Growth of Wikipedia Subprojects

These / Frage : Analogie zur Energieübertragung mittels Strahlung

Energie wird in z.B. in Form von Strahlung in ein System übertragen. Die Effizienz der Übertragung hängt von vielen Faktoren ab. Vernachlässigt man Wirkungsquerschnitt, Wellenlänge und Pulsform und betrachtet nur die Menge an Energie, die tatsächlich vom System aufgenommen wurde, dann bleibt dennoch zu unterscheiden, welche Form der inneren Energie erhöht wurde. Verschiedene Prozessbeschreibungen oder Modellvorstellungen helfen dabei, solche Situationen zu erklären. Die Erhöhung der Temperatur ist eine recht einfache Vorstellung, die Anregung von Rotationsmoden eine andere, mit einem komplizierteren Modell verbundene.

Das Ziel dieser Analyse ist es, zu betrachten, ob die Aktivität der Wikipedia Editoren, die sich in Form von Edit Ereignissen zählen lässt, zur messbaren Strukturveränderung des Systems und zum Volumen Wachstum in Beziehung zu setzen ist. Gibt es Phasen, in denen der eine odere Anteil dominiert? Wie kann man solche Phasen erkennen?

1. The Growth process

The random graph model [] is used to create new links between existing nodes with an equal probability for all possible nodes. The model of preferential attachement [] connects new nodes to an existing network, based on the properties of the nodes which are already available. Nodes with some neighbors have a higher chance to get new linked nodes, but nodes can also be added withaout any link. The model for the process of network growth has to cover the creation of new nodes as well as the creation of new links. In the case of wikipedia we can count the number of edit events. But what is going on during such an edit event? New text can be added and the overall amount of data grows (see ….), new nodes are added and new links are added. All edit activity lead to one, two or all of the mentioned results.

Network Growth Process = Link Creation + Node Creation + Content Creation

While the creation of links and the creation of new pages is primary a structural change, the text creation or content creation leads to more information within wikipedia. But based on structural information one can also derive new information. This means, also the creation or reorganization of the network structure leads to more information (citation ….). If a large page is just spilt into smaller but linked liked pages, it is much easier to retrieve information and relations to other nodes in the network can be found automatically. Therfore the context or the meaning of a certain text has to be known. With such information, the wikipedia pages can be used like a semantic network (citation to : DBPedia, …).

1.1 Growth Models

Linear, Algebraic, Exponential Growth

Logistic growth

Gombertz Model and Extended Growth Mode

Geometric

1.2 Network Metrics

Degree distribution

Average path length and Diamater

Average Cluster coefficient

2. Dataset

Here we work with the data set from the German wikipedia, which is available via http://de.wikipedia.org. The rawdata was obtained from the list of all available links and all revisions. For each link the pageid of the source and the destination are stored together with the a timestamp which describes the link creation event. The time resolution is one second. For the analysis the data is aggregated to a time resolution of one hour.

Based on this time series the nr of new created pages ***nN*** and the number of new created links lN per hour is calculated for each selected wikipedia subproject (currently in progress: 60,68,109,222). Fig 1 show the both time series for the German wikipedia with hourly resolution and Fig. 2 shows the same data, but with a binning of 24, which gives the dayly data.

2.1 Data collection

- Structure o our tables

2.2 Data extraction

- Procedure of processing

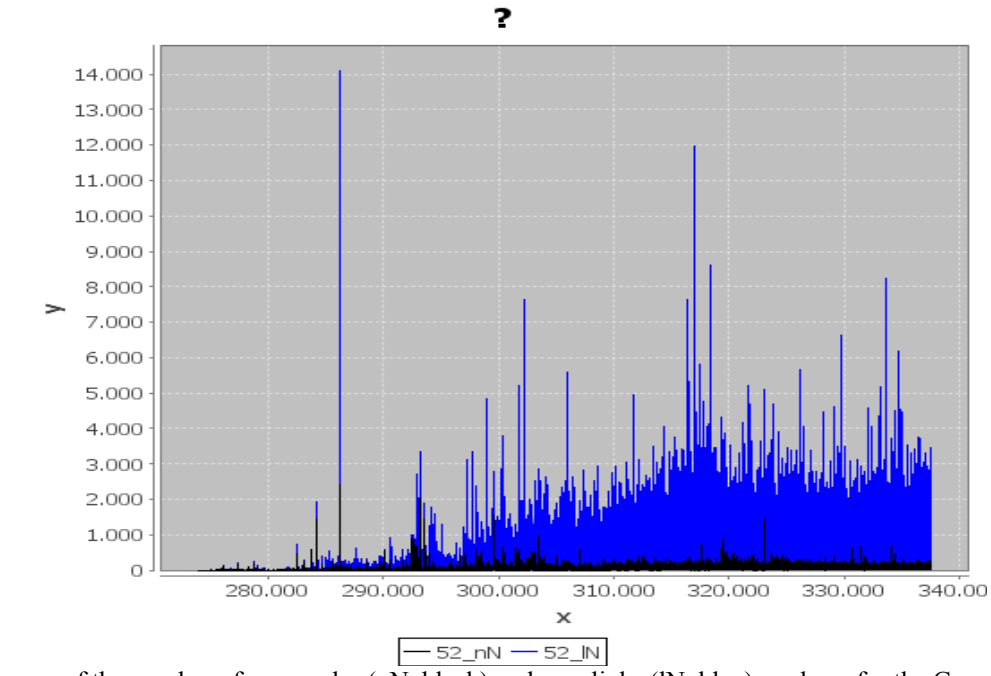


Fig: 1 – Change of the number of new nodes (nN, black) and new links (lN, blue) per hour for the German wikipedia in the time interval t_0= , t_last=

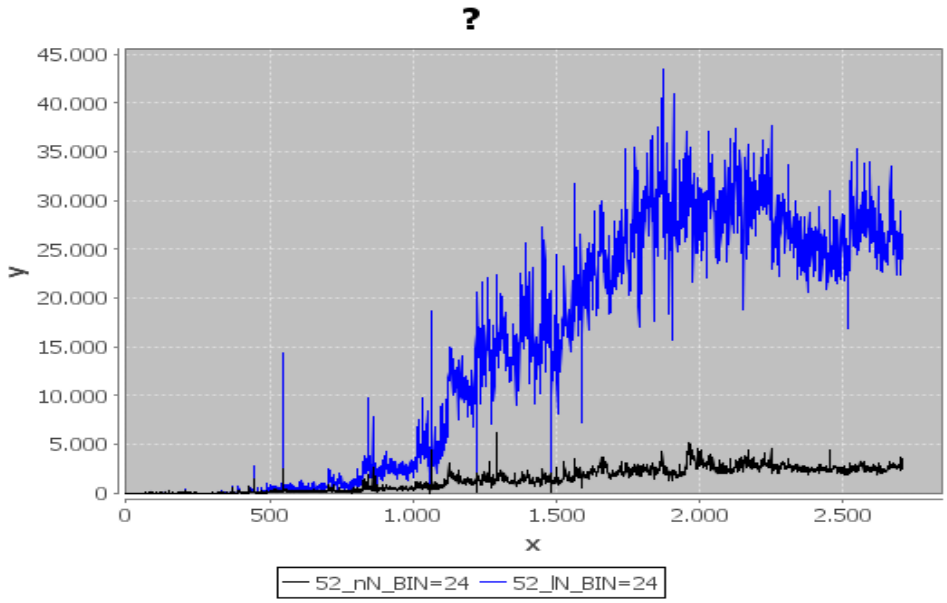


Fig: 2 – Change of the number of new nodes (nN, black) and new links (lN, blue) per day for the German wikipedia in the time interval t_0= , t_last=

The total size of the system ist shown in Fig. 3. For the number of links (blue curve) one can see ranges with different but relatively constant slopes. In Fig 4. the same behaviour is shown also for the number of pages (black curve).

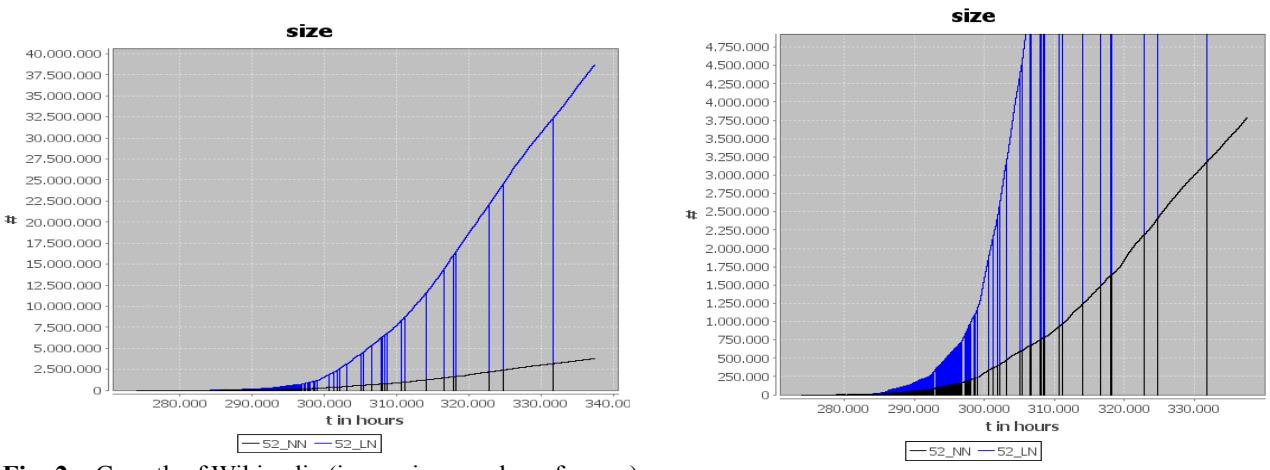


Fig: 2 – Growth of Wikipedia (increasing number of pages)

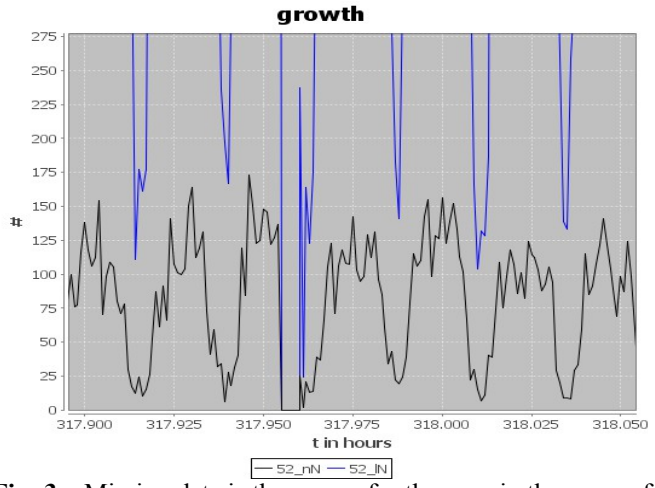


Fig: 3 – Missing data is the reason for the gaps in the curves from figure 2.

The edit activity leads to new information within the wikipedia system. Either new text is added to existing pages or new pages ore just new links are added. What is the ratio of the different type of new information which is created during a periode of time?

We compare the total edit activity (count of edit events per time unit) with the number of new pages, the number of new links and the change in textvolume as a function of time.

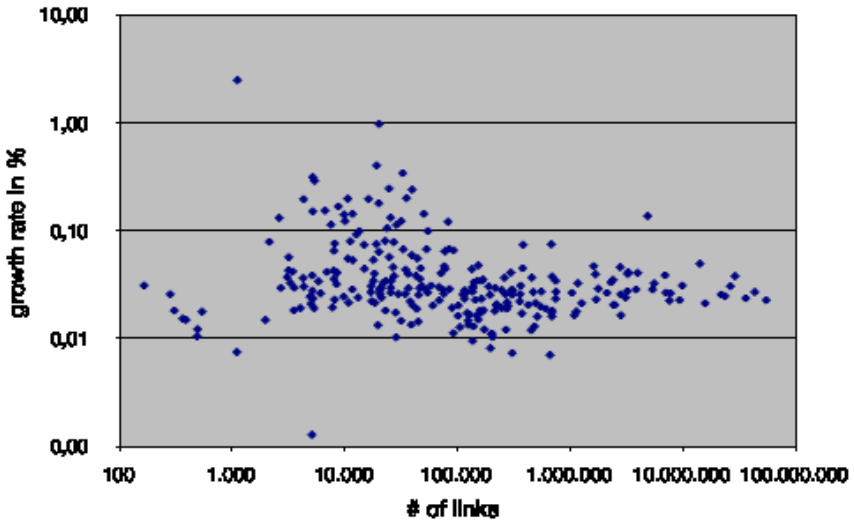
The result are the following time series:

number of events	*nE(t) per hour*
number of nodes	*nN(t) per hour*
number of links	*lN(t) per hour*
change of text volume	*tV(t) per hour and bytes*

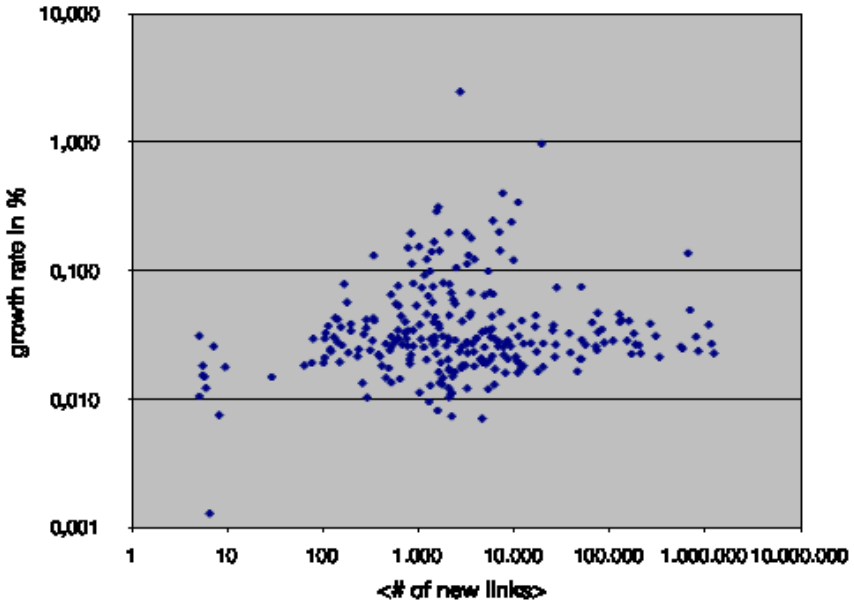
Based on this time series we want to identify times with more structural change and times and with more content changes. In order separate different parts of the overall contirbution.

3. Results

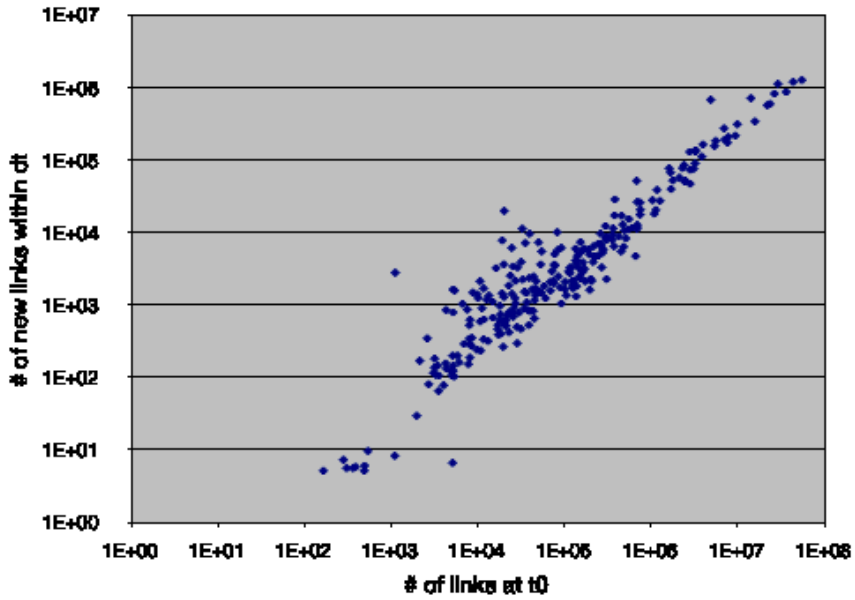
Growth rate as a function of the nr of links:



Growth rate as a function of the average number of new links :



Nr of new links as a function of the number of existing links :

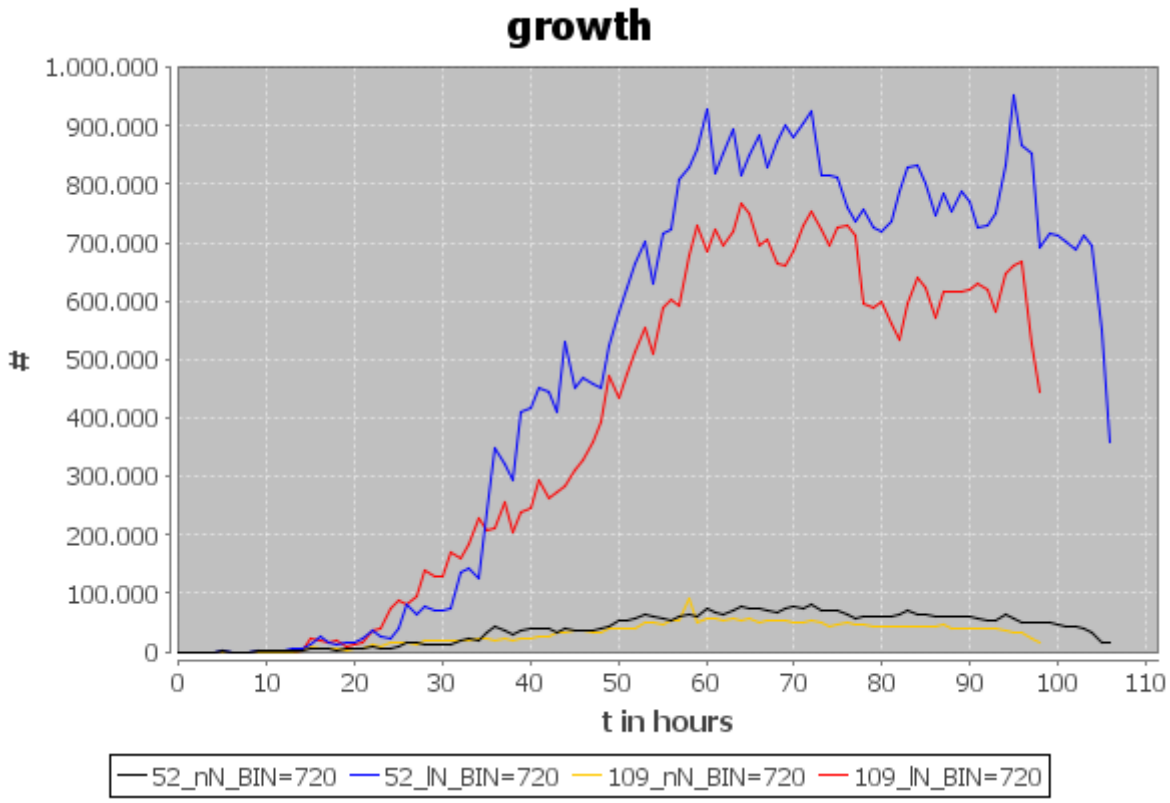


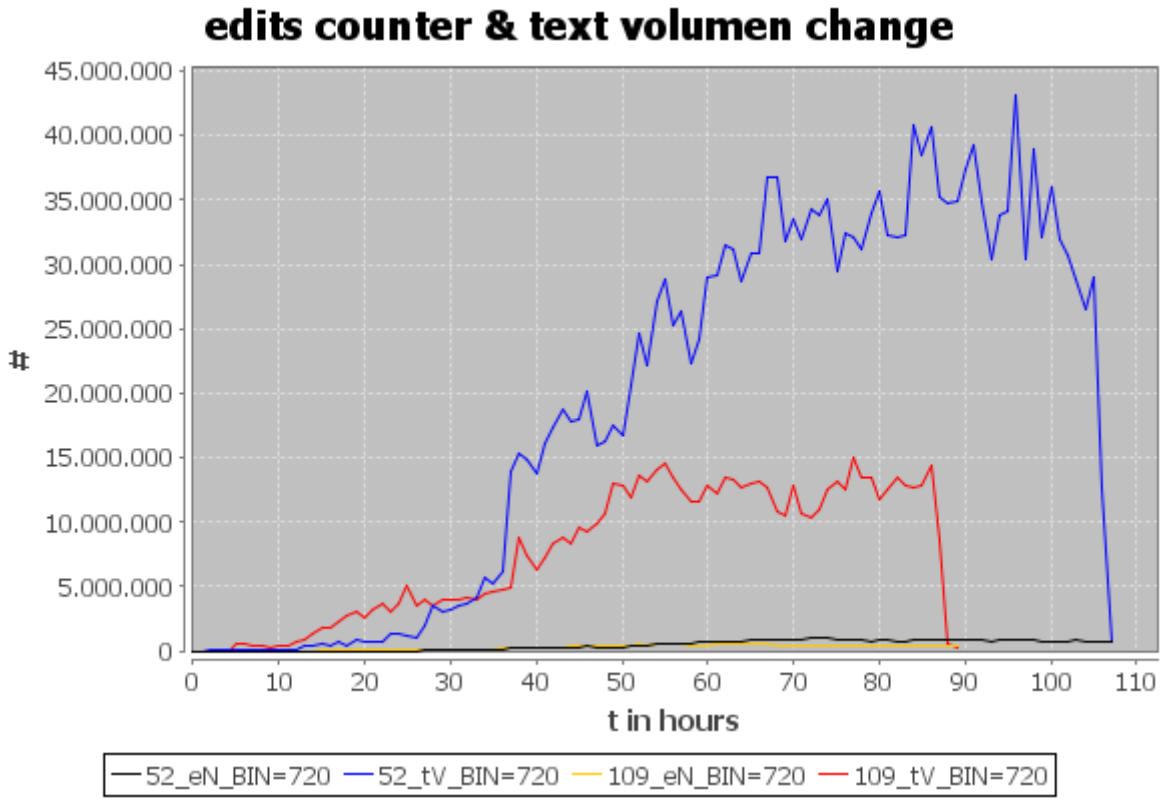
=> We have a linear growth for large Wikipedia projects, but for mid size projects we have a stronger growth.

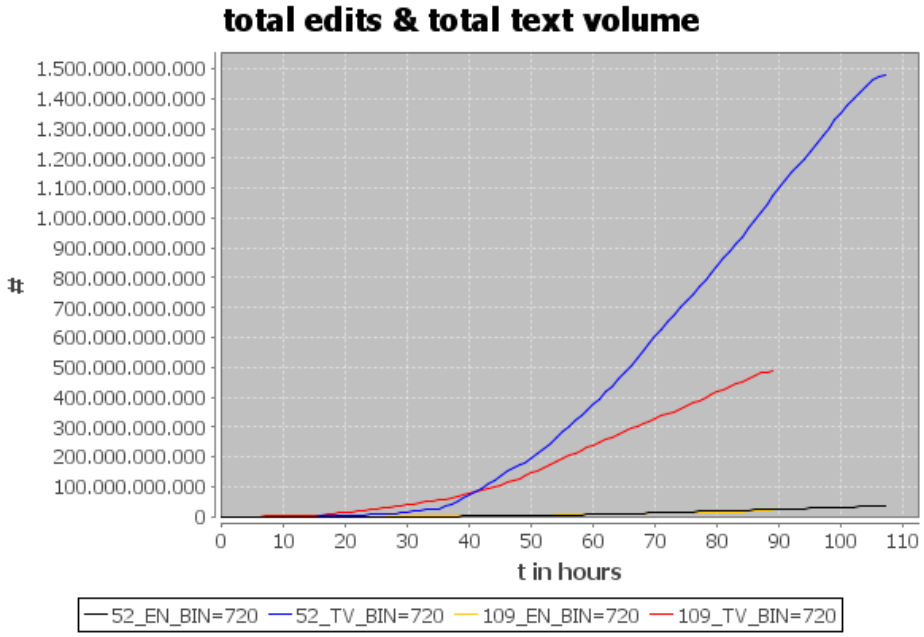
3.1 Information Flow Analysis

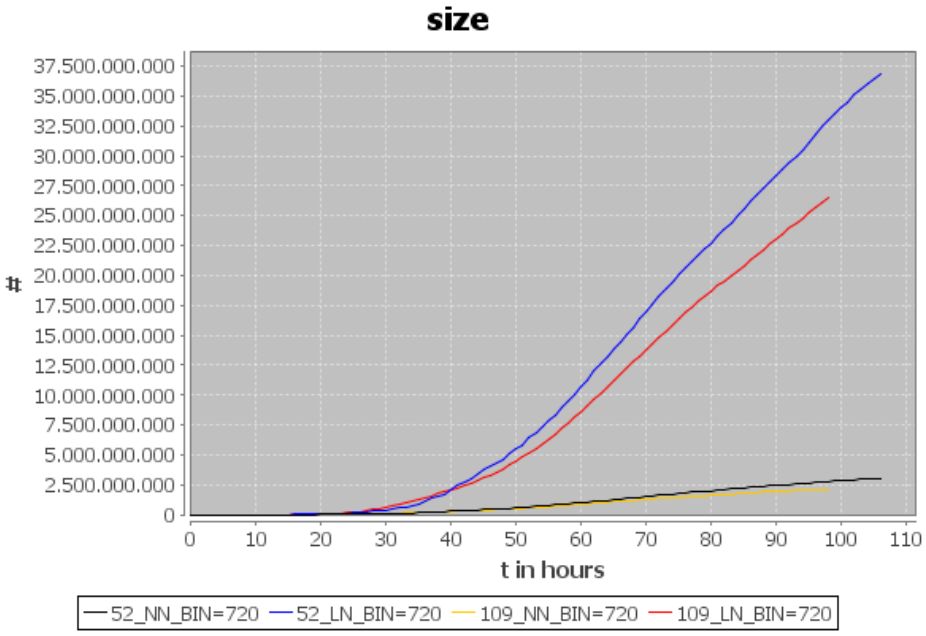
- Welche Anteile der Edit-Aktivität führen zu welchen Änderungen in :

  - Struktur & Inhalt






total edits & total text volume
1.500.000.000.000
1.400.000.000.000
1.300.000.000.000
1.200.000.000.000
1.100.000.000.000
1.000.000.000.000
900.000.000.000
800.000.000.000
700.000.000.000
600.000.000.000
500.000.000.000
400.000.000.000
300.000.000.000
200.000.000.000
100.000.000.000
0
#
0 10 20 30 40 50 60 70 80 90 100 110
t in hours
52_EN_BIN=720
52_TV_BIN=720
109_EN_BIN=720
109_TV_BIN=720


size
37.500.000.000
35.000.000.000
32.500.000.000
30.000.000.000
27.500.000.000
25.000.000.000
22.500.000.000
20.000.000.000
17.500.000.000
15.000.000.000
12.500.000.000
10.000.000.000
7.500.000.000
5.000.000.000
2.500.000.000
0
#
0 10 20 30 40 50 60 70 80 90 100 110
t in hours
52_NN_BIN=720
52_LN_BIN=720
109_NN_BIN=720
109_LN_BIN=720

3.2 Network Structure Analysis

3.2.1 Degree distribution

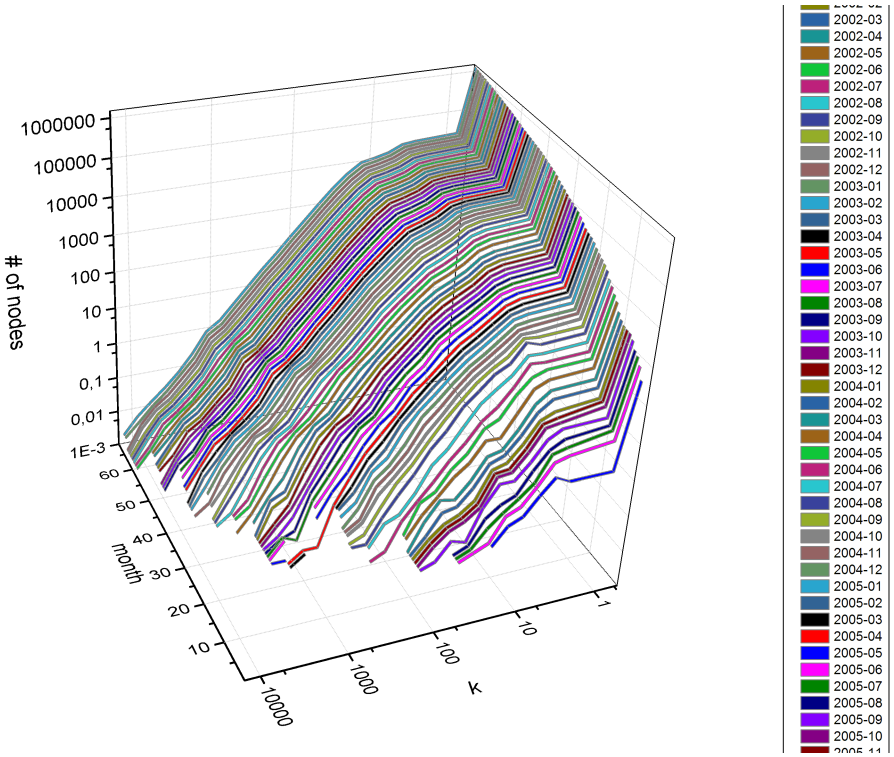


Fig. X : Degree distribution for the German Wikipedia project (langid=52).

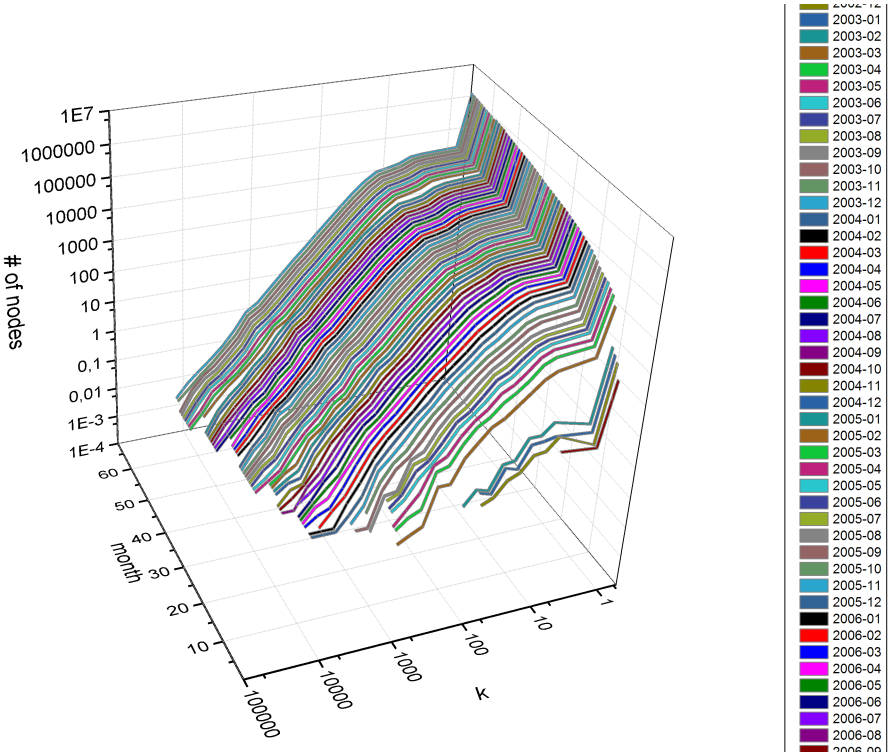
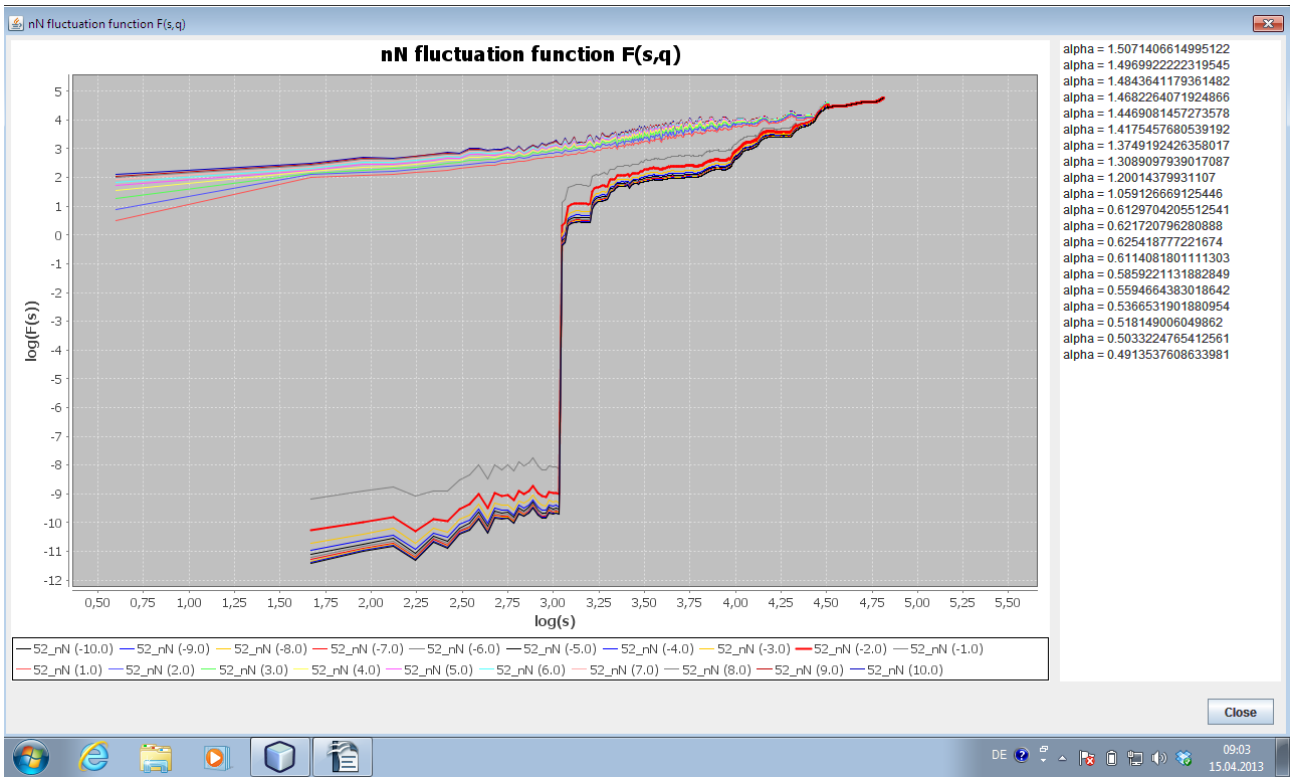


Fig. X : Degree distribution for the Japanese Wikipedia project (langid=109).

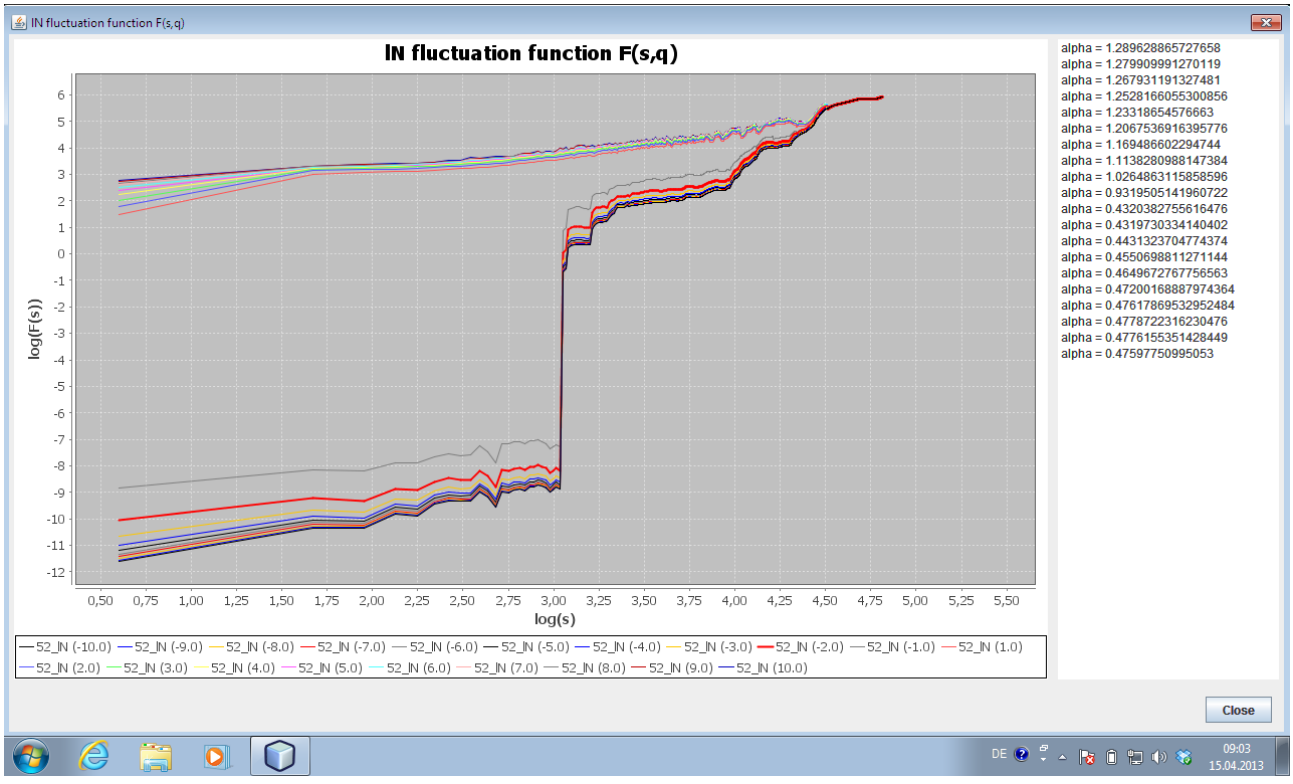
3.2.2 Average Path Length, Diamater, and Average Cluster Coefficient

3.3 Time Series Properties

3.3.1 MFDFA for number of new nodes



3.3.2 MFDFA for number of new links



What measures are usefull?

Check suggestions in section 1 ...

What time series should be used for DFA and MFDFA?

Find ranges in raw data and calc DFA / MFDFA for different ranges, based on degree distribution and activity plots

Should I calculate the attachement probability …
and compare it with the theoretical one in BA-Model?

Conclusion & Outlook

References