Meta Correlation Analysis with Hadoop.TS

Author: Mirko Kämpf
Version: 1.0.0
Date: 23.12.2013

Correlation analysis is a very common technique which is applied in physiological, financial and social media data analysis. The Hadoop.TS project provides some fundamental time series algorithms, e.g. for detrended fluctuation analysis (DFA and MFDFA), return interval statistics (RIS), cross-correlation analysis (CC), and the event-synchronisation (ES). A special type of cross-correlation analysis is the meta correlation analysis.

This hands on tutorial shows, how to do a large scale time series analysis project in a Hadoop cluster with the Hadoop.TS libraries.

In part one we explain and implement the an analysis tool which performs a meta correlation analysis for time series from Wikipedia and stock market data, retrieved from Yahoo financial services. This particular analysis technique was inspired by the paper "Evolvement of Uniformity and Volatility in the Stressed Global Financial Village" published in PLOS ONE.

The prototype works as a standalone application on a single cluster node and in part two we will show, how to apply the analysis method to a really large data set in parallel mode. This two step procedure is recommended as a best practice for time series algorithm development. Especially if you plan to implement new analysis procedures you might save a lot of time if you develop and test in a local mode.

The tutorial assumes you have a linux box with CentOS (6.4), the latest Oracle JDK and the Netbeans IDE (version 7.3). Beside this, a Github client is required.

Now it's a good time to start reading our first paper about the Hadoop.TS package (published by IJCA). Results from our previous Wikipedia usage analysis projects have been published in Physica A and presented at the European Conference for Complex Systems.

- Fluctuations in Wikipedia access-rate and edit-event data (Physica A, 2012)
- From Time Series to Co-Evolving Functional Networks: Dynamics of the Complex System 'Wikipedia' (presented: ECCS 2012, Bruseles)
- Comparing the usage of global and local Wikipedias with focus on Swedish Wikipedia (student's project report, 2013)

Section 1: Research Report

Introduction

Many studies have focused on properties of social networks and systems, using methodologies from complex systems research. Financial markets are such a type of system, and as such many studies have investigated their structural and dynamical properties using such tools. As more and more large data sets are available for research, a convergence of such methods can be seen. However, while many social systems seem intuitively connected, little research exists on these interconnections. Financial markets can't function without information flow.

Trading decisions are usually based on information available to the trader. Such information is distributed via several channels, e.g. by news articles, journals, and during the last decade increasingly via online services. But on the other side, an increase of the amount of information which is in the news is also triggered by events like business decisions and big movements in the market. Such coupled channels might lead to waves of events like extreme events in wikipedia access time series or to a change in the volume of information which is delivered by news magazines and channels.

Furthermore, people consume a lot of information via Internet applications like messaging services, provided by many communication networks (Email, Twitter, Facebook) or public web pages like the encyclopedia Wikipedia. By analyzing the correlation between stock market data and the access rates to special groups of the Wikipedia articles we study the role of the social network in the economic cycle. Can an increase of interest in a special topic be used as an indicator for changes in demand in financial markets? Questions like this one are important for individual decisions but can we also identify a global state of systems and their dynamics based on time series analysis?

A recent study by Preis et. al. [1] showed that following query volume for financial search terms on Google could predict stock market movement and another study by Alanyali et. al. [2] demonstrates a significant change in correlation between the daily mentions of companies in the Financial Times in the morning and how much they were traded on the stock market during the day. Those results support the hypothesis of an existing mutual influence between financial markets and the news.

We are interested in the role of Wikipedia as a source for news in the context of market activity. In our study, we are seeking an evidence for a or a qualitative description of the relationship between movements in financial markets and changes in the user activity in wikipedia.

The aim of this work is to study whether it is possible to gain insight into early stages of economic decision making in the stock markets, by linking large scale records of online information gathering and trading actions. We have previously shown that increases in the number of searches for a company name made on Google are correlated with increases in trading volume for that company's stock [ref]. More importantly, it has been demonstrated that during the period 2004-2011, increases in searches for financially related terms tended to be followed by decreases in the price of the Dow Jones Index Average [ref]. This finding is in line with the proposal that Google search data may provide insight into the process of traders

seeking information to help them determine optimum future decisions.

However, other online data sources may also possibly provide insight into trader information gathering processes. Whilst many Internet users rely on Google to locate a range of different useful information sources online, the online encyclopedia Wikipedia is a widely-used central reference source for information across a number of subjects. As such, one can consider Google as a provider of data that gives insights into what information Internet users are looking for, whereas Wikipedia data provides insights into what information Internet users in fact find. Thus, we have investigated whether changes in frequency of views of certain Wikipedia pages also anticipate subsequent changes in stock market prices.

A few key practical differences exist between data on Wikipedia usage and data on Google usage can be identified. Firstly, some search terms have multiple meanings. For example, the term "Apple" is widely recognized both as the fruit and as the technology company. Google data, as retrieved from Google Trends [ref], provides little insight into which meaning was of interest to the Internet user. In contrast, a Wikipedia page, other than those designed specifically for disambiguation, is about one topic only. This makes it possible to consider changes in the number of views of the page about Apple the company separately to changes in the number of views of the page about apple the fruit. Secondly, where data on Google usage largely relates to per-week changes in search volume, we are able to access data on hourly changes in Wikipedia usage. Thirdly, Wikipedia data describing access of all pages across the Wikipedia encyclopedia since 2007 is freely available, whereas some restrictions exist on accessing large volumes of Google usage data.

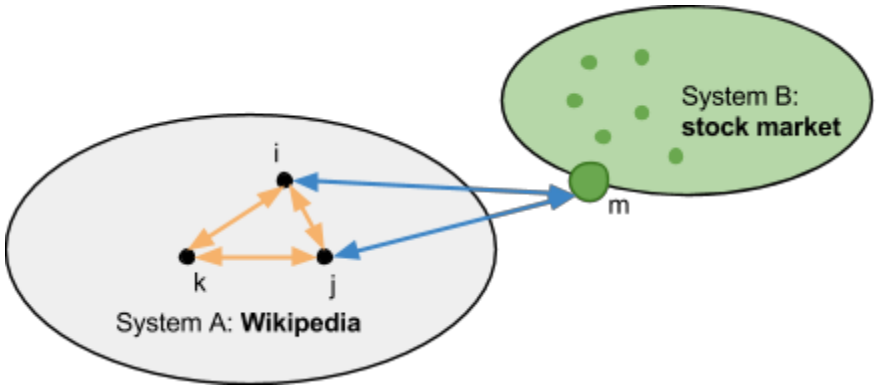


Fig. 1:

Therefore we use data which can easily be measured or obtained from existing databases as time series, to analyse individual properties of lots of single elements which form a complex system. Such time series are also used to reconstruct correlation networks which expresses the strength of intra component correlations. Inter component correlations are

calculated between different types of elements and the results are Bipartite networks, which expresses the coupling of different subsystems forming a complex system. One important questions is: How to define links between elements and subsystems? How are time delays handled correct and how are artifacts identified? In our work the Pearson correlation as well as the event synchronisation are used to reconstruct an adjacency matrix which represents the underlying system as a network. This allows us to apply and compare several filter techniques and to characterize properties of these networks as a function of time. Our approach is in line with time series analysis methods, which often depend on the correct filtering of raw data. Our results show, that these aspects can not be handled in a unified method - it depends on the properties of the measured data and the kind of effects that should be analyzed.

Methods

Description of the Analysis Procedure

Two systems A and B are considered to be bi-directional coupled interacting systems which consist of elements e_A and e_B. For each element a set of properties is measured and stored as a time series. For continuously measured data a time series has a length of n values at a sampling rate s_r with an equal distance dt. Values can be accessed by a numerical index and the time series has a start time t_0. The procedure works with a sliding window of length $dt * s_w$. which is shifted by dt_s. In the case $dt_s > s_w$ we have non overlapping windows otherwise windows do overlap. Event time series are not considered explicitly. In such a case the event time series would be filled up with zero if no event occurs during a given time frame defined by the sampling rate. This approach is not very efficient but for the beginning and especially with continuously measured data it will work well.

We want to measure[1] the influence of system B on the intra correlations measured for system A. Therefore the average intra correlations are calculated for all pairs of nodes in system A. We avoid self loops and because the cross-correlation function is symmetric, we calculate the correlation strength $C(i,j)$ only for pairs $a_{i,j}$ there $i > j$

$$C(i,j) = \frac{<(ri-<rj>)(rj-<rj>)>}{\sigma i \, \sigma j}$$

and according to [3] the partial correlation ρ between wikipedia pages i and j, using the time series **m**, which represents system B as the mediating variable is defined by [4], [5], [6]

$$\rho(i,j|m) = \frac{C(i,j) - C(i,m)C(j,m)}{\sqrt{(1-C^2(i,m))(1-C^2(j,m))}}.$$

As weak correlations between user access-rate time series also occur randomly (because of limited statistics), a filter or threshold has to be applied to eliminate unreliable or too weak links.

[1] What does measure mean here?

We thus re-normalized computed link strengths. The calculation of $C(i,j)$ was repeated ten times for randomly shuffled time series $\mathbf{a}_i(t)$ and $\mathbf{a}_j(t)$ to determine normalization factors for each pair of nodes. The average correlation value is:

$$C^{intra} = \frac{1}{N}\sum_{i=1}^{N}\overline{C(i)}$$

.

For each episode or time period we get one such value where ***t*** is the index of the time period. The calculation of PC requires three correlation values. For tuples of three time series which consist of pairs of nodes from system A and the average value which represents System B we calculate the partial correlation:

$$\overline{PC(i)} = \frac{1}{N-1}\sum_{j\neq i}^{N}\rho(i,j|m).$$

and finally the average partial correlation for each episode:

$$PC^{intra} = \frac{1}{N}\sum_{i=1}^{N}\overline{PC(i)}.$$

A measurement of the influence of system B on system A is given by the ratio of C and PC and was introduced by [7], [8] as the “Index Cohesive Force” (ICF).

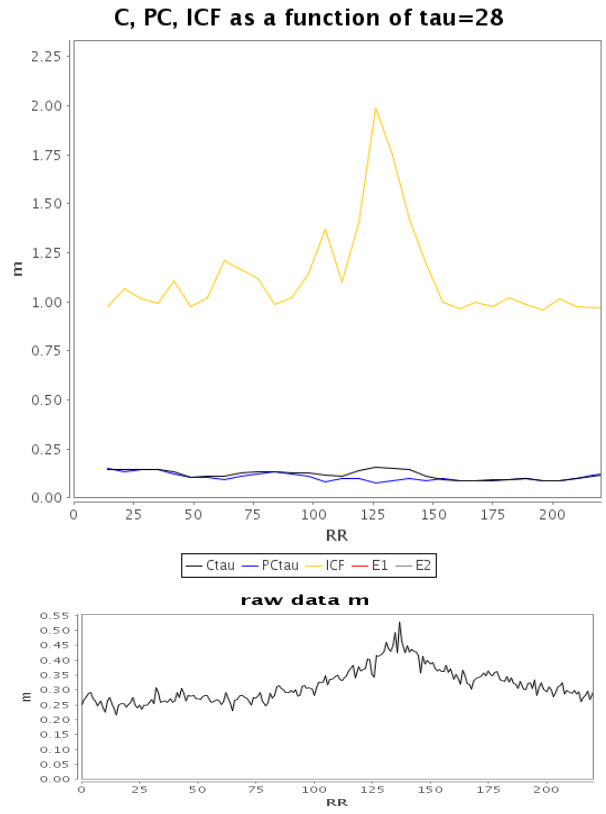


Fig. 2: The yellow curve shows the ICF values which increase in the presence of endogenous event which happens in the mediating tim series **m**. This clearly indicates that the intra correlations within the local article network increased before the peak in the core access counts was measured. This way it might be possible to predict the time of a change in a local trend.

In comparison to the observed effect the index has on stock correlations we analyse the correlations between the wiki page which represents the stocks index and the companies pages access rate time series. Kennet et. al. found that larger changes of the index results in higher stock correlations. Based on those findings we conclude with the following hypothesis:

If movements in stock markets cause an increase of interest in financial topics in Wikipedia one would measure a) an increase of intra-wiki correlations between pages, related to a given market and b) an increase in partial correlations between the stock market data and the wikipedia access rate data.

So far we did not calculate meta correlations between different groups of Wikipedia articles. It was shown, that several markets are interconnected, even the Wikipedia pages of several markets are interconnected like shown in the structural network in fig. . *Can we find such a connectivity also for the correlation and dependency networks of Wikipedia pages?* Therefore a

meta correlation, which is the cross correlation between the average intra correlations from two markets ***i*** and ***j*** are calculated:

$$MC(d) = \frac{\sum_{t=1}^{N-d} \left(\overline{C^i(t)} - \langle \overline{C^i} \rangle\right)\left(\overline{C^j(t)} - \langle \overline{C^j} \rangle\right)}{\sqrt{\sum_{t=1}^{N-d} \left(\overline{C^i(t)} - \langle \overline{C^i} \rangle\right)^2} \sqrt{\sum_{t=1}^{N-d} \left(\overline{C^i(t)} - \langle \overline{C^i} \rangle\right)^2}}$$

and compared with the correlation of the average access-rate for all groups, like shown in the previous section as we defined the time resolved relevance and representation index for a single Wikipedia page.

Data Set

Our data set consists of three major components. The Wikipedia pages, hosted by the Wikipedia Foundation are used as an online data source to retrieve data as neede, as well as the Yahoo Financial services. We load the local page networks and the financial time series once and hold this data in a local cache, implemented in an HBase cluster which allows fast random access to any individual fact. The access rate time series are provided as aggregated file with all click counts for all Wikipedia pages from one hour. This type of representation requires a transformation. Each file contains only one data point for each page. Construction of time series is done after the download for each individual group for which an analysis is done. In a later stage we would expand the whole access time data in an HBase table, which allows a much more flexible time series creation procedure.

Preparation of stock market data

The following table shows seven selected indices from five important globally relevant markets. Only for markets (highlighted in green) we can download the daily closing price (+ in column c) and the daily trading volume (+ in column t). The S&P 500 was not considered in this first case study, because of it's large number of companies it contains. This would lead to a relatively long running analysis procedure. As we have limited resources for this project we skipped that index.

Market	**Stocks**	**Index** (Yahoo symbol)	**CN in Wikipedia Network**	**c**	**v**
1 US	S&P 500	S&P 500 (**^GSPC**)	http://de.wikipedia.org/wiki/S%26P_500	+	+
1 US	NASDAQ	NASDAQ-100 (**^NDX**) NASDAQ-Composite (**^IXIC**)	http://de.wikipedia.org/wiki/NASDAQ-100 http://de.wikipedia.org/wiki/Nasdaq_Composite	+ +	- +
2 EU	FTSE 350	FTSE 350 (**^FTLC**)	http://de.wikipedia.org/wiki/FTSE_350_Index	+	-
3 EU	DAX Composite	DAX (**^GDAX**)	http://de.wikipedia.org/wiki/Composite_DAX http://de.wikipedia.org/wiki/DAX	+	+
4 India	BSE 200	BSE SENSEX (**BSESN**) BSE 100	http://de.wikipedia.org/wiki/BSE_Sensex	+	+
5 China	SSE Composite	SSE Composite (**000001.SS**)	http://de.wikipedia.org/wiki/SSE_Composite_Index	+	-

6 Japan	**Nikkei 225** Nikkei 500	Nikkei 225 (**^N225**) Nikkei 500	http://de.wikipedia.org/wiki/Nikkei_225	+	+

Table 1: List of selected markets, their symbol to identify the data series in Yahoo financial services and the data sets which are available for each market.

The financial data crawler works well with Yahoo web services [9]. The daily closing price is shown in the chart. Data is provided only for trading days so we store such data as a set of key value pairs. The key is a timestamp and the value represents the closing price or the trading volume respectively. Because the data points in the trading data series are not equidistant we have to fill the missing values (during weekends or bank holidays) with the last available value. Doing this, we can compare the daily acces count time series and the daily trading data directly and we do not destroy the daily and weekly cycles in Wikipedia usage data series, like we would have done if non trading days would have been removed from wikipedia data set.Figure 3, 4 and 5 show the preprocessed raw data series.

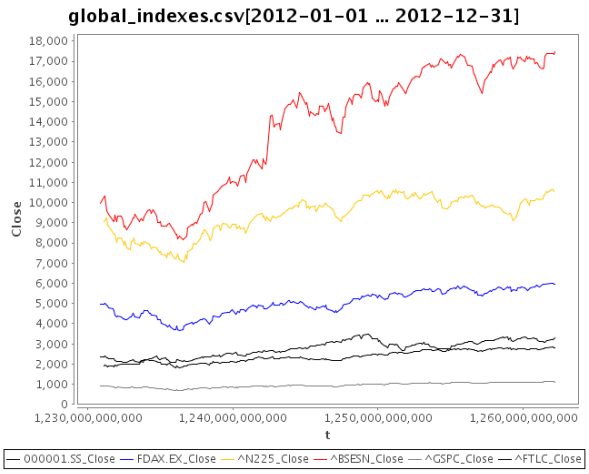


Fig. 3: raw data (daily close price) for six markets in 2009

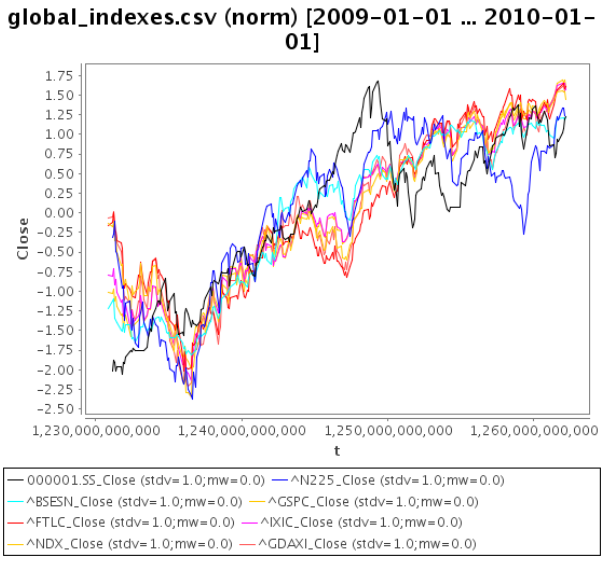


Fig. 4: normalized time series for closing price of 8 stock market indices (mean=0 and sigma=1).

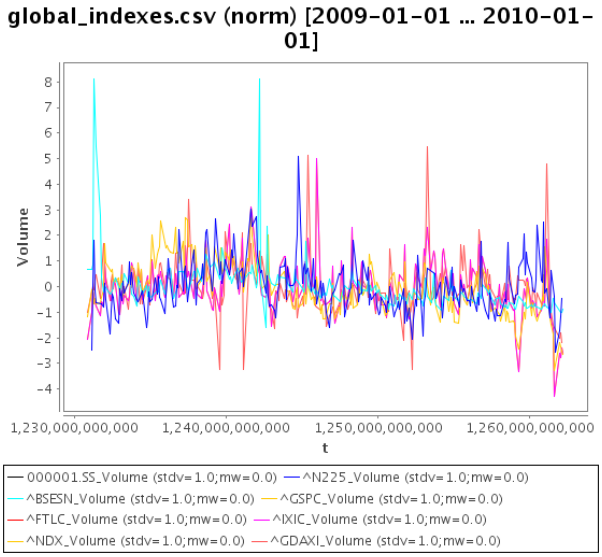


Fig. 5: normalized time series for trading volume of 4 stock market indices (mean=0 and sigma=1).

trading volume was not available for all indices.

Preparation of Wikipedia data

The local networks of pages around the pages of interest (central nodes, CN) and the corresponding click count data have to be extracted from different sources, provided by the Wikimedia foundation. Wikipedia is a multilingual system and it contains different types of links with a different semantic meaning. Pages about the same semantic concept are linked by so called inter-wiki links.

In this use case we start with a list-pages instead of single pages about a specific topic. The list-

page about a stock index contains links to the pages about companies included in that index. Traversing the inter-wiki links guides the crawler to list pages about the same index in different languages. So we can easily collect core nodes of the local networks of pages which cover the stock markets. Beside this we also collect all pages which are linked to the core nodes because those pages are directly related to the topic of interest. Here we have a specific situation because list pages and their related nodes are covering the same topic, while the links from normal pages usually cover the close neighborhood, which might be related but usually the neighborhood contains also many not obviously related aspects.

Such a purely data driven extraction method uses the known properties about the structure of local networks around wiki pages and so called list pages. Such implicit information lowers the barriers especially in a multilingual global environment like Wikipedia. This approach can also be extended to any type of web resources, especially to the the linked data cloud [10].

Selection of representative Wikipedia pages

Depending on the subject and objectives of a research or analysis project one has to select an characterize the data sources carefully.

Univariate analysis can be done on any group of time series and result can be grouped and sorted depending on several properties which might have an influence on any measured result one gets from an analysis procedure, which usually is of a linear complexity.

Bivariate or multivariate procedures require a precise definition and characterization of the chosen input data. One has to think about the number time series pairs or tuples which have to be processed. A slightly more expensive operation of the order O(N^2) is the creation of simple correlation matrix, which is a symmetric matrix, self-loops can be omitted and for each pair, only one direction has to be calculated. A dependency network is created from a filtered dependency matrix, which is not symmetric and which is calculated for all pair of pages in the presence of a third time series which can be each time series within the set. This leads to an order of (O(N^3).

It is important to classify analysis algorithms in that way because one can find out if or how well the procedures would behave on a large scale with thousands or more nodes. The time at what the grouping is done differs for univariate and multivariate procedures and this has a strong impact on the overall data management procedures.

Beside this considerations regarding required computational resource and procedures the content drives the definition of subsets, especially in the context of large data collections like in our case.

Any wikipedia page can be selected as the starting point for data collection. Starting from this central node all connected pages can be retrieved and depending on the depth of the crawling procedure a small or even a very large network can be collected.

SHOW a table of network size for our 3 reference nodes for link depth of d=1, …

A restriction to the next neighbors is often useful but sometimes an inclusion of the second neighborhood might enable a comparison or a relative measurement of a certain property within a given well defined context. In Wikipedia we find different types of links. Beside normal page links, external links to resources hosted outside Wikipedia we can work with so called inter-wiki links. We consider all page, available via such inter-wiki links as pages in the so called IWL group. The central node defines a local neighborhood (LN) and all pages linked to any IWL page define the global neighborhood (GN). Because this group contains not the initially selected central node the segregation allows a language dependent analysis. According to figure 6 we define and work with four groups: CN, IWL, LN, and GN. If one is interested in a semantic concept independently of language specific differences one can define a core which consists of (CN+IWL) and a hull which is formed by (LN+GN). The neighborhood is one step away from the center or the core.

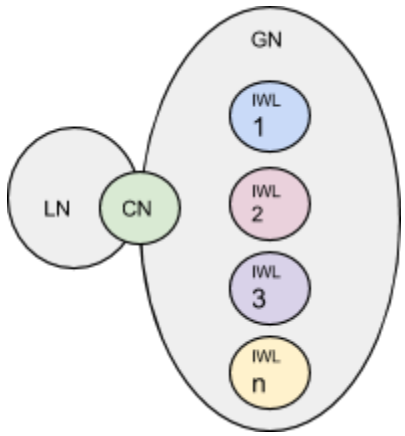


Fig. 6: The central node CN is a list-page which describes a stock market index. It has links to related pages in the same language (group LN). The group LN is the local neighborhood and contains pages about the companies which belong to that index. Inter-wiki links connect CN to many IWL pages in different languages covering the same index. Such META pages contain links to their local neighborhood. All of those groups define the global neighborhood GN. In our case the pages from LN and GN represent the pages which people are using in order to get details about the companies listed in an index while CN and IWL pages contain more general information and can be considered a hub in the local network.

Data about page, metadata, structural information and time series data has to be inspected and characterized before more expensive analysis procedures are started. This purely data driven approach gives a better understanding of and allows identification of artefacts. Figure 7 shows a "*relevance plot*" for crawl results from six local networks (for deetaisl see table 1). The size of the circles shows that for the German index DAX the most pages have been collected and the page representing the Chinese index SSE Composite has the lowest number of linked pages.

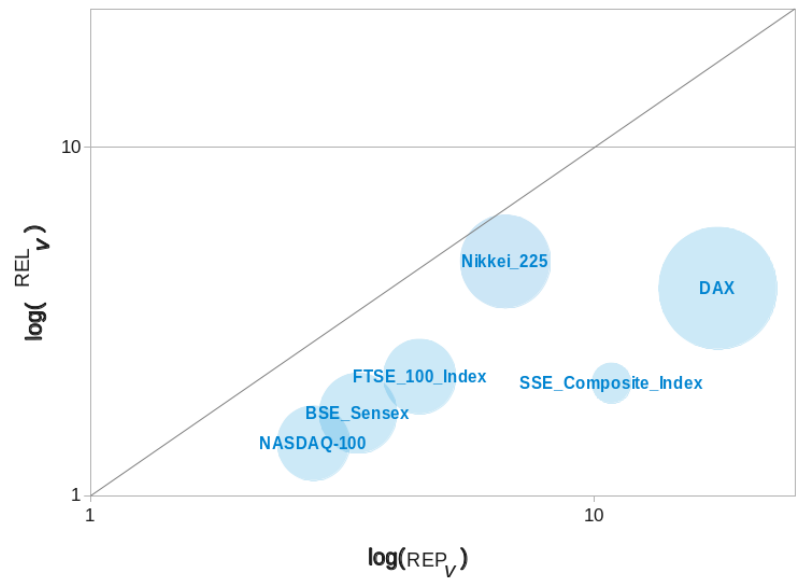


Fig. 7: The relevance plot shows the representation index (x-axis) and the relevance index (y-axis) calculated from text volumes for six stock markets. The size of the circles is proportional to the number of links which are going out from the chosen list-pages, which define the centre of the local networks and represent each market in Wikipedia.

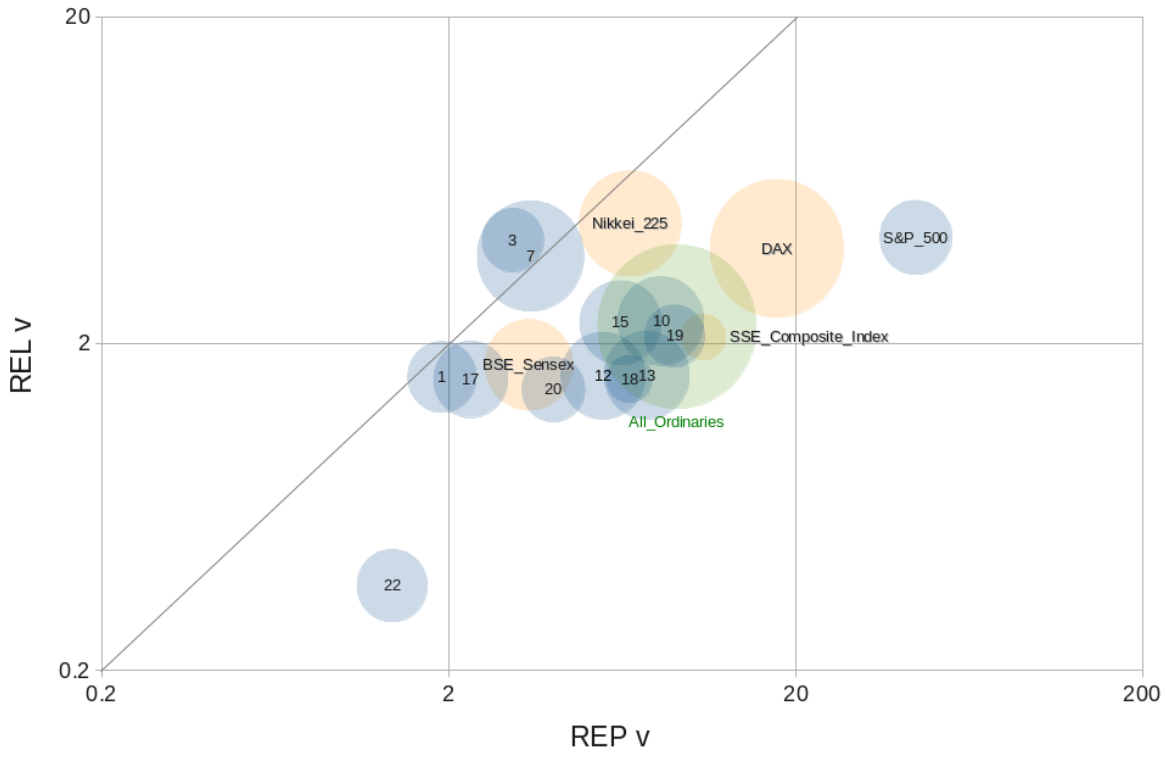


Fig. 8: The relevance plot shows the representation index (x-axis) and the relevance index (y-axis) calculated from text volumes for six stock markets. The size of the circles is proportional to the number of links which are going out from the chosen list-pages, which define the centre of the local networks and represent each market in Wikipedia.

1	Indice_de_Precios_y_Cotizaciones
2	S&P_500
3	S&P_Africa_40_Index
4	Nikkei_225
5	SSE_Composite_Index
6	DAX
7	FTSE_350_Index
8	BSE_Sensex
9	CAC40
10	Austrian_Traded_Index
11	AEX
12	Swiss_Market_Index
13	Athex_Composite_Share_Price_Index
14	MICEX
15	Hang_Seng_Index
16	All_Ordinaries
17	NZX_50_Index
18	KOSPI
19	Taiwan_Capitalization_Weighted_Stock_Index
20	Straits_Times_Index
21	IBOVESPA
22	TA-100_Index

How are those pages and those groups interconnected in Wikipedia? The page network is plotted in two color codings. Fig. 8a shows pages by index they are linked to and in Fig. 8b the color represents the language of the page.
SHOW a new network with better layout and another color coding by lanugag

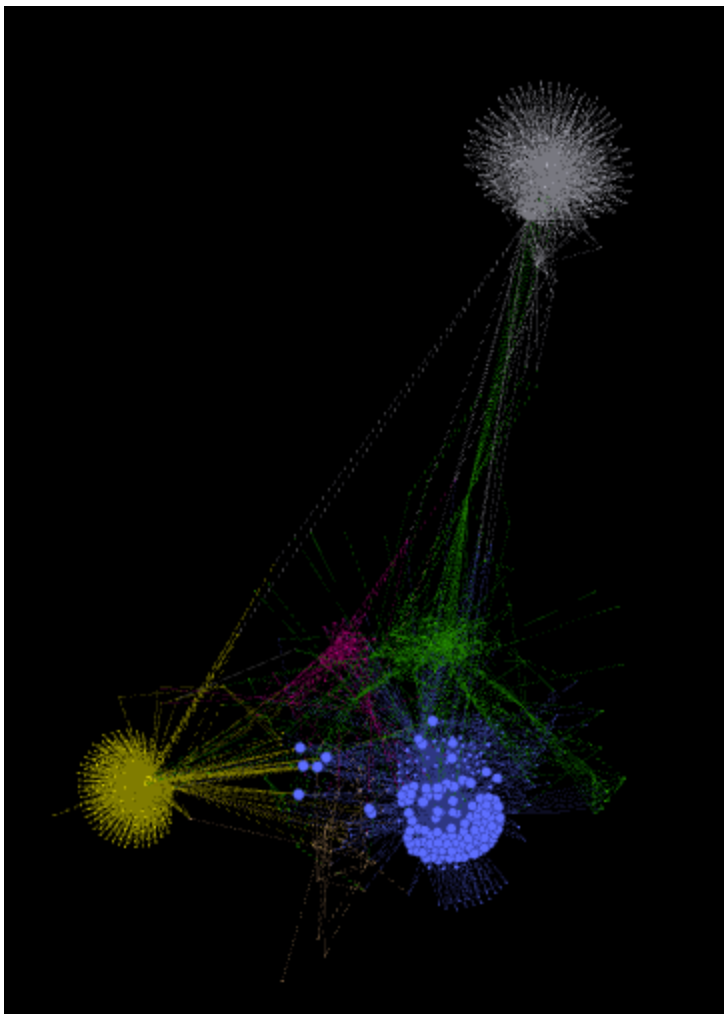


Fig. 8a: Wikipedia can be used as a source for information about financial markets. Multiple Wikipedia projects in several languages are inter-connected by so called “inter wiki” links (geen). This network shows the first neighborhood (violett links) around four central nodes for the stock market indices Nikkei 225, DAX, NASDAQ 100, and BSE 200.

The highest link density is found in the local network around the Japanese index Nikkei 225 (red). The list-pages for the two asian markets (red: Japan, orange: India) are linked directly via one intermediate page while the pages for the German index DAX (blue) and the American index NASDAQ 100 are not connected directly to each other.

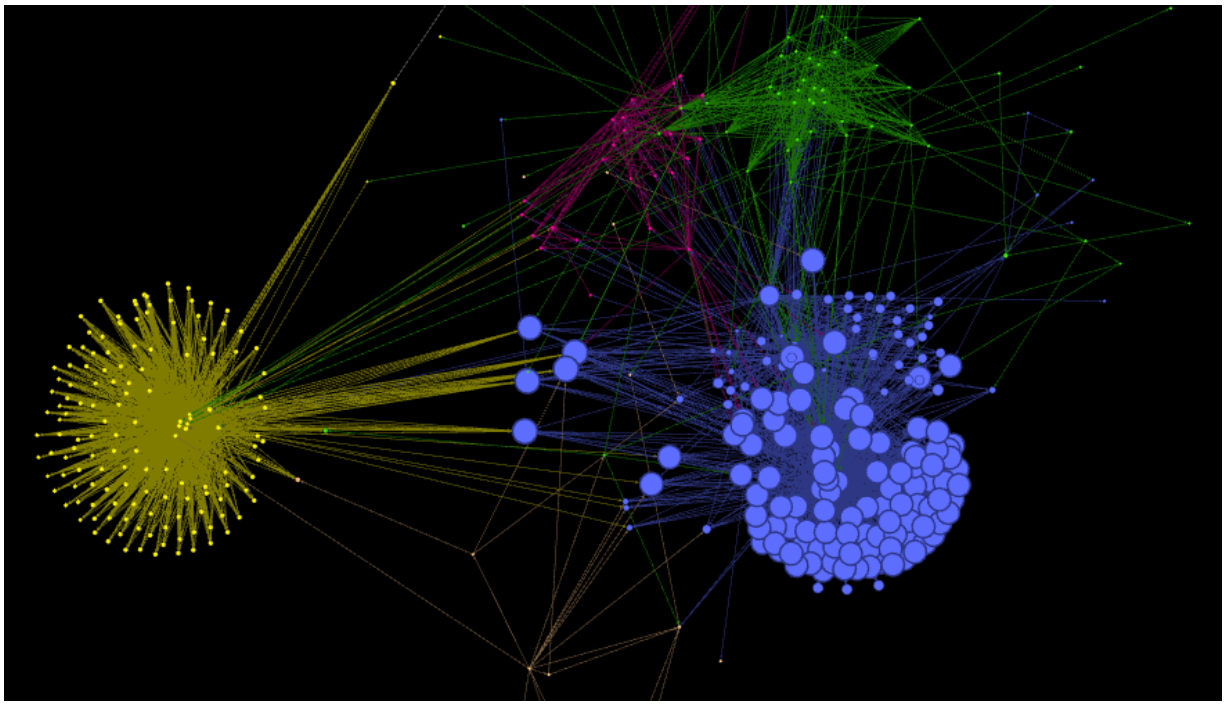


Fig. 8 b: Wikipedia can be used as a source for information about financial markets. Multiple Wikipedia projects in several languages are inter-connected by so called “inter wiki” links (geen). This network shows the first neighborhood (violett links) around four central nodes for the stock market indices Nikkei 225, DAX, NASDAQ 100, and BSE 200.

Figure 9 shows the access-rates per group for the Wikipedia pages representing the selected indices for 2009.

SHOW a Time Series Dashboard for 4 groups like we did in the relevance pape

Fig. 9:

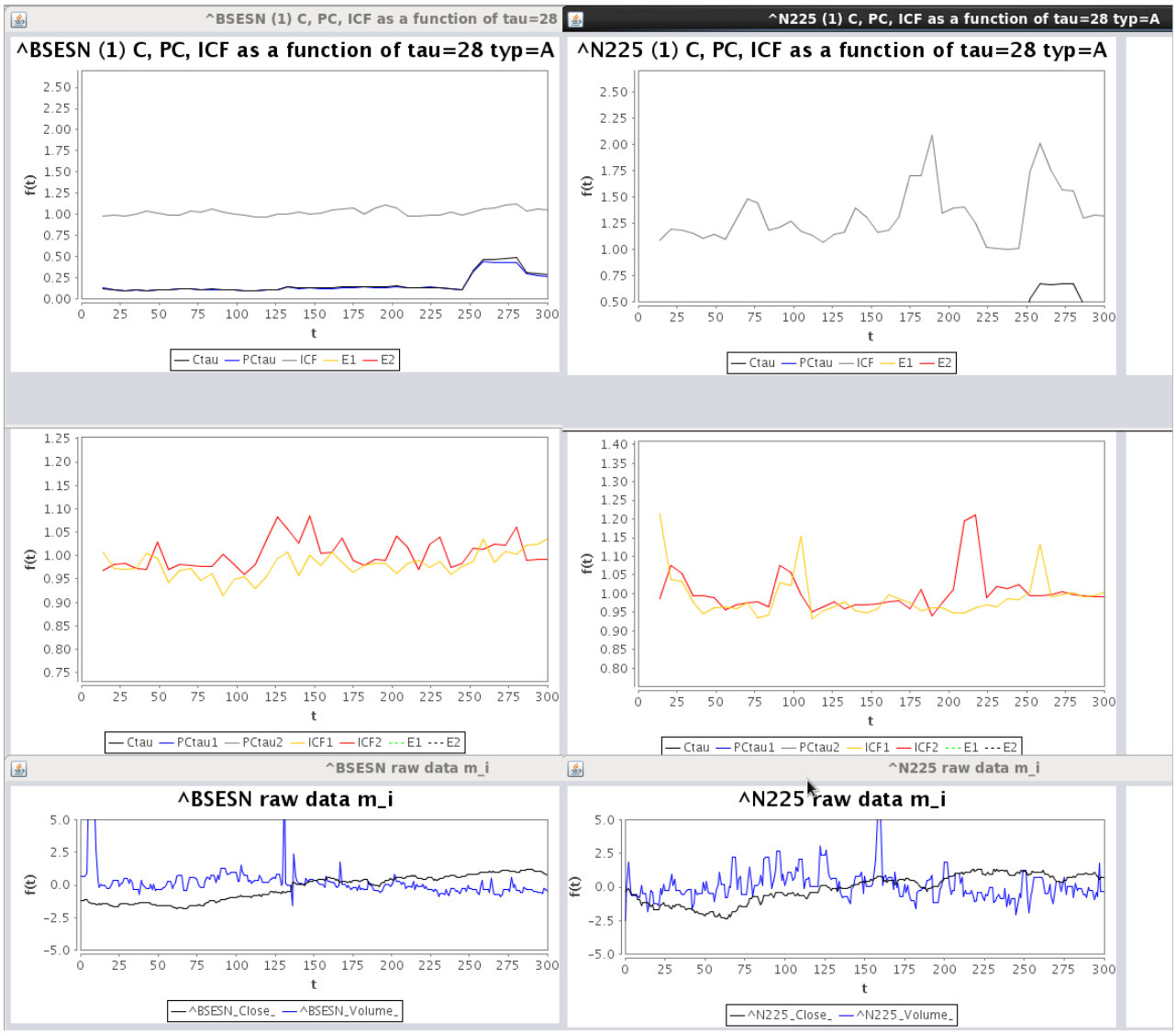
TRI for 3 years

Results

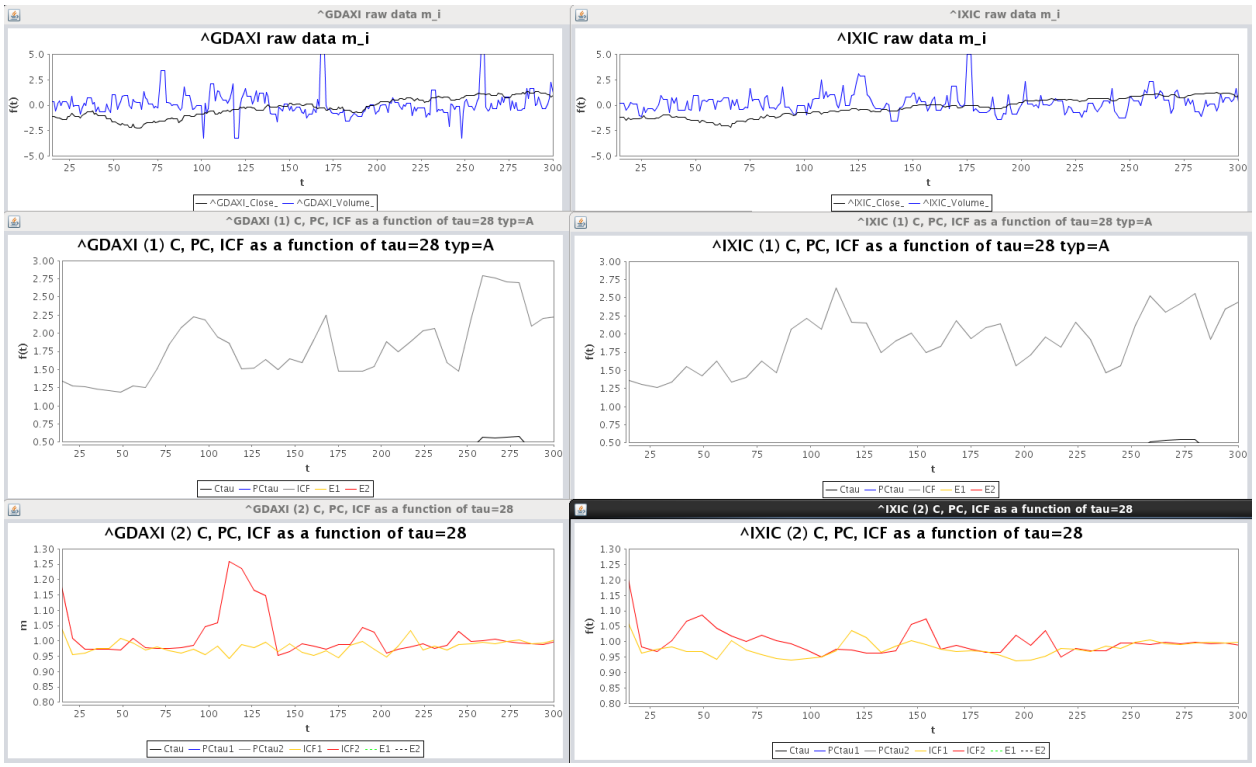
1.) We compare the **intra wiki correlations** (between the access rate time series of the core and the neighborhood in a local network of Wikipedia pages)

2.) We analyse the **meta correlation** between all Wikipedia pages representing a market with corresponding stock market data.

Top row: gray graph, analysis of **Intra-Wiki correlations**
Middle row: red / yellow graph, analysis of **Wiki-stock-market meta correlations**
Bottom row: raw data from stock market (**closing price** and **trading volume**)



Top row: raw data from stock market (**closing price** and **trading volume**)
Middle row: gray graph, analysis of **Intra-Wiki correlations**
Bottom row: red / yellow graph, analysis of **Wiki-stock-market meta correlations**



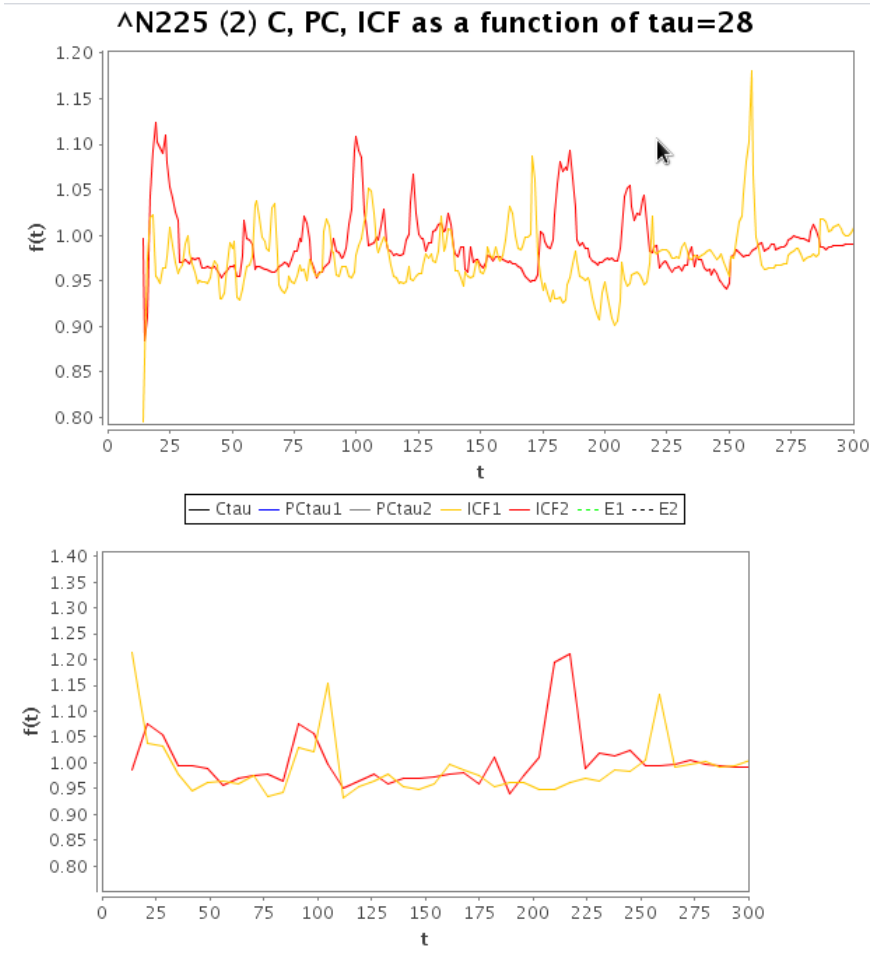
An increase in intra correlations (gray) comes before the increase in inter-correlations (red and yellow).

A clear signal in intra correlations depends on the volume of data available for an analysis. In case of BSESN and NASDAQ-100 the representation of the stock market index in Wikipedia is on a very low level (see fig. 7) this might cause the noisy plot in the intra correlations.

Test with higher resolution:

Meta-Correlation analysis: red/yellow graph, analysis of **Wiki-stock-market meta correlations** for the Nikkei Index.

Top panel:	overlapping sliding windows of length tau=28 days are shifted by dtau=1 day
Bottom panel:	tau = 28 days, dtau = 7 days



Intra-Correlation analysis: gray graph, analysis of **Intra-Wiki correlations** for the local network around the Nikkei Index.

Top panel:	overlapping sliding windows of length tau=28 days are shifted by dtau=1 day
Bottom panel:	tau = 28 days, dtau = 7 days

Meta correlation for markets

TRI-Correlation for markets

Dependency networks

How to test a dependency network?

Create a directed random graph
N = number of nodes
p = link probability
Find a random weight for all nodes (amplitude)

Find a random weight for all edges (frequency)
Export the graph
create a random time series for all nodes which consist of sine wave and some added noise
for all links:
calculated the contribution of the source on the destination mode (weight of the node)
increase the destinations amplitude on the frequency the link defines.

=> same frequenca froms clusters

Discussion

Do stock markets drive interest in financial topics in Wikipedia?

Our mid term study of Wikipedia access-count data has found that stronger correlated access-activity in Wikipedia might be used as an indicator of trend changes in stock markets. During periods of stronger trend changes in stock markets, the correlation in the measured access activity increases also. We looked at time series with a length of 300 days from January 1, 2009 to October 28, 2009.

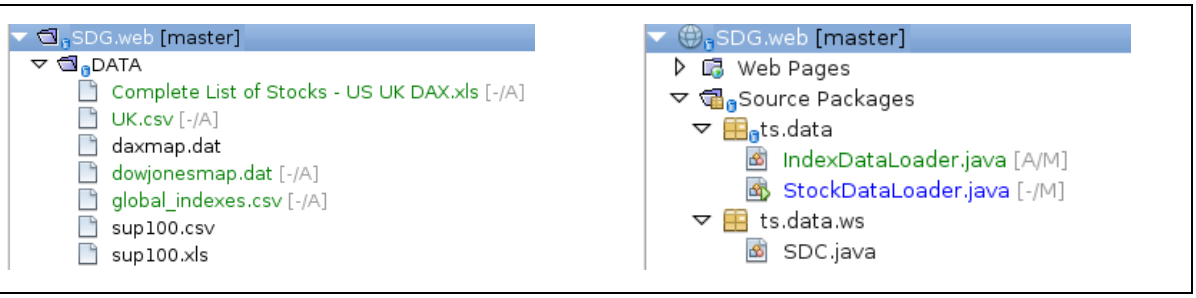
Conclusion & Outlook

Section 2: Hands on tutorial

Step 1: Collect data

1.1 Financial time series

In order to load the financial time series from Yahoo we need the **stock data loader tool**.



and the following list file which contains the URL of the central node in Wikipedia and the symbol of the indices to request data via Yahoo API:

```
# Wiki Link    Stock Symbol (Yahoo)    Index Name
# attention !!!! REPLACE (^) WITH (%5E)
http://de.wikipedia.org/wiki/Nikkei_225,%5EN225,Nikkei 225
http://de.wikipedia.org/wiki/SSE_Composite_Index,000001.SS,SSE Composite
http://de.wikipedia.org/wiki/DAX,FDAX.EX,DAX Performance
http://de.wikipedia.org/wiki/FTSE_350_Index,%5EFTLC,FTSE 350
http://de.wikipedia.org/wiki/S%26P_500,%5EGSPC,SuP500
http://de.wikipedia.org/wiki/BSE_Sensex,%5EBSESN,BSE SENSEX
```

The event series expansion will transform the raw data which is just a set of key value pairs. We fill up all gaps with the last known value from last trading days.

1.2 Local networks from Wikipedia

We prepare a study descriptor which contains the entry pages. For large study it is convenient to import a prepared list from a crawl definition file which contains the language code of the chosen wiki project in the first line and names of selected entry pages, one per line.

```
de
Nikkei_225
SSE_Composite_Index
DAX
FTSE_350_Index
S%26P_500
BSE_Sensex
```

Selection of entry pages:

CN lang	CN pagename
de	Nikkei_225
de	SSE_Composite_Index
de	DAX
de	FTSE_350_Index
de	S%26P_500
de	BSE_Sensex

1.3 Wikipedia access-rate time series

...

Step 2: Checkout the Hadoop.TS project from Github

...

Step 3: Create and visualize random time series

Hadoop.TS offers the following time series generators:

1. SineWaveGenerator,
2. DistributionGenerator,
3. FFTPhaseRandomizer

SineWaveGenerator:

```
Messreihe mr = TSGenerator.getSinusWave(fre, time, SR, ampl);
```

We create a test time series with a daily cycle and one with a weekly cycle.

fre	frequency in Hz f_{day}=1 / 86400 Hz $f_{week} = f_{day} / 7$
time	length in seconds
SR	sampling rate SR_{hourly}=1 / 3600
ampl	amplitude of the signal

DistributionGenerator

```
Messreihe mr = TSGenerator. ...
```

FFTPhaseRandomizer

```
Messreihe mr = TSGenerator. ...
```

Step 4: Group the data and define sliding windows, binning etc.

Step 5: Calculate the time dependent measure for choosen groups

Step 6: Export and visualize the results

Section 3: Text blocks for other documents

Example data

Our example data set contains the click count time series for a local network, which is a set of pages linked to a central node which is the page of our interest.

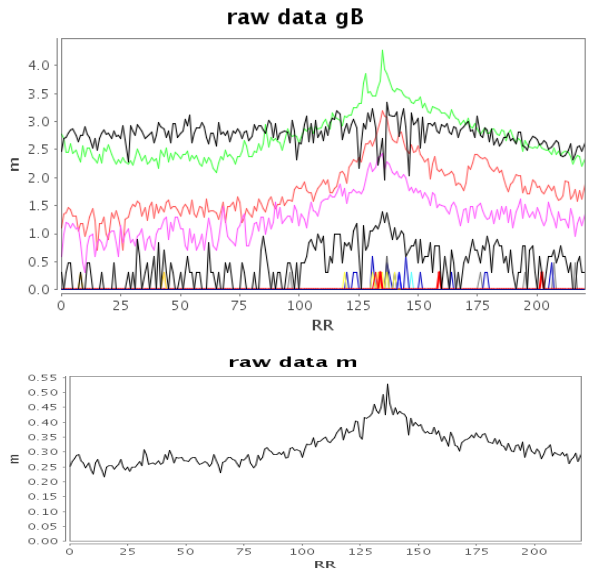


Fig. 1: For all pages which have an inter-wiki link to the central page ***Illuminati_(Book)*** the logarithm of the access count time series (aggregated by day) are shown in the top panel. The bottom panel is the average value for all time series from group gB showing the total interest in this topic during the first 220 days in 2009.

The reference time series ***m*** describes the overall interest in the selected topic, here it is the wikipedia page *Illuminati_(Book)* (see Fig. 1). We are interested in the influence of this set of pages on their neighborhood. All pages in group ***gB*** have wiki links to other pages which define group ***gA***. How strong is the core of this local network influencing the local neighborhood? For comparison we also show the raw data for group ***gA*** in Fig. 2.

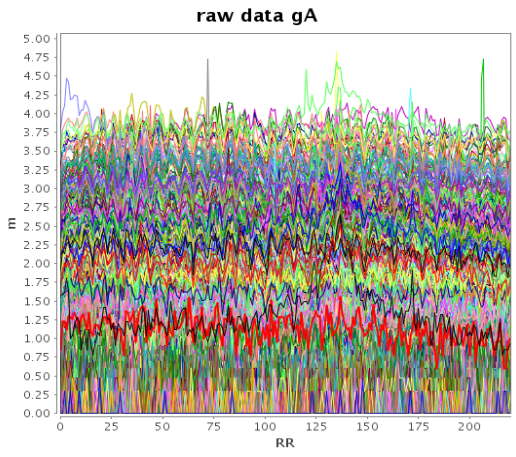


Fig. 2: For all pages which have a wiki link to any of the core pages around the central page ***Illuminati_(Book)*** the logarithm of the access count time series (aggregated by day) are shown for the first 220 days in 2009.

Quantify completeness: Count all non existing pages per group

In many wiki projects not all pages exist already. So it might be an indicator of completeness to compare the ratio of non existing pages beside the numbers of links. On a high level, there might be some links available but the detail pages are not created. This indicates general interest but less activity.