Spatiotemporal patterns of Sick twitters

Gianrocco Lazzari

December 3, 2015

1 Exploratory data analysis - sick_users

Overview of the work-flow:
From S3 amazon servers, get binary files of tweets (binary storage allow to save more space) → transform in CSV → R data analysis

Folder structure from Amazon S3 bucket *salathegroup-publications-supplemental*:

1. one_hundred: 19 files
2. all_tweets: 1202 files
3. sick_users: 2 files

(after merging the two files: "0000000" and "0000001" → sick_df) I found:

1.1 tweets

- the data-frames have 6 columns, as indeed set in Todd's scripts for parsing binary files: `print(data[0],data[1],data[2],data[3],sick,state,sep=",")`
- 4131650 *tweets* - `nrow(sick_df)`
- 23677 sick *tweets* - `sum(sick_df$sick==1)`

1.2 data[0]: user-ID

- 222446 **users** - `length(unique(sick_df$userID))`. Users activity distribution is reported in fig. 1(a) - please, note that *the distribution is not power-law*, as one can see from the same plot, in log-log scale 1(b)
- 23386 sick **users** - `length(unique(sick_df$userID[sick_df$sick==1]))` this means that there are only, on average 23677/23386≈1.012 tweets per sick user....very difficult to see patterns.. (Todd says: *Having ≈one tweet per sick period sounds reasonable*)

1.3 data[1,2]: longitude and latitude

- 117104 different longitude positions - `length(unique(sick_df$longitude))` - fig. 2(b)
- 118276 different latitude positions - `length(unique(sick_df$latitude))` - fig. 2(a) ...here there might be something weird since there are *9 tweets from the south hemisphere* and even 1 tweet at the South pole: `sick_df$latitude[sick_df$latitude<0]`

  [1] -34.834248 -15.970366 -47.289993 -12.413708

  [5] -89.999901 -22.420408 -22.420408 -15.458479

[9] -4.767236

See in fig. 3(a) the geographical distribution (12 tweets are out of the map) and in fig. 3(c) an example of sick tweets from a single user.

1.4 data[3]: Unix time

- Definition: `https://en.wikipedia.org/wiki/Unix_time#Encoding_time_as_a_number`
- 4036965 different tweet events - `length(unique(sick_df$time))`
  this means that there are on average tweets/events = 4131650/4036965 ≈ 1.023 tweets/event (at the Unix time resolution).
- 217 tweet events happening at time: '0' that we have to eliminate! - `time_cnt<-table(sick_df$time)[1]`.
  This is probably an error from the recording, since all other events have 1 or 2 tweets happening at the same time...
  Todd suggests: "Yes, ignore any tweets from before March 2011."
  Indeed, after the events at time=0, the first tweets happened on 2011-03-02
- of course, this leads to one event less: 4036964 different tweets events - `length(unique(sick_df$time[sick_df$time!=0]))`, for which this is the distribution of tweets (we didn't plot the histogram as it's difficult to read):

num. of tweets/event:	1	2	3	4	5
num. of events:	3944379	90743	1801	40	1

  this is of course the effect of resolution sampling, that for the Unix time = 1 sec.
- Tweets per day are reported in fig. 4(a). Before 2015 we found 11 days in total of no recording:
  2011: "2011-03-31" "2011-04-01" "2011-04-02" "2011-04-03"
  2014: "2014-03-26" "2014-08-08" "2014-11-25" "2014-11-26" "2014-11-27" "2014-11-28" "2014-11-29".

  The recording stopped for a long period, on "2015-03-02" and restated again on "2015-07-05" (112 days gap - in the middle there was recording in few days). All gaps (123 days in total) are highlighted in fig. 4(e).

  Todd said, "The missing data in 2011 and 2014 are due to outages on the collector"
- **Sick** tweets across all recording periods are reported in fig. 4(c). There is something weird as all the 'bumps' in this sick signal are around Christmas. Indeed these are the days with > 100 tweets: "2012-12-25 CET" "2013-12-24 CET" "2013-12-25 CET" "2013-12-26 CET" "2014-12-24 CET" "2014-12-25 CET" - `as.POSIXct(sick_hist$breaks[sick_hist$counts>100],origin="1970-01-01")`. The relative abundance of 'sick tweets' is shown in fig. 4(d).
- 1470 days of recording, in total - `length(unique(all_tweets_date))`.
  As one can see from the daily activity histogram (median activity: 2921 tweets/day), in fig. 4(b), there are some "extreme event". We could consider for instance, a threshold of 5000 tweets/day and look for some event in the world that could have triggered the extreme tweeting on those days. Above such a threshold, we find 15 days, sorted hereby by number of tweets:

2013-12-25	2013-12-01	2013-11-30
5193	5211	5221
2013-11-29	2011-03-04	2013-11-27
5569	5577	5756
2013-11-26	2013-11-28	2013-11-25
6018	6069	6141
2013-11-24	2011-03-03	2013-11-23
6645	6729	7409
2013-11-22	2013-11-21	2013-11-20
7735	8349	8449

we tried to match some events on those dates:

- 2013-11-20 until 2013-12-01: Putin releases Greenpeace activists?[1] or maybe the earthquake in Ohio?[2]. There is not evidence of spacial geographical pattern whatsoever - see fig. 3(b)
- 2011-03-[03-04]: ?

1.5 States

- 52 different us states present - the variable spans in the range [0,56] (by Todd) - `length(unique(sick_df$state))` - histogram of 'tweets activity' is shown in fig. 1(c). Activity per state in fig. 1(d)
- 323282 tweets outside the USA - `sum(sick_df$state==56)` → *Todd says: "For the geographical analysis, I ignored any tweets not in the united states (stateid = 56)."*

2 Further statistical analysis

2.1 Waiting-time distributions

Given the gaps in recording (fig 4(e)), the histogram of inter-events intervals will have a 'long tail'. The longest time intervals are of course just an artifact of recording gaps. Therefore we first restrict the histogram to all tweets before "2015-03-02" (3983794 tweeting events) - see fig. 5(a). In the histogram, there are still signature of quite large waiting-times - $\Delta t > 1$ day. For this reason, we zoom the histogram for intervals $\Delta t < 1$ h (3983629 events), in fig. 5(b).
In order to improve the fit, we further decreases the cut-off down to 20 min (3983414 events) . From fig. 5(b) (mind that it's a *semi-log* plot) is clear that the distribution is neither an exponential, nor a Gaussian. The possible presence of a power-law-like scaling might be suggested by the correspondent log-log plot, in fig. **??**, more likely for intervals $\Delta t > 1$ min. Using the `poweRlaw` R package [1], we try to fit few heavy-tailed distributions (maximum likelihood:'ml'; Kolmogorov-Smirnov statistic: 'KS') - see fig. 5(c) :

distribution	x_{min}	parameters	sd	ml	KS
power-law	169	$\alpha = 3.58651$	–	353206.2	0.018
log-normal	78	$\mu = 2.403108, \sigma = 1.147314$	–	–	0.0014

From the plot, a log-normal distribution has a much better fit, reflected also by the KS statistic.

Now, a question naturally arises: whether there is difference in the waiting-times when users get sick. For this reason we compare the inter-events distribution, for 'sick' (23632 events) and 'non-sick' (3961238 events) tweets, in fig. 5(d).

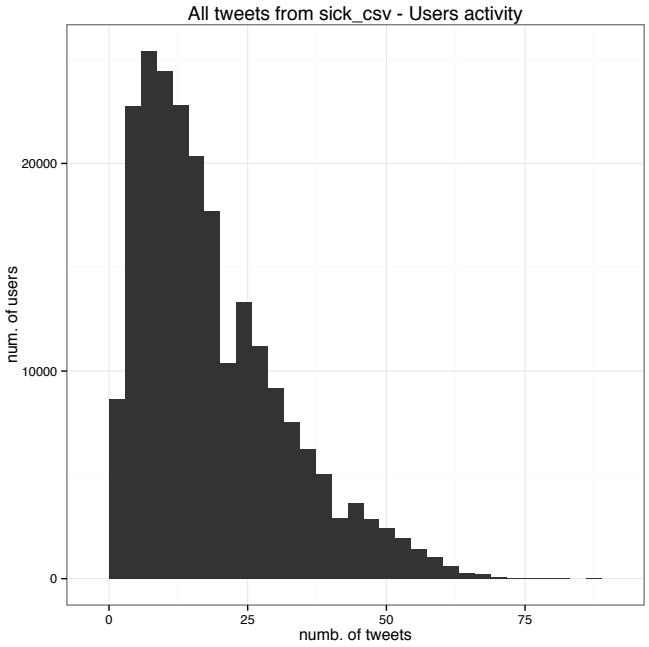
2.2 Spatio-temporal patterns

Compute all the geographical distance between two consecutive tweets of the same users, for all the users, it's quite time consuming (on 1 core - with no parallelization - it took ca. 98 min).
For all the histograms, where indicated, "binwidth: FD" refers to Freedman-Diaconis rule[3].
The distribution of tweets/user and mean-distance, for the 22330 sick users are shown in fig. 6(a) and fig. 6(b) (this number is less then the overall num. of sick users , as we are counting tweets before 2015-03-02; furthermore, we consider only users with more than 1 tweet of course, in order to compute at least 1 distance). On average they moved 270 km (average s.d. $\sim$ 367 km), over a total number of 'inter-tweets distances' = 342222 (`sum(sapply(sick_dist,length))`).
The distribution of num. sick tweets/sick user is better represented in a table:

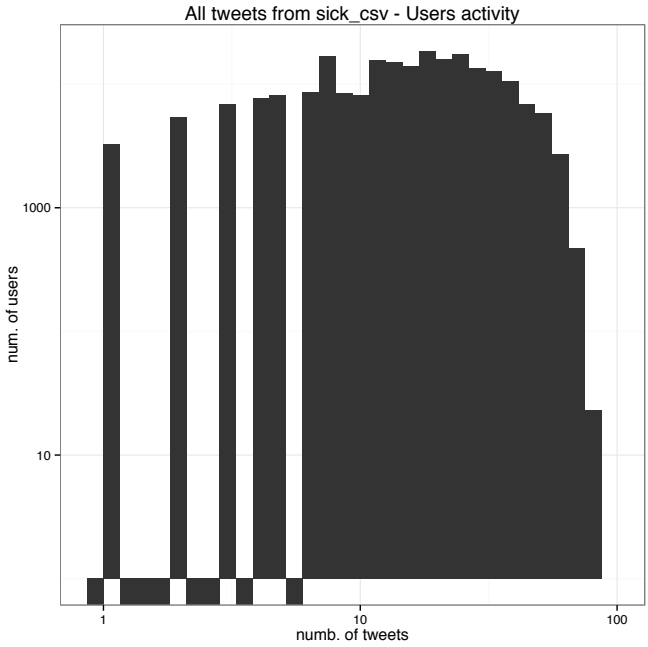
[1] `http://www.theguardian.com/environment/2013/nov/20/british-greenpeace-activist-alexandra-harris-freed-bail`
[2] `http://www.upi.com/blog/2013/11/20/35-magnitude-earthquake-shakes-southeast-Ohio/9361384995325/`
[3] `https://en.wikipedia.org/wiki/Freedman%E2%80%93Diaconis_rule`

num of tweets/user	1	2	3	4	5	9
num of user	23091	252	9	2	1	1

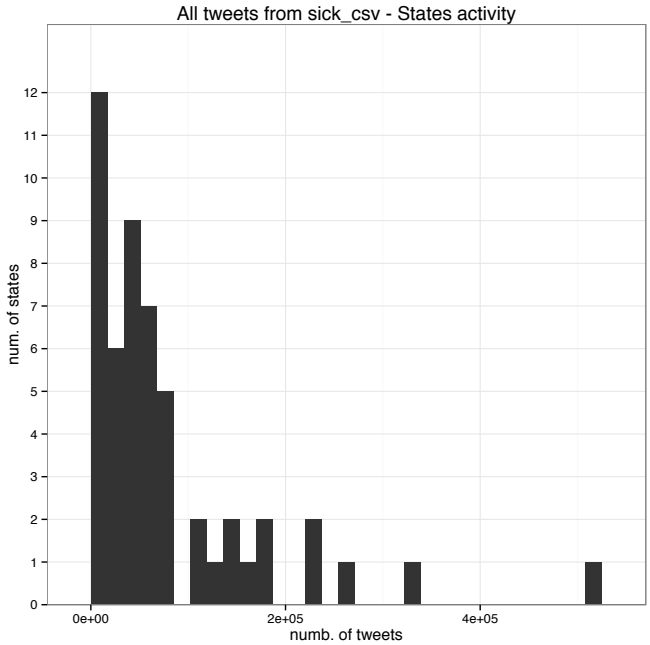
As you can see, we don't have a good statistic for the sick tweets distances, as most of the sick users have only 1 tweet!



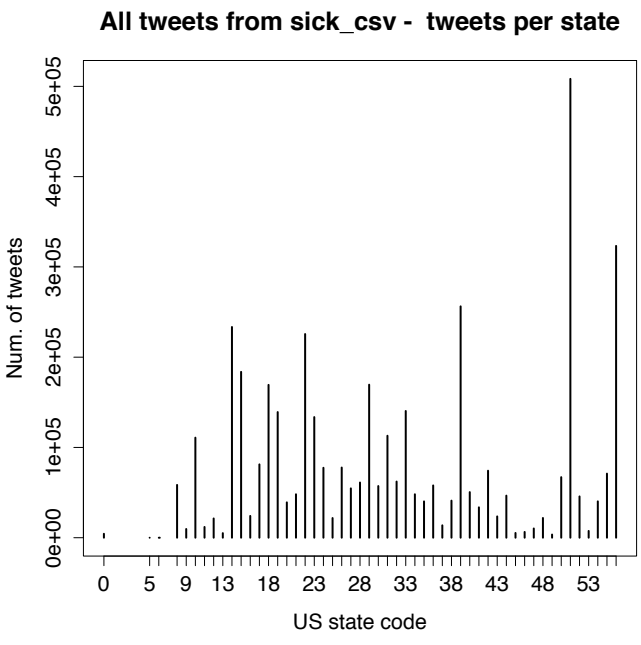
(a) Histogram of tweets per user



(b) Log-log histogram of tweets per user

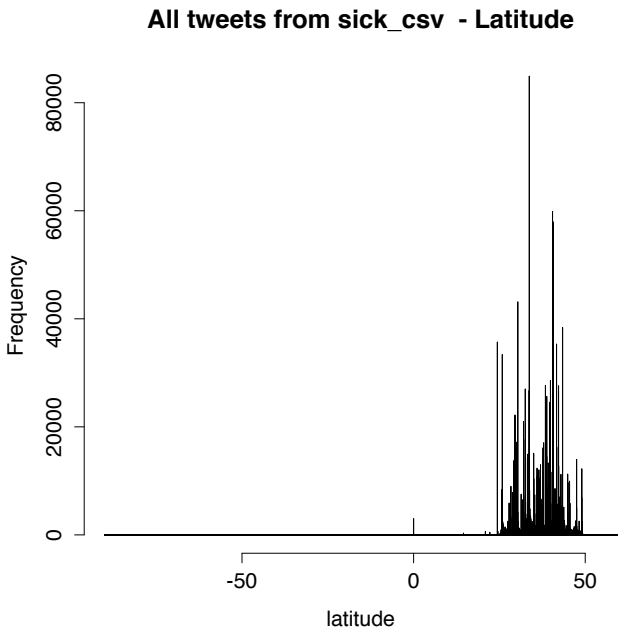


(c) Histogram of state activity

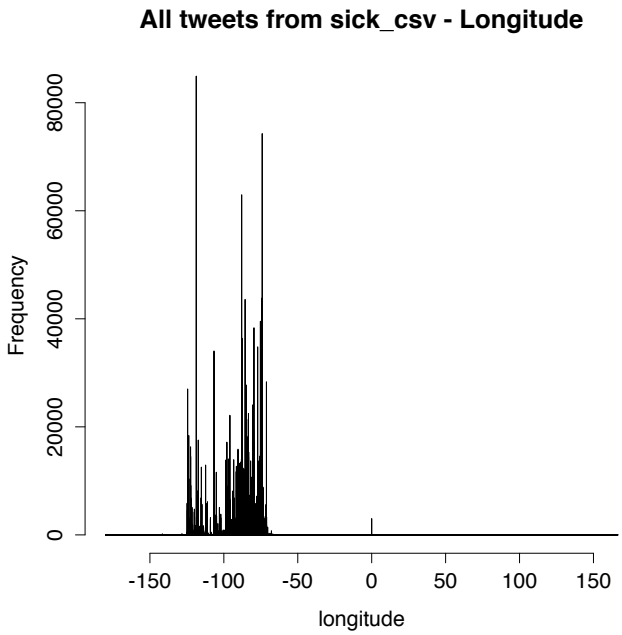


(d) Histogram of tweets per state

Figure 1: Activity distributions

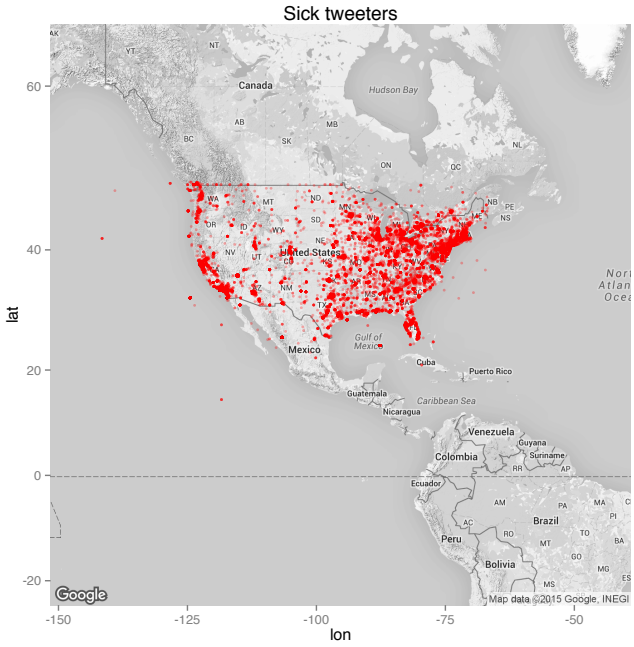


(a) Histogram of latitude

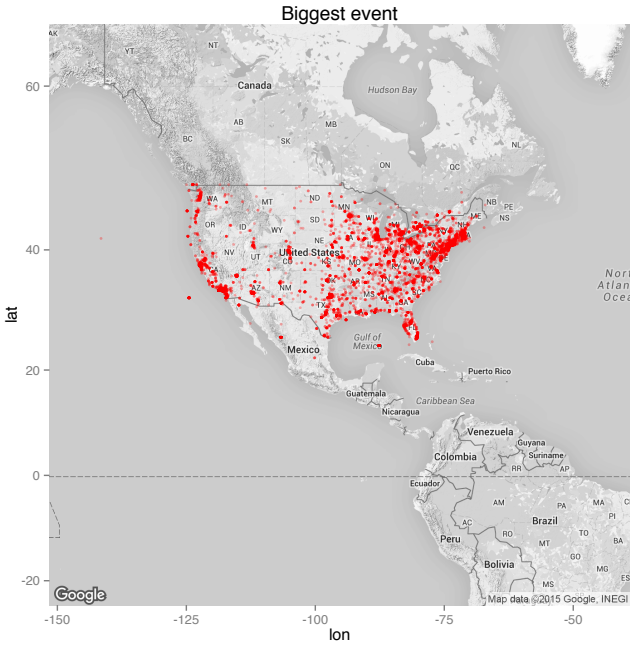


(b) Histogram of longitude

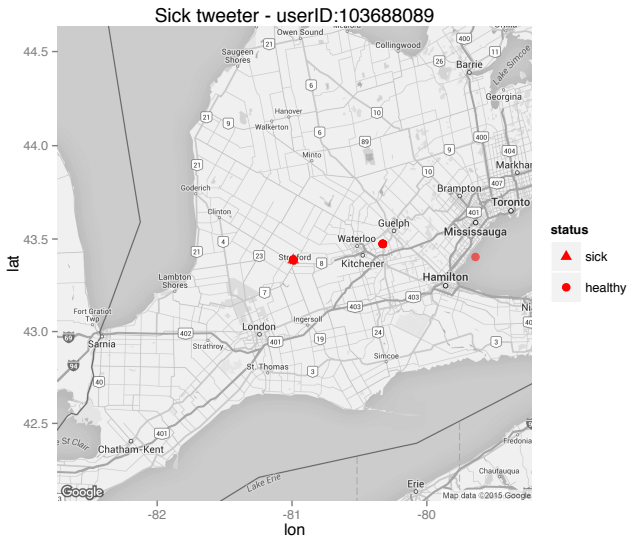
Figure 2: Geographical coordinates distribution



(a) Map of tweets from sick users

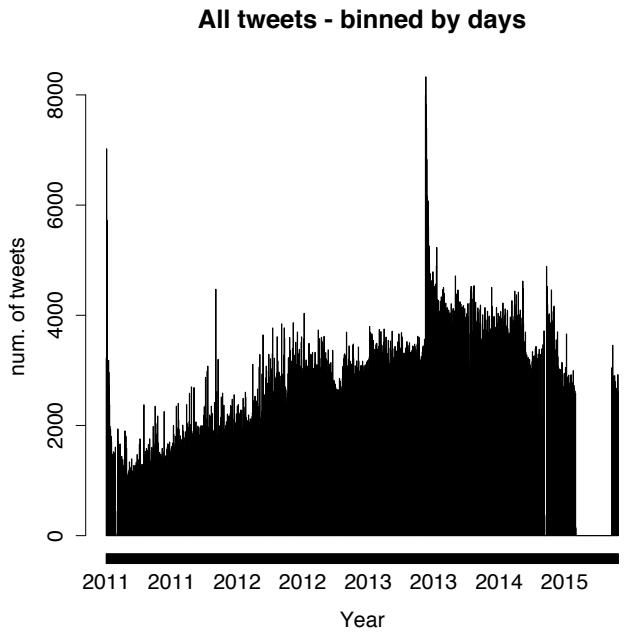


(b) Map of all tweets during 2013-11-20

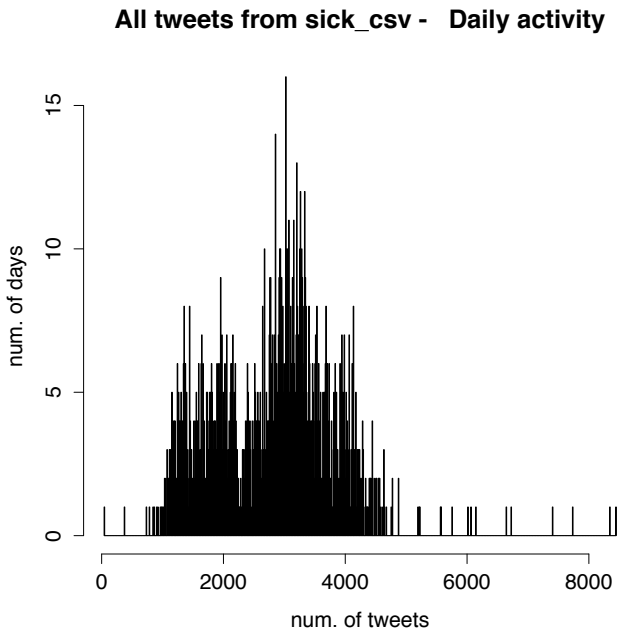


(c) Map of tweets from single sick user

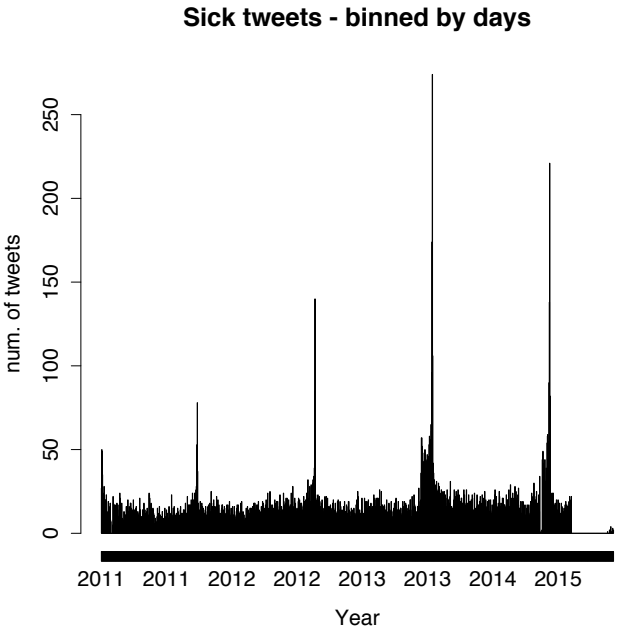
Figure 3: Geographical distribution of tweets



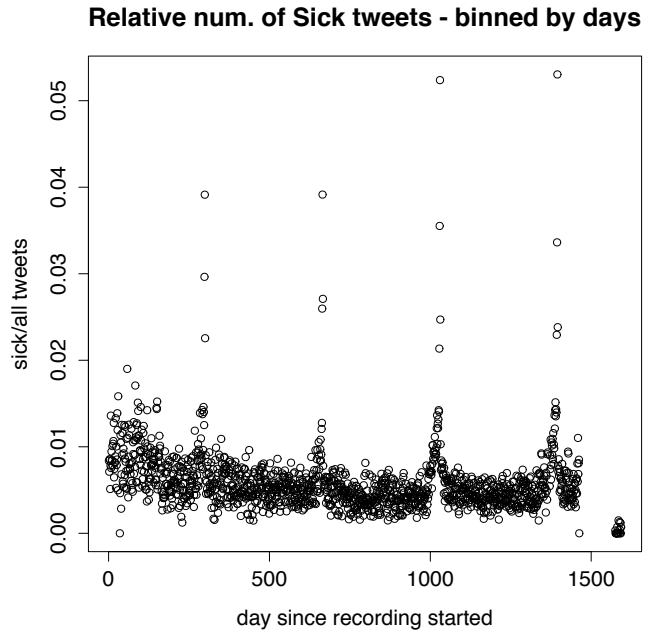
(a) All tweets across time - tweets at unix time 0 are excluded



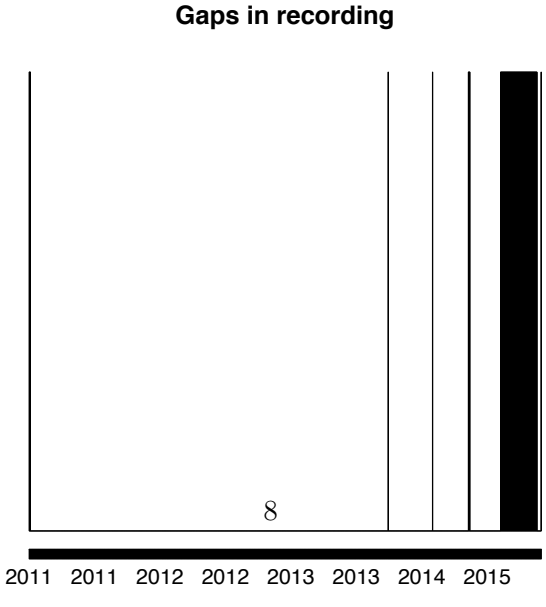
(b) Histogram of tweets per day

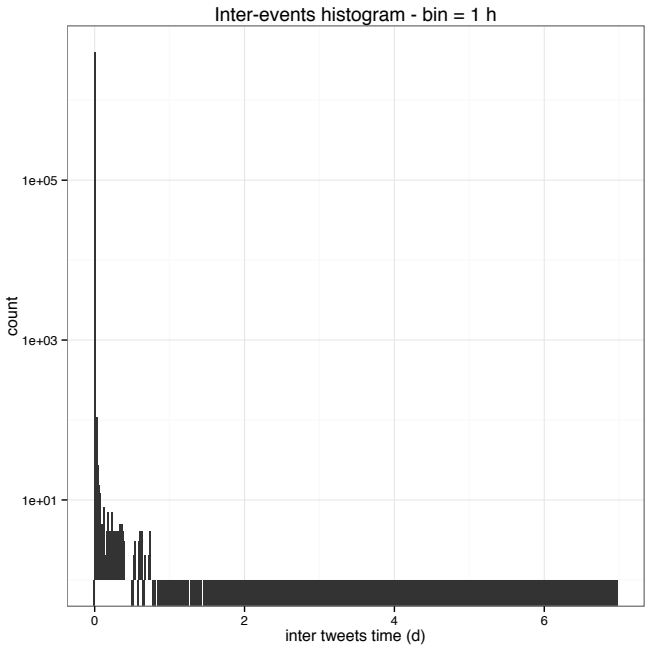


(c) Sick tweets across time

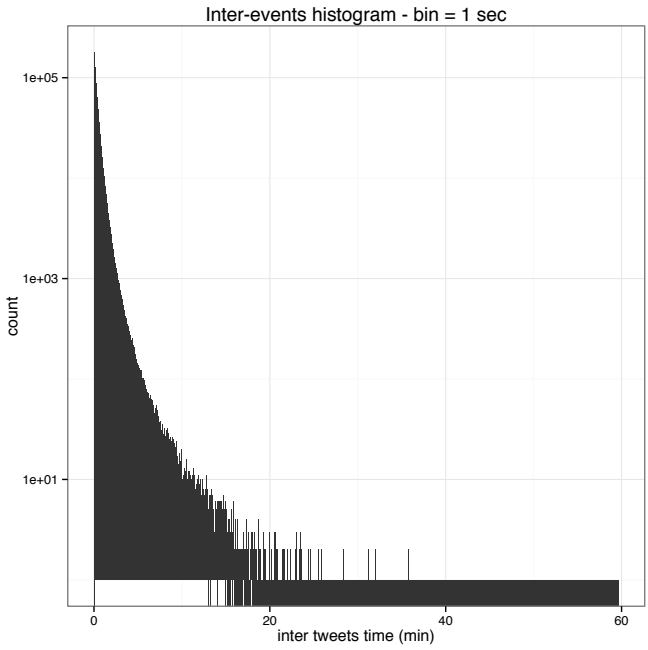


(d) Relative number of sick tweets across time

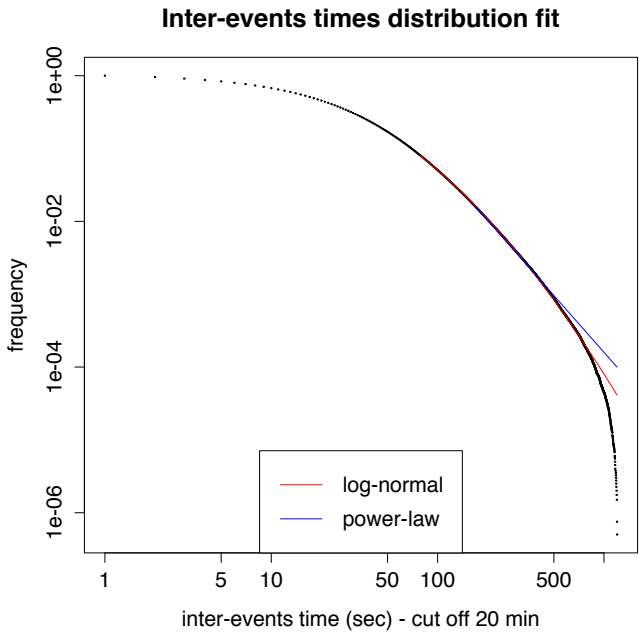




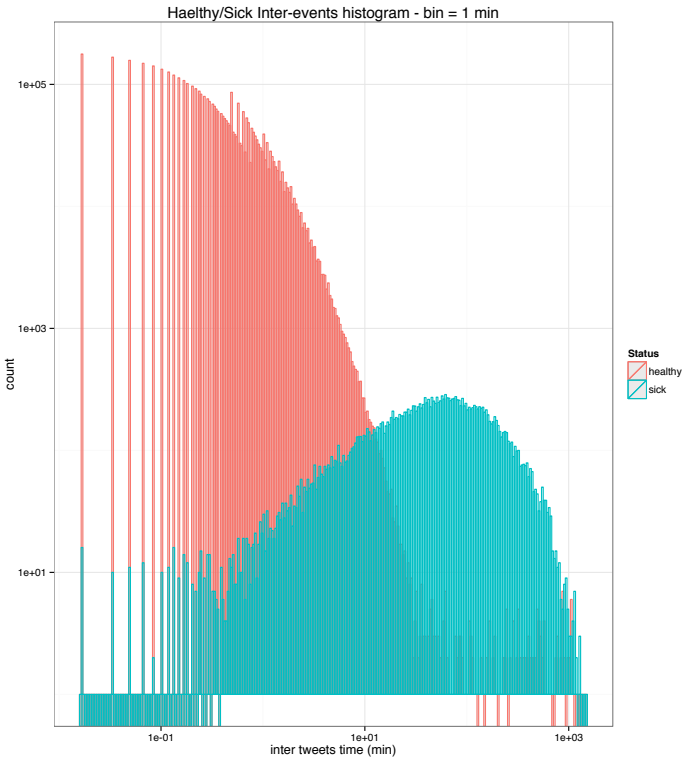
(a) Histogram of inter-tweets intervals



(b) Inter-tweets distributions, cut-off at 1 h - semi-log plot

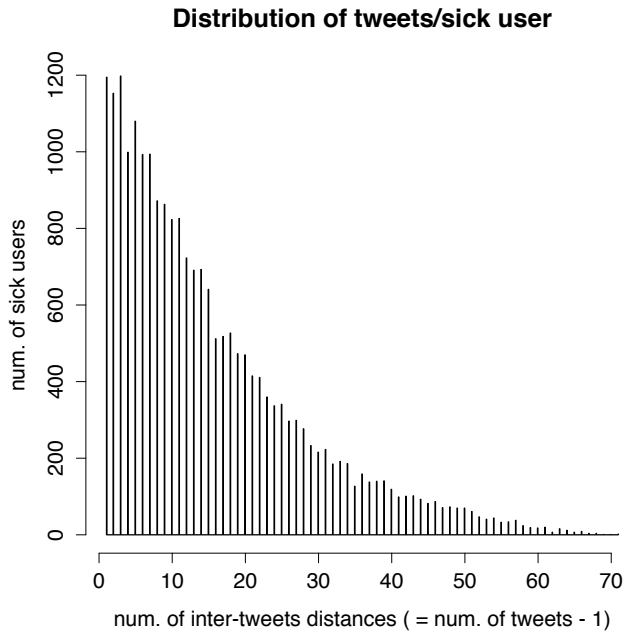


(c) Fit of inter-events distribution

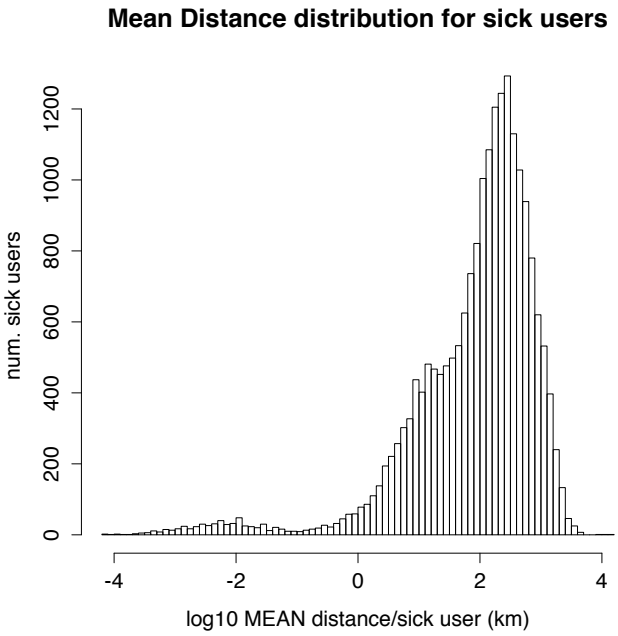


(d) inter-events distribution - cut-off: 1 day

Figure 5: Inter-events time distributions - tweets ma before 2015-03-02



(a) Distribution of num. of tweets/sick user - median = 11



(b) Mean distance distribution, for sick users - binwidth: FD

Figure 6: Distance and tweets-number distributions, for *sick users* before 2015-03-02. Mind that here all their tweets are present, *before and after getting sick.*

References

[1] Colin S. Gillespie. Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software*, 64(2):1–16, 2015.