Centre for
Data Analytics

Insight

From Smart Cities to Smart Neighbourhoods: Detecting Local Events from Social Media

Yang Li and Alan F. Smeaton

Insight Centre for Data Analytics

Dublin City University

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN SFI Science Foundation Ireland

Event Detection

Research topic across many application areas

Early work in detecting **news** events leveraged NLP, named entity recognition, operating on well-structured text

Nowadays, we're interested in event detection from social media

Twitterstand – breaking news from Twitter by clustering similar tweets

Sakaki et al. do likewise using a SVM

Twitcident enables management of tweets during events as they happen

These successfully detect global events based on significantly increased tweet volume

Our interest ?

Twitter often posts tweets about events which are more local, community-based … local flood, a fire, road closure

Can we detect unusual events at a *local* level, within a city … a smart neighbourhood ?

More challenging because volume is less, but very localised and representing semantic consistency, yet semantic deviation from normal

We focussed on geotagged tweets from Dublin city

Assumption

We assume a periodicity and consistency in tweeting behaviour

We assume local events, which are reported, cause semantic irregularities more recognisable than visitors, holidays, or one-off tweets

Approach is to determine normal crowd behaviour in a geographic region of the city, monitor sudden increases in the number and then focus on the topic

Data Used

English-only tweets, 2 month period, geotagged and in a bounding box in Dublin … 387,800 from 14,533 unique users … availability ?

City-wide is too big, we divided into (25) sub-areas, finding users tweet from few locations …

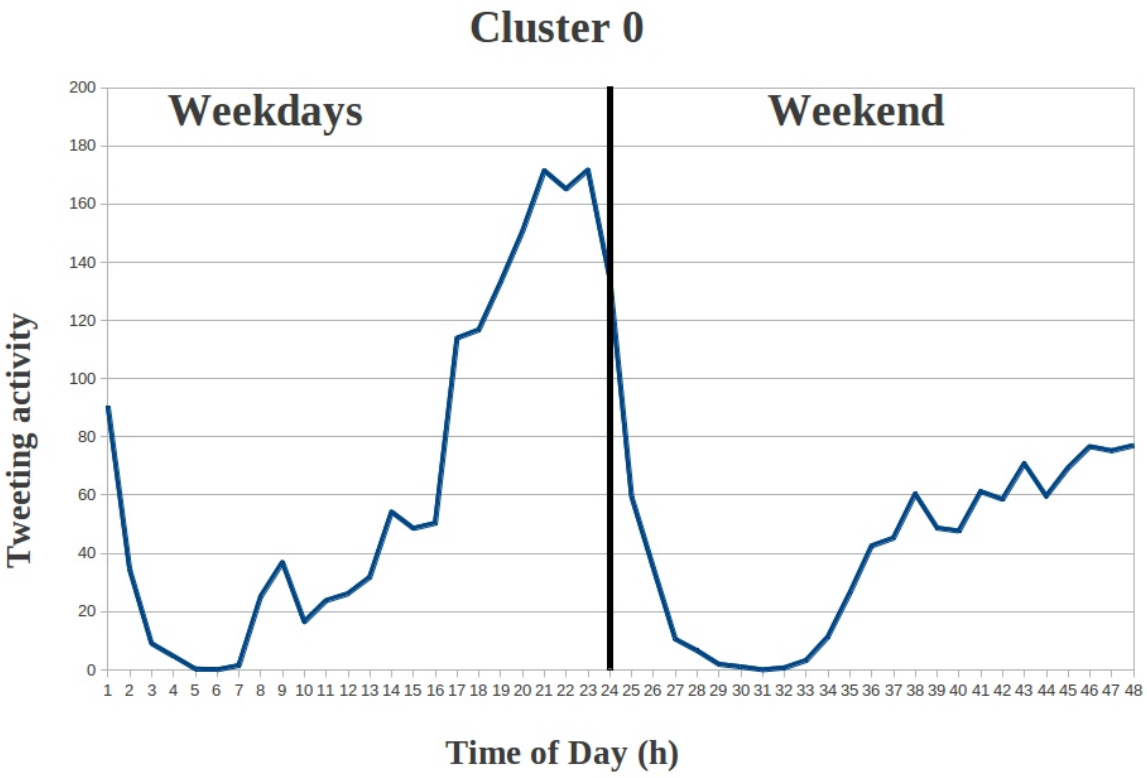
Based on 5,875 users generating 95% of our tweets, 44% tweet from only 1 or 2 (of 25) partitions

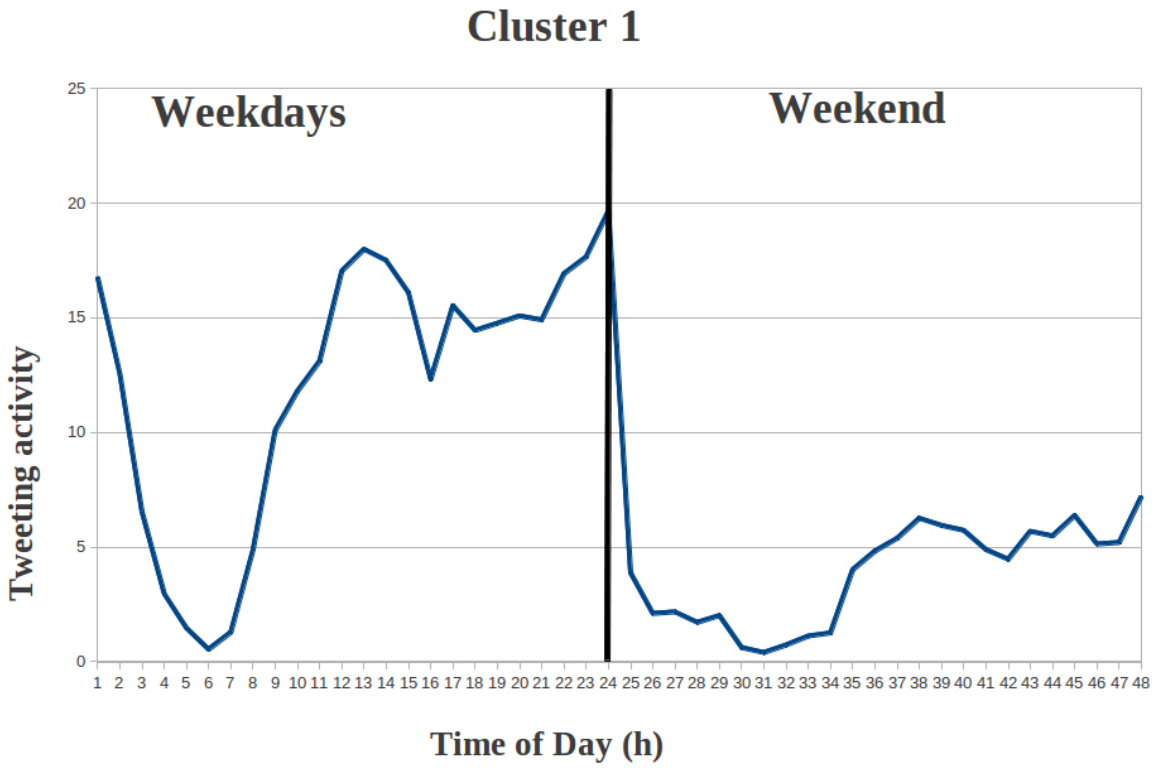
23% users tweeted across +5 partitions with a Power Law distribution, and these “random” zones are of interest for detecting local events

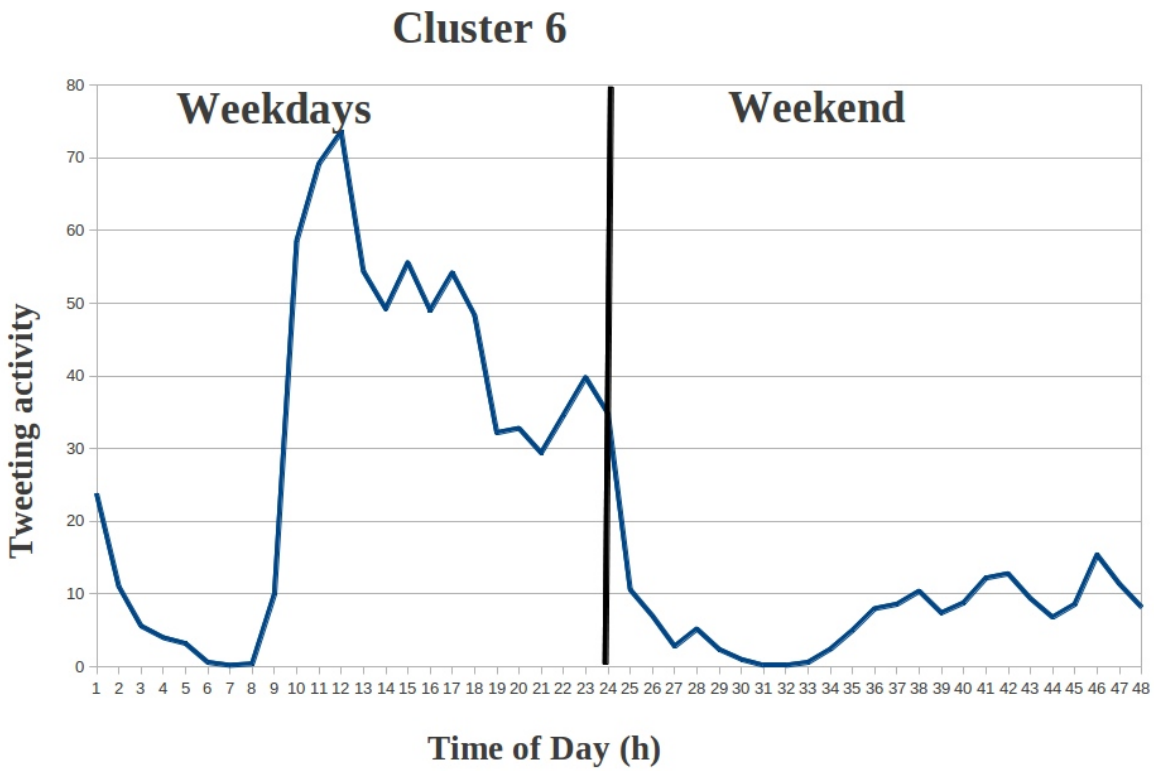
Users tweet at regular times

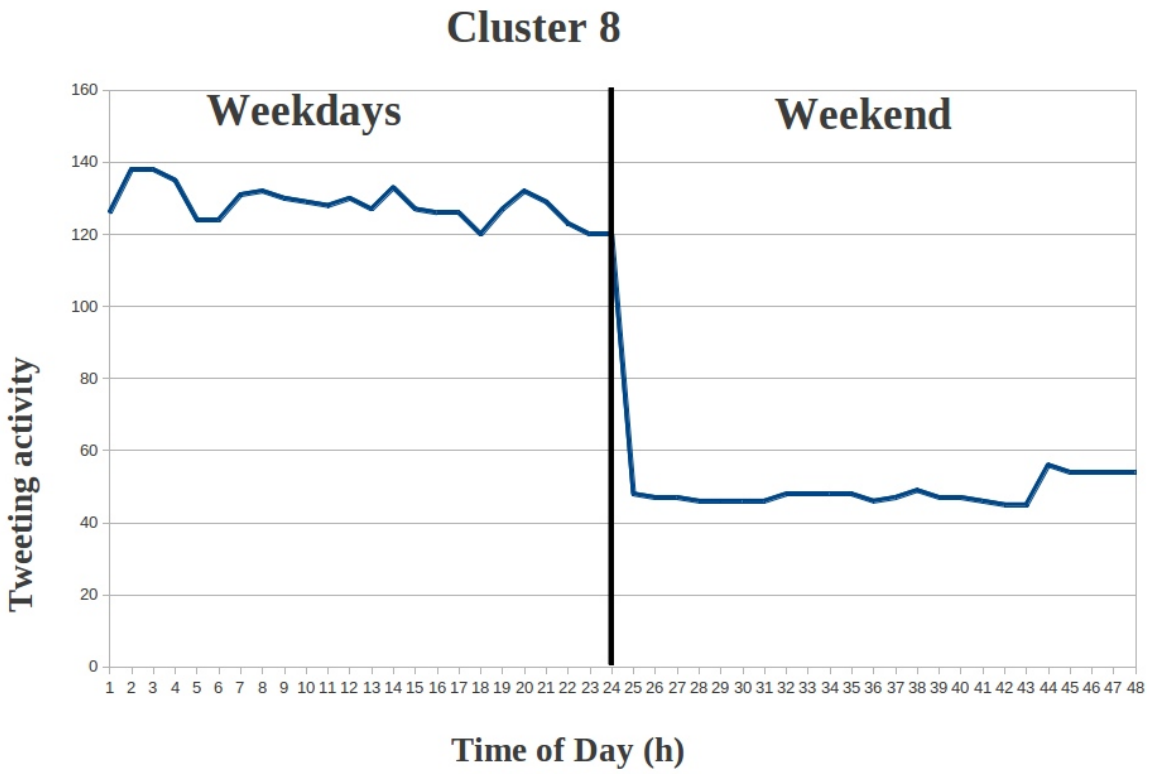
Focusing on 805, our most active users (+100), clustered them using time-of-day and weekday/ weekend into 10 clusters

We observed recurring temporal patterns of when people tweet


Cluster 0
Weekdays
Weekend
Tweeting activity
Time of Day (h)
0
20
40
60
80
100
120
140
160
180
200
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48


Cluster 1
Weekdays
Weekend
Tweeting activity
Time of Day (h)
0
5
10
15
20
25
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48


Cluster 6
Weekdays
Weekend
Tweeting activity
Time of Day (h)
0
10
20
30
40
50
60
70
80
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48


Cluster 8
Weekdays
Weekend
Tweeting activity
Time of Day (h)
0
20
40
60
80
100
120
140
160
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48

Users tweet at regular times

Focus on 805, our most active users (+100), clustered them using time-of-day and weekday/ weekend into 10 clusters

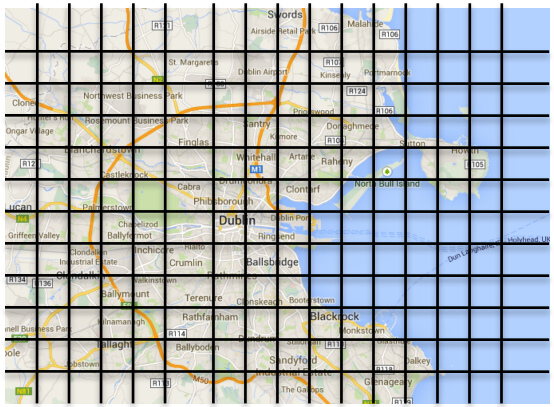
We observed recurring temporal patterns of when people tweet

So people exhibit temporal patterns of when, and where they tweet

Partitioning the city

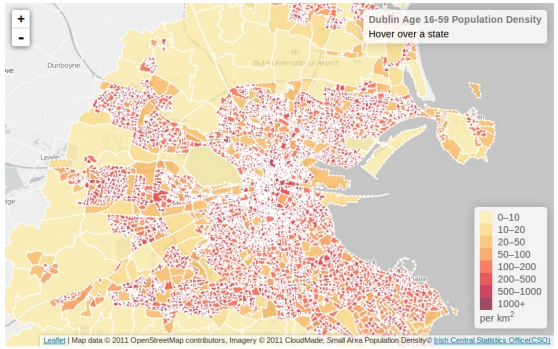
Dividing by grid ?

-> imbalance in population distribution



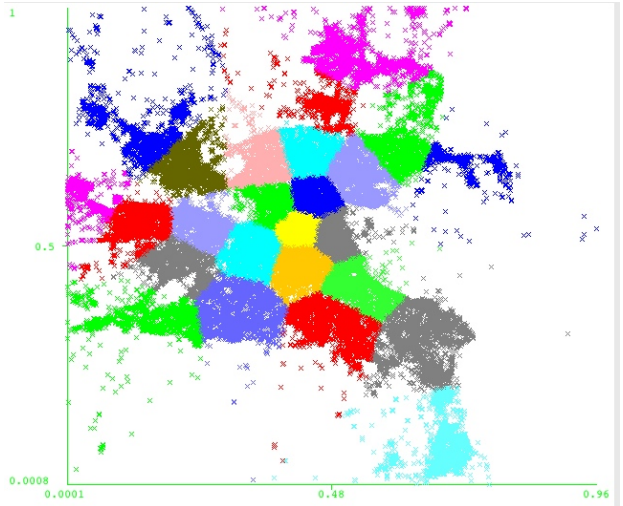
Dividing by population ?

-> imbalance in tweet usage



K-means clustering based on geographical occurrences of tweets

Partitioning into 25 regions



Partitioning the city

Dividing by grid ?

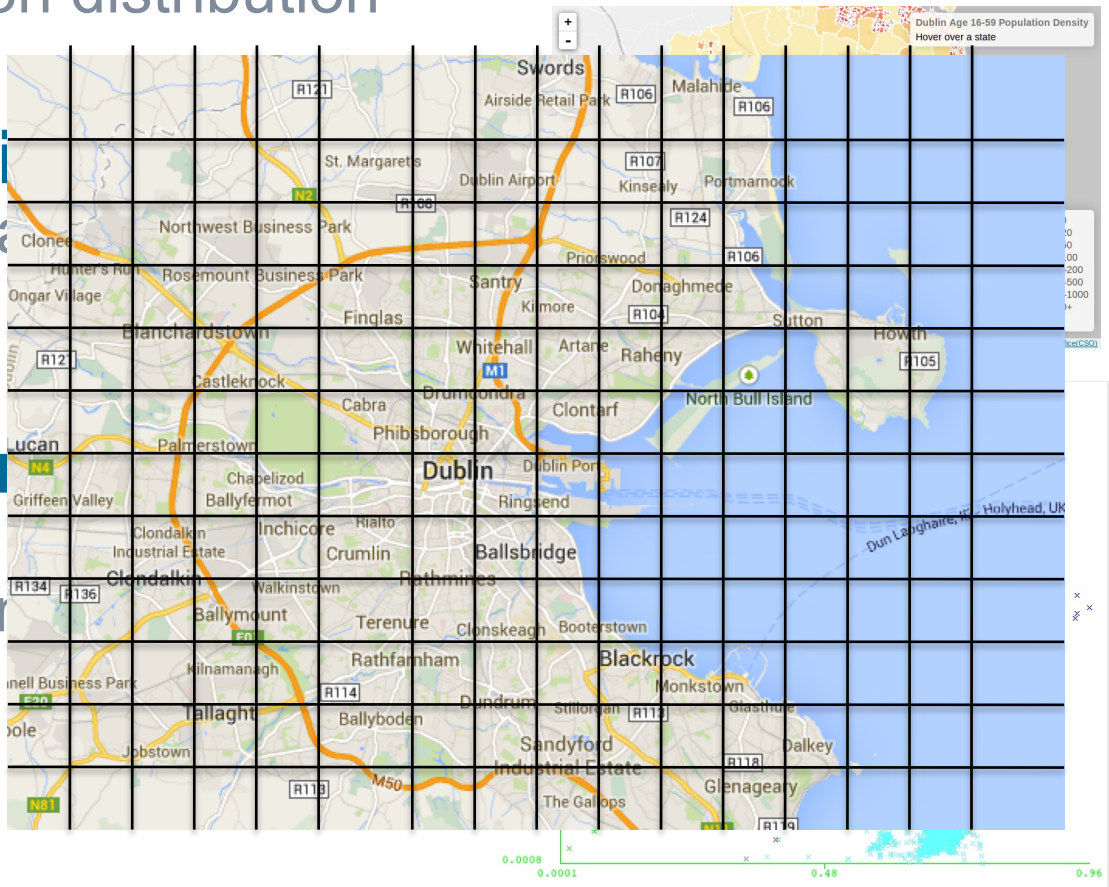
-> imbalance in population distribution

Dividing by populati

-> imbalance in tweet usa

K-means clustering
geographical occurr
tweets

Partitioning into 25 regio



Partitioning the city

Dividing by grid ?

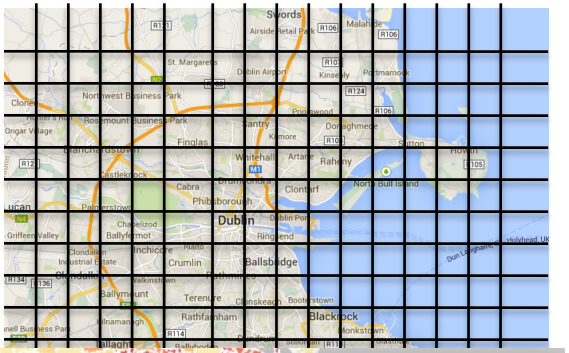
-> imbalance in population distr

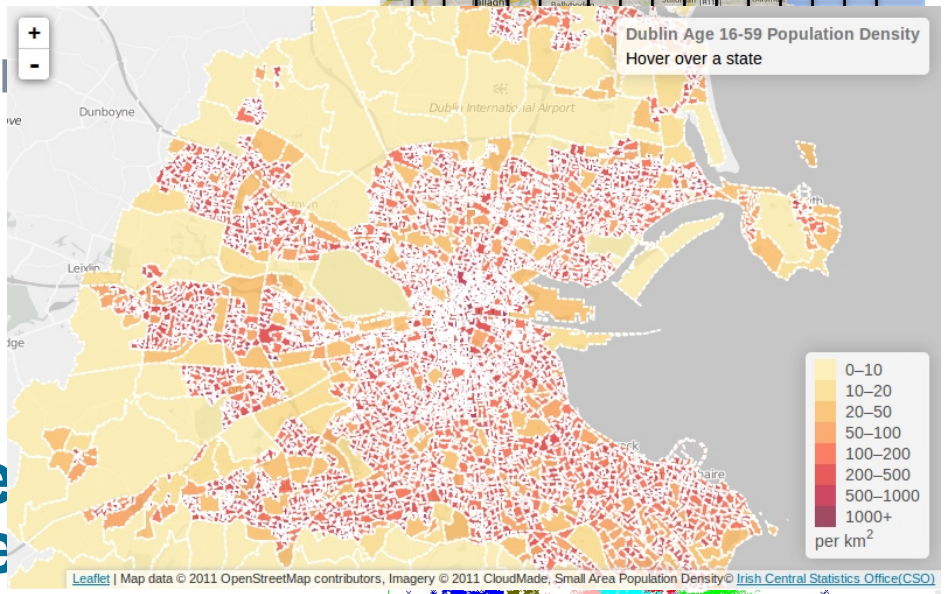
Dividing by population ?

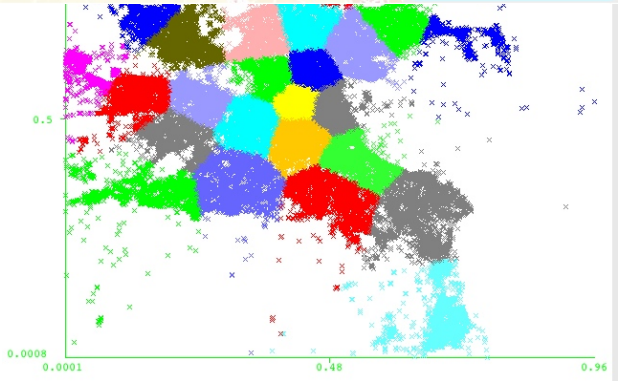
-> imbalance in tweet usage

K-means clustering base geographical occurrence tweets

Partitioning into 25 regions







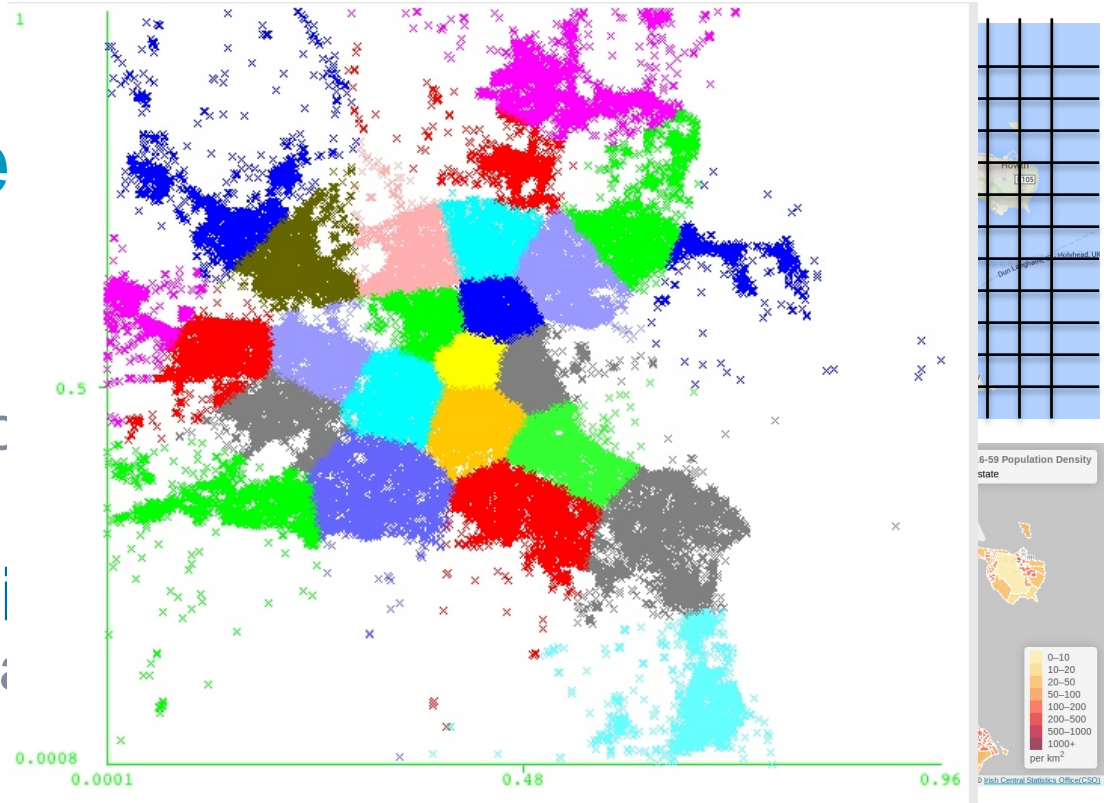
Partitioning the

Dividing by grid ?

-> imbalance in populatic

Dividing by populati

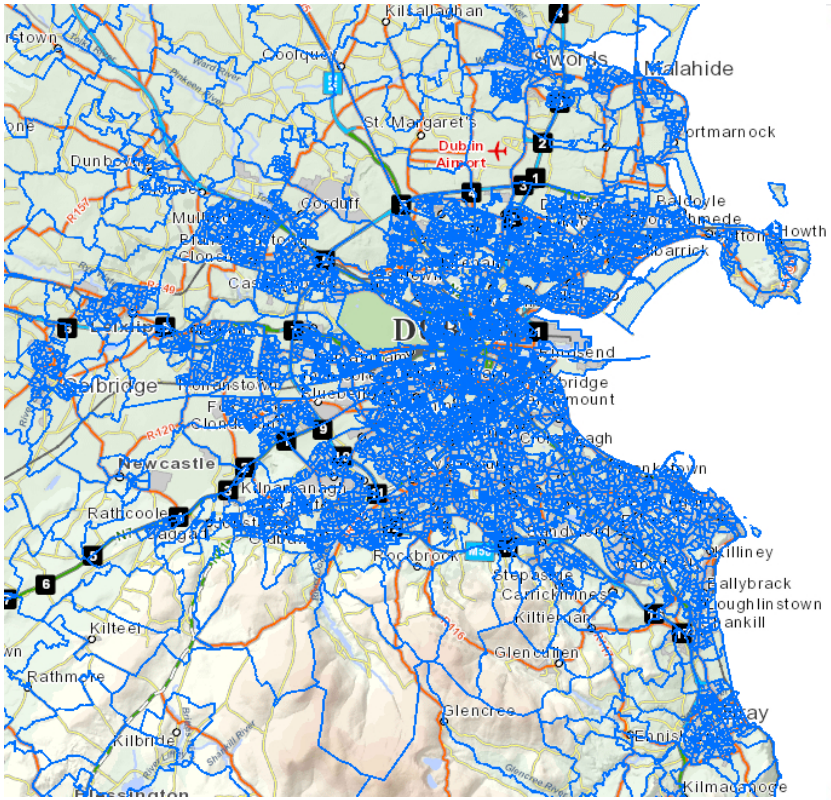
-> imbalance in tweet usa

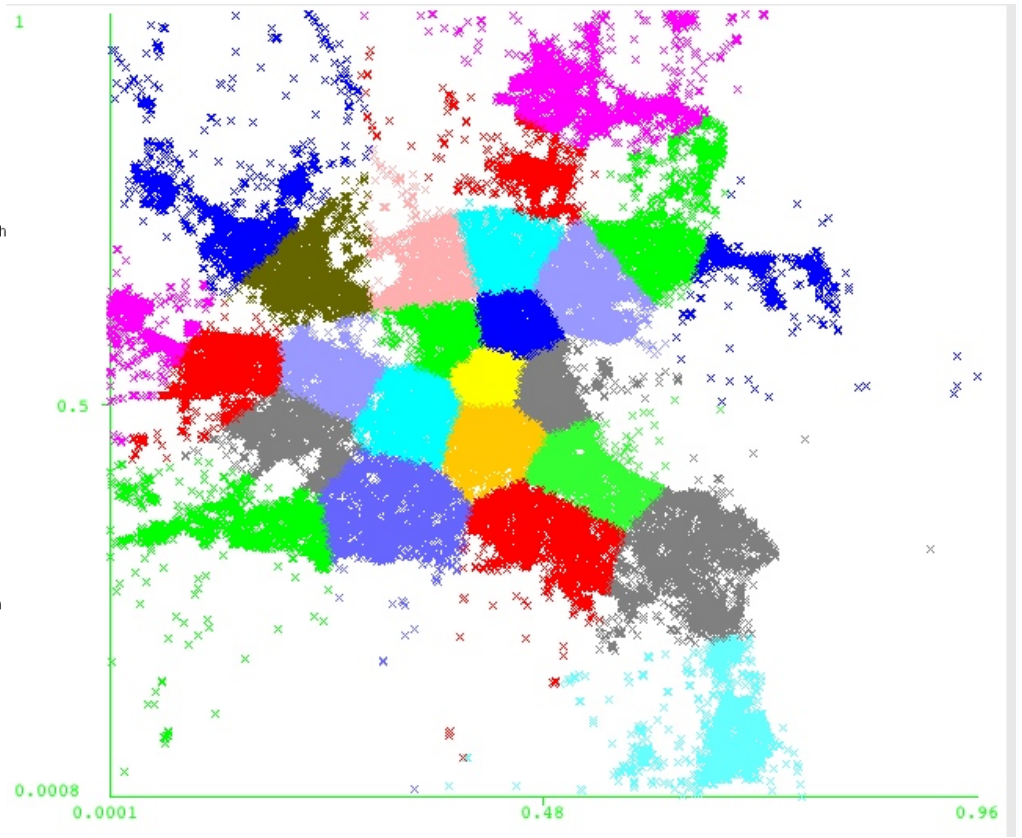


K-means clustering based on geographical occurrences of tweets

Partitioning into 25 regions

Are partitions reasonable ?





Population distribution (CSO) vs. Partitions

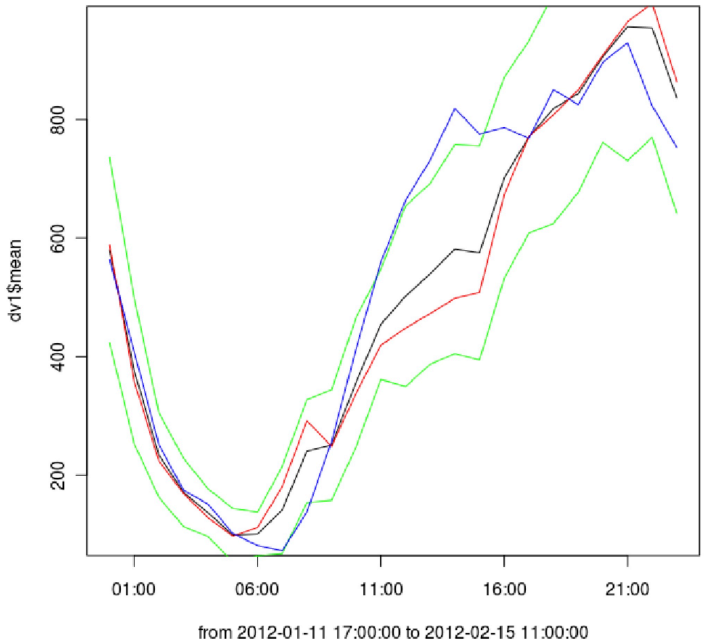
Measurements of Regularity (1)

Time of tweeting within partitions

We analyse weekday / weekend separately

Regularity calculated based on 24x hourly bins each with a rolling one-month window

Standard deviations from this could indicate a local event



Measurements of Regularity (2)

Location of regular Tweets

Can be compounded by visitors, away from home for work / vacation

For each partition we maintain a set of regular active tweeters

If many visitors tweet from a partition could indicate a local event

Measurements of Regularity (3)

Semantic regularity of Twitter content, per partition

Using Lemur, we built a language model for each geo-tagged tweet in each partition to represent semantic consistency

For each incoming geotagged tweet we rank partitions by P of generating the tweet, use KL divergence

Comparing predicted vs. actual partition, Mean Reciprocal Rank = 0.429, 33% of predictions are correct

Measurements of Regularity

We then combine them ..

$$F = \alpha.NT + \beta.NU + \gamma.SR$$

Evaluation …

Boo !

There is no standardised test collection and few standardised tasks on harvested Twitter content, except TREC

But who is to know about slow traffic on M50 near Blanchardstown exit on morning of 5th March 2013 ?

Instead we have anecdotal examples of local events which occurred

Anecdotal events

Event and Date	Time	GPS Coordinates	Related Twitter Content
Local flooding in Glencree Valley Jan 25, 2013	16:45:10	53.1809595,-6.1887448	The flooding around #Glencreevalley #Enniskerry is crazy! Watch out drivers! #Aaroadwatch
	16:50:08	53.182842,-6.191808	my car is like a floating boat #Enniskerry #flooding
Car crash on O'Connell Street caused by heavy rain, Jan 25, 2013	17:28:32	53.1809595,-6.1887448	@aaroadwatch bus and car collision on o'Connell street sb
	17:30:32	53.348604,-6.2597	@RobbieH46 slowly....it's a fecking car crash!!!!
	17:30:50	53.347887,-6.259207	Poor man or women in car crash.. #sayapray dangerous driving in this weather #5wordweather @spin1038
Heavy traffic jam Blanchardstown, Mar 09, 2013	17:17:11	53.3948484,-6.3912147	massive traffic jam in blanch won't be home till Christmas
	17:21:49	53.394718,-6.389326	traffic freaks me out!!!
	17:05:01	53.393323,-6.393317	Caught in a traffic jam
Pipe burst, cut off water supply Clongriffin Jan 07, 2013	14:22:16	8. 53.404341,-6.158719	@DonnieWahlberg its raining we have no water because of a burst pipe I am bogged down in housework but I am happy and having fun anyway :-)
	22:32:06	53.2853,-6.22825	@seanm91 apparently while attempting to fix the water pipe they damaged the gas line #incompetence

Conclusions

We examined dynamics of small, local areas within a city through social media

Focus on consistencies across Twitter behaviour covering location, time, and content for each of 25 city regions

Experiments inconclusive but anecdotal evidence of detection of local events

Thanks to …

Science Foundation Ireland

IBM