# 10. Life Cycle of Social Networks

> Look deep into nature, and then you will understand everything better.
>
> *(Albert Einstein)*

Different Wikipedia projects grow differently. This is not surprising, because they are maintained by different communities. Thus they are influenced by different economical, political and cultural conditions. Because each Wikipedia sub project is created in a different language, one can say, that each Wikipedia project represents a different cultural context wherein different topics are important. This view is based on differences in how languages are used and how different cultural aspects are reflected within community driven large projects.

A second perspective exist. On a higher abstraction level, all Wikipedias can be unified, by ignoring the cultural and lingual differences. In this case one can say: Wikipedia is the global encyclopedia. It is a crowd based information storage and knowledge creation system. A growing system with inherent memory. For multiple languages, there exist several Wikipedia instances. All may have comparable properties. If a reasonable approach of normalization of the data exists, one could compare the project life cycles or phases.

To verify this hypothesis we analyze the growth procedure of four Wikipedias. We selected English, Swedish, Dutch, and Hebrew Wikipedia projects. A preliminary report was published by Schreck *et al.* [**?** ].

## 10.1. Cultural Aspects of Global Online Networks

The detailed impact of different cultures on Wikipedia content and on the processes, such as content creation and consumption are not in the scope of this work. In our data driven studies we can follow a natural path and differentiate by language. This is easy, because all Wikipedia projects for all languages coexist and they are interrelated already. Each language can define one dimension if a global analysis is desired. An advanced approach uses more specific dimensions. They were derived by Hofsted *et al.* [**?** ] from global survey data. As they show, it is possible to apply factor analysis to determine the predominant cultural dimensions. Hoftsted *et al.* defined four cultural dimensions regarding fundamental anthropological problem fields. The dimensions are named: power distance (PDI), individualism (IDV), uncertainty avoidance (UAI) and masculinity (MAS). Long-term orientation (LTO) and Indulgence versus restraint (IVR) were added later. Although this approach is data driven it is not applicable to a global system like Wikipedia. Wikipedia covers many different topics. A clear segregation between cultures is not possible. Many people speak multiple languages. Even if they have a different cultural background, they may contribute to the same Wikipedia project or to a different one, depending on their current background or

context. Culture is one context, but obviously not the only one which influences the representation of topics within Wikipedia (see figure 13.4.2 for an illustration of the impact of the lingual context on a topic's representation in different languages). Therefore, Wikipedia seems to be a good source for advanced studies on cultural differences in knowledge formation and knowledge sharing which is related to cultural contexts as well.

## 10.2. Growth of Wikipedia Projects

The Wikipedia projects are more than just networks of pages, to which a new page is added at a given time. Pages provide information, they innervate new ideas, lead to questions, and as a consequence, new pages are added by different people. The editorial process can be highly controversial as Yasseri *et al.* [28] and Eckstrand *et al.* [**?** ] show. The system is embedded within user communities which consist of editors and readers . Not all people contribute to Wikipedia, but a critical mass of users seams to be required by a Wikipedia project in order to survive. The evolution of Wikipedia project sizes was already analyzed by Ortega *et al.* [**?** ]. They found that the contributions to Wikipedia are dominantly made by several so called "power users". Based on a calculation of the Gini coefficients for the top ten Wikipedias they state that approximately 90% of all users are responsible for less than 10% of the content. A comparable distribution of user activities in several other wikis - non of them are Wikipedia projects - was found by Stuckman and Purtilo [**?** ]. Such a strong bias towards some very active users has to be taken into account. Analysis on the life cycle should not just be build on the editorial activity as presented by Gorgeon and Swanson [**?** ]. They studied the evolution of the topic or concept "Web 2.0" in Wikipedia based on article size, number of editorial actions and number of contributors. As a result, they define four phases for an article: Seeding, Germination, Growth, Maturity (for details see section 5 in [**?** ]). The life cycle phases already take the activity and controversial character of editorial events into account. One can clearly conclude, that editorial activity does not always lead to an increase of content, because higher quality can be achieved by clear statements which are often the result of shorter sentences. Too long articles are sometime misleading or distracting. Different category classes exist in Wikipedia. [**?** ] show, that article size distributions are bi-modal for English and Polish Wikipedia projects. These studies do not care about the network structure of the articles. Based on the node degree or on a centrality measure one can differentiate between leaf nodes, which contain definitions and well accepted facts and more central pages which are related to many topics and which define context as they aggregate several leave nodes. Such additional aspects show, that edit activity is not only related to a change of words or sentences. Furthermore, the embedding of a page is important. In many cases it is even not possible to work with just one page, because the selected topic is represented by different pages within the same language. Aggregation over all pages of the topic - or even a full category - and contextual normalization within the local embedding was developed as a part of this work. Such aggregated measures can contribute to advanced life cycle models on the microscopic scale.

A social network is defined by interactions between many people. No matter if the final result is creation of a new resource in a content network or of it leads to a specific temporal state of minds of all connected participants, one can analyze the underlying structure. According to Borge-Holthoefer *et al.* [**?** ] the evolution dynamics of a

social community can be described by the size of the giant component, plotted as a function of time. Changes of growth rate can be interpreted as an indicator of existence of such a particular social network which has no physical representation, such as Wikipedia topics. The calculation of the giant component uses already more details, such as the link structure instead of counting words only.

In case of Wikipedia we have a mixture of both. The social ties of interconnect users and editors influence the process of content creation. An edit war is something which goes on without explicitly being announced. But this mental state within the community is measurable and contributes to the life cycle of the articles. Such procedures bind or require energy and information can be lost over time.

## 10.3. Towards an Integrated Growth Model for Social Content Networks

According to section **??** the **random graph model** is used to create new links between existing nodes with an equal probability for all possible nodes. A second important model, called **preferential attachment**, connects new nodes to an existing network, based on the properties of the existing nodes (node degree). Thus, nodes with some neighbors have a higher chance to get new nodes attached to it, but in this model, nodes can also be added with out any link. Such simple models are helpful if the final structure of a generated network or the measured state of system should be analyzed. They can not be used to describe the evolution of systems like Wikipedia entirely because the neglect the change of internal state and structure. They do not represent changes in the growth rate nor do they handle different phases within the systems life cycle.

To formalize this idea better, I use the concept of Emissivity as an analogy. Although the analogy is weak it helps to understand the many facets within one coherent framework. First, I compare Wikipedia with a physical body. It consists of matter and has a given structure. In Wikipedia we can not find this. Content in digital documents can easily be copied and one has not to care about conversation of mass. Because we try to describe the flow of information, this is not a critical issue as it belongs only to the description. Lets assume we have contribution to a system, like Wikipedia, we can clearly say, the more information it contains, and the better the structure supports easy access to this information the higher the impact or usefulness of it may be. With this in mind I compare Wikipedia with a solid body, which exists in a field of radiation. The incoming energy flow leads to an increase of internal energy and to internal heating. The body emits energy according to its internal state. In an equilibrium state it emits the same amount of energy as it absorbs.

**Explain Strahlungsgleichgewicht** *Energie wird in z.B. in Form von Strahlung in ein System bertragen. Die Effizienz der bertragung hngt von vielen Faktoren ab. Vernachlssigt man Wirkungsquerschnitt, Wellenlnge und Pulsform und betrachtet nur die Menge an Energie, die tatschlich vom System aufgenommen wurde, dann bleibt dennoch zu unterscheiden, welche Form der inneren Energie erhht wurde. Verschiedene Prozessbeschreibungen oder Modellvorstellungen helfen dabei, solche Situationen zu erklren. Die Erhhung 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 Aktivitt der Wikipedia Editoren, die sich in Form von Edit Ereignissen zhlen lsst, zur messbaren Strukturvernderung 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?

A new integrated model is required. Inspired by the previously mentioned idea of Emissivity we use equation Eq. (10.1) to describe the process of network growth based on information aggregation.

$$\Delta I_{\text{system}} = I_{\text{link creation}} + I_{\text{node creation}} + I_{\text{node changes}} + I_{\text{link changes}} \approx a_{\text{edit}} \cdot v_{\text{contrib}} \tag{10.1}$$

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. Structural information also contains information. This means, also the creation or reorganization of the network structure leads to more information. If a large page is just split into smaller but interlinked pages, it is much easier to retrieve information. Relations to other nodes in the network can be found automatically. Therefore 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. Finally, the link structure and the content within the neighborhood define context. A formal representation of Wikipedia data is available as a semantic graph, as provided by the DBPedia project [**?** ].

Our growth model covers the creation of new nodes as well as the creation of new links beside changes to the existing content and structure rather than the networks topology. In the case of Wikipedia we can easily count the number of edit events. But what goes on exactly during such an edit event is not measured in our current study. Although each edit event is different and different activity leads to different detail results, we unify this to one contribution. Such a contribution covers one, two or all of the mentioned changes on a different level which is not further resolved at this time.

For the selected Wikipedia projects we extracted all link creation events. Each time a new link appears, also a new page can be created, if both do not already exist. All events are grouped by language and sorted by time stamp. 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.

If appropriate computational resources are available, one should also calculate the topological properties of the network as a function of time. This can be done at global or at local scale. The global topology is calculated for the full graph, while a local topology takes only the local neighborhood around some pre selected nodes which represent the topic one is interested in well at much lower cost.

Time resolved calculation of topological properties requires an incremental update of a large growing in data structure. During each update step, real data is used, and in parallel, simulations are possible within the same framework. This allows inspection of the influence of time dependent attachment probabilities. Figure 10.1 shows the integrated analysis and simulation procedure which is currently under development. Preliminary results are presented in the remaining pages of this chapter.

## 10.4. Comparison of Growth of Article Number and Article Quality

Several models are used to describe the growth process of networks. Two very popular network models are the random graph model and the scale free network. Both models describe how the internal structure evolves in time,

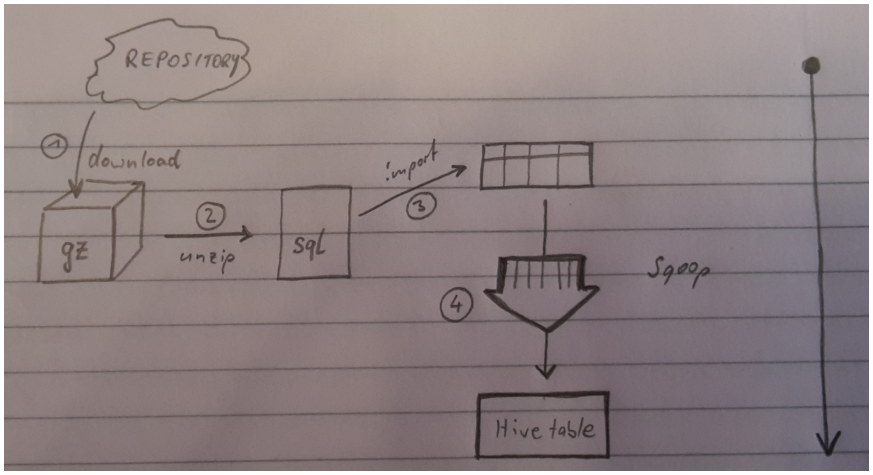


Figure 10.1.: **AnalysisDetailsGrowthAnalysis**. SHOW THE FULL PROCEDURE AS A SKETCH ... using the DSPM.

based on the degree distribution. In the first case, one assumes that all nodes (pages) already exist, and the growth process consists of adding links, one in each time step. In the second case, one page is added in each time step. This means a new link and a new page are created at the same time.

In a real network, like the Wikipedia content network, both processes of adding new pages and adding new links between pages are coupled and cannot be separated from each other. In order to describe the growth of the Swedish Wikipedia project in more detail we analyze the growth rates for the number of pages and the number of links. Because we have several types of links we also compare the growth rates of the number of links for those types. *Internal links* are links within the same Wikipedia (same language) and redirects to another page of the same Wikipedia. Internal links represent semantic relations between the terms the pages are about or just relations between topics or concepts which are used within a certain page. If the meaning of a term is ambiguous, special pages help to show users all possible meanings (based on other pages). Such pages do not contribute much text, but this structural information is of a high value and increases the usability of Wikipedia. *External links* are links to another language (Interwiki links) and links to pages outside the Wikipedia project (e.g., references). The frequency of such links represents an important quality indicator for Wikipedia articles.

### 10.4.1. Evolution of the Degree Distribution

All links contribute to the Wikipedia's structure which evolves over time. The creation of a new link is a result of an edit activity of an user. Figure 15 shows the temporal evolution of the internal link degree distribution for all pages of the Swedish Wikipedia. Redirects and external links are disregarded in this plot. Already since the beginning in 2002 the degree distribution can be described by a power law, with the exception of pages with a very low degree (low number of links). While pages are added over time, the distribution changes and its power-law shape becomes more obvious, since the range of degrees becomes wider. Actually, most of the pages have much more than ten internal links and are well described by a power-law degree distribution. Only the number of pages with less than ten internal links is smaller than assumed in the scale-free model that predicts power-law degree distributions.

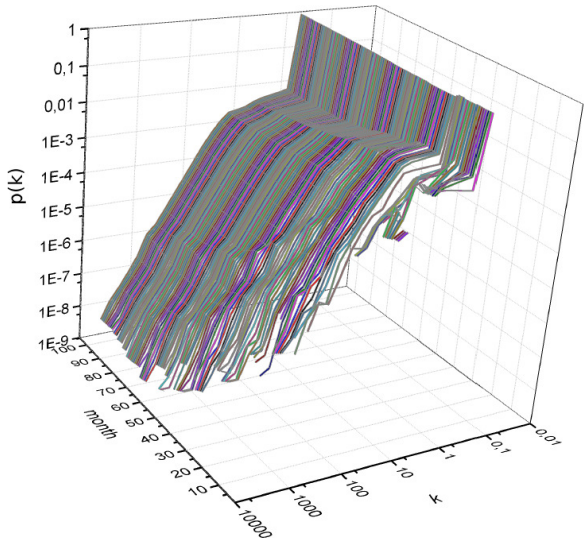


Figure 10.2.: **Evolution of Degree Distribution**. Degree distribution $p(k)$ (i.e., distribution of the number $k$ of links per page) for internal links in the Swedish Wikipedia project. One curve is shown for each month from January 2001 till December 2009.

### 10.4.2. Growth of the Content Network and Structural Changes

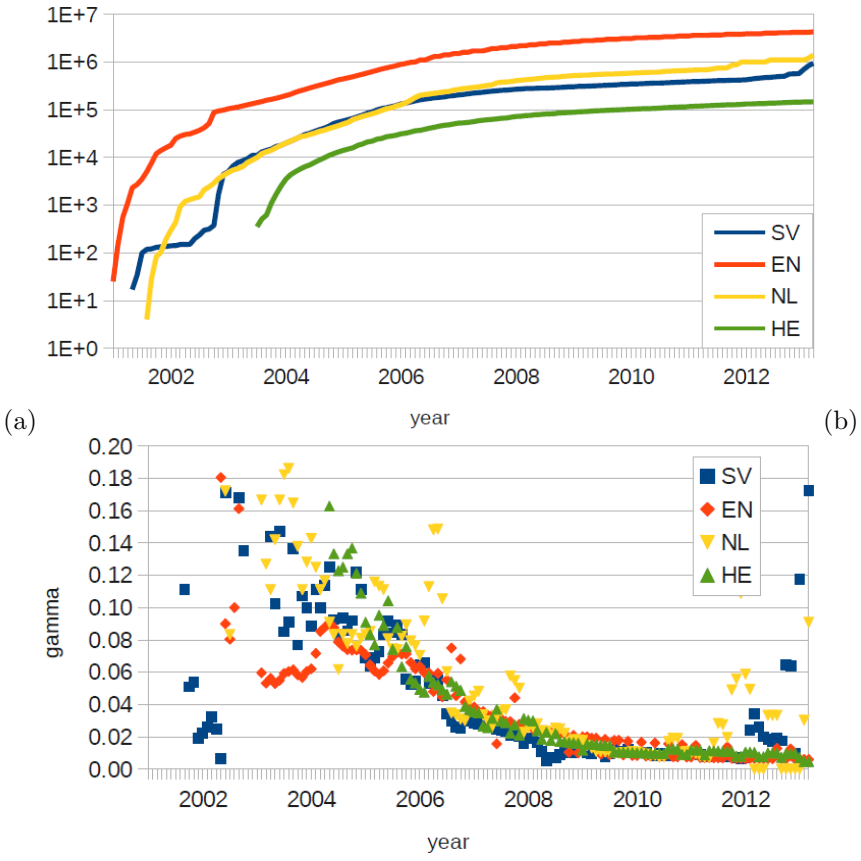


Figure 10.3.: **Comparison of Growth Rate for four languages**. For the Wikipedia projects in Swedish (blue), English (red), Dutch (yellow), and Hebrew (green) (a) the number of pages and (b) the exponential growth rate $\gamma$ is shown.

Figure 16(a) shows the total number of pages for four Wikipedia projects (Swedish, English, Dutch, and Hebrew). The number of pages $N_P(t)$ is growing by the number of new pages $n_P(t) = N_P(t) - N_P(t-1)$ per time interval $\Delta t = 1$ month. Figure 16(b) shows the growth rate $\gamma$ for an exponential growth model $N_P(t) = N_P(t-1)\exp(\gamma)$,

which has been determined by $\gamma \approx n_P(t)/N_P(t)$. Note that an increased $\Delta t$ has been used if $n_P(t) = 0$.

In the beginning the growth rate $\gamma$ is quite large. Later, a tendency towards saturation can be identified. This shows that the character of edit events changed over time. In the early stage of a Wikipedia project most of the edit events are related to the creation of new pages, while later on the internal structure evolves. For the English Wikipedia project, one can see an intermediate regime with an exponential growth ($\gamma \approx 0.7$). Such an exponential growth cannot be unambiguously identified for the Swedish Wikipedia. Interestingly, the page-growth rate has been drastically increasing during the last few months (in 2013) for the Dutch and – even more dramatically – for the Swedish Wikipedia. Acutally, the Swedish and the Dutch Wikipedia started to create articles using bots.

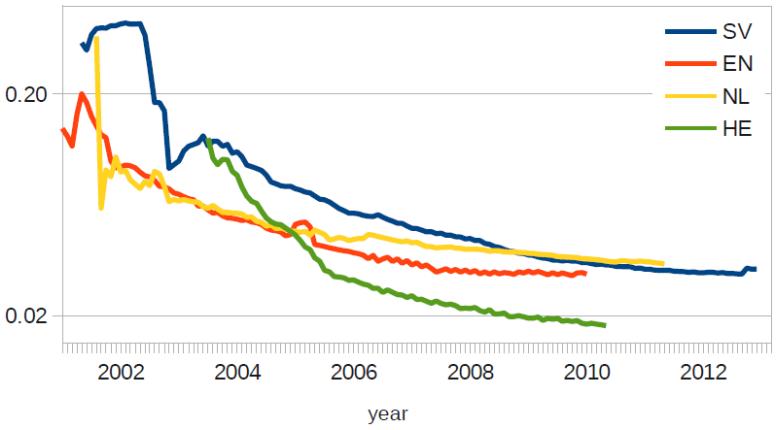


Figure 10.4.: **Adding content or change the structure**. Ratio of total number of pages $N_P(t)$ and total number of links $l_P(t) = l_{\text{int}}(t) + l_{\text{ext}}(t)$ (internal and external) as fuction of time between January 2001 and April 2013 for the Wikipedia projects in Swedish (blue), English (red), Dutch (yellow), and Hebrew (green). Note that only Swedish data was available till 2012 and that the vertical axis has a logarithmic scale.

Figure 17 confirms that editorial activity tends to focus more on the addition of links than the creation of new pages during later states of Wikipedia evolution. It shows the ratio of the total number of pages $N_P(t)$ and the total number of links $l(t)$ as function of time. For all languages this ratio decreases during most of the time after a relatively large value (around 0.2, i.e., approximately five links per article) in the beginning. The final values are between 0.015 and 0.04, i.e. at approximately 25-60 links per article. For the Dutch and the Swedish Wikipedia the initial change (between 2001 and 2003) is quite sudden. In general, all four languages show a stronger decay of the page number to link number ratio in the beginning and a much slower decay later on. This behavior suggests that an exponential decay model may also be appropriate. However, we cannot find any regimes with unique or approximately constant decay rates for any of the considered four languages. The different decay rates of the page number to link number ratio might also be indicators for two different network growth processes.

The Swedish Wikipedia has initially $\approx 5$ links per page and later the number of links per page increases to an average of $\approx 25$. This is in line with the change in the degree distribution, which is shown in Fig. 15. Here one can see a continuous shift towards a dominating structural growth process, while the growth of content volume – measured in number of pages – becomes less important. The current ratio of page number to link number for the Swedish Wikipedia is quite similar those for the English and the Dutch version, while the Hebrew Wikipedia has about twice as many links per article. During the quick growth of the Swedish Wikipedia article number in the past few months (see Figs. 16(a,b)), the article to link number ratio has slightly grown, which may indicate a slight

change of the structure towards properties typical for Wikipedias at earlier stages of evolution. Although, this weak growth is still comparable with typical fluctuations of the ratio (just about twice as large), it may indicate that creating articles by bots leads to a step back in the quality of content.

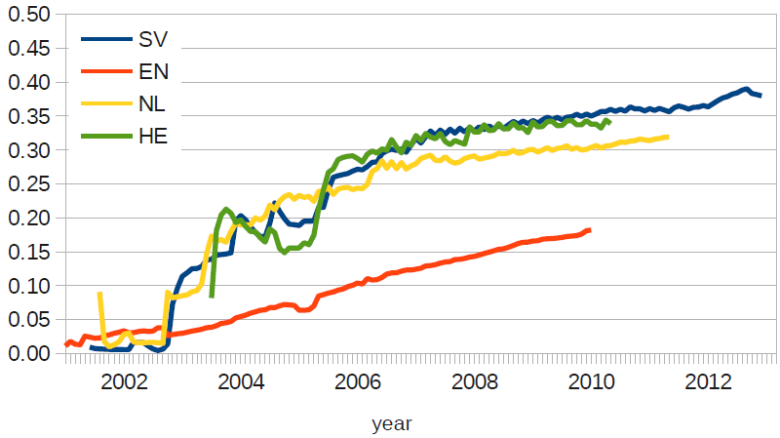


Figure 10.5.: **Change of internal structure vs embedding**. Ratio of number of external links $l_{\text{ext}}(t)$ and total number of links $l_P(t) = l_{\text{int}}(t) + l_{\text{ext}}(t)$ (internal and external) as fuction of time between January 2001 and April 2013 for the Wikipedia projects in Swedish (blue), English (red), Dutch (yellow), and Hebrew (green). Note that only Swedish data was available till 2012.

Next we separate the changes of internal and external link numbers. External links (to other language versions or references outside Wikipedia) are particularly important for confirmation of the article content and can thus be regarded as an important quality indicator for the articles. Figure 18 shows the ratio of the number of external links to the number of all links (internal and external). The increasing curves show, that the ratio of external links grows for most of the time in all four Wikipedias. We note that there are two major groups of external links: just 'further reading' links (often in bad articles) and references (more likely in good articles). The habit of adding references increased in the last years, while one got more discouraged adding just simple links; they are usually also limited to 3-5 per article.

In Fig. 18 one can also find indicators for two regimes for each Wikipedia project except the English: fast growth from approximately 2003 to 2005 followed by a behavior close to saturation after 2005. Table 2 shows the times where the qualitative behavior changes. The transition time A ($t_A$) is determined from Figure 17, which shows the ratio of total number of pages and total number of links while the transition time C ($t_C$) is based on the plot in Figure 18, which shows the ratio of external and internal links. For all languages we find that $t_A$ is before $t_C$, but the differences vary from 4 to 52 months depending on the languages.

| Language | $t_A$ | $t_C$ | $t_C - t_A$ |
|---|---|---|---|
| SV | 05/2003 | 10/2007 | 43 |
| EN | 12/2001 | 12/2002 | 12 |
| NL | 01/2002 | 05/2006 | 52 |
| HE | 09/2005 | 01/2006 | 4 |

The Swedish Wikipedia has already reached the largest ratio of external links among all links in 2010 and has continued to increase this ratio during the last 2.5 years (see Fig. 18). This is an indication for a very good reference quality of average articles in the Swedish Wikipedia, better than in the Dutch Wikipedia. Note that

the ratio of external links is very much lower in the English Wikipedia, just approximately half as large as in the Swedish Wikipedia (data from 2010). The slight drop of the Swedish curve in Fig. 18 in the last months is probably associated with the drastic increase of the total number of articles (see Fig. 16). However, it is too weak to be considered as an indication of a drop in article reference quality, and there was a significant larger increase during 2012 just before the slight drop. We note that bot generated articles usually have a quite high density of references, meaning just one sentence but 2-3 references to publications which, however, may not be linked using a web link. Overall, we can conclude that the article reference quality of the Swedish Wikipedia has not decreased significantly in the past months despite the quite drastic content increase with nearly doubled article number probably due to article creation by bots.

### 10.4.3. Attachement Probability

A link is created to a node with degree k=x. For all link creation events we count the number of nodes with degree k within a given range $k_min...k_max$ to draw a histogram which shows the distribution function of the attachment probability.

**We have to compare this result with the theoretical distribution used in in the BA-Model ...**

Link creation in the presence of edit-activity and access-activity has to be studied if the model should be modified for per link creation predictions.

So far we investigated the system properties only. But for a prediction of a probability for link creation between two nodes, a separate approach is required. For a given time, one wants to calculate a link creation probability for a selected node pair. Therefore the following data is taken into account: the short term history of both nodes regarding edit and access activity.

The large amount of Wikipedia data allows application of a machine learning algorithms such as decision trees and rule-based models for classification.

But one has to be careful. Because the system is not stationary, the models may not behave well unless they are trained with data from within the same phase. Here we investigate data from one regime only. A detailed study of the impact of non stationary conditions has to be investigated in a separate project.

Here we consider the logistic regression to classify the event under consideration as one of type A, which means, a link is probably created within the next time window or of type B, which means, no link will be created. Other properties, such as correlation between time series, existence of peaks in both series can be used as additional features which may have an influence on the quality of the model.

**More work is required here ...**

**Further research (maybe later on) ...**

### 10.4.4. Phases and Phase Transitions

We distinguish different phases in a less restricted way, compared to the clear definition of a phase transition in physics. If a particular property is dominating, we consider this a phase in the life cycle of the system. A

more precise term could be *regime*. More research on more data is required before real phase transitions can be propagated.

Figure ... shows quite stable growth rates within several time ranges. Even a negative growth rate could be measured, this is a clear indicator for internal structural changes.

Can we differentiate between phases with more change in content from such with more structural changes?

## Conclusion

As shown in this section, one can study the life cycle properties of Wikipedia projects, based on the pages edit history and simple system parameters, such as number of pages, links and their changes. Each new page and each link creation event can be extracted from this data set. Even real time studies would be possible. Although Wikipedia does not provide aggregates of edit events on a daily base, it seams to be reasonable to provide such data. This would enable the public research community to analyze the coupled processes on a large complex system with less overhead, which is currently caused by the pre-processing procedures.