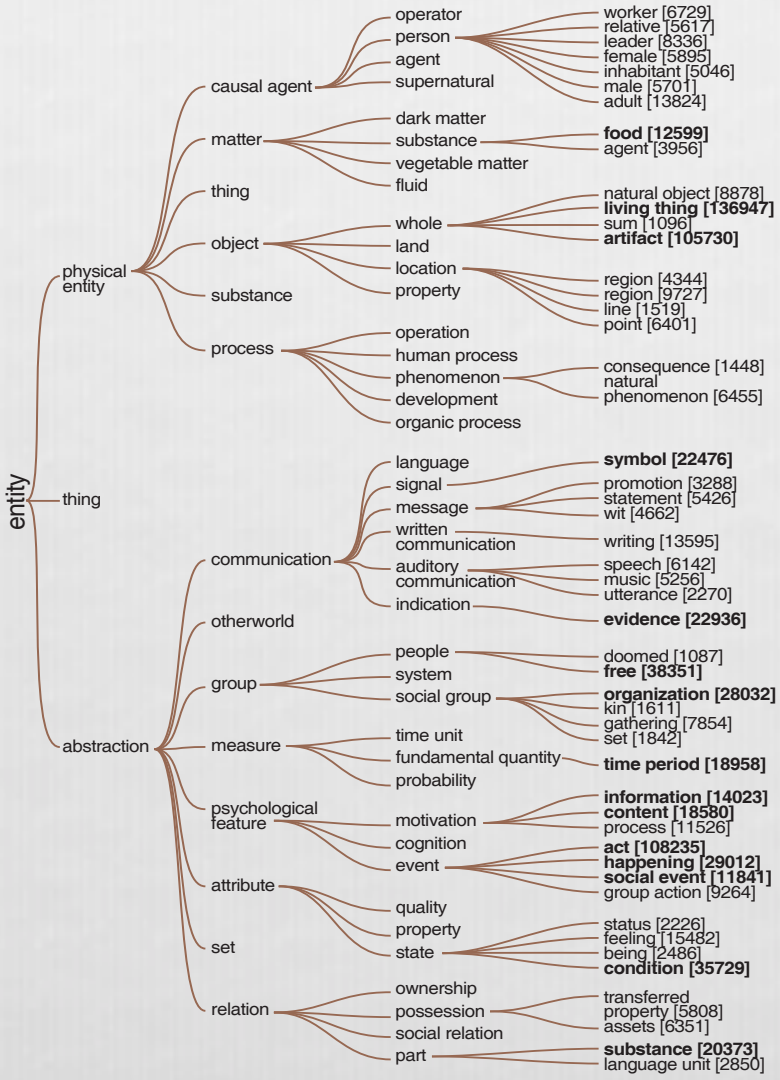


edge weight >= 0.1
(only clusters with more than one node)



edge weight >= 0.2
(only cluster: different topics)

WordNet Tree



About myself

- **Master in Computer Science at the University of Potsdam, Germany**
- ***Wikipedia, Text Mining***

- **Research assistant in ISI, Turin, Italy**
- ***Twitter, Wikipedia***

- **PhD at UPF and Yahoo! Research, Barcelona**
- ***User engagement***