Analysis of the non stationary growth process of Wikipedia projects

Abstract

Introduction

Wikipedia is a very well known global information store, which is used by people from different cultures together and within their community engagement they create interconnected sub projects, one for each language. The content evolution of a global system like Wikipedia depends on cultural aspects [???], the existence of the necessary infrastructure to access the system [???], and on the acceptance of the system by its users.

It is quite easy to measure the overall interest in Wikipedia content based on a monthly or quarterly "click-count" data set. Such a data set is provided by the Wikipedia Foundation [...]. The contributions to Wikipedia can be measured based on the "Edit-History". This is an inherent feature of the Mediawiki software [www.mediawiki.org], which drives all Wikipedia projects. Both processes, Wikipedia content creation and Wikipedia usage were studied in our previous work [WIKIPAPER 1]. A comparison of language specific properties, especially related to the Swedish Wikipedia project were analyzed in [final report on WIKI Project 2]. The map in figure 1 shows the percentage of people with internet access per region in color coding. The Wikipedia "edit-activity" is ilustrated in the tree-map. Because many languages are spoken in many regions we can not give an exact mapping between both images and because English is relevant in all, we show this tree map for all non English Wikipedia articles.

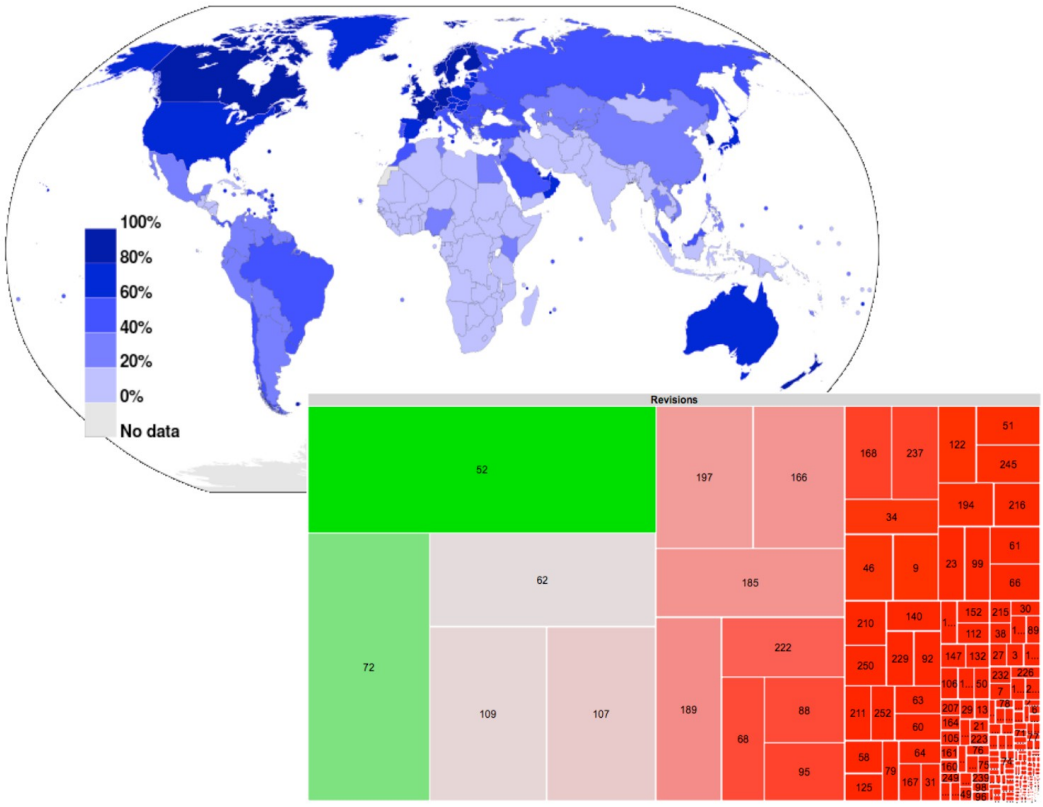


Fig. 1 : ...

Usage and Creation of content are two coexisting processes. Content changes just if users decide to create new pages, or if a user adds new text, images or even links to existing pages. Edit activity leads to new information, stored in the system. But how are edit-activity and the growth process related to each other?

Energie wird z.B. in Form von Strahlung von einem System auf ein anderes 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 Ziel-System aufgenommen wurde, dann lässt sich unterscheiden, welche Form der inneren Energie erhöht wurde. Verschiedene Prozessbeschreibungen oder Modellvorstellungen helfen dabei, solche Prozesse und Wechselwirkungen zu erklären.

Die Erhöhung der Temperatur ist eine recht einfache Vorstellung, die Anregung von Rotationsmoden eine weitere , mit einem komplizierteren Modell verbundene.

An Stelle von Energie wird in Wikipedia Information gespeichert. Die eingebrachte Menge an Information verteilt sich intern auf mehrere Kanäle (Erstellung neuer Seiten, verlinkung von Seiten, Erweiterung bestehender Seiteninhalte). Die Analyse zeigt die Grenzen der einfachen Wachstumsmodelle. Das Wachstum des sozialen Netzwerks der Wikipedia Nutzer und das Wachstums des “Inhalts-Netzwerks” stellt eine Kopplung zweier dynamischer Netzwerke dar. Wir beschreiben daher ein einfaches Informationsfluss-Modell, welches die Kopplung der beiden Netzwerke beschreibt. While user-activity is the source of new information which will be stored in the content network the distribution of this new information within this network is characterized by different quantities.

The question is: Can we see a “life-cycle” of a wikipedia project like it is shown by […] for single wikipedia pages? How does the growht process change within this life-cycle phases?

In section one, we start with a short review on existing studies on Wikipedia growth.

(… summary of existing studies …)

(… select properties for comparison ...)

In section two we describe the analysis procedure, raw dataset, and intermediate data which are used for model comparison. A discussion of our results is given in section three.

1. The Growth Process of a Complex Networkems

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. The Wikipedia Growth process can be described by the following equation:

$$\text{Stored Information} = \#\text{of Nodes} + \#\text{of Links} + \text{Content}(\text{ Nodes })$$

While the creation of links and the creation of new pages is a structural change which lead to a semantic structure, also the text or content creation process leads to a more obvious increase of information within wikipedia. But based on structural information one can also derive new information. This means, that also the creation or reorganization of the network structure leads to more information. If a large page is just split down into smaller but linked pages, it is much easier to retrieve information. Relations from such pages to other nodes in the network can be found automatically based on context analysis. Therefore the context or the meaning of a certain text, which represents the page, 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

Link Density

Diamater

2. Dataset & Preprocessing

In this study we work with a data set which is available as a MySQL database dump. A table containing all revisions (size) and a table of records which represent the events of creation of links within wikipedia pages (size) are loaded into the distributed file system of a small Hadoop cluster to allow efficient filter and join queries. This intermediate tables are filtered by the language id (see [ref id-list] for a list of all languages and related ids). From the result set, which is a CSV file containing an ordered list of event records at a time resolution of milliseconds a first set of time series is generated. For selected Wikipedia projects, the number of edit events per hour is counted as $nE_{lang}(t)$. Each time a link is created a new page is created if one of the linked pages not already exists. This events are counted as the number of new pages $nN_{lang}(t)$ per hour but all link creation events are counted as $nL_{lang}(t)$ per hour. The change in text-volume is measured as $tV_{lang}(t)$ per hour and bytes. Figure 1 shows the number of new nodes (black) and the number of new links per hour for the German wikipedia a) per hour and b) per day.

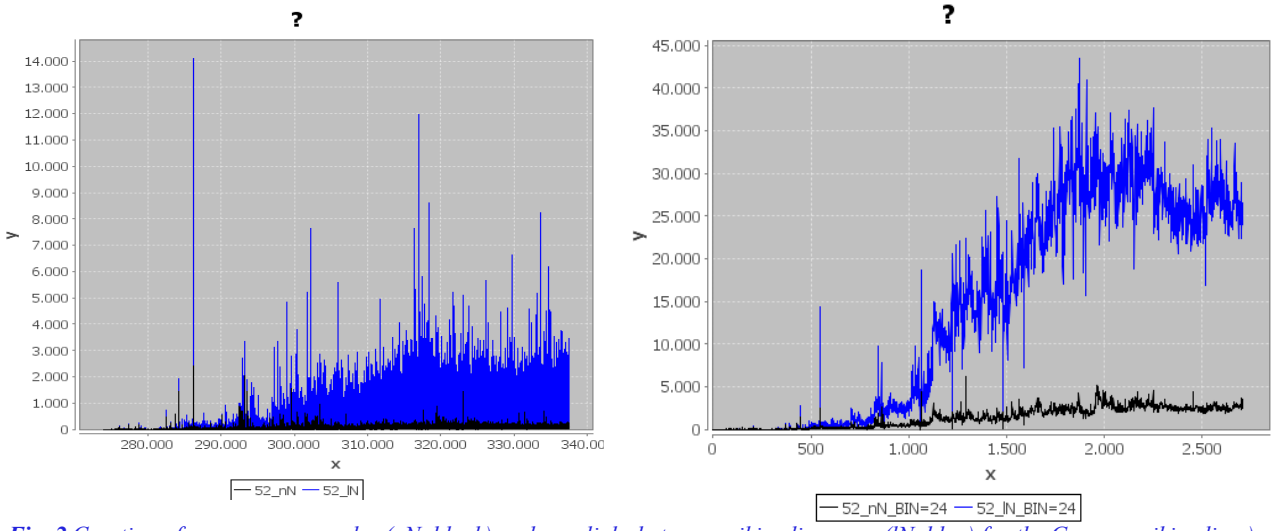


Fig. 2 *Creation of new pages or nodes (nN, black) and new links between wikipedia pages (lN, blue) for the German wikipedia a) per hour, and b) per day.*

The total size of the network is shown in figure 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).

TODO : Illustrate the ranges of different slopes and give the value of the slope.

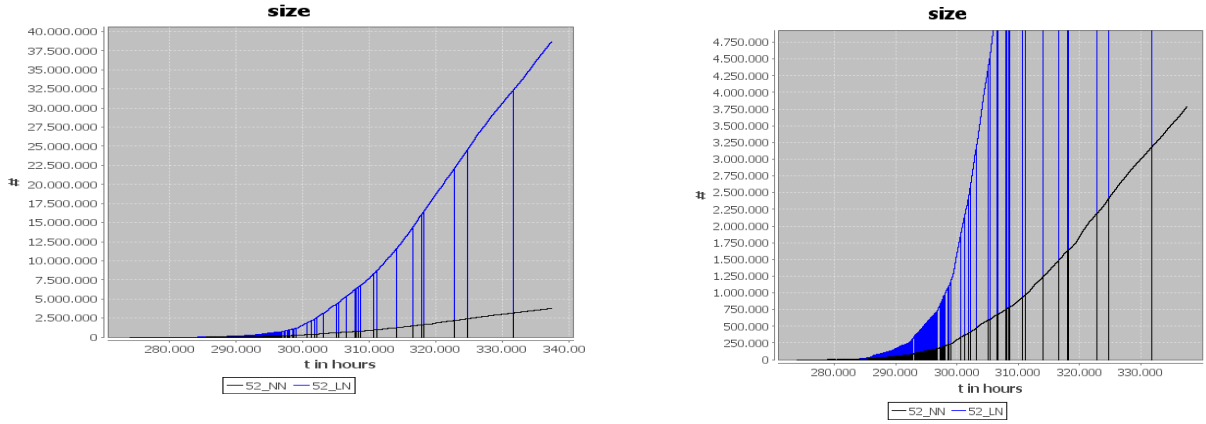


Fig. 3 *– Growth of Wikipedia, in the classical sense, is given by an increasing number of pages (black curve). But not just the number of pages, also the number of links and the total text-volume are indicators for a growing amount of information in the wikipedia conten-network.*

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. The following analysis is based on a set of eight time series for each of eight selected languages. Table 1 shows some properties of this wikipedia sub projects.

lang					
52	de				
60	en				
68	fa				
88	fi				
109	ja				
122	ko				
166	nl				
222	sv				

Tab. 1 – Properties of selected Wikipedia projects.

3. Results

3.1 Growth of Wikipedia in 2009

We analyze the growth of all wikipedia projects in 2009 between January and September in the time interval from t_0 = 01-01-2009 to t_1=20-09-2009. Therefore we count all existing links z_0 at t_0 (label ALL), the number of new links per month g_{a_i} (for i=1 … 9 as JAN, FEB …). The average absolute growth rate is $\langle g_{a_i} \rangle$ (label MW) and the relative average growth rate $\langle g_{r_i} \rangle$ is given by the quotient $\langle g_{a_i} \rangle$ / z_0.

	ALL	JAN	FEB	MAR	APR	MAI	JUN	JUL	AUG	SEP	MW	Ratio
52	55.210.823	1.159.352	1.361.695	1.429.308	1.120.953	1.433.584	1.197.250	1.180.016	1.169.041	1.134.972	1.242.908	2,25 %
60	9.935.826	330.514	310.774	314.870	272.869	360.211	285.067	273.355	309.287	292.290	305.471	3,07 %
68	7.656.830	179.061	153.357	198.586	169.917	185.382	165.593	167.285	168.682	149.033	170.766	2,23 %
88	7.802.475	201.602	168.931	183.030	271.871	201.137	198.591	177.637	234.001	198.692	203.944	2,61 %
109	36.320.830	897.484	763.418	937.098	807.082	796.760	833.611	820.327	902.618	913.101	852.389	2,35 %
122	3.288.424	133.684	127.519	129.838	125.311	139.202	133.911	128.148	149.569	142.532	134.413	4,09 %
166	15.848.923	343.318	304.500	360.274	340.670	340.838	327.896	301.382	357.277	328.539	333.855	2,11 %
222	9.422.728	218.844	202.939	233.377	203.114	236.762	197.575	188.800	212.143	219.116	212.519	2,26 %

Tab. 2 – Growth rates of selected Wikipedia projects.

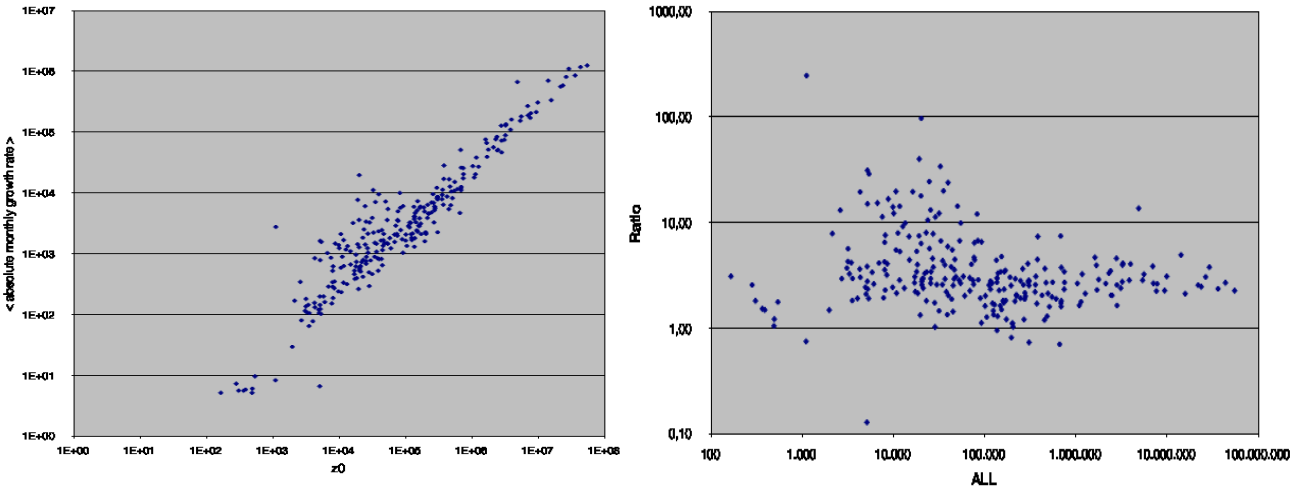
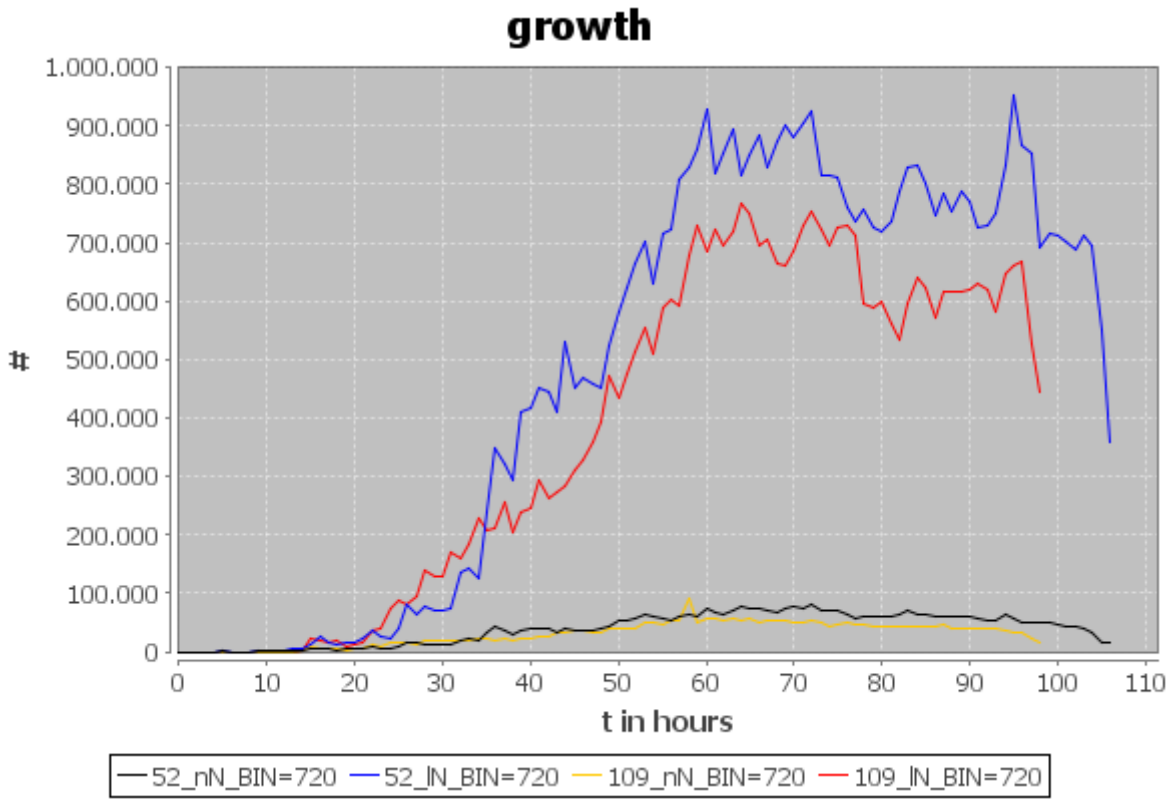


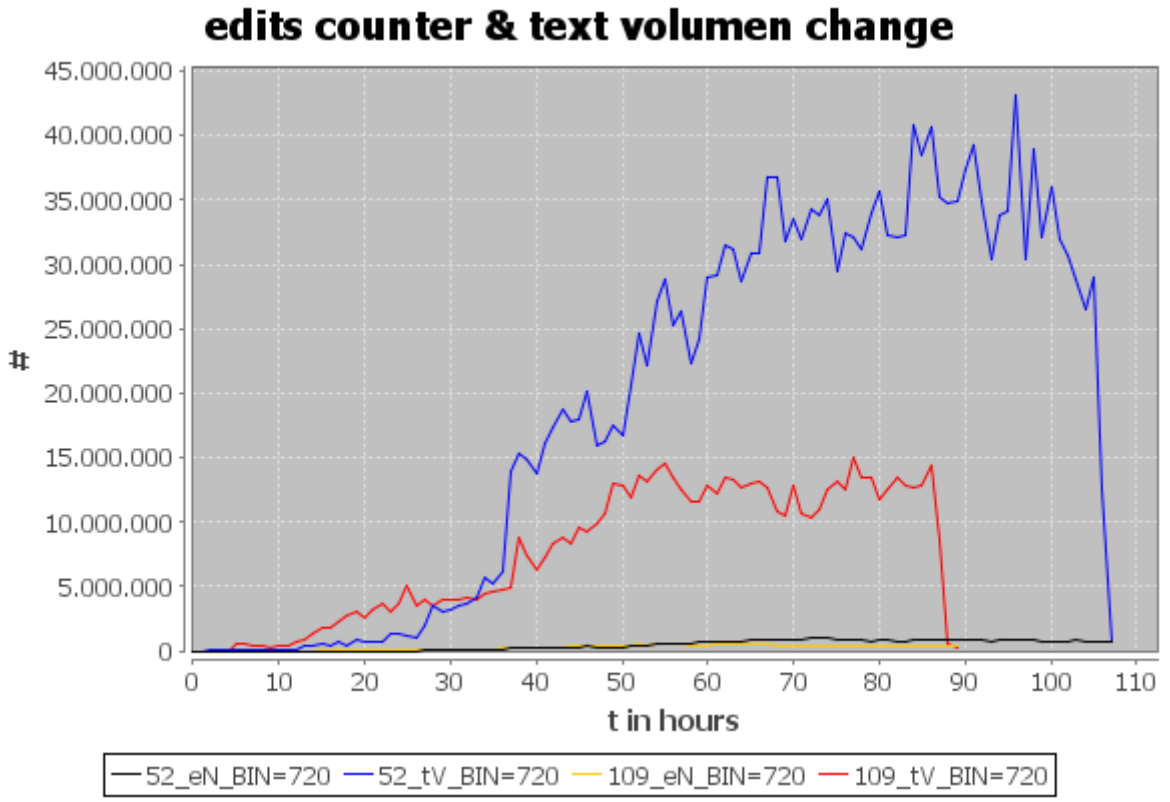
Fig. 4 – Absolute growth rate and relative growth rate as a function of the initial number of links in wikipedia projects of different languages.The left panel shows the linear dependency between network size and growth rate for the majority of wikipedia projects. Projects with a size between 1000 and 100.000 links grow more than ten times faster than large wikipedia projects.

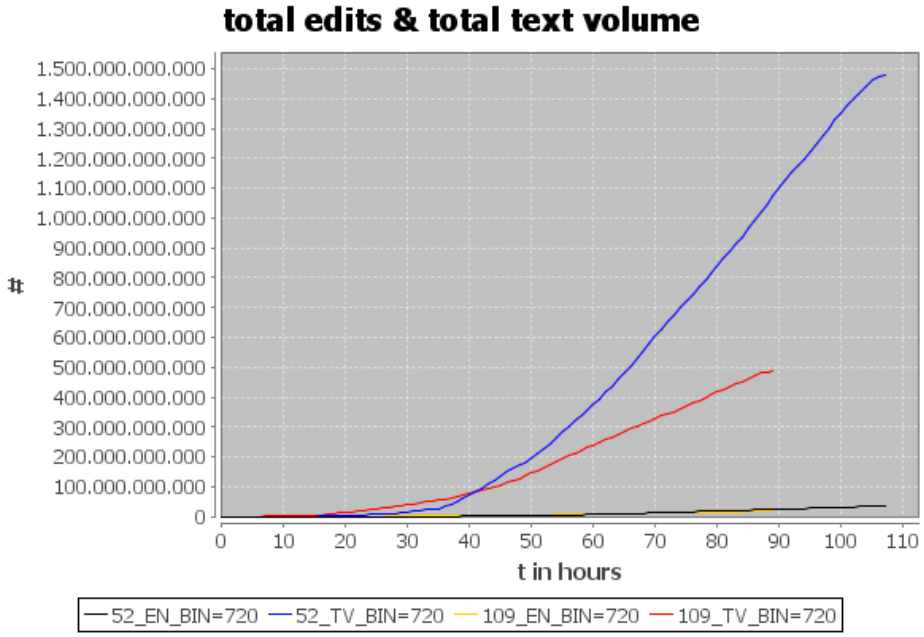
3.1 Information Flow Analysis

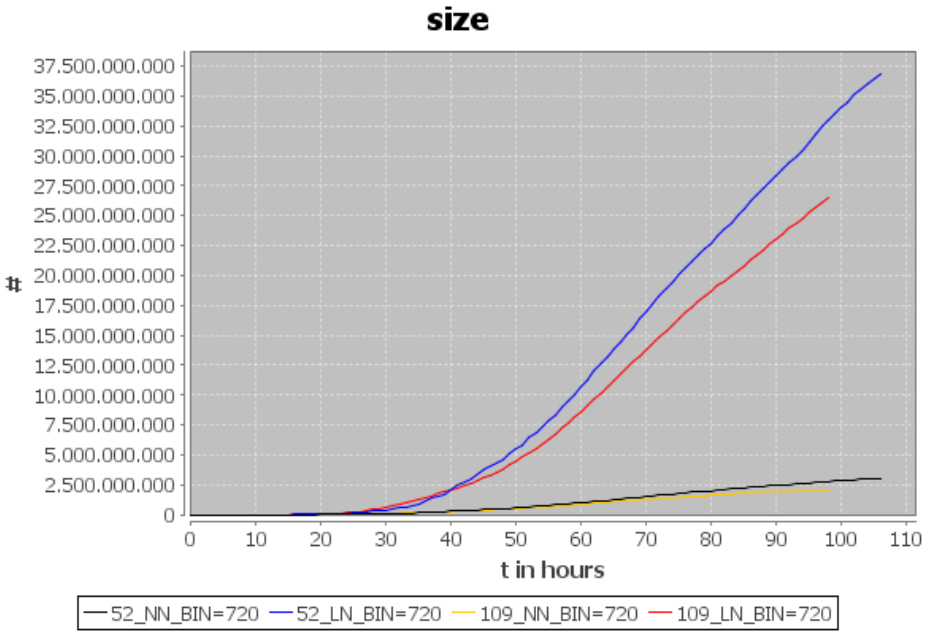
- Welche Anteile der Edit-Aktivität führen zu welchen Änderungen in :

  - 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 Analysis

3.2.1 Degree distribution

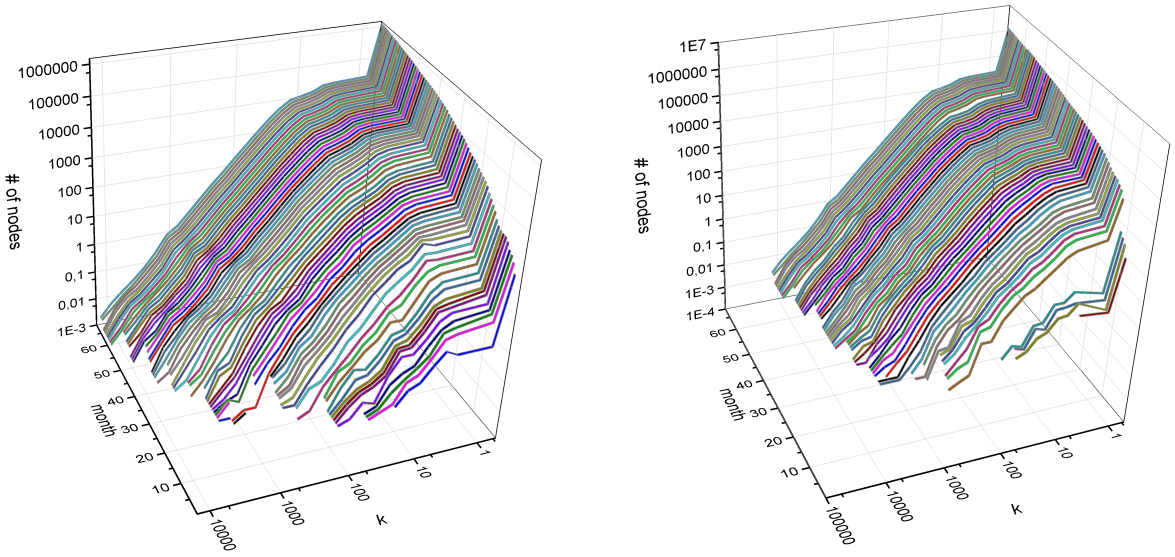


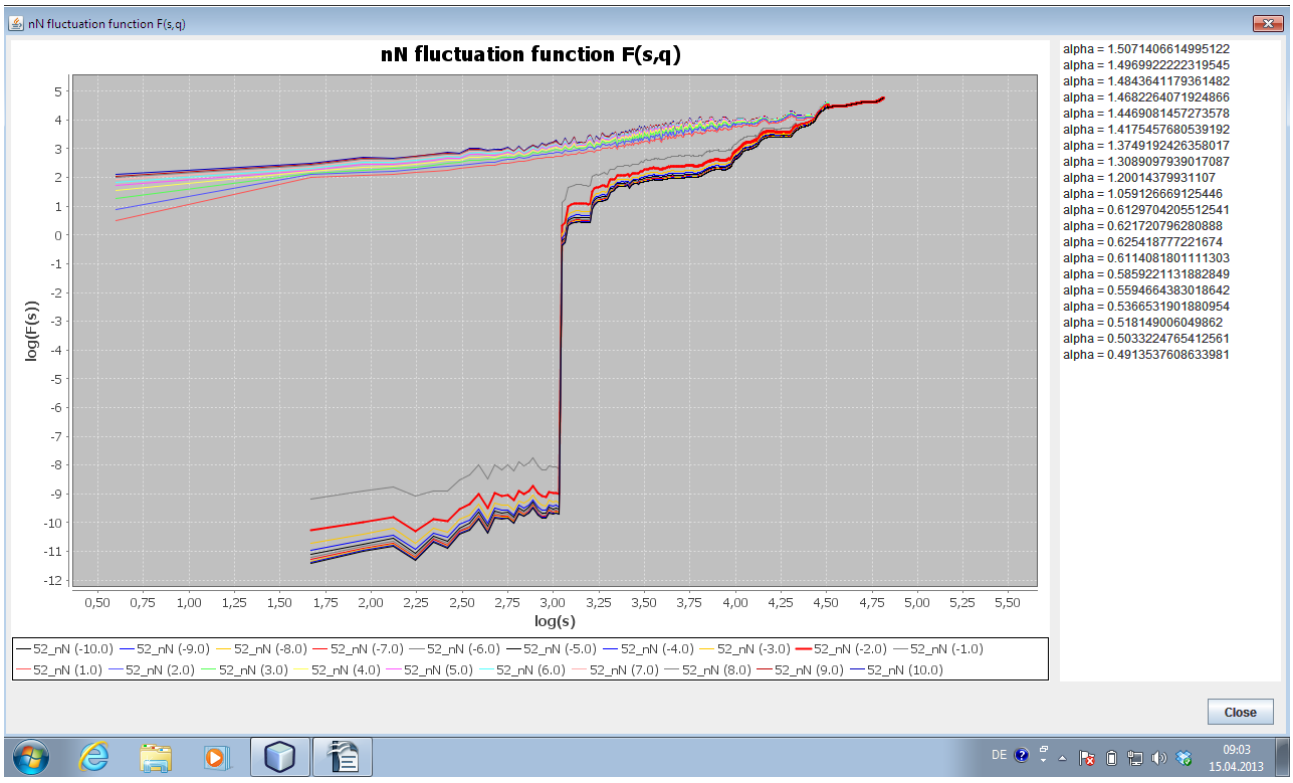
Fig. X : a) Histogram of node degree for German Wikipedia project (langid=52), b) histogram of node degrees for the Japanese Wikipedia project (langid=109).

Fig. X : Degree-distribution of for a) German Wikipedia project (langid=52), b) Japanese Wikipedia project (langid=109).

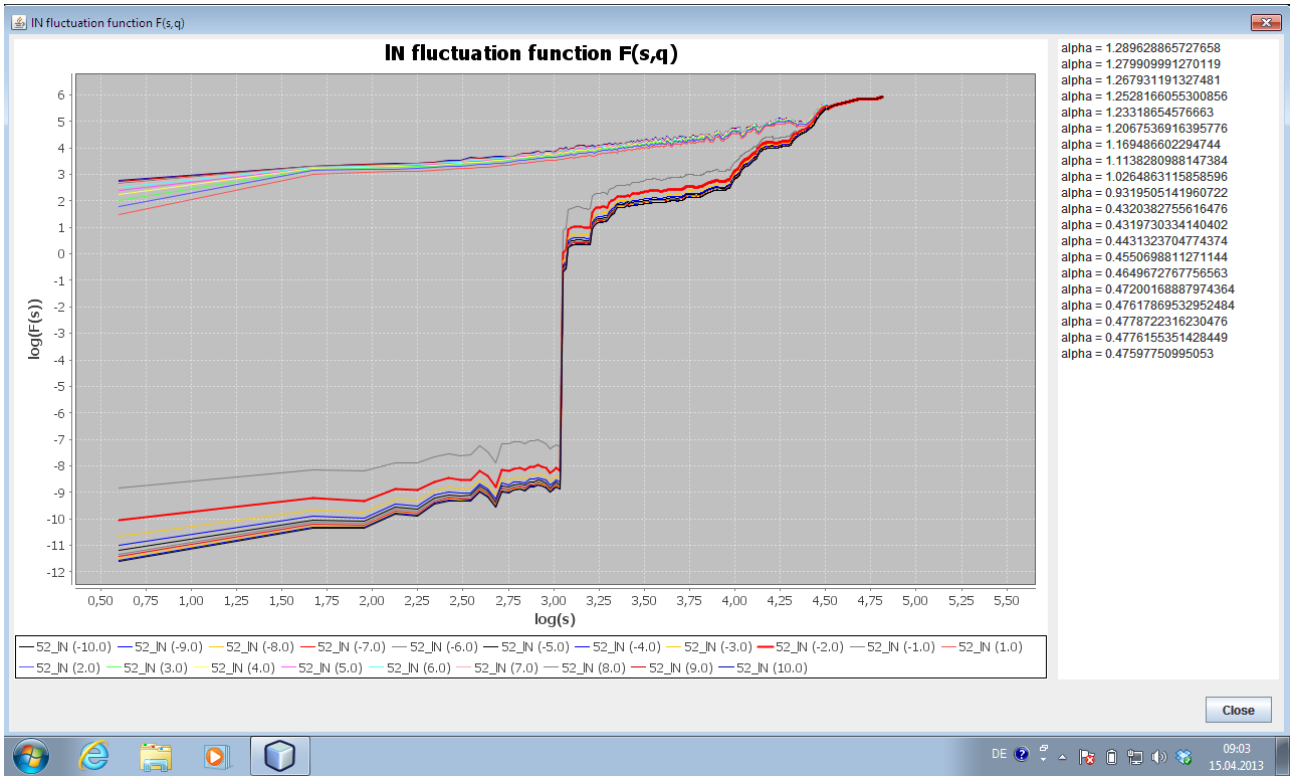
3.2.2 Link Density and Diamater

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