Quantifying activity through repository mining: the case of Moose

J. J. Merelo

25 de agosto de 2016

Abstract

This is a report on the characteristics of the changes in a project from using the number of lines changed in every commit that affects those files. Our intention, by analyzing repositories that hold projects of different type such as books, research papers or simply software, is to show that they present remarkably similar characteristics, main diferences based on the fact that they are collaborative or not and the number of commits present. In this report we will visualize patterns in creativity bursts by examining the size of changes in projects in different ways, trying to prove that they fullfill the characteristics of a self-organized system in a critical state.

This is a report for a particular repository whose data is contained in the file /home/jmerelo/proyectos/literaturame/data/softwar
This file contains a sequence of change sizes for every commit that affects those parcitular files. Since changes include insertion and deletion of files, the biggest of these values is taken; in particular, this means that the addition of all changes will not be equal to the sum of the sizes of all files.

Not all files in the repository are considered; some care has been taken to include only files written in the same language; also, in general only files that have actually been written by the user have been included. For that reason, the size of the first commit is eliminated from the sequence.

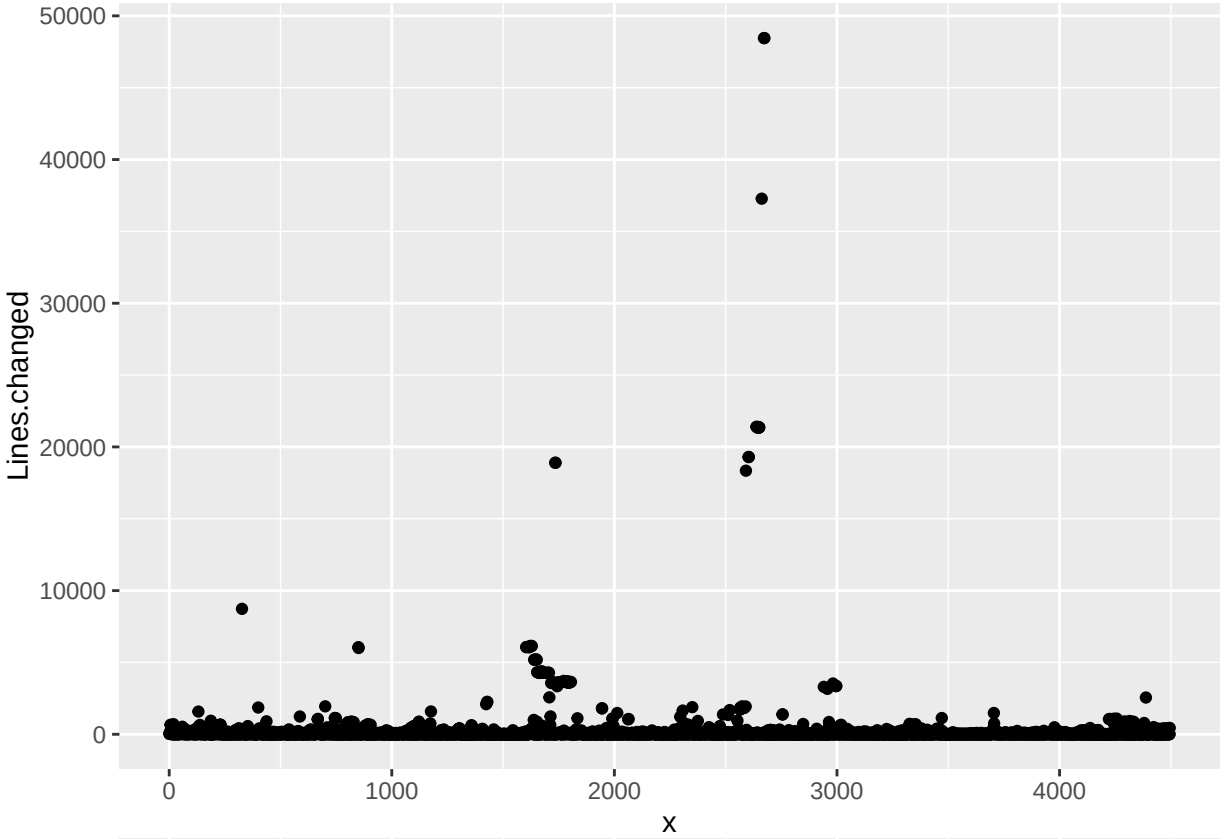
What I intend with these report is to show that *writing is writing*. It is a creative process, with the same characteristics, or similar, cutting accross types of writing, be it software, fiction or other kind of *literary* writing.

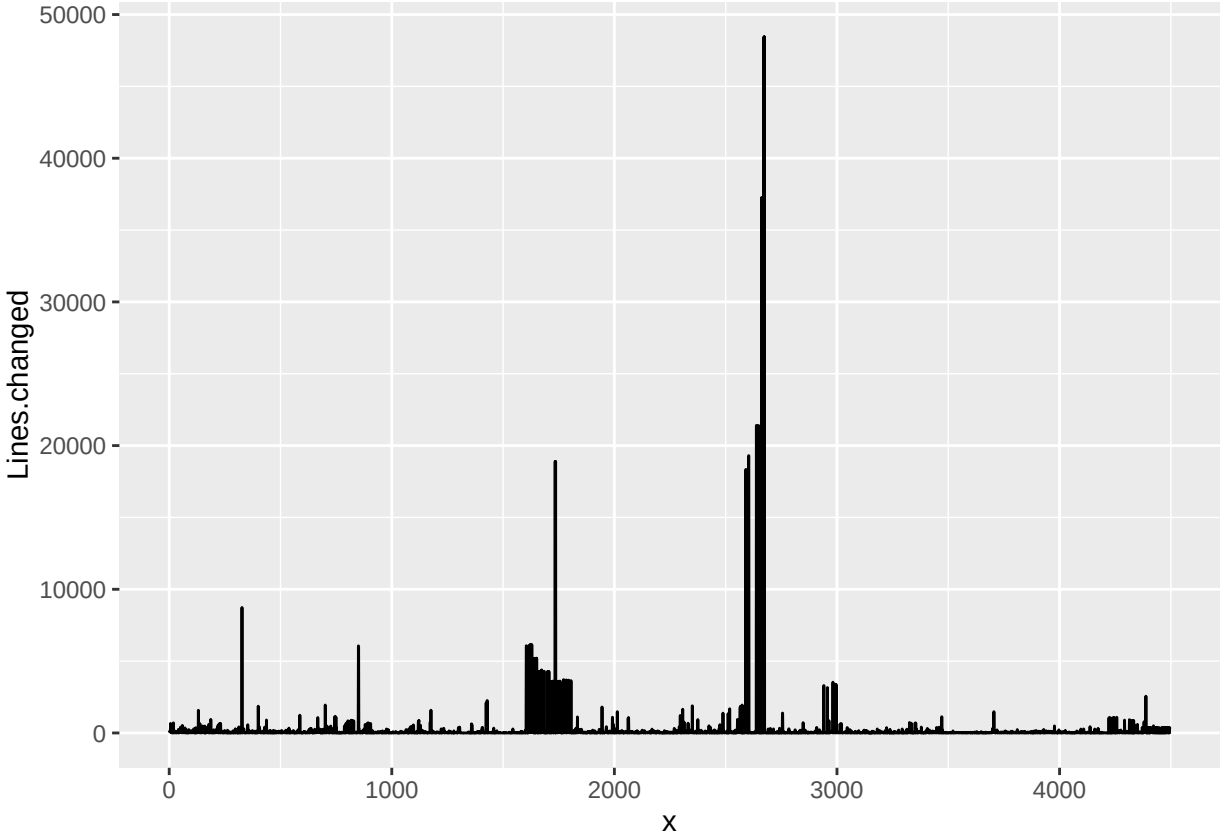
This is a summary of the characteristics of commit sizes.

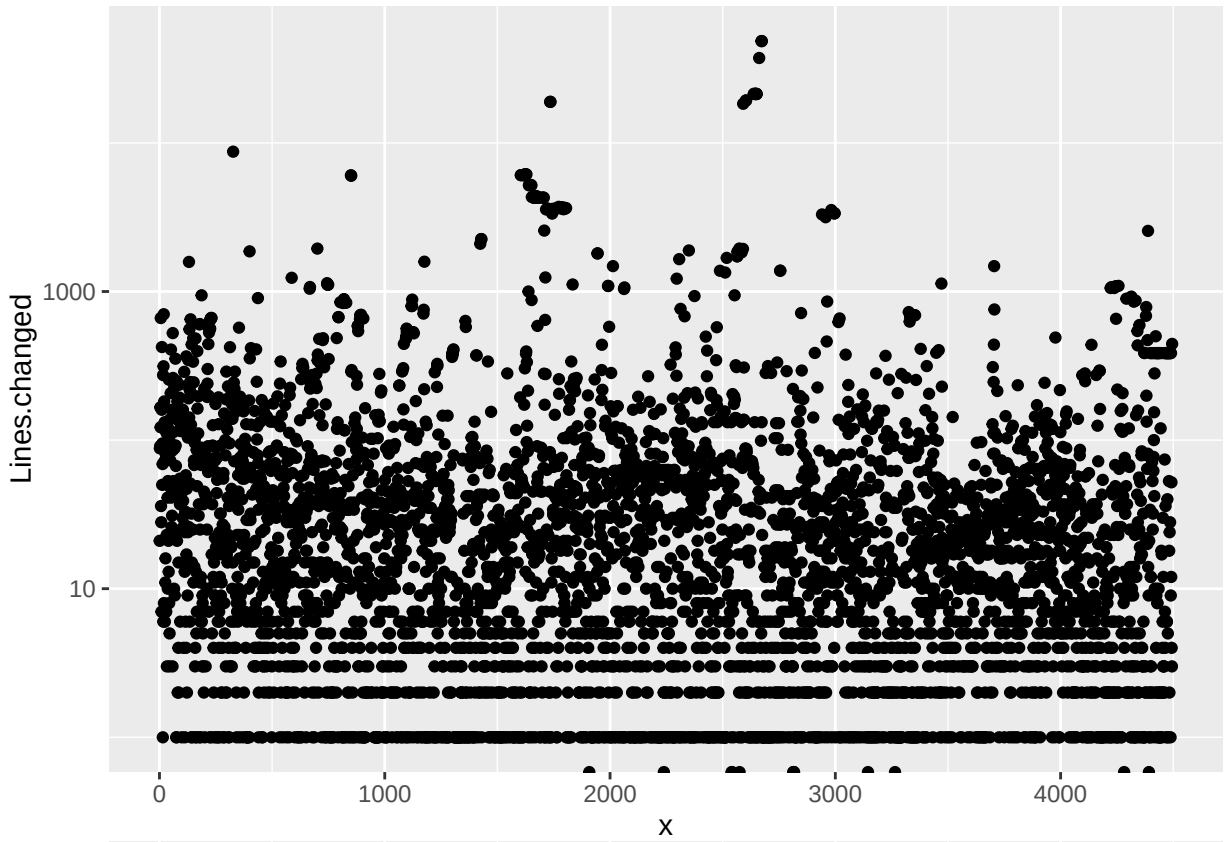
```
summary(lines)
```

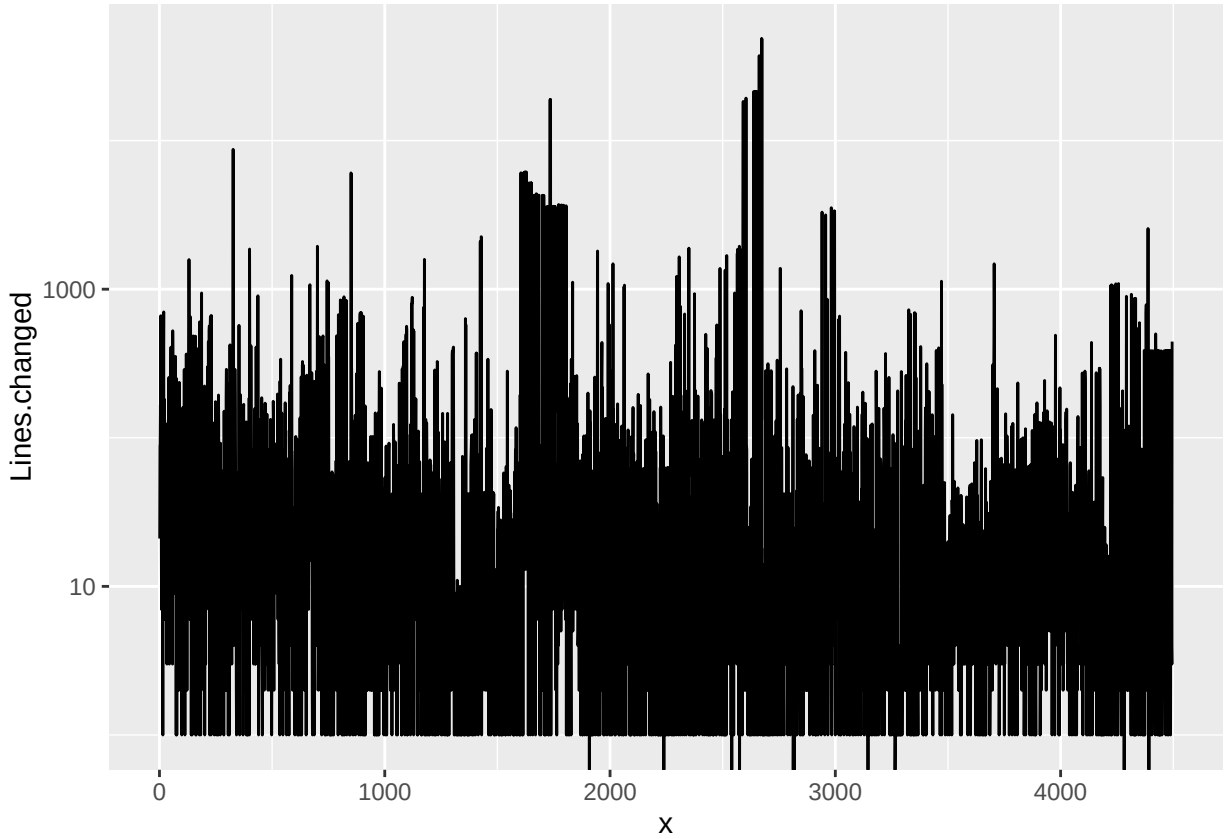
```
##  Lines.changed          SMA5             SMA10
##  Min.   :    0.0   Min.   :    0.6   Min.   :    1.00
##  1st Qu.:    4.0   1st Qu.:   17.2   1st Qu.:   23.73
##  Median :   16.0   Median :   35.2   Median :   41.75
##  Mean   :  218.4   Mean   :  218.5   Mean   :  218.59
##  3rd Qu.:   53.0   3rd Qu.:   84.6   3rd Qu.:  106.30
##  Max.   :48454.0   Max.   :19428.6   Max.   :14962.40
##                    NA's   :4         NA's   :9
```

The timeline of the commit sizes is represented next, in different ways, with logarithmic or decimal y scale. The serrated characteristics is the same, as well as the big changes in scale. x axis is simply the temporal sequence of commits, while the y axis is the absolute size.


50000
40000
30000
20000
10000
0
Lines.changed
0
1000
2000
3000
4000
x


50000
40000
30000
20000
10000
0
Lines.changed
0
1000
2000
3000
4000
x

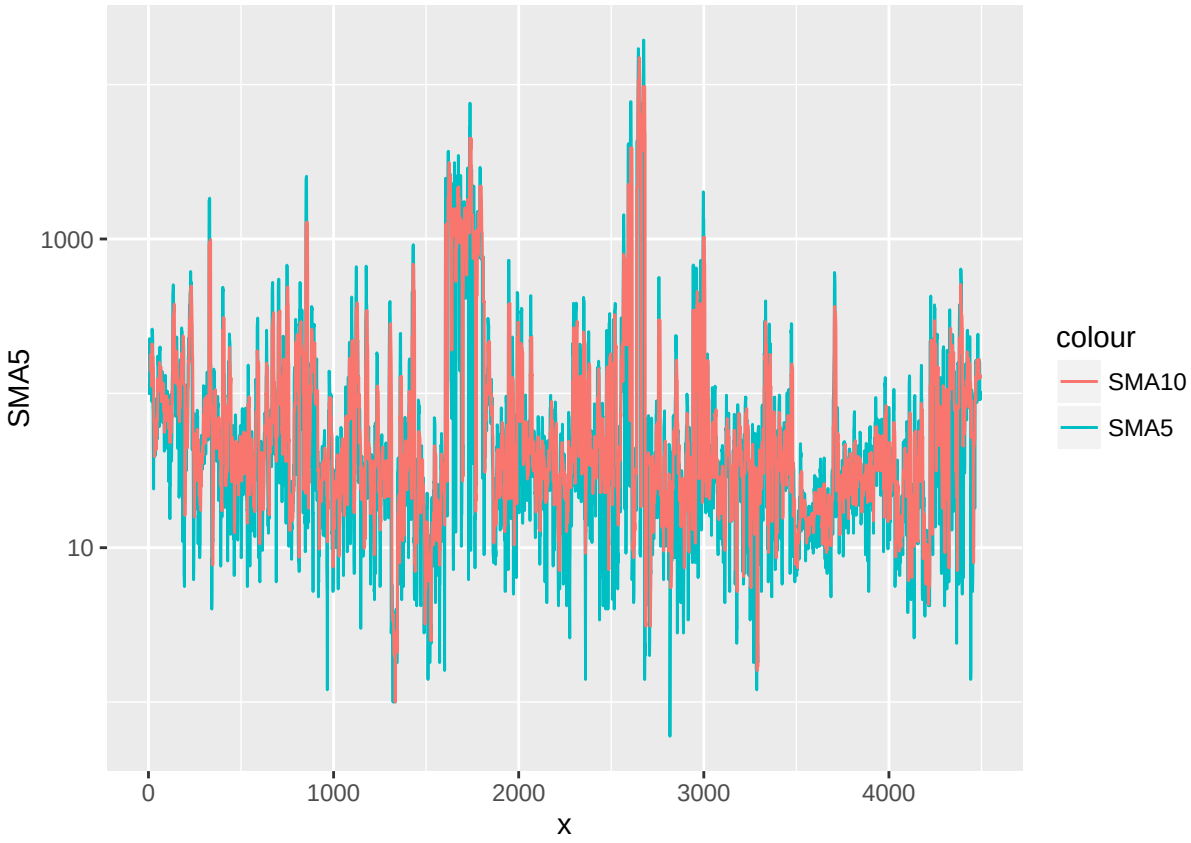




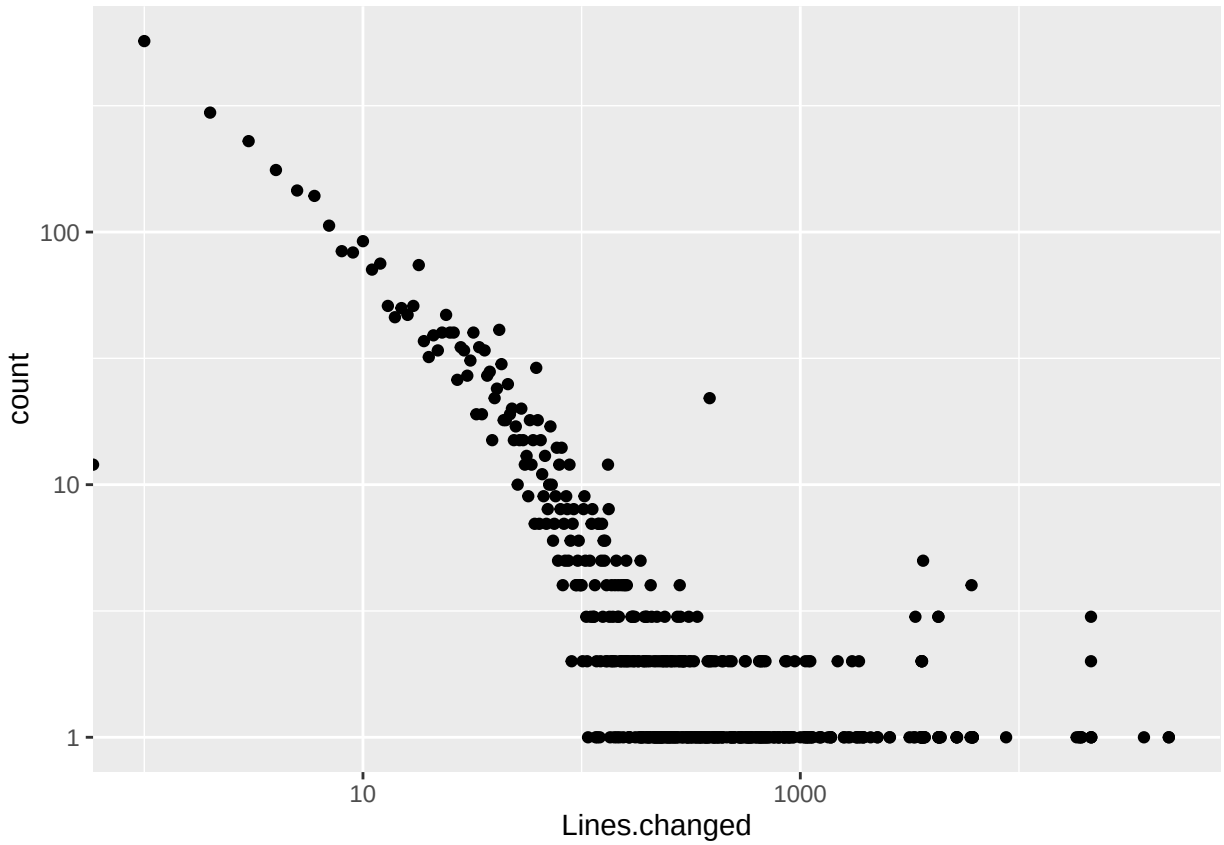
Probably it will be a bit more visible if we do a order 5 and then an order 10 smoothing

```
## Warning: Removed 4 rows containing missing values (geom_path).
```

```
## Warning: Removed 9 rows containing missing values (geom_path).
```

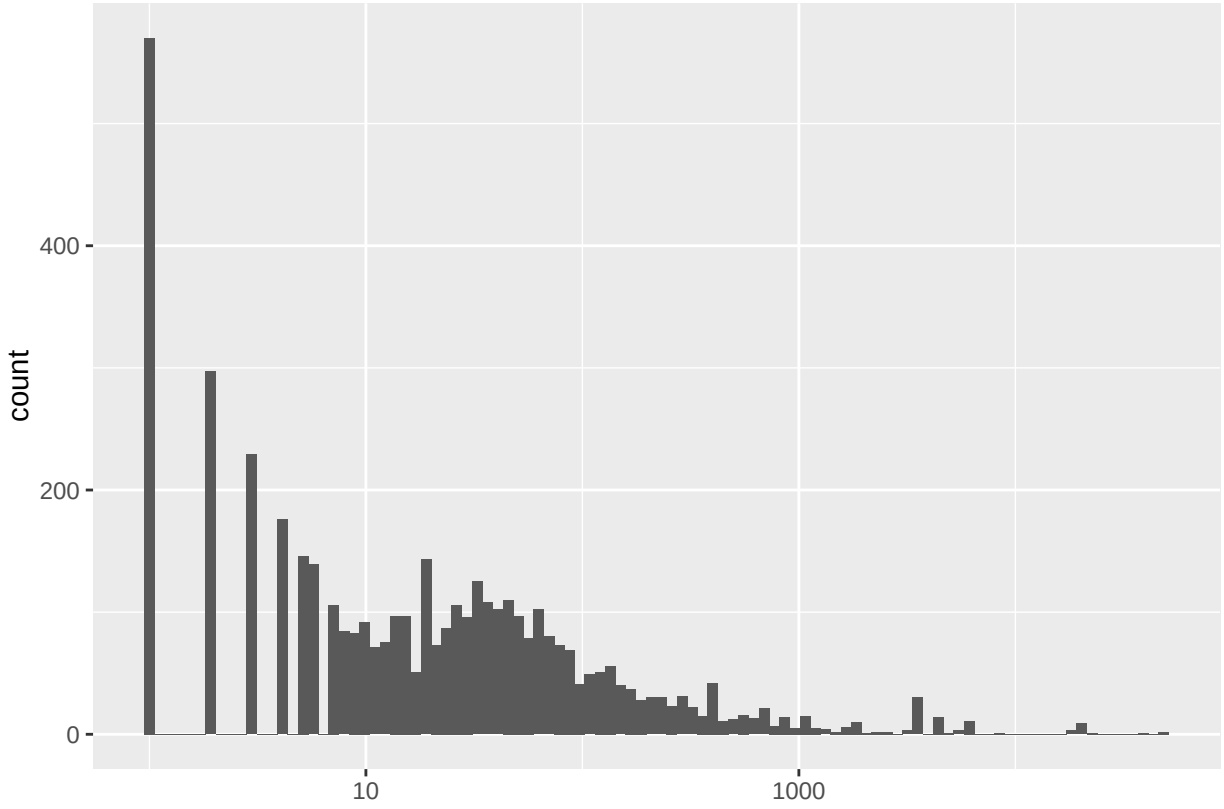


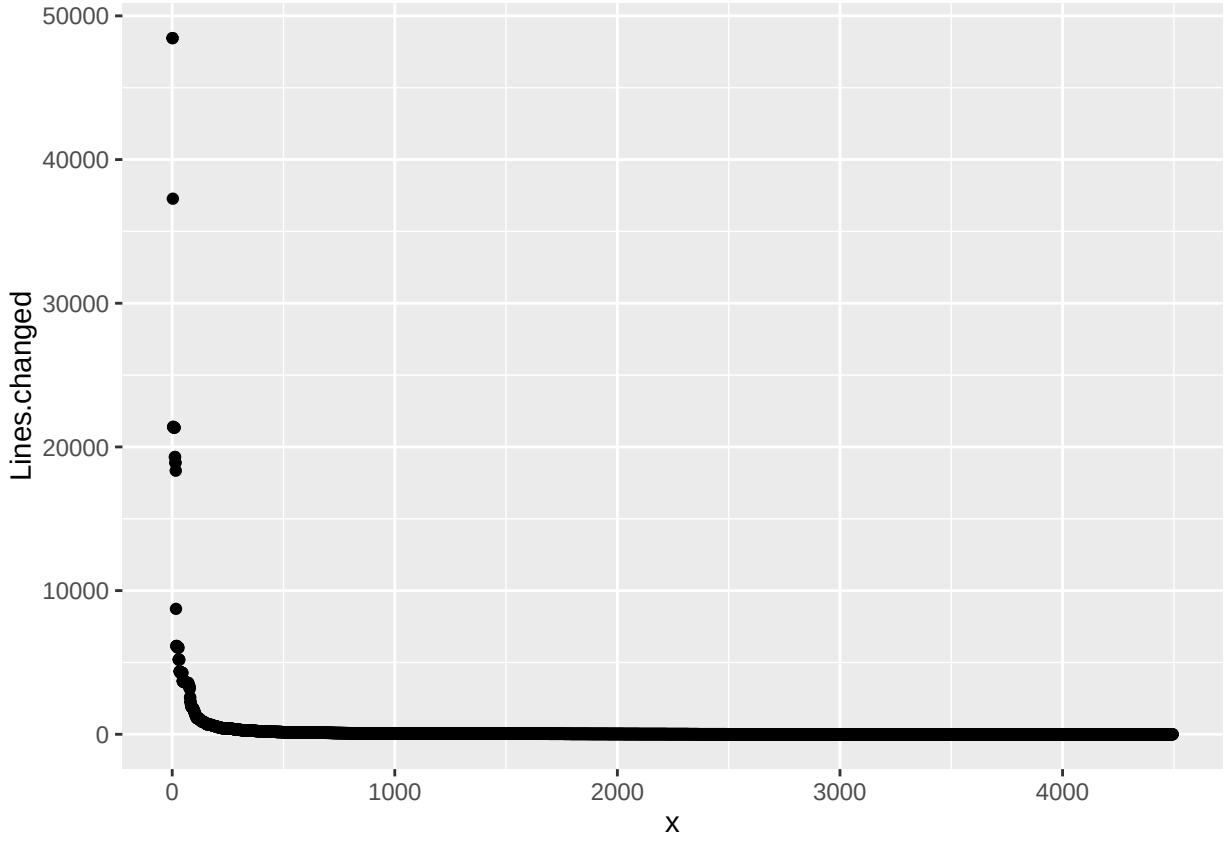
The change in scale might mean that it is distributed using a Pareto distribution. Next we represent the number of changes of a particular size in a log-log scale.

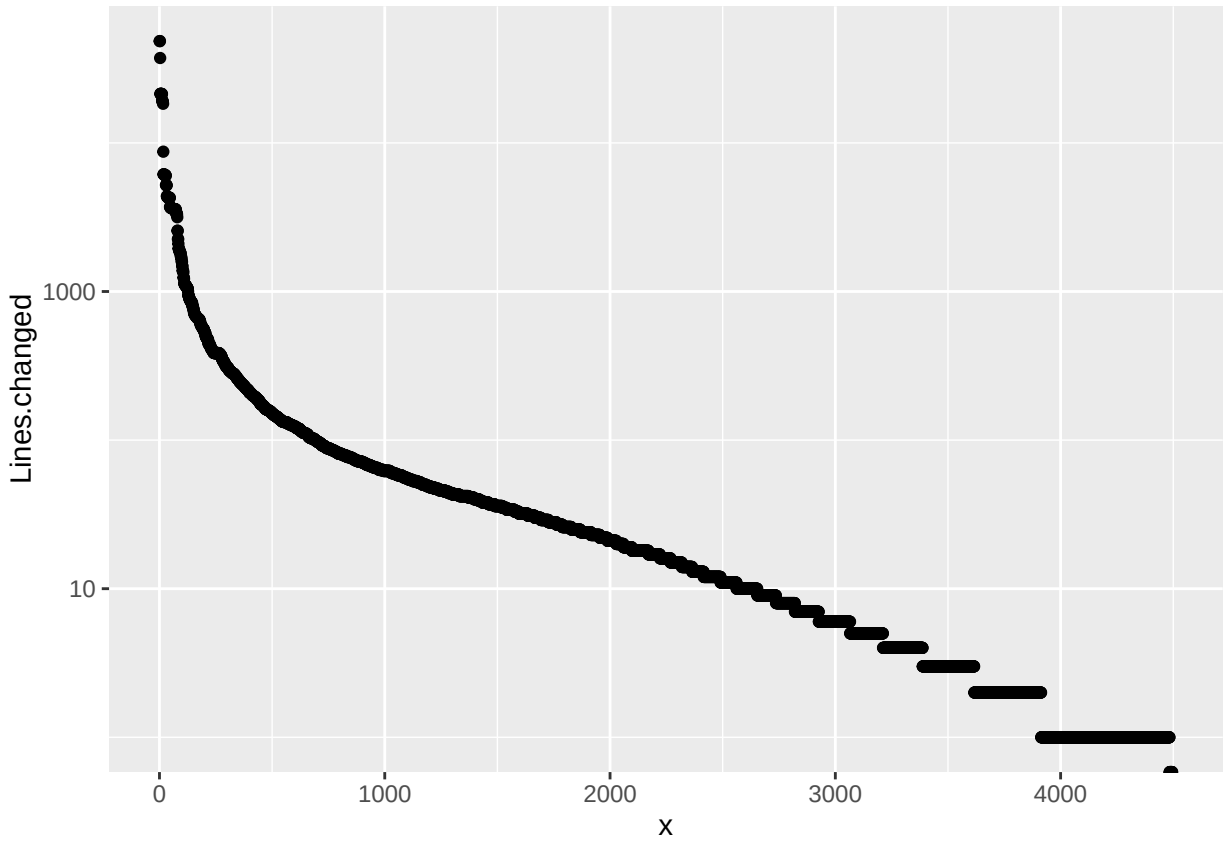


This distribution also appears as a Zipf distribution, with the commit sizes ranked in descending order and plotted with a logarighmic y axis.

```
## Warning: Removed 12 rows containing non-finite values (stat_bin).
```

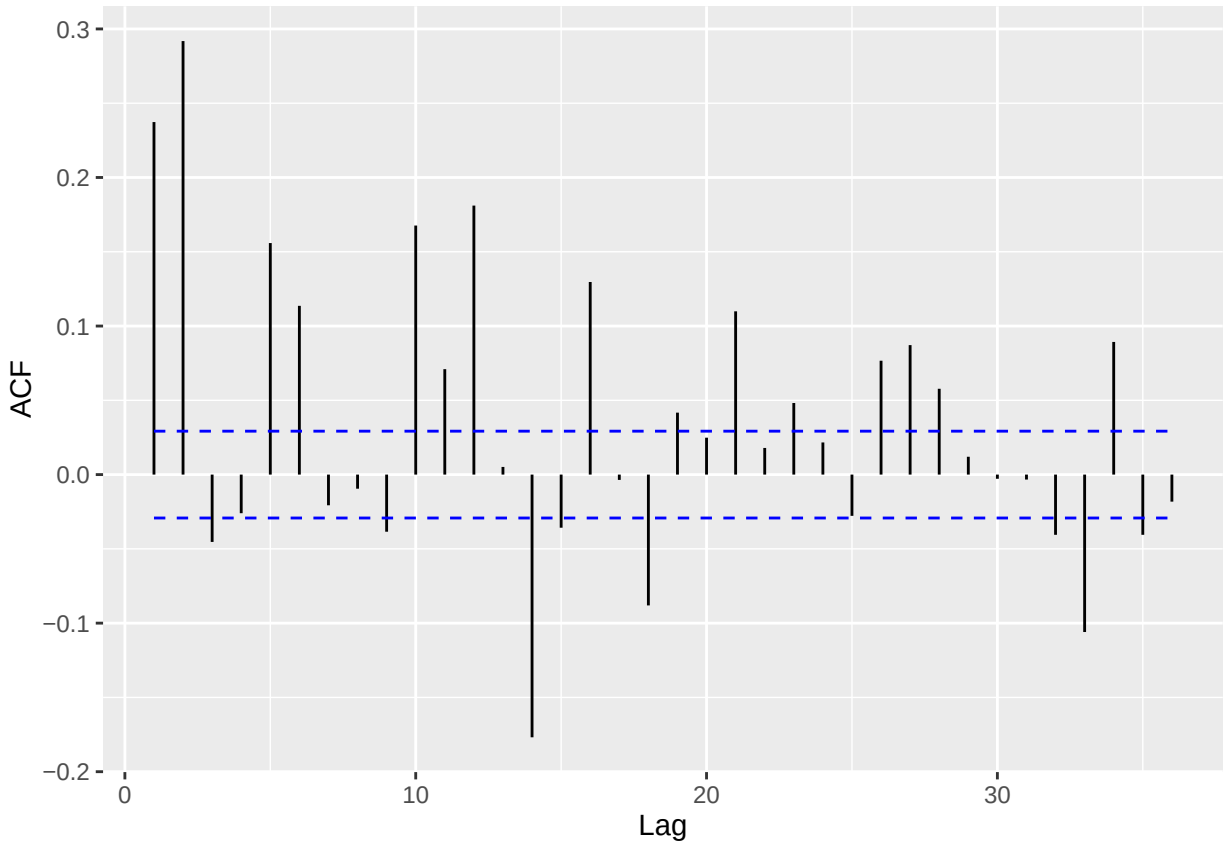

count
400
200
0
10
1000
lines$Lines.changed


Lines.changed
50000
40000
30000
20000
10000
0
0
1000
2000
3000
4000
x

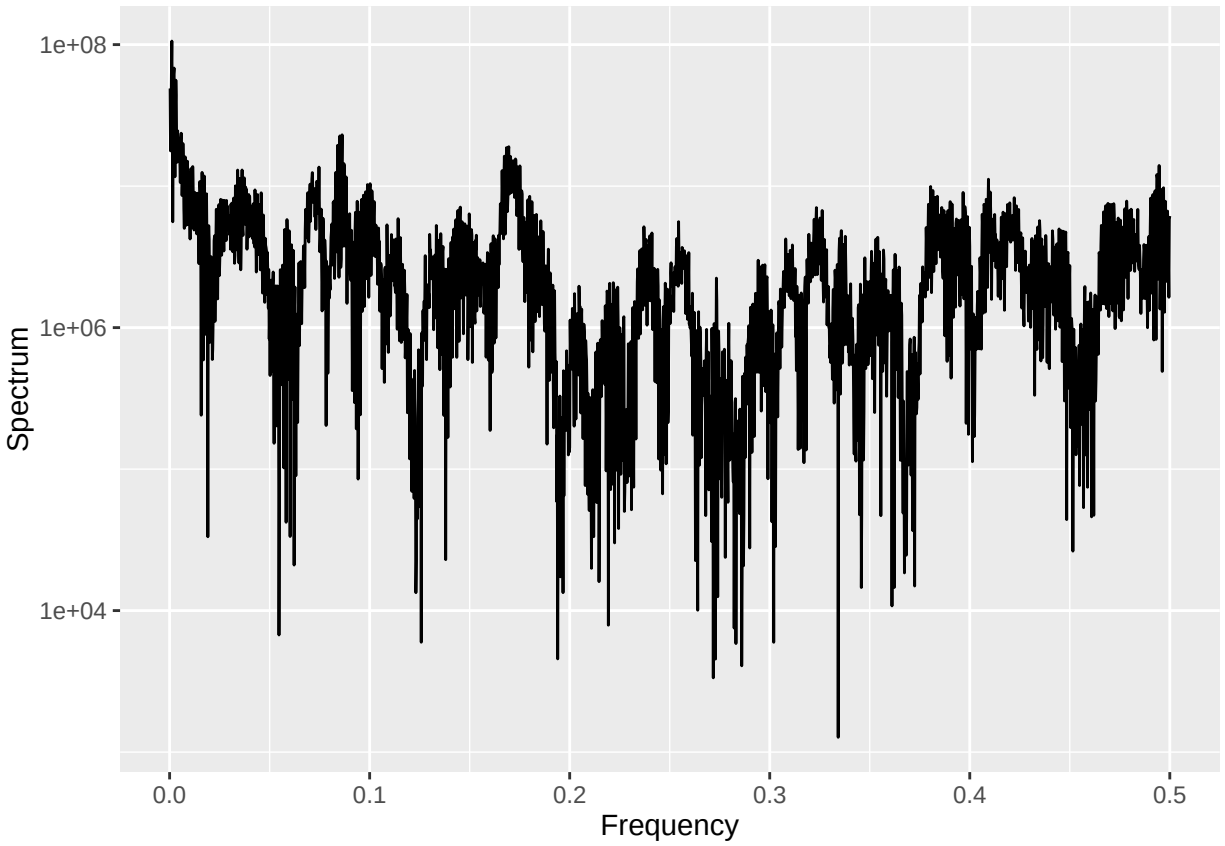


The Zipf exponent for this graph is 6.0282936, -0.0014443

Finally, these scale distributions hints at the possibility of long-scale correlations. The partial autocorrelation is plotted in the next figure.



These possible long range autocorrelations point also to a state of self-organized criticality. One of the characteristics of this state is the presence of *pink* noise, as measured by the power spectral density, shown below



Conclusion

This files has been generated automatically from the data file, so it is difficult to draw a conclusion sight unseen. From what I have seen, there are in many cases long-distance correlations centered around 20 commits of distance, and they show mainly in repositories with a collaborative activity, but they might show up too in some others developed mainly by a single person. You can draw your own conclusions on your own repos by running the Perl script in http://github.com/JJ/literaturame