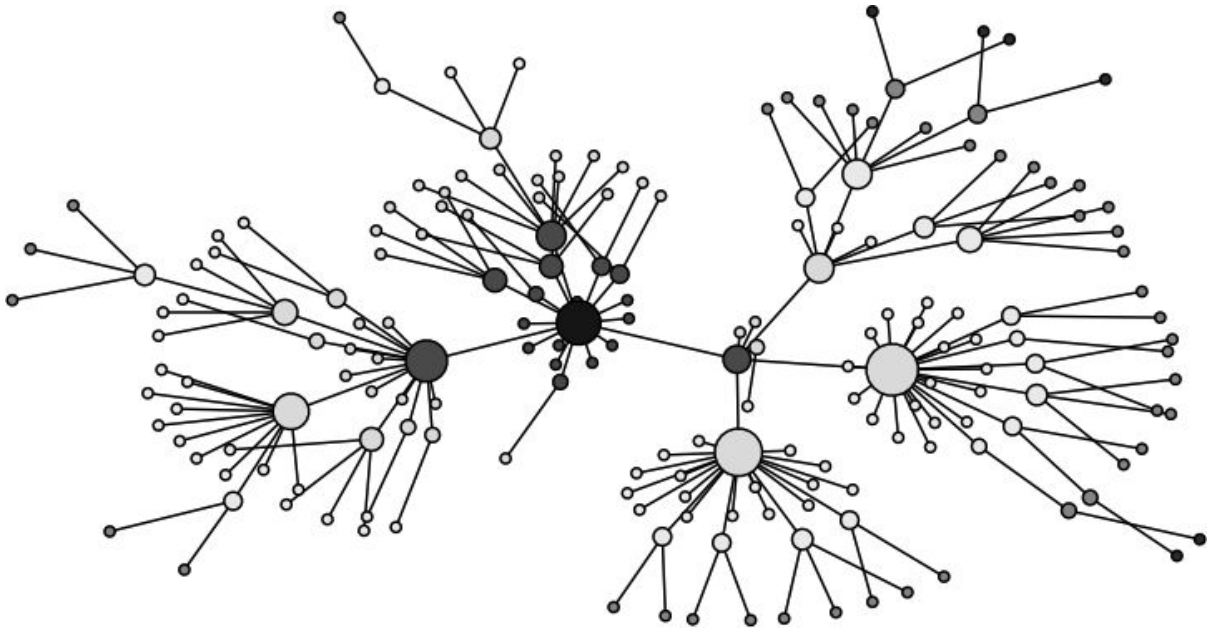
Project Report week 2



Group 22

Wim florijn	s1503251
Rob van Emous	s1470647

Table of contents

Introduction

The provided graph of the UTwente network structure is very large. That is why the calculations required to answer Question 1 could not be done using the full graph. These calculations would take too much time while at the same time needing an extraordinary amount of memory. These calculations have been done on a stripped down version of the UTwente graph which contains only about 100 000 nodes or 33% of the original graph. The full Utwente web graph is pruned to this size (threshold) in two different ways. Afterwards the effectiveness of both methods will be compared.

- **First Trimming technique**
  Only pages with a number of incoming edges larger than the threshold are included in the stripped down graph. The idea behind this method is that the important pages: with the most incoming links will still be included in the stripped down graph.
- **Second Trimming technique**
  The nodes are sorted to the amount of slashes contained in the matching URL. After this, only the first ‘n’ nodes (up to the threshold) are used. This should be equivalent to pruning the tree of nodes at a certain depth and therefore should keep the structure of the tree in mostly intact.

The calculations needed for question 2 are done using the full graph because the algorithms used are less hardware intensive. This provides a much more detailed solution. The interpretation of both questions is described below.

Question 1 - What is the (conceptual) structure of the UTwente web?
To answer this question, the Giant Strongly Connected Component, and the IN, OUT, Tendril, Tube and Disconnected sets are generated using the strongly connected components of the stripped down UTwente graph. This way the global structure of the UTwente web will be made visible and can be be compared to the structure of the entire internet.

Question 2 - What are the important pages of the UTwente web, based on the analysis of the web graph?
This question can be answered by sorting the nodes according to the number of incoming edges or calculating the PageRank of all nodes in the full UTwente graph. After that, the nodes can be sorted and compared by using these two variables to gain insight in the importance of particular nodes. Next to this, the strongly connected components and the giant strongly connected component will be generated to provide a view on the layout and the connection of components in the UTwente network.

Alternatives and justification of chosen approach

There are alternatives in the way of how the stripped down UTwente graph can be generated. The graph can for example be generated by taking a random sample from the dataset, by only taking the first n nodes from the dataset or leaving out a certain amount of subdomains: www.studentunions.utwente.nl for example. Instead, we have thought of smarter ways to sample the dataset.

The first method we have chosen, involves sampling the dataset based on the amount of incoming edges per node. This type of sampling will probably only keep the most important nodes, and therefore will disturb the structure of the full graph as least as possible.
The second method takes only nodes up to a certain nesting depth. It is assumed the amount of slashes indicate this depth and the more nested a page is, the less it is of importance. It will therefore tend to keep the structure of the web graph/tree intact.

The strongly connected components and the giant strongly connected component of the full UTwente graph can be compared with the strongly connected components and the giant strongly connected component of the stripped UTwente graph to determine the possibly negative effects of the stripping.

In general, the process for generating the results for each question is the following:

What is the (conceptual) structure of the UTwente web?

1. Generate the stripped down UTwente graph (by either of the two methods)
2. Generate the strongly connected components of the graph
3. Determine the giant strongly connected component of the graph
4. Determine the in & out sets of the graph
5. Determine the tendril & tube and disconnected sets of the graph

Details of generating the IN, OUT, Tube, Tendril and Disconnected sets
First, the giant strongly connected component is removed from the set of strongly connected components.

- **IN & OUT**
  The IN- and OUT-sets are generated as described by David Easley and Jon Kleinberg in their book ‘Networks, Crowds, and Markets: Reasoning About a Highly Connected World’. Thus, if a strongly connected component can reach the giant strongly connected component, it is an IN-set. If it is the other way around, it can be reached by the giant strongly connected component, it is an OUT-set
- **Tubes**
  If a strongly connected component is not in either the In-set or the OUT-set, but can be reached from the In-set and can reach the OUT-set, it is a tube.
- **Tendrils**

If a strongly connected component is not in either the In-set or the OUT-set, but can be reached from the In-set while it cannot reach the OUT-set, it is a tendril. Furthermore, if a strongly connected component is not in either the In-set or the OUT-set, but can reach the OUT-set while it can't reach the In set, it is a tendril.

- **Disconnected**
  If a strongly connected component does not belong to the IN, OUT, Tube or Tendril sets, it is disconnected.

What are the important pages of the UTwente web, based on the analysis of the web graph?

1. Generate the full UTwente graph
2. Generate the strongly connected components of the graph
3. Determine the giant strongly connected component of the graph
4. Calculate the page-ranking of the graph
5. Calculate the number of incoming edges for each node

Results

The results of both research questions will be discussed separately.

What is the (conceptual) structure of the UTwente web?
The UTwente web consists of a large amount of strongly connected components which all belong to either the IN, OUT, Tube, Tendril or Disconnected sets. As already discussed above, only about 100.000 nodes of the total graph were used in the calculations. Therefore, all percentages mentioned in the results relate to this total amount of nodes.

Results based on first trimming technique

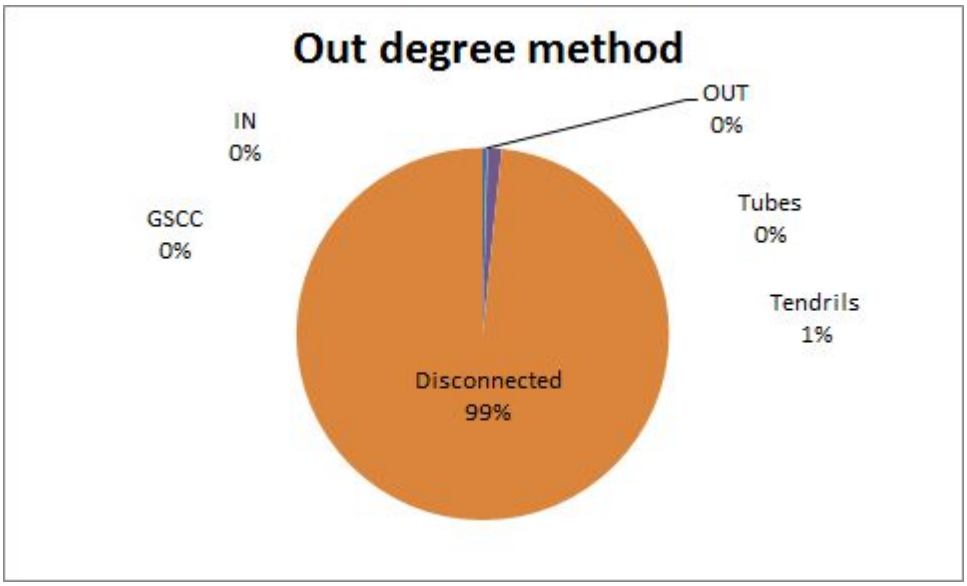


Figure 1: The distribution of nodes between the GSCC, IN, OUT, Tube, Tendril and Disconnected sets using the first trimming technique

As shown in Figure 1, 99% of the nodes in the graph belong to the Disconnected set. 0.35% of the nodes belong to the GSCC. 0.09% of the nodes belong to the OUT-set, and 1.19% belong to the Tendrils set. Both the IN- and the Tubes-set are empty.
The fact that almost all nodes are in the disconnected set, is probably caused by the trimming technique cutting away most links between clusters of nodes. These links acted as bridges between those clusters, while at the same time having only a small amount of in-links.

Results based on second trimming technique

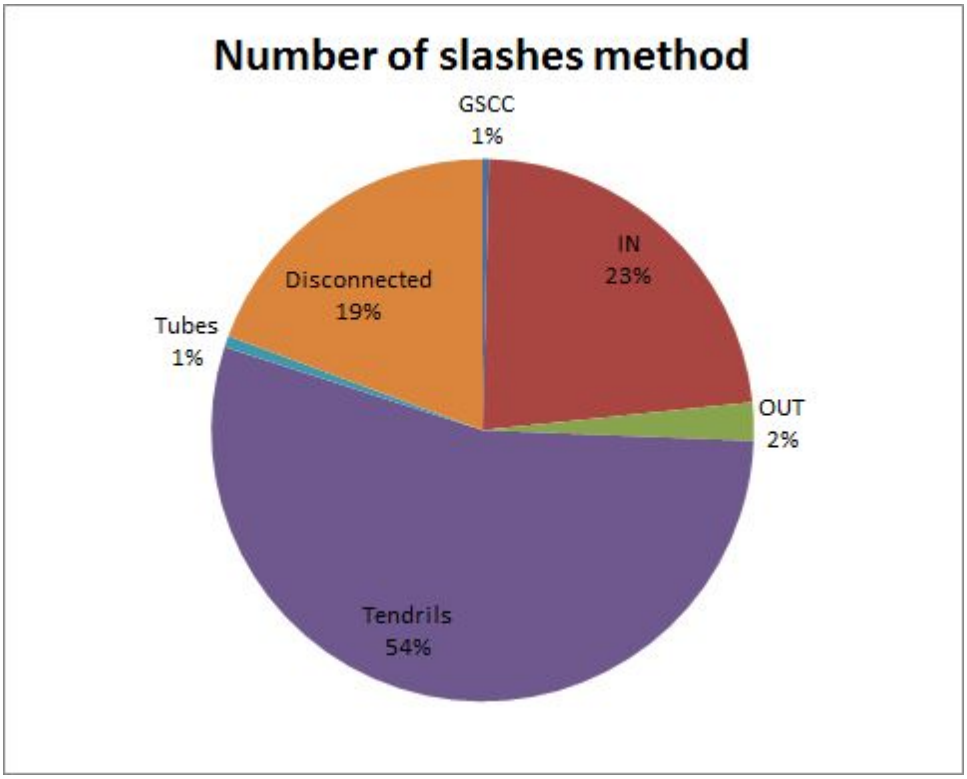


Figure 2: The distribution of nodes between the GSCC, IN, OUT, Tube, Tendril and Disconnected sets using the second trimming technique

As shown in Figure 2, 23% of the nodes in the graph belong to the IN-set, 2% to the OUT-set and 1% to the giant strongly connected set (GSCC). 1% of the nodes are tubes and 54% are tendrils.

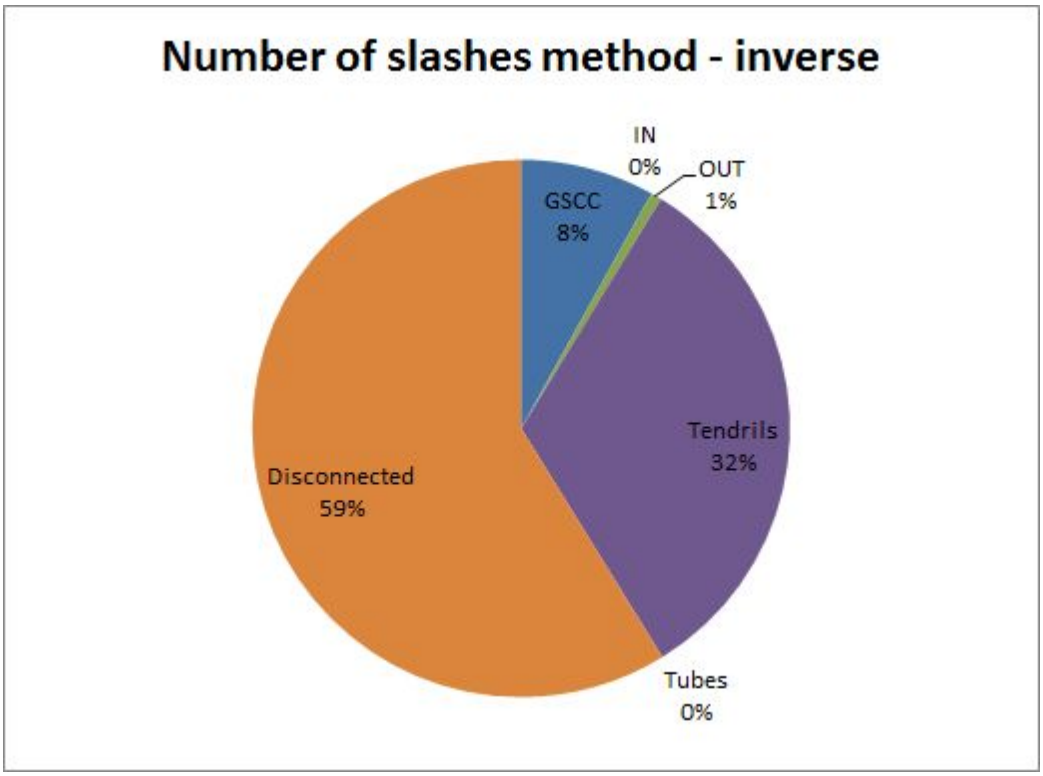


Figure 3: The distribution of nodes between the GSCC, IN, OUT, Tube, Tendril and Disconnected sets using the inverse of the second trimming technique

Finally, if the second trimming method would be valid, its inverse would not be: The inverse would keep the nodes the normal method would throw away. Therefore the bow-tie structure of the graph trimmed after sorting in the reverse order of the number of slashes per URL has been calculated (Figure 3). It has about three times (19% ⇒ 59%) the amount of disconnected components compared to the normal trimming method

What are the important pages of the UTwente web, based on the analysis of the web graph?

The calculations have been done on the full UTwente graph, instead of a trimmed down one. The full UTwente graph contains 184833 strongly connected components with a giant strongly connected component size of 124897 nodes. The UTwente network consists of a total of 312554 nodes.

The top 10 of most important pages of the UTwente web based on their PageRank (the value of this rank is visible in column four), is shown in figure 4. As one could expect, the english homepage has the highest PageRank. An interesting phenomenon is that all pages in the top 25 are nodes in the giant strongly connected component of the graph. On top of that, a correlation between the in-degree and the pagerank can be seen in the data: the best node based on PageRank also has the highest in-degree. A visualisation of this relation using a list of the 1000 best ranked nodes is shown in Figure 5.

Page	inDegreeRank	PageRankValue	inDegreeValue	Node id	Url	Type
0	1	0.02338017	201228	76	http://www.utwente.nl/en/index.html	gscc
1	113	0.015371659	11601	73	http://my.utwente.nl/index.html	gscc
2	43	0.01458656	47454	1019	http://www.utwente.nl/en/about-our-website/disclaimer/index.html	gscc
3	44	0.014030222	46259	1020	http://www.utwente.nl/en/about-our-website/index.html	gscc
4	6	0.01369465	103484	168	http://www.utwente.nl/eng/entrepreneurs.html	gscc
5	46	0.012311575	41324	117	http://www.utwente.nl/en/prospective-students/phd_programmes/index.html	gscc
6	120	0.011880886	6509	769	http://www.utwente.nl/en/contact/index.html	gscc
7	45	0.011851954	41394	161	http://www.utwente.nl/en/organization.1.html	gscc
8	47	0.011759692	41315	133	http://www.utwente.nl/en/research.1.html	gscc
9	50	0.011734689	40846	181	http://www.utwente.nl/en/campus/index.html	gscc
10	52	0.011484048	39330	1	http://utwente.nl/index.html	gscc
11	48	0.011336033	41158	979	http://www.utwente.nl/telefoongids/en.html	gscc
12	124	0.010045286	5774	968	http://www.utwente.nl/en/news.1.html	gscc
13	127	0.010045071	5077	77	http://www.utwente.nl/de/bildung/index.html	gscc
14	125	0.010033354	5486	969	http://www.utwente.nl/en/events.1.html	gscc
15	6302	0.010015186	39	5643	http://www.utwente.nl/en/contact/alarmnumber/index.html	gscc
16	9	0.009184815	98433	747	http://www.alumnus.utwente.nl/en.html	gscc
17	101	0.008248581	26476	434	http://www.utwente.nl/onderwijs/ewi.html	gscc
18	5	0.007982587	123271	2132	http://www.ewi.utwente.nl/en/index.html	gscc
19	12	0.007952434	94451	72	http://www.utwente.nl/telefoongids.html	gscc
20	3665	0.007530862	88	770	http://www.utwente.nl/contact/alarmnummer.docx/index.html	gscc
21	82	0.006786154	32446	3	http://www.utwente.nl/onderwijs/index.html	gscc
22	16	0.006684064	87462	5153	http://www.ewi.utwente.nl/en/research/index.html	gscc
23	15	0.006604026	87935	2445	http://www.ewi.utwente.nl/en/education/index.html	gscc
24	17	0.00658673	87454	20712	http://www.ewi.utwente.nl/en/jobs/index.html	gscc

Figure 4: The top 25 most import pages of the UTwente web

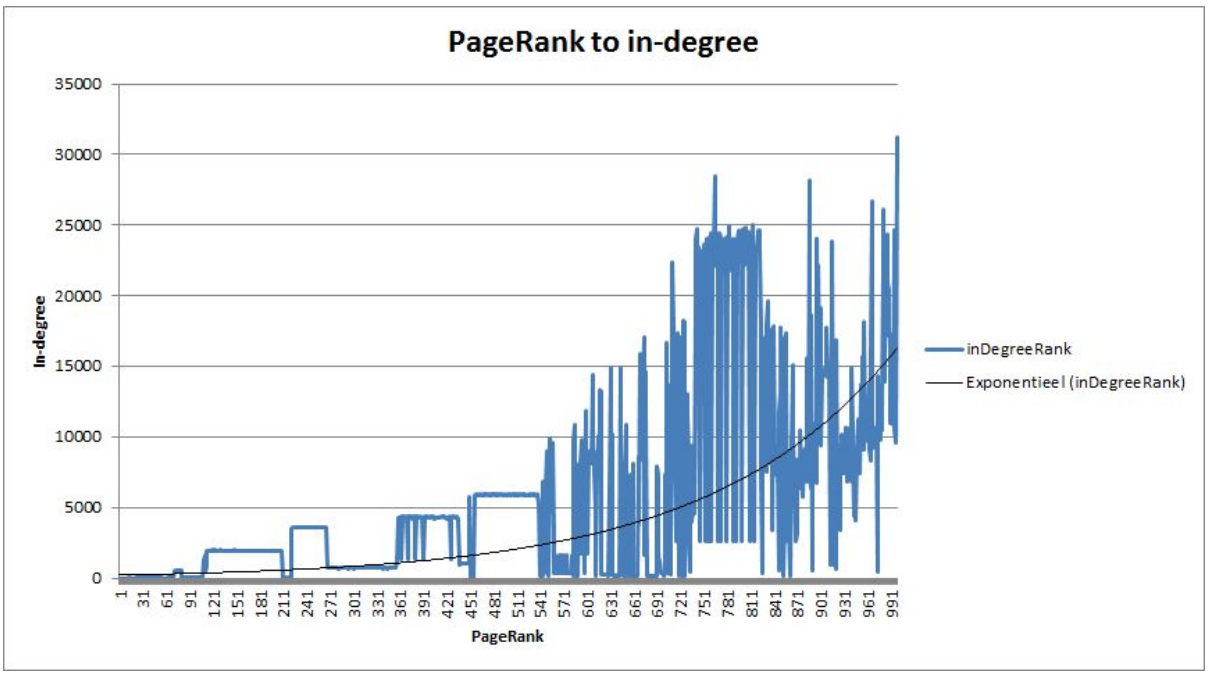


Figure 5: The relation between PageRank and InDegree

Conclusion

A conclusion can be made about there being a huge difference in size of the In/Out/Tubes/Tendrils/Disconnected sets using the two pruning techniques on the UTwente graph. Clearly, the most efficient trimming technique for the graph is the second one. To further justify this opinion, the second method has been compared to its inverse. The inverse performed much worse, having a performance somewhere between the first and second method judging on the amount of disconnected components (99% ⇒ 59% ⇒ 19%).
Using the chosen technique the IN-set contains 23% of the nodes which similar to the 21,5% of the IN-set in the bow-tie graph. On the contrary, the OUT-set containing just 2% of the nodes is nowhere near the expected 21,5%. The 54% of the nodes in the Tendrils and 19% in the Disconnected-set are more than double the 21,5% respectively 8% which they should have been. Finally the GSCC containing just 1% of the nodes is much smaller than the expected 27,5%.
By comparing the size of the GSCC to that of the exact calculation: 1% versus 40%, it is clear that the reduced graph has lost a rather large amount of connections between strongly connected components. On the other hand, it does have similarities to the bow-tie-structure. If more time and computer memory would have been available, a larger part of the total grap could have been used in the calculation. This would have made the evidence towards the UTwente web having a bow-tie structure stronger.

It is interesting that the most important pages of the UTwente network all belong to the GSCC and thus are strongly connected. This means that the UTwente network contains a small set of pages which are highly connected and are very important in the network. It is very important for the UTwente to maintain these pages because they can be seen as the popular-backbone of the UTwente network.