



Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose

**Fred Morstatter, Jürgen Pfeffer,
Huan Liu, Kathleen M. Carley**

Introduction

- Twitter is a social media giant.
 - 140M+ active users
 - 400M+ tweets/day
- Timely.
- Important tool for protests and real-time information during crisis.



Introduction

- Twitter shares its data.
- “Firehose” feed - 100% - costly.
- “Streaming API” feed - 1% - free.
 - Streaming API takes parameters from user.
 - Returns tweets matching parameters.
 - Samples data when volume reaches 1%.


Problem Overview



- We don't know how Twitter samples data.
- Is the sampled data from the Streaming API representative of the true activity on Twitter's Firehose?

Background

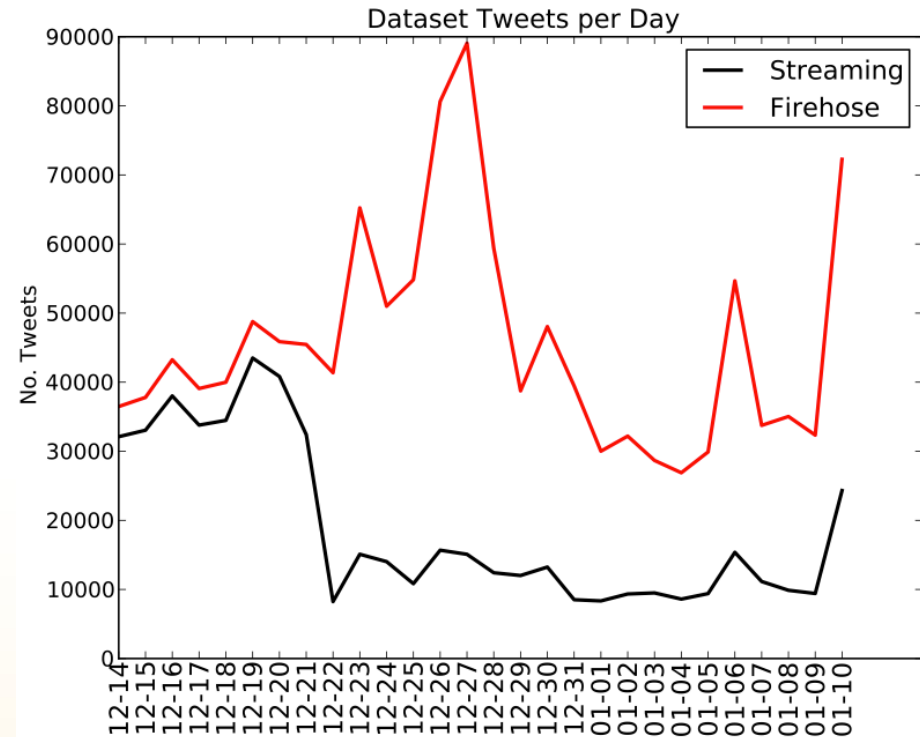
- Studying Arab Spring activity in Syria.

Keywords	Geoboxes	Users
#syria, #assad, #aleppovolcano, #alawite, #homs, #hama, #tartous, #idlib, #damascus, #daraa, #aleppo, #سوريا*, #houla	 <p>(32.8, 35.9), (37.3, 42.3)</p>	@SyrianRevo

- Given brief access to Firehose.
- Collected data from both the Streaming API and Firehose for 28 days (12/14/2011 to 01/10/2012).

Our Dataset

- 500k from Streaming API
- 1.2M from Firehose
- 42% Overall Coverage
- Daily Coverage from 17% to 89%.



Analysis

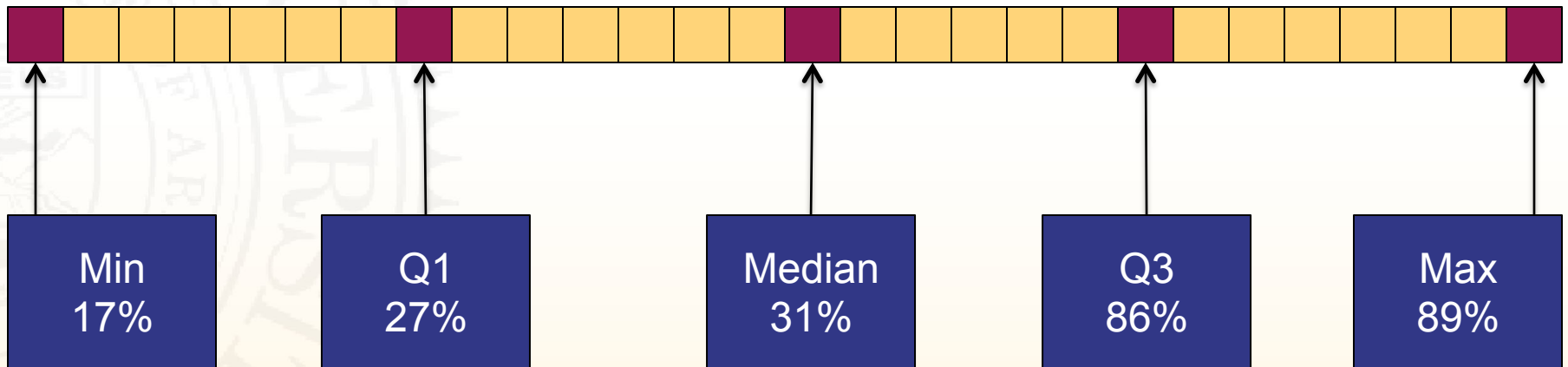


- Compare facets of the tweet data from Streaming API and Firehose.
 - Hashtags
 - Topics
 - Network Topology
 - Geographic Distribution

Days of Interest



Coverage →

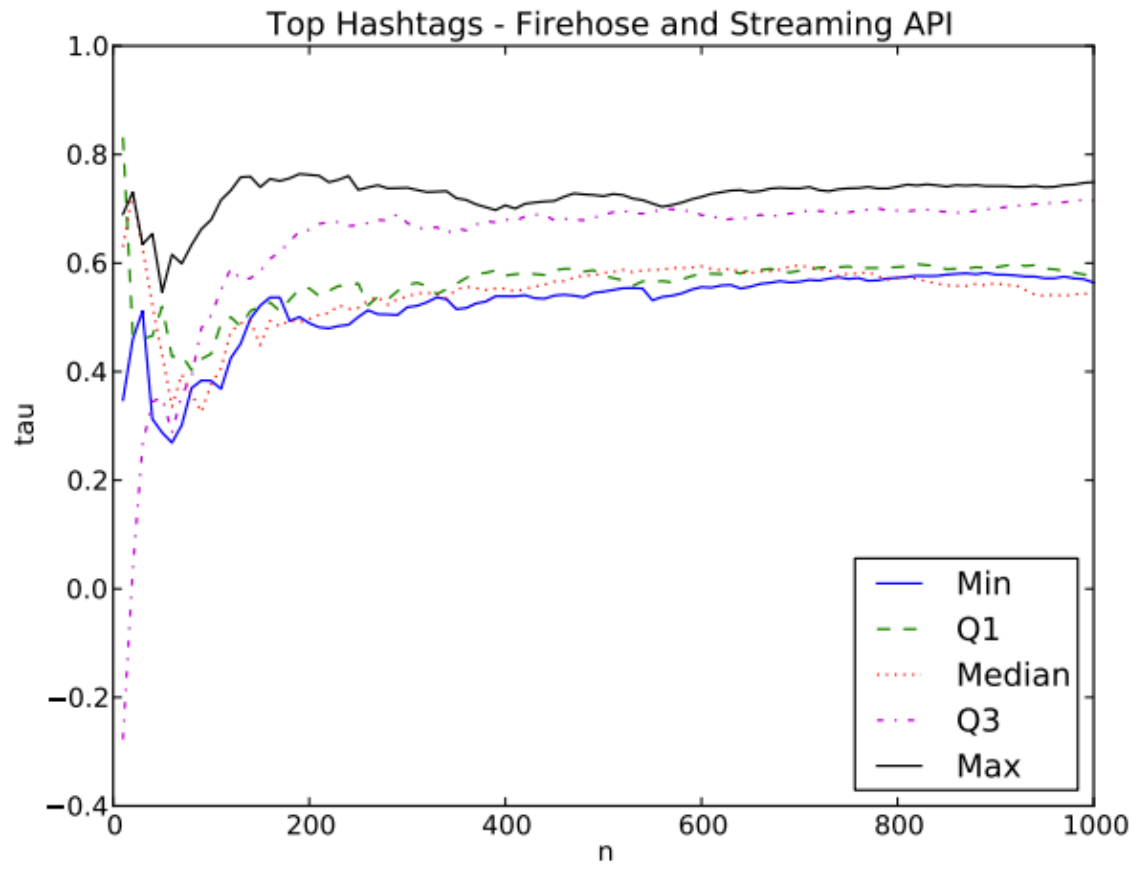


Top Hashtags



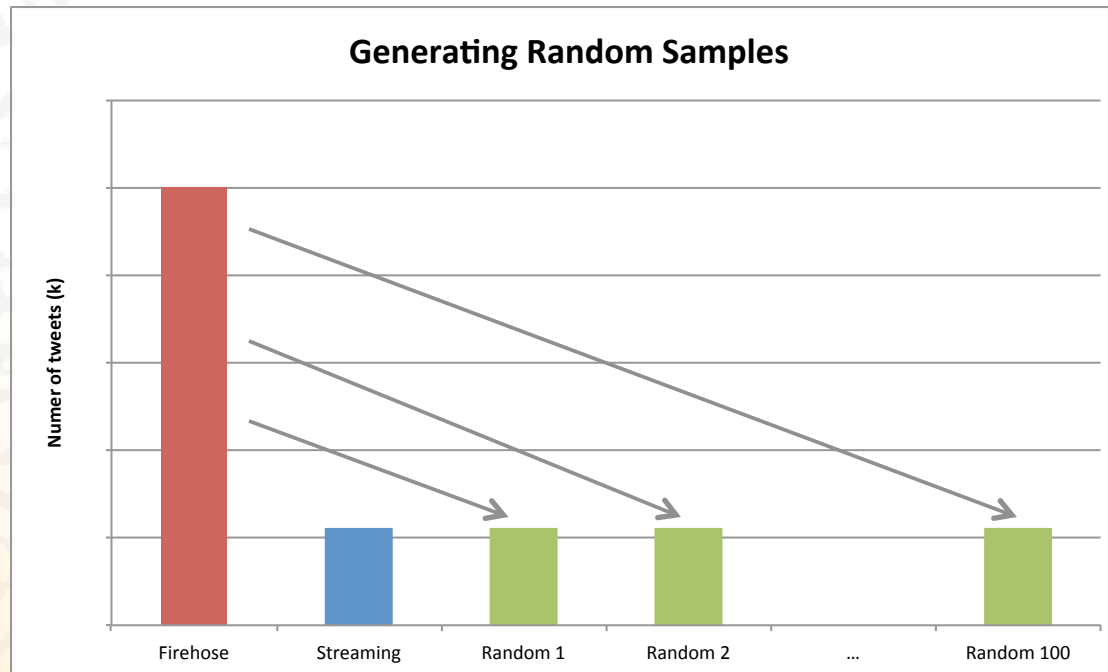
- **Question:** Are the most frequent hashtags found in the Streaming API the same as those in the Firehose?
- **Approach:** Rank the top n hashtags from each source and study the correlation between the lists.

Top Hashtags



Verification

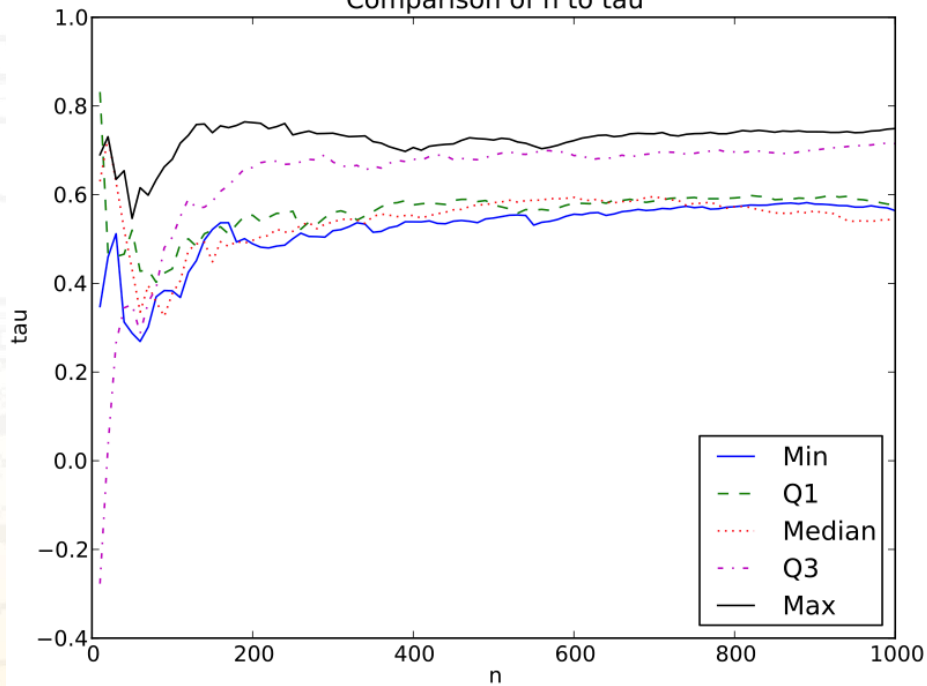
- Created 100 of our own “Streaming API” results by sampling the Firehose data.



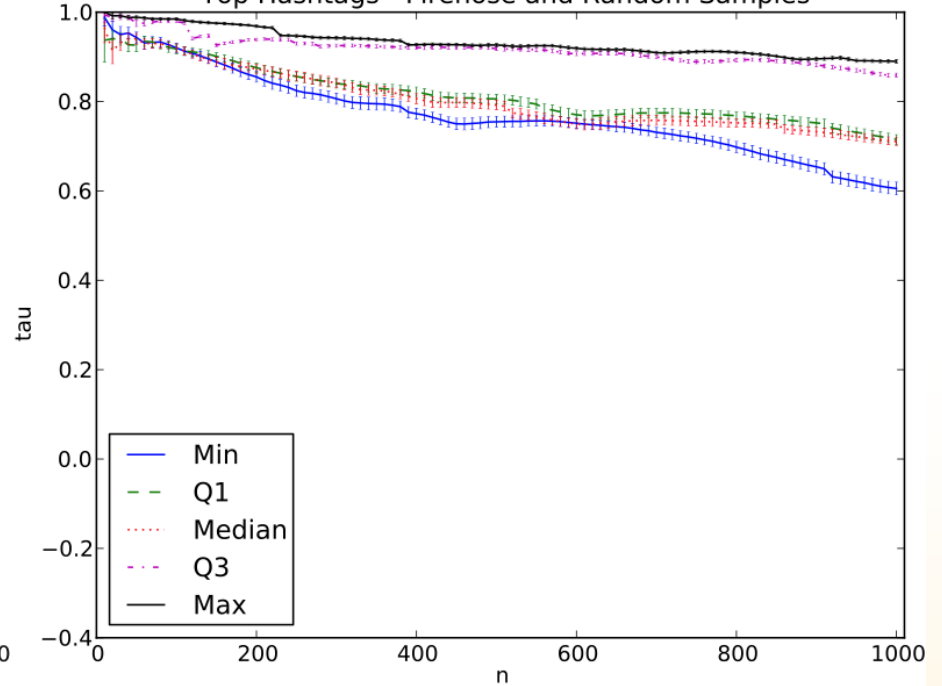
Results



Comparison of n to τ



Top Hashtags - Firehose and Random Samples

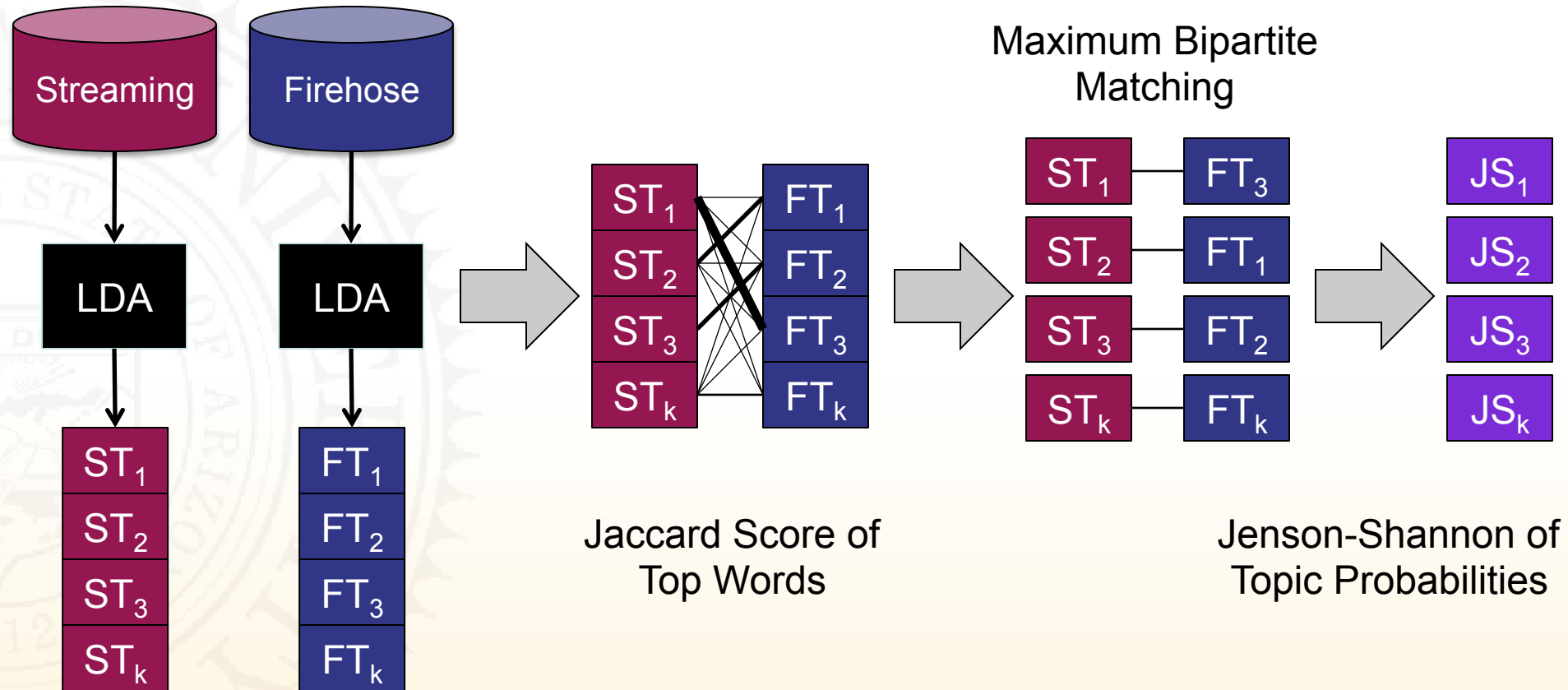


Topic Extraction

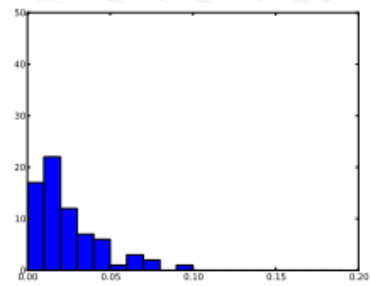


- **Question:** Are the topics extracted from the Streaming API data the same as those in the Firehose?
- **Approach:** Use LDA configured with identical parameters to generate topics from each source.

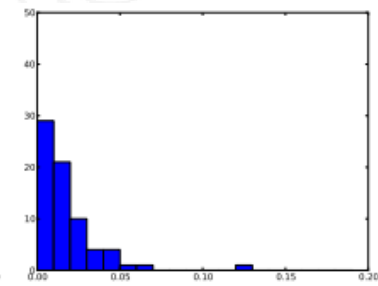
Topic Comparison



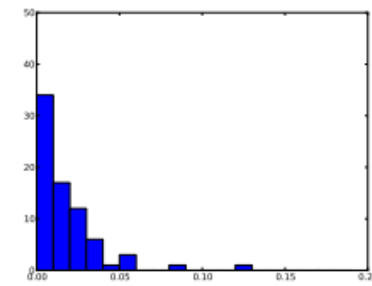
Histogram of JS Distances



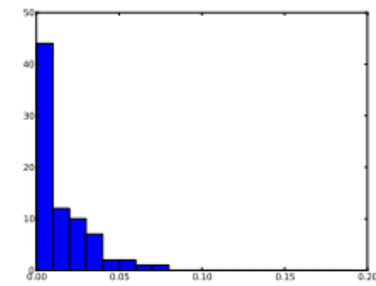
(a) Min. $\mu = 0.024$,
 $\sigma = 0.019$.



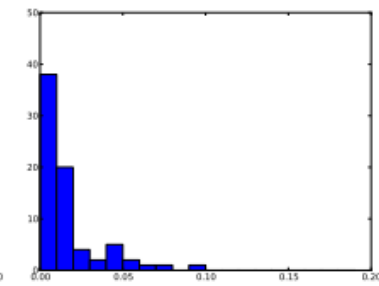
(b) Q1. $\mu = 0.018$,
 $\sigma = 0.018$.



(c) Median. $\mu = 0.018$,
 $\sigma = 0.020$.

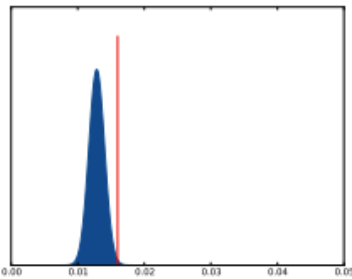
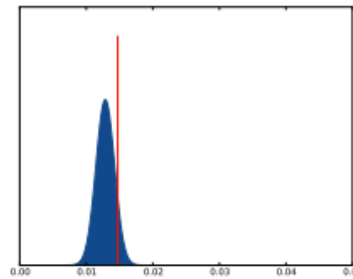
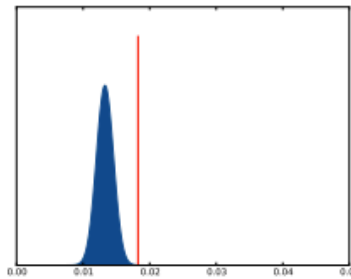
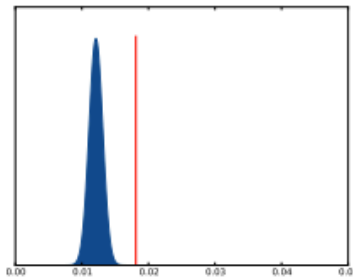
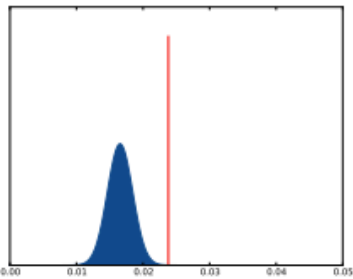


(d) Q3. $\mu = 0.014$,
 $\sigma = 0.016$.



(e) Max. $\mu = 0.016$,
 $\sigma = 0.018$.

Comparison with Random Samples



(a) Min. $S = 0.024$,
 $\hat{\mu} = 0.017$,
 $\hat{\sigma} = 0.002$,
 $z = 3.500$.

(b) Q1. $S = 0.018$,
 $\hat{\mu} = 0.012$,
 $\hat{\sigma} = 0.001$,
 $z = 6.000$.

(c) Median. $S = 0.018$,
 $\hat{\mu} = 0.013$,
 $\hat{\sigma} = 0.001$,
 $z = 5.000$.

(d) Q3. $S = 0.014$,
 $\hat{\mu} = 0.013$,
 $\hat{\sigma} = 0.001$,
 $z = 1.000$.

(e) Max. $S = 0.016$,
 $\hat{\mu} = 0.013$,
 $\hat{\sigma} = 0.001$,
 $z = 3.000$.

All days but Q3 see *significantly* better topics with random samples.

Network Topology



- **Question:** Do the Streaming API and the Firehose agree upon the most central users in the retweet network?
- **Approach:** Extract the retweet network and find the agreement through common centrality measures.

Network Topology

- User-User Retweet Networks, aggregated by day.
 - In-Degree Centrality
 - Betweenness Centrality
 - Potential Reach Centrality
- Compare agreement of central users between the two datasets.

Network Topology

Measure	Top-k	Average Agreement (min-max)	All 28 Days
In-Degree	10	4.21 (0-9)	4
In-Degree	100	53.4 (36-82)	73
Betweenness	100	54.8 (41-81)	55
Potential Reach	100	59.2 (32-83)	80

On average, the Streaming API finds ~50% of the key users.

Geographic Distribution

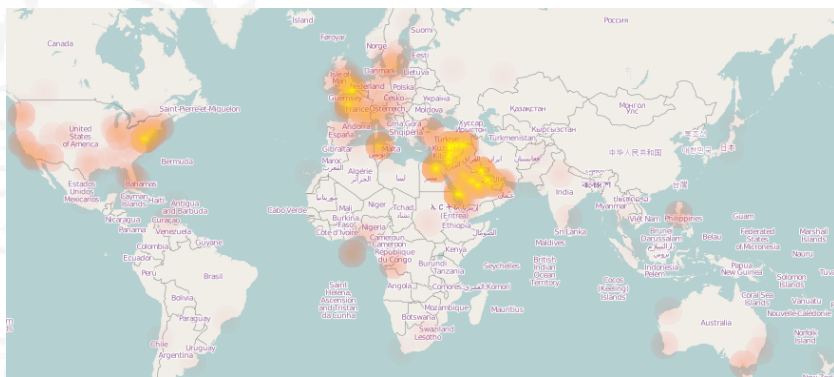


- **Question:** Is the distribution of geotagged tweets in the Streaming API data the same as in the Firehose?
- **Approach:** Analyze the difference in the continental distribution of geotagged tweets.

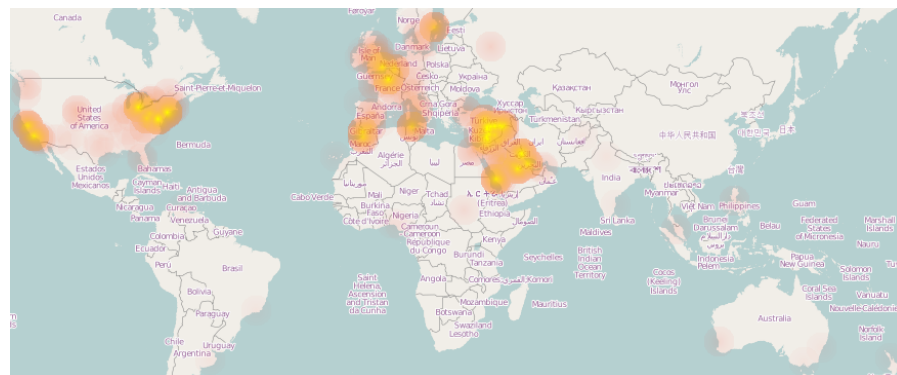
Geographic Distribution



Firehose



Streaming API



- We get **>90%** of all geotagged tweets!
- No significant difference in location distribution.

Geographic Distribution

Continent	Firehose	Streaming	Error
Africa	156 (5.74%)	33 (3.10%)	-2.64%
Antarctica	0 (0.00%)	0 (0.00%)	0.00%
Asia	932 (34.26%)	321 (30.11%)	-4.15%
Europe	300 (11.03%)	139 (13.04%)	+2.01%
Mid-Ocean	765 (28.12%)	295 (27.67%)	-0.45%
N. America	607 (22.32%)	293 (27.49%)	+5.17%
Oceania	54 (1.98%)	15 (1.41%)	-0.57%
S. America	3 (0.11%)	2 (0.19%)	+0.08%
Total	2720 (100.00%)	1066 (100.00%)	0.00%

Is the Sample Good Enough?

Success with Streaming API is strongly dependent on two factors:

- Analysis Performed.
 - Top hashtags and topics have some bias.
 - Network can give reasonable indication of top users.
 - The geographic facet is almost perfect.
- Amount of data from the Streaming API in relation to the Firehose.

Future Work



- Apply methodology to more datasets.
- Deeper study into how measures are altered by sampling.
- Discover bias automatically without Firehose data.

Acknowledgments



- We would like to thank the Office of Naval Research for their continued support through the grants N000141010091 and N000141110527.
- We would also like to thank the members of the DMML Lab.

Questions?



Network Sampling

	Firehose		Streaming API	
Metrics	avg.day	28 days	avg.day	28 days
nodes	6,590	73,719	2,466 (37.4%)	30,894 (41.9%)
links	10,173	204,022	3,667 (36.0%)	76,750 (37.6%)
$D_{in} > 0$	25.1%	19.3%	32.4%	20.5%
$max(D_{in})$	341	2,956	167.3	1,252
main comp.	5,609	70,383	2,069	28,701
main comp. %	84.6%	95.5%	82.5%	92.9%
Clust.Coeff.	0.029	0.053	0.033	0.050
DC_{in} Centr.	0.059	0.042	0.085	0.043
BC Centr.	0.010	0.053	0.010	0.050
$PReach$ Centr.	0.130	0.240	0.156	0.205

References



- Slide 2 Protest Image:
<http://www.dibussi.com/2011/03/the-digital-disconnect-and-misconceptions-about-revolution-20.html>



Network Topology

- User-User Retweet Networks.
- In-Degree, Betweenness, and Potential Reach Centrality, aggregated by day.
- Compare agreement of central users in the dataset.

@John: "RT @Bob: This is a tweet."

@Mike: "RT @Bob: This is another tweet."

