Hashtags on Twitter: Linguistic Aspects, Popularity Prediction

Pawan Goyal

CSE, IITKGP

July 24-30, 2015

Course Info

Course Website:

http://cse.iitkgp.ac.in/~pawang/courses/SC15.html

Course Info

Course Website:

`http://cse.iitkgp.ac.in/~pawang/courses/SC15.html`

Reference Books

- Matthew A. Russell. *Mining the Social Web: Data Mining Facebook, Twitter, Linkedin, Google+, Github, and More*, 2nd Edition, O'Reilly Media, 2013.
- Jennifer Golbeck. *Analyzing the social web*, Morgan Kaufmann, 2013.
- Charu Aggarwal (ed.), *Social Network Data Analytics*, Springer, 2011.

Course Info

Course Website:

`http://cse.iitkgp.ac.in/~pawang/courses/SC15.html`

Reference Books

- Matthew A. Russell. *Mining the Social Web: Data Mining Facebook, Twitter, Linkedin, Google+, Github, and More*, 2nd Edition, O'Reilly Media, 2013.
- Jennifer Golbeck. *Analyzing the social web*, Morgan Kaufmann, 2013.
- Charu Aggarwal (ed.), *Social Network Data Analytics*, Springer, 2011.

Lecture Material

- Lecture Slides
- Additional Readings

Course Evaluation Plan

- Mid-Sem : 20%
- End-Sem : 40%
- Project: 40%

Course Evaluation Plan

- Mid-Sem : 20%
- End-Sem : 40%
- Project: 40%

Project Groups

Each group should consist of 4 students

Adjunct Faculty

Dr. Monojit Choudhury, Microsoft Research

First set of lectures: August 19th and 20th, 5:30 - 7:30 PM

- Working with the Social Media Data
  - Language usage over social media: Basic challenges - 1 [spelling variation]
  - Language usage over social media: Basic challenges - 2 [transliteration & spelling]
  - Processing social media text: Entity Extraction
  - Processing social media text: POS tagging and parsing

Social Network Site (SNS)

"A web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system." (Boyd and Elison, 2007, p. 211)

What People Do in Social Media

What People Do in Social Media

- People make “friends” with others and build social relationships, connections and communities.

What People Do in Social Media

- People make “friends” with others and build social relationships, connections and communities.
- People ask and answer one another.

What People Do in Social Media

- People make “friends” with others and build social relationships, connections and communities.
- People ask and answer one another.
- People create, publish or distribute information in the form of text, photos, video, audio, or tweets.

What People Do in Social Media

- People make “friends” with others and build social relationships, connections and communities.
- People ask and answer one another.
- People create, publish or distribute information in the form of text, photos, video, audio, or tweets.
- People share bookmarks, presentation slides, or other files.

What People Do in Social Media

- People make “friends” with others and build social relationships, connections and communities.
- People ask and answer one another.
- People create, publish or distribute information in the form of text, photos, video, audio, or tweets.
- People share bookmarks, presentation slides, or other files.
- People provide feedback on or rate others’ information.

What People Do in Social Media

- People make “friends” with others and build social relationships, connections and communities.
- People ask and answer one another.
- People create, publish or distribute information in the form of text, photos, video, audio, or tweets.
- People share bookmarks, presentation slides, or other files.
- People provide feedback on or rate others’ information.
- People create social tags or folksonomies.

Social Networking Websites

Name	Description/focus	Date launched	Registered users
Google+	General	28 June 2011	1,600,000,000[135]
Facebook	General: photos, videos, blogs, apps.	February 2004	1,280,000,000[86]
Twitter	General. Micro-blogging, RSS, updates	15 July 2006	645,750,000[309]
Qzone	General. In Simplified Chinese; caters for mainland China users		480,000,000[247][248]
Sina Weibo	Social microblogging site in mainland China.	14 August 2009	300,000,000[262]
Instagram	A photo and video sharing site.	October 2010	300,000,000[162]
Habbo	General for teens. Over 31 communities worldwide. Chat room and user profiles.	August 2000	268,000,000[139][140][141]
VK	General, including music upload, listening and search. Popular in Russia and former Soviet republics.	September 2006	249,409,900[317]
Tumblr	Microblogging platform and social networking website.	February 2007	226,950,000[307]
LinkedIn	Business and professional networking	May 2003	200,000,000[184]
Renren	Significant site in China. Was known as 校内 (Xiaonei) until August 2009.		160,000,000[252]
Bebo	General	July 2005	117,000,000[16]
Tagged	General.	October 2004	100,000,000[288]
Orkut	General. Owned by Google Inc. Popular in India and Brazil.[229]	22 January 2004	100,000,000[230]

Reference for gathering social data

Larger Cover

Mining the Social Web, 2nd Edition

Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More

By Matthew A. Russell

Publisher: O'Reilly Media
Final Release Date: October 2013
Pages: 448

4.6

Read **11** Reviews | Write a Review

How can you tap into the wealth of social web data to discover who's making connections with whom, what they're talking about, and where they're located? With this expanded and thoroughly revised edition, you'll learn how to acquire, analyze, and summarize data from all corners of the social web, including Facebook, Twitter,...

Full description

Tools Used

IPhython Notebook

May use the following url: `http://smash.psych.nyu.edu/courses/spring12/modeling/ipythonhints.html`

Twitter

A microblogging service, allowing people to communicate with short, 140-character messages that roughly correspond to thoughts or ideas.

Twitter

A microblogging service, allowing people to communicate with short, 140-character messages that roughly correspond to thoughts or ideas.

Twitter's relationship model

Twitter

A microblogging service, allowing people to communicate with short, 140-character messages that roughly correspond to thoughts or ideas.

Twitter's relationship model

Allows you to keep up with the latest happenings of any other user, even though the other user may not choose to follow you back or even know that you exist.

Fundamental Twitter Terminology

Tweets are the essence of Twitter. In addition to the textual content, we get:

Tweet entities

User mentions, hashtags, URLs and media

Places

Locations in the real world, attached to a tweet

Fundamental Twitter Terminology

@ptwobrussell is writing @SocialWebMining, 2nd Ed. from his home office in Franklin, TN. Be #social: http://on.fb.me/16WJAf9

Tweet entities

- User mentions: @ptwobrussell, @SocialWebMining
- hashtag: #social
- URL: http://on.fb.me/16WJAf9
- Place: Franklin, Tennesee, in the tweet and a location metadata, where the tweet is authored

Fundamental Twitter Terminology

Timelines

Chronologically sorted collections of tweets

- **home timeline**: view that you see when you log into your account and look at all the tweets from users that you are following:
  `https://twitter.com`
- **user timeline**: collection of tweets only from a certain user:
  `https://twitter.com/timoreilly`
- **User's home timeline:** can be accessed with the additional *following* suffix, appended to the URL. :
  `https://twitter.com/timoreilly/following`

user timeline

Search Twitter

Have an account? **Log in**

TWEETS **34.7K** | FOLLOWING **1,441** | FOLLOWERS **1.96M** | FAVORITES **659** | LISTS **22**

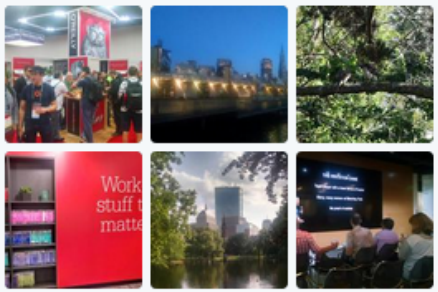
Follow

Tim O'Reilly
@timoreilly

Founder and CEO, O'Reilly Media. Watching the alpha geeks, sharing their stories, helping the future unfold.

Oakland, CA
radar.oreilly.com
Joined March 2007

310 Photos and videos



Tweets | Tweets & replies | Photos & videos

Tim O'Reilly @timoreilly · 3h

Things are hopping on the #oscon show floor!

6 8 View photo

Tim O'Reilly @timoreilly · 11h

Great suggestions

New to Twitter?

Sign up now to get your own personalized timeline!

Sign up

You may also like · Refresh


Eric Schmidt
@ericschmidt


Alex Howard
@digiphile

Mashable
@mashable

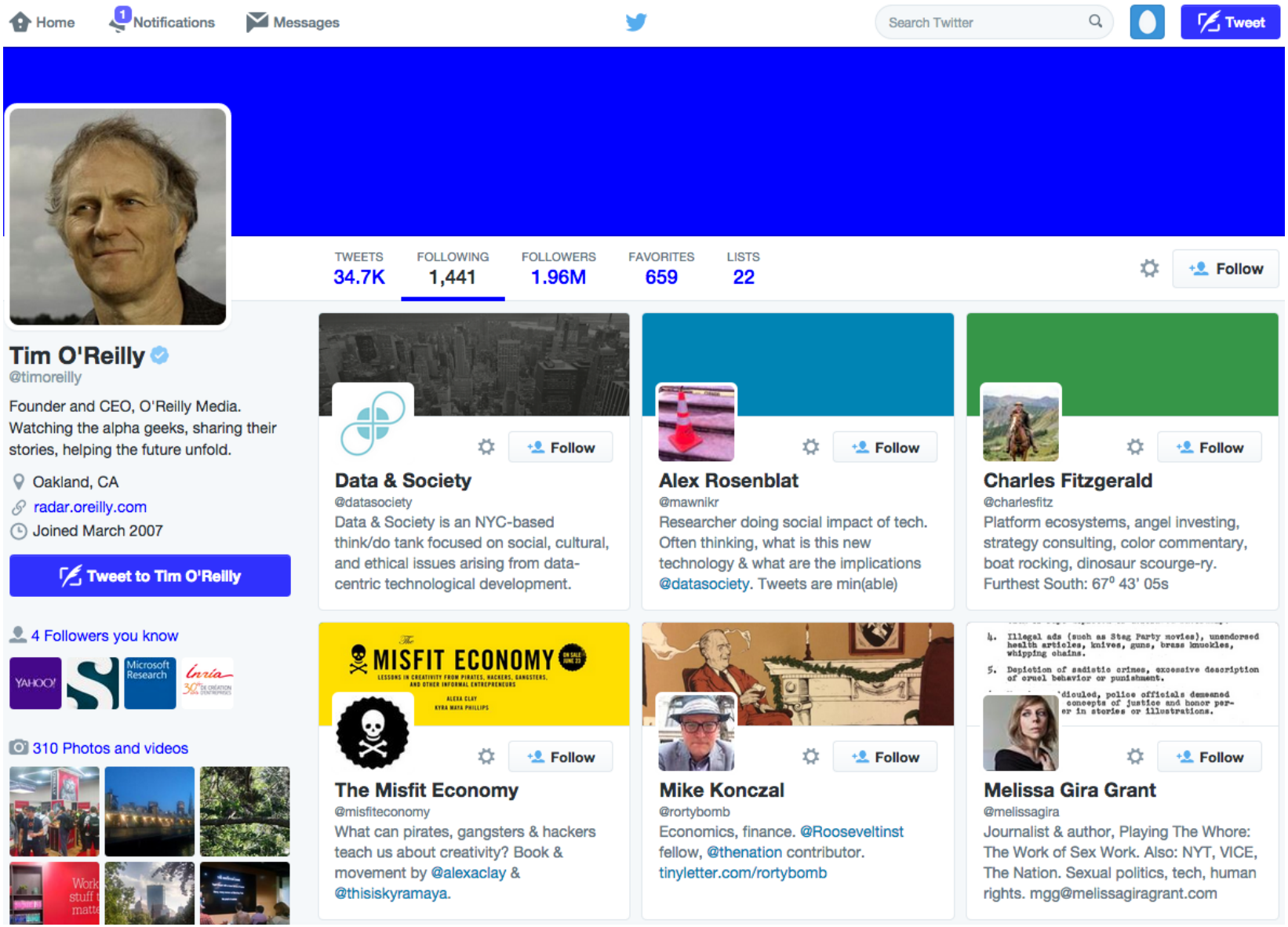

Clay Shirky
@cshirky


Fred Wilson
@fredwilson

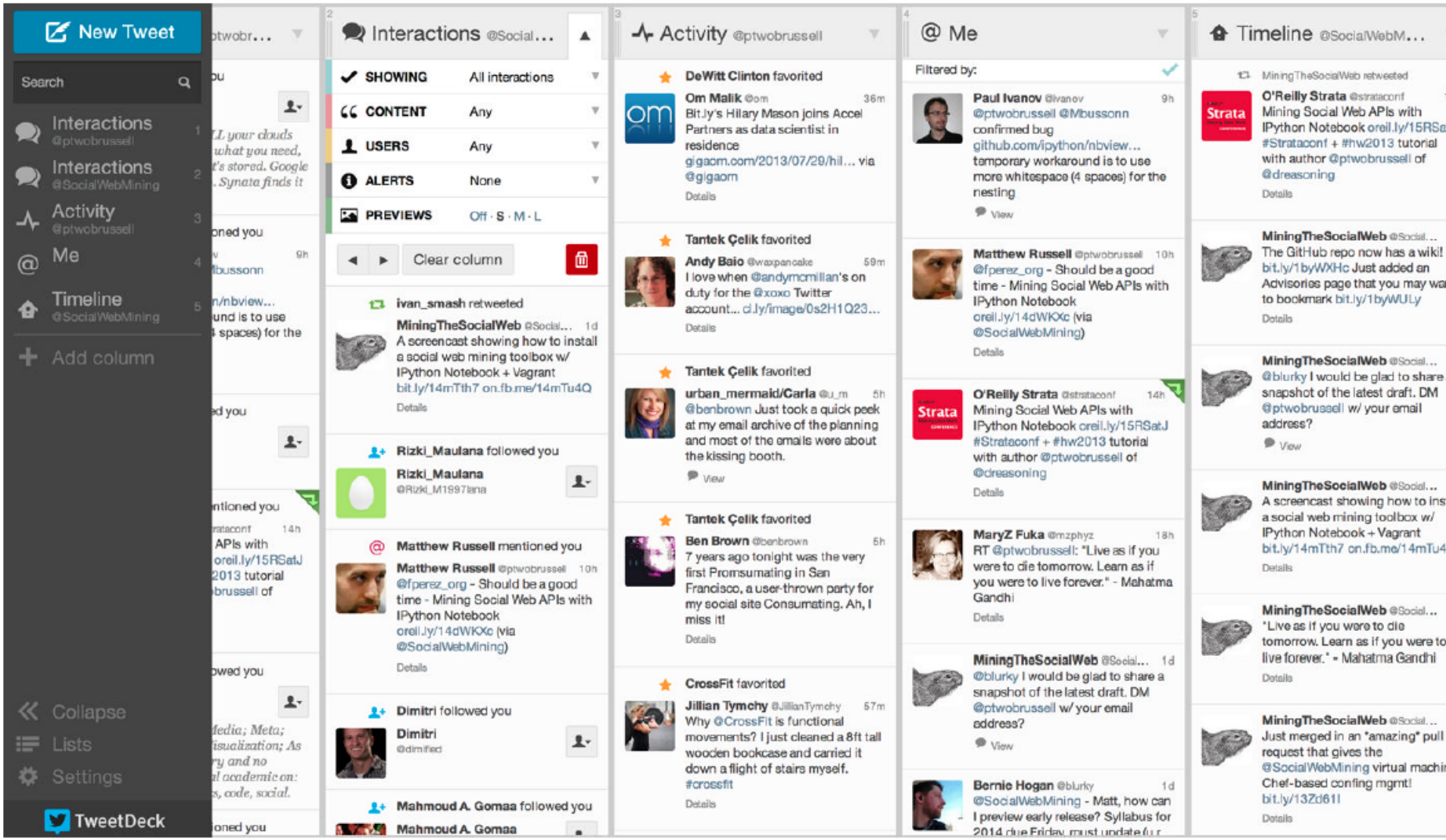
Trends

#Sharknado3

user home timeline



TweetDeck: a highly customizable user interface



Streams

Samples of public tweets flowing through Twitter in realtime

Public firehose

Known to peak at hundreds of thousands of tweets per minute during events with particularly wide interest.

Public timeline

A small random sample of the public timeline is available, that provides filterable access to enough public data for API developers

Exploring Twitter data

`https://github.com/ptwobrussell/Mining-the-Social-Web-2nd-Edition`

Requires : IPython Notebook

Things to try out

- Exploring Trending Topics
- Searching for tweets using a query
- Analyzing the 140 characters

Exploring Twitter data

`https://github.com/ptwobrussell/Mining-the-Social-Web-2nd-Edition`

Requires : IPython Notebook

Things to try out

- Exploring Trending Topics
- Searching for tweets using a query
- Analyzing the 140 characters

Other OSNs that the book discusses

- Facebook
- Google+
- LinkedIn

#Hashtags

What are #Hashtags?

Come under the general category of memes; a short unit of text that spreads from person to person within a culture.

#Hashtags

What are #Hashtags?

Come under the general category of memes; a short unit of text that spreads from person to person within a culture.

syntax

Adding a hash symbol (#) before a string of

- letters
- numerical digits or
- underscore sign (_)

Why are Hashtags useful?

Why are Hashtags useful?

- Hashtags are used to classify messages, propagate ideas and to promote specific topics and people.

Why are Hashtags useful?

- Hashtags are used to classify messages, propagate ideas and to promote specific topics and people.
- Allow users to create communities of people interested in the same topic, making it easier to find and share information related to it.

Social Phenomena

Social Phenomena

- A new social event can lead to the simultaneous emergence of several different hashtags, each one generated by a different user.

Social Phenomena

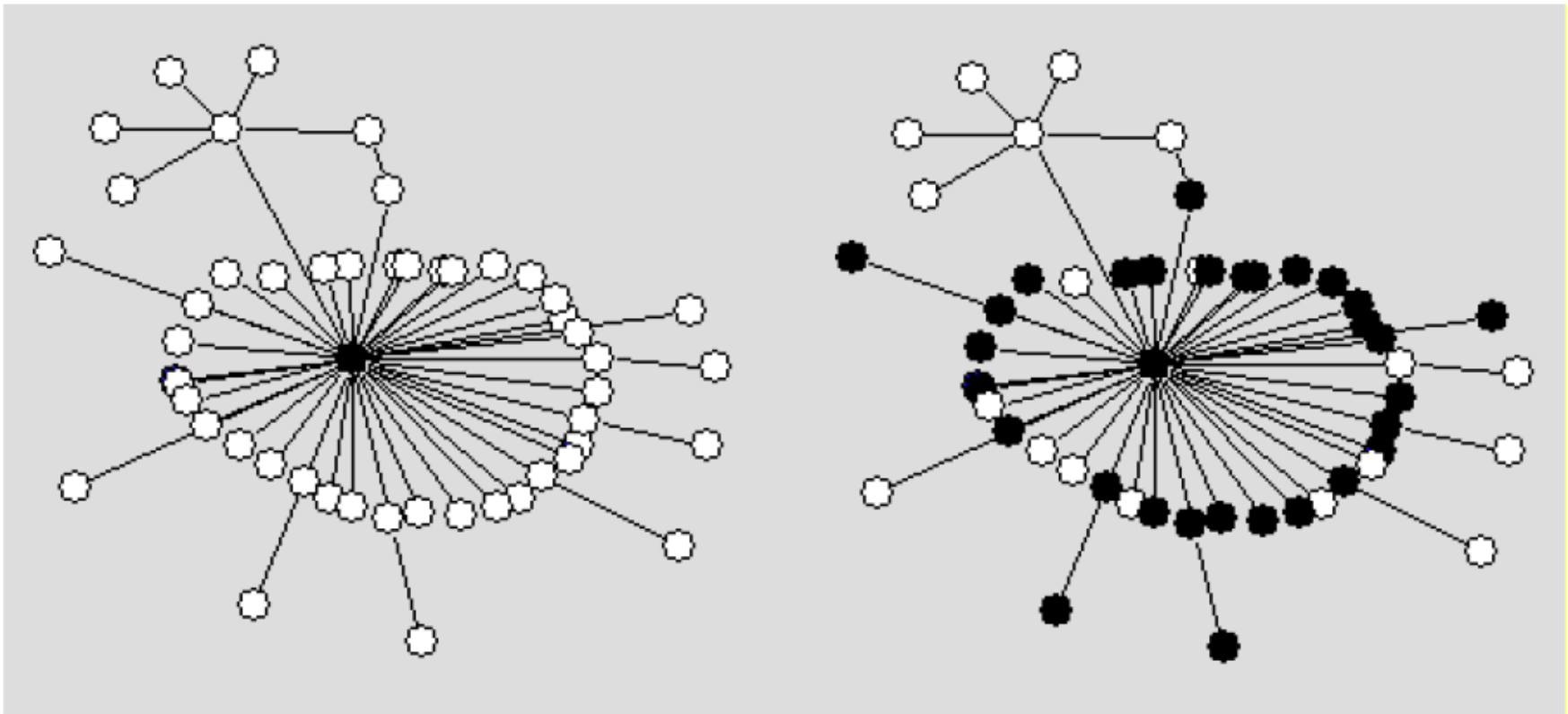
- A new social event can lead to the simultaneous emergence of several different hashtags, each one generated by a different user.
- They can either be accepted by other members of the network or not.

Social Phenomena

- A new social event can lead to the simultaneous emergence of several different hashtags, each one generated by a different user.
- They can either be accepted by other members of the network or not.
- Some propagate and thrive, while others die eventually or immediately after birth, being restricted to a few messages.

Novelty's propogation process

Subgraphs from Twitter dataset showing two distinct moments in the process of spreading #musicmonday



First Reference

Evandro Cunha, Gabriel Magno, Giovanni Comarela, Virgilio Almeida, Marcos André Goncalves, and Fabrício Benevenuto. 2011. *Analyzing the dynamic evolution of hashtags on Twitter: a language-based approach*. In Proceedings of the Workshop on Languages in Social Media (LSM '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 58 - 65.

Research Objectives

To understand the process of propagation of innovative hashtags in light of linguistic theories.

Interesting Questions

- Does the distribution of the hashtags in frequency rankings follow some pattern, as the words in the lexicon of a language?
- Is the length of a hashtag a factor that influences its success or failure?

Dataset Used

- 55 million users
- 2 billion follow links
- 8% users ignored because the profile was private

Dataset Used

- 55 million users
- 2 billion follow links
- 8% users ignored because the profile was private
- More than 1.7 billion tweets between July 2006 and August 2009 were analyzed

What aspect of the tweets would be important?

- Must find interchangeable hashtags, i.e. different tags used for the same purpose, to characterize messages on the same topic.

What aspect of the tweets would be important?

- Must find interchangeable hashtags, i.e. different tags used for the same purpose, to characterize messages on the same topic.
- For example, #michaeljackson, #mj, #jackson refer to the same subject.

Topics Selected

A minor base was built for each of the following topics:

- Michael Jackson (singer's death widely reported during that period) $\rightarrow$ MJ
- Swine Flu (H1N1 epidemic as major issue of 2009) $\rightarrow$ SF
- Music Monday (a very successful campaign in favor of posting tweets related to music on Mondays) $\rightarrow$ MM

Forming the bases

Filtering tweets that included

Forming the bases

Filtering tweets that included

- atleast one hashtag

Forming the bases

Filtering tweets that included

- atleast one hashtag
- at least one of the following terms that was thought to be related to the topics:
  - **MJ:** 'michael jackson'
  - **SF:** 'swine flu' or '#swineflu'
  - **MM:** '#musicmonday'

Summary information

- Number of tweets posted in that base
- Number of users who posted tweets
- Number of follow links among users of the base
- Number of different hashtags used in the tweets of the base

Summary information

- Number of tweets posted in that base
- Number of users who posted tweets
- Number of follow links among users of the base
- Number of different hashtags used in the tweets of the base

Base	Tweets	Users	Follow links	Different hashtags
MJ	221,128	91,176	3,171,118	19,679
SF	295,333	83,211	5,806,407	17,196
MM	835,883	196,411	7,136,213	16,005

Table 1. Summary information about the bases built.

Empirical Phenomenon

Rich-get-richer phenomenon

Also known as 'preferential attachment process': In some systems, the popularity of the most common items tends to increase faster than the popularity of the less comon ones.

Zipf's Law

Frequency of words in English or any other language follow a power law.

$$f \propto \frac{1}{r}$$

Empirical Phenomenon

Rich-get-richer phenomenon

Also known as 'preferential attachment process': In some systems, the popularity of the most common items tends to increase faster than the popularity of the less comon ones.

Zipf's Law

Frequency of words in English or any other language follow a power law.

$$f \propto \frac{1}{r}$$

$$log(f) = log(k) - log(r)$$

Distribution of Hashtags

- i-tweets hastags: hastags appearing in at most i tweets
- j-tweet hashtags: hashtags that appear in at least j tweets

Distribution of Hashtags

- i-tweets hastags: hastags appearing in at most i tweets
- j-tweet hashtags: hashtags that appear in at least j tweets

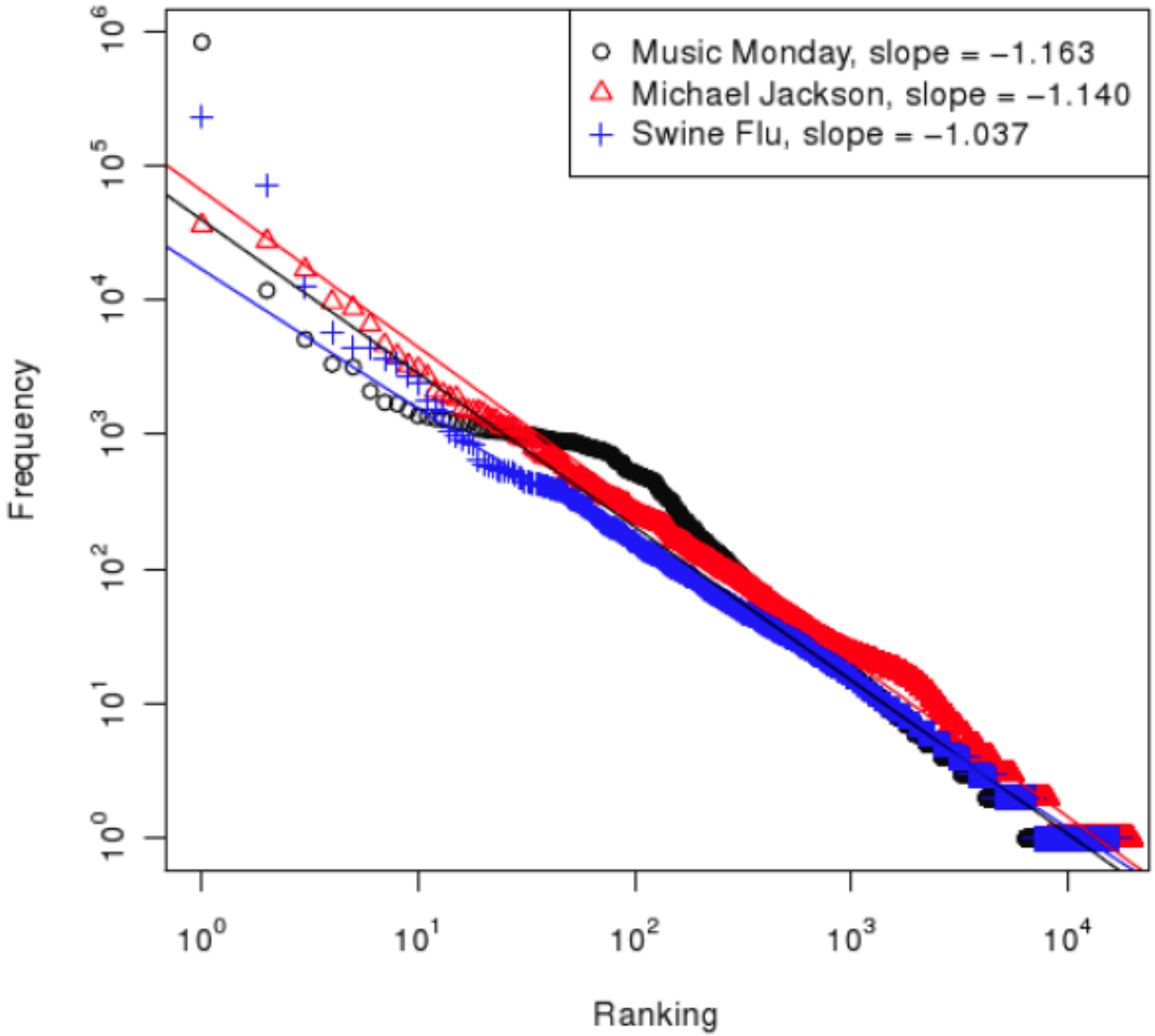
Base	% of i-tweet hashtags inside the base		
	i=1	i=2	i=10
MJ	59%	72%	88%
SF	59%	73%	92%
MM	60%	74%	91%

Table 2. Distribution of less common hashtags of each base.

Base	number of j-tweet hashtags inside the base		
	j=10,000	j=5,000	j=1,000
MJ	3	6	28
SF	3	4	14
MM	2	3	28

Table 3. Distribution of most popular hashtags of each base.

Data from most used Hashtags



Verification of Zipfian Law

Volume of Tweets vs. its position in popularity ranking

Base	Most used	2nd most used	3rd most used
MJ	#michaeljackson 35,861 12.3%	#michael 27,298 9.3%	#mj 16,758 5.7%
SF	#swineflu 230,457 51.5%	#h1n1 70,693 15.8%	#swine 12,444 2.8%
MM	#musicmonday 824,778 79.7%	#musicmondays 11,770 1.1%	#music 5,106 0.5%

Hashtag Length and Frequency

Zipf's Other Laws: Word length and word frequency

Word frequency is inversely proportional to their length.

The length of the most popular hashtags was compared with the the less popular ones.

Main Findings

- Most popular ones are simple, direct and short
- Among those with little utilization, many are formed by long strings of characters

Most Common Hashtags and most common 15-character hashtags

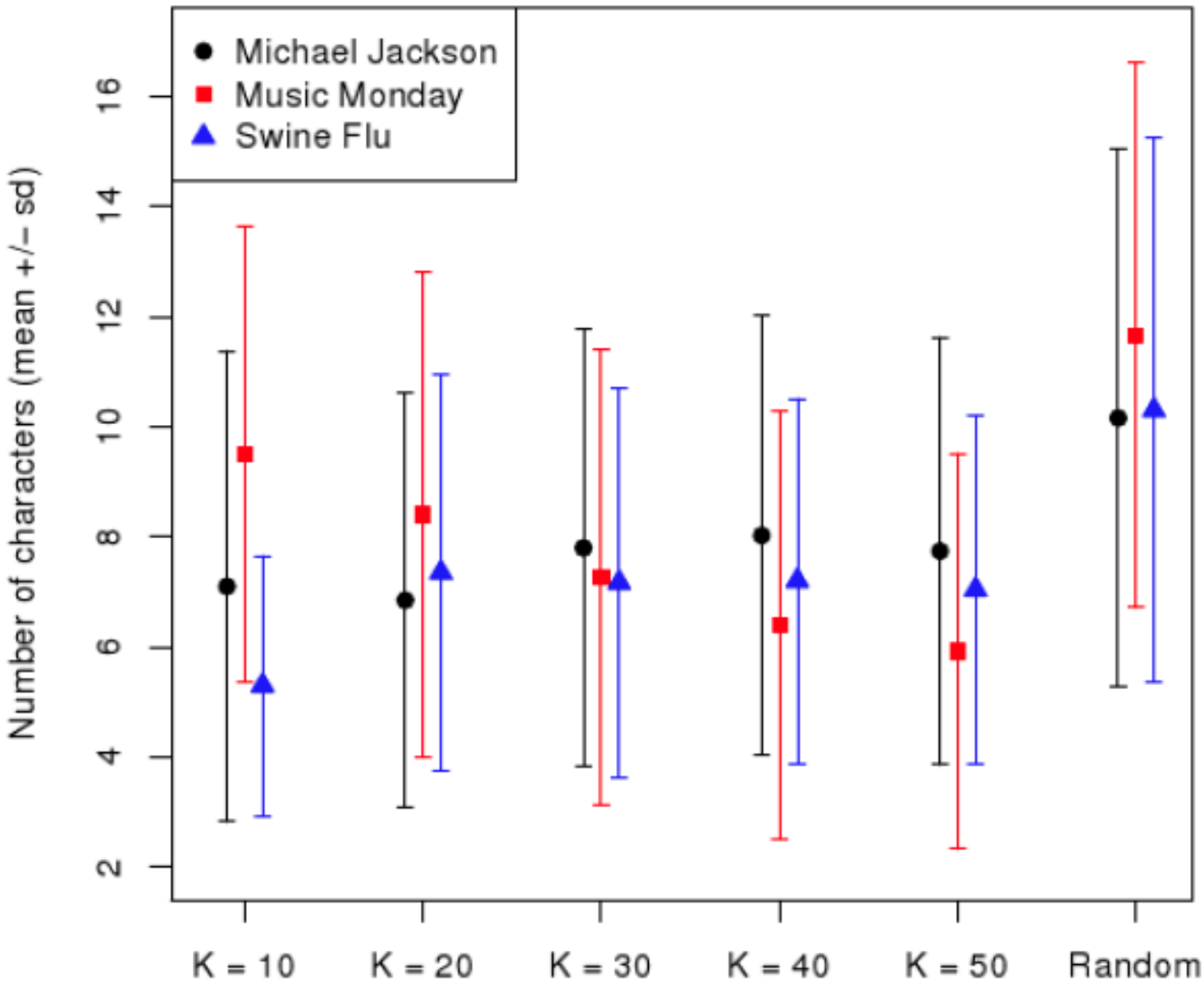
Most common hashtags (number of tweets)	Most common hashtags with 15 or more characters (number of tweets)
#michaeljackson (35,861)	#nothingpersonal (962)
#michael (27,298)	#iwillneverforget (912)
#mj (16,758)	#thankyoumichael (690)
#swineflu (230,457)	#swinefluhatesyou (1,056)
#h1n1 (70,693)	#crapnamesforpubs (145)
#swine (12,444)	#superhappyfunflu (124)
#musicmonday (824,778)	#musicmondayhttp (540)
#musicmondays (11,770)	#fatpeoplearesexier (471)
#music (5,106)	#crapurbanlegends (23)

Average Length

Topic	Average length of...					
	...the k most popular hashtags					...the less popular hashtags
	k=10	k=20	k=30	k=40	k=50	
MJ	7.1	6.85	7.8	8.02	7.74	10.16
SF	5.3	7.35	7.17	7.2	7.04	10.3
MM	9.5	8.4	7.27	6.4	5.92	11.66

Table 6. Average length of the most and the less popular hashtags. The samples with the less popular hashtags were formed by 50 randomly selected hashtags among those which appeared only in one tweet of each base.

Avergae Number of Characters



Underscores in Hashtags

Base	Number of _-hashtags	% of _-hashtags among *i*-tweet hashtags	
		i=2	*i*=10
MJ	251 (1.2%)	89%	97%
SF	155 (0.9%)	87%	97%
MM	143 (0.9%)	89%	98%

Underscores in Hashtags

Distribution of hashtags containing the sign "_"

- 97% of _-hashtags are used in 10 or less tweets
- #michael_jackson: position 248, only 128 tweets
- #swine_flu: position 67, only 246 tweets
- #music_monday: wasn't even used

User behavior seems to indicate rejection of this sign

Second Reference

Oren Tsur and Ari Rappoport. 2012. *What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities*. In Proceedings of the fifth ACM international conference on Web search and data mining (WSDM '12). ACM, New York, NY, USA, 643-652.

Content-based Popularity Prediction

Objective

Given an idea/meme m, and a time frame t, can we predict the acceptance of m in the community (social network)?

Content-based Popularity Prediction

Objective

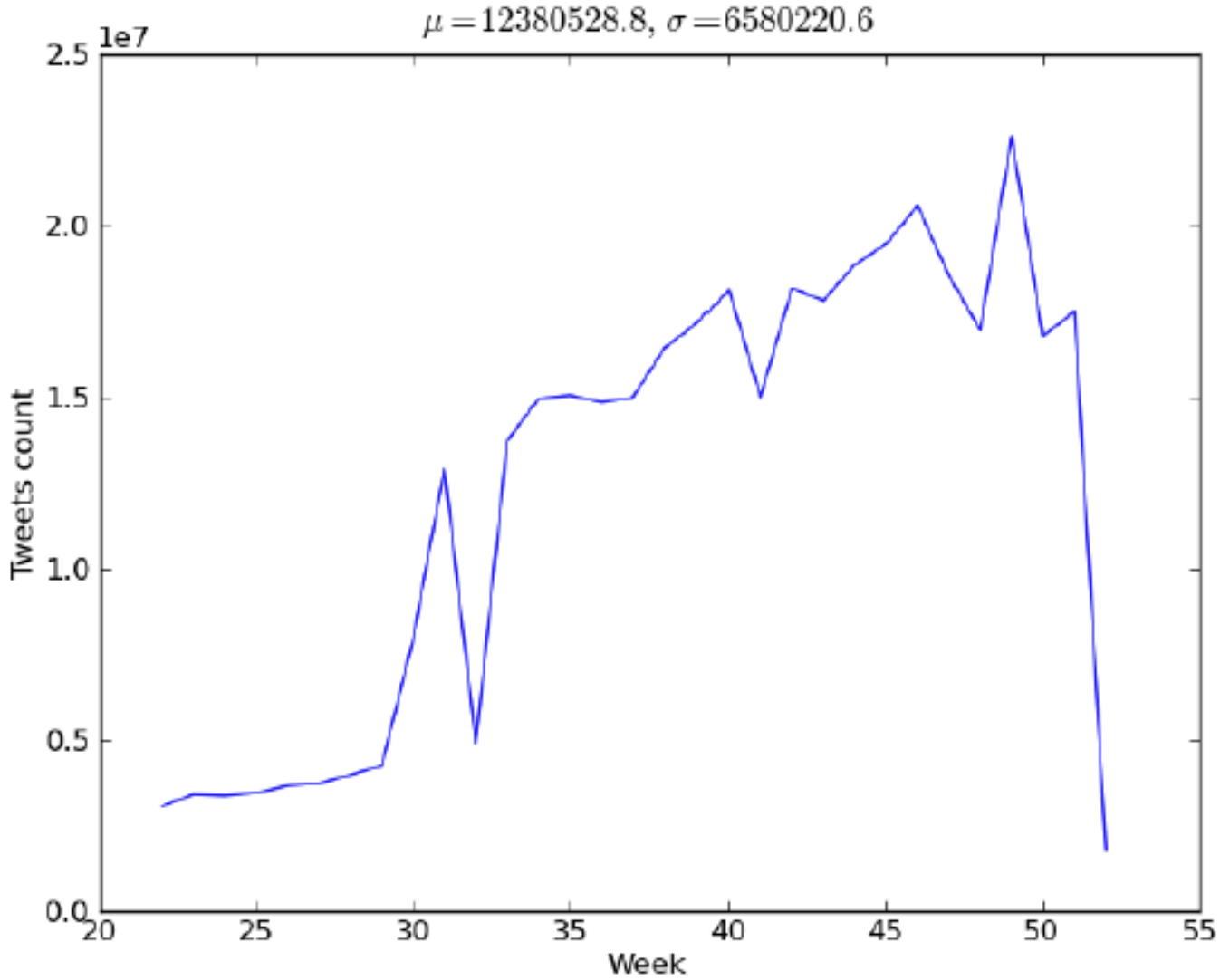
Given an idea/meme m, and a time frame t, can we predict the acceptance of m in the community (social network)?

Interesting Questions

- Can we accurately predict the acceptance of a meme based solely on the meme's content?
- Does the meme's context improve the prediction?
- Relation between graph topology and the content and how do they integrate for efficient propagation?

Corpus Used

400 million tweets, tweeted between June-December, 2009.



Filetering and Normalization

Filtering

- Filtered tweets containing non-Latin characters, to maintain a corpus of English tweets only
- Hashtags that appear over 100 times

Filetering and Normalization

Filtering

- Filtered tweets containing non-Latin characters, to maintain a corpus of English tweets only
- Hashtags that appear over 100 times

Normalization

The same hashtag could have got different counts because of being introduced in a different week.

$$N(ht^i) = \sum_{j \in weeks} count(ht^i{}_j) \frac{w_1}{w_j}$$

Fresh Hashtags

Hashtags that did not get popular before the corpus was collected.

Fresh Hashtags

Hashtags that did not get popular before the corpus was collected.

Definition

A hashtag is defined as *fresh* if it did not appear in the first week or if its normalized count in the first week is less than 10% of its normalized count in its peak.

Fresh hashtags

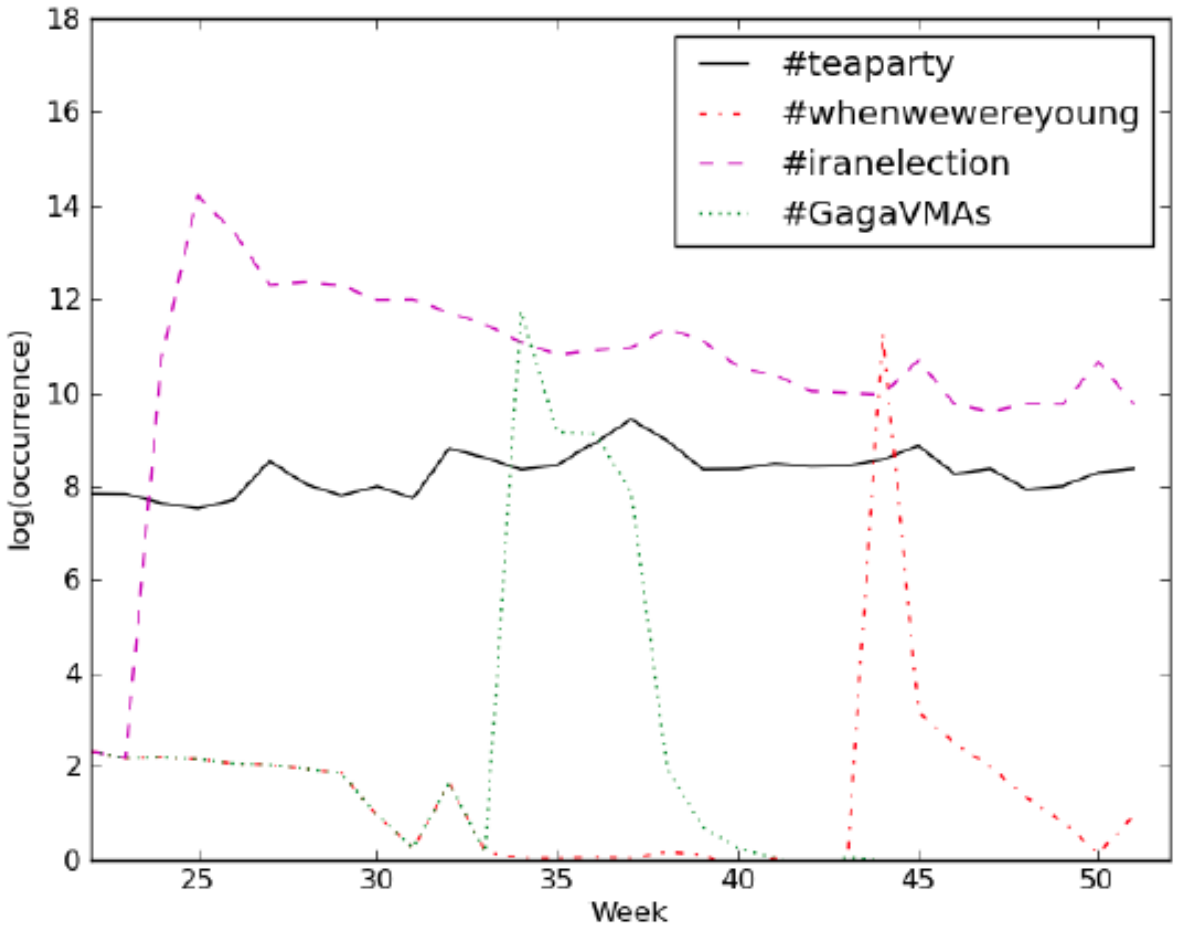


Figure 2: Four typical temporal trends (unnormalized counts).

Identifying the distinct words in a hashtag

#thankyousachin

Preprocessing

Identifying the distinct words in a hashtag

#thankyousachin

Issues with hashtags: #savethenhs, #weluvjb

- Matching hashtags against a lexicon of English words
- Exploiting redundancy of hashtags that differ only orthographically

Preprocessing

Identifying the distinct words in a hashtag

#thankyousachin

Issues with hashtags: #savethenhs, #weluvjb

- Matching hashtags against a lexicon of English words
- Exploiting redundancy of hashtags that differ only orthographically

matches tuples like #freeiran, #FreeIran; performs segmentation as per the capital letters

Prediction Model

The Target Function

$$f(ht) = n$$

ht is a vector space representation of a given hashtag

n is the normalized count of its occurrences in a time frame.

Prediction Model

The Target Function

$$f(ht) = n$$

ht is a vector space representation of a given hashtag
n is the normalized count of its occurrences in a time frame.
Transformed target function: $f'(ht) = log(n)$.

Prediction Model

The Target Function

$$f(ht) = n$$

ht is a vector space representation of a given hashtag
n is the normalized count of its occurrences in a time frame.
Transformed target function: $f'(ht) = log(n)$.

Regression Model

Training set: $(X, Y) = \{x_i, y_i\}$, where for a hashtag ht_i:
x_i: feature vector representation of ht_i
$y_i = log(n_i)$, where n_i is the normalized count of occurrences of ht_i

Prediction Model

The Target Function

$$f(ht) = n$$

ht is a vector space representation of a given hashtag
n is the normalized count of its occurrences in a time frame.
Transformed target function: $f'(ht) = log(n)$.

Regression Model

Training set: $(X, Y) = \{x_i, y_i\}$, where for a hashtag ht_i:
x_i: feature vector representation of ht_i
$y_i = log(n_i)$, where n_i is the normalized count of occurrences of ht_i

$$Y = b + w^T X$$

L1 Regularization with Stochastic Gradient Descent

$$L_r(b, w) = \frac{1}{2} \sum_i \left(y_i - (b + \sum_j w_j^T x_i^j) \right)^2 + \frac{1}{2} \lambda \|w\|$$

L1 Regularization with Stochastic Gradient Descent

$$L_r(b, w) = \frac{1}{2} \sum_i \left(y_i - (b + \sum_j w_j^T x_i^j) \right)^2 + \frac{1}{2} \lambda \|w\|$$

Parameter update for Stochastic Gradient Descent (SGD)

$$\begin{aligned} \Delta b &= \eta_t \left(y_i - (b + w^T x_i) \right) \\ \Delta w_i &= \eta_t \left(y_i - (b + w^T x_i) x_i - \lambda w_i \right) \end{aligned}$$

Model Features

- Hashtag content :- features that can be extracted from the hashtag itself.
- Global tweet features:- features related to the content of the tweets containing the hashtag.
- Graph topology features:- features related to graph topology and retweet statistics.
- Global temporal features:- features related to temporal pattern of the use of the hashtag.

Hashtag Content Features

Character Length

7 bins were used: 2, 3, 4, 5, 6-9, 10-14, >14 characters

Hashtag Content Features

Character Length

7 bins were used: 2, 3, 4, 5, 6-9, 10-14, >14 characters

Number of words

55% of the hashtags were compounds of more than one word, e.g. #freeIran, #GoogleGoesGaga.

Four bins: 1 word, 2-3 words, 4 words, >4 words

Hashtag Content Features

Orthography

Hashtags can be written in capital letters, contain some capital and/or digits, e.g. #myheart4JB.

'*right*' writing style may make it readable: savethenhs vs. saveTheNHS

Attributes: no caps, some caps, all caps, contain digits

Hashtag Content Features

Lexical Items

Hashtag/its words are matched against five predefined lists:

Hashtag Content Features

Lexical Items

Hashtag/its words are matched against five predefined lists:

- a general lexicon containing all words from a large portion (612MB) of English Wikipedia

Hashtag Content Features

Lexical Items

Hashtag/its words are matched against five predefined lists:

- a general lexicon containing all words from a large portion (612MB) of English Wikipedia
- a list of proper names taken from the name list compiled at the US census of 1995

Hashtag Content Features

Lexical Items

Hashtag/its words are matched against five predefined lists:

- a general lexicon containing all words from a large portion (612MB) of English Wikipedia
- a list of proper names taken from the name list compiled at the US census of 1995
- a list of celebrity names compiled from Forbes' 'The Celebrity 100' lists of 2008-2010.

Hashtag Content Features

Lexical Items

Hashtag/its words are matched against five predefined lists:

- a general lexicon containing all words from a large portion (612MB) of English Wikipedia
- a list of proper names taken from the name list compiled at the US census of 1995
- a list of celebrity names compiled from Forbes' 'The Celebrity 100' lists of 2008-2010.
- a short list of holidays and days of the week

Hashtag Content Features

Lexical Items

Hashtag/its words are matched against five predefined lists:

- a general lexicon containing all words from a large portion (612MB) of English Wikipedia
- a list of proper names taken from the name list compiled at the US census of 1995
- a list of celebrity names compiled from Forbes' 'The Celebrity 100' lists of 2008-2010.
- a short list of holidays and days of the week
- a list of all the world's countries

Each of the five attributes is an attribute in the vector

Hashtag Content Features

Location

Location of a hashtag can give an indication of the way it is used:
For instance, if located in the middle of the tweet, hashtag also serves as part of the sentence and not only as a meta tag.

Hashtag Content Features

Location

Location of a hashtag can give an indication of the way it is used:
For instance, if located in the middle of the tweet, hashtag also serves as part of the sentence and not only as a meta tag.
Three locations: *prefix, infix, suffix*
Ex: “AP: Report: #Iran’s paramilitary launches cyber attack http://is.gd/HiCYJU #iranelections #freeiran”

Hashtag Content Features

Location

Location of a hashtag can give an indication of the way it is used:
For instance, if located in the middle of the tweet, hashtag also serves as part of the sentence and not only as a meta tag.
Three locations: *prefix, infix, suffix*
Ex: “AP: Report: #Iran’s paramilitary launches cyber attack http://is.gd/HiCYJU #iranelections #freeiran”
Last two hashtags are considered suffixes

Hashtag Content Features

Location

Location of a hashtag can give an indication of the way it is used:
For instance, if located in the middle of the tweet, hashtag also serves as part of the sentence and not only as a meta tag.
Three locations: *prefix, infix, suffix*
Ex: “AP: Report: #Iran’s paramilitary launches cyber attack http://is.gd/HiCYJU #iranelections #freeiran”
Last two hashtags are considered suffixes
#Iran is considered Infix

Hashtag Content Features

Location

Location of a hashtag can give an indication of the way it is used:
For instance, if located in the middle of the tweet, hashtag also serves as part of the sentence and not only as a meta tag.
Three locations: *prefix, infix, suffix*
Ex: “AP: Report: #Iran’s paramilitary launches cyber attack http://is.gd/HiCYJU #iranelections #freeiran”
Last two hashtags are considered suffixes
#Iran is considered Infix

Collocation

Whether it collocates with other hashtags?
Value 1 if more than 40% of the occurrences are collocated with other hashtags.

Hashtag Content Features

Cognitive Dimension

Some words trigger specific emotions and encourage specific behavior and this psychological dimension can influence its spread.

Hashtag Content Features

Cognitive Dimension

Some words trigger specific emotions and encourage specific behavior and this psychological dimension can influence its spread.
LIWC project assigns words to a number of emotional and cognitive dimensions.

Hashtag Content Features

Cognitive Dimension

Some words trigger specific emotions and encourage specific behavior and this psychological dimension can influence its spread.
LIWC project assigns words to a number of emotional and cognitive dimensions.
Ex: positive sentiment, negative sentiment, optimistic, self, anger etc.

Global Tweet Features

Global Tweet Features

- 1000 most frequent words the hashtag co-occurred with were extracted.
- This list is mapped to the 69 LIWC categories

Graph Topology Features

Graph Topology Features

Average Number of followers

Average number of followers of users, who used the hashtag, is divided to 19 bins on a logarithmic scale

Graph Topology Features

Average Number of followers

Average number of followers of users, who used the hashtag, is divided to 19 bins on a logarithmic scale

Max Number of followers

Max number of followers of users, who used the hashtag, is divided to 19 bins on a logarithmic scale

Graph Topology Features

Average Number of followers

Average number of followers of users, who used the hashtag, is divided to 19 bins on a logarithmic scale

Max Number of followers

Max number of followers of users, who used the hashtag, is divided to 19 bins on a logarithmic scale

Retweets ratio

Tendency of a hashtag to appear in retweeted messages.

Graph Temporal Features

- The normalized weekly count of each hashtag was sampled in four time stamps: $w_i, i \in \{t, t+1, t+2, t+6\}$

Graph Temporal Features

- The normalized weekly count of each hashtag was sampled in four time stamps: $w_i, i \in \{t, t+1, t+2, t+6\}$
- t is the first week of occurrence and $t+j$ is the j-th week after the first occurrence.

Graph Temporal Features

- The normalized weekly count of each hashtag was sampled in four time stamps: $w_i, i \in \{t, t+1, t+2, t+6\}$
- t is the first week of occurrence and $t+j$ is the j-th week after the first occurrence.
- Three lag values are otained $d_{k \in 1,2,3}$, where d_k is the ratio of change from the previous time stamp. (stickiness and persistence)

Graph Temporal Features

- The normalized weekly count of each hashtag was sampled in four time stamps: $w_i, i \in \{t, t+1, t+2, t+6\}$
- t is the first week of occurrence and $t+j$ is the j-th week after the first occurrence.
- Three lag values are otained $d_{k \in 1,2,3}$, where d_k is the ratio of change from the previous time stamp. (stickiness and persistence)
- 17 bins on a logarithmic scale (-200% to 200% change)

Experimental Setup

Learning three aspects in the prediction

- what is the attribute combination that yields the best prediction?
- what are the strongest attributes and how do they complement each other?
- how does the prediction accuracy change given different time frames?

Experimental Setup

Learning three aspects in the prediction

- what is the attribute combination that yields the best prediction?
- what are the strongest attributes and how do they complement each other?
- how does the prediction accuracy change given different time frames?

Performance measurement

Experimental Setup

Learning three aspects in the prediction

- what is the attribute combination that yields the best prediction?
- what are the strongest attributes and how do they complement each other?
- how does the prediction accuracy change given different time frames?

Performance measurement

- Experiments were executed in a 10-fold cross validation manner.

Experimental Setup

Learning three aspects in the prediction

- what is the attribute combination that yields the best prediction?
- what are the strongest attributes and how do they complement each other?
- how does the prediction accuracy change given different time frames?

Performance measurement

- Experiments were executed in a 10-fold cross validation manner.
- Performance is measured by the mean square error (MSE).

Results

Model	$\mathbf{MSE_{10}}$	$\mathbf{MSE_{15}}$	$\mathbf{MSE_{20}}$	$\mathbf{MSE_{25}}$
baseline	4.988	3.796	3.125	2.698
$\mathbf{HT_{all}}$	4.380	3.410	2.902	2.565
$\mathbf{TW_{content}}$	4.776	3.509	2.743	2.221
Graph	4.295	3.144	2.404	1.923
Temporal	3.294	2.893	2.507	2.112
$\mathbf{Hybrid_{all}}$	2.584	2.098	1.685	1.315

Table 1: MSE of basic models and the hybrid model in horizons. MSE_n indicates results for acceptance prediction in an n weeks time frame.

Results

Model	MSE	Corr–coeff
baseline	3.796	-0.021
$\mathbf{Ht_{all}}$	3.410	0.319
$\mathbf{TW_{cont}}$	3.509	0.275
Graph	3.144	0.414
Temporal	2.893	0.487
$\mathbf{HT_{cont} + TW_{cont}}$	2.967	0.467
$\mathbf{HT_{cont} + TW_{cont}}$ + Graph	2.546	0.573
$\mathbf{HT_{cont} + TW_{cont}}$ + Temp	2.321	0.6234
Graph+Temporal	2.450	0.594
$\mathbf{Hybrid_{all}}$	2.098	0.669

Table 3: MSE and correlation coefficient for various combinations of feature types for a 15 weeks time frame.

Model	MSE	Corr–coeff
$\mathbf{d_1}$	3.236	0.383
$\mathbf{d_2}$	3.39	0.326
$\mathbf{d_3}$	3.44	0.303
$\mathbf{d_1 + d_2}$	3.088	0.431
$\mathbf{d_1 + d_3}$	2.97	0.464
$\mathbf{d_2 + d_3}$	3.19	0.398
$\mathbf{d_1 + d_2 + d_3}$	2.893	0.487

Table 5: MSE and correlation coefficient for different number of lags and different distances between sampling points in 15 weeks horizon. d_i indicates the the i-th lag described in Section 4.3.4.

Our Extension: Stratified Learning

Suman Kalyan Maity, Abhishek Gupta, Pawan Goyal and Animesh Mukherjee. *A stratified learning approach for predicting the popularity of Twitter Idioms*, ICWSM 2015.

What are Twitter Idioms

Our Extension: Stratified Learning

Suman Kalyan Maity, Abhishek Gupta, Pawan Goyal and Animesh Mukherjee. *A stratified learning approach for predicting the popularity of Twitter Idioms*, ICWSM 2015.

What are Twitter Idioms

- Twitter hashtags that are primarily used for conversational and personal resons

Our Extension: Stratified Learning

Suman Kalyan Maity, Abhishek Gupta, Pawan Goyal and Animesh Mukherjee. *A stratified learning approach for predicting the popularity of Twitter Idioms*, ICWSM 2015.

What are Twitter Idioms

- Twitter hashtags that are primarily used for conversational and personal resons
- e.g., #10ThingsAboutMe, #4WordsAfterABreakup, #WhatMadeMeMadAsAKid
- 17-25% of the Twitter hashtags correspond to this category
- We stratify the hastags into two categories: Idioms and non-Idioms and learn the regression model for each strata

Our Extension: Stratified Learning

- Given a hashtag, we first categorize it as an Idiom or a non-idiom.

Features used for classification

- Hashtag character length, number of words: #IfICouldLiveMyLifeInMyFavoriteMovieItWouldBe, #TweetAPictureThatDescribesYourFriendship
- Presence of days of the week and numerals: #FlashbackFriday, #20ThingsIDontLike
- Presence of hashtag words in the dictionary
- Common personal pronouns, verbs
- ...

Our Extension: Stratified Learning

- Idiom specific features were used for popularity prediction

Idiom specific features

- Frequency of the n-grams in the English texts
- Early coinage: #ThingsAboutMe has multiple variants #1ThingsAboutMe, #10ThingsAboutMe, #20ThingsAboutMe etc.
- ...

Our Extension: Stratified Learning

- Idiom specific features were used for popularity prediction

Idiom specific features

- Frequency of the n-grams in the English texts
- Early coinage: #ThingsAboutMe has multiple variants #1ThingsAboutMe, #10ThingsAboutMe, #20ThingsAboutMe etc.
- ...

Improves the correlation coefficient by 19.5%