WP1: Dataset for characterizing distributed OSN

4th TREND Plenary Meeting
Ghent, February 14th-15th

Roberto González
Universidad Carlos III de Madrid

Summary of tools

- Twitter Crawler
  - Geographical Location and Relationship
  - Terms
  - Consecutive Tweets
- Google Plus Crawler
  - Geographical Location and Relationship
- Facebook Crawler
  - Geographical Location and Relationship
- Youtube Crawler
  - Statistics

Twitter Datasets

	Collected	Geolocated
# Friends	5M	973K
# Followers	60M	16.5M
# Relationships	450M	100M

10-01-2011 until 28-04-2011

	Collected
# Tweets	400K
# Users	22K
# Popular Tweets	1.3K

12-03-2011 until 24-06-2011

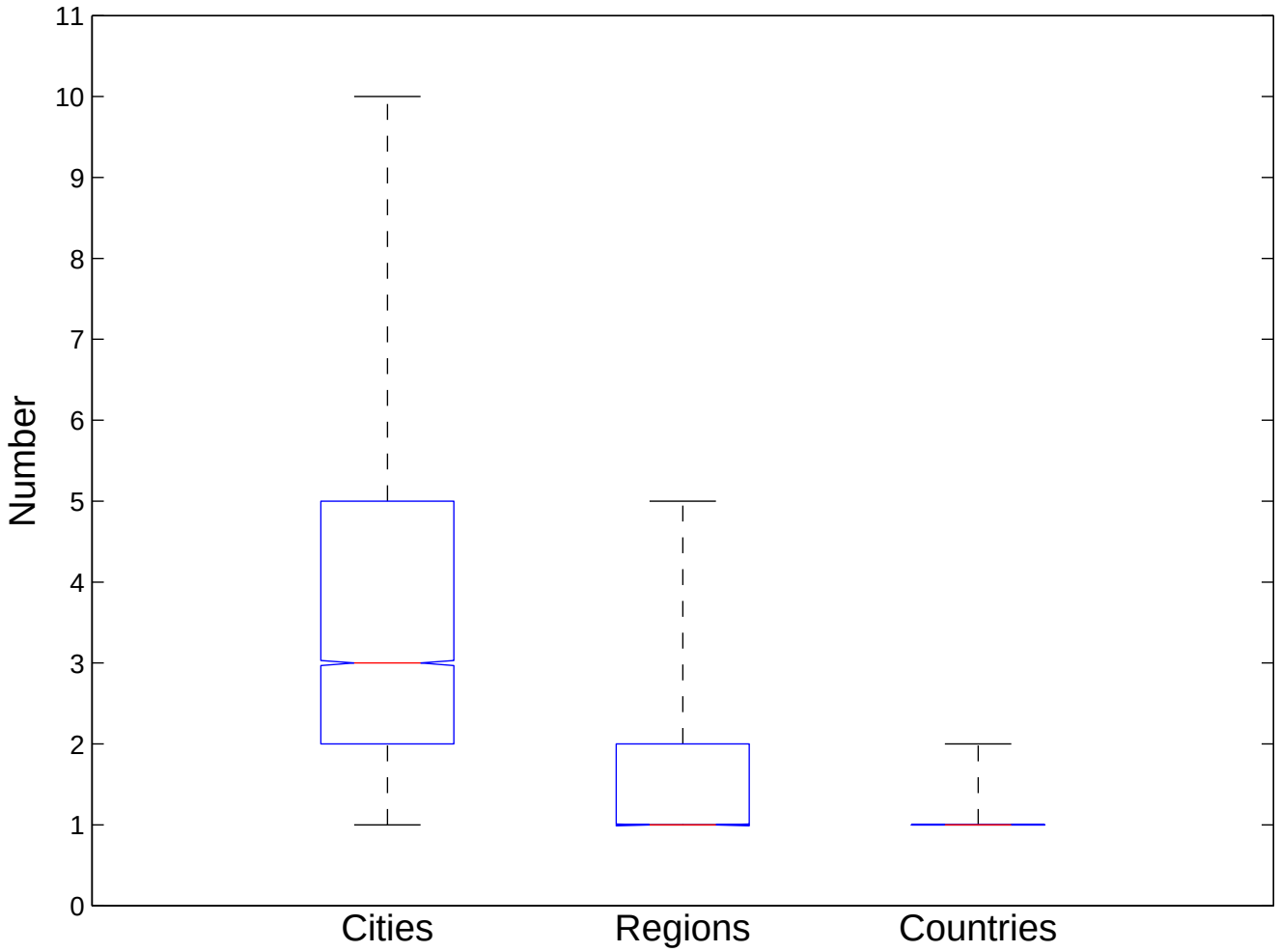
	Consecutive
# Tweets	>1M
# Tweets/hour	>30K

17-01-2012 until 11-02-2012

Decentralized Twitter

- Improve Twitter service to save energy
  - Where to place the servers?
  - How many datacenters?
- Locality in Twitter
- Traffic in Twitter
- User behaviour

User locations



- 22K users >10 tweets GPS
- 20%->1 city
- 90%->1 country

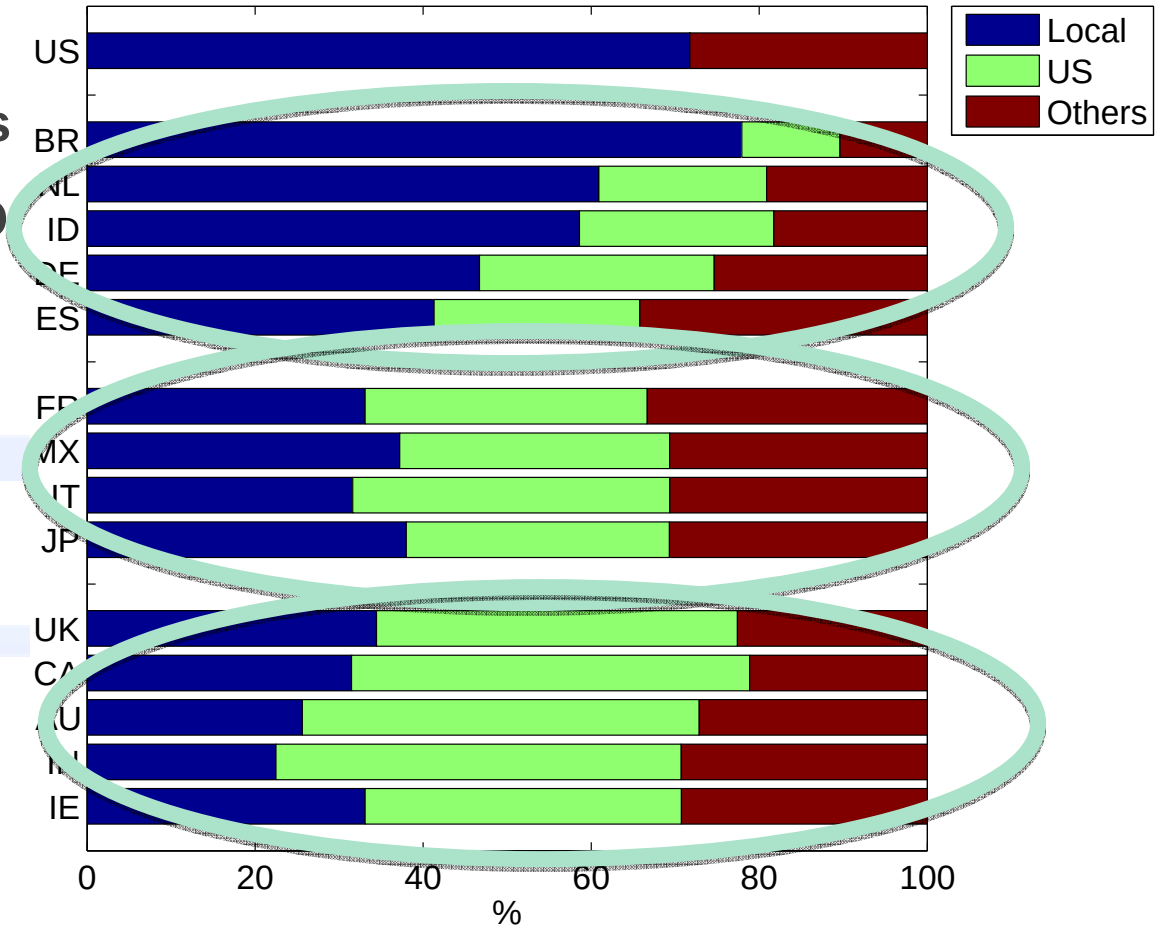
Follower locality

- ~ 1M users geolocalized
- > 16M of their followers
- ~ 50% in US

Country	Friends		Followers	
	Number	%	Number	%
US	528 K	54.25 %	7.37 M	44.59 %
UK	70.6 K	7.27 %	987 K	5.98 %
BR	61.7 K	6.34 %	1.81 M	10.94 %
CA	39.4 K	4.05 %	565 K	3.42 %
DE	21.7 K	2.23 %	331 K	2.00 %
AU	20.3 K	2.09 %	232 K	1.40 %
IN	18.8 K	1.93 %	442 K	2.67 %
NL	14.9 K	1.53 %	334 K	2.02 %
ID	12.1 K	1.24 %	862 K	5.22 %
FR	10.8 K	1.11 %	232 K	1.41 %
ES	8.7 K	0.89 %	277 K	1.68 %
IT	7.1 K	0.73 %	159 K	0.96 %
JP	6.9 K	0.71 %	192 K	1.16 %
IE	6.5 K	0.67 %	95.4 K	0.58 %
MX	5.5 K	0.56 %	234 K	1.41 %
TOP 15	833 K	85.60 %	13.13 M	85.44 %
ALL	**973 K**	**100 %**	**16.53 M**	**100 %**

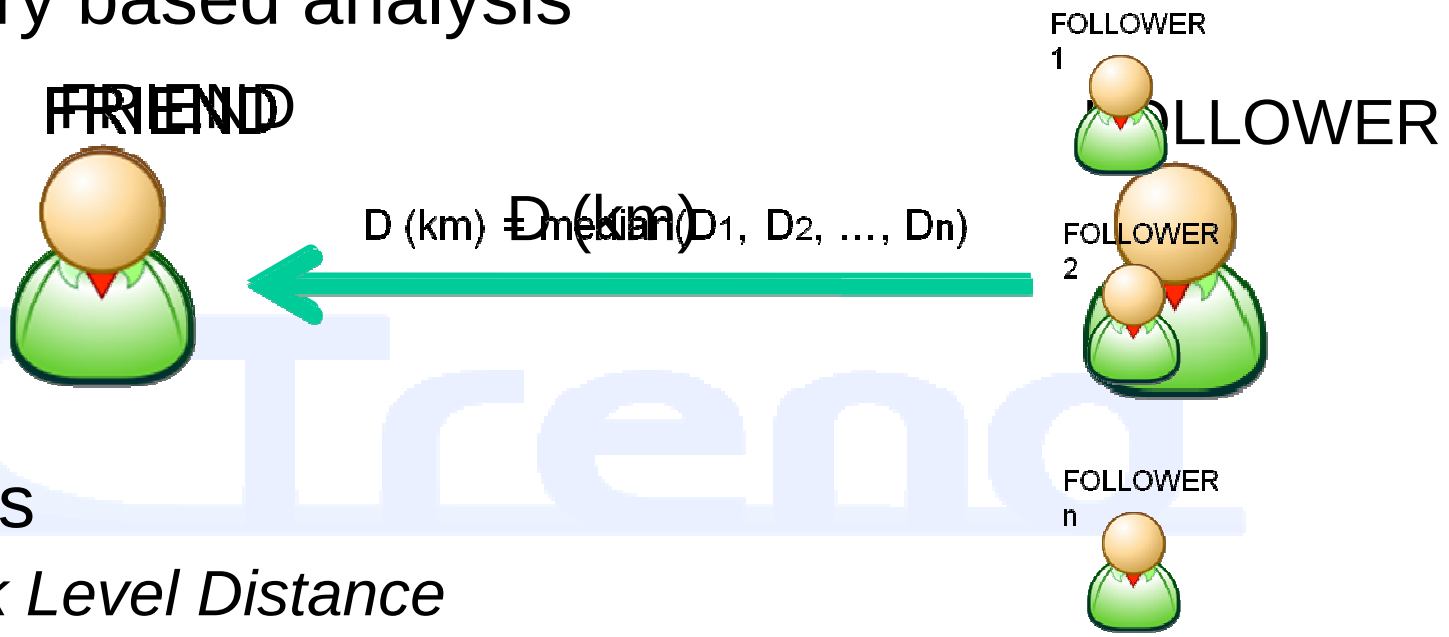
Countries Profiles

* US
  * 70% intra-country links
* Local Profile
  * US predominance
  * Local > US & Local > O
  * Strong local culture
  * Strong Official Language
* Shared Locality Profile
  * Local ≈ US ≈ Others
  * Typically countries where Twitter is less popular in our dataset
* External Locality Profile
  * English-speaking countries
  * Majority of links go to US (e.g. 48% IN, 47% AU)



Methodology and Metrics

Country based analysis



Metrics

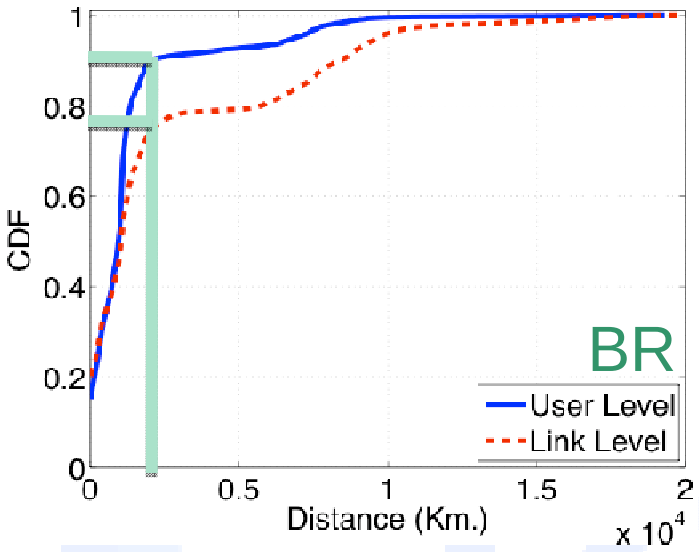
Link Level Distance

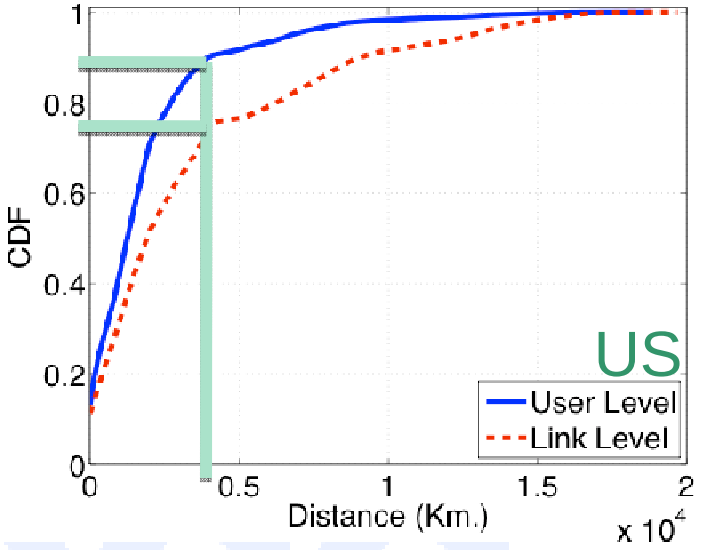
Geog. distance for each friend → follower link

User Level Distance

Median distance from a friend to its followers

Link and User Level Distance

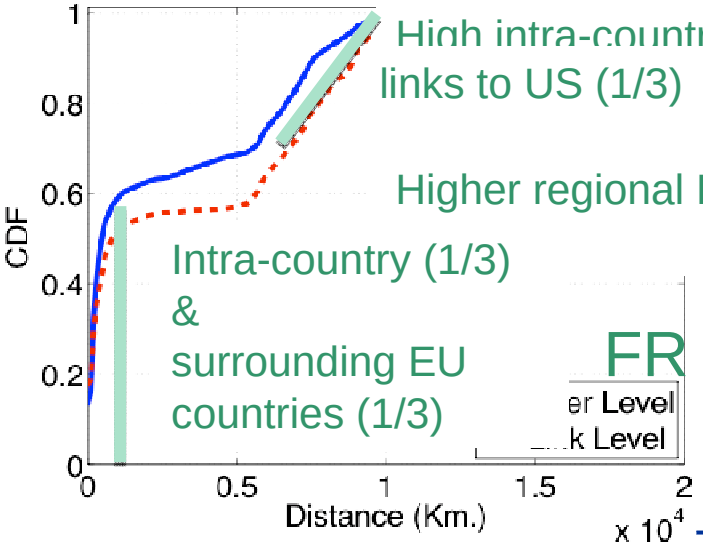




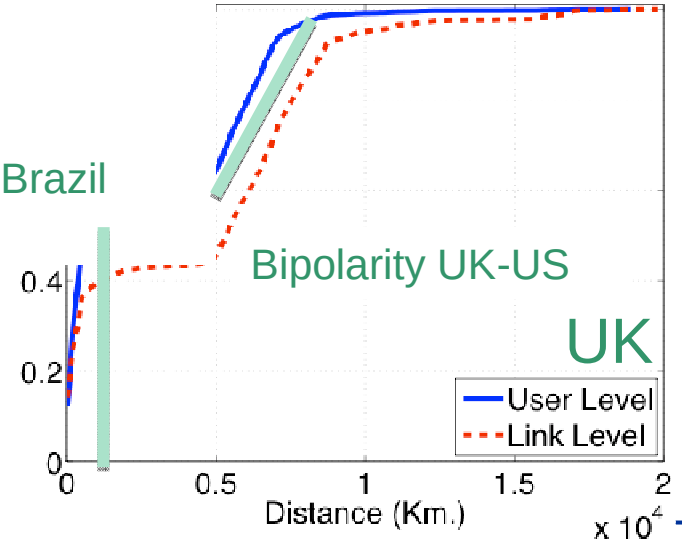
High intra-country locality links to US (1/3)

Higher regional Locality in Brazil

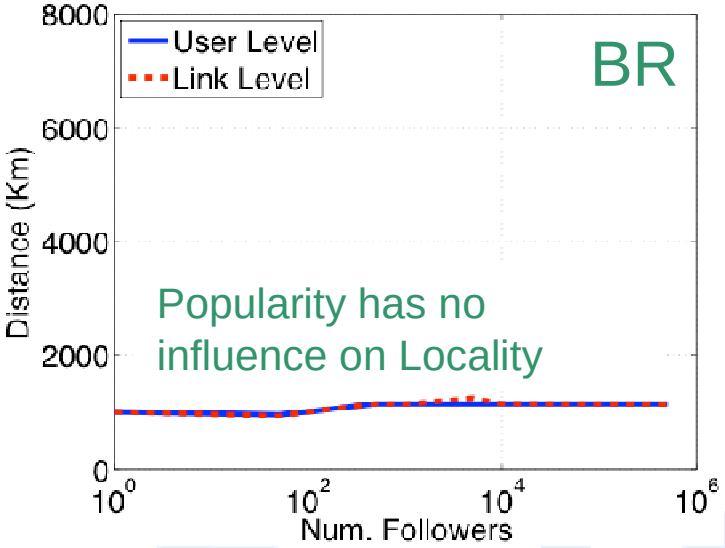
Intra-country (1/3) & surrounding EU countries (1/3)

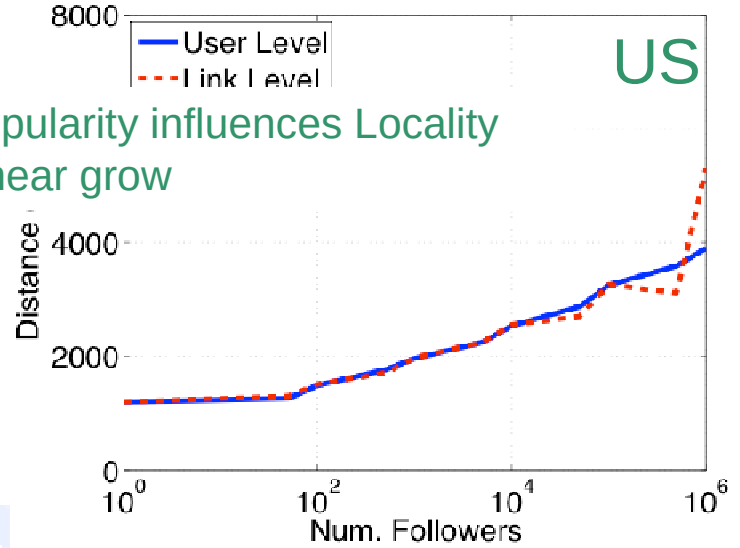


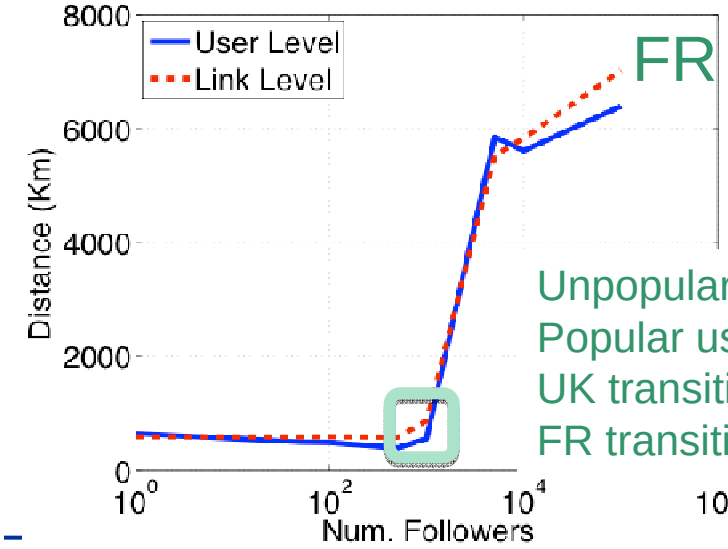
Bipolarity UK-US

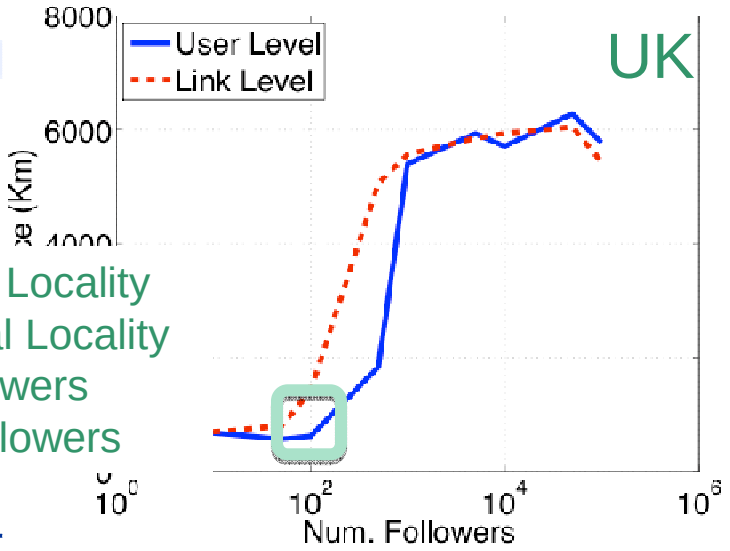


Locality vs Popularity







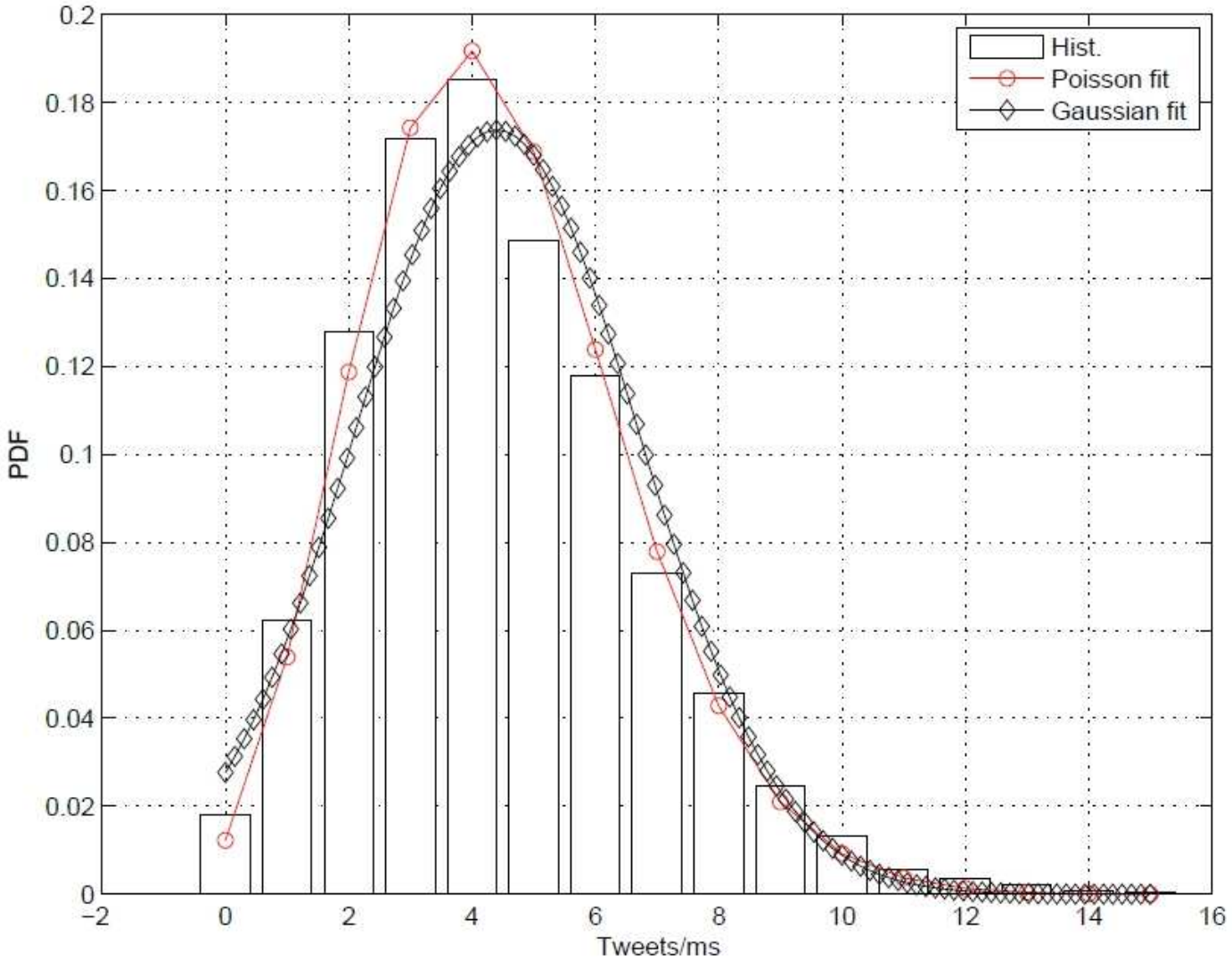


Unpopular users → High Locality
Popular users → External Locality
UK transition → 100 followers
FR transition → 1000 followers

Locality conclusions

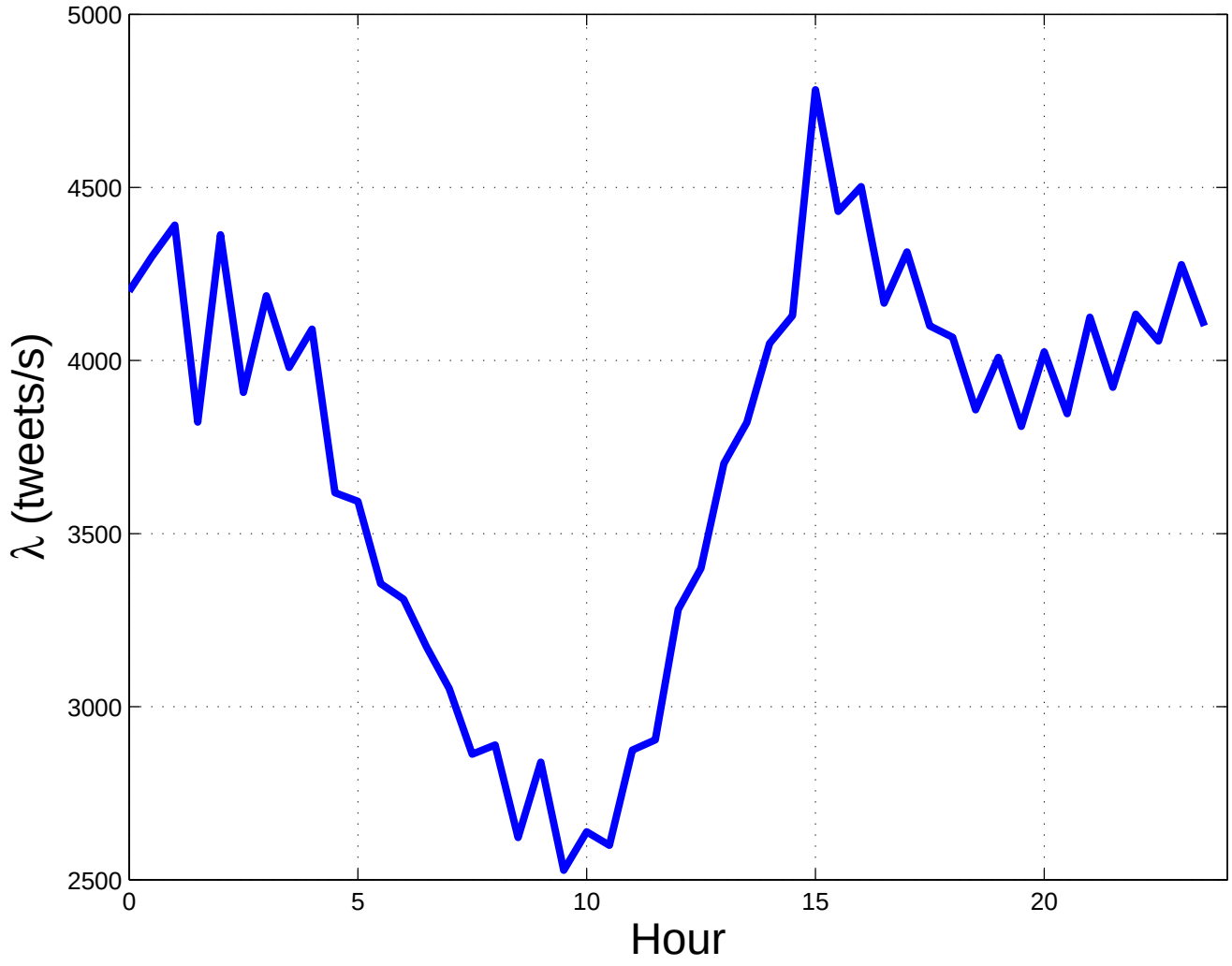
- User post tweets from a single country.
- Follower Locality trends are different for different countries
  - High intracountry Locality in countries with an own language (i.e. Brazil)
  - External Locality in countries where English is the official o co-official language (i.e. Australia)

Traffic in Twitter



- Tweets arrival -> Poisson process
- 2-5 tweets/ms

Tweets per hour



- 50% at 10 than in busy hour
- Is it possible to turn off some servers?

User behavior

- Crawlers
  - User location
  - User relationship
  - Traffic pattern
    - Only tweet arrival!!!
- Mobility action
- Tstat
  - Real traffic
  - Traffic patter for tweet consumption