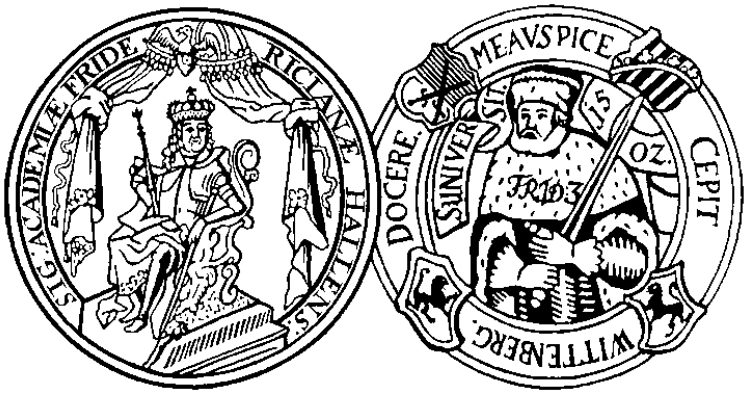


Martin-Luther University
Halle-Wittenberg

Bachelor Thesis

Reconstruction of complex networks based on event time series

---

Author:
Arne Böker

Supervisor:
PD Dr. Jan W. Kantelhardt

Institute of Physics, Faculty of Science II
Martin-Luther-University Halle-Wittenberg

September 25, 2012

Contents

1 Introduction

In the course of the past decade, the term 'Web 2.0' has come into existence. It is used to describe recent functions of the Internet, which allow the users to communicate faster than before through short messages, commenting, blogs, *et cetera*. Especially in combination with the novel development of mobile Internet, some Web 2.0 media have taken over a large part of our everyday lives and of our interaction with other people.

This development combines well with the field of socio-economic physics, which is not very old itself. The new working group SOE of the German physical association DPG has been founded in 2001 and gained the status of a division in 2009. Web 2.0 provides many large social networks with large amounts of data. These networks are of great interest for sociology and for sociophysics, which uses methods of network physics (graph theory) to describe and analyze these social networks theoretically. We will use this approach too, and take a look at the online encyclopedia Wikipedia as an example for an online social network. Its users interact by reading and editing encyclopedia entries and by discussing them on special 'Talk' pages, making Wikipedia an encyclopedia (content network) and a communication network at the same time.

A network consists of nodes (vertices) and edges. If the nodes have a property which is changing with time, these changes can be compared for a pair of nodes by mathematical methods. For changes behaving similarly, we draw an edge between the nodes. Doing this for many pairs of nodes, we can reconstruct a network from the 'property time series'. The network can usually also be constructed as a weighted network because the mathematical comparison methods give continuously distributed results, and possibly as a directed network because the methods may yield interaction directionality. There are many social networks and various mathematical methods to use, so the idea can be implemented in very different ways. As already mentioned, we are using Wikipedia as a social network. Each Wikipedia page is a node, time stamps of page edits form the time series. These time series are compared using the Event Synchronization method, which will be introduced in section 3. We will end up with two networks, one 'static network' which uses the links between Wikipedia pages (coded in the page content) as edges, and one 'dynamic network', which is our reconstructed network from the page edit time series[1].

In order to work with Wikipedia's structure and not only with the encyclopedic content, it is necessary to understand basic characteristics of Wikipedia. The second chapter, after this introduction, gives an overview of the relevant features of Wikipedia and of the dataset we are working with. Even more important is a deeper understanding of the Event Synchronization algorithm. Event Synchronization has been developed just ten years ago and has not been used very widely yet, so the third chapter introduces the method itself before treating its general properties and our usage of it. Afterwards we can step to network reconstruction itself. Chapter four shows our calculated results and some network graphs. In the final chapter we will give a round-up of what has been done for this thesis and a thematically similar parallel thesis as well as an outlook to possible future work regarding network reconstruction and Wikipedia.

[1]Note that both networks may change with the progressing development of Wikipedia. They are both snapshots of a limited time.

2 About Wikipedia

We are using data collected from the online encyclopedia Wikipedia between 01/2009 and 10/2009. This section gives a short introduction into relevant features of Wikipedia and our dataset.

2.1 What is Wikipedia?

The name *Wikipedia* is a portmanteau of the hawaiian *wiki* meaning 'quick' and *encyclopedia*. Wikipedia is, in its own words, 'a free, collaboratively edited, and multilingual Internet encyclopedia' [Wikipedia2012a] founded by Jimmy Wales and Larry Sanger in 2001. 'Free' means that the content is accessible for anyone and the organization behind Wikipedia (the Wikimedia Foundation) is a non-profit organization, 'collaboratively edited' means that any user is allowed to edit or create articles - which is vital for us: Because of this collaborative character we can regard Wikipedia as a social network. Wikipedia is 'multilingual' not as a single encyclopedia in several languages but as multiple encyclopedias. There are 285 Wikipedias in different languages or dialects with very variable numbers of articles. The largest version is the English Wikipedia with more than four million articles and 24 million other pages [Wikipedia2012b] (see also section 2.3), the German, French and Dutch Wikipedias also contain more than a million articles apiece. Other more exotic language versions can have less than 100 articles, in some cases these articles are only automatically generated. The contents of different Wikipedias are generally developed independently from each other, making each language version an encyclopedia of its own.

2.2 Characteristics of our data

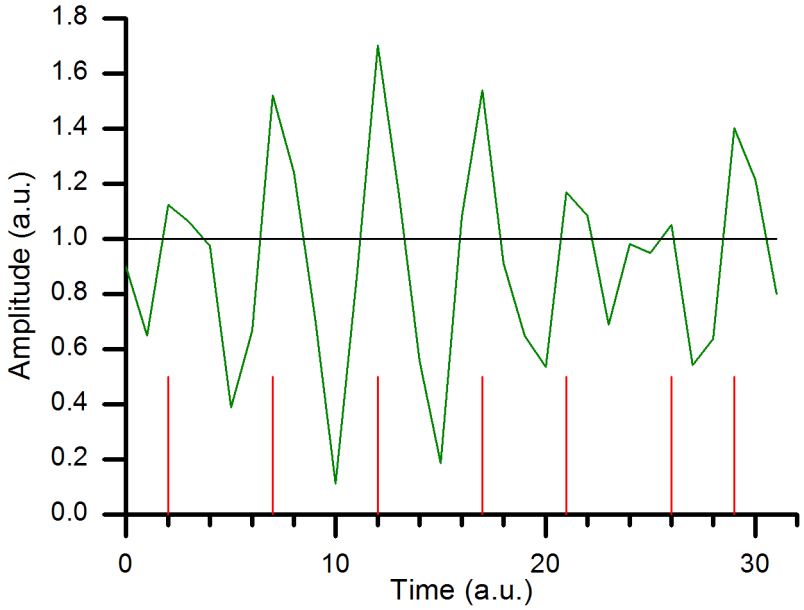


Figure 1: Continuous time series plot $f(t)$ (blue line) with a threshold at 1.0 and events (maxima exceeding 1.0) marked as spikes.

Our dataset was collected by Domas Mituzas from 2008 to 2010 [Mituzas2010] and prepared in 2009 by Lev Muchnik, who made it available to us. Beside other data, it contains date and time for every edit that happened during a total time of 40 weeks. This way we get an edit time series for each Wikipedia page to work with. These time series are different from usual time series in format (see Figure 1 - the red spikes visualize the event time series format), so to quantify synchronicity we need a suitable algorithm, which we will introduce in section 3.1.

2.3 Namespaces

Wikipedia pages are grouped by their general functions into *namespaces*. They can be recognized by title prefixes on Wikipedia pages. For example the page 'Category:Physics' is part of the namespace Category while the page Physics belongs to the namespace Main[2]. There are 16 main namespaces and 14 respective 'talk' namespaces for discussions among editors (see Table 1 on page 4 for details), each assigned a number from -2 to 15 and 100 to 111.

Among these namespaces, the *Main* namespace (0) is the most interesting for us as it contains the Wikipedia articles. Looking at the distribution of namespaces in Wikipedia, it becomes clear that it is the most common too, but not dominant enough to ignore the other namespaces.

Figure 2 shows an example namespace distribution: The complete bar signifies the 1000 most actively edited pages from the Hebrew Wikipedia, different colours stand for the different namespaces. The large blue bar is namespace 0, but 1, 2, 3 and 4 still make up about one fourth of the total pages. The numbers can vary: some Wikipedias consist almost solely of articles, in others they make less than 50%. Because of this discrepancy it is necessary to focus on articles and exclude all other pages, so our results will always refer to those.

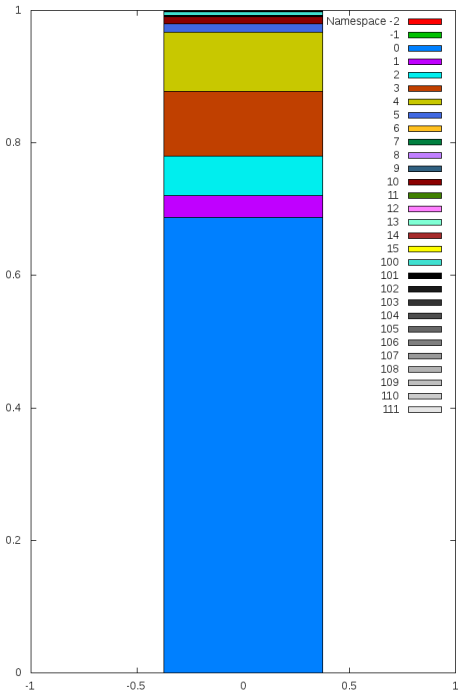


Figure 2: An example namespace distribution from the Hebrew Wikipedia in 2009. The large blue bar is the namespace 'Main', taking up 70% of all pages.

[2]The 'Main' namespace is not displayed in Wikipedia.

Table 1: Overview over the namespaces of Wikipedia with index numbers. Wikipedia itself only lists namespaces -2 to 101, 107 and 108, with the last two namespaces corresponding to our 102 and 103 [Wikipedia2012d]. The remaining namespaces are probably outdated, but have to be listed here because they are part of our dataset.

Index	**Title**	**Index**	**Title**
0	Main	1	Talk
2	User	3	User talk
4	Wikipedia	5	Wikipedia talk
6	File	7	File talk
8	MediaWiki	9	MediaWiki talk
10	Template	11	Template talk
12	Help	13	Help talk
14	Category	15	Category talk
100	Portal	101	Portal talk
102	Book	103	Book talk
104	Word	105	Word talk
106	Text	107	Text talk
108	Quote	109	Quote talk
110	News	111	News talk
-1	Media		
-2	Special		

2.4 Nodegroups

To construct a network, we need vertices (nodes) and edges. In our networks Wikipedia pages will serve as nodes. Each node is assigned a time series, the time series' calculated synchronicities give the weight for the edges. This way we construct a weighted dynamic network, which then is compared to the static (unweighted) link network.

Since our computational power does not suffice to calculate the synchronicity of all pairings between several millions of Wikipedia pages, we choose *nodegroups* with a maximum of a few thousand nodes. Until now we have performed our analysis on 10 such nodegroups, all of which are static networks centered around one article. Two of the central articles are about the financial indices **DAX** and **S&P 500**, another two are **lists of cities** in Germany and the United Kingdom. Note that German cities generally have more than 2,000 inhabitants and UK cities more than 10,000. Also city status in the UK is granted by the monarch and because of this harder to achieve than in Germany [Wikipedia2012c]. So there are a lot more cities in Germany than in the UK. The remaining six nodegroups are selected cities of different size in Germany and the UK, namely **Heidelberg, Berlin, Oxford, Birmingham, Sulingen** and **Bad Harzburg im Harz**. The articles regarding Oxford, Birmingham, the UK city list and the financial articles are from the English Wikipedia, the five articles referring to German cities are part of the German Wikipedia. The nodegroups can partly be contained in other Wikipedias because links between different language versions are possible. Especially links between two articles about the same topic in different languages are common.

3 Event Synchronization

We apply a method called *Event Synchronization* (short form ES), a fast and simple algorithm to calculate the synchronicity of two time series. Event synchronization was first proposed by Quiroga et al. in 2002 and applied to EEG signals [Quiroga2002] and later to monsoonal rainfall [Malik2010, Malik2011]. Since hardly any work beyond studying neural spike trains (following Quiroga's work) and monsoonal rainfall has been done using the method and especially investigating its properties, we need to understand some features of this method before using it.

3.1 Why apply event synchronization?

Most wikipedia pages are edited less frequently than once a day, which makes typical methods of time series analysis unusable. We use a format called *event time series*, where not the numbers of events per time step are given (which would be the usual format of a discrete time series) but instead the exact times of every single event. It can be visualized in form of the red spikes in Figure 1 on page 2, in this case we see seven events at time points 2, 7, 12, 17, 21, 26 and 29.

Event synchronization is a method which requires little activity to give reasonable results and makes use of this type of time series. In previous works applying ES [Quiroga2002, Malik2010] it was necessary to define *events* out of the behaviour of time series, usually values exceeding a certain threshold were considered an event (Figure 1). As described, our data already have this format, making ES the best method to use.

3.2 The algorithm of event synchronization

Let i and j be two event time series with events occuring at times t_l^i and t_m^j. The total numbers of events in these time series be s^i and s^j respectively, so l ranges from 1 to s^i and m from 1 to s^j. The basic idea is to calculate how many times an event in time series i precedes an event in time series j or vice versa closely enough to be considered synchronous.

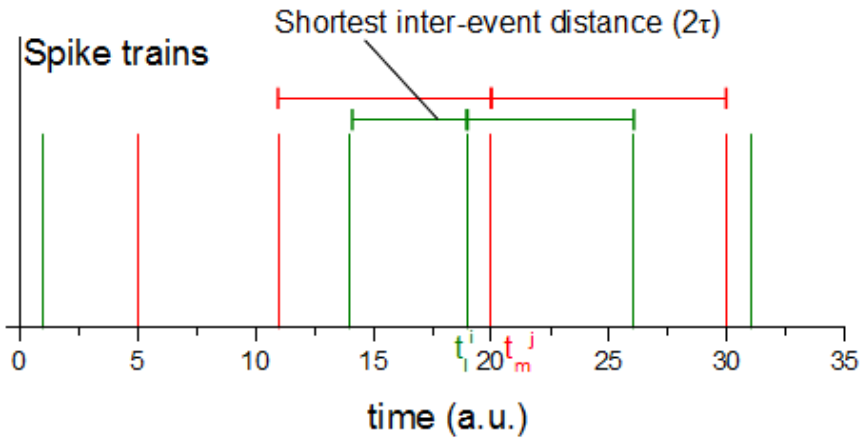


Figure 3: Two event time series i (green) and j (red) with two freely chosen events and the shortest distance to another event, in our case called 2τ. The two events will be counted as synchronous because their distance is less than τ.

To quantify the term 'closely enough' we define a time lag τ_{lm}^{ij} as

$$\tau_{lm}^{ij} = \frac{1}{2} \min \left(t_{l+1}^{i} - t_{l}^{i}, t_{l}^{i} - t_{l-1}^{i}, t_{m+1}^{j} - t_{m}^{j}, t_{m}^{j} - t_{m-1}^{j} \right) \tag{1}$$

The definition of τ can vary depending on the context. In our case the above definition is the most suitable one, but Quiroga also mentions a global time lag τ_g, which is independent from the choice of l and m and can be more useful in other cases. It can be defined as a minimum or average of all τ_{lm}^{ij}, keeping in mind that the parameter is meant to avoid double-counting and thus should be sufficiently small.

Now, if the difference between t_l^i and t_m^j is less than τ_{lm}^{ij}, the two events are considered synchronous. In the algorithm this is expressed by a quantity J defined as

$$J_{lm}^{ij} = \begin{cases} 1 & \text{if} \quad 0 < t_l^i - t_m^j < \tau_{lm}^{ij} \\ \frac{1}{2} & \text{if} \quad t_l^i - t_m^j = \tau_{lm}^{ij} \\ 0 & \text{else} \end{cases} \tag{2}$$

Note that t_l^i must occur before t_m^j for $J_{lm}^{ij} > 0$, so $J_{lm}^{ij} + J_{lm}^{ji}$ can only be 0 or 1, reflecting that every pair of events can only be regarded asynchronous or synchronous.
In the next step all J_{lm}^{ij} are cumulated:

$$c(i|j) = \sum_{l=1}^{s^i} \sum_{m=1}^{s^j} J_{lm}^{ij} \tag{3}$$

Adding the parameters for both directions, $c(i|j)$ and $c(j|i)$, leads to a result Q_{ij} which is called **strength of event synchronization**, subtracting them gives a **delay** value q_{ij}:

$$Q_{ij} = \frac{c(i|j) + c(j|i)}{\sqrt{s^i s^j}} \tag{4a}$$

$$q_{ij} = \frac{c(i|j) - c(j|i)}{\sqrt{s^i s^j}} \tag{4b}$$

The denominator $\sqrt{s^i s^j}$ normalizes the values, so the strength of ES ranges from $Q_{ij} = 0$ to $Q_{ij} = \frac{\max(s^i, s^j)}{\sqrt{s^i s^j}}$ (the maximum value is 1 and can only be reached if $s^i = s^j$), while the delay value ranges from $q_{ij} = -Q_{ij}$ to $q_{ij} = +Q_{ij}$. $Q_{ij} = 1$ means complete synchronization (i.e. for every event in time series i, there is a synchronous event in j and vice versa), $Q_{ij} = 0$ complete desynchronization, the meaning of which we want to discuss in section 3.3. $q_{ij} = -Q_{ij}$ means that for all pairs of synchronous events, the event in time series i precedes the one in j, $q_{ij} = +Q_{ij}$ the opposite direction. If $q_{ij} = 0$, no direction is favoured.

3.3 General properties

To interpret the calculated results, we need to learn more about the properties of ES. It is especially important to know which value is calculated for statistically independent time series so we can identify unsignificant links in our results. We performed two experiments with randomly generated time series: the first one simulates a transition between identical and independent time series, the second one covers the case of different densities[3].

Experiment 1 - Transition from identical to independent

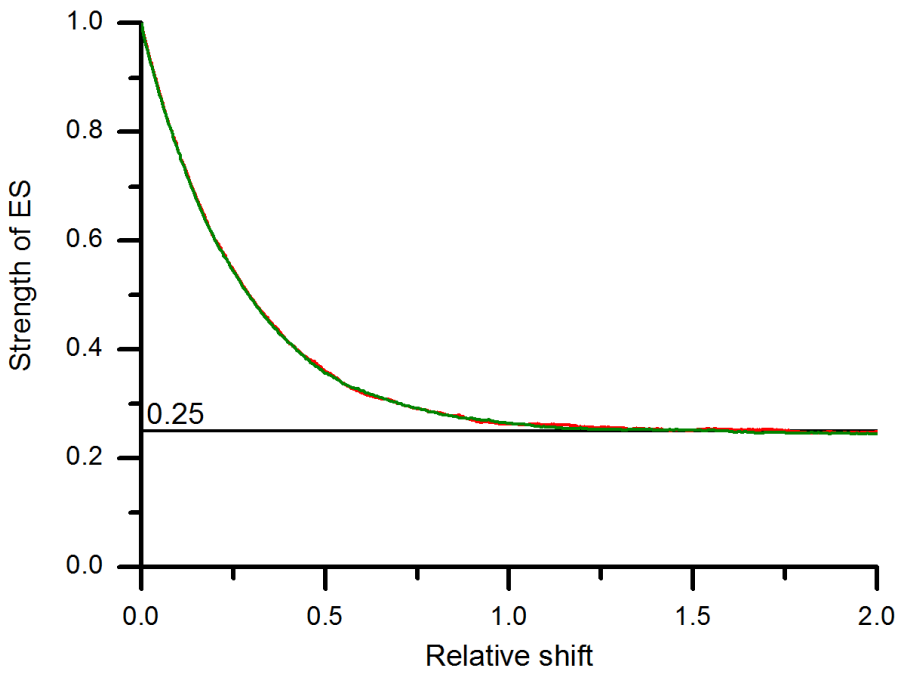


Figure 4: Strength of ES with respect to the relative shift, which is the quotient of time shift Δt and average inter-event time (reciprocal density). The two colours correspond to two different densities. The results start at 1 for identical time series, then decay to a value around 0.25. For a relative shift greater than 1.5 there is no more significant decrease.

[3] *Density* should be the number of events divided by total time. Since the total time is equal for all time series in our data set, "density" will be used synonymously with "number of events" or "event count".

If event synchronization is calculated for two identical time series with random events, the result is 1 because for every event in one time series there is a synchronous match in the other time series at the exact same time. Now we successively shift one of these time series by a small time Δt, which means every event t_l is replaced by $t_l + \Delta t$ and re-calculate event synchronization.

As shown in Figure 4, the strength of event synchronization does not reach 0, but rather stagnates close to 0.25. If we subtract 0.25 from the result values and use a logarithmic scale for the vertical axis, the graph is a straight line (Figure 5), so the decay is exponential. It follows the function

$$Q(t) = 0.75e^{-4 \cdot \Delta t/T} + 0.25 \tag{5}$$

where $\Delta t/T$ is the relative time shift between the time series and Q the strength of ES.

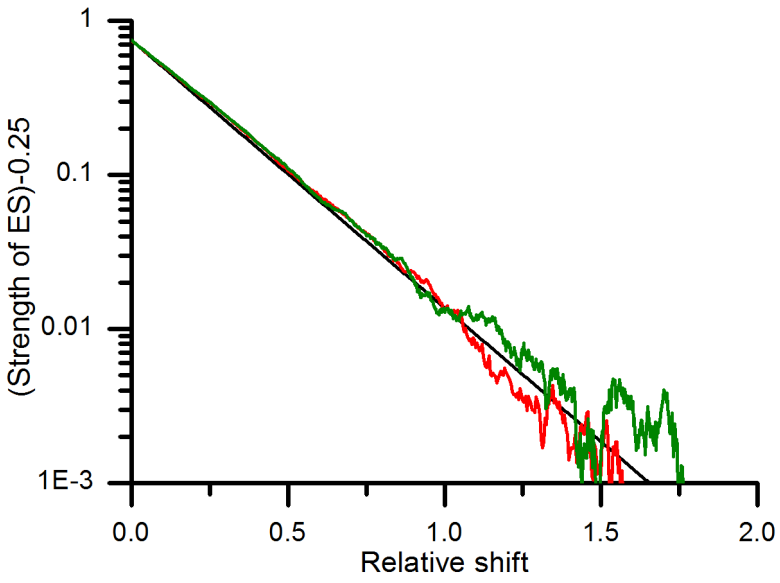


Figure 5: Strength of ES minus 0.25 with respect to the relative shift, in a semi-log plot. The two colours correspond to two different densities. The straight line displays the fit function $y = 0.75e^{-4x}$.

Why do we get this function?
In generating of the time series, every point of time is equally likely to contain an event. This results in a Poisson process, so the distances between two events follow the exponential distribution

$$p(t) = \frac{1}{T}e^{-t/T} \tag{6}$$

where T is the mean inter-event time (distance).

Looking at one pair of events $(t_l^i,\ t_m^j)$, there are four distances forming the value τ (Eq. (1)), all of which are independent from the time shift between t_l^i and t_m^j. The probability for this pair to be synchronized is equal to the probability for $\delta = |t_l^i - t_m^j|$ to be smaller than each of these four distances:

$$p_{sync}(\delta) = (p(\delta))^4 = \left(\frac{1}{T}e^{-\delta/T}\right)^4 \tag{7}$$

This leads to $Q \propto e^{-4\delta/T}$, because Q counts the number of synchronized event pairs, which is obviously proportional to the probability for a single pair to be synchronized.

With this model we are able to explain the exponential part of the fit function (Eq. (5)), but not the vertical offset of 0.25. The reason is that we only consider the behaviour of initially synchronized event pairs $(t_l^i,\ t_m^{\mathbf{j=i}})$. If the time shift is large, the synchronization of these two events vanishes, leading to the overall $e^{-4t/T}$ function, but for a large time shift t_l^i will synchronize with the next event $t_m^{j=i+1}$, leading to the observed aberrations. This effect causes the strength of ES to stay above 0, empirically we have already seen the 0.25.

Experiment 2 - Effects of density differences
The above experiment has shown that the strength of ES between two independent time series with equal density is close to 0.25. But what happens if the densities are not equal?

Figure 6 shows the strength of ES (blue crosses) with respect to the ratio of densities on two logarithmic axes. The function peaks at $x = 1$, which means equal density, nearly reaching the expected 0.25 (The more exact result is 0.238), then decays symmetrically[4]. For a density ratio of $x = 100$ the synchronization is as low as 0.04 and not to be expected to increase[5].
The red line in Figure 6 shows an empirically determined function which fits the values well:

$$Q_{rand} = \frac{0.7}{\left(5 + \max\left(x, \frac{1}{x}\right)\right)^{0.6}} \tag{8}$$

where x is the density ratio and $\max\left(x, \frac{1}{x}\right) \geq 1$. For any two given densities we can either calculate the strength of ES of independent time series directly or approximate it using the above equation, which is more convenient.

[4]Note that the axes are logarithmic, so the symmetry is not $f(1-x) = f(1+x)$, but rather $f\left(\frac{1}{x}\right) = f(x)$

[5]Due to symmetry density ratios 100 and 0.01 result in the same strength of ES. In what follows, we will only mention ratios above 1, keeping in mind that results for the reciprocal ratio behave the same.

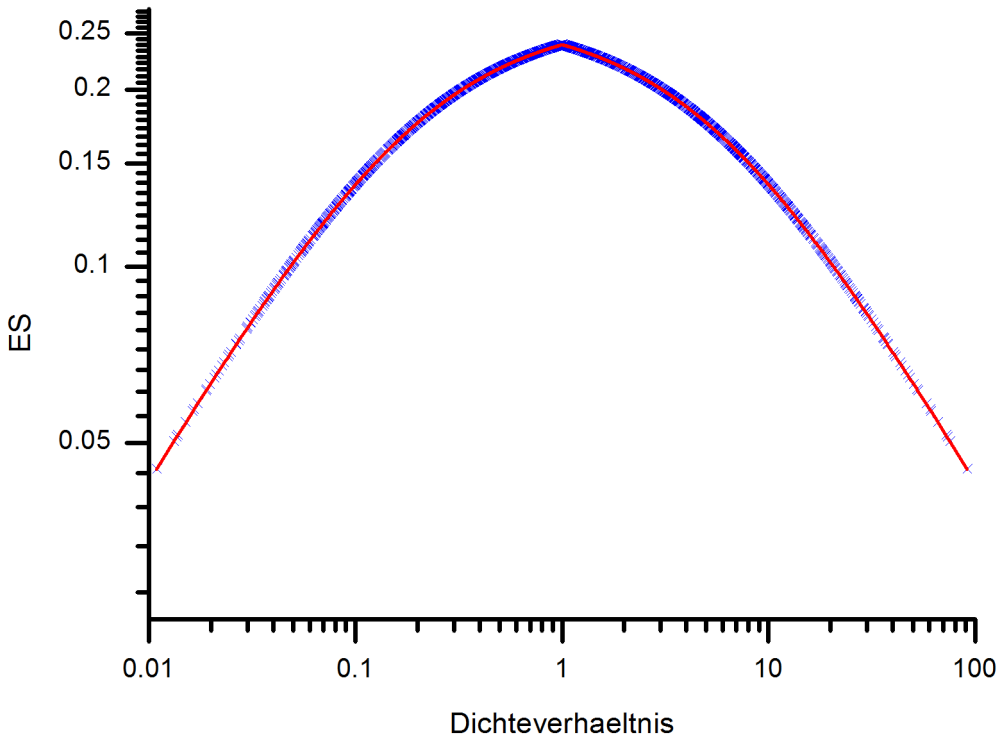
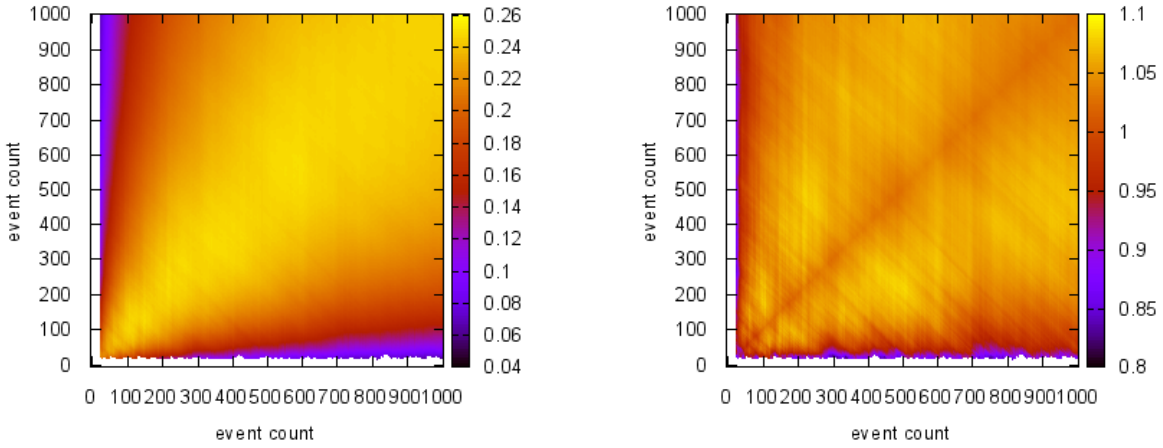


Figure 6: Event synchronization of independent time series with respect to the ratio of densities on log-log scales. The line shows the fit function $Q = \frac{0.7}{(5+x)^{0.6}}$ where $x \geq 1$ is the density ratio.

This result helps us to quantify the influence of density differences. For example for two time series with very different densities, a strength of ES around 0.2 can signify a strong synchronicity as it is higher than this random result, while the same result means no synchronicity if the time series have similar densities. We are going to use this as a calibration, dividing the calculated ES result by the respective value from Eq. (8). We are then able to define thresholds to do further filtering, for example taking only node pairs with a calibrated result $Q_{cal} = \frac{Q}{Q_{rand}} \geq 1.0$ into account for network reconstruction. This way we spare out insignificant links based on random synchronization. If we choose a higher limit for Q_{cal}, we can similarly sort out weak and probably still insignificant links. As a side note it is to mention that the original results Q are limited to 1, but the calibrated results Q_{cal} are not. Most commonly the values are no larger than 4, but in theory they can be arbitrarily large.

Because this experiment has become quite important through its use as a calibration, we want to take another look at the results, now taking into account not only the ratio of densities, but also the absolute densities to see if they change the results.



(a) Calibration values. The axes represent event counts in two time series, colours the strength of ES.

(b) The same results divided by the approximate random value from Eq. (8)

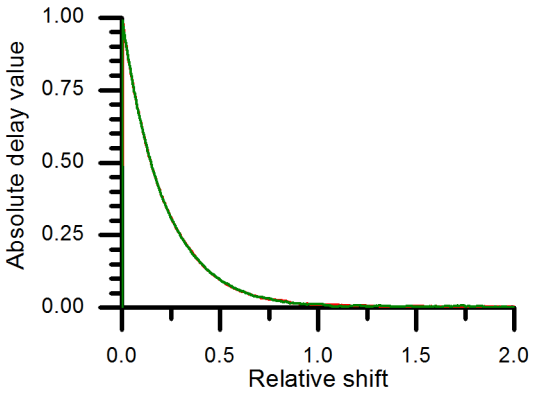
Figure 7: Colour plot of the calibration values.

Figure 7 shows the calibration values for all combinations of event counts from 30 to 1000 in two independent random time series with identical duration. The axes represent the event counts while the result value is coded in the colours.

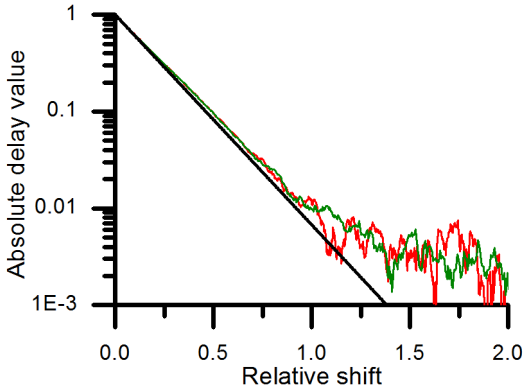
Figure 7(a) shows the raw result values. The symmetry to the diagonal (which corresponds to a density ratio of 1) is clearly visible, with the maximum value again near 0.25. The colour changes do not deviate from straight lines through the origin, which means they appear always at the same density ratios. If the result value depended on the absolute density, we would see deviations from the straight lines, so this diagram answers the question asked before: The calibration is independent from the event counts, it only depends on their ratio.
Furthermore the diagram contains vertical and diagonal streaks. These are effects from time series generation because the random number generator does not produce sufficiently independent numbers. Though visible, this effect does not influence the observations because we are only describing them qualitatively. Further calibration values are always calculated individually for each pair of time series, so the exact numbers used in this figure will not appear again.
Figure 7(b) shows the result values divided by the approximation from Eq. (8). Though the mentioned streaks are even more clearly visible, the qualitative behaviour is still recognizable: The quotient is 1 on the diagonal, which means the fit function is equal to the actual results here, then for density ratios from 1.1 to 5 the approximation is slightly higher than the results. For higher ratios the approximation stays below the results, but still reasonably close. The deviation can be expected to increase further for very high density ratios, but these ratios are not to be expected in this work, so the function provides a legitimate approximation.

3.4 Properties of the delay value

Now that we understand the basic behaviour of the strength of ES Q, we can perform the same experiments for the delay value q (see Eq. (4b)). Of course, calculating the mean value $\langle q \rangle$ for independent time series is not necessary because it can only be 0. Independence of two time series implies that there is no preferred direction of delay barring noise. So we consider the absolute value $|q|$ instead of q.



(a) Dependence of the delay value on the relative shift on linear scales. For high shifts the values are approximately zero.



(b) The same with a logarithmic y-axis. The black line shows the function $y = e^{-5x}$, which seems to be a reasonable approximation.

Figure 8: Results of Experiment 1 for the delay value

Experiment 1 - Transition from identical to independent
As before, in the first experiment we calculate ES on two identical time series, then successively shift one of them and recalculate ES. Figure 8 shows the results of this experiment. As before (see Figure 4 on page 7), the axes represent the relative shift and the result value (here $|q|$ instead of Q), and the red and green lines are the results for two different average inter-event times.
In Figure 8(a) we can see that the result tends towards 0 for high shifts, which was not necessarily expected because $|q|$ can be anywhere between 0 and Q ($= 0.25$), but is not surprising either because the time series are nearly statistically independent at that point. A mean value of $|q| = 0.25$ would signify that there is always a clear direction of the delay (but not always the same direction because the average q is still 0), which is counter-intuitive.
Figure 8(b) shows the approximately exponential behaviour, like we already know it regarding the strength of ES. The black line in the figure is the graph of $q(t) = e^{-5t/T}$, which is a reasonable fit for low shifts. The function can be understood as an overlap of the decay function for Q and an additional effect of the relation between q and Q: For very low shifts, all pairs of events have the same direction of delay as long as they are still synchronized, thus $|q|$ is close to its maximum: $|q| \approx Q$. Then some pairs of events start to newly synchronize, as described previously in explanation of the aberrations from $Q = e^{-4t/T}$ behaviour. These newly synchronized pairs' delay direction is opposed to the general direction, which is given by the shifting direction between the time series, so they

decrease the result instead of increasing it like in the case of Q, leading to a fifth $e^{-t/T}$ factor in place of the 0.25 summand. As before, this explanation is limited to not too high shifts, as some effects are not taken into account, especially the newly synchronized pairs changing their respective delay directions after some time or synchronization of events with greater distance.

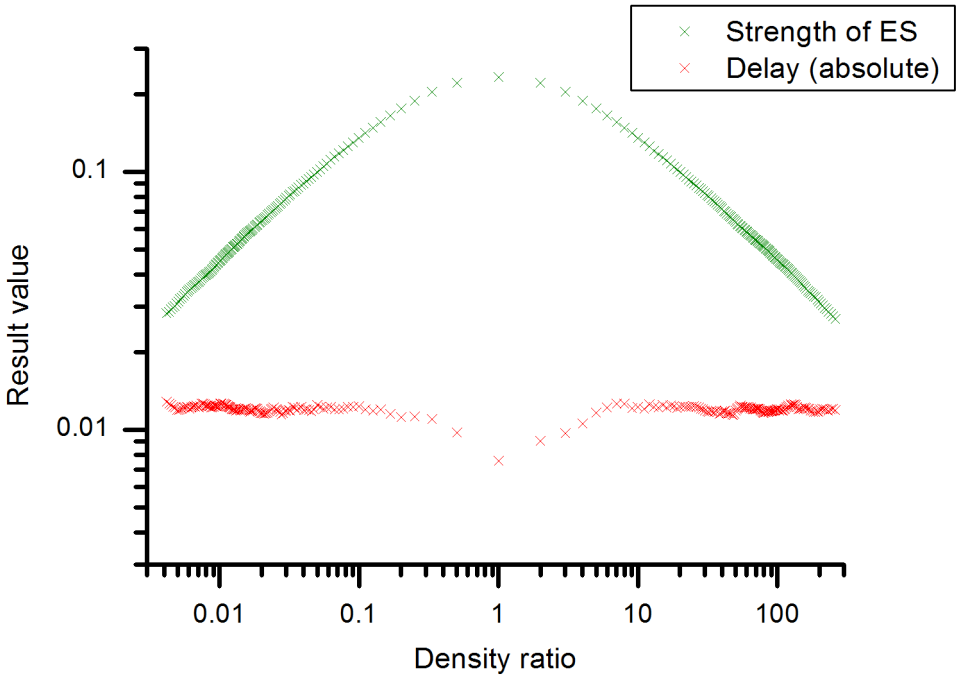


Figure 9: Calibration curve for strength of ES and absolute delay value. The delay results are significantly lower and mainly constant, except for a dip at a density ratio of 1. The delay results can be considered a fluctuation around zero, which increases with higher density ratios due to lack of data points, but also decreases together with the strength of ES.

Experiment 2 - Effects of density differences
In experiment 2 we calculated ES for two independent time series with given densities. Figure 9 shows the results in a log-log plot like before. The top curve is the calibration of the strength of ES which was calculated using the same method (see Figure 6 on page 10). The bottom curve is the result for the delay value $|q|$. Apparently it behaves differently from Q: While the symmetry to a density ratio of 1 is still there, the results increase instead of decreasing close to 1, and further away they stagnate.

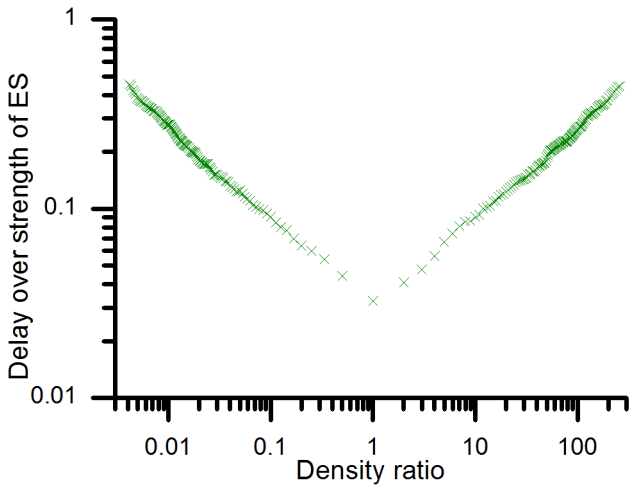


Figure 10: Calibration curve for the delay value divided by strength of ES. The values increase monotonously with increasing density ratio.

The calculated value $\langle |q| \rangle$ represents a fluctuation of q. We assume this fluctuation to depend on the number of countable event pairs, which is reasonably high for time series with equal densities, but clearly lower for higher density ratios because then one of the time series contains only few events. For lower numbers of event pairs the fluctuation increases, but the increase is countered by the decreasing strength of ES. To prove this explanation, we divide the delay by the strength of ES, expecting a monotonous increase of q/Q with increasing density ratio.

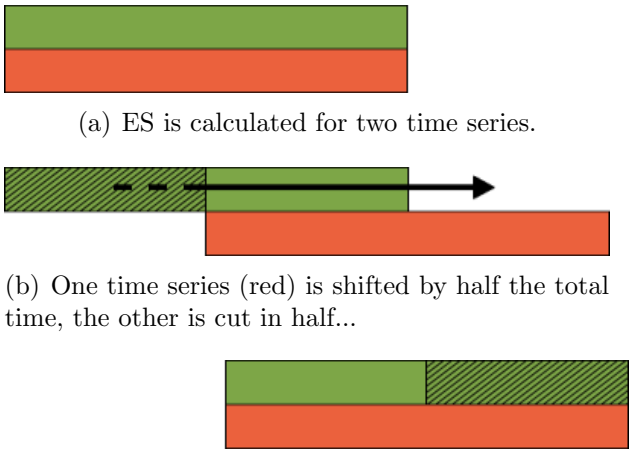
Figure 10 shows the respecting curve: In the plot it is linear with a slope of 0.5, which means q/Q follows a power-law with regard to the density ratio: $\frac{q}{Q} \propto x^{0.5}$. More importantly, the increase is monotonous, which supports the explanation in the preceding paragraph.

At this point we have seen and investigated the behaviour of both result values of the ES algorithm. In the following sections, especially in the application of ES, we will concentrate on the strength of ES (Q), but the delay value may be of use in future projects.

3.5 Surrogate Data Test

To judge the relevance and significance of ES results, we introduce a surrogate data test. This is an additional calculation designed to provide a value to compare the real results to.

To perform the surrogate data test for a given pair of time series, we shift one of the time series by half of the total time (i.e. 20 weeks, as the dataset comprises 40 weeks), then restore time congruence by adding the 'free' first half of the other time series to its end (Fig. 11). This way every point has a distance of 20 weeks to its former position relative to the respective other time series. The surrogate data test result is the strength of ES of these shifted time series. The shift is supposed to eliminate dependencies between the time series, so the test results can be expected to be equal to the calibration values because in both cases we calculate ES for a pair of independent time series with defined densities.



(a) ES is calculated for two time series.

(b) One time series (red) is shifted by half the total time, the other is cut in half...

(c) ...and placed congruently to re-calculate ES.

Figure 11: The surrogate data test concept.

Figure 12 negates this expectation. The surrogate data test results are not equal but actually much lower than the calibration values. This is the consequence of a misconception: Assuming equal results also means assuming similar time series. Remembering the generating of random time series, every point of time had the same probability of featuring an event, which led to a Poisson process and exponential distribution of the event distances (page 8). In reality, events are not equally likely to happen at any time. Our time series are Wikipedia edit times. Explanations for temporarily changing distributions are weekly trends [Yasseri2012a] or bursts of many events during short time intervals due to external events as investigated in [Kämpf2012] or editing responses (*reverts*) up to 'edit wars' [Sumi2011, Yasseri2012b].

We have seen that the surrogate data test does not give the same results as the calculation with shuffled data (calibration). Though it was not expected this way, it gives us additional information, if interpreted accordingly. Also note that the dependence on density ratio is very similar for both curves in Figure 12. At this point we are able to calculate ES, calibrate it using random data, and add an interpretation using the surrogate data test, which is enough to proceed to looking at real data.

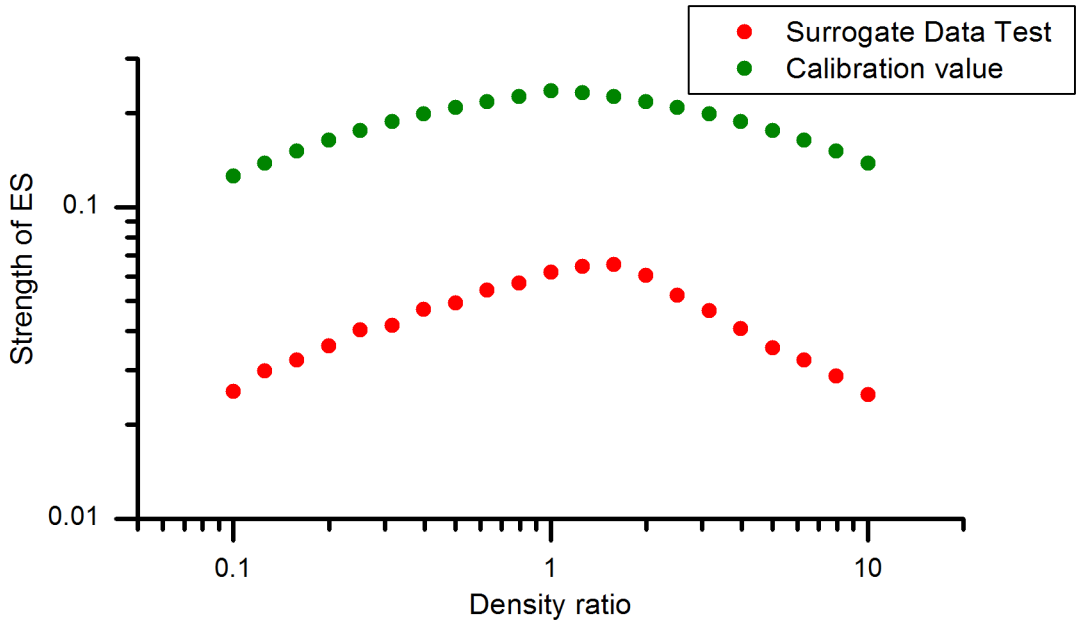


Figure 12: Surrogate data test in comparison to the calibration curve after logarithmic binning. The surrogate data test unexpectedly gives significantly lower results, uncovering aberrations from a time-independent distribution used in time series generation for the calibration.

3.6 Behaviour for Wikipedia data

We calculate ES using edit time series from Wikipedia. Working on this, it has become apparent that most wikipedia pages are edited very rarely. This can lead to surprising results, which we will address in this section. It should be noted that the algorithm cannot work if there are less than four events in at least one of the time series, so we disregard all pages with fewer than four edits. This is a large part of the data, but in turn it means that our results are based only on pages of relatively high public interest.

Figure 13 shows all results for one example group of Wikipedia pages (the DAX node-group), sorted by the lesser density among the two time series. Each point is the calibrated result for one pair of time series (left axis) - black for the strength of ES, green for the strength of ES in the surrogate data test. The red line (right axis) marks the lower event count among the two time series, i.e. the number of edits of the less edited article. We can see that for low event counts (low densities) there are only few possible result values while for higher event counts the number of possible results is increased, observable by the clear horizontal lines on the left side, whose number increases with increasing density until they are not separately recognizable anymore. Furthermore, the results for an event count of 4 are always either 0 or larger than 1, decreasing with higher densities. This can be explained via the calibration, which is generally larger for high event counts because in that case both time series have similarly high density, while the low event counts only apply to one of the time series and the other one may contain any number of events, in many cases leading to high ratios, small calibration values and accordingly large calibrated results.

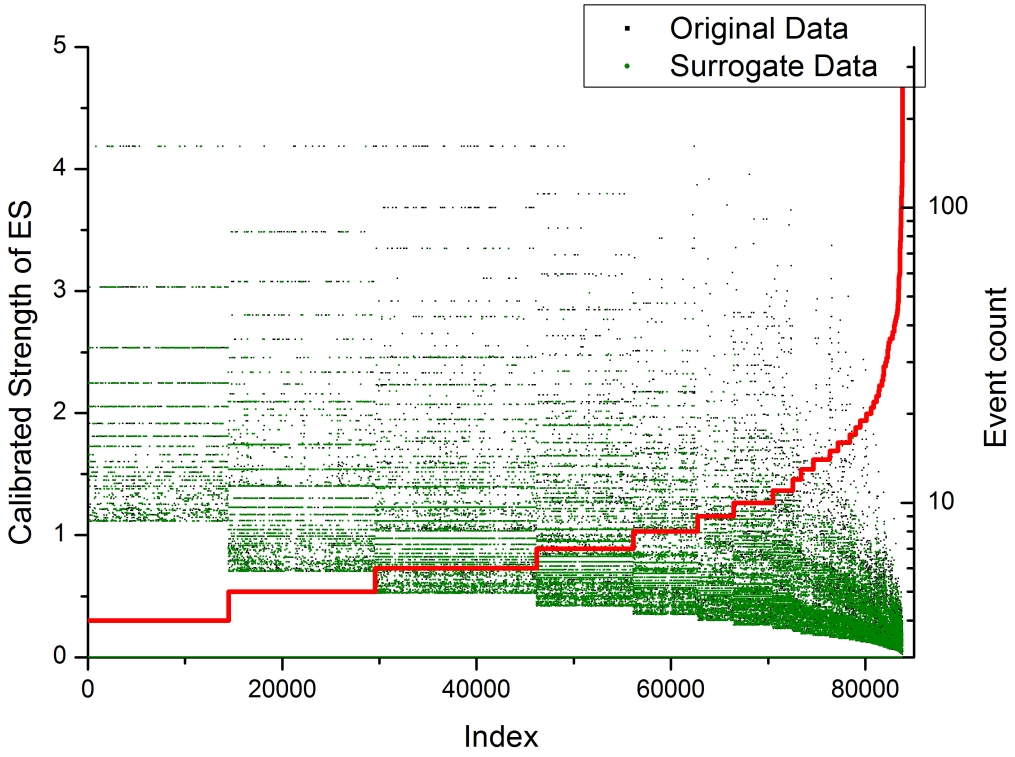


Figure 13: All results for one example dataset (the DAX nodegroup as defined in section 2.4), sorted by density of the less active time series. Black dots are calibrated ES results, green dots calibrated results of the surrogate data test. The red line (right axis) marks the event count in the less active time series. The bottom axis is only an index for the result points. For very low event counts the ES results can only take clearly discrete values, for higher event counts they are continuous and clearly lower.

The effect of discrete values for low densities leads to a second effect, which can be seen in Figure 14. The graph depicts the relative frequencies of values from 0 to 5 for the quotient of strength of ES and surrogate data test. The lines represent different filters: the black line contains all results; the red, blue and green lines only results from time series with an event count greater than 50, 100 or 150. The effect illustrated by this is a preference of integer quotients like $\frac{1}{1}$, $\frac{2}{1}$, $\frac{3}{1}$. For most quotients the green line shows the highest values, followed by blue, red and black, but at such integer positions the graphs contain peaks and the order is inverted, so the black line is on top. For quotients like $\frac{1}{2}$, $\frac{1}{3}$, ... this effect can be seen if the reciprocal x axis is chosen. The behaviour of the graph remains unchanged.
Going back to the ES results this means that "simple" quotients are generally more common than others, but preferredly for low event counts, while the values for higher event counts are more evenly distributed.
Additionally, Figure 14 also hints at the value 0 for the strength of ES being very high for low event counts. We take a closer look at this in Figure 15.

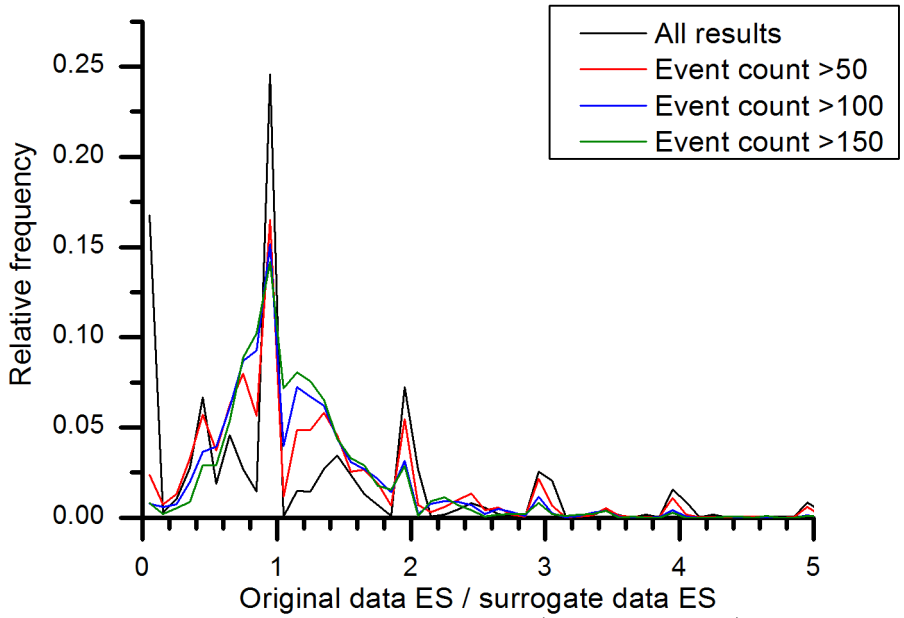


Figure 14: Strength of ES over surrogate data test (horizontal axis) in a histogram (vertical axis: relative frequencies). The black line contains all results, the red, blue and green lines only results with event counts greater than 50, 100, or 150. All lines peak when the strength of ES is a multiple of the surrogate data test. At the peaks the order of the lines is inverted, the black line being highest. This means that these numbers are most common for low densities.

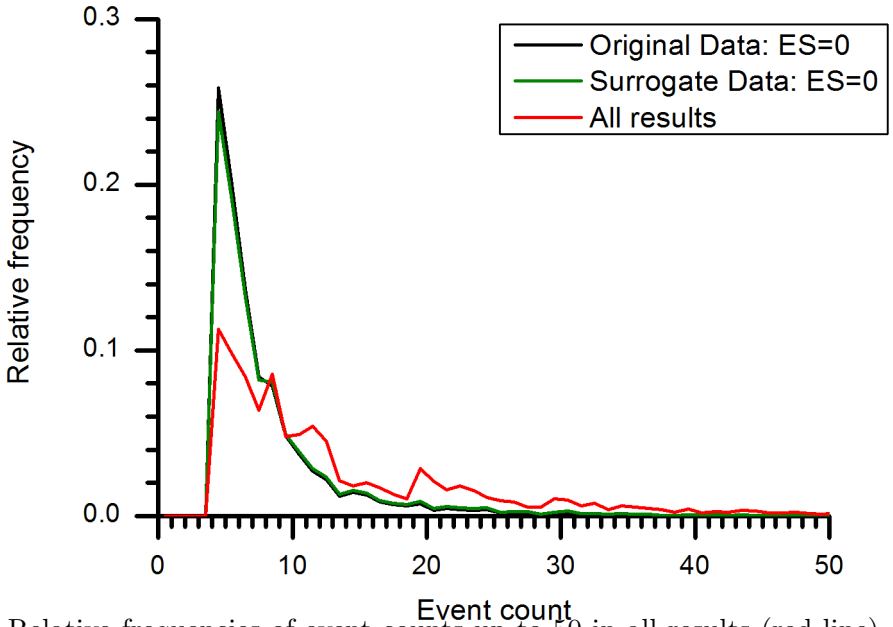


Figure 15: Relative frequencies of event counts up to 50 in all results (red line) or only results with a strength of ES $Q = 0$ for the original/surrogate data (black/green lines). While only 12 percent of all used time series contain exactly four events, 26 percent of the zero results are calculated from time series with four events. For event counts above 10 the red line is above the other two, which means that zero results are less common than for lower densities.

Figure 15 shows the distribution of event counts among all used pairs of time series (red) or only pairs for which the result was zero (black line: real data, green line: surrogate data). For low event counts (from 4 to 10) the black and green lines are clearly higher than the red line, which means that a greater part of the zero results than of all results lies in this area. For higher event counts the red line is on top. Because the green and black lines do not deviate significantly from each other, this cannot be a consequence of synchronicity or asynchronicity.

If two time series contain very few events, these events are likely to be far away from any events in the respective other time series, especially because of possible response edits or reverts, as mentioned earlier. This way a result of zero can be explained as a kind of noise effect due to the very low event counts. For higher event counts this kind of noise is unlikely to appear, so zero is less common as a result.

Though it is important to take note of these effects, their influence on the results is not strong enough to justify the effort of mathematically exact comprehension in this case, so we will not go beyond the above descriptions and explanatory approaches and instead advance to the original aim of network reconstruction.

4 Network reconstruction

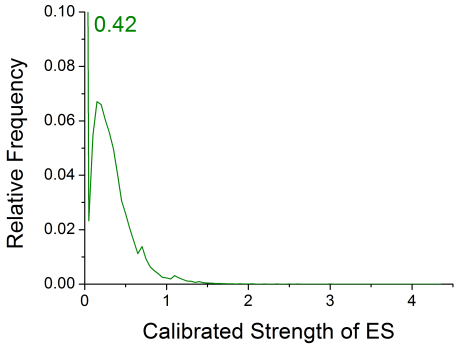
4.1 Result histograms

As mentioned before, we have applied ES to 10 nodegroups of Wikipedia articles, 8 of them related to cities and 2 related to economy:

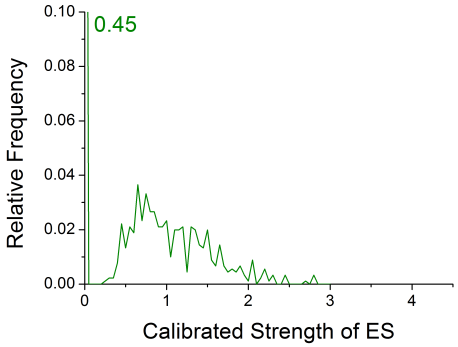
1. All articles in the German wikipedia about cities in Germany
2. All articles in the English wikipedia about cities in the United Kingdom
3. All articles with a direct link to the German article Berlin
4. All articles with a direct link to the German article Heidelberg
5. All articles with a direct link to the German article Sulingen
6. All articles with a direct link to the German article Bad Harzburg (Harz)
7. All articles with a direct link to the English article Oxford
8. All articles with a direct link to the English article Birmingham
9. All articles with a direct link to the German article DAX
10. All articles with a direct link to the English article S&P 500

Figure 16 on page 21 depicts the relative frequencies of ES strengths: The horizontal axis represents the calibrated strength of ES as defined on page 9 (section 3.3), the vertical axis the relative frequencies. Green numbers are the frequencies of the result '0', which are in most cases too large to display. The only exception is the group 'Berlin', for which only 5.5 percent of all results are zero. However, the large ES values are much more important because they are used for network reconstruction, so our focus has to be these results.

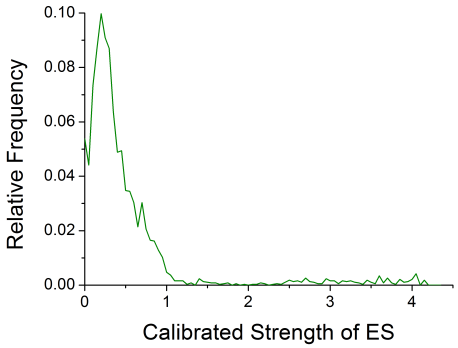
Looking at the histograms, we can identify two distinct types: In the first type (for example Figure 16(a)) most results are concentrated between 0 and 1, forming a maximum near 0.25 and then declining steeply so only very few results are above 1 and can possibly be considered significant. The second histogram type features a maximum around 0.75, then stays moderately high. The maximum is weaker than the one of the first type. Results larger than 4 are very rare in both cases, but between 1 and 4 the number of results is clearly larger in type 2 than in type 1.



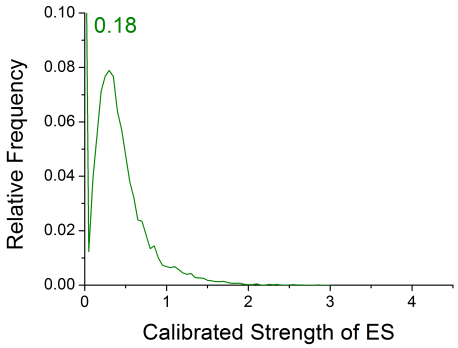
(a) German Cities



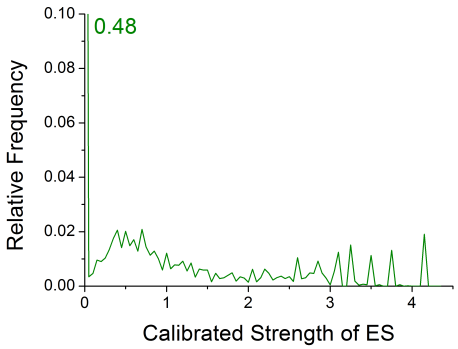
(b) UK Cities



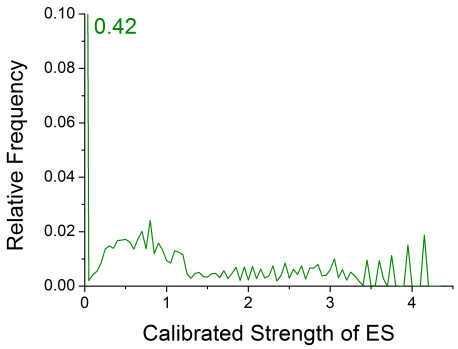
(c) Berlin



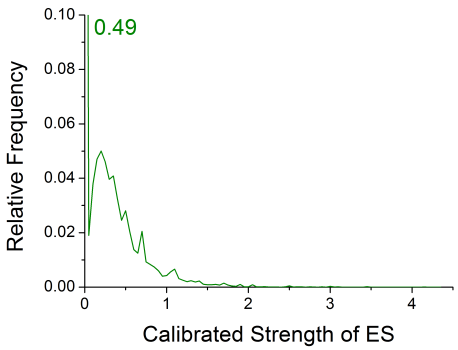
(d) Heidelberg



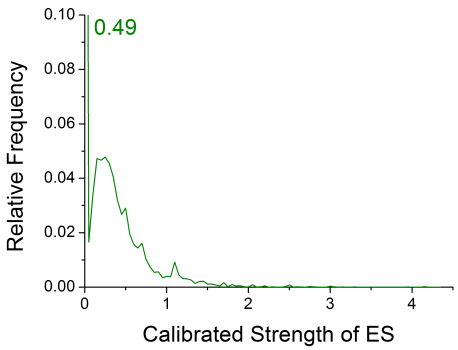
(e) Oxford



(f) Birmingham

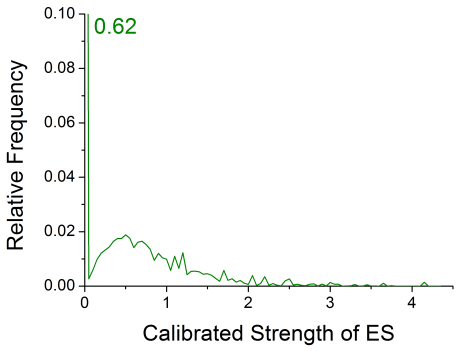


(g) Sulingen

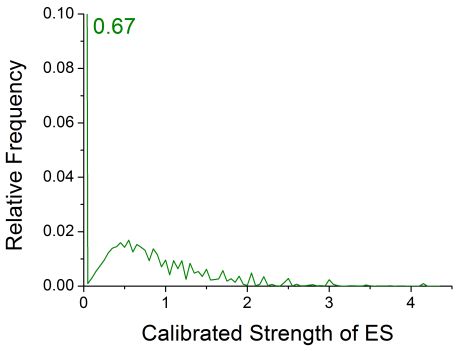


(h) Bad Harzburg

Figure 16: Result histograms for the eight city-themed groups. Calibrated strengths of ES are on the horizontal axis with a bin width of 0.05 and relative frequencies on the vertical axis.



(a) DAX



(b) S&P 500

Figure 17: Result histograms for the two groups of financial data. The behaviour has similarities to both types seen in the city context, with a visible, but weak maximum below 1, but not too few larger results.

The financial data histograms (Figure 17) show a third type of behaviour, which can also be seen as a mixture of the first two: There is a weak maximum close to 0.5, calibrated results larger than 1 occur in moderate numbers. The frequencies decay monotonously (barring noise) following the maximum. This is true for type 1 too, but not for type 2. The peak at zero is even higher than in the city histograms.

While we do not know much about the backgrounds, it is clear that the different histogram forms are connected to different ways of using Wikipedia. Looking at Figure 16 again it becomes evident that histogram form 1 occurs for the five groups based on German cities, all of which consist mainly of articles in the German Wikipedia, and form 2 for the other three nodegroups from the English Wikipedia. Apparently users of the English Wikipedia tend to edit several thematically related pages in a short time, which leads to high correlation, while German users more often edit one article at a time and possibly disregard thematically similar articles, creating low synchronicity. The financial data are of interest for a more specific group of people with its own editing patterns, which can see further investigation in future work.

The differences are not caused by different sizes of the networks. The total numbers of results vary between 1000 and 200 000, but a relation to histogram behaviour can not be found. The size of the user communities might be a factor. The English Wikipedia has a lot more users than the German Wikipedia; the wide spreading of the English language through time zones and cultures can have influence too. But all of this is speculation and can only remain so without deeper knowledge of Wikipedia communities and of sociology. Instead of searching for explanations for the histogram shapes, we can take the last step toward network reconstruction, which was the goal we set at the beginning of this thesis, and try to find a connection to the histogram shapes.

4.2 Network graphs

Figure 18 depicts the reconstructed network of the nodegroup 'Berlin'. Line shading and strength represent link strengths. The graph contains 176 links between a total of 88 nodes (only counting links with a calibrated strength of ES $Q \geq 1$), which equals an average of 2 links per node, in graph theory known as a node's *degree*. For comparison, Figure 19 shows the network for the cities in the UK. Only the 176 strongest links are visible, but in total the network contains 46 nodes and 1000 links, meaning an average degree of 21.7. These numbers reflect the two different histogram types described earlier. The networks from the English Wikipedia contain a higher percentage of significant links than the German networks, though the total numbers of linked node pairs are similar, which means that there need to be higher degrees too.

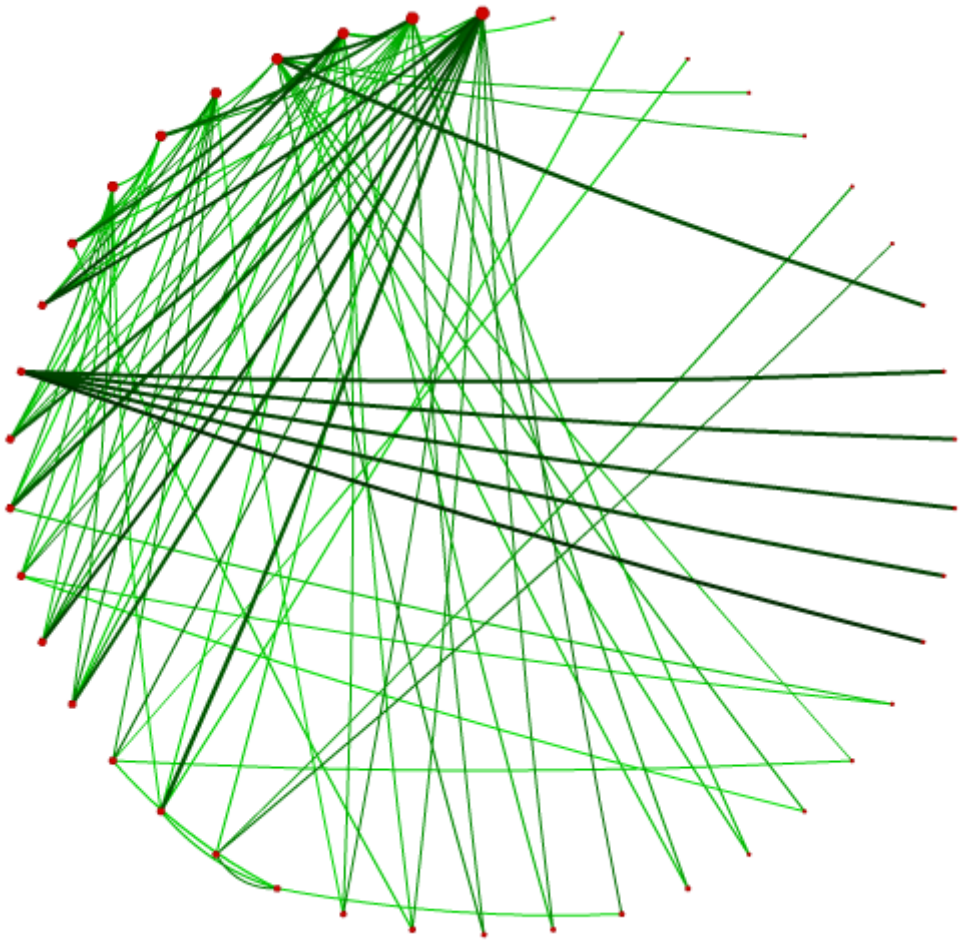
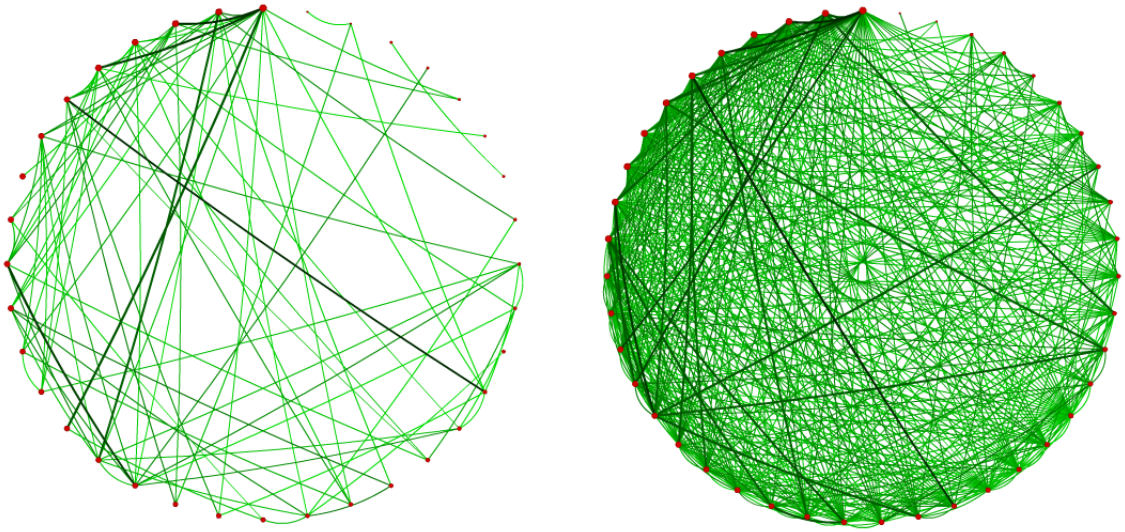


Figure 18: Edit network for the 'Berlin' nodegroup.

Aside from these numbers, the networks cannot be distinguished well by eye. The node degrees seem to be more evenly distributed in the UK cities network than in the Berlin network. In the latter, most links connect one out of very few nodes with high degree to another node. These nodes are placed in the top left area, so most links lead to that area. The cities network is sorted in the same way, but it contains more links outside the top left area, so there are more nodes with a relatively high degree. This is also true if all 1000 links are drawn. The graph of the German cities nodegroup is similar to the UK cities one, while the graphs around single cities look more like the Berlin network, so the two types of networks seem to be caused by the different ways of choosing the nodegroups.



(a) Network with 176 links

(b) Complete network with 1000 links

Figure 19: Edit networks for the 'UK Cities' nodegroup, one with only the 176 strongest links to make it comparable to Figure 18, the other one with all 1000 calculated links.

At this point we can characterize the networks more detailedly if we introduce network measures from graph theory like the node degree mentioned earlier. This has not happened yet and is rather a prospect for future work, which will be discussed in the final chapter.

The two groups of city lists invite to a further way of drawing the graph. We place all cities on a map in their real position, then connect them using the calculated links. The result of this idea can be seen in Figure 20. The German cities are marked as blue circles on a map, links as red lines with different strength to reflect link strengths. As before, it is hard to find a good interpretation as backgrounds are missing, but some properties are observable.

The link strengths are variable, with many weak links and a few very strong ones, like we also observed in the histogram (Figure 16(a)). The strong links are scattered through the map with no evidence for reasons. One single user with large activity and interest in two specific cities can already be enough to produce a strong link, but there are many possible reasons. A direct comparison of the time series could provide more information,

another thing which has not been done yet.
Some cities are clearly visible as central points. Berlin and Dresden can be easily identified as big cities with a central position in the network; more surprisingly Jena seems to have a large degree too, though it is not that well-known as a city. Other big cities such as Munich or Hamburg are nearly missing in the network. So we see that the degree of a city in the network does not reflect its size or status, which is less surprising than it seems at the first glance. Public interest is expressed in form of large numbers of page views, but not necessarily in form of edits. Edit synchronicity rather hints at co-evolution of articles and possibly of the cities the articles refer to.

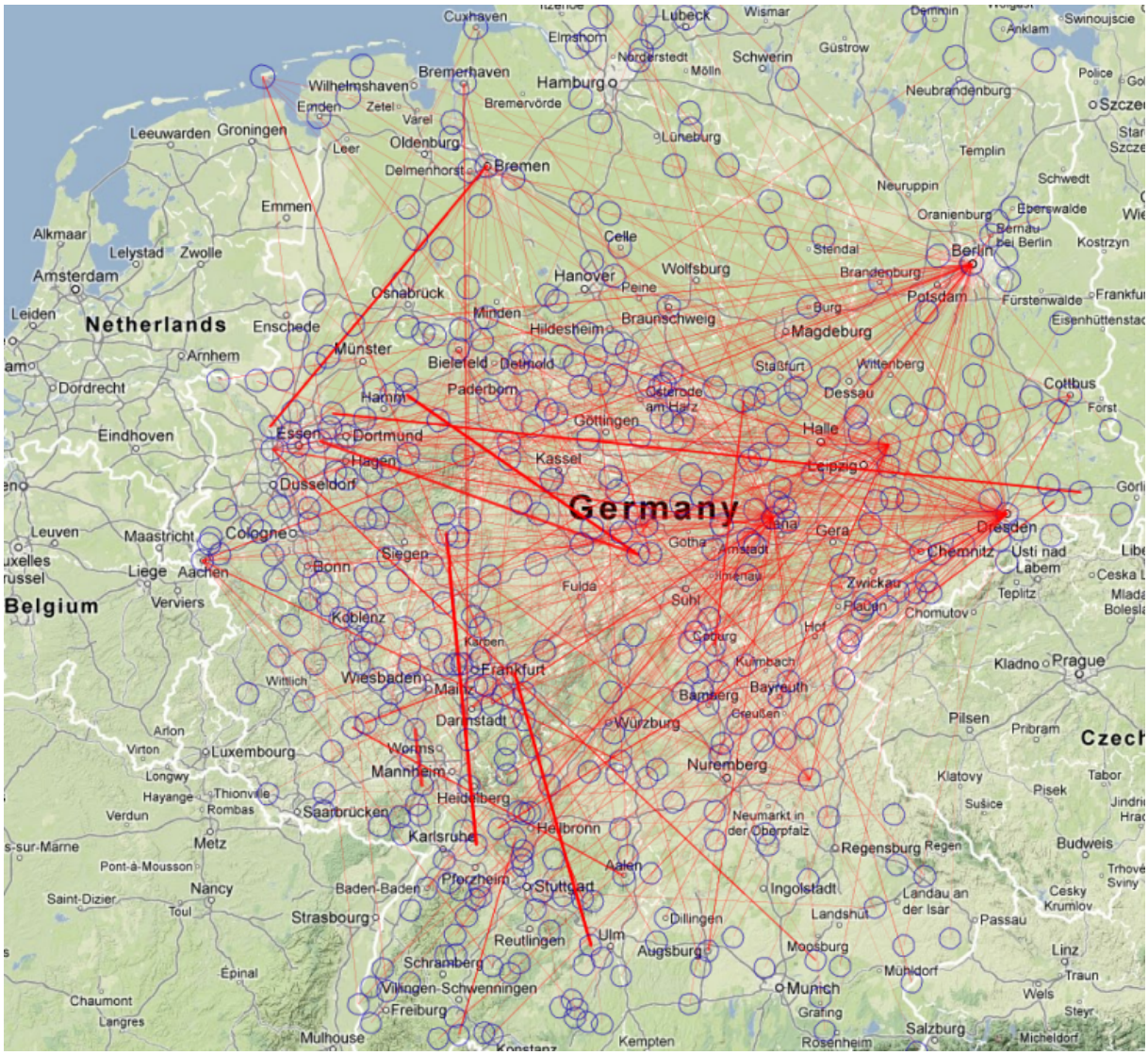


Figure 20: Result network for the group 'Cities in Germany' on a map of Germany. Line thickness mirrors link strength. Some very strong links are distributed all across the map with no directly evident reasons; also some cities can be identified as important centers, for example Berlin, Dresden and Jena. Other large cities have nearly no links, like Munich or Hamburg.

5 Conclusion and prospect

5.1 Cross correlation analysis for access data

The studied Wikipedia dataset (see section 2.2) does not only contain page edit times, but also access rates. These are much higher and can be used as discrete time series with an hourly binning. This has been done by Berit Schreck parallelly to this project [Schreck2012]. She has applied a cross correlation algorithm to the time series and used the results as link strengths.

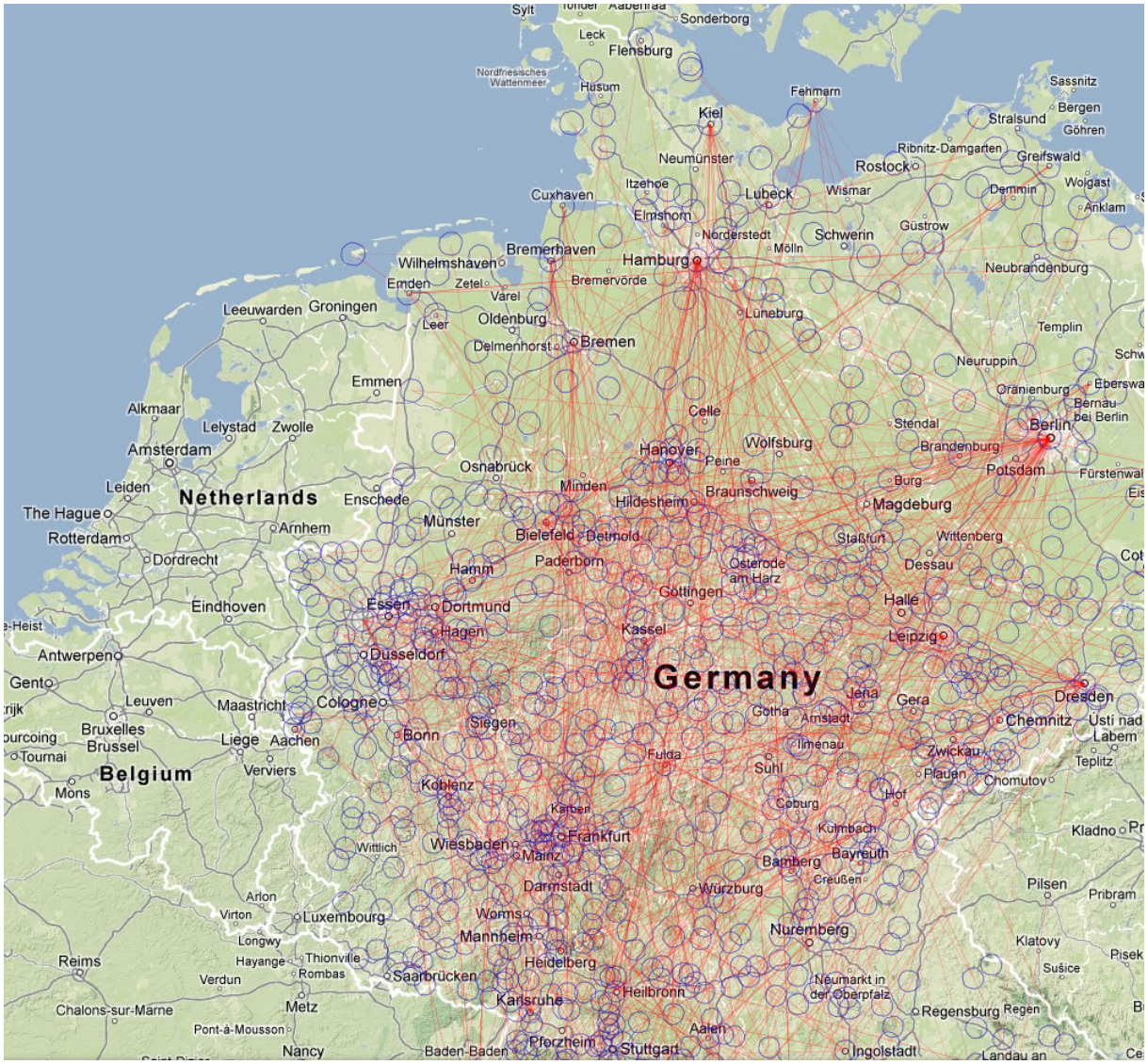


Figure 21: German cities' network calculated from access time series using cross correlation. Figure taken from [Schreck2012] for comparison.

The access network (Figure 21) appears to have properties quite different from our edit network. Link strengths are more homogenously distributed, also the south of Germany seems to contain more links than before, visible by the higher fraction of red colour.

Measures from graph theory, such as node degrees, have been mentioned earlier. While we did not use them for the edit networks, some analysis has already been done regarding the access networks. More is to be done in the future, the aim being a comparison between the static link network and the reconstructed access and edit networks.

5.2 Event Synchronization analysis for access data

Our Wikipedia access data are different from the edit data in their format: While the access time series are time series with a value $f(t)$ for every point of time t, the edit time series are event time series, as described earlier. To improve comparability, it might be useful to transform the access time series into the event time series format.
Figure 22 is an example for how this can be achieved. We choose a threshold value f_0 and then generate a new event time series by placing events at all t_i with $f(t_i) > f_0$. This method has already been used in the previous works applying ES [Quiroga2002, Malik2010] and could be useful in this context again.

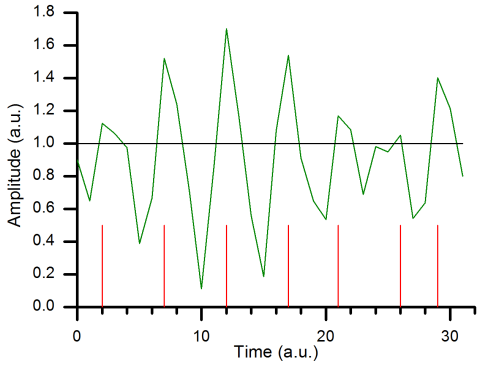


Figure 22: A continuous (or discrete) time series can be transformed into an event time series by marking maxima exceeding a certain threshold as events.

5.3 Conclusion

The aim of this work was to use the algorithm of Event Synchronization (ES) to reconstruct complex networks from event time series. To do so, we needed to understand the recently developed algorithm first and to learn how to interpret the results calculated by it. We achieved this by running experiments with randomly generated time series and finding a calibration function, which helped us to sort out insignificant links. We also introduced a test of significance and compared its results to the calibration function. The same analyses have been performed not only for the strength of ES, which we used as link strengths, but also for the delay value, which can be useful as a directed link strength.
We applied the algorithm to 10 groups of articles from the online encyclopedia Wikipedia, especially from its English and German versions. The results could be drawn as histograms and classified in three different types with close connection to the Wikipedia language the respective nodegroups are based on. We drew network graphs and described them optically, keeping in mind that more detailed mathematical descriptions are possible and necessary to understand the results.

References

[Kämpf2012] M. Kämpf, S. Tismer, J. W. Kantelhardt, L. Muchnik. *Burst event and return interval statistics in Wikipedia access and edit data.* Physica A **391, 23**, p. 6101–6111 (2012).

[Malik2010] N. Malik, N. Marwan, J. Kurths. *Spatial structures and directionalities in monsoonal precipitation over South Asia.* Nonlin. Processes Geophys. **17**, p. 371-381 (2010).

[Malik2011] N. Malik, B. Bookhagen, N. Marwan, J. Kurths. *Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks.* Clim. Dyn. **39**, p. 971-987 (2011).

[Mituzas2010] D. Mituzas. *Page view statistics for Wikimedia projects.* `http://dammit.lt/wikistats` - last visit 2012/09/16

[Quiroga2002] R. Quian Quiroga, T. Kreuz, P. Grassberger. *Event Synchronization: A simple and fast method to measure synchronicity and time delay patterns.* Physical Review E **66**, 041904 (2002).

[Schreck2012] B. Schreck. *Rekonstruktion komplexer Netzwerke mittels Kreuzkorrelationsmethode.* Bachelor thesis, MLU Halle-Wittenberg, 2012.

[Sumi2011] R. Sumi, T. Yasseri, A. Rung. *Edit wars in Wikipedia.* WebSci Conference 2011, Koblenz. `http://journal.webscience.org/523`

[Wikipedia2012a] `http://en.wikipedia.org/wiki/Wikipedia` - last visit 2012/09/10

[Wikipedia2012b] `http://meta.wikimedia.org/wiki/List_of_Wikipedias` - last visit 2012/09/10

[Wikipedia2012c] `http://en.wikipedia.org/wiki/City` - last visit 2012/09/16

[Wikipedia2012d] `http://en.wikipedia.org/wiki/Wikipedia:Namespaces` - last visit 2012/09/15

[Yasseri2012a] T. Yasseri, R. Sumi, J. Kertész. *Circadian patterns of Wikipedia editorial sctivity: A demographic analysis.* PLOS ONE **7**, e30091 (2012).

[Yasseri2012b] T. Yasseri, R. Sumi, A. Rung, A. Kornai, J. Kertész. *Dynamics of conflicts in Wikipedia.* PLOS ONE **7**, e38869 (2012).

Acknowledgements

First of all I would like to thank my supervisor PD Dr. Jan W. Kantelhardt for giving me the opportunity to write this thesis as a part of his working group and for various forms of support during the process.

As parts of the working group, Berit Schreck and Patrick Wohlfahrt have been very supportive too, giving advice when needed and creating a very good atmosphere throughout the time.

Maybe most importantly I want to give thanks to Mirko Kämpf for keeping the necessary technical structure working and providing technical as well as methodical help. His willingness to help others and his ability to stay calm and rational amazed me on many occasions.

Finally thanks to everyone who gave this thesis its form by proofreading or small hints. I appreciated their help and flexibility.

Declaration

I declare that I wrote this thesis autonomously, that no resources and auxiliaries were used apart from those mentioned in the thesis, direct or indirect excerpts from other sources were marked and the thesis was not submitted to another examination institution before in identical or similar form.

Arne Böker
Halle (Saale), 2012/09/25