

Contents

2

- Processing Twitter Data
 - ▣ Visualization (Exploratory Data Analysis)
 - ▣ In-depth Analysis
- Tools for analysis
- Interesting sources

Processing Twitter Data

Visualization (Exploratory Data Analysis)

Find more at :

<http://www.slideshare.net/kristw/kristw-hackshackers>

Twitter Data Visualization

4

- Visualization is for story telling, exploratory data analysis and result illustration
- Extracts from twitter data
 - User Who?
 - Text (+media) What?
 - Geo-location Where?
 - Time When?
 - Generator How?
 - Amount How much? (Aggregate Data)
- To visualize, combine these extracts together

User + Time

5

- An interactive timeline based on when your friends started using Twitter

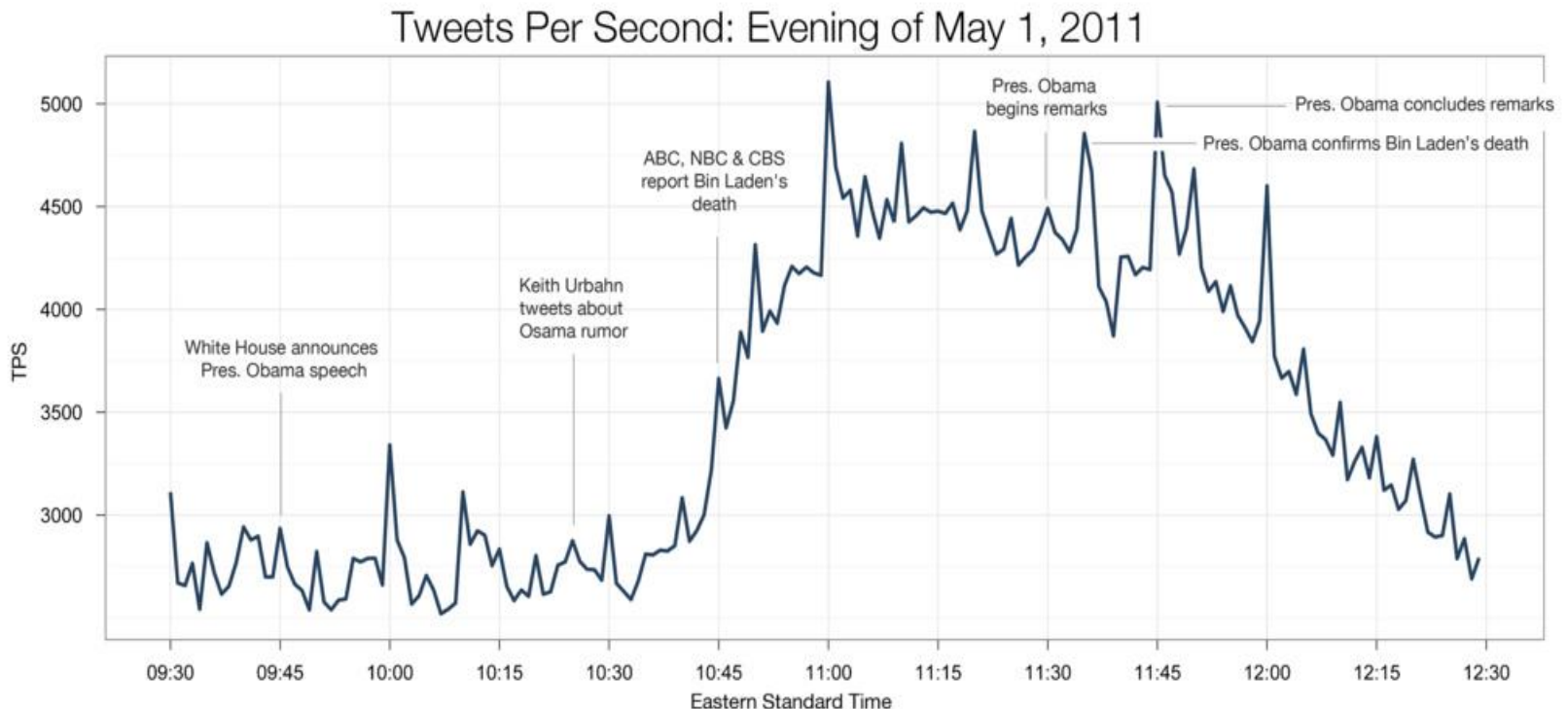


<https://blog.twitter.com/2014/visualizing-your-twitter-conversations>

Time + Amount

6

- A graph of the Tweet activity on the evening of Sunday May 1, 2011.



<https://www.flickr.com/photos/twitteroffice/5681263084>

Geo + Amount

7

- Twitter Heat Map of “f*ck you” and “Good Morning”

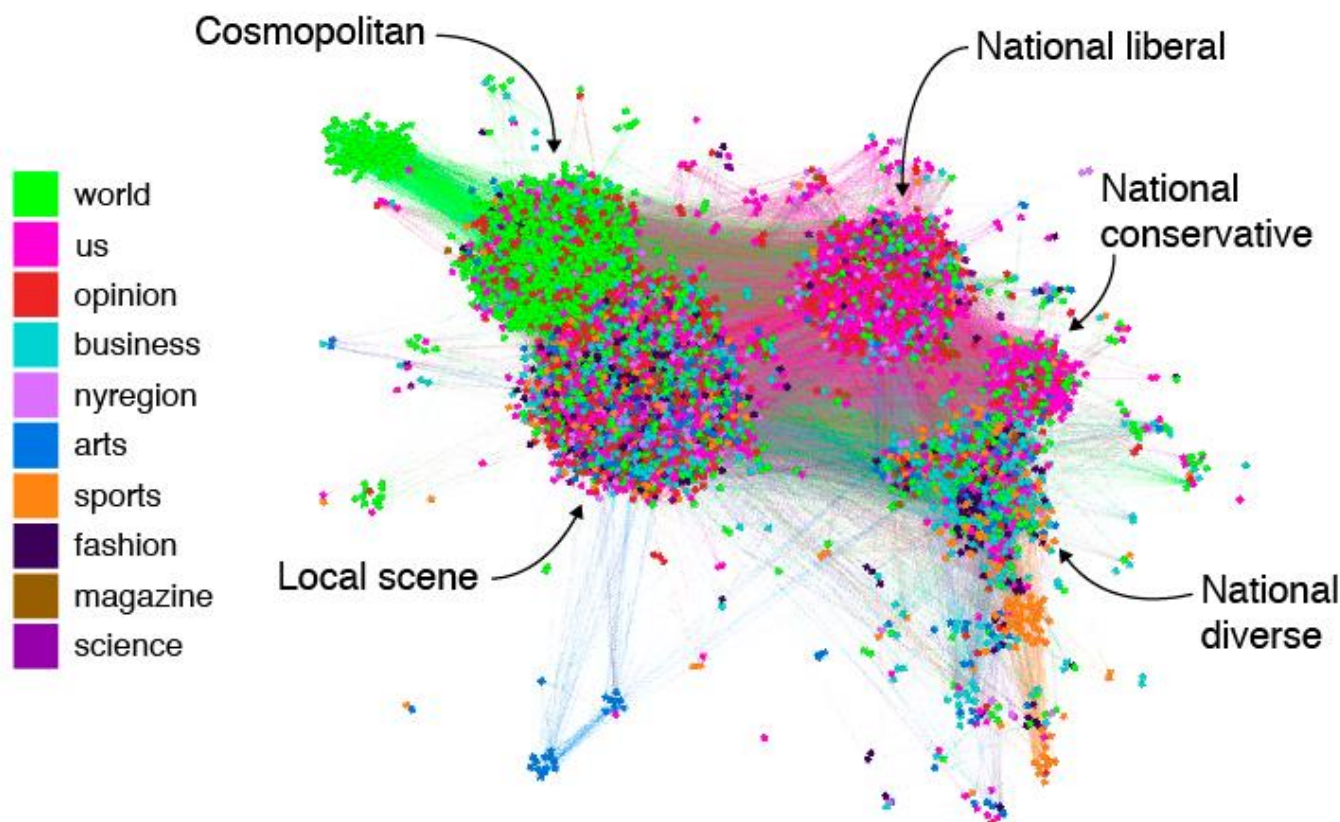


http://www.huffingtonpost.com/2012/08/20/twitter-heatmap-good-morning-fck-you_n_1811065.html

User + Text

8

- While Twitter brings many users together, we typically connect with like-minded souls online

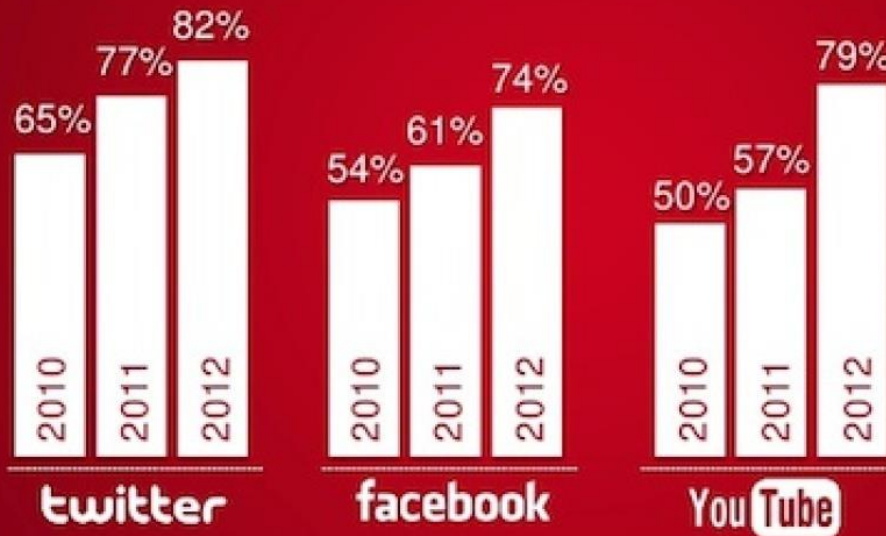


<http://necsi.edu/research/social/nytwitter/nyt.pdf>

User + Amount

Twitter is Most Popular Platform Among Global Companies

Percent of Fortune Global 100 Companies with...



Nearly Half of Companies Have Google+ Accounts



And a Quarter of Companies Have Pinterest Accounts

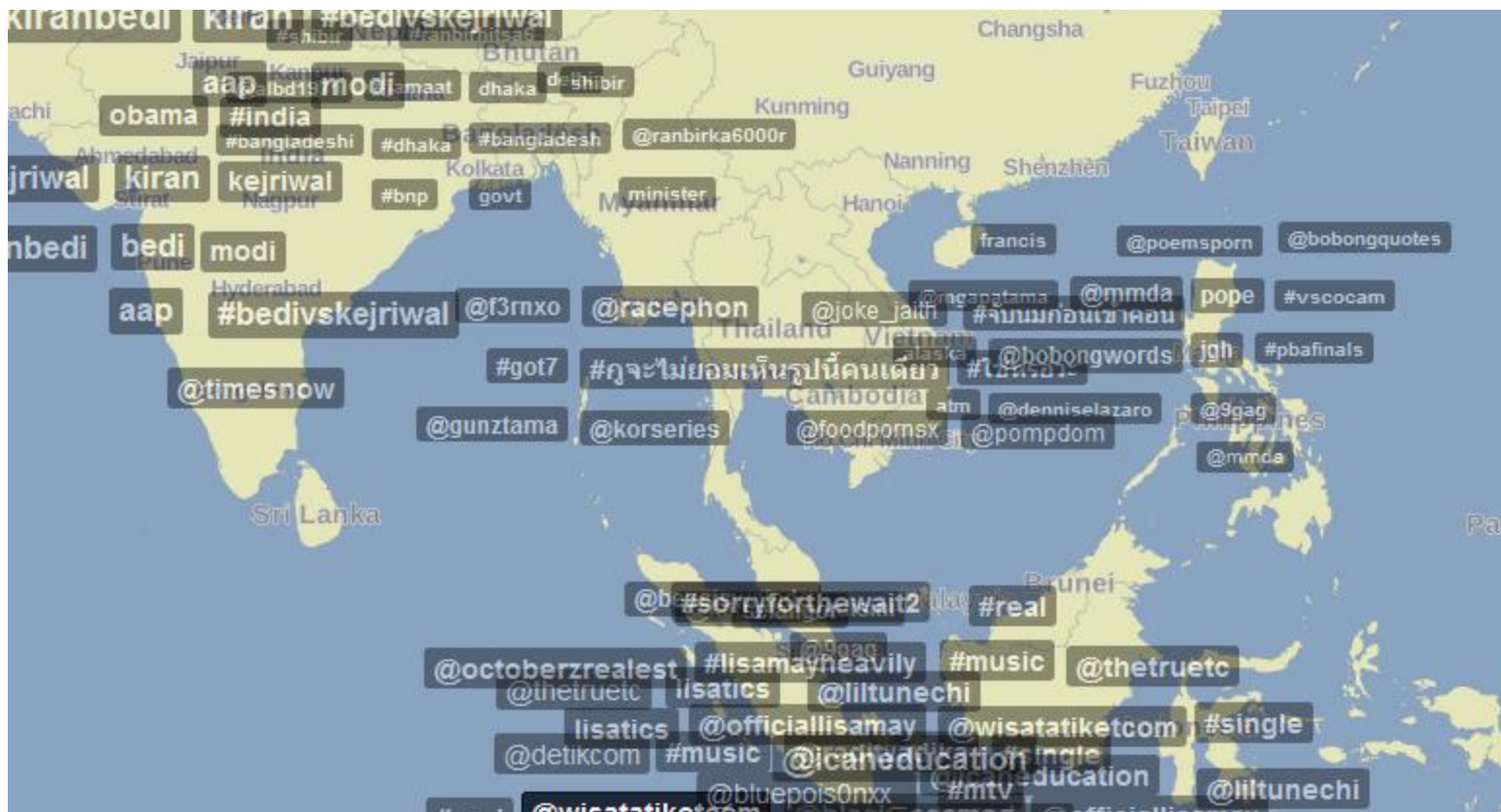


Burson-Marsteller

Geo + Text

11

Real-time Tweet Maps

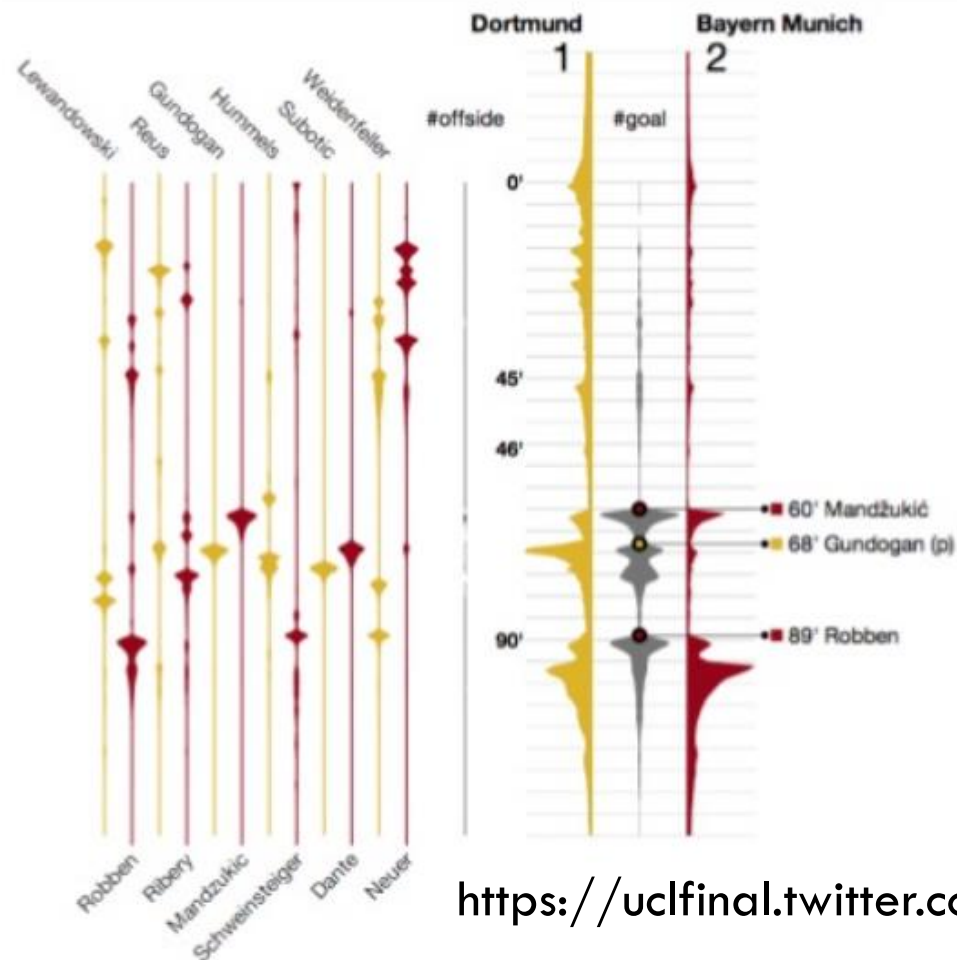


<http://trendsmap.com>

Text + Time + Amount

12

□ UEFA Champion League

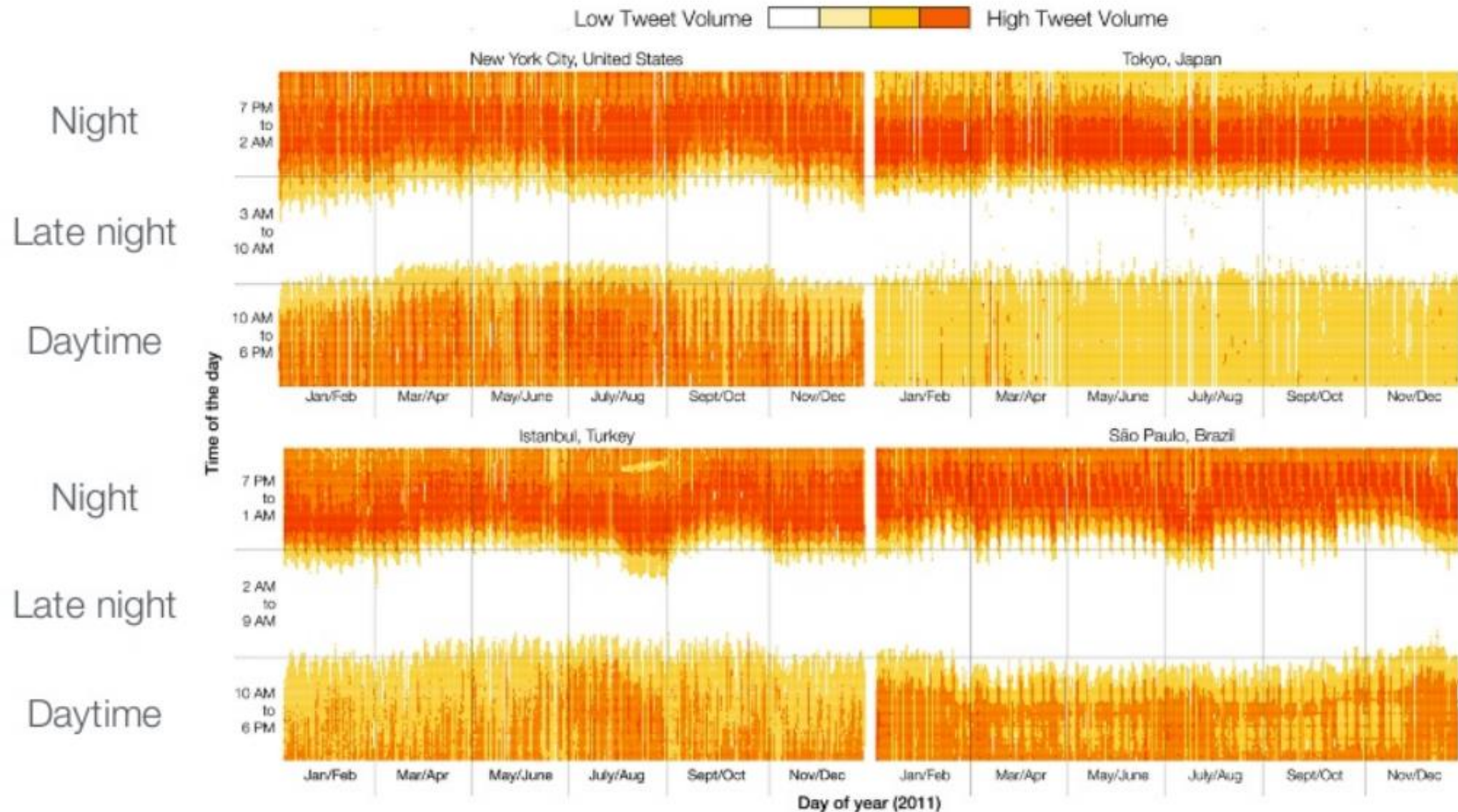


<https://uclfinal.twitter.com/>

Geo + Time + Amount

13

□ Tweet Pattern

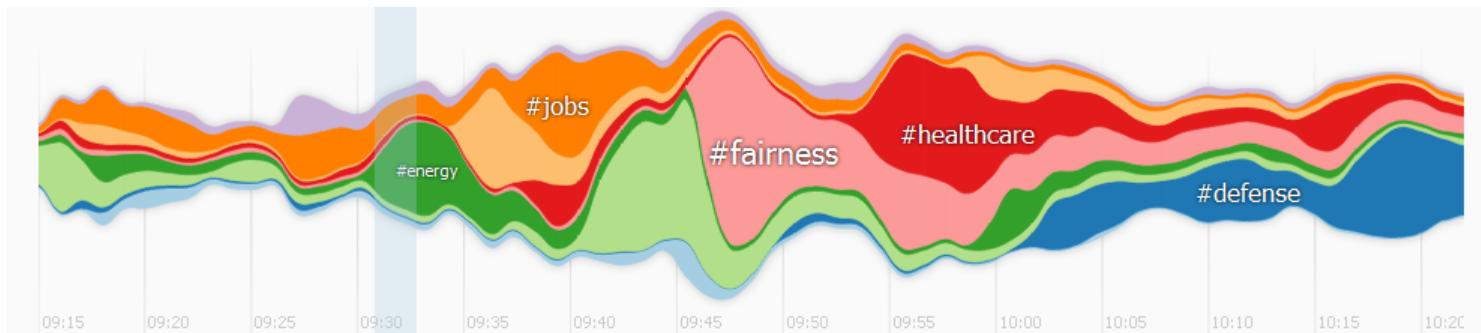


<https://blog.twitter.com/2012/studying-rapidly-evolving-user-interests>

Text + Time + Geo + Amount

14

□ State of The Union 2014

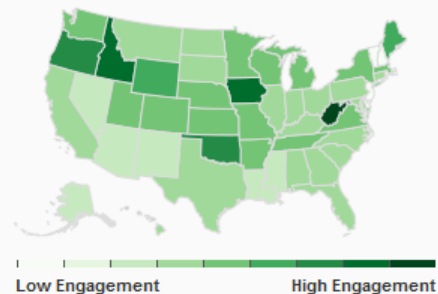


businesses to stay focused on innovation, not costly, needless litigation.

Now, one of the biggest factors in bringing more jobs back is our commitment to American energy. The all-of-the-above energy strategy I announced a few years ago is working, and today, America is closer to energy independence than we've been in decades.

One of the reasons why is natural gas – if extracted safely, it's the bridge fuel that can power our economy with less of the carbon pollution that causes climate change. Businesses plan to invest almost \$100 billion in new factories that use natural gas. I'll cut red tape to help states get those factories built, and this Congress can help by putting people to work building fueling stations that shift more cars and trucks from foreign oil

Real-time engagement distribution on Twitter for this paragraph



Map for #energy

#budget

<http://twitter.github.io/interactive/sotu2014/>

Processing Twitter Data

In-depth Analysis

Find more in :

- *Social Media Mining An Introduction*

By Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu

- *Twitter Data Analytics*

By Shamanth Kumar, Fred Morstatter, and Huan Liu

Social Media Mining (1)

16

- There are three groups of questions we want to answer
- Group 1: General Activities
 - ▣ Who are the most important people in a social network?
 - ▣ How do people befriend others?
 - ▣ How can we find interesting patterns in user-generated content?

Social Media Mining (2)

17

- **Group2: Communities and Interactions**
 - ▣ How can we identify communities in a social network?
 - ▣ When someone posts an interesting article on a social network, how far can the article be transmitted in that network?
- **Group3: Real-world problems**
 - ▣ How can we measure the influence of individuals in a social network?
 - ▣ How can we recommend content or friends to individuals online?
 - ▣ How can we analyze the behavior of individuals online?

Twitter Analysis

18

- Text Measures
 - Trending Topics
 - Sentimental Analysis
- Network Measures
 - User Influence
 - User Behavior

Trending Topics

19

□ Count occurrences of Specific Words



<http://yearinreview.twitter.com/en/hottopics.html>

Latent Dirichlet Allocation (LDA)

20

- Every topic in LDA is a collection of words
- Each topic contains all of the words in the corpus with a probability of the word belonging to that topic.
- For example,
 - ▣ Sports 40% “basketball”, 35% “football”, 15% “baseball”,
 ..., 0.02% “congress”, and 0.01% “Obama”
 - ▣ Politics 35% “congress”, 30% “Obama”, ..., 1% “football”, 0.1%
 “baseball”, 0.1% “basketball”
- LDA finds the most probable words for a topic, associating each topic with a theme is left to the user

Preprocessing before LDA

21

- In order to using MALLET library in JAVA for LDA, we have to preprocess data with these five steps

Step	Products
0. Raw Data	No more media blackout hiding #OCCUPYWALLSTREET! :)
1. Lowercase	no more media blackout hiding #occupywallstreet! :)
2. Tokenize	[no, more, media, blackout, hiding, #occupywallstreet]
3. Stopword Removal	[no, media, blackout, hiding, #occupywallstreet]
4. Stemming	[no, media, blackout, hide, #occupywallstreet]
5. Vectorization	a vector that contains a sequence of numbers for each word in the vocabulary

Typology of Trending Topics [1]

22

- **News**
- **Ongoing events:** real-time information sharing
 - E.g. A soccer game, A keynote presentation by Apple
- **Memes:** triggered by viral ideas initiated by either an individual or an organization
 - E.g. Ice Bucket Challenge
- **Commemoratives:** the commemoration of certain person or event that is being remembered in a given day
 - E.g. New Year, Father Day, PrincessDiana

[1] Arkaitz Zubiaga et.al., Real-Time Classification of Twitter Trends, Journal of the American Society for Information Science and Technology 2013

Sentimental Analysis

23

- “Sentiment analysis” seeks to automatically associate a piece of text with a “sentiment score”, a positive or negative emotional score
- Using natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

Sentimental Analysis Approaches

24

- Existing approaches to sentiment analysis can be grouped into four main categories ^[1]
 - ▣ Keyword spotting: based on the presence of unambiguous affect words such as happy, sad, afraid, and bored
 - ▣ Lexical affinity: not only detects obvious affect words, it also assigns arbitrary words a probable “affinity” to particular emotions
 - ▣ Statistical methods: leverage on elements from machine learning such as latent semantic analysis, support vector machines, "bag of words" and *Semantic Orientation*
 - ▣ Concept-level techniques: leverage on elements from knowledge representation such as ontologies and semantic networks

[1] http://en.wikipedia.org/wiki/Sentiment_analysis

Dictionary-based Approach

25

- Sentiment analysis framework using dictionary-based approach
 - ▣ There are words together with its sentimental score in the specific dictionary
 - ▣ Apply Porter stemmer to dictionary terms and tweets
 - ▣ Compute Value $[1,9]$ and then minus 5
 - ▣ Words not contained in the dictionary \rightarrow neutral
 - ▣ Total Score = Sum of the score from each word in each metric

Naïve Bayes Approach (1)

26

- Sentiment analysis framework using Naïve Bayes Classification
 - ▣ Enumerating each Tweet in the dataset
 - ▣ Building a lexicon from the Tweets that use an emoticon
 - ▣ Calculating a sentiment score for each Tweet that does not have an emoticon

Sentimental Scale Visualization

28

- A graph showing sentimental tendency of tweets containing a word “tea”



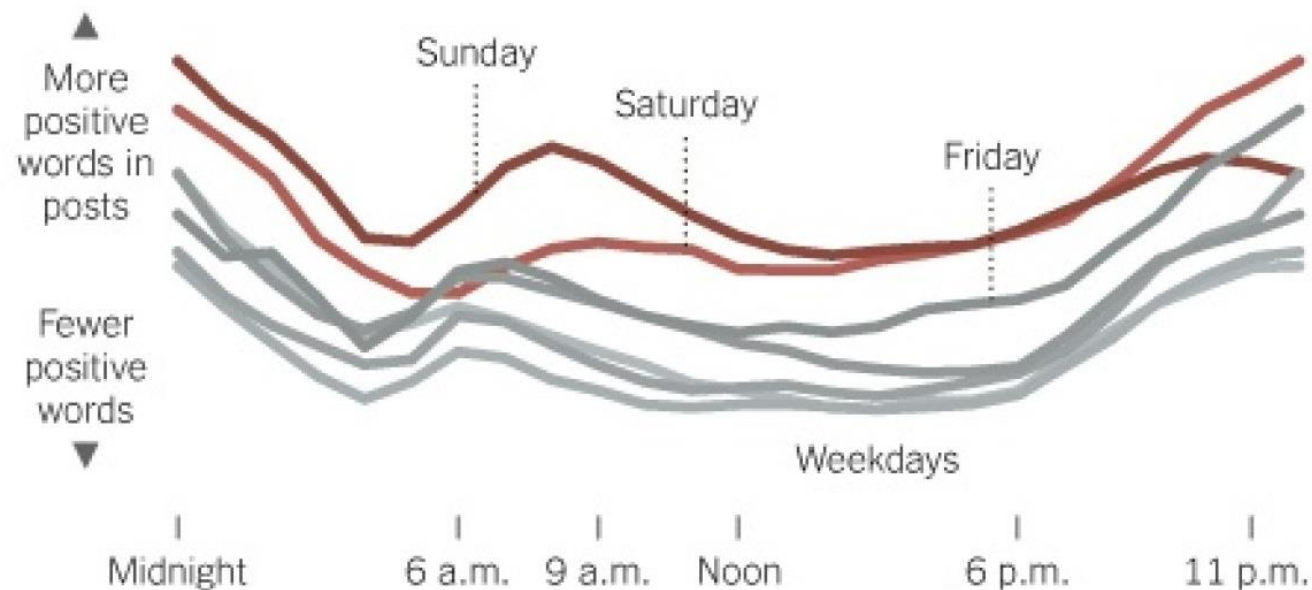
http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Sentiment + Time

29

Studying Moods Through Twitter

A textual analysis of more than 500 million Twitter messages found people around the world tend to express more positive emotions in the morning and evening, and are most positive on weekends. The recurring daily pattern suggests moods are influenced by sleep and circadian rhythms.



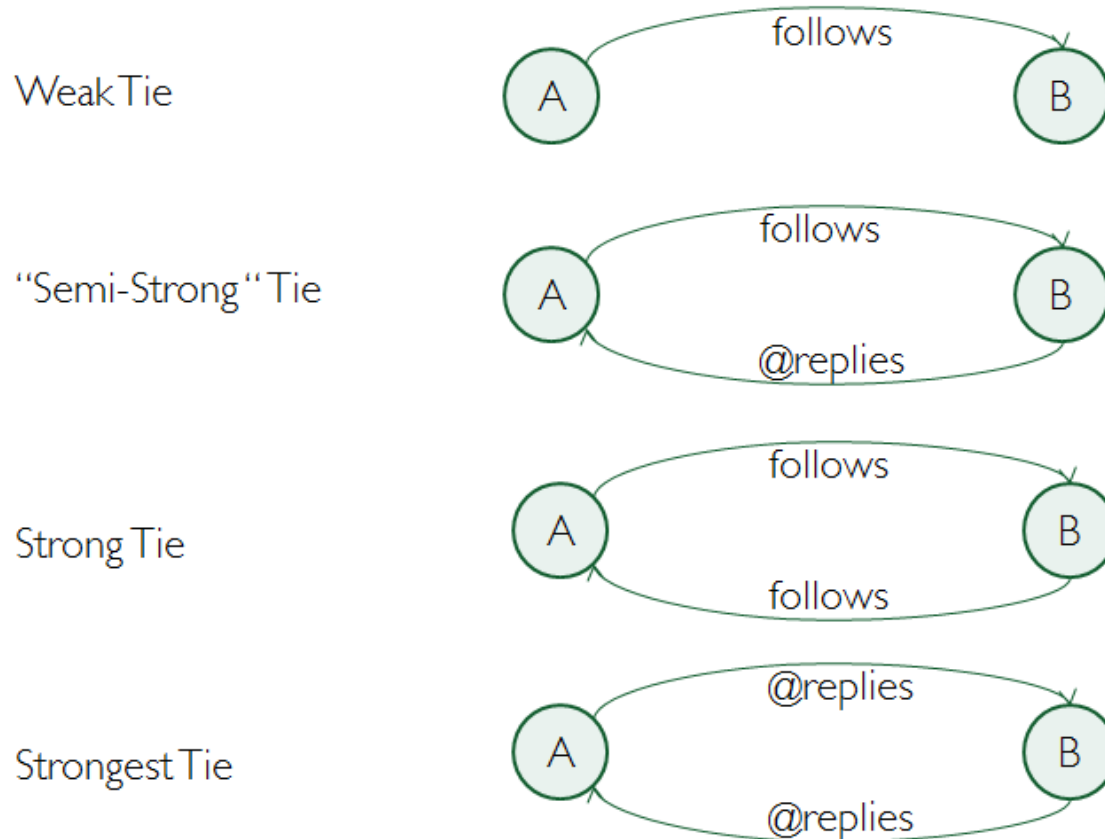
Source: Science

THE NEW YORK TIMES

http://www.nytimes.com/2011/09/30/science/30twitter.html?_r=0

Tie Strength in Twitter

30



Networks from Twitter Data

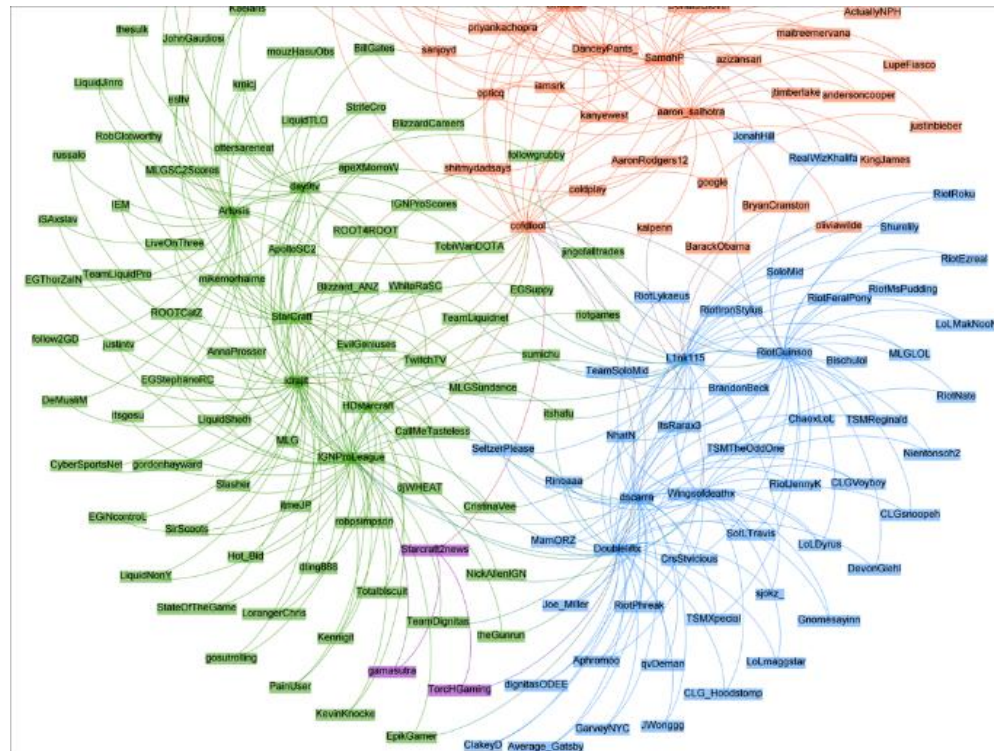
31

- Interest Graph friend – follower
 (Twitter Social Graph)
- Conversation Graph mention (reply)
- Retweet Graph retweet

Twitter Social Graph

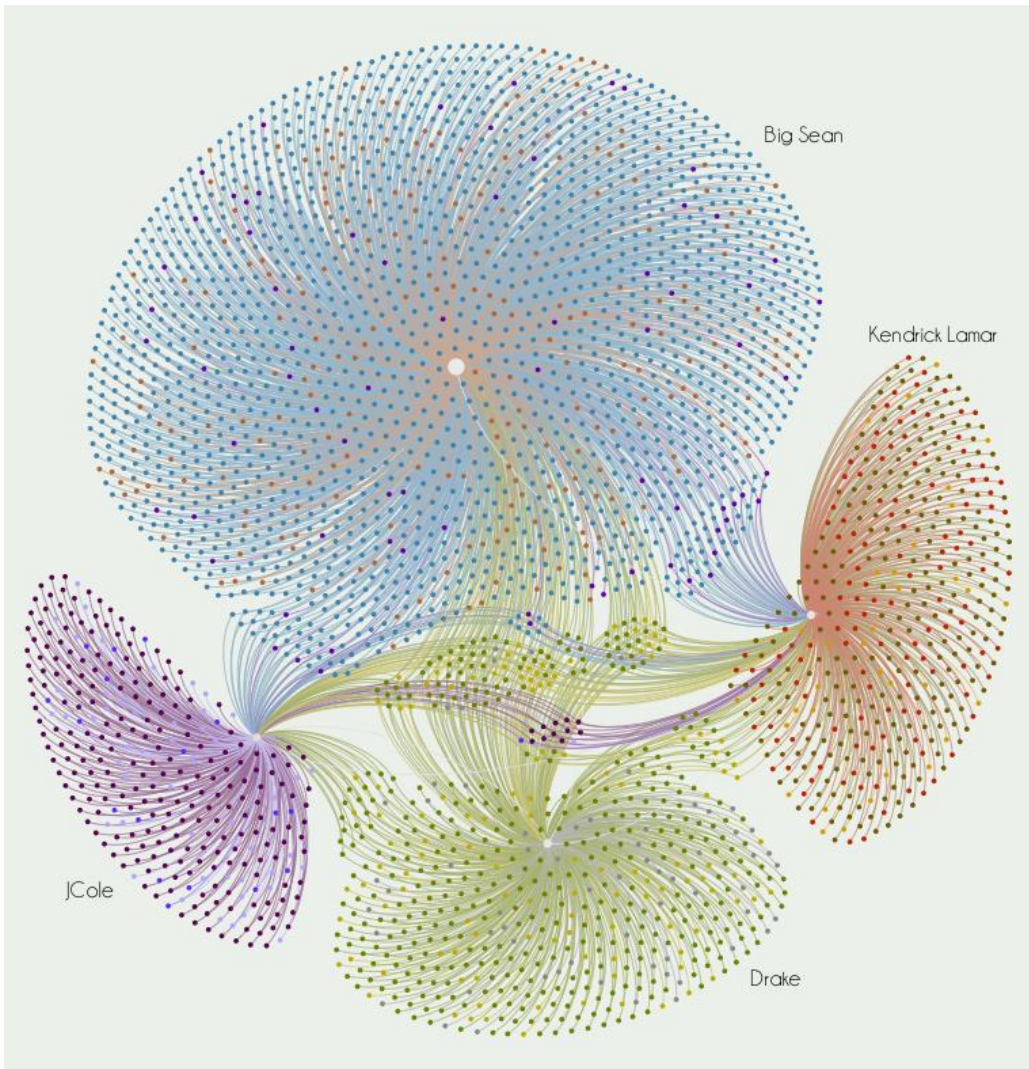
32

- Try to find independent communities within a graph; assign modularity score based on connections from individual nodes to “hub” nodes (gephi)



Conversation Graph

33

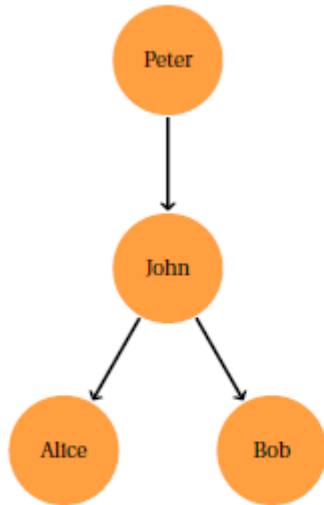


- From 3000 tweets for 4 rappers (Drake, Kendrick Lamar, J Cole, and Big Sean)
- Created By Achal Soni (Gephi)

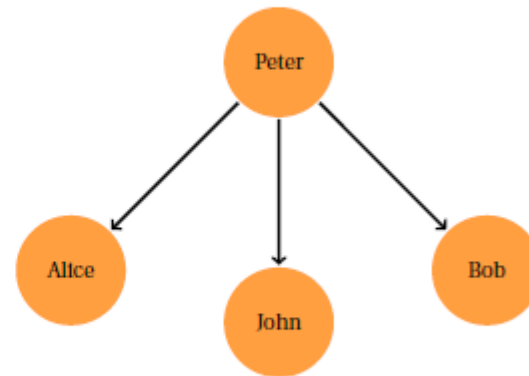
Retweet Graph

34

- One can only identify the original source of the information and not the intermediate users along the information propagation path.



(a) Actual propagation path



(b) Path extracted from Twitter API

Network Measures

35

- Centrality
 - ▣ How important a node is within a network
 - ▣ User Influence
- Transitivity and Reciprocity
 - ▣ How links (edges) are formed in a social graph
 - ▣ Link Prediction
- Similarity (Structural, Regular)
 - ▣ Compute similarity between two nodes in a network
 - ▣ Community Analysis, Behavior Prediction

Degree Centrality

36

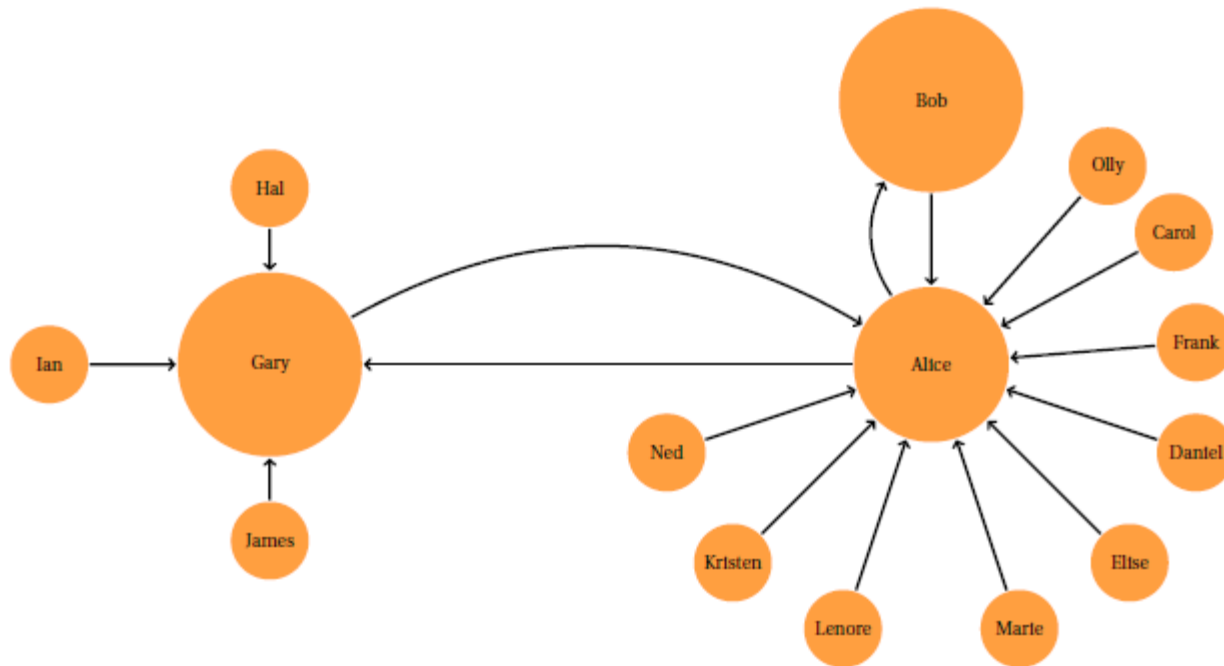
- Count the number of links attached to the node
- The key question was “how many people retweeted this node?”



Eigenvalue Centrality

37

- Eigenvector Centrality builds upon this to ask “how important are these retweeters?”



Centrality Measures

38

- Degree Centrality
- Eigenvector Centrality
- Katz Centrality
- PageRank
- Betweenness Centrality
- Closeness Centrality
- Group Centrality

Collaborative Filtering

39



Memory-based Approach

40

□ E.g. Movie Ratings

User-based Filtering

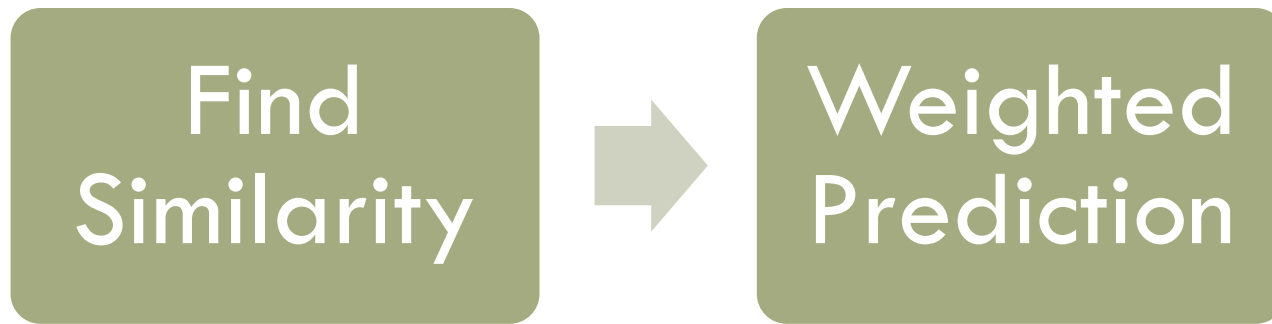
	Amy	Jeff	Mike	Chris	Ken
The Piano	-	-	+		+
Pulp Fiction	-	+	+	-	+
Cliffhanger	+		-	+	-
Fargo	-	+	+	-	+

Item-based Filtering

Memory-based Approach

41

- A prediction is normally based on the weighted average of the recommendations of several people.



Tools for analysis

Find more at :

http://en.wikipedia.org/wiki/Social_network_analysis_software

Mining Twitter with R (1)

43

- Package “twitterR” (R based Twitter client) provides an interface to the Twitter web API

Function	Short Description
<code>decode_short_url</code>	A function to decode shortened URLs
<code>favorites</code>	A function to get favorite tweets
<code>friendships</code>	A function to detail relations between yourself & other users
<code>getCurRateLimitInfo</code>	A function to retrieve current rate limit information
<code>getTrends</code>	Functions to view Twitter trends
<code>registerTwitterOAuth</code>	Register OAuth credentials to twitter R session
<code>twListToDF</code>	A function to convert twitterR lists to data.frames

<http://cran.r-project.org/web/packages/twitterR/twitterR.pdf>

Mining Twitter with R (2)

44

- The examples of other useful packages for text mining using R

```
library(tm)           # Framework for text mining.
library(SnowballC)   # Provides wordStem() for stemming.
library(qdap)        # Quantitative discourse analysis of transcripts.
library(qdapDictionaries)
library(dplyr)       # Data preparation and pipes %>%.
library(RColorBrewer) # Generate palette of colours for plots.
library(ggplot2)     # Plot word frequencies.
library(scales)      # Include commas in numbers.
library(Rgraphviz)   # Correlation plots.
```

NodeXL (1)

45

- Network Overview Discovery Exploration for Excel
- A free and open-source network analysis and visualization software package for Microsoft Excel 2007/2010
- Intended for users with little or no programming experience to allow them to collect, analyze, and visualize a variety of networks

NodeXL (2)

46

The screenshot displays the Microsoft Excel interface with the NodeXL add-in. The main window shows a network graph with nodes and edges, overlaid on an Excel spreadsheet. The spreadsheet columns are labeled with graph metrics: Degree, In-Degree, Out-Degree, Betweenness Centrality, Closeness Centrality, Eigenvector Centrality, PageRank, Clustering Coefficient, Reciprocated Vertex Pair Ratio, and Add Your Own Columns. The graph is divided into several clusters, each labeled with a G-number and a description of the cluster's content. The clusters are:

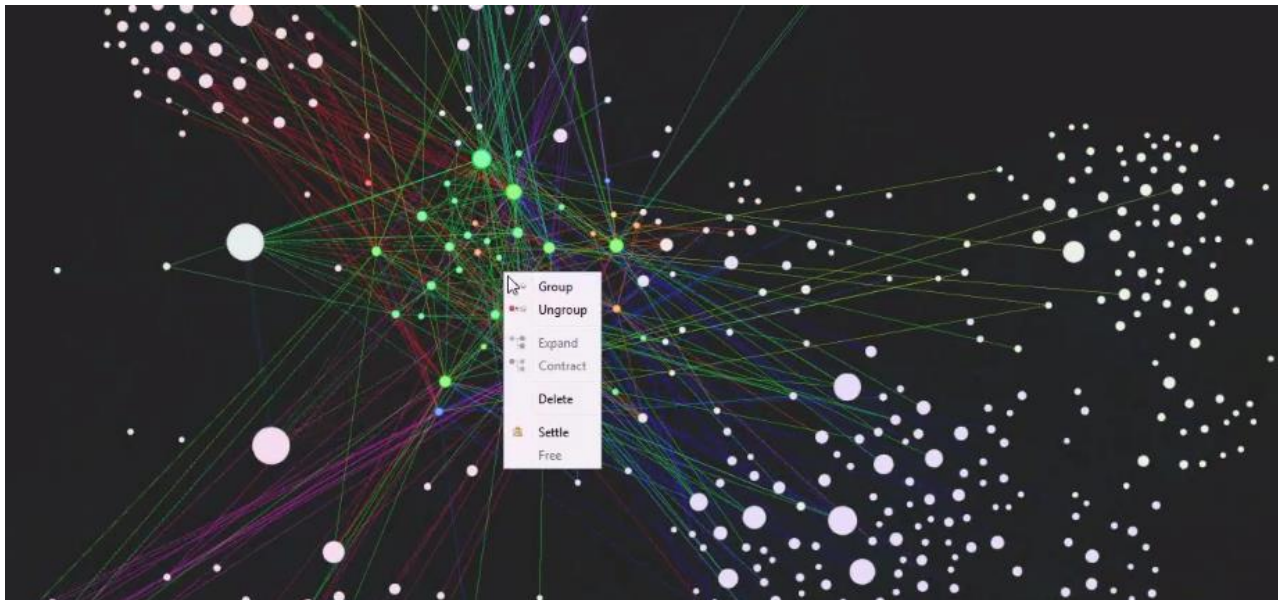
- G2: pewinternet use online adults 18 twitter facebook 71 17 instagram
- G3: susannahfox pewresearch online adults use s 91 u cell phone
- G4: pew internet releases libraryjournal americans new libraries value public communities
- G5: usage l des les étude pew internet 5 enseignements réseaux
- G6: pew internet scoopit
- G7: use adults online 2morrowknight sept '13 71 18 twitter 22
- G8: mobile social irainie needle thread humans mantra cl
- G10: sai pew internet

The interface includes a ribbon with tabs for File, Home, Insert, Page Layout, Formulas, Data, Review, View, Developer, NodeXL, and Design. The NodeXL ribbon contains various toolbars for graph manipulation, such as Refresh Graph, Summary, Automate, Color, Vertex Shape, Vertex Size, Edge Width, Dynamic Filters, Graph Metrics, Subgraph Images, Groups, Import, Export, and Workbook Columns. The spreadsheet shows data for various users, including pewinternet, scoopit, libraryjournal, frouellet, tinkuy, stephane_ozil, _ig_up_, pewresearch, viuzfr, queerspawn, and cynthiajabar.

Gephi (<https://gephi.github.io/>)

47

- An open-source network analysis and visualization software package written in Java on the NetBeans platform
- See video: <http://vimeo.com/9726202>



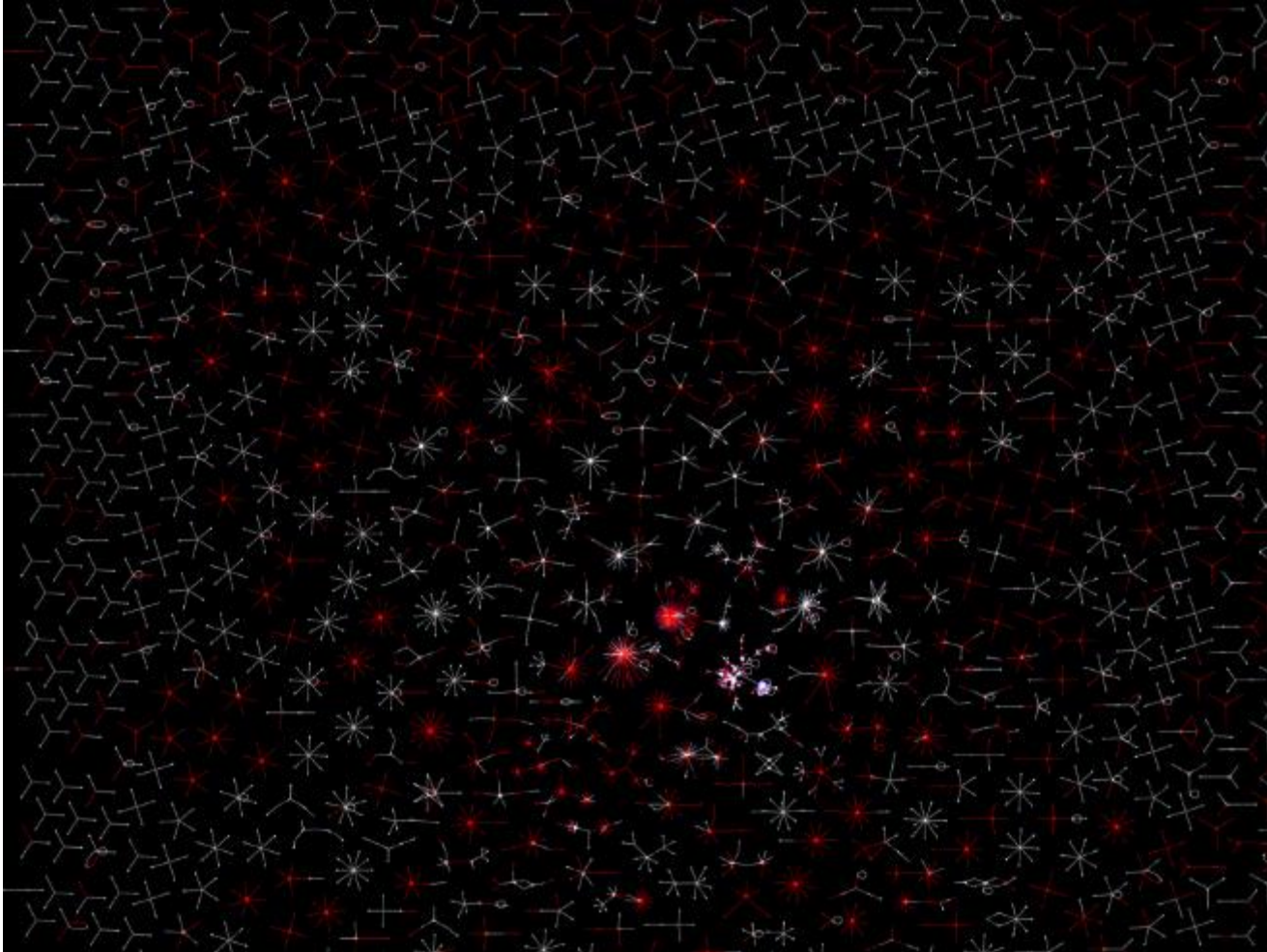
Graphviz (www.graphviz.org)

48

- An open source graph visualization software
- A simple text language → Diagrams
- Output formats e.g. images and SVG for web pages; PDF or Postscript for inclusion in other documents; or display in an interactive graph browser
- Useful features for concrete diagrams, such as options for colors, fonts, tabular node layouts, line styles, hyperlinks, and custom shapes.

Graphviz (www.graphviz.org)

49



50

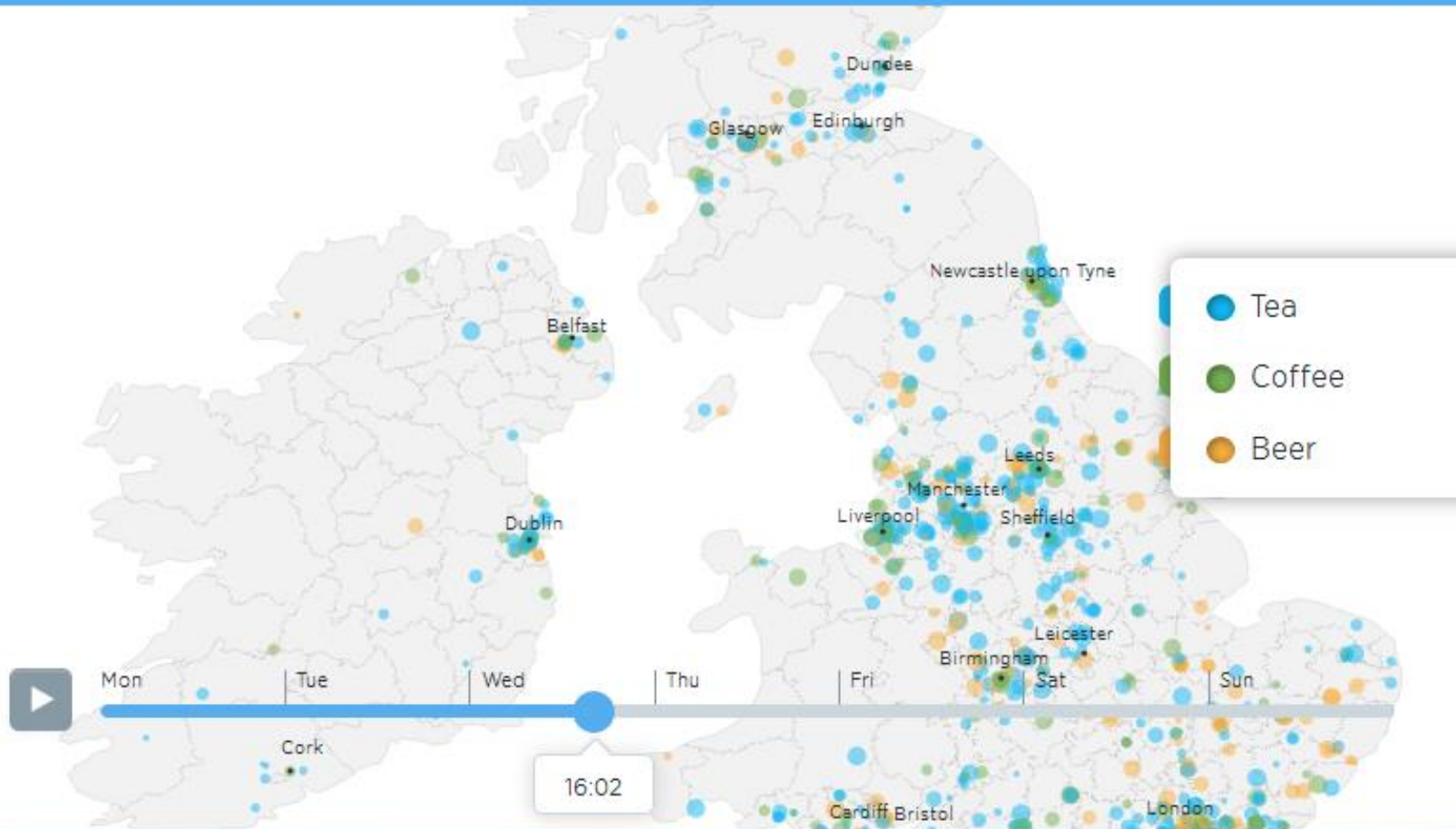
Interesting sources

Moments.twitter.com/uki/

51

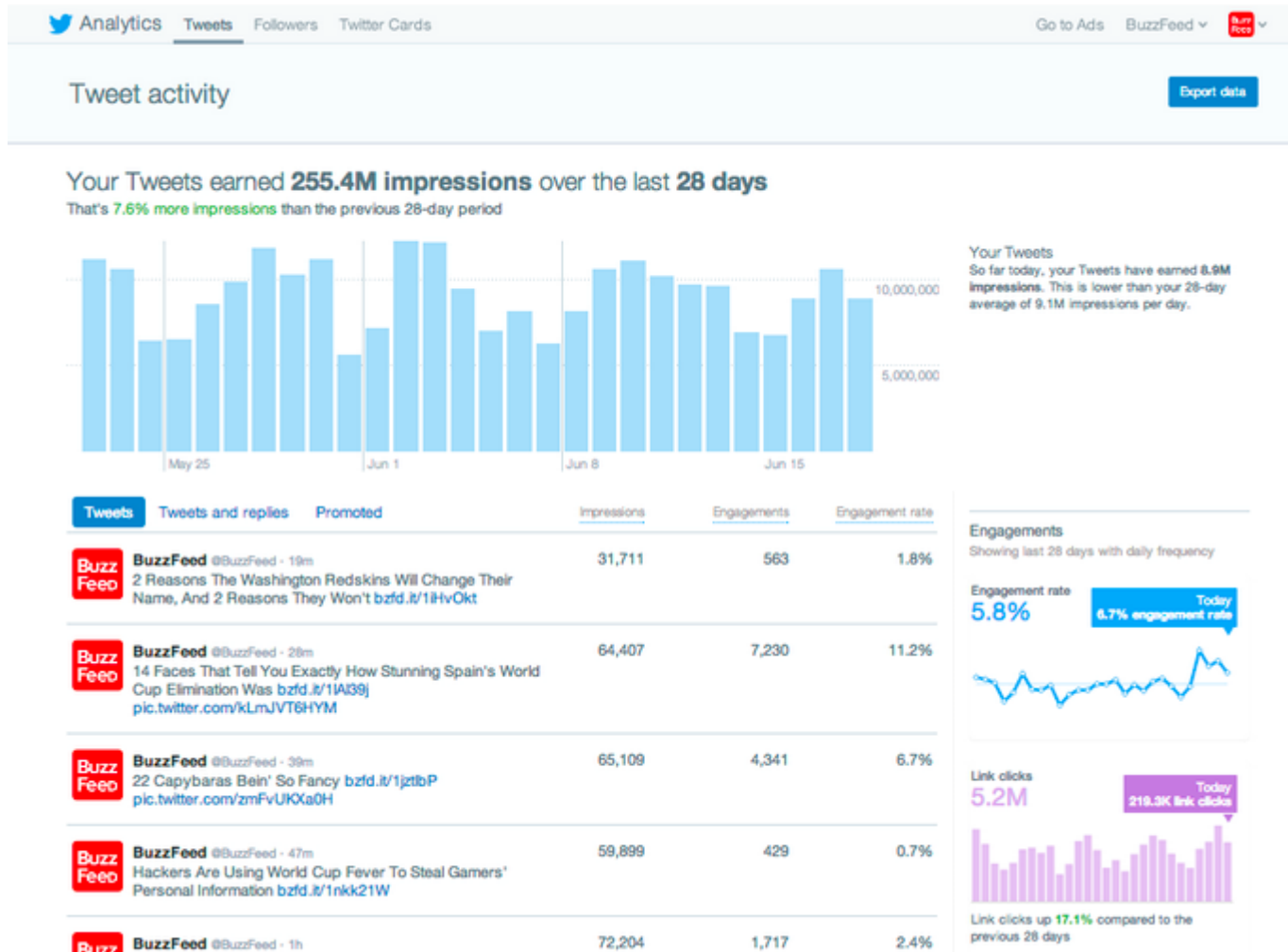


#EverydayMoments Beta



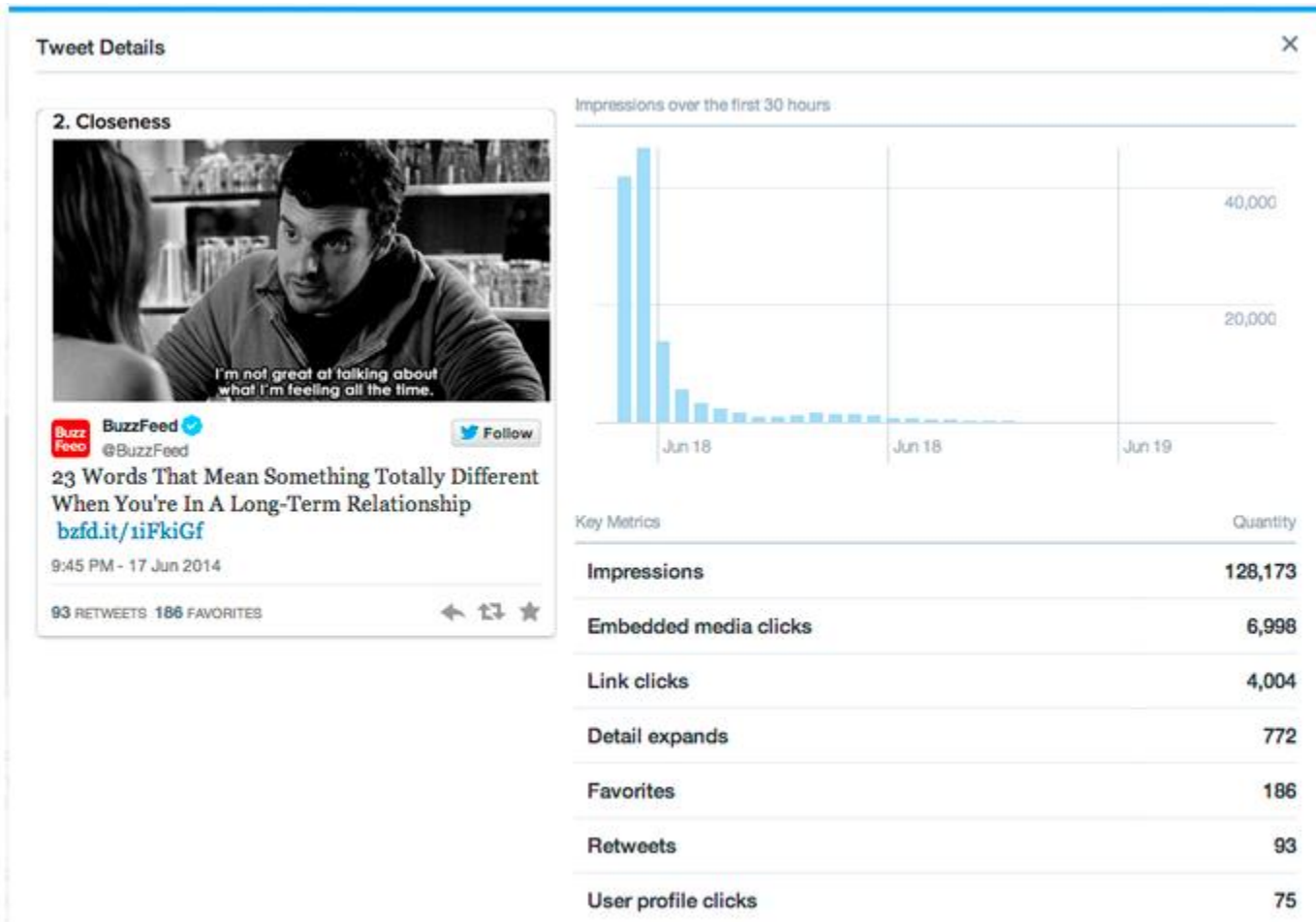
Analytics.twitter.com

52



Analytics.twitter.com

53



<https://blog.twitter.com/2014/new-tweet-activity-dashboard-offers-richer-analytics>

The Official Twitter Blog

Keeping you connected to everything from Twitter.

Big night for #SOTU on Twitter

Wednesday, January 21, 2015 | By Bridget Coyne (@bcoyne), Government & Elections Team 01/21/2015 - 04:25

Tags: [civic](#), [live events](#), [politics](#), and [Twitter data](#)

A look back at the real-time State of the Union conversation. [Read more...](#)

#NFL Conference Championships recap

Tuesday, January 20, 2015 | By Brian Poliakoff (@brianpoliakoff), Sports Communications Manager 01/20/2015 - 19:04

Tags: [live events](#) and [sports](#)

How the #NFL conference championships played out on Twitter. [Read more...](#)



The Official Twitter Blog

Your source for company news, stories and updates.



Media

Tracking how Twitter is used in TV, sports, music, government, news and more.



Advertising

Your official source for Twitter Ads product updates, tips, events and success stories.



Engineering

Information from Twitter's engineering team about our technology, tools and services.



Developers

Connecting the Twitter developer community through best practices and tutorials.

Interactive.twitter.com

56

- <http://interactive.twitter.com>
- <http://twitter.github.io/interactive/>



#FRACTILE: SPACE-FILLING VINES




#NBA: WHERE ARE YOUR TEAM'S FOLLOWERS?

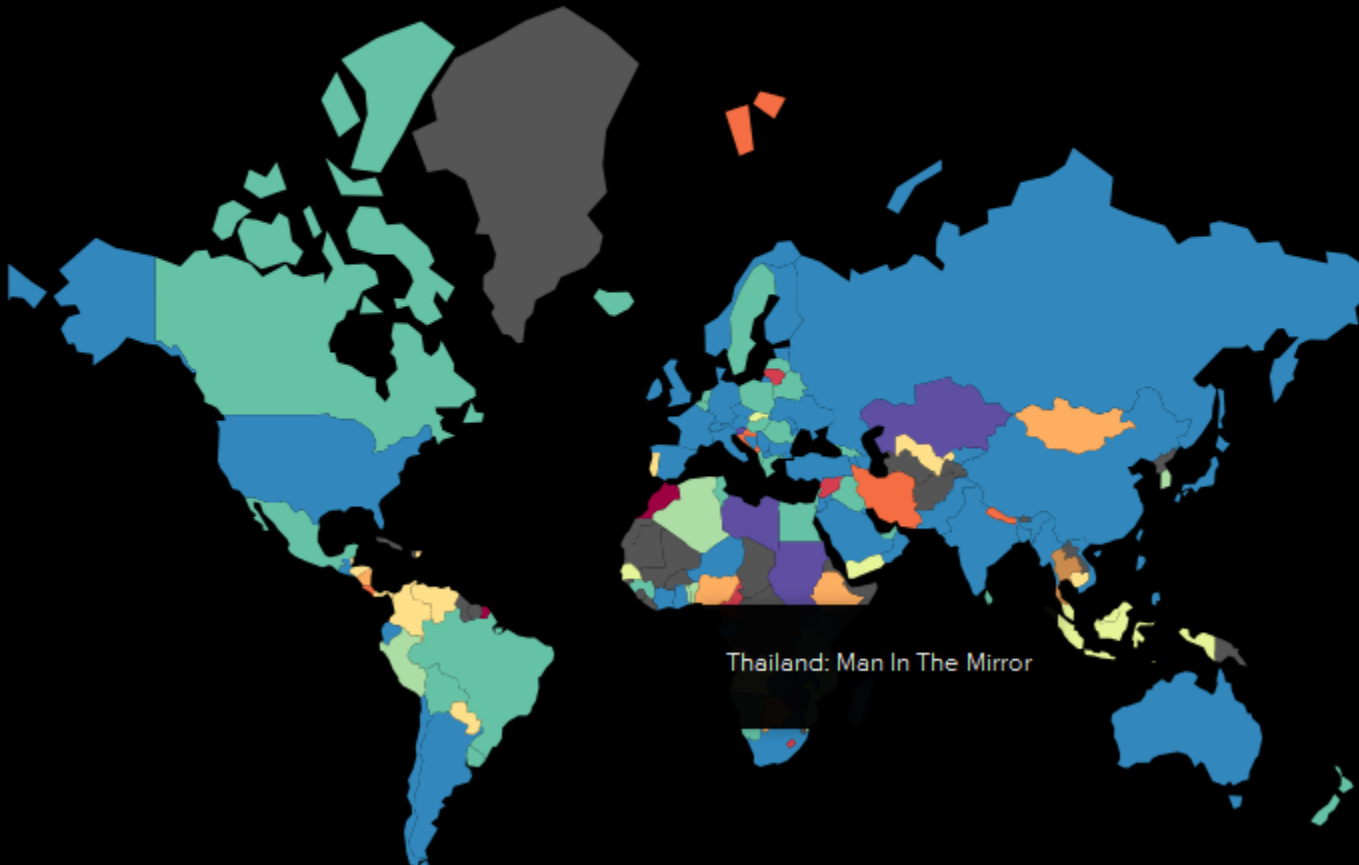


Interactive.twitter.com

57

 The all-time top-tweeted Michael Jackson songs

@MichaelJackson's top-tweeted singles around the world



Analyzing Big Data With Twitter

58

- Special course in Fall 2012 from UC Berkeley School of Informatics by Marti Hearst
- Cooperating with Twitter Inc.
- Taught Topics
 - ▣ Twitter Philosophy; Twitter Software Ecosystem
 - ▣ Using Hadoop and Pig at Twitter
 - ▣ The Twitter API
 - ▣ Trend Detection in Twitter's Streams
 - ▣ Real-time Twitter Search
 - ▣ Correlating Twitter Data with Other Data
 - ▣ Graph Algorithms for the Twitter Social

Analyzing Big Data With Twitter

59

- Taught Topics (Cont.)
 - ▣ GraphLab: Big Learning with Graphs
 - ▣ Large-scale Anomaly Detection at Twitter
 - ▣ Recommendation Algorithms at Twitter
 - ▣ Security at Twitter
 - ▣ Information Diffusion and Outbreak Detection at Twitter
 - ▣ Etc.
- Find more on the course webpage
<http://blogs.ischool.berkeley.edu/i290-abdt-s12/>
- Youtube Playlist of the lectures
<https://www.youtube.com/playlist?list=PLE8C1256A28C1487F>

Bibliography of Research on Twitter & Microblogging

60

□ <http://www.danah.org/researchBibs/twitter.php>

Bibliography of Research on Twitter & Microblogging

1. Jennifer Golbeck, and Derek Hansen. (2011). **Computing political preference among twitter followers.** *Proceedings of CHI 2011.* ACM. (conference paper)
2. Aditi gupta, hemank Lamba, and Ponnurangam Kumaraguru. (2013). **\$1.00 per RT #BostonMarathon #PrayForBoston: Analyzing Fake Content on Twitter.** *Eigth IEEE APWG eCrime Research Summit (eCRS).* (pp. 12). (conference paper)
3. Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. (2013). **Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy.** *Proceedings of the 22nd international conference on World Wide Web companion.* (pp. 8). (conference paper)
4. Aditi Gupta, and Ponnurangam Kumaraguru. (2012). **Credibility Ranking of Tweets during High Impact Events.** *Workshop on Privacy and Security in Online Social Media Co-located with WWW 2012.* (conference paper)
5. Adrien Guille, and Cécile Favre. (2014). **Mention-anomaly-based Event Detection and Tracking in Twitter.** *Proceedings of the IEEE/ACM International Conference on Advances in Social Network Analysis and Mining.*
6. Adrien Guille, and Hakim Hacid. (2012). **A Predictive Model for the Temporal Dynamics of Information Diffusion in Online Social Networks.** *International Workshop on Mining Social Network Dynamics at the 21st World Wide Web Conference (WWW 2012) .*
7. Ampofo, Lawrence, Anstead, Nick, and O'Loughlin, Ben. (2011). **Trust, confidence, credibility: Citizen responses on Twitter to opinion polls during the 2010 UK general election.** *Information, Communication & Society, 14(6), 850-871.* (journal article)
8. Anja Rudat, Jürgen Buder, and Friedrich W. Hesse. (2014). **Audience design in Twitter: Retweeting behavior between informational value and followers' interests.** *Computers in Human Behavior, 35, 132-139.* (journal article)

Thank You

Hope you enjoy this twitter data analysis tour