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 provided us with a limited amount of information on which authors are wrongly clustered by one or both of the algorithms. We intended to get more insights on this with Gephi, however, when adding larger files to Gephi or even trying to compare 2 years with each other it has proven very hard not to crash Gephi. Of all the results that we did manage to obtain some images were exported and used for further analysis in part 3.

We also compared the results of both these scripts using a tool we came upon for task 3. This tool resulted less valuable for the comparison of these scripts than the changes in part 3 however so no clear conclusions can be drawn from this, as can be seen in the following visualisations. Sankey results will be discussed in more detailed in task 3.

http://app14.cf/static/chart/sankey_baseline.html
http://app14.cf/static/chart/sankey_text.html
http://app14.cf/static/chart/sankey_graph.html

2. Influence analysis

Results and analysis

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

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

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.

After observing these results, it seemed interesting to look if this meant that there is a correlation between the community pagerank of an author, and its h-index on Google Scholar. In order to investigate this, we filtered on the different communities and plotted the points for each author with its PageRank and h-index values in a graph. We used the Excel trendline option to see if we could derive any sort of correlation in our results for the results shown in *Table 2.2*.

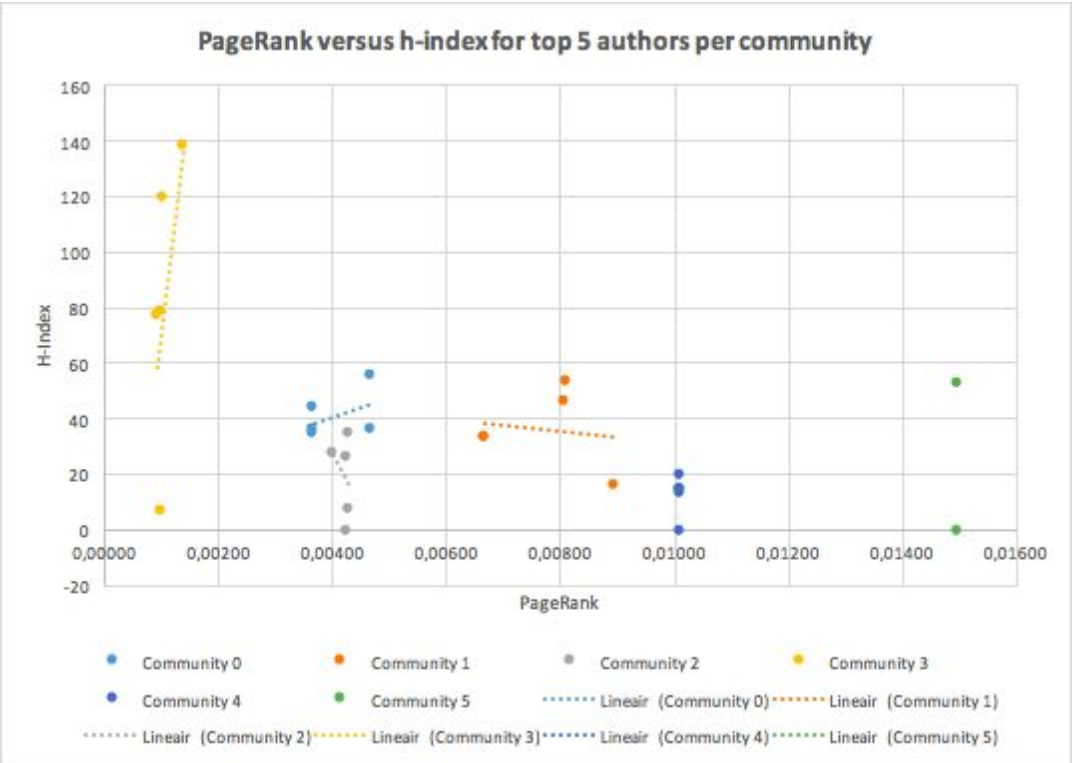


Figure 2.1: Results of PageRank versus h-index for top 5 authors per community.

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

From the above graph, we could not find any clear evidence of there being any sort of correlation (positive, or negative). Some of the trendlines seem to show a positive correlation between the h-index and the PageRank, but others again show a negative one. Perhaps this is because we have only a limited amount of data. Since the PageRank values are determined within a community, this means we can not compare all of them at once and we will need to compare them grouped by community. The h-index values for the authors had to be looked up manually on Google Scholar because their robots.txt do not permit scripts doing this automatically on such a large scale. We only looked up the top 5 for all of the authors since it is very time consuming to look up all the others. Obviously, with only 5 data points for each of the communities, it does not seem plausible that we would be able to find any reliable results from our graphs. This was confirmed by the amount of variation observed in the above figures. So, from our results, we were not able to conclude the presence or absence of a correlation between the h-index of an author and the observed PageRank within a community. However, when looking at the results it does seem like there would be a correlation and this remains an interesting point for further research when the Google h-index values can be imported on a large scale.

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. This gave us a good first idea of how the popularity of groups as a whole changed over the years, but it is lacking in detail about how many new people get in exactly because others may have left in the same year.

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.

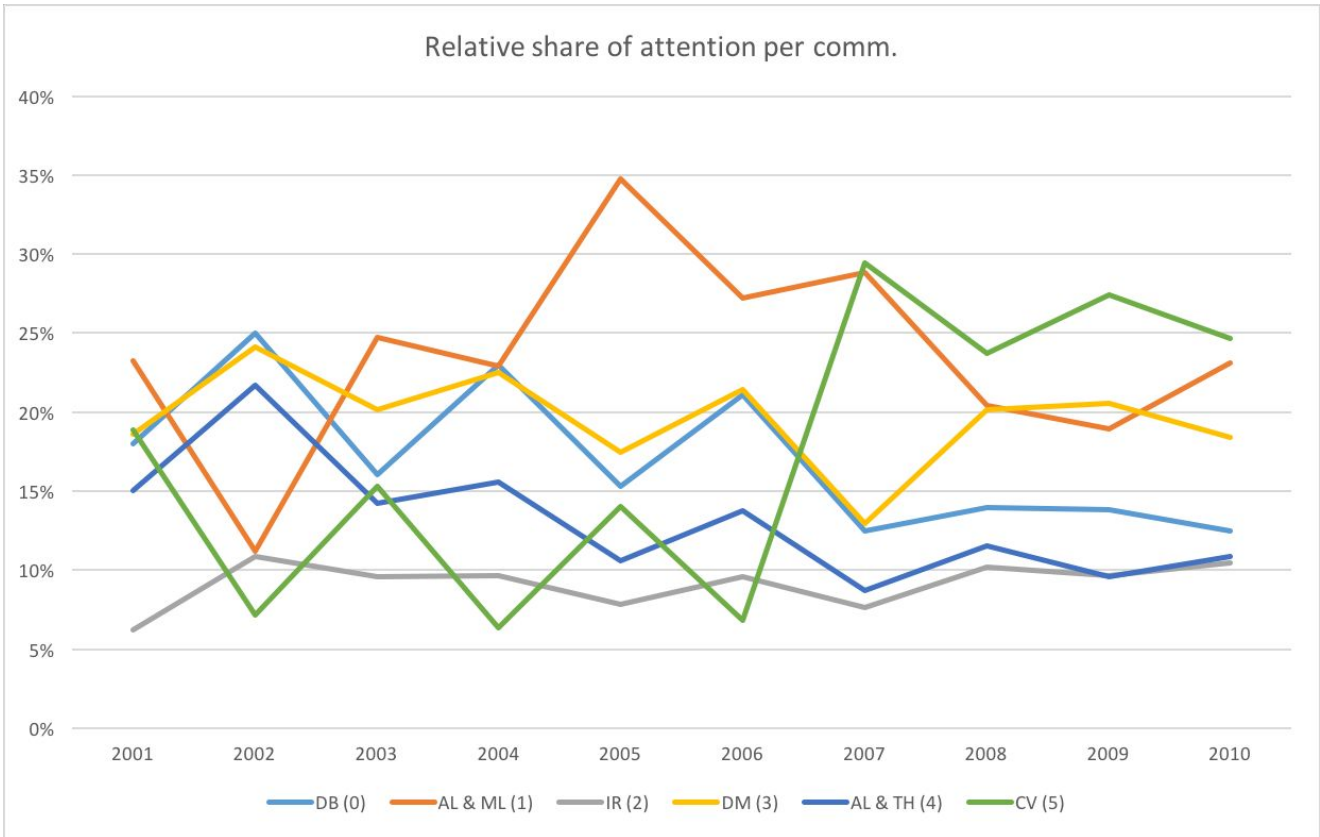


Figure 3.3: Relative share of authors within each community per year.

A problem we encountered was that we were simply summing over the years as if the same authors participate in multiple years. This is a problem, because by doing this, it was not clear how many members in communities per year are either newcomers, returnees, existing members or members switching in from another community. To tackle this problem, we created a more detailed table. This table displays changes of community members over each year. Described changes include newcomers, quitters, returnees, stayers, members that were part of another community in the previous year and joined this community (“switch in” members) and members that have quit this community to join another community (“switch out” members). We matched the numbers in this new table with the ground truth values used for the earlier table to check if they would match. The number of completely new authors + the number of switch-in authors - the number quitters - switch-out authors exactly match the match the difference in community size calculated earlier, confirming the script ran as intended.

The script uses a list of global id’s that is created using the unique names that come from the DBLP input. This way we can make sure we are comparing the exact same person moving around communities for several years. These global id’s matching with the unique names are stored in a separate file. Then it computes exactly per community which people have left,

entered, etc. and the results are visualised in a table and even better using Google Sankey. The Google Sankey visualisation gives an interactive result so it is worth opening the file named “sankey_interactive_req-internet” or visiting:
http://app14.cf/static/chart/sankey.html

The first notable insight that we obtained from table 3.4 and figure 3.5 is that the biggest changes are in authors being new entrants to the community or being completely gone from not only the last community they were in but all of them. Most notably in 2006 and 2008 some communities had lost a lot of contributors that did not contribute in a different community that year. This corresponds with the earlier line chart showing large dips in members in these communities those years. This information, combined with the result that most authors that contributed in some community in a year did not do so in the year before leads us to think that it is not very normal to contribute each year. Considering the amount of effort required to get your work recognized and published this sounds like a plausible conclusion.

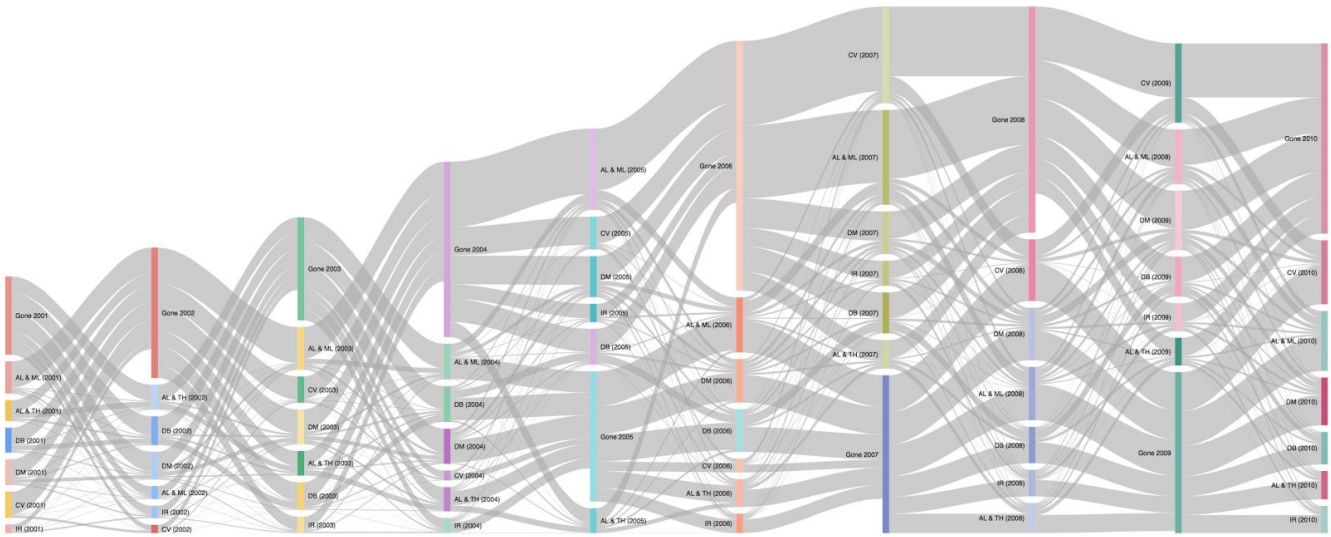
Another insight which is most visible in the Sankey visualisation is that authors are not very likely to belong to a different community in a new year. The number of authors doing this is even smaller than the amount of authors that are in a community two years in a row. Both of these observations seem logical considering our previous conclusion.

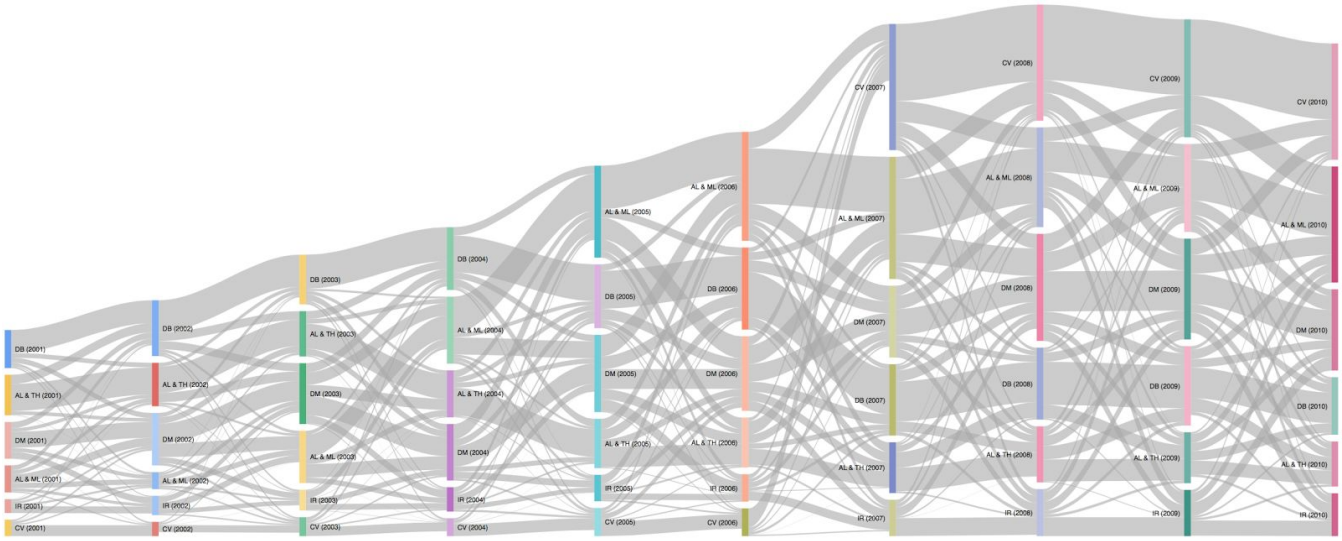
One more interesting observation is that amount of returnees (people that were already in any community in any of the years before) is still lower in the final year than the amount of new authors in that year. So in other words an author contributing to any of the fields is more likely to not have contributed anything in the 10 years before than him being a multiple contributor. This leads to think that, even stronger than before, contributors are not only not likely to publish every year, many of them are not even likely to publish more than once in ten years.

Finally, omitting the “gone” information from Sankey allowed us to ‘zoom in’ on only the people that stayed or switched communities in image 3.6. The interactive visualisation is to be found here: http://app14.cf/static/chart/sankey_only_switchers_baseline.html

This shows that of all the people that publish two years in a row, most of them do this in the same community. The rare exceptions are to be found for example in DM in 2005 and 2008, where about an equal amount of authors came from AL&ML in the year before as the amount of stayers in DM. There are also often nearly as many people going from DM to AL&ML. These facts lead us to conclude that these fields must be quite strongly connected with each other. Also, some of the thinnest lines can be seen between IR and DB, which would suggest that these are very weakly connected.

Community	Type	2001 -> 2002	-> 2003	-> 2004	-> 2005	-> 2006	-> 2007	-> 2008	-> 2009	-> 2010
DB (0)	New	424	335	398	377	418	370	253	309	259
AL & ML (1)	New	212	682	477	1118	611	1125	508	498	442
IR (2)	New	193	245	173	237	221	292	250	266	263
DM (3)	New	436	449	424	470	458	458	521	556	452
AL & TH (4)	New	354	269	227	221	275	256	254	221	235
CV (5)	New	116	409	129	429	170	1386	621	795	594
DB (0)	Quit	372	373	378	496	491	558	570	464	556
AL & ML (1)	Quit	586	206	700	475	1367	704	1521	799	794
IR (2)	Quit	126	185	268	219	281	306	382	364	401
DM (3)	Quit	397	366	482	511	541	609	604	656	869
AL & TH (4)	Quit	269	347	331	317	317	384	392	401	402
CV (5)	Quit	502	135	496	135	596	176	1546	839	1196
DB (0)	Returnee	0	77	115	137	181	199	212	197	187
AL & ML (1)	Returnee	0	117	66	333	163	489	199	338	335
IR (2)	Returnee	0	48	52	85	79	113	120	133	136
DM (3)	Returnee	0	88	104	168	166	198	193	281	219
AL & TH (4)	Returnee	0	69	89	116	106	160	175	150	177
CV (5)	Returnee	0	103	29	196	41	491	195	404	271
DB (0)	Stay	109	153	162	175	181	215	191	205	146
AL & ML (1)	Stay	25	27	111	175	207	264	263	145	194
IR (2)	Stay	33	22	28	23	30	42	85	90	73
DM (3)	Stay	76	95	92	97	94	84	124	191	144
AL & TH (4)	Stay	129	125	119	132	97	76	113	105	65
CV (5)	Stay	52	34	43	57	69	91	367	360	362
DB (0)	Switch In	106	50	120	106	169	125	151	173	128
AL & ML (1)	Switch In	50	123	140	179	242	225	212	232	359
IR (2)	Switch In	51	52	82	62	101	108	133	130	131
DM (3)	Switch In	104	140	160	174	246	206	328	289	243
AL & TH (4)	Switch In	71	84	105	83	141	142	126	138	150
CV (5)	Switch In	16	40	19	48	28	179	186	199	192
DB (0)	Switch Out	72	113	75	124	123	176	148	138	182
AL & ML (1)	Switch Out	104	54	138	144	231	255	319	238	225
IR (2)	Switch Out	33	70	71	93	96	83	88	134	145
DM (3)	Switch Out	99	155	198	172	274	271	218	319	304
AL & TH (4)	Switch Out	64	82	97	91	138	159	129	162	147
CV (5)	Switch Out	26	15	47	28	65	41	234	170	200




DB (2001)
AL & TH (2001)
DM (2001)
AL & ML (2001)
IR (2001)
CV (2001)
DB (2002)
AL & TH (2002)
DM (2002)
AL & ML (2002)
IR (2002)
CV (2002)
DB (2003)
AL & TH (2003)
DM (2003)
AL & ML (2003)
IR (2003)
CV (2003)
DB (2004)
AL & ML (2004)
AL & TH (2004)
DM (2004)
IR (2004)
CV (2004)
AL & ML (2005)
DB (2005)
DM (2005)
AL & TH (2005)
IR (2005)
CV (2005)
AL & ML (2006)
DB (2006)
DM (2006)
AL & TH (2006)
IR (2006)
CV (2006)
CV (2007)
AL & ML (2007)
DM (2007)
DB (2007)
AL & TH (2007)
IR (2007)
CV (2008)
AL & ML (2008)
DM (2008)
DB (2008)
AL & TH (2008)
IR (2008)
CV (2009)
AL & ML (2009)
DM (2009)
DB (2009)
AL & TH (2009)
IR (2009)
CV (2010)
AL & ML (2010)
DM (2010)
DB (2010)
AL & TH (2010)
IR (2010)

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 this experimental software we have not included all results in our report, but only the comparison between 2001 and 2002.

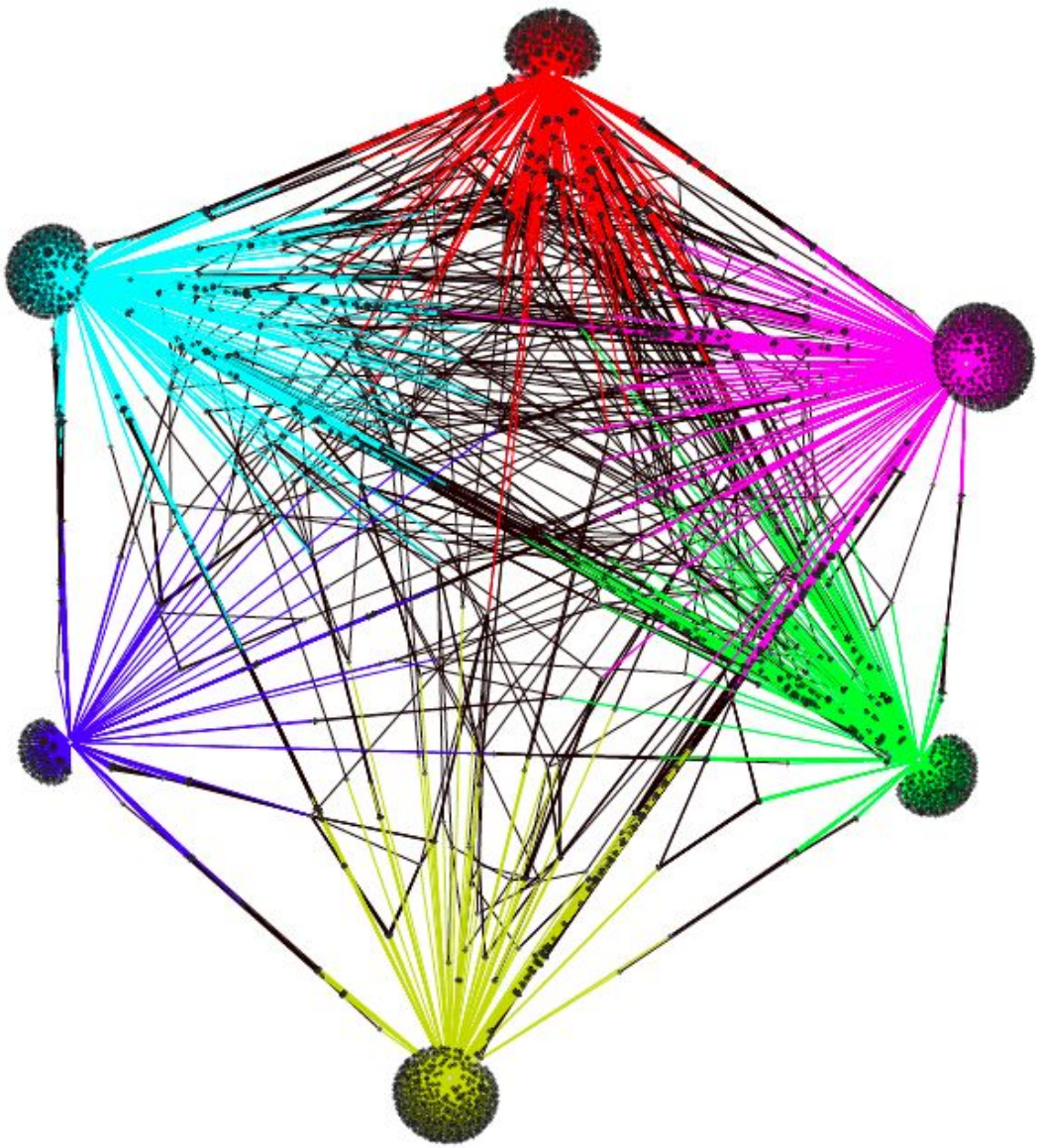


Image 3.: Graphical overview of the clustering into communities as extracted from the 2001 ground-truth.

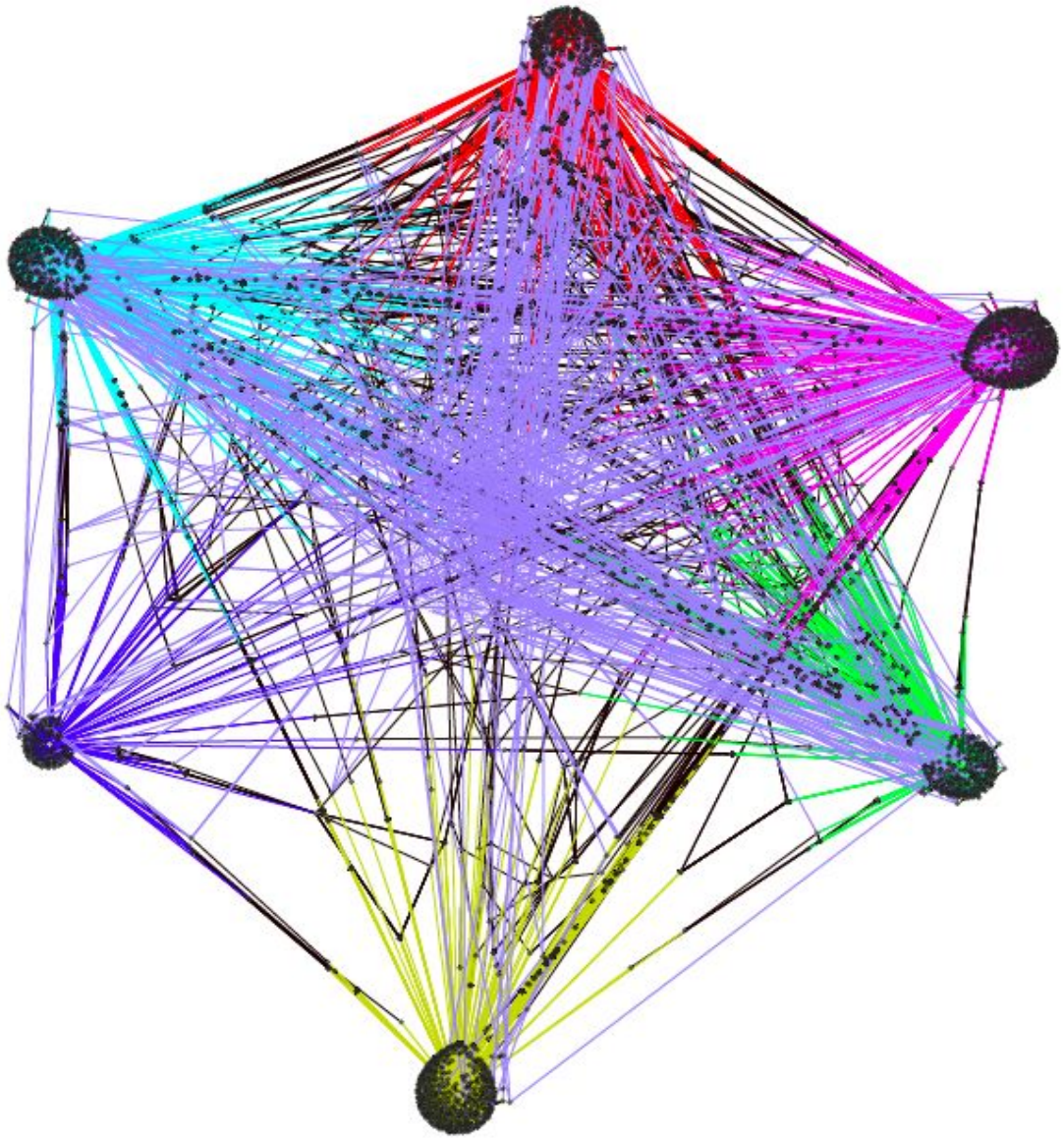


Image 3.5: Graphical overview of the clustering into communities in 2002 drawn on top of Image 3.1 for comparison.

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.

After our exams, during the last days we all worked on clarifying our findings further, obtaining more graphical output for our results and getting answers to more of the questions.