What Does the Twitter Say

Sheng Cai
Data Incubator Fellow
sheng.cai@mg.thedataincubator.com
Fall 2015

1 Introduction

In this project, a **2.53GB Twitter database** (part of Twitter7 database) is downloaded from Stanford Large Network Dataset Collection (SNAP) and analyzed using **NetworkX** module and **Gephi**.

In this 2.53GB Twitter database, **18,572,084 tweets** during the period of **June 1, 2009 to June 30, 2009** are recorded. Due to possible recoding issue, **98.5% of** the tweets are recorded between **June 12, 2009 to June 30, 2009**.

2 Data format

Original database is recorded as one single TXT file: “tweets2009-06.txt”.

Each entry in this database is recorded with **tweet time, original post user ID, tweet content**. Using these information and the aforementioned tools, several trend will be unveiled, along with Twitter user habbits and Twitter celebrities.

Each entry is recorded as:

Tweets

```
...

T    2009-06-11 16:56:39
U    http://twitter.com/joshuaculling
W    RT @NTC2009: #ntc2009 attendees on Twitter: @CIRAME,
@idahofreedom, @MassCLT, @NathanBenefield, @joncaldara, @pgessing,
@CarlaHowell

T    2009-06-11 16:56:40
U    http://twitter.com/birdsblooms
W    Urban gardens go up the wall http://snipurl.com/jwoe6#gardening
(via WoodwardGardens)

...
```

To facilitate further processing, original database is divided into 30 separate database based on tweet date. Their document size is as follow:

day	1	2	3	4	5	6	7
size	1	1	1	2	3	2	4
day	8	9	10	11	12	13	14
size	12	6	8	38527	170665	151487	142740
day	15	16	17	18	19	20	21
size	155748	146493	128458	123844	113983	120101	144999
day	22	23	24	25	26	27	28
size	121903	135787	89607	174125	151398	156763	131731
day	29	30					

size	165891	169106					

Unit of size is KB. As one can easily see that most of the data is collected from June 12 to June 30. In the following analysis, we will analysis both the temporal and spatial features of the given data in June 12 to June 30 period.

3 Data feature extraction and basic statistic analysis

There are various ways to interpret Twitter data. In this project, tweets are analyzed based on 2 features:

1. **retweet and reply** (@): People retweet other poster's tweet or reply to other poster's tweet by using at sign (@). Using the @ sign means current user either knows or subscribes to the original poster. For famous Twitter account, the more people retweet its post, the more famous this twitter account is.
2. **sharing same topic** (#): People express their concern and manually tag their tweet under a certain community sharing the same topic by using hashtag sign (#). Using the # sign is a convenient and effective way to manually partition the whole Twitter network. Counting the number of related Twitter user using same hashtag is viewed as a direct way to uncover trend and concern of general public. Twitter encourages the use of hashtag and nowadays hashtag is a social network phenomenon.

Based on these two features, each Twitter user is abstracted to the following class:

pseudocode

```
class Twitter_user:
     ID: Twitter user id
     Reference User: @ed user id
     Topic: #ed topic list
```

To facilitate analysis and reveal the underlying temporal pattern, data visualization will first be performed per day. Temporal patterns will then be revealed.

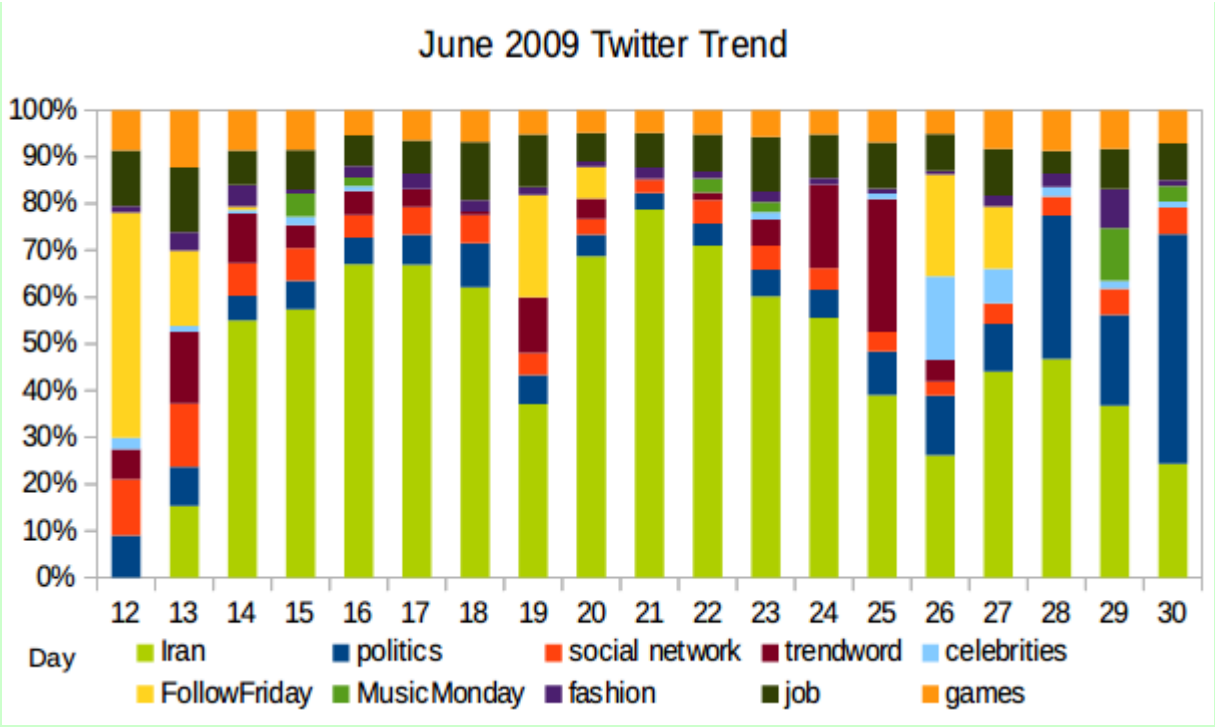
As mentioned in above paragraphs, hashtag is a natural way to partition all twitter users. Given the topics people are talking about everyday, it is hard to present all the topics in one graph. So the preliminary step before visualizing data is to extract daily hot topics.

Method graphBuilder.topTenHashtag is used to generate daily top ten hashtags. Every day ten most popular hashtags are extracted. From June 12 to June 30, there are 40 unique top ten hashtags. These hashtags are:

hashtags

```
squarespace          honduras       iremember      unfollowperez
11thcommandment      michaeljackson chupa          jobs musicmonday
inaperfectworld      tweetmyjobs    jtv            neda
followfriday         iphone         fx35           zensursula
140mafia             gr88           spymaster
unfollowperezhilton  forasarney     iran           tcot
helpiranelection     crisishn       haveyouever    iranelection
fb   ff              tehran         news           iranians
bsb                  cnnfail        goodpussy      dontyouhate
lolquiz              gilad          iran9
```

and are further abstracted to the following 10 classes. Overall trend is displayed in the following graph.



Overall, June marks the starting of Iran's Green Revolution and Twitter was suddenly bombarded with posts regarding situations in Iran. Meanwhile, Michael Jackson died on June 25. But MJ's death is not as influential as Iran situation and the ensuing Hunduras Crisis. No matter what the outside world changes, social gamers and FF/MM gangs never stop spamming Twitter. Social recruiting also never stops.

4 Visualize the social network

As discussed in Section III, Twitter users are connected by both retweets/reply and hashtags. In the following visualization, retweet/reply are used to calculate social network edge weight, hashtags are used as node properties. Using social network graph, we can discuss the following topics:

1. Does gamers form their own community considering their numbers are stable?
2. Who is the most retweeted/replied Twitter user (MVP users)? Does subscribers of MVP users share the same topic as the MVPs?

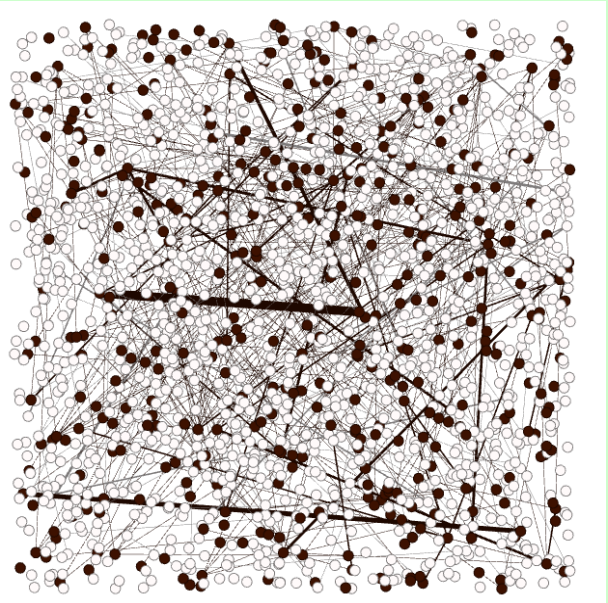


Figure 1: 140mafia(black) vs Spymaster(white)

4.1 Social gaming society (140mafia vs Spymaster)

June 21, 2009 is a Sunday, and is the peak day of social gaming hashtag counts. On this day 566,812 users posted at least one tweet. 329,184 edges are built based on Twitter user replies/retweets. Among them 1720 users posted tweet related to social gaming, and their communications with each other are

expressed using 434 edges.

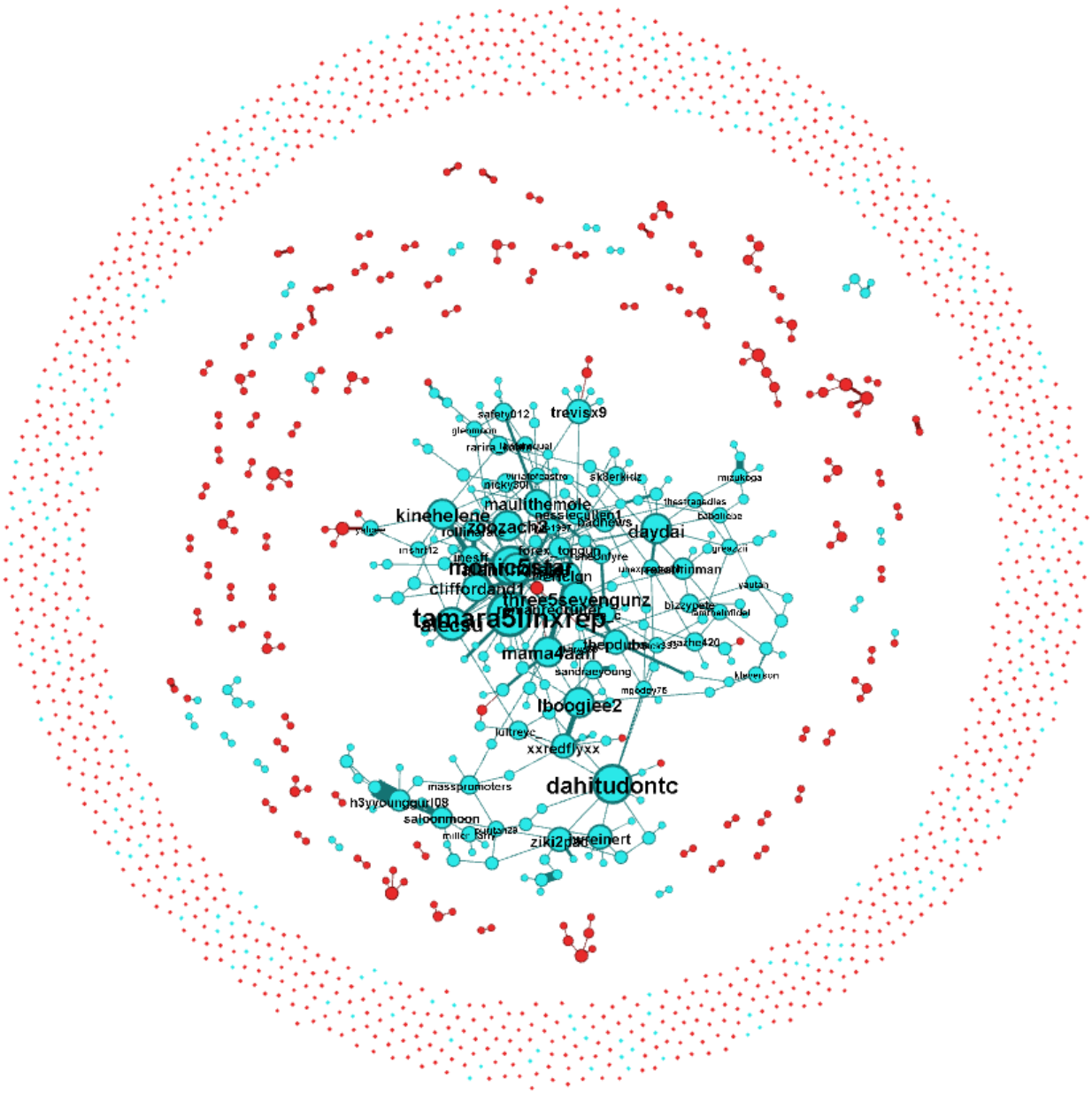


Figure 2: 140mafia network vs Spymaster network

Two mainstream social games are discussed: *140mafia* and *spymaster*. These two games use retweet/reply in totally different ways:

1. *140mafia*: The more followers you have, the greater the chances to succeed since a key component of this game is to recruit more twitter users. Thus more communications between "gangsters" are required.
2. *Spymaster*: This game allows you to "assassinate" your fellow followers. In this game, info sharing indeed brings you in-game cash, but most of the times, only event information such as "Assassinating a user," "Securing a new Safe House," and "Spymaster level increase," are

broadcasted.

The differences between these two games affects different communication style of their user. Apply “Yifan Hu” layout and “Force Atlas” layout on the network, one can get the clustered network as in figure 2.

In figure 2, blue dots are *140mafia* players, while red dots are players not engaging in *140mafia* game. Notice that there are very rare cases one player engages in both game simultaneously based on graph analysis, we can use figure 2 red dots to approximate *Spymaster* players. Node size represents user communicate activity (node degree).

The center community is filled with *140mafia* players, while *Spymasters* orbit around *140mafia* in small groups. Players of both games without a community formulate the outer asteroid belt. One can easily tell the different play styles of those two social gaming groups: Most *140mafia* games formulate a big community and communicate a lot, while *Spymaster* gamers tends to formulate small communities of less than 10 users. This phenomenon is a good reflection of those two games play style.

4.2 Twitter MVPs and their influence on subscribers

In this section we will find out the most replied/referred user on Twitter and check to see whether they serve as information dispatcher by examine their topics and their follower's topics.

`mvp.py` is used to generate June 21 top ten most referred Twitter user. Along with their topic, they are listed here:

hashtags

```
tweetmeme:    referred 2488 times, no active hashtagged topic
mashable:     referred 1791 times, no active hashtagged topic
addthis:      referred 1197 times, no active hashtagged topic
firefox:      referred 951  times, no active hashtagged topic
aplusk:       referred 891 times,  no active hashtagged topic
cnnbrk:       referred 883 times,  topic: 'michael', 'honduras',
'yemencrash', 'minnesotasenate'
songzyuuup:   referred 798 times,  topic: 'treysongzanticipation'
mileycyrus:   referred 796 times,  no active hashtagged topic
breakingnews: referred 696 times,  no active hashtagged topic
cardoso:      referred 612 times,  topic: 'twpiratafail',
'forasarney', 'ficasarney', 'genteburra'
```

Let's check user “songzyuuup”. This is a Twitter ad account for singer *Tremaine Aldon Neverson*, aka '*Trey Songz*'. Songz released a mixtape titled Anticipation in June 2009 through his blog, which featured songs from his third album. Let's see how his subscriber react to this topic.

Notice that one topic can have multiple expression. Topic **#treysongzanticipation** has the following variations:

Tweets

```
#treysongzanticipation:
     #treysongzmusicmonday          #twittition
     #treysongzanticipation         #treysongzanticipatn
     #anticipation                  #treysongzanticipationhttp
     #treysongzant                  #antiscipation
     #treysong                      #tre
     #treysonza
```

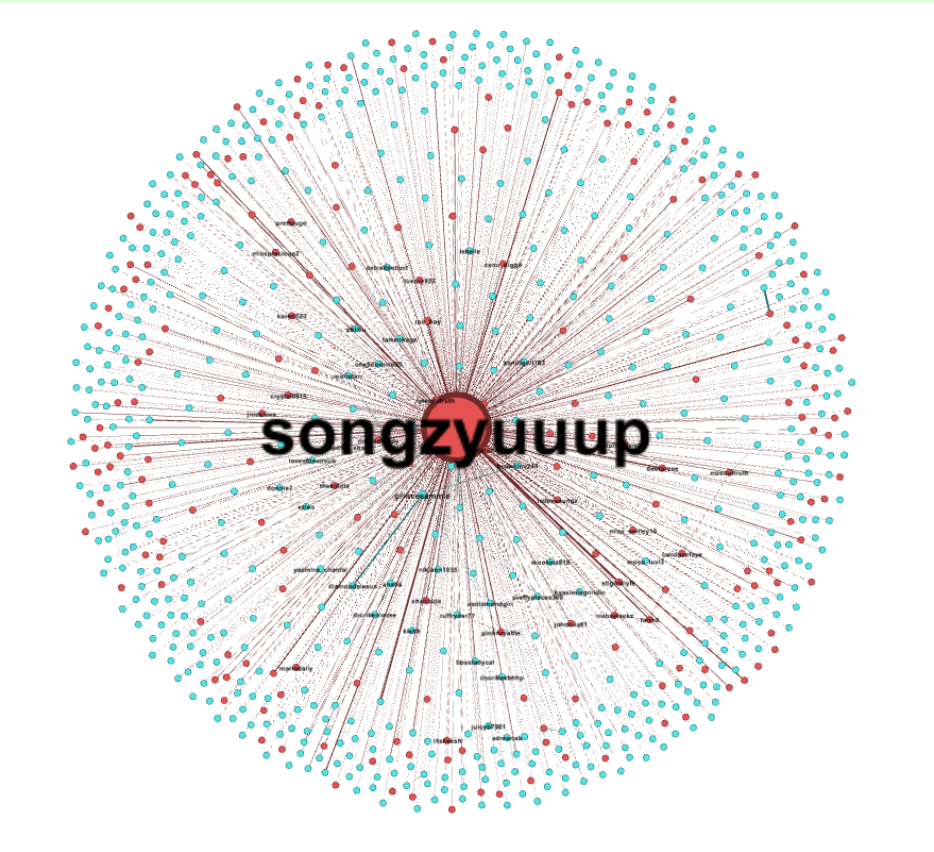


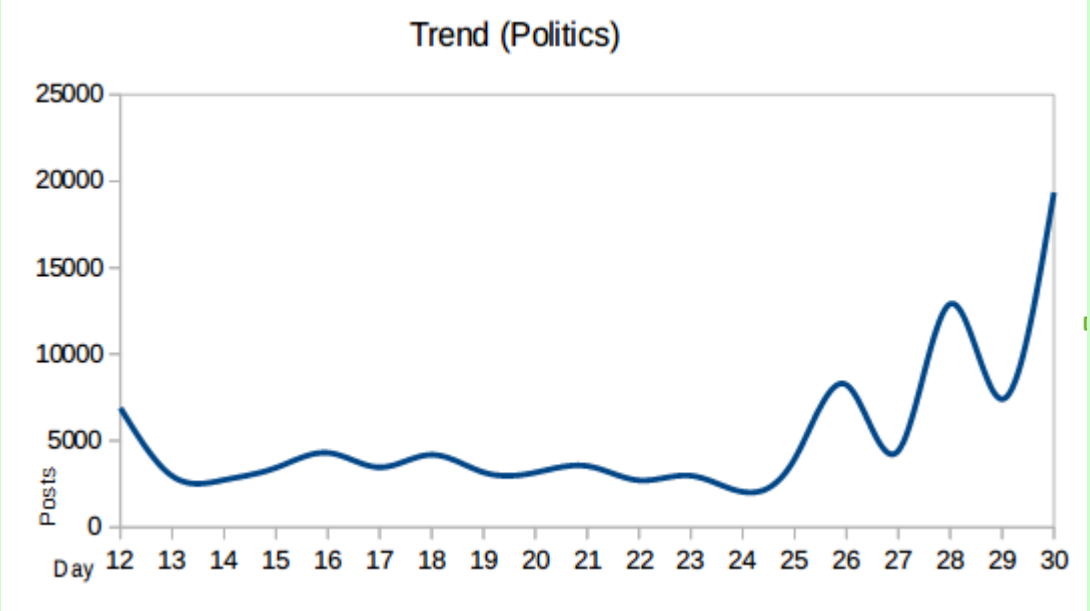
Figure 3: Influence of 'songzyuuup'

Figure 3 demonstrate the influence of user 'songzyuuup'. Nodes with red color are tweeting *Trey Songz*'s new album, while blue dot does not explicitly tweeted the new album. Gephi reports that 25.47% nodes in Figure 3 are red, while 74.53% are blue. This is a very high number considering the time delay in information propagation.

5 Conclusion

Networkx and Gephi are used on June 2009 Twitter data and uncovered some interesting conclusion underlying 2.6GB data. Even without using temporal pattern, we are able to identify different play style of social gamers, and identify the influence of a singer when his new album is released.

6 Supplement materials



Trend of the top ten topics on Twitter during June 2009

1. Politics:

1.1. **#Honduras** and **#crisishn**: On June 28, 2009, Honduran president *Manuel Zelaya* was ordered to be sent to exile but refused to comply, starting 2009 Honduran constitutional crisis. **#crisishn** means crisis in Honduras.

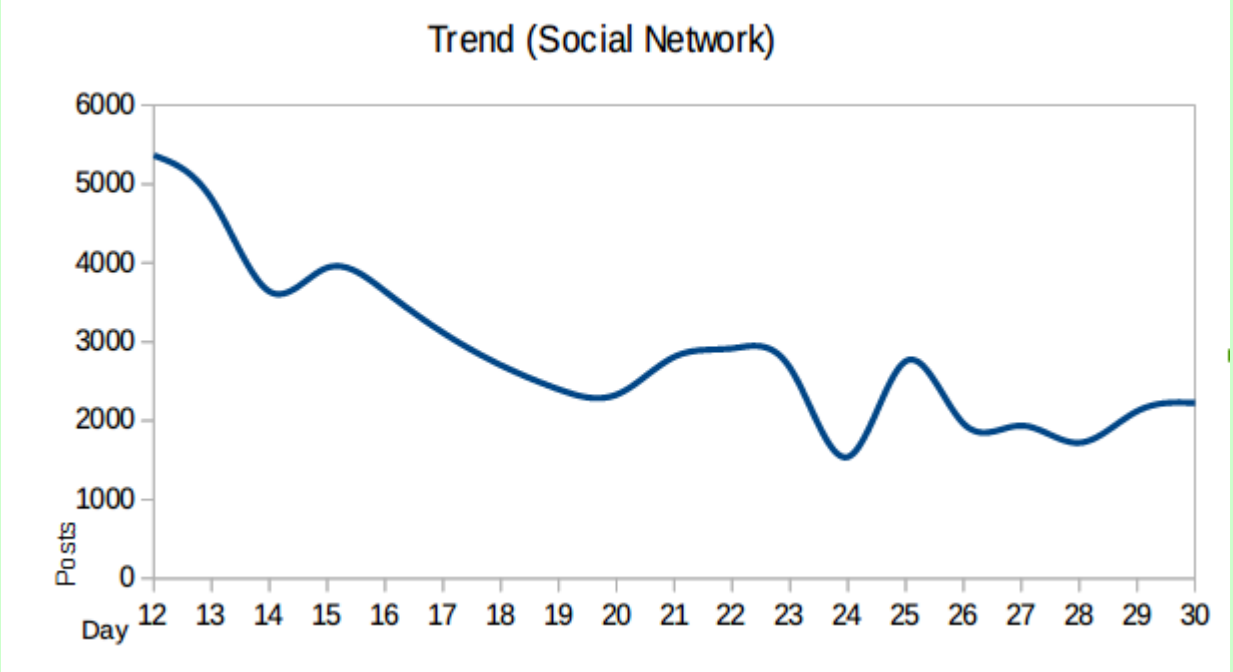
1.2. **#zensursula**: Online nickname given to *Ursula von der Leyen*, the former German Minister for Family Affairs. She enforces censorship of child pornography on the Internet around 2009. Now (2015) it's more focused on general internet censorship

1.3. **#forasarney** and **#chupa**: It's a call for *José Sarney*'s resignation. He is the president of the Brazilian senate and he is being accused of corruption, with lots of irregularities. **#chupa** is a Mexican swear word.

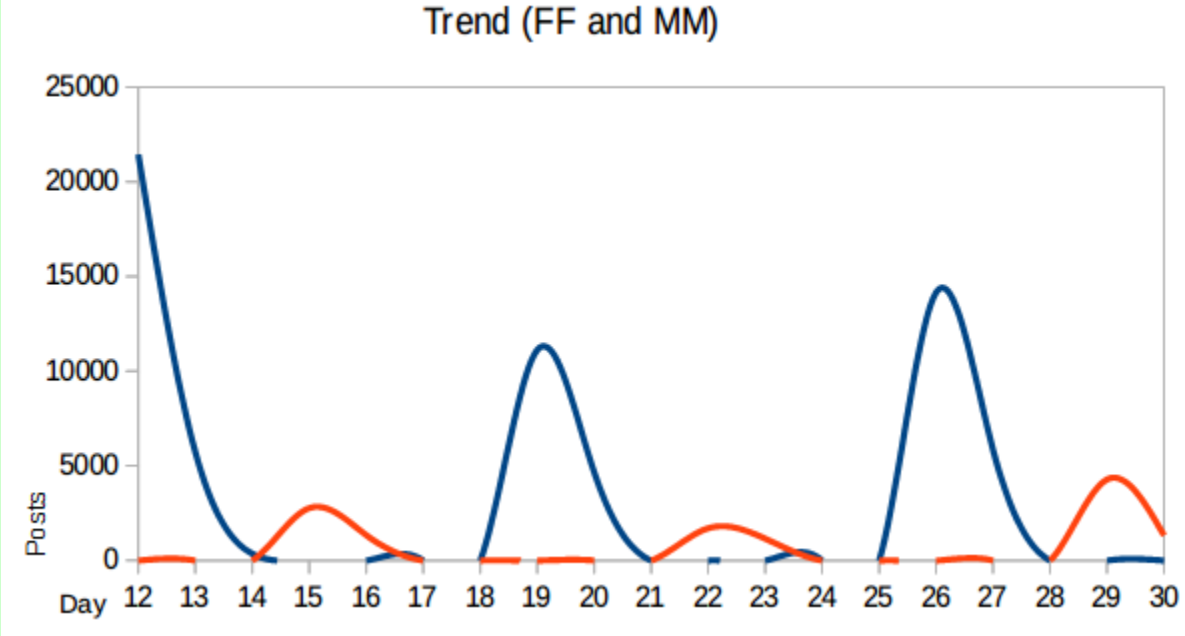
1.4. **#tcot** and **#p2**: In one sentence, the left is **#p2** progressive side, the right **#tcot** conservative side. This marks user's political view.

1.5. **#gilad**: *Gilad Shalit* is an Israeli sports columnist. She was captured by Hamas on June 25, 2009.

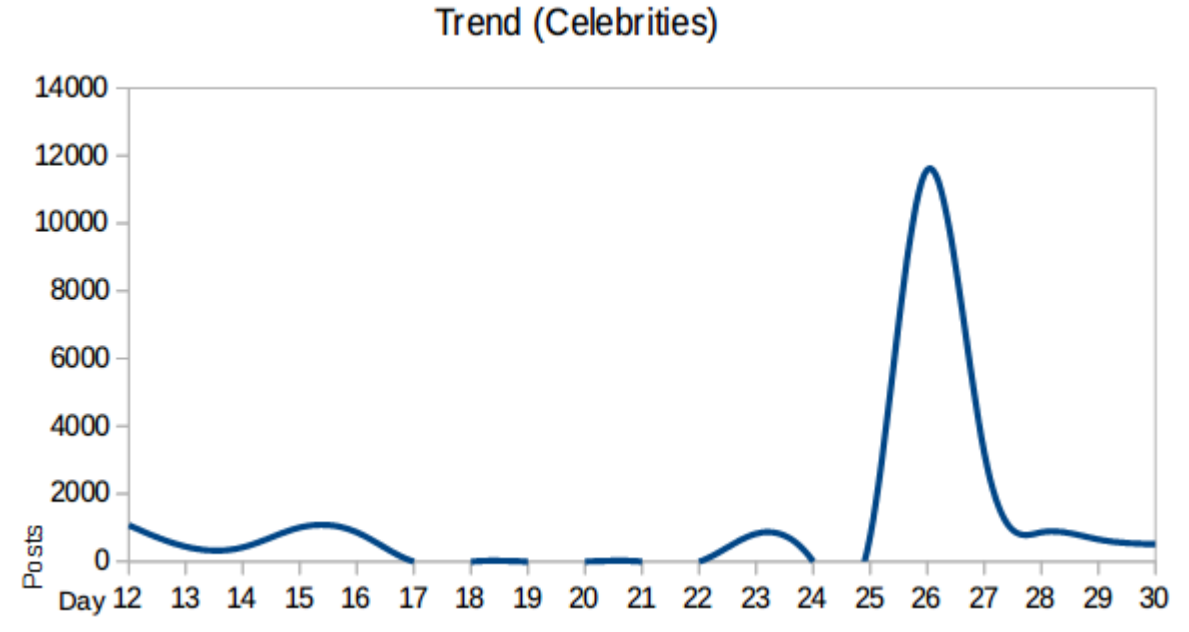
1.6. **#news:** general news reports on Twitter.



2. Social network:
    2.1. **#squarespace**: Automatic posts posted to Twitter from Squarespace.
    2.2. **#fb**: Automatic posts posted to Twitter from Facebook.



3. Follow Friday: **#ff** and **#followfriday**: A way to introduce your friend to your other friend. Usually happens on Friday.
4. Music Monday: **#musicmonday**: Another way to know more friends.

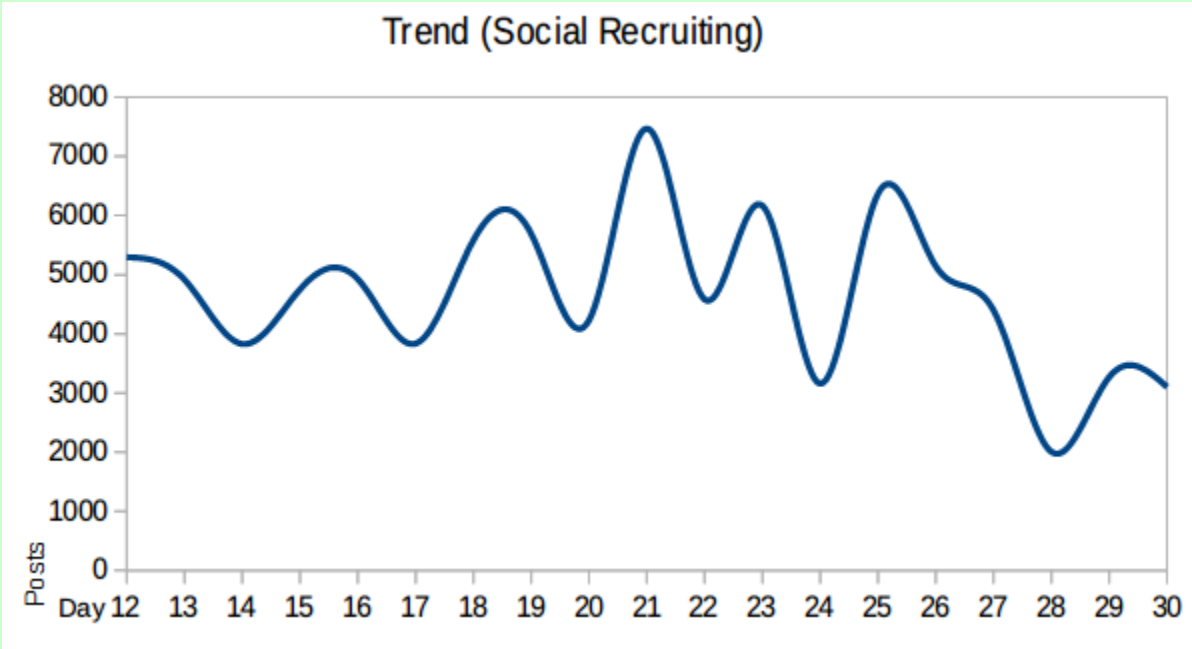


5. Celebrities:

   5.1. **#michaeljackson**: June 25, 2009, MJ died.

   5.2. **#unfollwperez** and **#unfollowperezhilton**: June 2009 Perez Hilton started the so called “Post-MMVA incident” in Toronto. He received little sympathy in the media and general public.

   5.3. **#bsb**: Backstreet Boys “Unbreakable” tour. They are about to release their album “This Is Us” on October 2009.



6. Job:

   6.1. **#jobs** and **#tweetmyjobs**: social recruiting.

7. Iran:

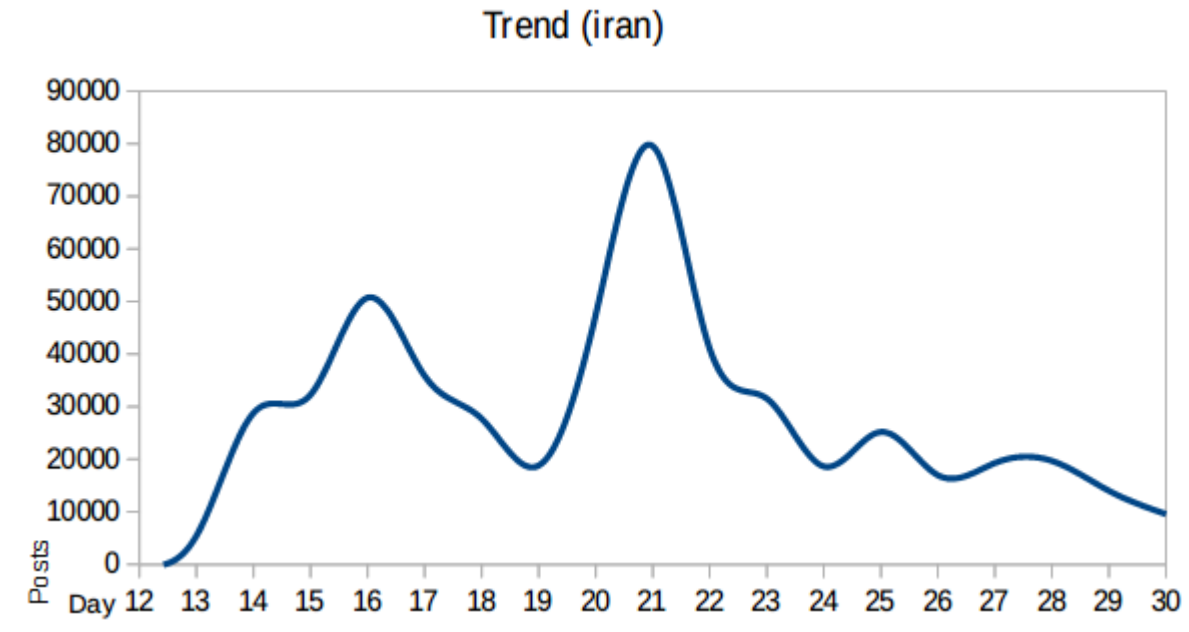
Background story: 2009 Iran president election was a total mess. People got killed, poll results got modified by corrupted government officials, press releases were biased, riots started on Iranian street. Iran was in hell by then. This riot is now dubbed "Green Revolution",[1] or "Twitter Revolution"

7.1. **#neda**: Philosophy student *Nedā Āghā-Soltān* was shot dead during the election protest on June 20. 2009.

7.2. **#iran**, **#iranelectrion**, **#helpiranelection**, **#iranians**, #**tehran** and **#iran9**: People are talking about Iran election. **#helpiranelection** marks their determination to support democracy for Iran. **#tehran** is the capital city of Iran.

7.3. **#gr88**: This event is named as "Green Revolution".

7.4. **#cnnfail**: Twitter users blasted CNN around June 15. 2009 for a lack of coverage of the Tehran protests, with Iranian citizens claiming ballot fraud and taking to the streets.



8. Fashion:

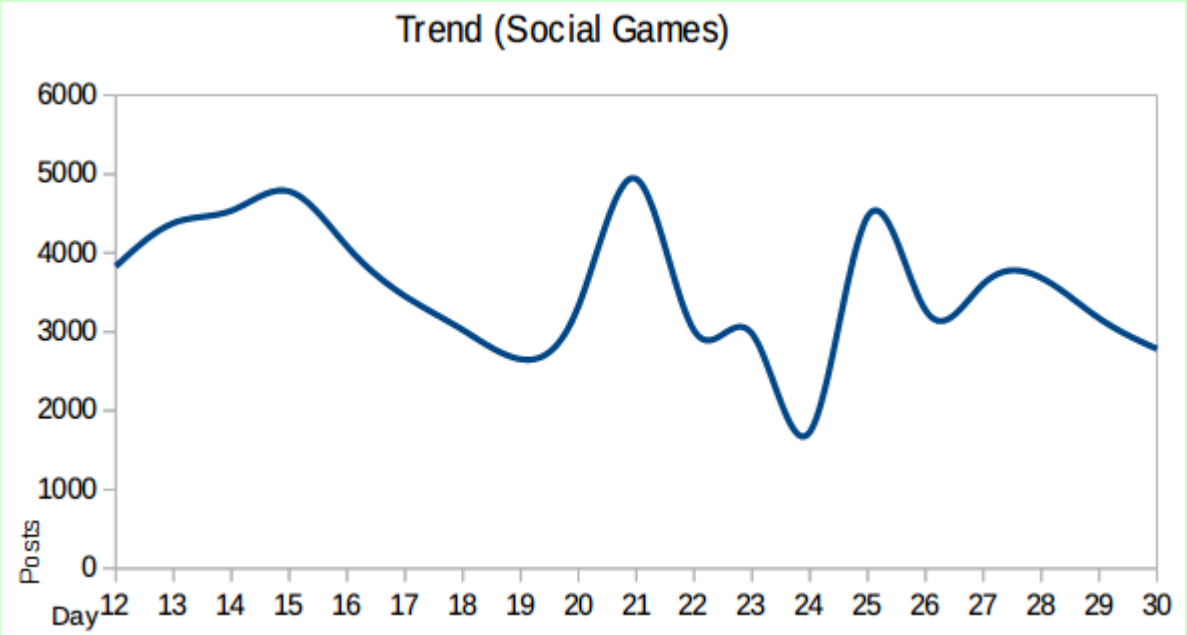
8.1. **#jtv**: Jewelry television is a representation of fashion.

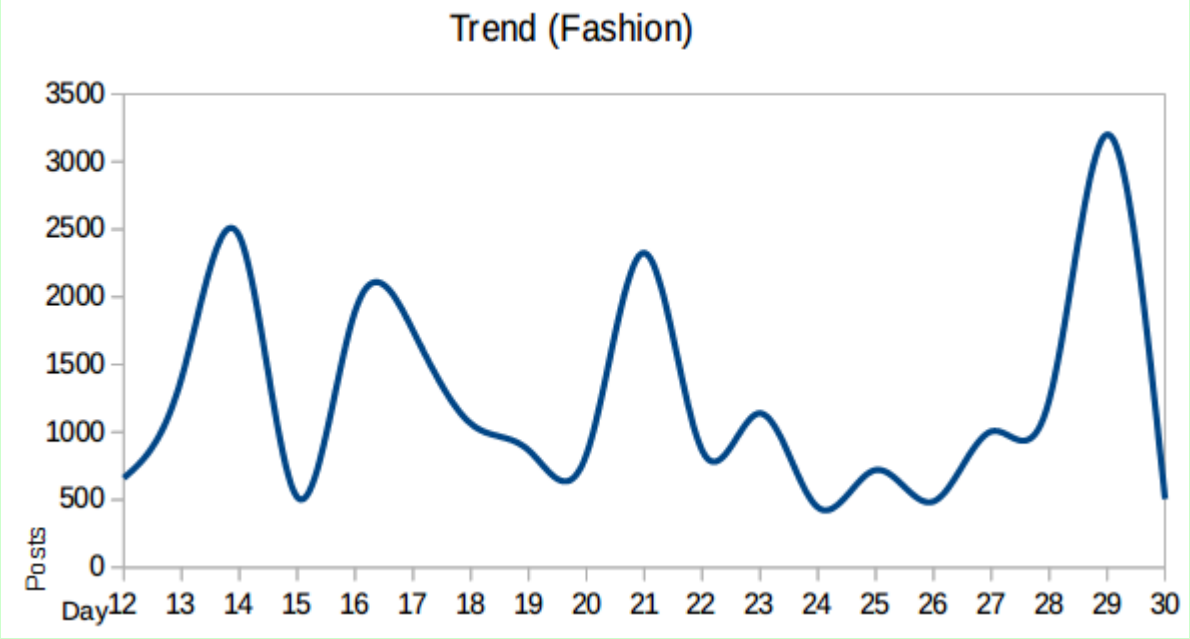
8.2. **#iphone**: so does iphone.

9. Social games:

9.1. **#140mafia** and **#spymaster**: Social network based browser games. Just like Candy Crush and Farmville.

1 2009 Iranian presidential election and the ensuing chaotic society: https://en.wikipedia.org/wiki/2009_Iranian_presidential_election_protests and https://en.wikipedia.org/wiki/Twitter_Revolution#Case_Study:_Twitter_Revolution_in_Iran


Trend (Social Games)
6000
5000
4000
3000
2000
1000
0
Posts
Day
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30


Trend (Fashion)
3500
3000
2500
2000
1500
1000
500
0
Posts
Day
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30