Data Scientist's Approach to Social Data

Scott Hendrickson
Principal Data Scientist, Gnip
@DrSkippy27

April 22, 2014

Social Data

Social data

created by
interactions
among people

Social data

form and content
shaped by
people's behavior

Where Does Social Data Come From?

GNIP

Community of Users

- Common interests
- Cultural commonalities
- Topic interest/focus
- Common assumption about sharing/privacy interaction
- Contemporaries/peers

Platform

- Interaction modes
- Habits
- Connections
- Recommendations
- Engagement
- Storage/Structure

Social Data Analysts

- Privacy
- T.O.S.
- Enablers of advertising network analysis

Sources

- firehoses
- APIs
- scraping

Firehose

Continuous stream
of activities
in near-real time

Activity

people interacting
on social media platform

Firehose volumes

Publisher	Daily Activity
Twitter	520M
Tumblr	110M
Foursquare	4.2M
Wordpress Posts	1M
Wordpress Comments	1.7M
Disqus	1.9M
Engagement (likes, votes)	>60M

Every day @Gnip

$\frac{3}{4}$ Billion IN

4 Billion OUT

Analysis

Analysis considerations

- Technology - interfaces, tools, infrastructure for accessing
- Latency - how soon after activity as created?
- Uniformity - how hard/costly to normalize data formats?
- Coverage - do you need it all? a defined sample?
- Meta-data - how much and what kind of data about the data?

Business considerations

- Licensing - do you have the right to analyze, display, store data?
- Terms-of-Service Compliance - violating publishers terms of service, privacy protections?
- Cost - data collection costs? licensing costs? processing and storage costs?
- Analysis mode - batch vs. real-time? event vs. background? time, structure, language, people?

Models …

- domains - events, time series, language analysis, graph structures, etc.
- drives storage, analysis and access strategies, etc.
- analysis objectives - detection, alerting, status dashboard, discovery projects, validation, etc.

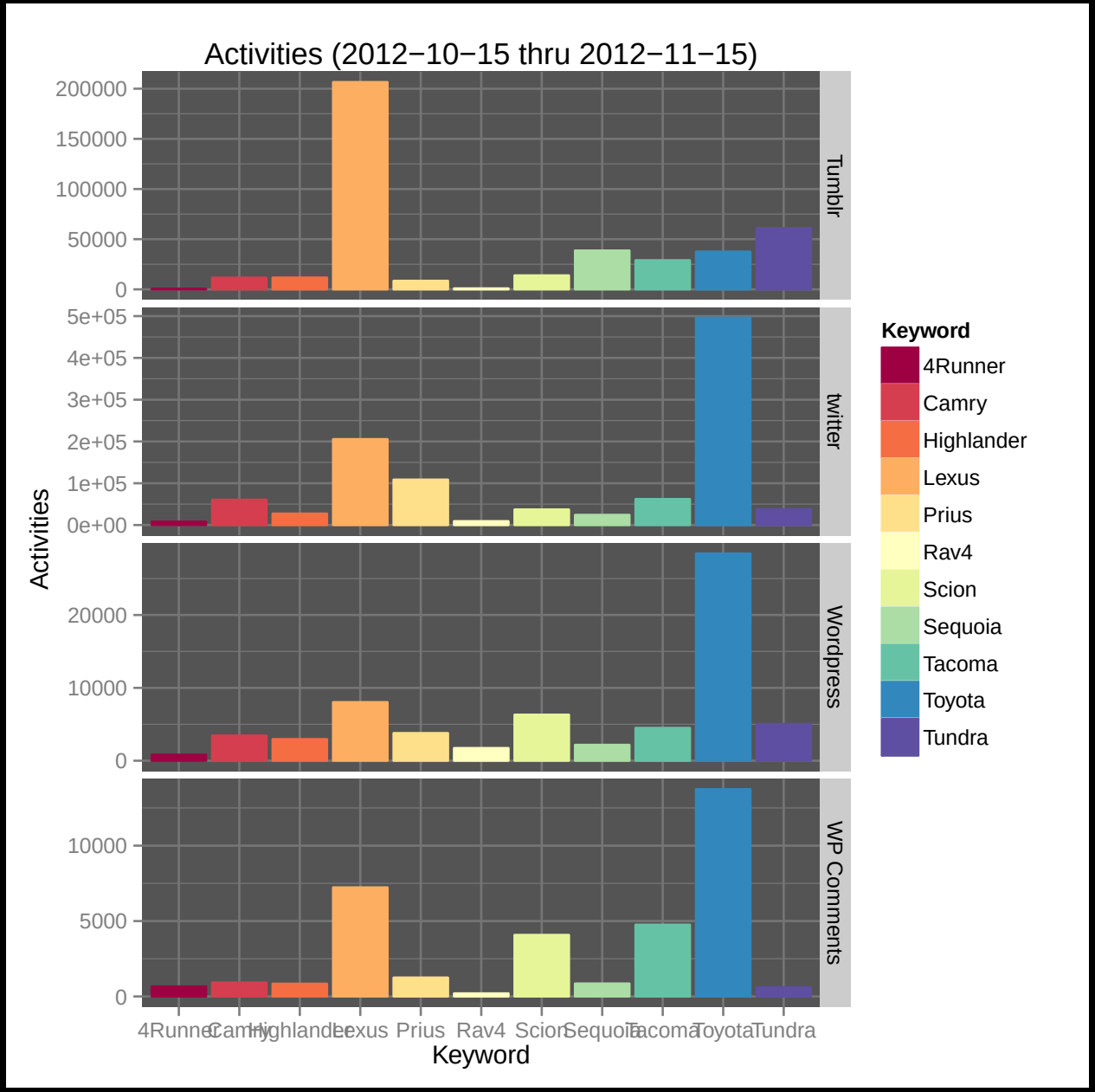
Darren Aronofsky's "Noah" Delayed Due to Flooding

Posted: October 31st, 2012 by WorstPreviews.com Staff

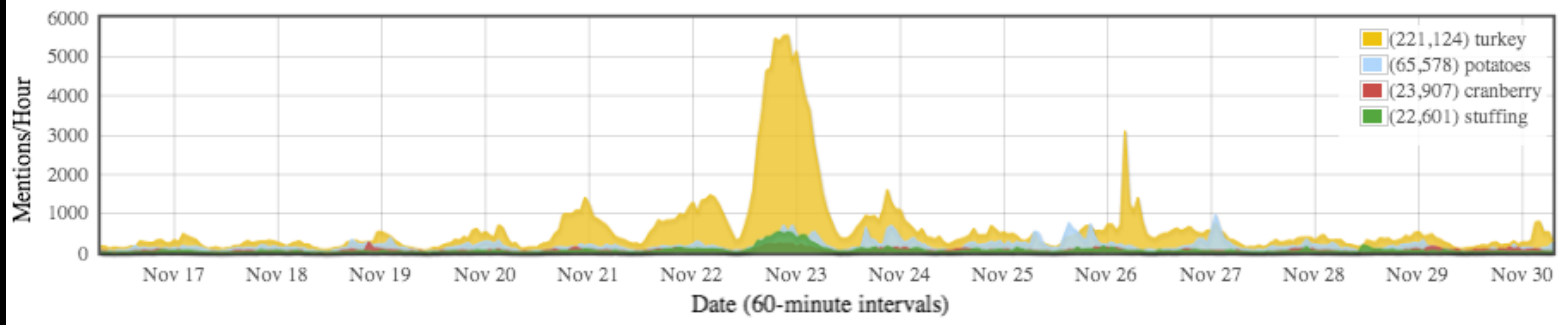
SUBMIT COMMENT

Darren Aronofsky (Black Swan, The Wrestler) has been filming his "Noah" film, based on the Biblical tale of Noah's Ark, at Oyster Bay, NY. To make it as realistic as possible, the director built a massive ark, which measures 450 feet long, 75 feet tall and 45 feet wide. Unfortunately, it was never meant to be sailed.

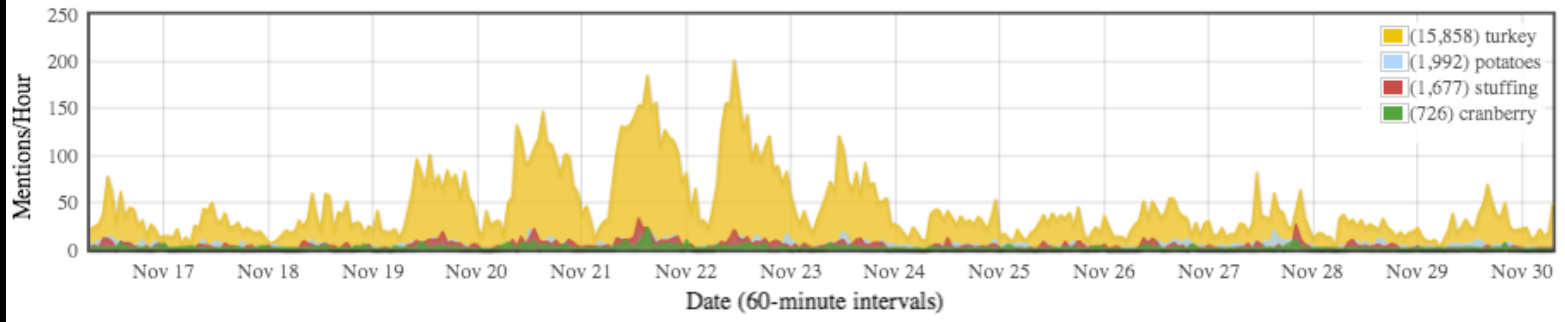
Audience


Activities (2012−10−15 thru 2012−11−15)
Activities
Keyword
Tumblr
twitter
Wordpress
WP Comments
200000
150000
100000
50000
0
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
20000
10000
5000
4Runner
Camry
Highlander
Lexus
Prius
Rav4
Scion
Sequoia
Tacoma
Toyota
Tundra
Keyword

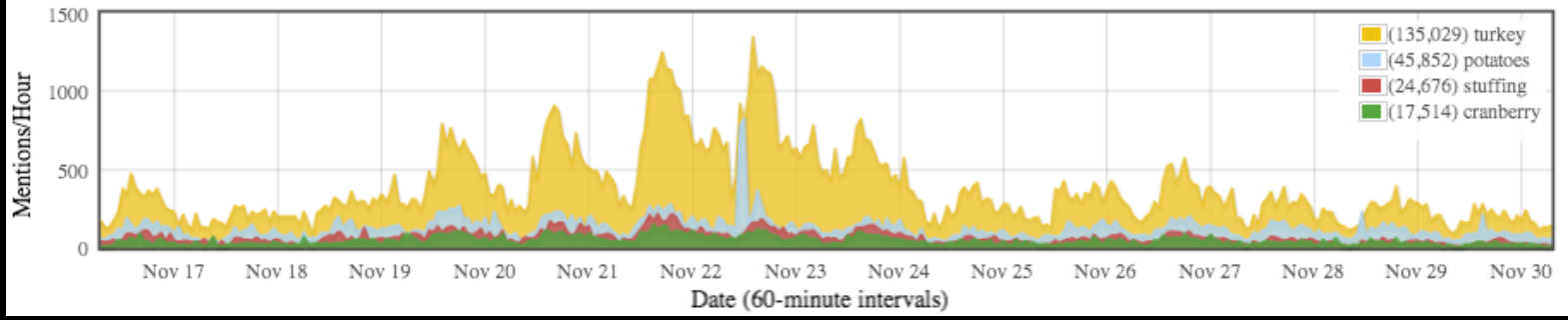
Tumblr



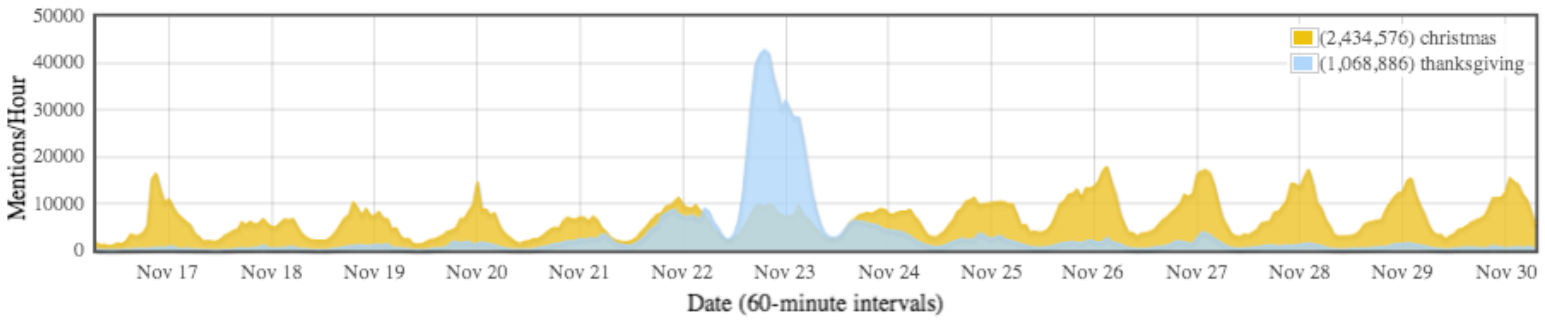
Disqus



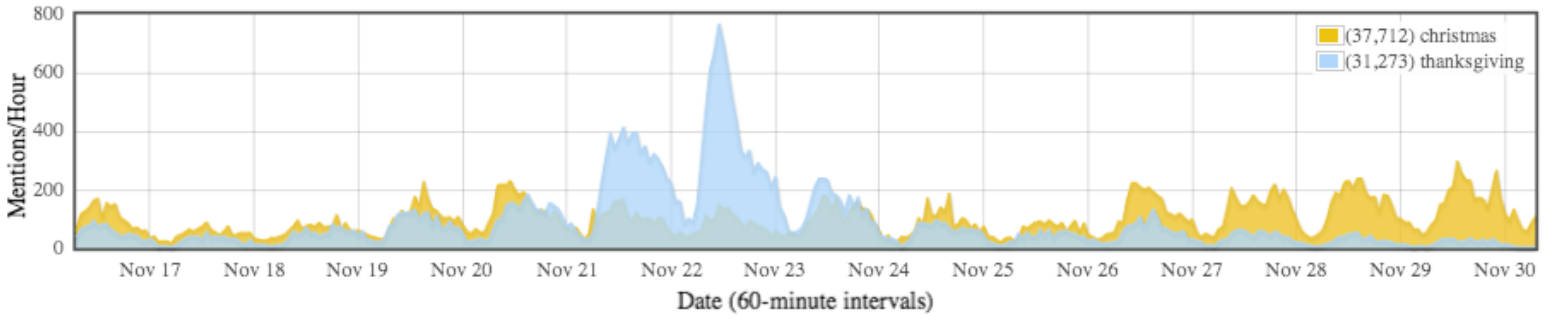
Wordpress



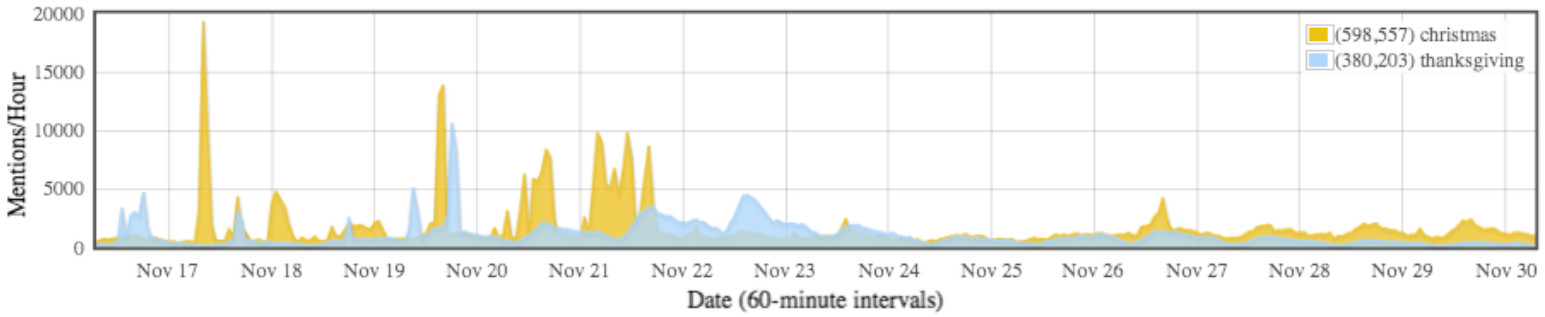
Tumblr



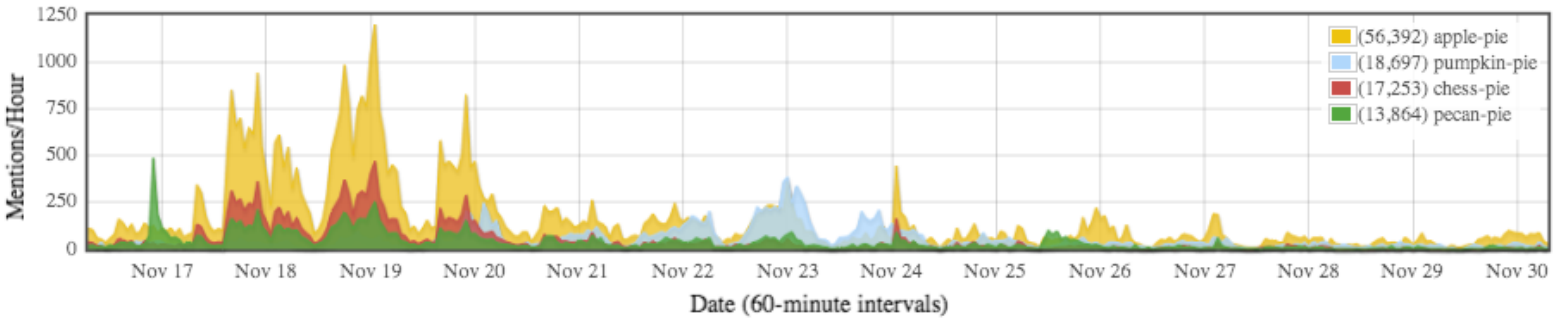
Disqus



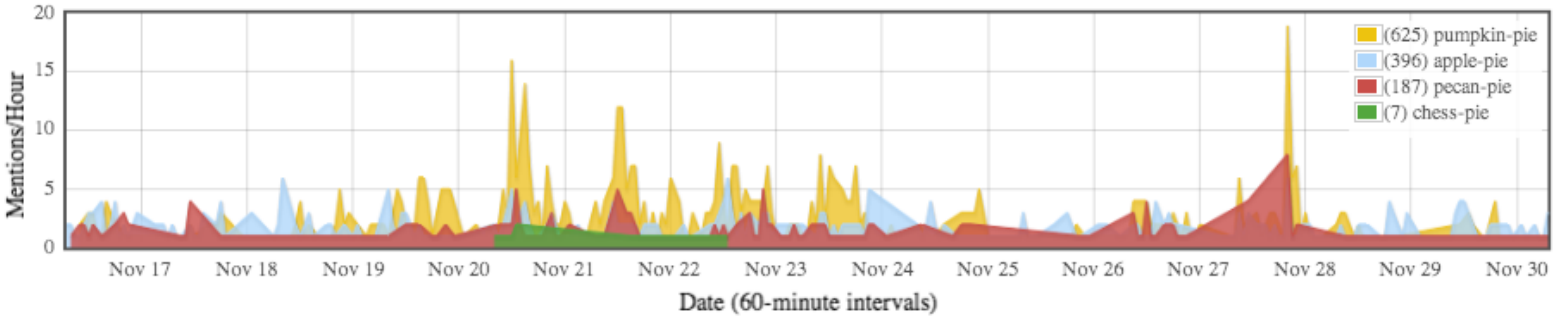
Wordpress



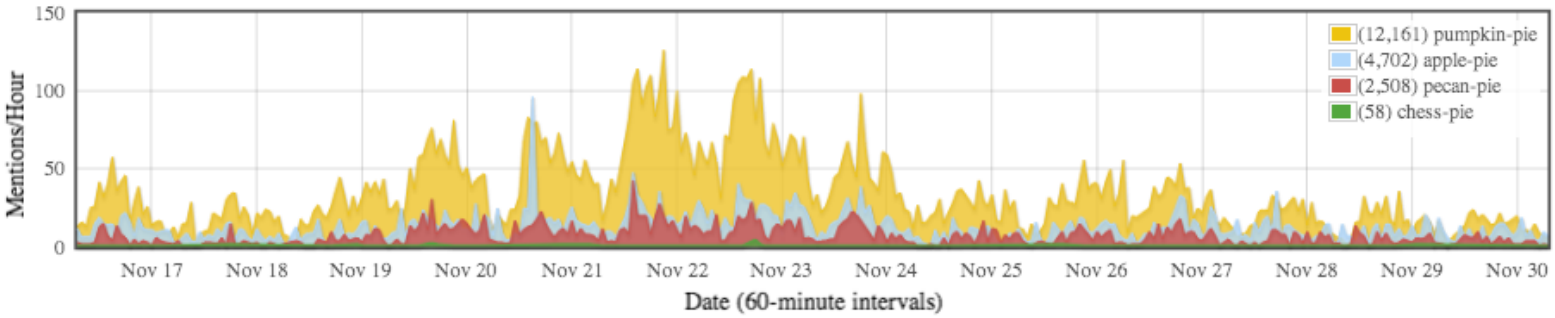
Tumblr



Disqus

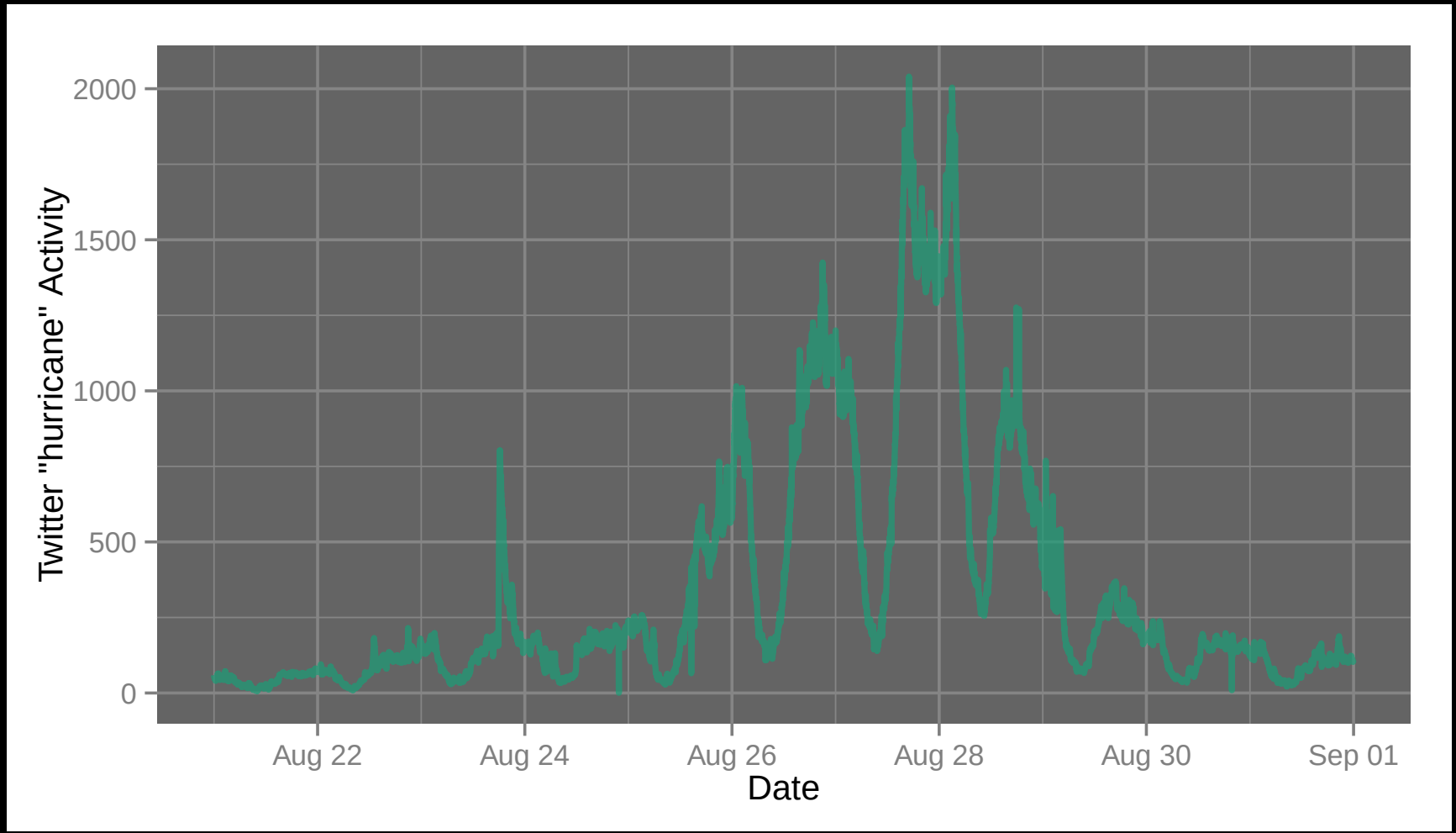


Wordpress

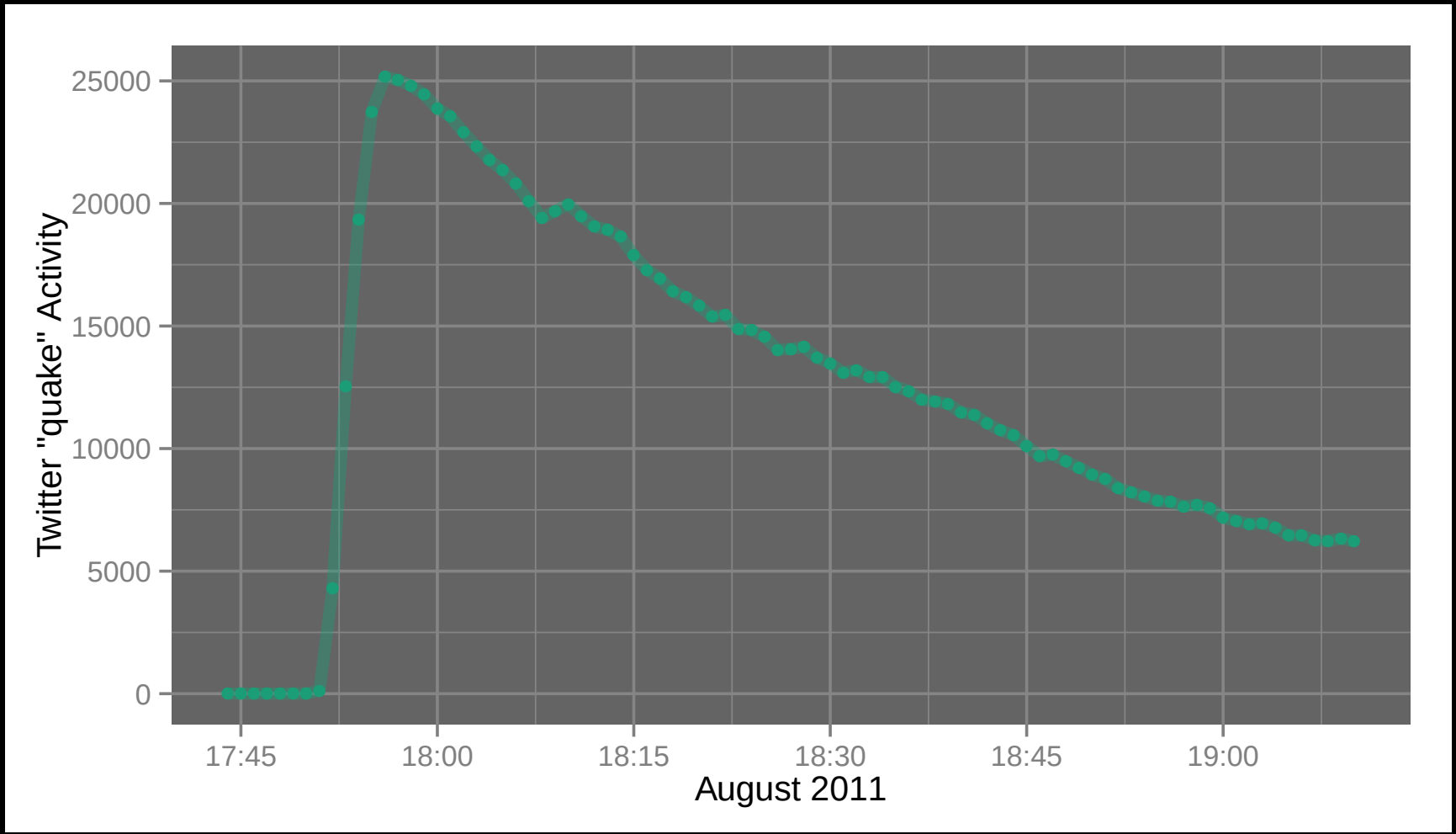


Social Medial Pulse

Expected: hurricane



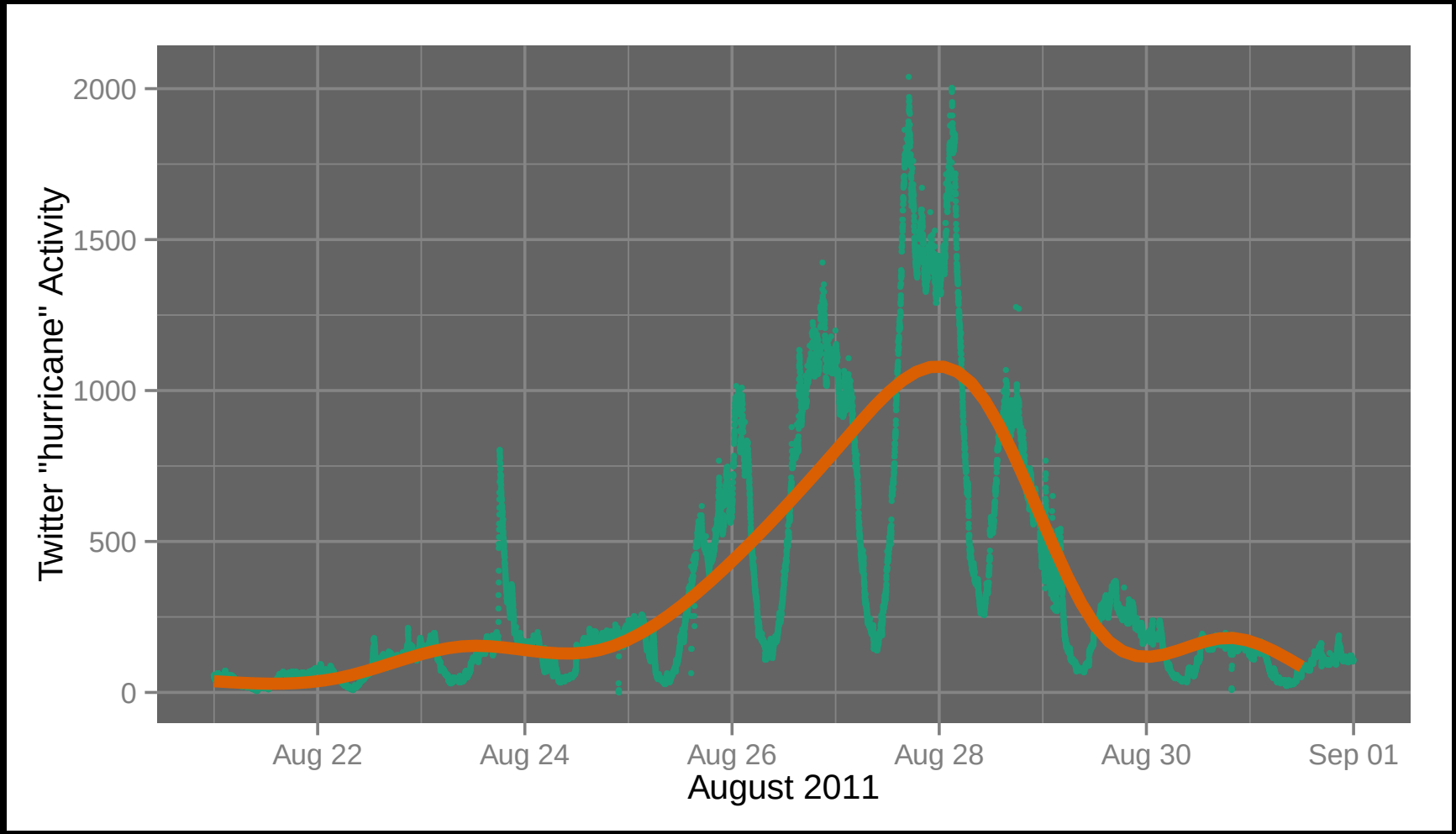
Unexpected: earthquake



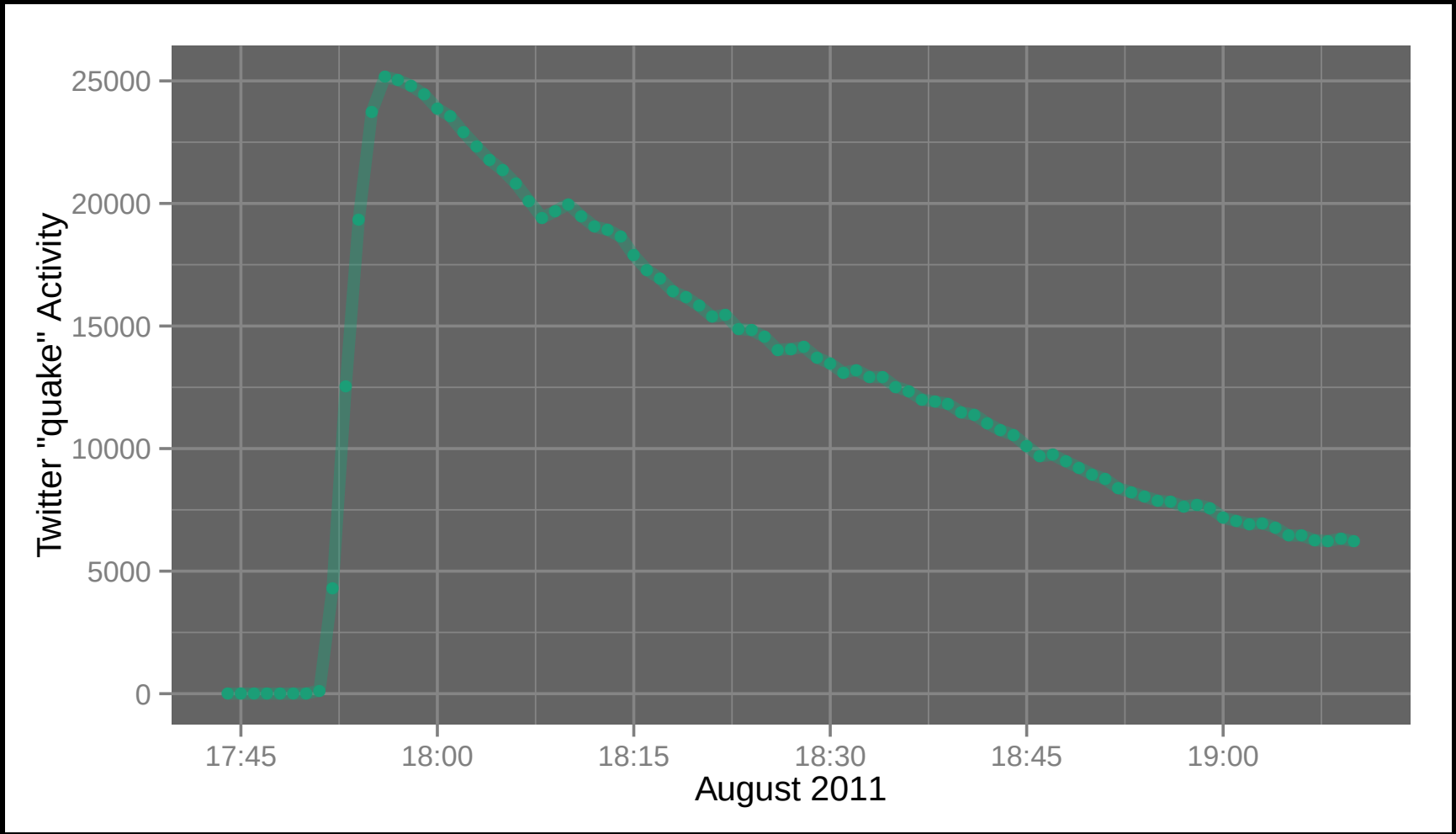
Classifying events

Type	Response	Examples
Expected	Build-up/Decay	Hurricane Sandy Olympics
Unexpected (many obs.)	Social Media Pulse	Beyoncé VMAs Mexico earthquake Steve Jobs
Unexpected (network spread)	Network Models	Osama bin Laden Whitney Houston Syrian dissidents

Expected: hurricane



Unexpected: earthquake



Half-life

time to observe
$\frac{1}{2}$ of the activities
for an event

Social media pulse

Given an event, the probability of an activity from one person,

$$f(t) = \lambda \exp(-\lambda t), \text{ for } t \geq 0.$$

Many people posting on same cue; so sum of random variables

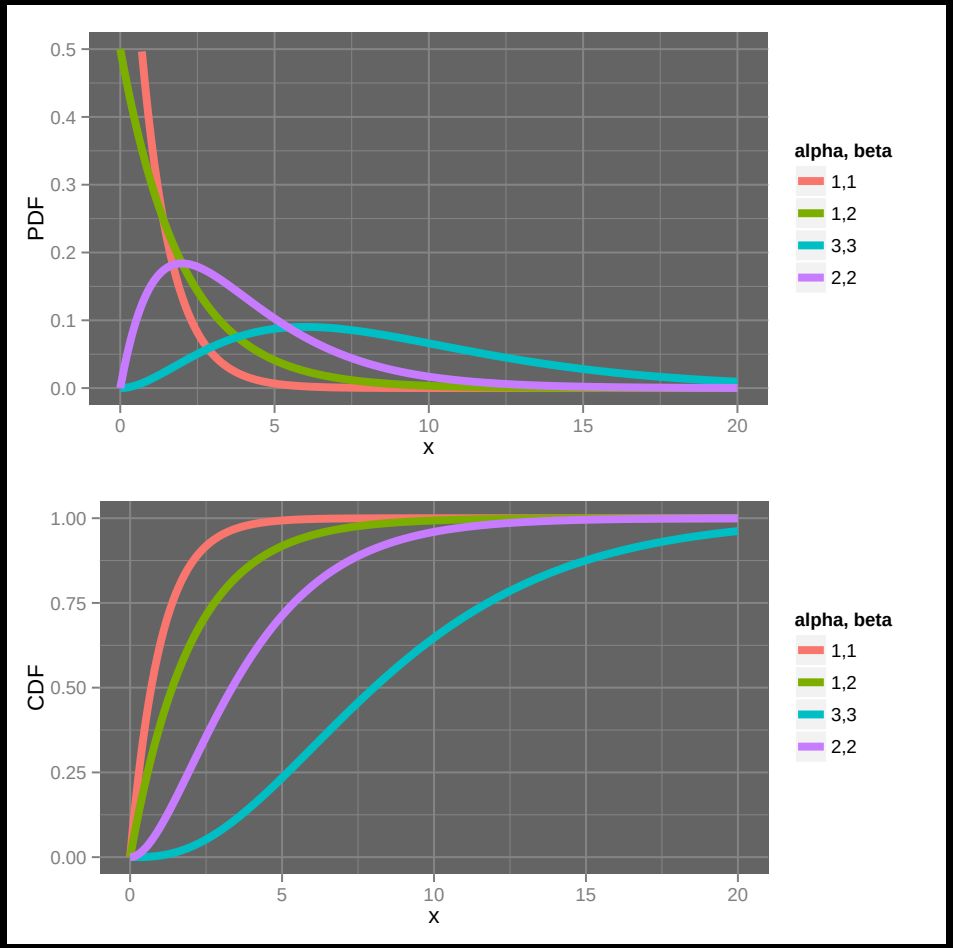
$$S = X_1 + X_2 + \ldots + X_{n \text{ posters}}$$

Gamma probability distribution function,

$$f_S(t) = \frac{\beta^{-\alpha} t^{\alpha-1} \exp(\frac{-t}{\beta})}{\Gamma(\alpha)}$$

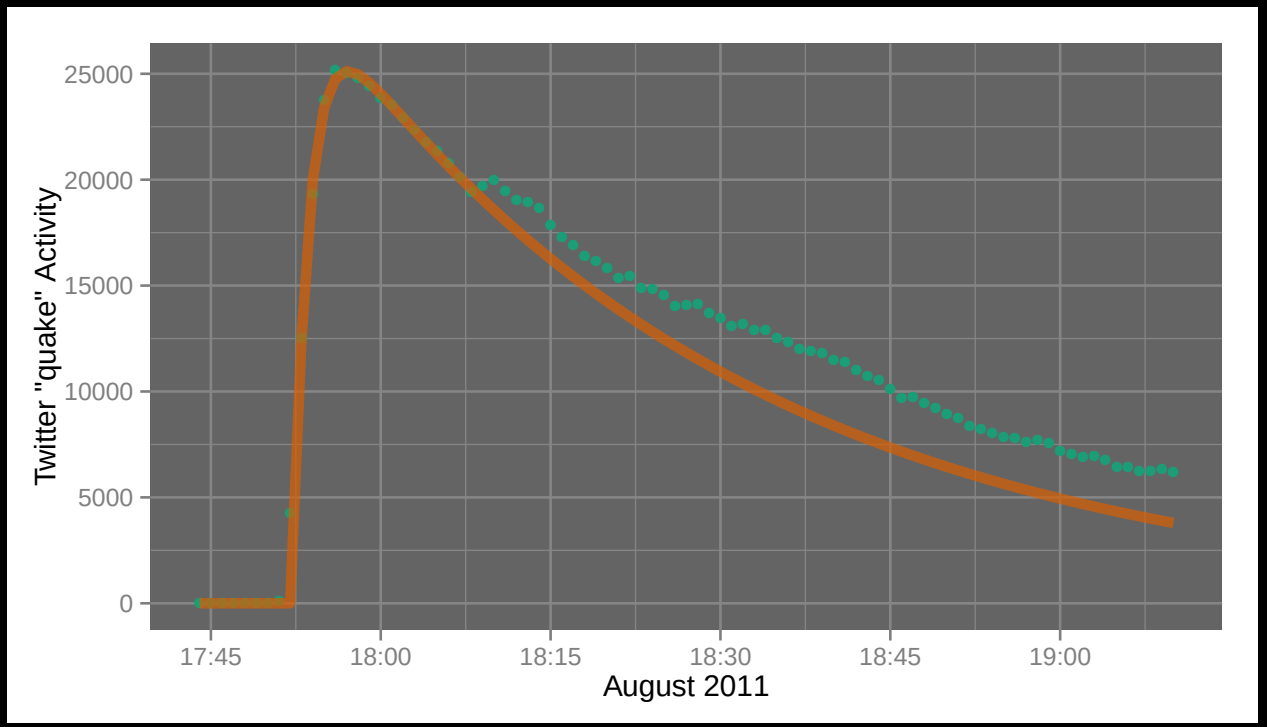
Cumulative distribution is the “generalized regularized incomplete gamma function”,

$$F_S(t) = Q(\alpha, 0, \frac{t}{\beta})$$

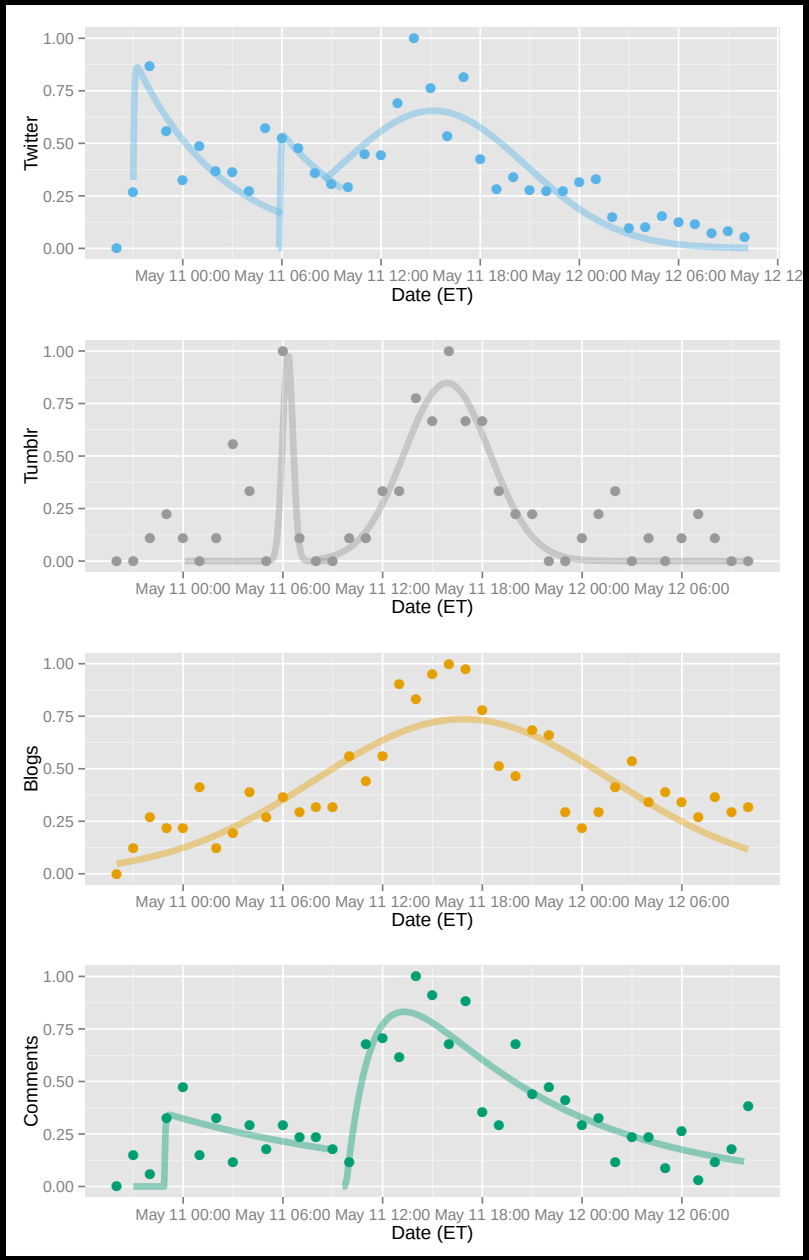

PDF
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
x
alpha, beta
1,1
1,2
3,3
2,2
CDF
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
x
alpha, beta
1,1
1,2
3,3
2,2

Why model half-life?

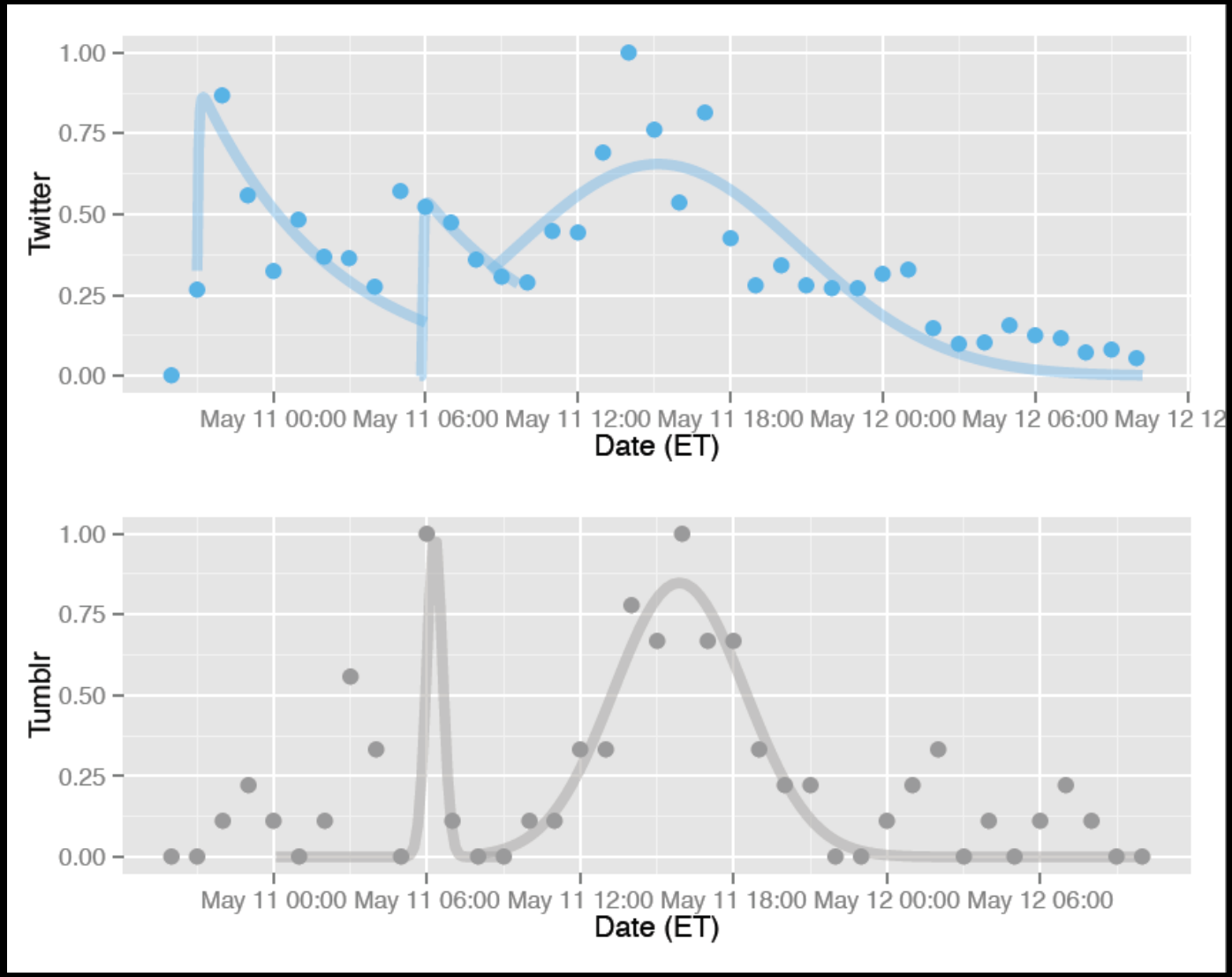
- predict total story volume
- compare half-lives
- anomalous story evolution


25000
20000
15000
10000
5000
0
Twitter "quake" Activity
17:45
18:00
18:15
18:30
18:45
19:00
August 2011

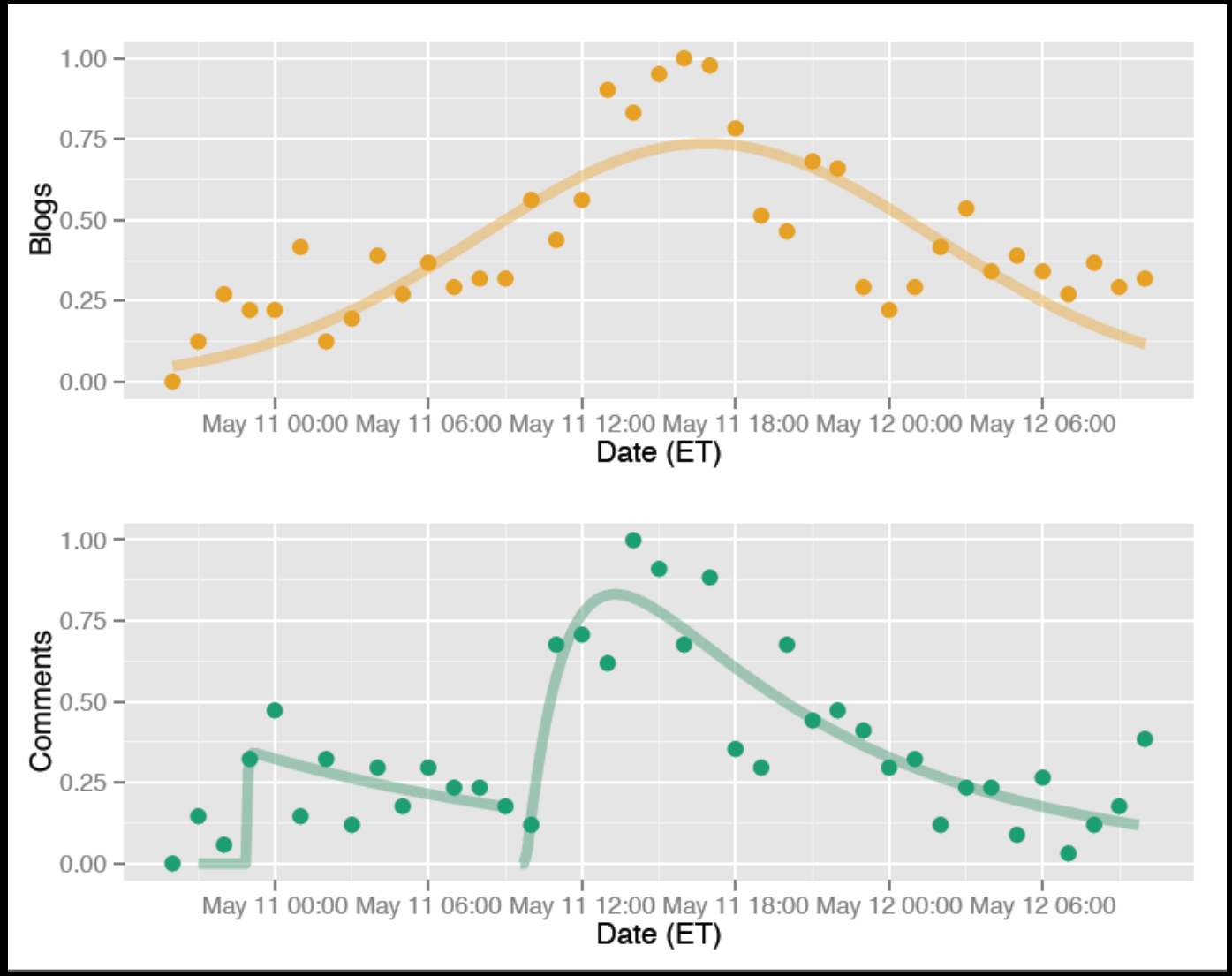
JPMorgan (-$2B)


Twitter
1.00
0.75
0.50
0.25
0.00
May 11 00:00 May 11 06:00 May 11 12:00 May 11 18:00 May 12 00:00 May 12 06:00 May 12 12
Date (ET)
Tumblr
1.00
0.75
0.50
0.25
0.00
May 11 00:00 May 11 06:00 May 11 12:00 May 11 18:00 May 12 00:00 May 12 06:00
Date (ET)
Blogs
1.00
0.75
0.50
0.25
0.00
May 11 00:00 May 11 06:00 May 11 12:00 May 11 18:00 May 12 00:00 May 12 06:00
Date (ET)
Comments
1.00
0.75
0.50
0.25
0.00
May 11 00:00 May 11 06:00 May 11 12:00 May 11 18:00 May 12 00:00 May 12 06:00
Date (ET)

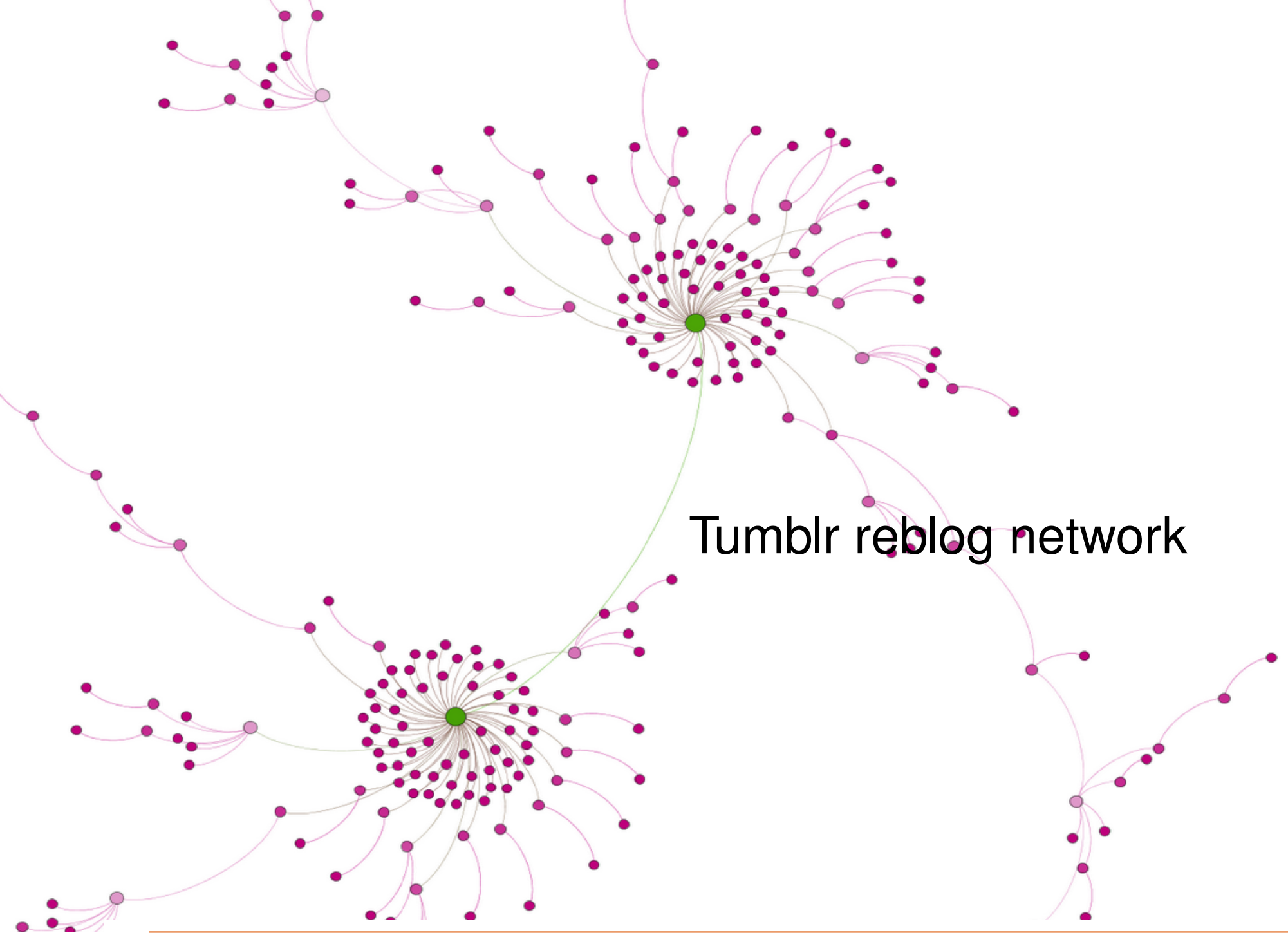
JPMorgan (-$2B)

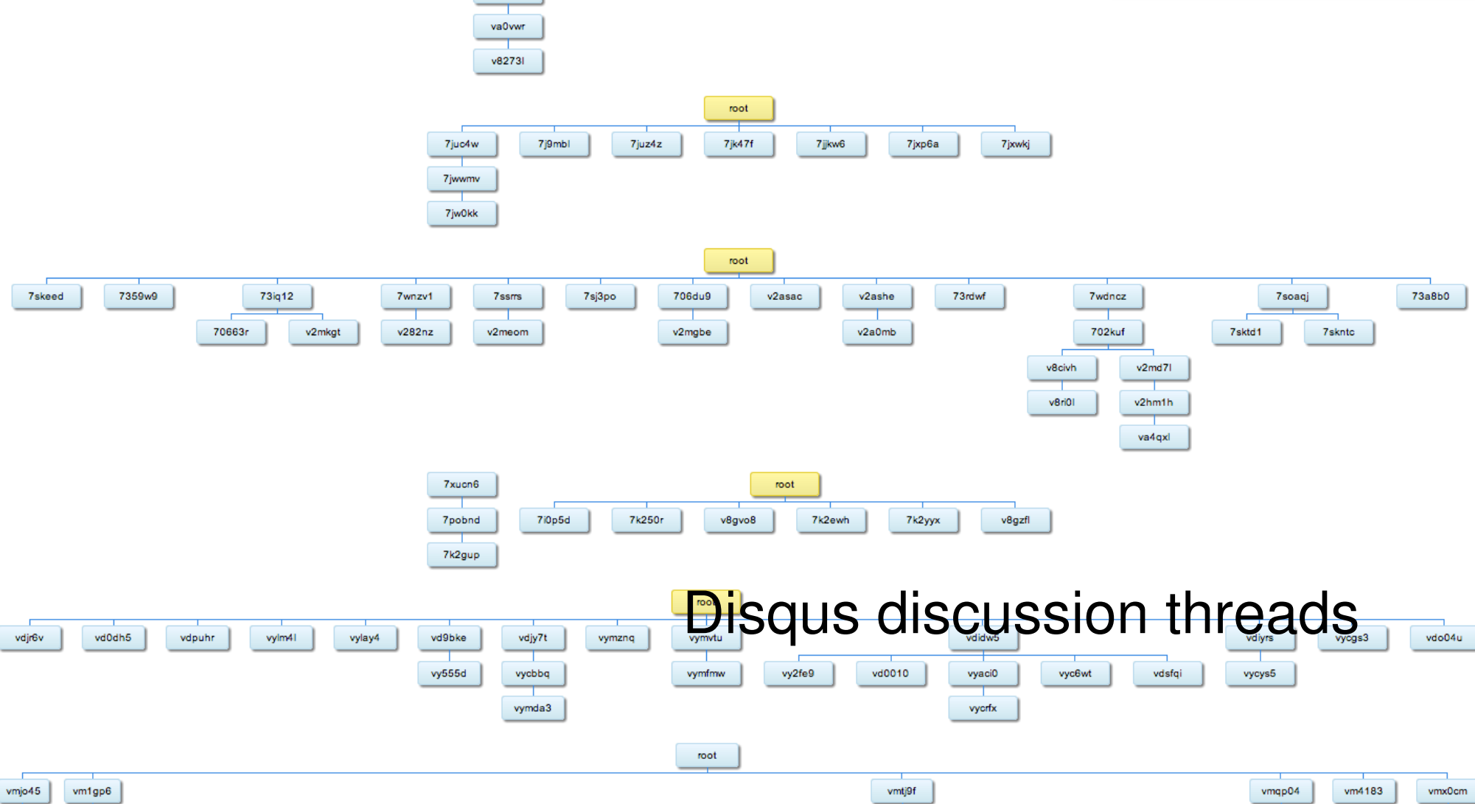

1.00
0.75
0.50
0.25
0.00
Twitter
May 11 00:00 May 11 06:00 May 11 12:00 May 11 18:00 May 12 00:00 May 12 06:00 May 12 12
Date (ET)
1.00
0.75
0.50
0.25
0.00
Tumblr
May 11 00:00 May 11 06:00 May 11 12:00 May 11 18:00 May 12 00:00 May 12 06:00
Date (ET)

JPMorgan (-$2B)


Blogs
1.00
0.75
0.50
0.25
0.00
May 11 00:00
May 11 06:00
May 11 12:00
May 11 18:00
May 12 00:00
May 12 06:00
Date (ET)
Comments
1.00
0.75
0.50
0.25
0.00
May 11 00:00
May 11 06:00
May 11 12:00
May 11 18:00
May 12 00:00
May 12 06:00
Date (ET)

Structure


Tumblr reblog network



Disqus discussion threads

Structure and Language

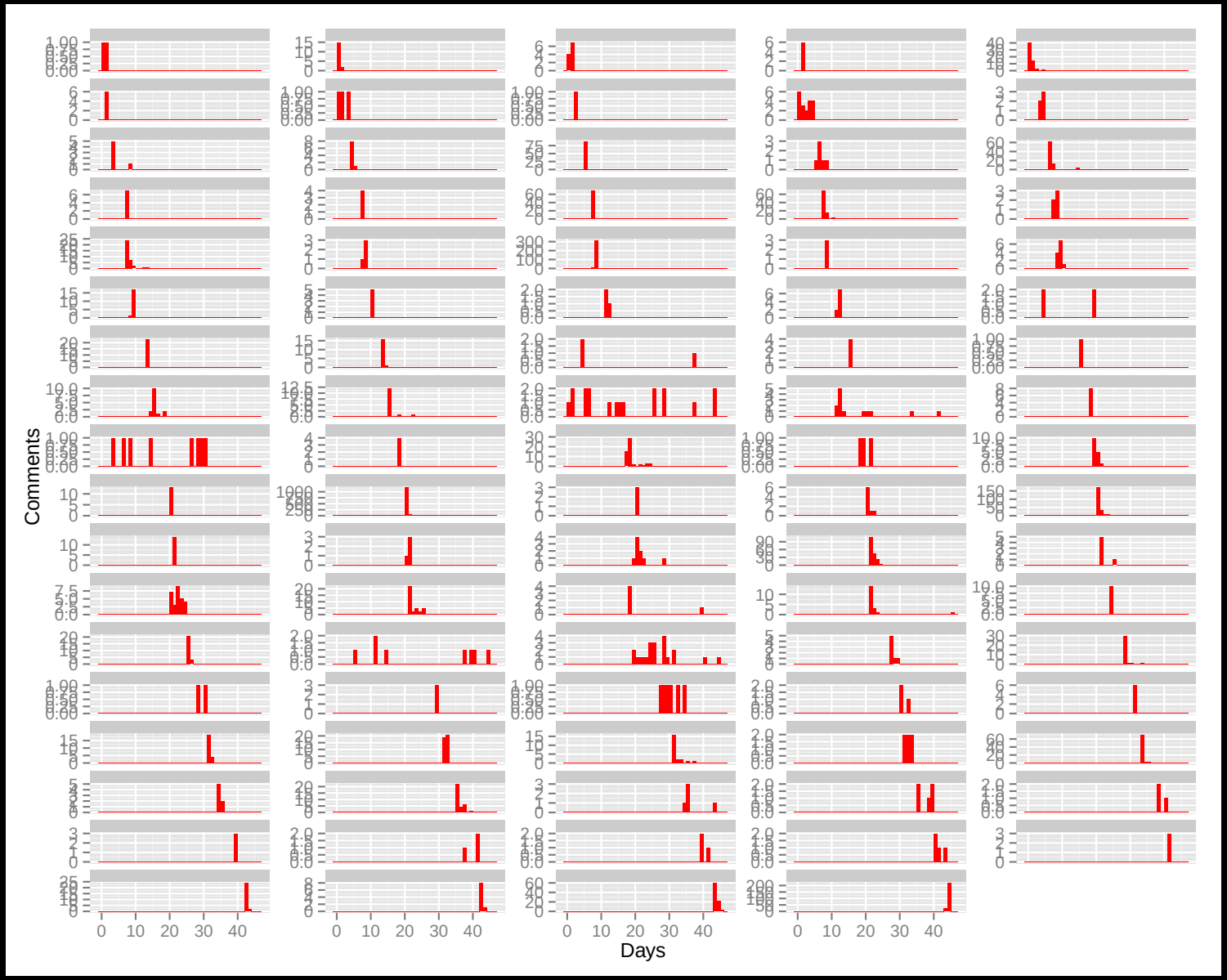
what do we talk about when they talk about X?

Apologies: Raymond Carver

Disqus threads and topics

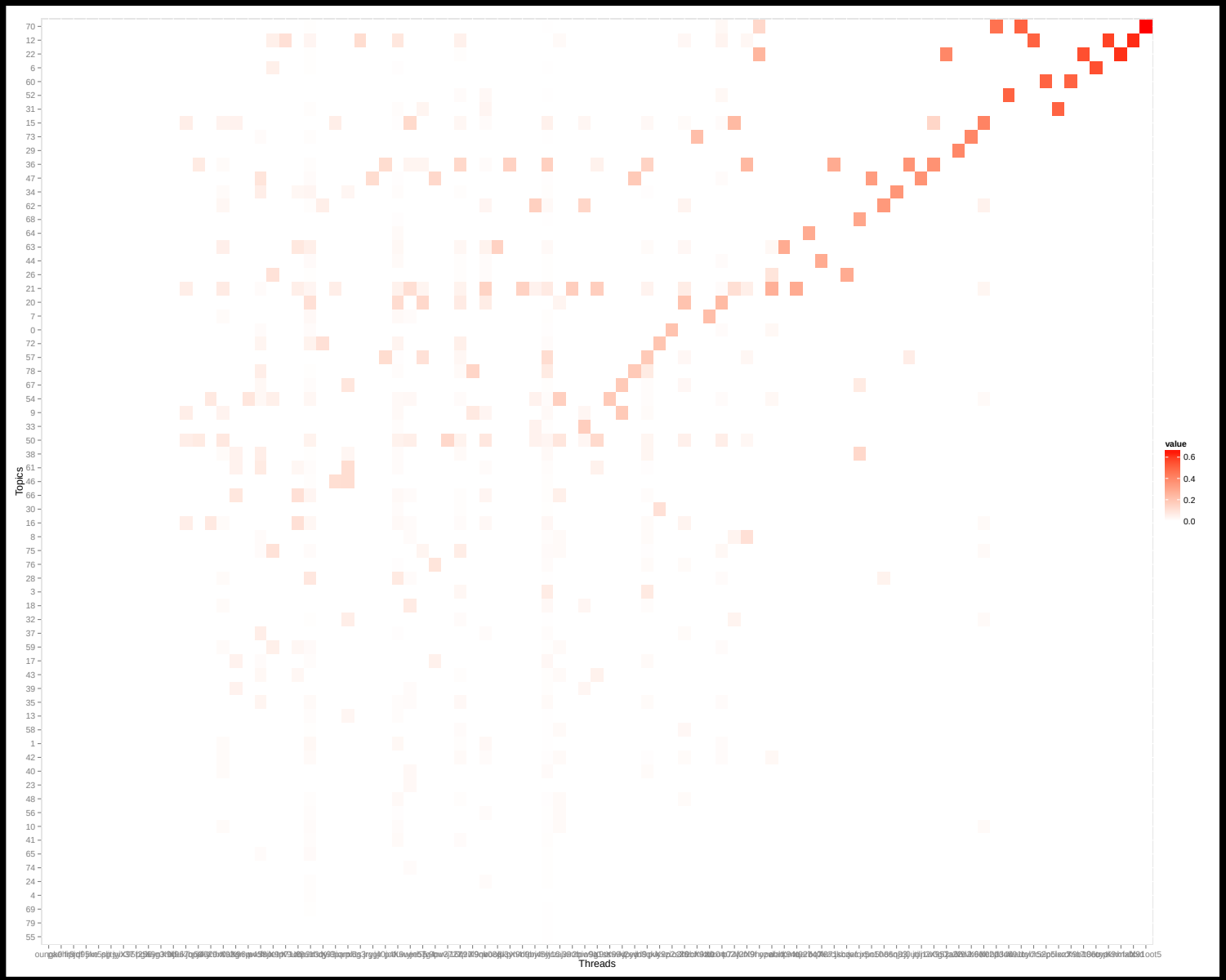
- 7 weeks data
- Key words: “texting,” “driving” and variants
- Select top threads based on mentions
  61,406 comments from 365 threads
- Select comments based on mentions 32,856 comments from 16,886 threads
- LSI: 500 features $\rightarrow$ 80 features; 80 clusters

Thread activity



Topics activity

Topics × Threads

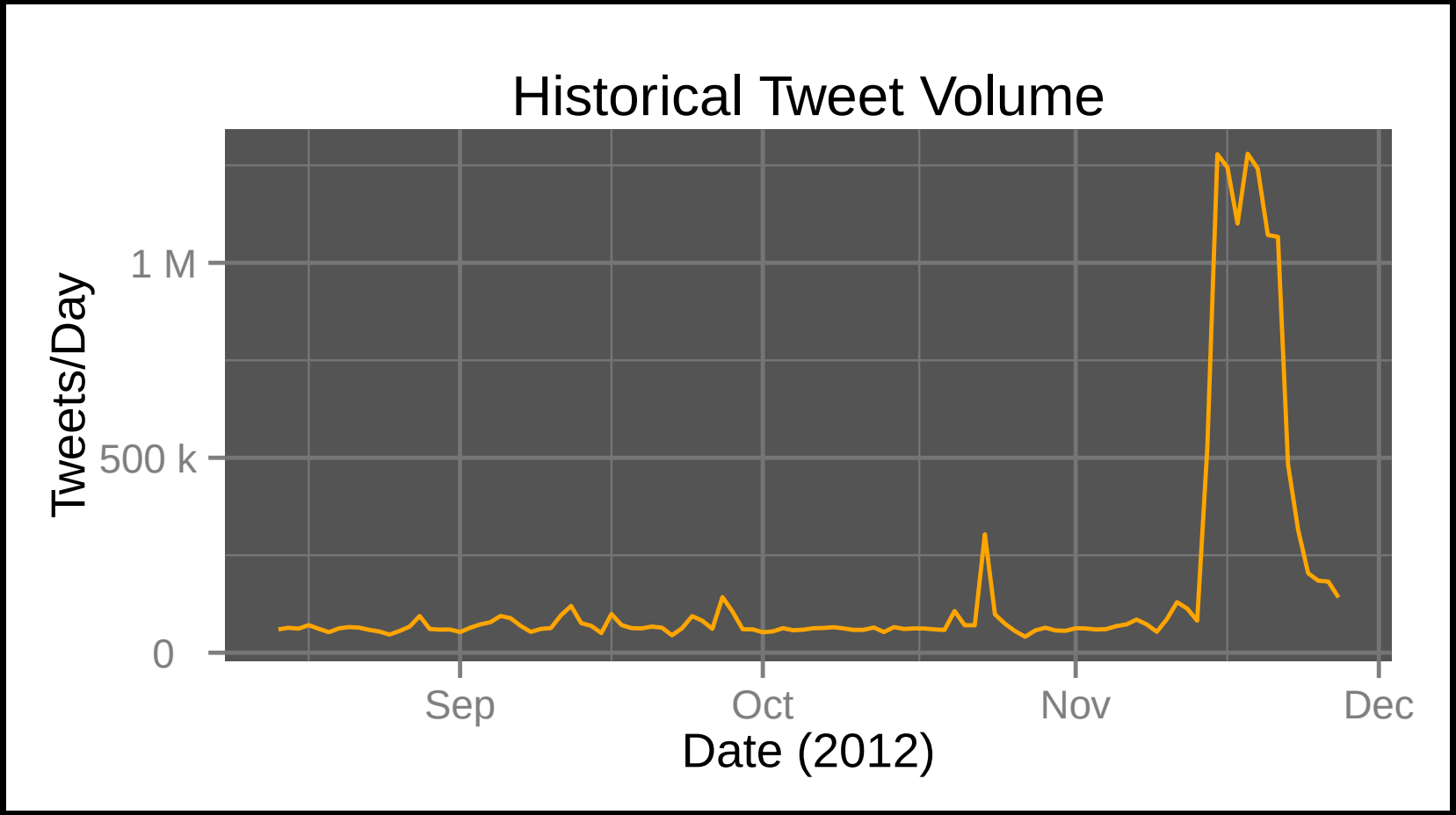


When we talk about texting and driving, we talk about …

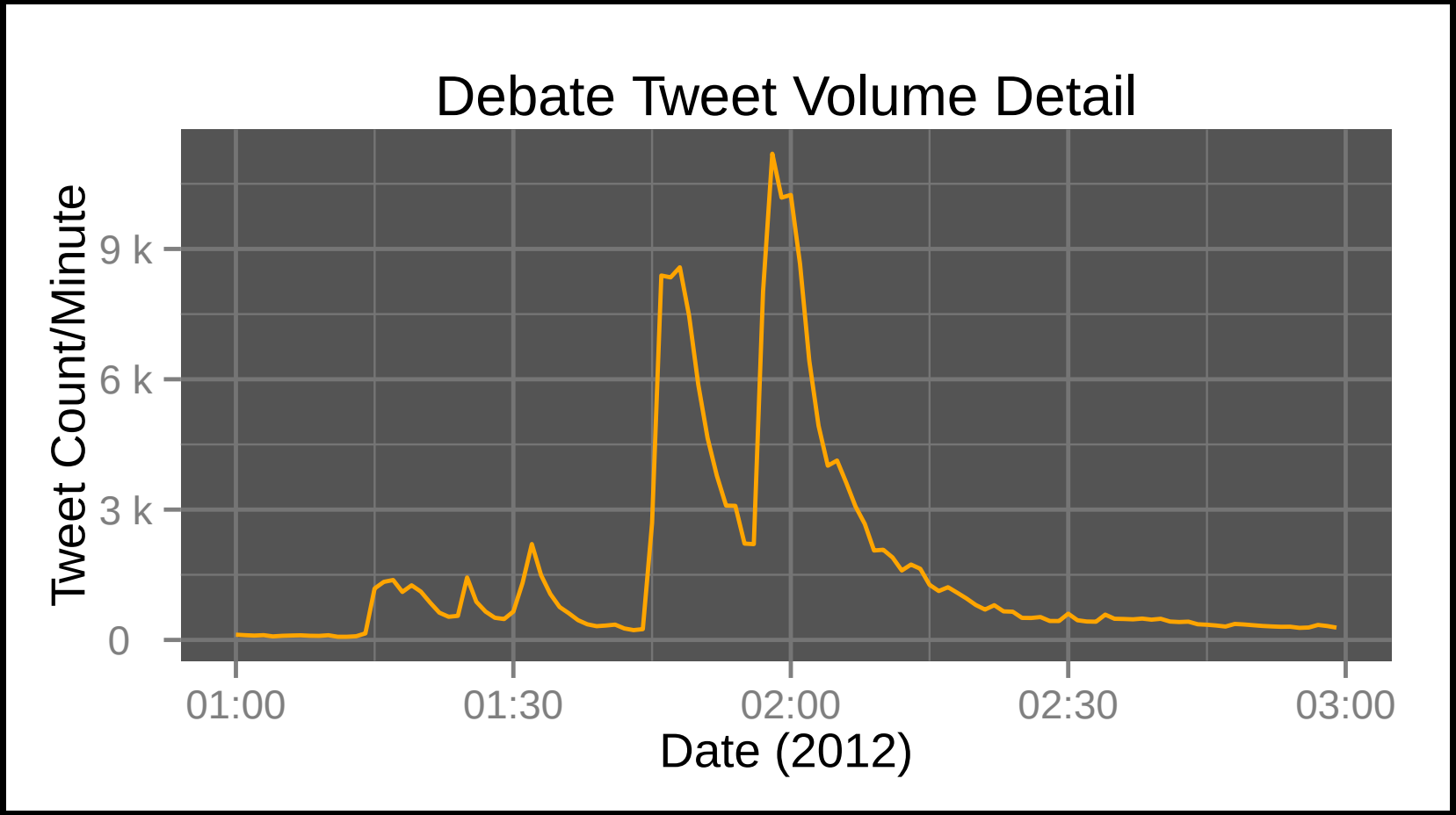
- Topic 12: poor graphic design
- Topic 50: fake ids and fake drivers licenses
- Topic 58: health/accident insurance
- Topic 62: drunk drivers
- Topic 64: buses and bus drivers
- Topic 67: bikes, bike lanes
- Topic 68: trucks and truck drivers

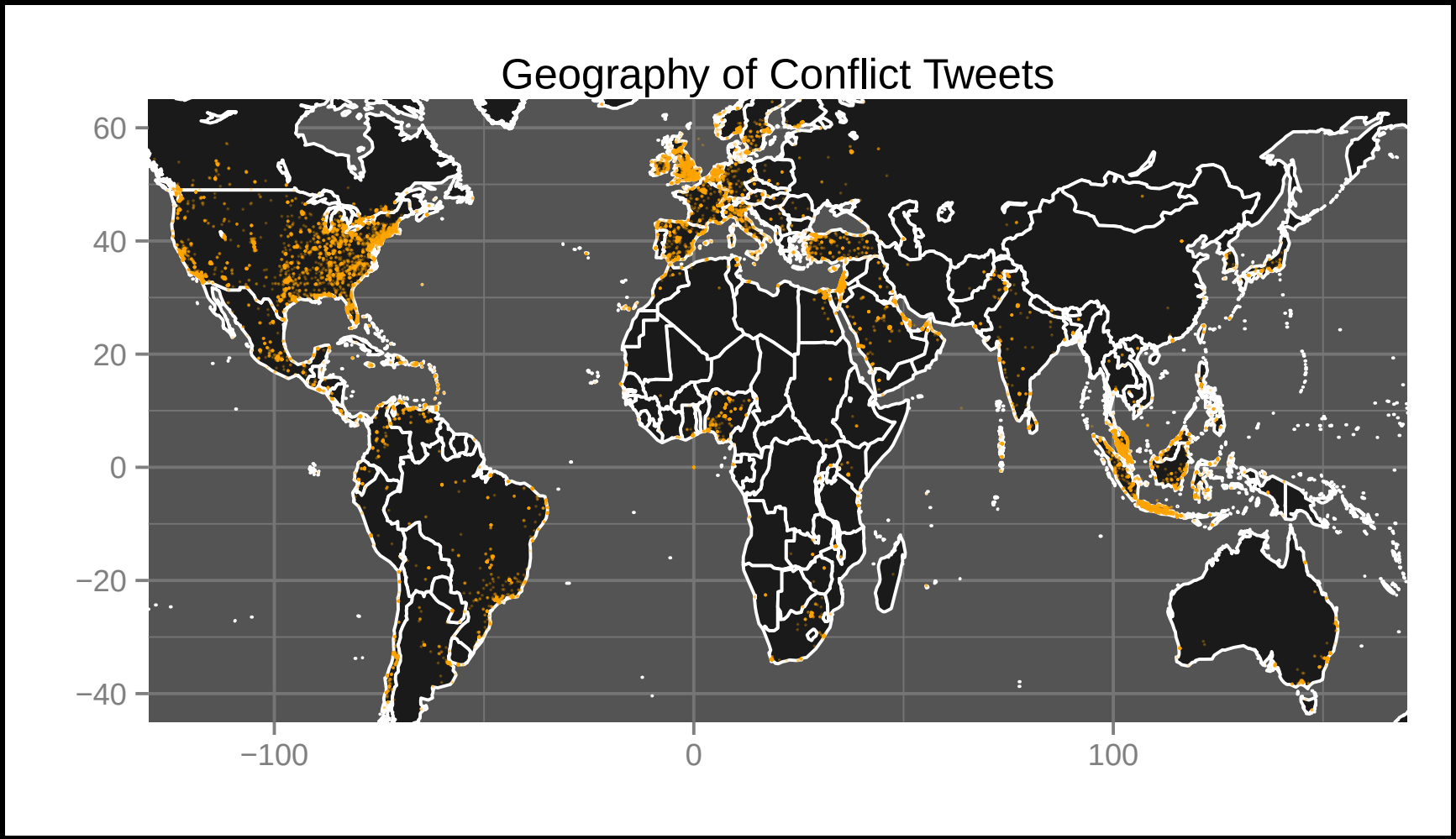
Synthesis

Hamas vs. Israel in Gaza 2012

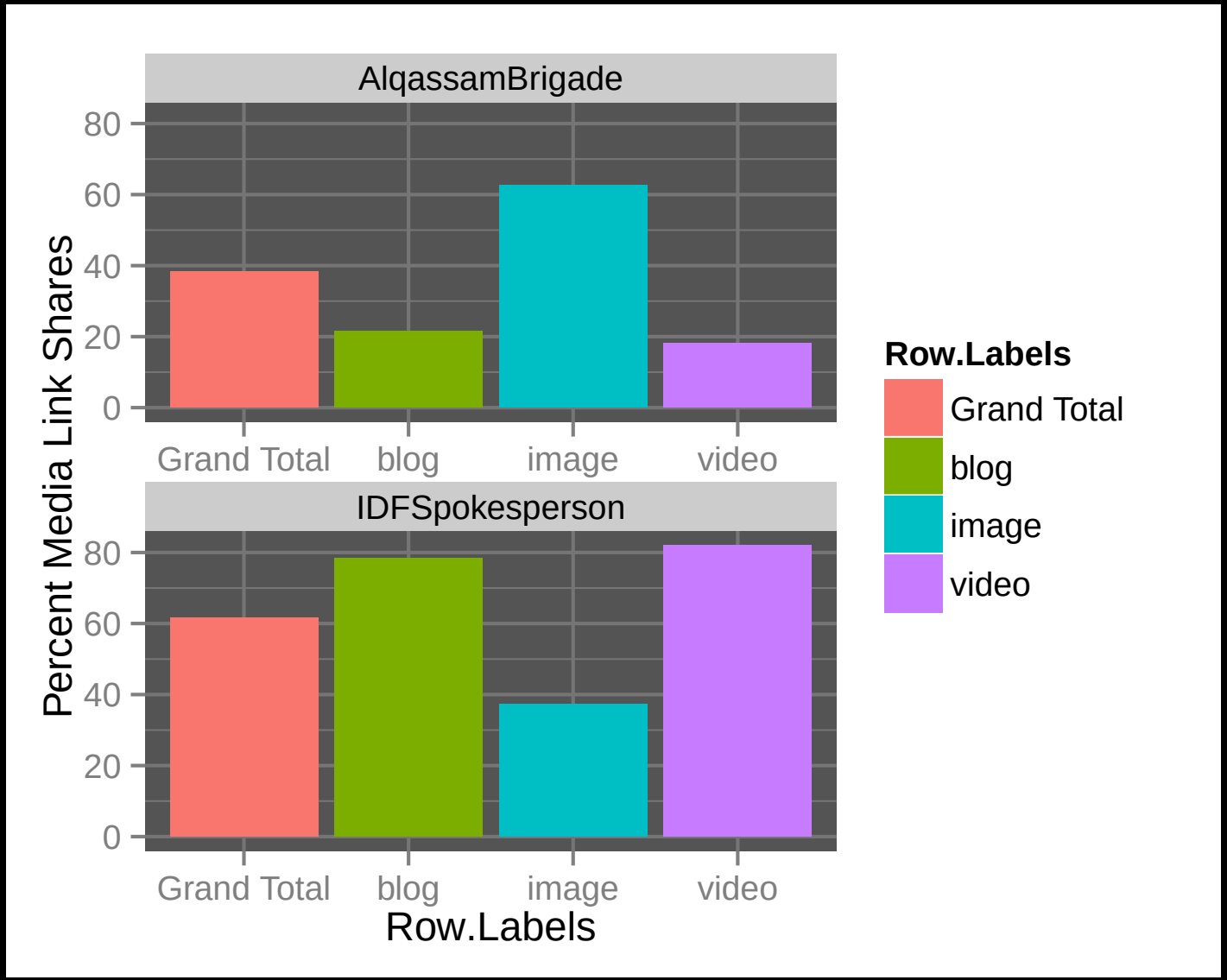


What's that leading spike?

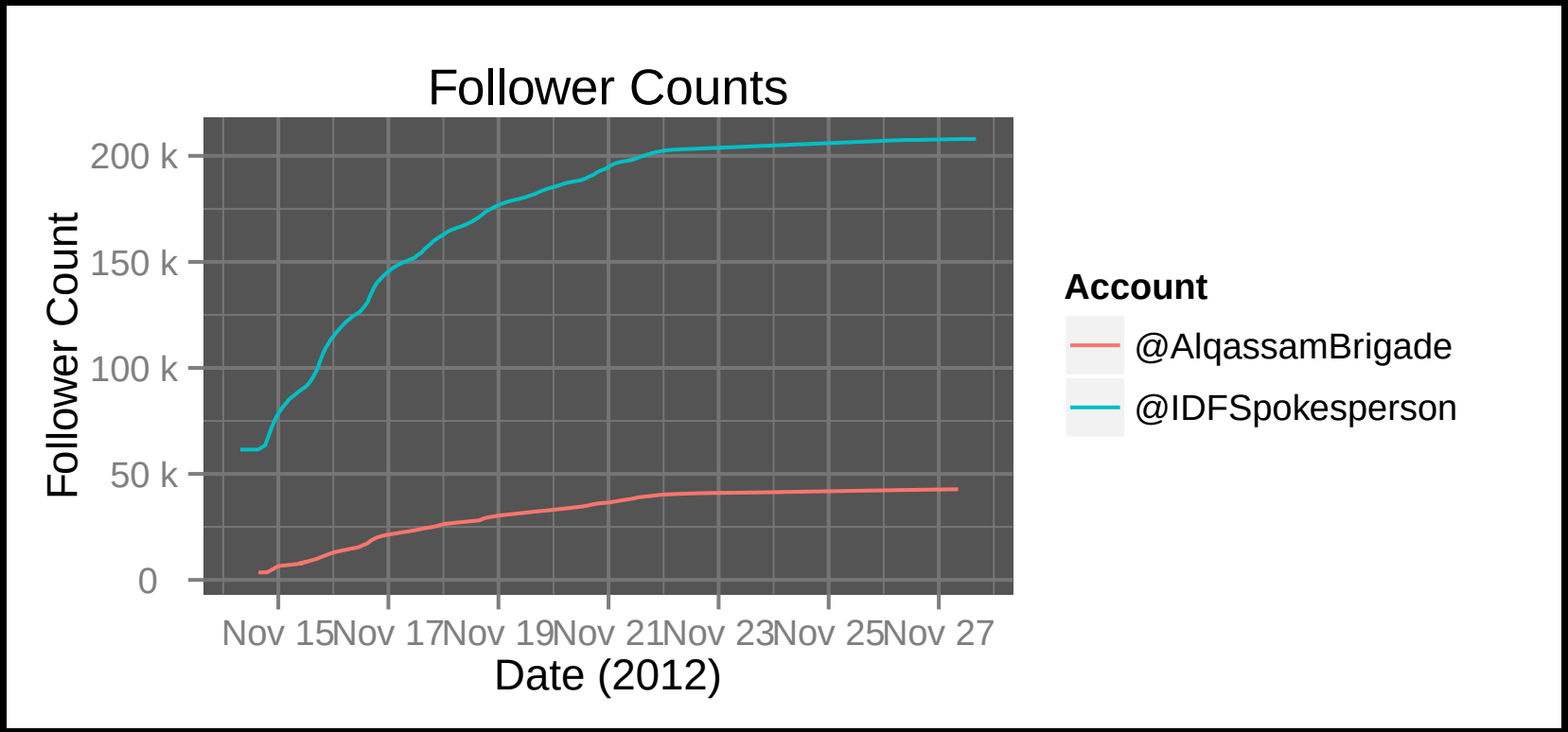



Geography of Conflict Tweets
60
40
20
0
−20
−40
−100
0
100

Key accounts use of media



Conflict grows followers



Thank you!

- Presentation at: http://github.com/DrSkippy27/Approach-to-Leveraging-Social-Data_2013