The Tweets They are a-Changin': Evolution of Twitter Users and Behavior

Yabing Liu†, Chloe Kliman-Silver§, Alan Mislove†

†Northeastern University
§Brown University

ICWSM 2014

Twitter

Twitter: Popular microblogging platform

- Started in 2006 as SMS service
- Over 200 million monthly active users today
- Used by many organizations and individuals

Result: Significant amounts of Twitter research

- Twitter makes data easy to access
- Significant public data available

Examine how human society functions at scale

What have people studied?

Tweeting behavior

over 768,000 tweets in 1 month -- retweets [Macskassy and Michelson, ICWSM'11]

over 650,000 tweets over 1 month -- tweet contents [Macskassy, ICWSM'12]

over 476 million tweets over 7 months -- hashtags [Yang et al., WWW'12]

1.6 million deleted tweets over 1 week -- deletion of tweets [Almuhimedi, et al., CSCW'13]

Twitter user demographics

about 100,000 users from 3 datasets -- user lang [Krishnamurthy, et al., WOSN'08]

about 32 million English tweets over 1 month -- user location [Hecht et al., CHI'11]

The talk

Goal: How Twitter changes over time?

Collect over 37 billion tweets spanning over 7 years

Examine the evolution of the (public) Twitter ecosystem

- Whether prior results still hold
- Whether the (often implicit) assumptions of proposed systems are still valid

Outline

First Twitter dataset (2006-2009)

Dataset	Date range	Users	Tweets	Date collected	Tweets	Users
Crawl	21/03/2006 – 14/08/2009	25,437,870	1,412,317,185	14/08/2009	~100%	~100%

Crawl:

Collected by previous work [Cha et al. 2010]

Iteratively download the 3,200 most recent tweets of all public users alive at the time

Notes:

Does not include any tweets deleted before August 14, 2009

The user information is as-of August 2009.

Second Twitter dataset

Dataset	Date range	Users	Tweets	Date collected	Tweets	Users
Gardenhose	15/08/2009 – 31/12/2013	376,876,673	36,495,528,785	Time of tweet	~10–15%	~30.61%

Gardenhose:

Twitter 'Gardenhose' public stream

https://stream.twitter.com/1.1/statuses/sample.json, with elevated access.

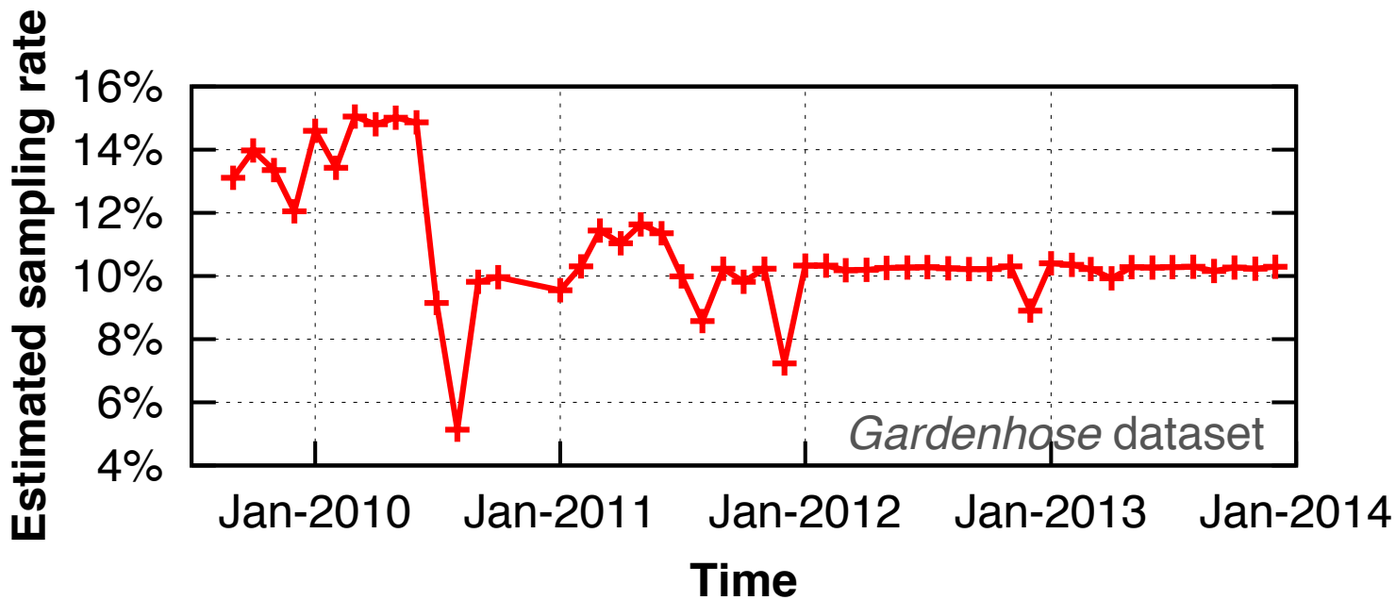
A random sample of all public tweets(tweet + user)

Notes:

With a bias towards more active users

Twitter does not inform us when user leave the network.

The sampling rate of



Notes:

Reason: Twitter does not state the rate.

A sampling rate of ~15% until July 2010, and ~10% since then

Our measurement infrastructure was down between Oct. 18, 2010 and Dec. 31, 2010.

Third Twitter dataset

Dataset	Date range	Users	Tweets	Date collected	Tweets	Users
UserSample	21/03/2006 – 31/12/2013	1,210,077	–	12/31/2013	~0.1%	~0.1%

UserSample:

- A random sample of users
- Generate 2 million random user_ids between 1 and 1,918,524,009
- Query Twitter in Jan 2014 for the most recent info on each user
- Both via the Twitter API and the web site
- 1,210,077 (60.51%) user_ids were ever assigned to a user.

Together:

- We have over 388 million unique users and over 37 billion tweets.
- For each analysis, we use the most appropriate dataset.

Outline

~~1 Motivation~~

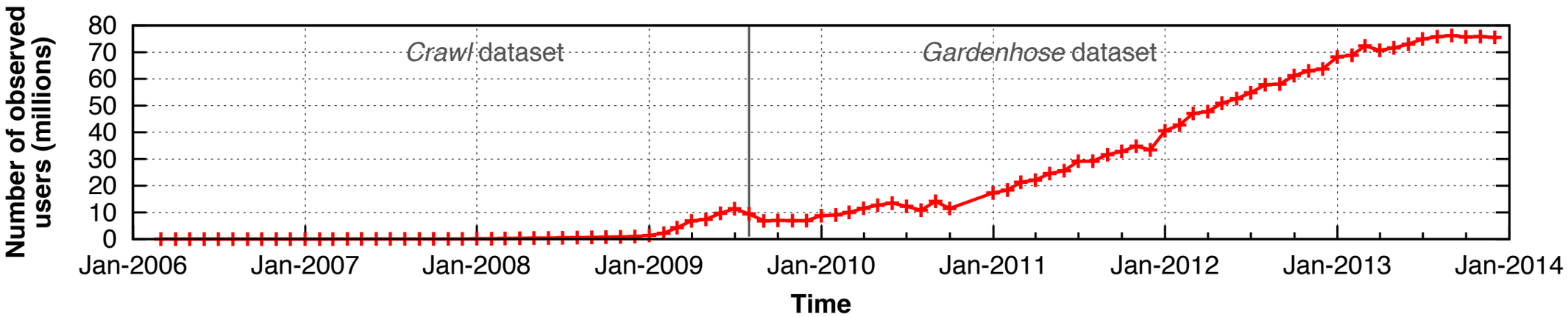
~~2 Goals~~

~~3 Twitter Datasets~~

4 User characteristics

5 Tweeting behavior

How is Twitter growing?



Observations:

Rapid growth from 2009 through 2012 and a leveling-off of the number in 2013

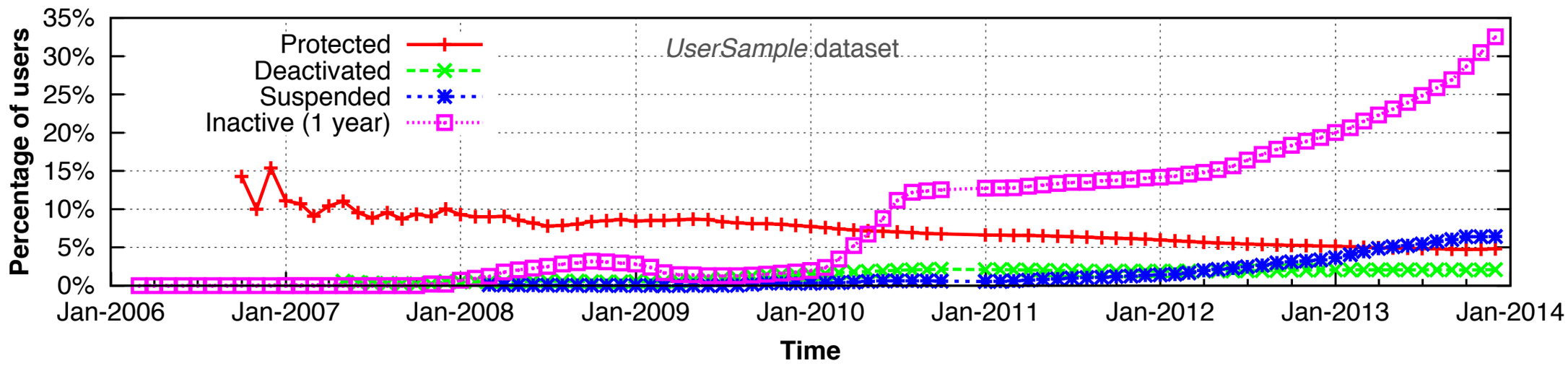
June 2013: Over 73 million users tweet VS. 218 million reported active users

Reasons:

Users from a random 10% sample of tweets

Twitter's definition of an active user: login activity, not tweeting activity

How many users are leaving



Observations:

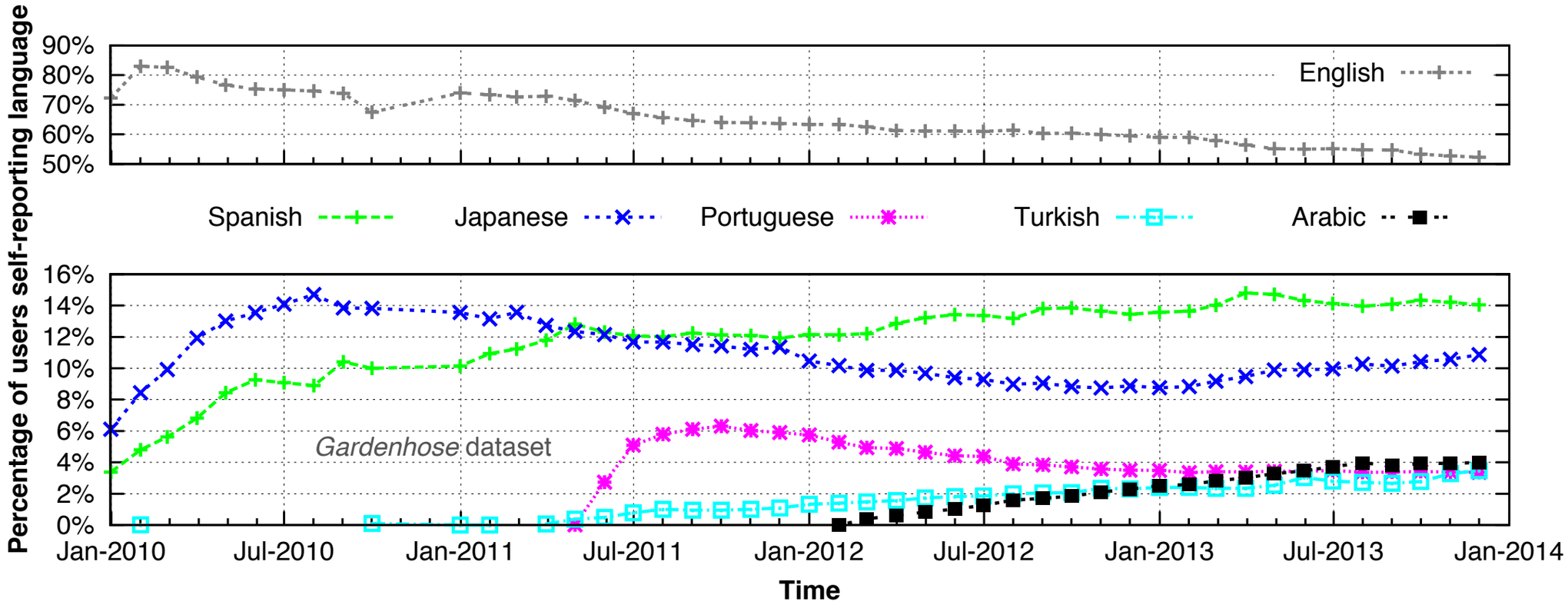
Protected accounts: goes down to 4.8% by 2013 -- most new accounts are public

Deactivated accounts: a relatively stable 2% of users

Suspended accounts: over 6% of entire Twitter users by 2013

Inactive accounts: up to 32.5% of all accounts by the end of 2013

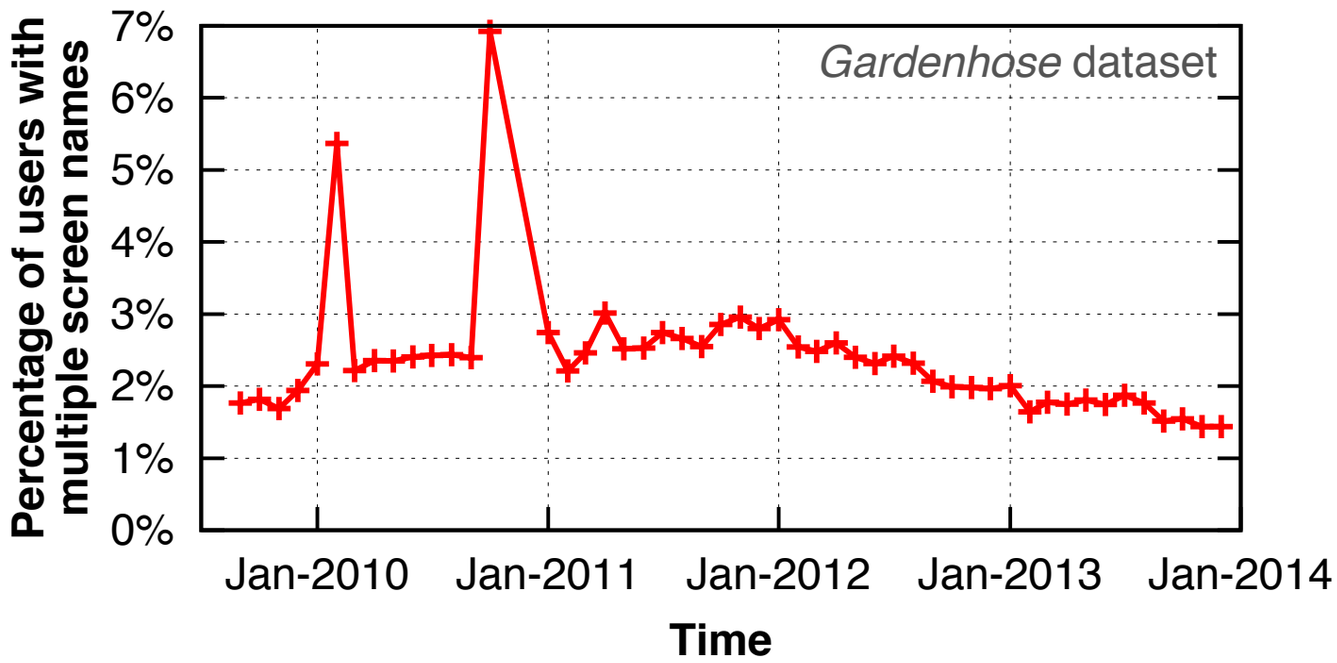
What languages do users speak?



Observations:

- The self-reported lang field since Jan.12, 2010
- English: a steady and continuing decrease of users from 83% to 52%
- Spanish and Japanese: approximately 10%
- More diverse and global

When do users change screen name?



Observations:

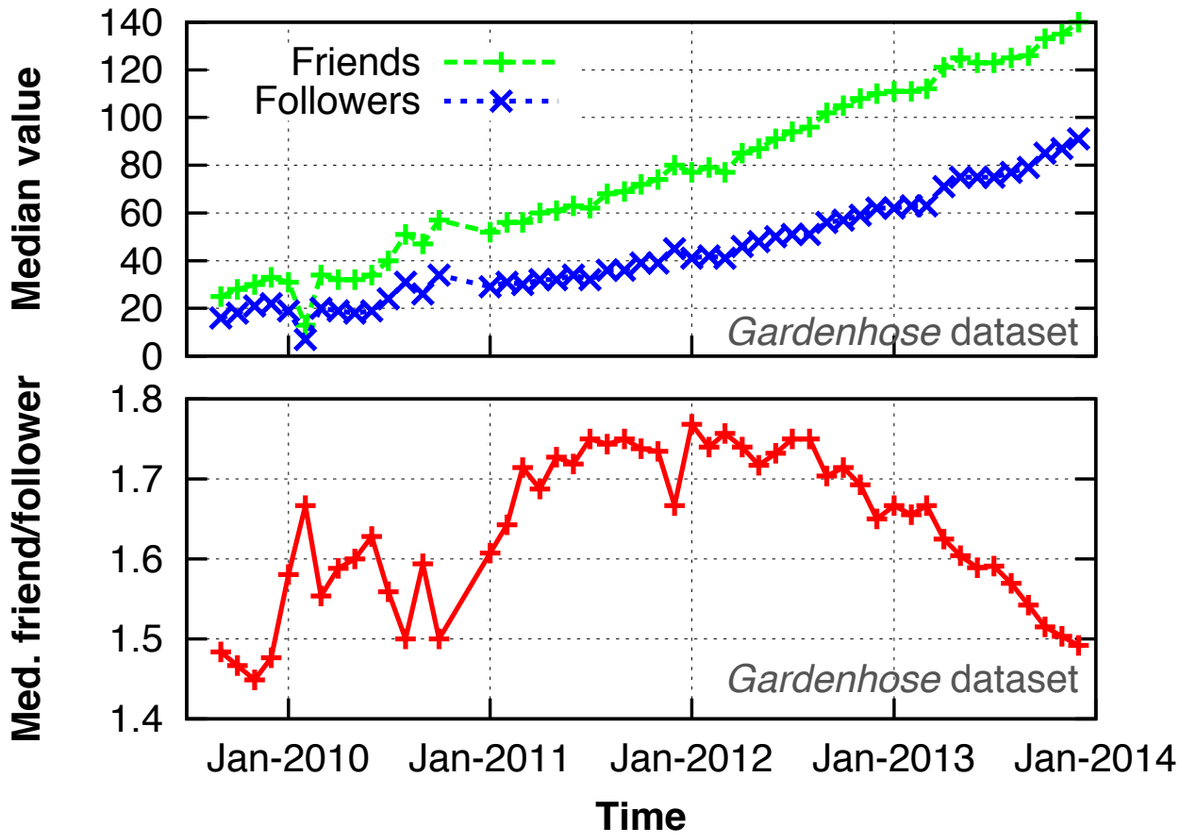
Up to 3% of users change their screen names every month.

Example: @Barack to @BarackObama

The "spikes" in Feb and Oct 2010: Twitter opened up old, inactive screen names to be reclaimed by active users.

To track users: user_id

How social are Twitter users?



Observations:

A dramatic increase in the median followers/friends count of almost 400% from 2009 to 2013

The distribution of followers is much more biased than the distribution of friends. => Twitter is disassortative.

The rise of Twitter follower spam in 2010 and 2011

Outline

~~1 Motivation~~

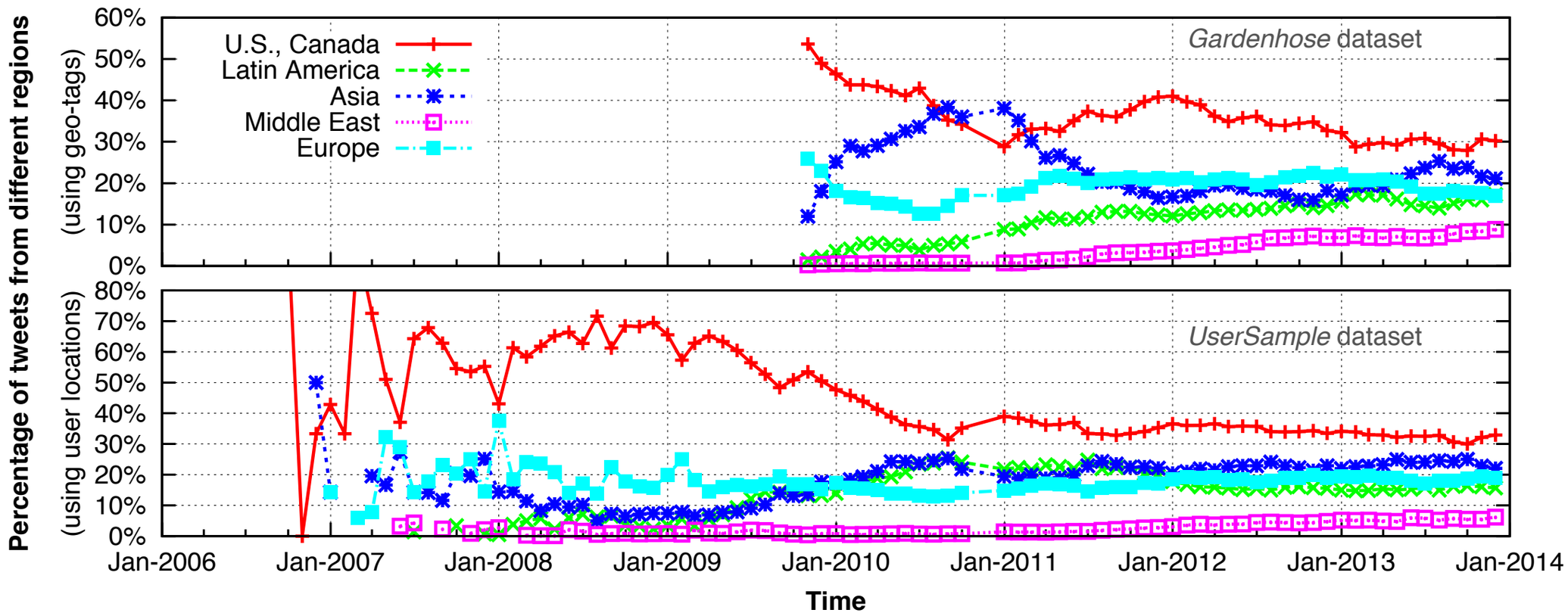
~~2 Goals~~

~~3 Twitter Datasets~~

~~4 User characteristics~~

5 Tweeting behavior

Where are the tweets coming



Information:

The self-reported, unformatted location field attached to user profile [Bing Maps]

The geo field(lat/lon) attached to some tweets since Nov. 2009 [GIS shape files]

42.4% of users provide a location string interpretable by Bing.

1.23% of tweets have included geo-tags.

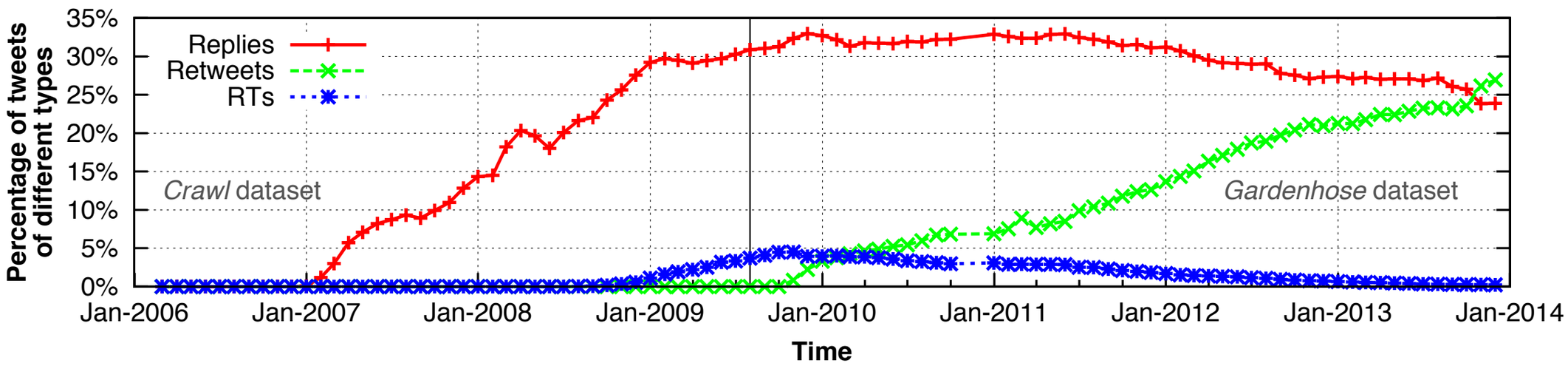
Observations:

U.S. and Canada: decline from 80% to 32%

Middle East and Latin America: a substantial increase of tweets

Europe: stable at 20%

What induces users to tweet?



Information:

Retweets: natively supported by Twitter since Nov 2009

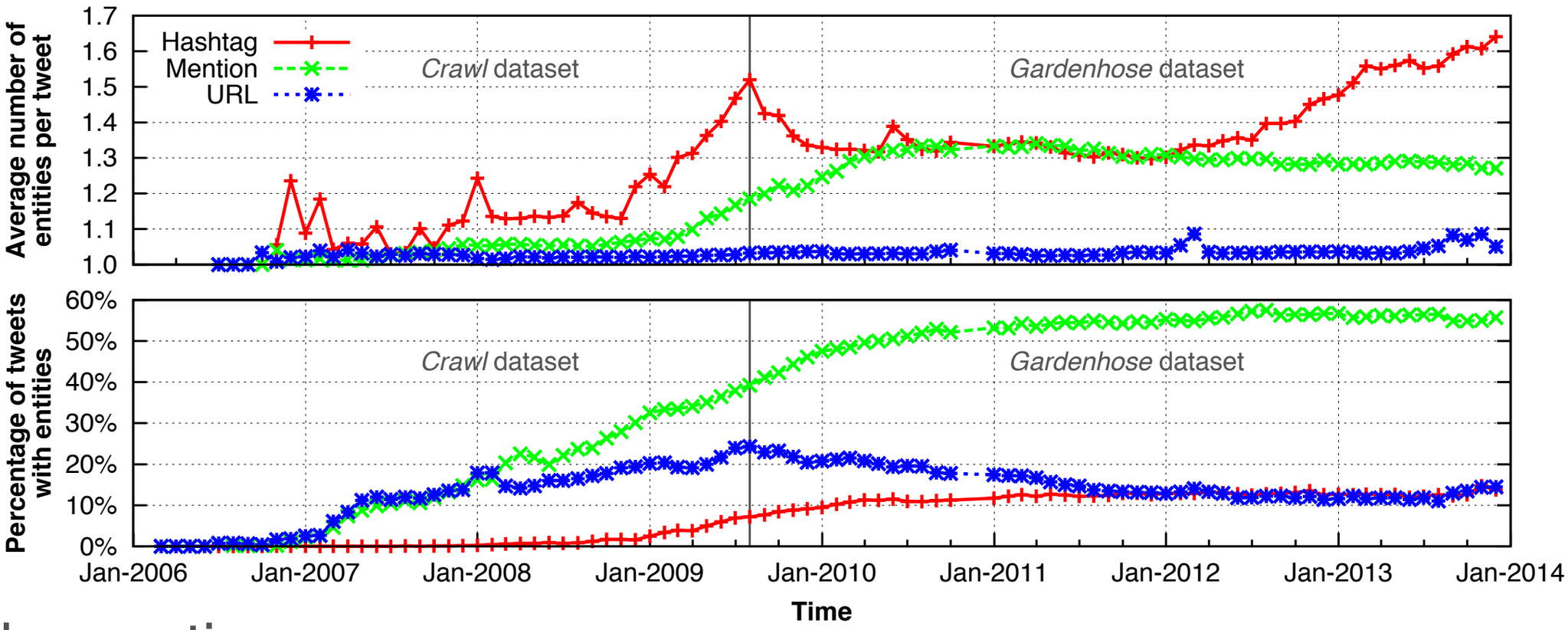
RTs: manually copied the tweet and added a "RT @username" at the beginning

Observations:

Retweets: the percentage increases rapidly afterwards.

Reply: a rapid adoption of the mechanism, peaking at ~35% of all tweets in 2010 and declining slightly afterwards

What do tweets contain?



Observations:

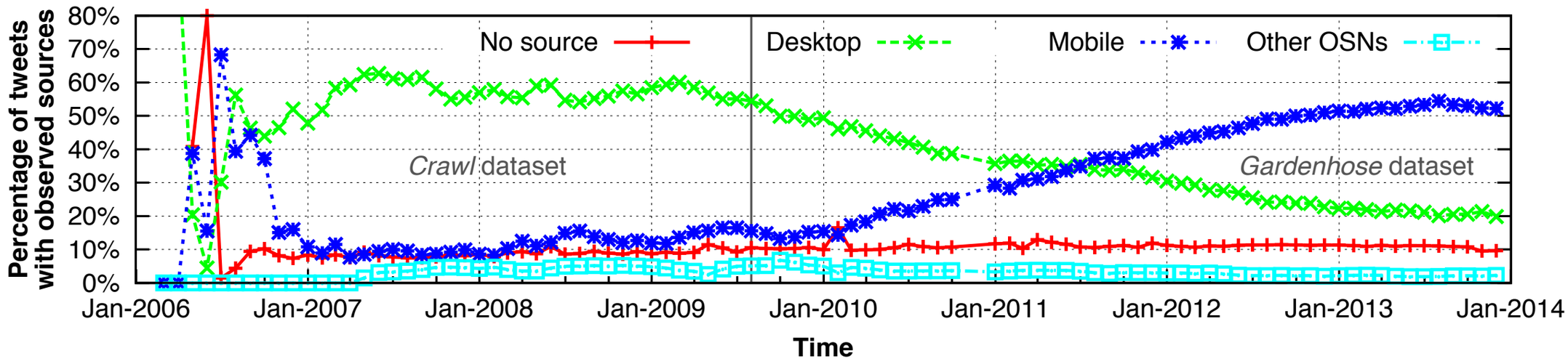
The percentage of tweets with mentions has increased substantially since 2009.

The percentage of tweets with URLs has decreased to stabilize at 12%.

URLs and mentions have stabilized around 1.0 and 1.3, respectively.

The average number of hashtags shows a continuing increase beyond 1.6.

What device are users tweeting from?



Information:

The source field attached to each tweet

Manually classify all 54 unique sources that represented at least 1% of tweets in any month

Observations:

A consistently decreasing trend for desktop clients and a corresponding increasing trend for mobile clients

Tweets created by Other OSNs: consistently ~3% of the overall tweets

Conclusions

Collect dataset of over 37 billion tweets from 7 years

- Examine the evolution of Twitter itself
- Focus on the Twitter users and their behavior

Quantify a number of trends

- the spread of Twitter across the globe
- the shift from a primarily-desktop to a primarily-mobile system
- the rise of spam and malicious behavior
- the changes in users' tweeting behavior

Aid researchers in understanding the Twitter platform and interpreting prior results

Questions?

We make all of our analysis available to the research community (to the extent allowed by Twitter's Terms of Service) at

http://twitter-research.ccs.neu.edu/

Email: ybliu@ccs.neu.edu

Backup slides

Determine user_id status in *UserSample* dataset

Query Twitter in Jan 2014 for the most recent info on each user

Both via the Twitter Rest API and the web site

https://twitter.com/intent/user?user_id="+userid

User_Id Statuses	Via
Public	Twitter API
Protected	Twitter API
Suspended	Web Site
Deactivated	Web Site + Tweet
Unknown	Web Site + NoTweet

Comparison of findings

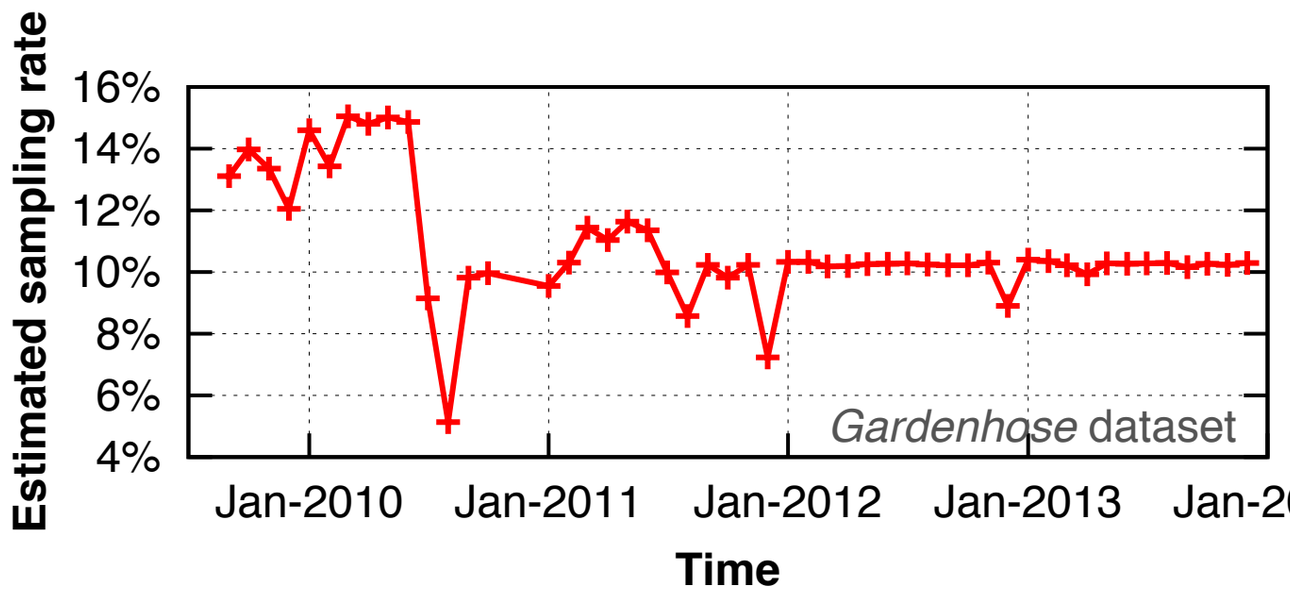
Examples:

[Macskassy and Michelson 2011] report that 32% of tweets are retweets, contradicting our measurement of 10% at the same time. The mismatch is likely caused by the authors' snowball sampling method.

[Petrovic, Osborne, and Lavrenko 2013] and [Almuhimedi et al. 2013] find that around 2-3% of tweets were deleted in their 2012 dataset, which is consistent with our results (2.35%) for the same time period.

In terms of lang, our findings supports the previous findings by [Krishnamurthy, Gill, and Arlitt 2008] about the top 10 languages on Twitter in 2008. However, we also show that this situation has changed significantly, with English today covering barely half of the user population.

The sampling rate of



$$rate = \frac{obs}{sc_{last} - sc_{first}}$$

The average value of rate across all users with $SC_{last} - SC_{first} > 1000$

The first observed value of statuses_count SC_{first}

The last observed value of statuses_count SC_{last}

The number of tweets we observed obs

JSON Example: {"created_at":"Fri Nov 01 00:00:40 +0000 2013","id":
396064209307303936,"text":"RT @HentaiUchi: 17 Like it? RTVRetweet it!
http:\/\/t.co\/VKiS2ceBuvf",user":{"id":
1639501730,"id_str":"1639501730","name":"Momo Velia
Deviluke","screen_name":"MomoVeliia","followers_count":

Users joining and leaving