2IID0 - Web Analytics
Advanced homework
January 2016

Community detection and dynamics analysis

Dennis van den Brand - 0772180
Cengizhan Can - 0872958
Amber van der Heijden - 0862713
Guido Santegoeds - 0890429

https://github.com/den1den/web-analytics/tree/final

1. Community detection

Results and analysis

(a) Cluster the authors using only the attributes and then analyze the communities on the graphs.

In order to do this we used `Kmeans_text.py` on the text features files, which gave us a clustering of our data based on the titles of the publications by each author included in the `text_features` file for each of the 10 years. The results of this clustering method are included in the zip file.

(b) Cluster the authors using only the graphs and then analyze the communities on the attributes.

In order to do this we used `Agglomerative_graph.py` on the graph adjacency lists, which gave us a clustering of our data based on which authors worked with each other during each of the 10 years included in the files. These results are also included in the deliverables zip.

After using the basic scripts to come up with the different clusters we then created a Python script that would combine the results of all these years into files that include all the 10 years together. This enables us to better run dynamic analyses on the data using tools such as Gephi later on.

Interpretations about our findings

Purity results (k-means) **on** 20xx_pred_label_text.txt	**Purity results (agglomerative)** **on** 20xx_pred_label_graph.txt
0.256018217306	0.37410540013
0.281188893234	0.310129057489
0.291710114703	0.373044838373
0.279734411085	0.342956120092
0.35051943055	0.350904193921
0.29417000445	0.344459279039
0.330819851933	0.408143679737
0.248442906574	0.355190311419
0.277283372365	0.388602654176
0.268542643738	0.383011985409

(Table 1.1: Purity script results on results obtained through `Kmeans_text.py` and `Agglomerative_graph.py`.)

We used an updated version of the `purity.py` script to evaluate the validity of our obtained results through `Kmeans_text.py`. According to these scores, the clustering that was performed using the attributes does not provide a clustering that adheres to the ground truth. In fact, it does not even get one third of the authors into the right community. We used the updated version of `purity.py` script again to also evaluate the validity of the results from `Agglomerative_graph.py`. This shows us that the script based on which authors work together consistently does a better job clustering the authors into the right community. The purity scores for clustering on the graph are higher than for the clustering based on the titles for each of the years that we have data on. However, while it does a better job than the title based script, it is still far from perfect, scoring a little under 40% in getting authors in the right cluster. These results show us that to obtain certain valuable results in other parts of the assignment it will be best to work, at least initially, with the ground truth values and then gain interesting insights by comparing those with results based on the graph and the text data.

Comparing these two based on the purity results seems to indicate the authors working together is a better way to test to which community someone belongs than the words used in the titles of their publication. To shed some more light on the differences between these two however, we have been inspecting the results using a graphical tool called Gephi. This can provide better insights on which authors are wrongly clustered by one or both of the algorithms and perhaps even tell us why this is. We intend to elaborate on this further when one of our computers has Gephi running correctly for a bit longer as currently its continuous crashes have been costing us many hours of extra time and frustration.

2. Influence analysis

Results and analysis

author_id	comm_pr	h_index	comm_id
862	0.00325714120282	101	0
246	0.0028907627079	8	0
1367	0.00237892725022	30	0
526	0.00189298915823	49	0
72	0.00186464045378	58	0
522	0.00491036131849	138	1
1020	0.0046902332728	120	1
4477	0.003888867688	78	1
1	0.00387070201214	84	1
4	0.0033472416427	126	1
1349	0.00578478837525	-1	2
1350	0.00472890386061	119	2
1397	0.00413760773371	36	2
1364	0.00344346378039	61	2
3729	0.00334696416152	23	2
2502	0.00605506396086	-1	3
2457	0.0057844613838	41	3
2605	0.00540355416338	26	3
2607	0.00508627467366	17	3
2517	0.00492927711431	58	3
1725	0.00531368775358	77	4
1521	0.00476105066254	35	4
1321	0.00471563913153	31	4
1516	0.00459491222507	-1	4
1737	0.00452375933455	46	4
2452	0.0158080265957	27	5
2345	0.013396406523	17	5
2585	0.0121835869183	37	5
2565	0.0121228753812	47	5
2227	0.0121228753812	13	5

Table 2.1: Results of PageRank on communities on the graph of 2010.

author_id	comm_pr	h_index	comm_id
4600	0.00470336809378	55	0
2368	0.00469789489655	36	0
2061	0.00367367229786	36	0
3391	0.00367367229786	44	0
3567	0.00367367229786	34	0
196	0.00897190976861	16	1
1366	0.0081207740829	53	1
1670	0.00809106508768	46	1
1942	0.00668353098699	33	1
1944	0.00668353098699	33	1
457	0.00430016860219	34	2
2453	0.00430016860219	7	2
1589	0.00428527144484	-1	2
1138	0.00428527144484	26	2
482	0.00401295461481	27	2
522	0.00138831781343	138	3
1350	0.00104523022431	119	3
1365	0.00101765729907	6	3
4477	0.00100712562932	78	3
1725	0.000928920720718	77	3
184	0.0101214428554	14	4
1454	0.0101214428554	-1	4
236	0.0101214428554	14	4
237	0.0101214428554	19	4
4883	0.0101214428554	13	4
3996	0.5	-1	5
3997	0.5	52	5

Table 2.2: Results of PageRank on communities on the attributes of 2010.

In order to determine which authors are influencers in any of the computed clusters in part 1 of the assignment we have decided to use the supplied PageRank algorithm and the h-index values computed by Google Scholar. We have chosen to use the results of 2010, since we think that recent results will be better for accurately determining the influence using the h-index.

The results can be seen in the tables. We have used the results of the five authors with the highest PageRank within each of the six communities for both of the methods from part 1 of the assignment. Then we have sorted on the communities and the PageRank and have matched the `author_id` to their actual names from the `author_mapping` file of 2010. Using their real names, we can now find their profile page on Google Scholar and look up their h-index. All this information is shown in the above tables. An `h_index` value of -1 means that there was no information for the relevant author on Google Scholar, so also no h-index.

Interpretations about our findings

We can easily see that there are quite a lot of authors with significantly high values for the `h_index`. The average h-index (not taking the -1 values into account) for the computations based on the graph is 56, and for the computations based on the attributes this is 42. Hirsch estimates that after 20 years a "successful scientist" will have an h-index of 20, an "outstanding scientist" an h-index of 40, and a "truly unique" individual an h-index of 60.[1] For both sets of our results this would mean that the average author, within the top five in the communities, is an outstanding scientist. Using this information we think that it is reasonable to say that the top five authors in each of the clusters can be seen as influencers inside those particular communities, with just a few exceptions that could perhaps be due to time differences between Google's h-index and our input data. This means we have validated our results and we can conclude that the authors with the higher PageRank values within a community can be seen as influencers inside that particular community based on their h-index in Google Scholar.

[1] Meho, L.I. (2007) *The rise and rise of citation analysis.* Physics World, January 2007, 32-36

3. Dynamics analysis

How do these research communities change over years? How can you quantify this or locate concrete examples of evidence?

- Do some communities become larger or smaller?
- How much fresh blood gets in?
- Are some communities get stronger/weaker connected with some other communities than before?
- If a (part of the) community stays structurally similar over some years, does it focus on the similar or different topics (keywords) over these years?

We started off by simply looking at the numbers of authors belonging to the different communities in every year. The ground truth data was placed into tables that can give a better view of how communities are growing or shrinking over the years and which ones attract the most fresh blood at given times.

As for how much fresh blood is drawn into each community each year, table 3.1 shows the exact numbers of authors and therefore also hints towards the number of new arrivals to the community in that year. This number will be equal to the difference with the year before plus the amount of people that left the subject that year. Some notable differences can be seen in the CV community, which counted 1839 more authors in the year 2007 than the year before, and almost lost half of these authors again the year after. Looking at the total numbers we can also note that 2007 was a good year for all of these communities in general, but when looking at the relative scores in table 3.2 it becomes clear that the CV community did exceptionally well, even compared to the growth of the others. Another notable statistic is the continuous interest in the community of AL & ML, during the whole 10 years it has had the most interest of all the communities, however during the latest years, it has been losing heavily in the relative scores and is no longer leading the group. This is of course also due to the increasing popularity of CV. The smallest community during these years has clearly been IR, and when looking at the second table it becomes clear that the relative amount of interest in this subject has remained almost stable during this whole period. Finally, it can be seen in the absolute numbers that academical interest in these subjects has been increasing in general over all of these communities during the 10 years. Starting with around 3000 authors per year and ending at around 6000, 10 years later.

Number of authors	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	Total
DB (0)	553	639	615	795	795	949	909	807	884	720	**7666**
AL & ML (1)	715	287	949	794	1805	1223	2103	1182	1213	1330	**11601**
IR (2)	192	277	367	335	407	431	555	588	619	603	**4374**
DM (3)	572	616	772	780	909	964	946	1166	1317	1058	**9100**
AL & TH (4)	462	554	547	540	552	619	634	668	614	627	**5817**
CV (5)	580	184	586	220	730	308	2147	1369	1758	1419	**9301**
Total	**3074**	**2557**	**3836**	**3464**	**5198**	**4494**	**7294**	**5780**	**6405**	**5757**	**47859**

Table 3.1: The exact number of authors within each community per year.

Relative share of attention	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	Total
DB (0)	18%	25%	16%	23%	15%	21%	12%	14%	14%	13%	**16%**
AL & ML (1)	23%	11%	25%	23%	35%	27%	29%	20%	19%	23%	**24%**
IR (2)	6%	11%	10%	10%	8%	10%	8%	10%	10%	10%	**9%**
DM (3)	19%	24%	20%	23%	17%	21%	13%	20%	21%	18%	**19%**
AL & TH (4)	15%	22%	14%	16%	11%	14%	9%	12%	10%	11%	**12%**
CV (5)	**19%**	**7%**	**15%**	**6%**	**14%**	**7%**	**29%**	**24%**	**27%**	**25%**	**19%**

Table 3.2: Relative share of authors within each community per year.

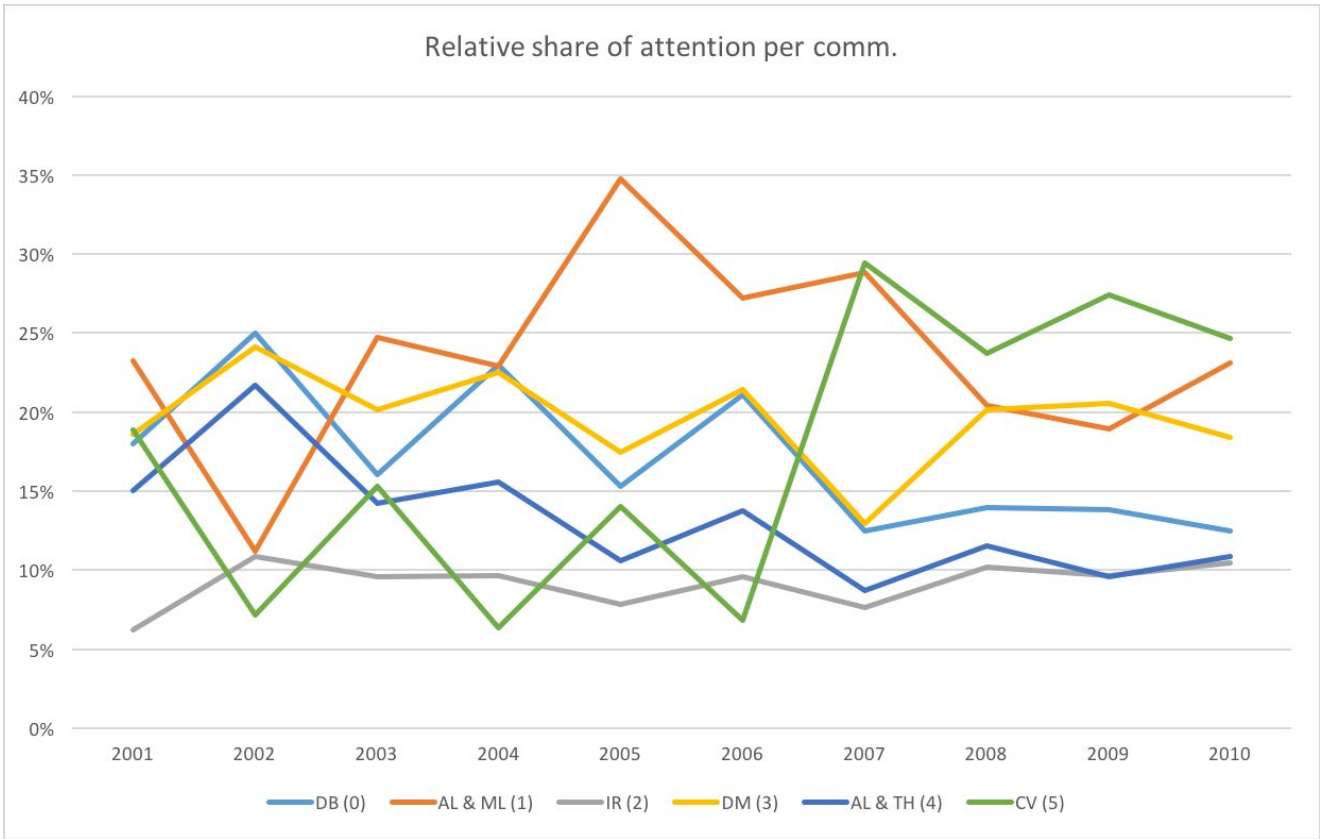

Figure 3.3: Relative share of authors within each community per year.

To analyse the obtained results further we used the graphical, computational and timeline functionalities of Gephi to perform a dynamics analysis on the ground-truth DBLP repository data that was supplied for the assignment. We used the ground-truth data since we felt this data would show a more accurate representation of the results. Included in our report are the comparisons between the nodes and edges in 2001 and 2002. We have first made a graphical representation of the division into communities in 2001, and then made a graphical representation that shows how all new nodes in 2002 are integrated into the new division. We did this by drawing the edges from the adjacency list for 2002 on top of the graphical representation of the 2001 baseline. By doing this we can extract a lot of information about how the research communities change over the years. Of course, this analysis can be extended to cover all years up to and including 2010. For the sake of brevity and due to extensive disagreements with Gephi we have not (yet) included all results in our report, but only the comparison between 2001 and 2002.

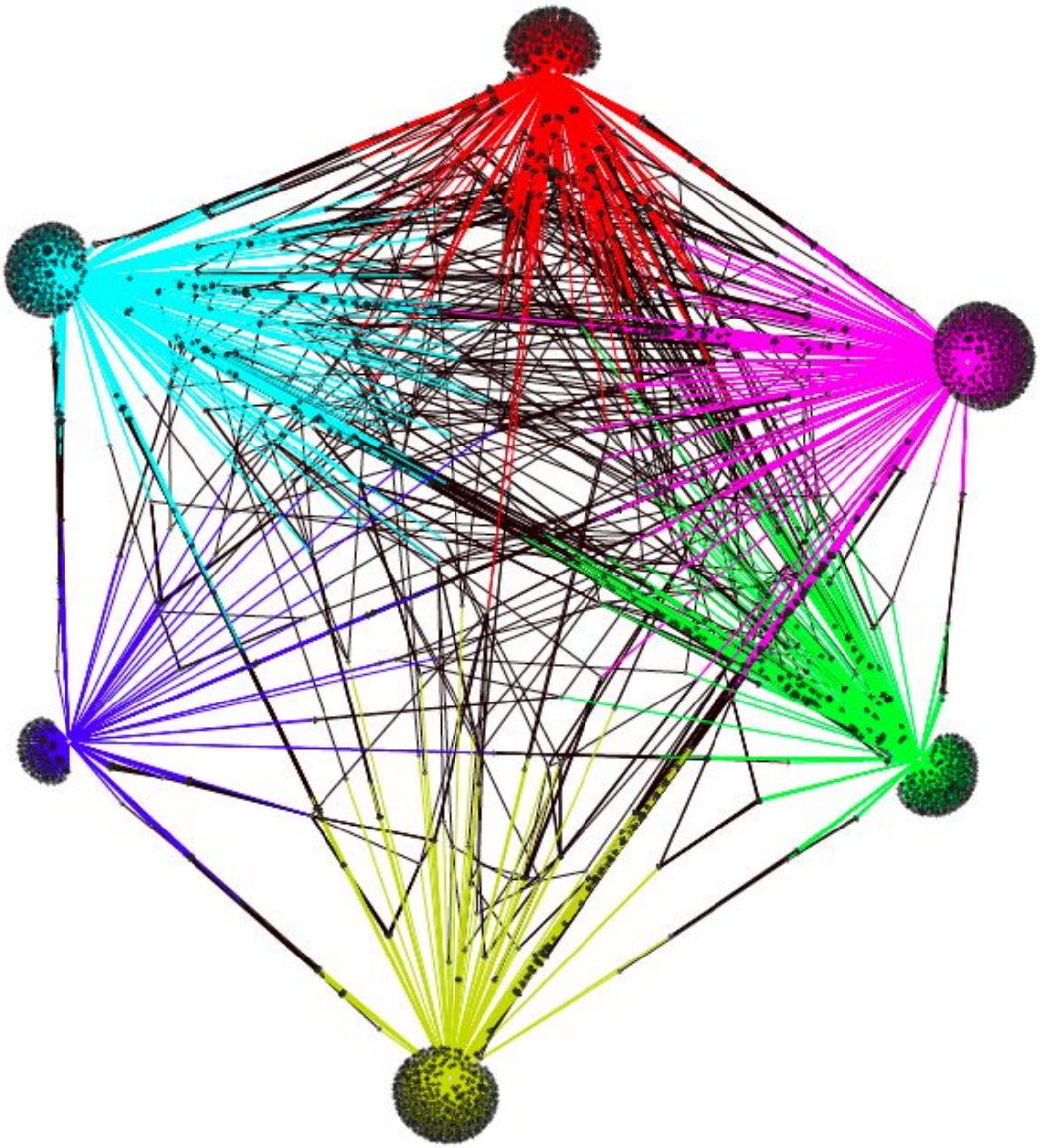

Image 3.4: Graphical overview of the clustering into communities as extracted from the 2001 ground-truth.

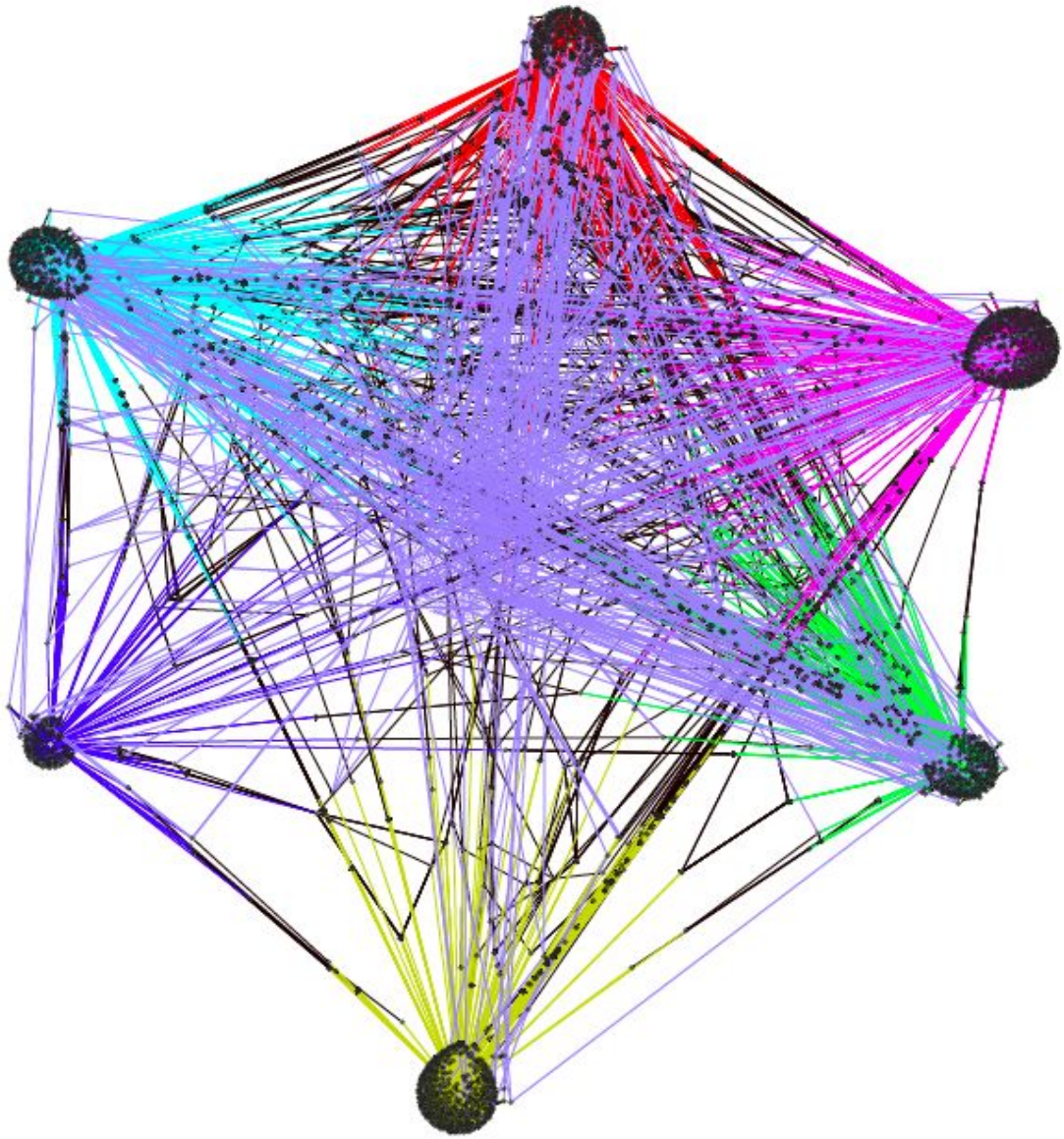

Image 3.5: Graphical overview of the clustering into communities in 2002 drawn on top of Image 3.1 for comparison.

The edges belonging to nodes in each of the communities are indicated by different colors. And the changes in the distribution of nodes when comparing 2001 to 2002 are highlighted by the many blue/purple-ish lines that crisscross between the different communities.

From the graphical representations in *Image 3.4* and *Image 3.5,* we can easily see that there are great differences in the amount of changes in connections that are made in different communities. Looking at the edges, we can see that the yellow and dark-blue communities clearly have significantly less influx than the other communities.

Peer review

We all contributed equally to this assignment.

First, Amber, Dennis and Guido started by playing around with the available input data, trying to get a bit of an idea what the data represented exactly and coming up with some ideas about how to work with it. The same group then managed to get some code running and giving useful results on the clustering based on both the text attributes and the graph inputs using `Kmeans_text.py` and `Agglomerative_graph.py`. We then also checked the results on their purity using an altered version of the provided code.
Later, we worked with all 4 members together on assignment 2 using PageRank. We actually thought to have some quite good results pretty quickly and efficiently but this proved a bit premature and Dennis together with Cengizhan had to work on this quite a bit more later on.

For assignment 3 Amber, Cengizhan and Dennis started by getting all the data in the right format for visualisation. This turned out to be the most challenging part of all but especially thanks to Dennis we eventually managed to get useful results that could be visualized. Dennis and Guido then visualised the data and the whole group analysed results to come up with the conclusions we could draw from this.

Finally, Amber, Cengizhan and Guido worked on textually reporting the information we had learned from the data as best as possible supported by the various numerical and graphical results that we had gotten by now.