Study of Citation Network In Information System

Prepared by:

Shen Yutang (A0179964E)

Wu Tong (A011953B)

Xiong Zhe (A0133927B)

Zeng Xi (A0133950J)

Zhou Jingyuan (A0133961E)

IS4241 Social Media Network Analysis

Content

1 Introduction

To enhance the content of an original article, it is a common practice in making references to articles published in other journals. Referencing is a common social behavior among researchers, with various patterns and findings to be studied, explored and analyzed. Undoubtedly, it would be difficult to find meaningful results by examining the journals at an individual level; instead, aggregating the references at the higher journal level would help us plot a clearer picture about the connections and relationships among great many entities.

To facilitate the research on the journal citation, the concept of network is applied, so that common network techniques and analytics tools could be leveraged, such as UCINET, R-Programming and Gephi. Such network analysis would be helpful in understanding the characteristics of the network, analyzing the citation behavior and deducing the reference patterns among the research sample.

2 Network Overview

The full dataset used in this project comes from the database Clarivate Analytics.

In the citation network, a journal is modelled as node and the citation from one journal to the other is modelled as the edge. We treat the graph as directed graph as two journals may cite publications from each other. Also, there will be self loops as publications in the same journal may cite among themselves. The weightage on the edge represents how many times the specific citation has been made. Besides, each journal has an attribute named “impact factor”, which is a measurement of the importance level of journals.

The network starts from three well-established journals in the area of information systems: MIS QUART, INFORM SYST RES, and J ASSOC INF SYST. By referring to the Journal Citation Report by Clarivate Analytics, the two-degree egocentric network would be plotted, but only strongly connected neighbours which cite at least 3% of ego would be present on the network. Through this criteria, five journals with strong connections could be identified as the “beginning” journals to start the network analysis, including MIS QUART, INFORM SYST RES, J ASSOC INF SYST, COMPUT HUM BEHAV and INFORM MANAGE-AMS. For convenience, this set of five journals would be referred as s-journals (selected journals) in the following sections.

Starting from the s-journals, dataset consolidating journals that cited or are cited by them are downloaded from the database directly. Treating them as egos, a web crawler was implemented and used for collecting connections among alters as well as their weightage. The crawler can be accessed in Appendix A.

The full set of citation data used in the project was extracted from the database Clarivate Analytics, so that the network chart with all connections starting from s-journals could be plotted. Although in actual world, the journal citation network should not be centered around a few selected journals and the citation network is much more complicated, the dataset used in this project can be used as a test data set for analysing the journal citation network.

3 Network Visualization & Characteristics

By following the steps mentioned above, the overall network could be plotted with the assistance of the application Gephi. The journal network described in detail in this paper is composed of 1444 journals. Each journal that has a in-degree larger than 1, including self-citation, is represented in Figure 3.1 by a node. Nodes are sized and colored in a green-blue scale proportional to the number of journals this node has been cited by. Between nodes, directed, weighted edges are drawn that account for the total number of citations that edge source has made to edge destination up to 2016. Edges are colored in a yellow-blue scale ranked by weight. Leaves are hidden on the diagram to highlight only active nodes.

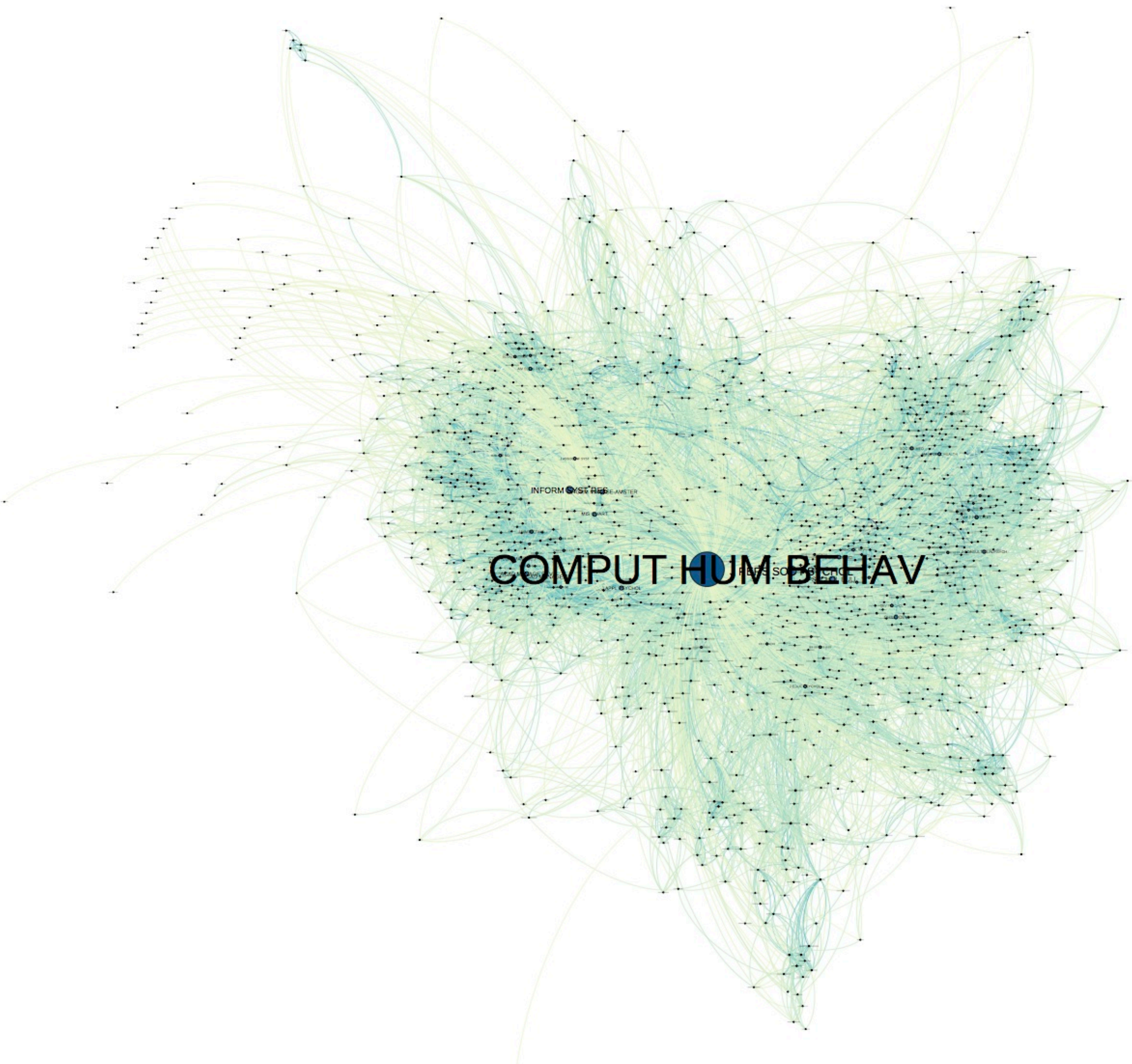

Figure 3.1 Netwoork Graph Overview

After plotting the citation network as shown above, to provide a clearer picture on the journal citation relationships, in-degree and out-degree centralities were calculated for each citation object. Due to space limitation, only a snapshot of citations with its network metrics are listed below, but the full table could be available at the appendix files, ranking by alphabetical order.

From	To	Weighted In-degree Centralities	Weighted Out-degree Centralities

ACAD MANAG ANN	AM ECON REV	8796	1916
ACAD MANAG LEARN EDU	EUR J PUBLIC HEALTH	653	441
ACAD MANAGE J	J MARKETING RES	5491	1375
ACAD MANAGE PERSPECT	SOC FORCES	1061	876
ACAD MANAGE REV	AM J PUBLIC HEALTH	6919	668
ACAD-REV LATINOAM AD	COMPUTING	0	4
ACCIDENT ANAL PREV	PSYCHOL MARKET	620	1478
ACCOUNT AUDIT ACCOUN	PSYCHOPHYSIOLOGY	2453	1334
ACCOUNT FINANC	PSYCHOLOGIA	36	2
ACCOUNT ORG SOC	QME-QUANT MARK ECON	119	156
…...			
WRIT COMMUN	J RES EDUC EFF	52	213
YALE LAW J	EDUC RESEARCHER	594	211
YOUNG	PSYCHIAT PRAX	63	137
YOUTH SOC	J APPL DEV PSYCHOL	347	778
YOUTH VIOLENCE JUV J	SUBST USE MISUSE	421	2171
Z KL PSYCH PSYCHOTH	J TELEMED TELECARE	0	2
Z PADAGOG PSYCHOL	INT J LANG COMM DIS	80	180
Z PSYCHOL	J CONSUM BEHAV	179	848
Z PSYCHOSOM MED PSYC	J BLACK STUD	27	41
ZDRAV VARST	J UNIVERS COMPUT SCI	2	0

Figure 3.2 Network Degree Centralities (Partial)

4 Network Measurement

In order to analyse the importance level of nodes in the network, we measured these network statistics: weighted and unweighted in degree centrality, weighted and unweighted out degree centrality, betweenness centrality, closeness centrality as well as eigenvalue.

4.1 Betweenness Centrality

Betweenness centrality is a measure of the frequency a given vertex lies on the shortest/geodesic path between two vertices. A journal with high betweenness centrality has more power in controlling the information flow. For instance, if journal A has high betweenness centrality, it may be easier to find another journal through it, and it will have higher possibility to be a key-player in the citation network. In our citation network, betweenness centrality can be used as an indicator of the "interdisciplinarity" of journals. It can also be used to measure the degree to which a certain journal can behave as a point of control in the communication of journals.

4.2 Degree Centrality

As the network is a directed network, four degree centrality was calculated: unweighted in-degree centrality, unweighted out-degree centrality, weighted in-degree centrality as well as weighted out-degree centrality. In-degree centrality can be used to analyse the number of journals that cite a certain journal. Journals with higher in-degree centrality may be more authoritative as many other journals cite them. Also, the content they cover in the publications might be reliable. On the other hand, journals with higher out-degree centrality cite other journals more frequently. Therefore, the content that they include may be broader or more comprehensive. Also, in the case that a journal has more citations from other journals, it may be more objective.

In our network, there are edges with high weight. This makes the weightage of links unbalanced with extreme values. Therefore, both weighted and unweighted degree centralities were calculated to analyze the network from different perspectives.

4.3 Closeness Centrality

Closeness Centrality measures the sum of shortest/geodesic distances a node to all other nodes in the network. In our network, a higher closeness centrality may indicate a stronger power to access to other nodes. In other words, from journals with high closeness centrality, it may be easier to access to a larger number of journals quickly, or with smaller average total effort. Compared with betweenness centrality which measures the local position of a journal, closeness centrality provides a global measure about the position of a journal in the whole citation network (Leydesdorff, n.d.). Also, closeness centrality depends less on the relations between individual pairs of journals, it can therefore provide a measure of "multidisciplinarity" of journals.

4.4 Eigenvector Centrality

Eigenvector centrality is a measurement of nodes in a network that considers both their number of edges and quality of edges. To illustrate, if a journal has a large amount of links to other journals, it may not have high eigenvector centrality when the quality of citations is not good. In contrast, a journal may have high eigenvector centrality if it has good citations even if the amount of citations is not high.

As introduced by Shankar (2015), the eigenvalue of a vertex in a network is dependent on the centralities of its neighbours as well as the leading eigenvector (the vector with the highest eigenvalue). However, our citation network is not suitable for analysis using eigenvector centrality as except the 5 selected journals, for all other journals', not all of their neighbours are included in the network (the fact is that most of them are neglected as they are not neighbours of the 5 selected journals). Therefore, eigenvector centrality will be used only for discussion and information in this report.

5 Network Analysis by Regression

Journal impact factor has been a significant tool to measure the quality of a journal for a long time. However, the use of the term “impact factor” has gradually evolved to include both journal and author impact, especially in Europe (Garfield, E), and thus it often cause ambiguity of comparing journals or comparing authors. Therefore, we are interested in figuring out if the impact factor the journal has direct impact on the citation numbers.

In order to verify if the difference in impact factors of the journals has impact on the citation numbers from a journal to another journal, we have done both node-level and edge-level regression analysis.

5.1 Edge-Level Regression Analysis

For edge-level regression analysis, we chose the difference in impact factors of citing and cited journals as the independent variable, and the citation numbers from citing journal to cited journal as the dependent variable. We set the null hypothesis that there does not exist a linear relationship between the two variables. Then we perform the linear regression analysis using the lm() method in RStudio (Figure 5.1).

```
> attach(data)
> regression3<-lm(Weight~ImpactFactorFromToDiff)
> summary(regression3)

Call:
lm(formula = Weight ~ ImpactFactorFromToDiff)

Residuals:
   Min      1Q Median      3Q     Max
-135.7   -59.5  -39.1     5.8  3663.3

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             73.0580     1.1665  62.627  < 2e-16 ***
ImpactFactorFromToDiff  -1.1117     0.2299  -4.835 1.35e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 144 on 15706 degrees of freedom
Multiple R-squared:  0.001486,	Adjusted R-squared:  0.001422
F-statistic: 23.37 on 1 and 15706 DF,  p-value: 1.346e-06
```

Figure 5.1: Linear Regression Between Edge Weight and Impact Factor From-to Difference

As indicated in the output, the estimated coefficient of ImpactFactorFromToDiff is -1.1117 with p-value(1.346*10^(-6)) which is much smaller than the cut-off significant value of 5%. Therefore we can reject the null hypothesis, meaning that the difference in impact factor of two journals does have impact on the citing numbers from one journal to another.

This regression result is actually in line with the reality. The journal with lower impact factor will cite the journal with higher impact factor more times in order to increase its own objectivity and credibility.

5.2 Node-Level Regression Analysis

From our intuition and understanding, we felt that the impact factor of a journal may be affected the times it is cited by other journals. Therefore, with weighted in-degree centrality and impact factor of all the nodes in the network, we would like to do a linear regression analysis to examine our hypothesis.

We chose the weighted in-degree centrality of the journal as the independent variable, and its impact factor as the dependent variable. We set the null hypothesis that there does not exist a linear relationship between the impact factor of the journal and its weighted in-degree centrality. Then we perform the linear regression analysis using the lm() method in RStudio (Figure 2).

```
> Regression2<-lm(ImpactFactor~WeightedInDegree)
> summary(Regression2)

Call:
lm(formula = ImpactFactor ~ WeightedInDegree)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.607  -1.000  -0.471   0.328  69.628 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.9852490  0.0870891  22.796   <2e-16 ***
WeightedInDegree 0.0004643  0.0000479   9.692   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.104 on 1442 degrees of freedom
Multiple R-squared:  0.06116,	Adjusted R-squared:  0.06051 
F-statistic: 93.94 on 1 and 1442 DF,  p-value: < 2.2e-16
```

Figure 5.2: Linear Regression Between Weighted In-degree Centrality of the node and its Impact Factor

As shown in the Coefficients table, the estimates for the model intercept is 1.9852490 and the coefficient measuring the slope of the relationship with weighted in-degree centrality is 0.0004643. We can see that the p-value is much smaller than the cut-off significant value of 5%, so we can reject the null hypothesis, meaning that the weighted in-degree centrality of a node (i.e. the weighted times it is cited by other journals) does have impact on its impact factor.

However, one thing to note is that the adjusted R-Squared is 0.06051, meaning that the model only fits 6.05% of the data. The result can be explained by the fact that the dataset we have obtained is limited. We obtained the dataset starting from the 5 journals in Information Systems field and tried to find its 1.5-degree ego network. For many of the journals in the network, the numbers it is cited by other journals outside the network could not be displayed, and thus its weighted in-degree centrality is not the real in-degree centrality. Therefore, the low adjusted R-Squared could be justified due to the limitation of the data collected.

Apart from examining the relationship between weighted in-degree centrality of the node and its impact factor, we are also interested in performing more regression analysis to find out if there exists relationship between impact factor of a journal and its other centralities such as betweenness centrality, closeness centrality, and eigenvalue. Therefore we have performed linear regression analysis for those centralities in RStudio as well, and set the respective null hypothesis as being that impact factor of a journal does not affect its betweenness centrality, closeness centrality, and eigenvalue. The regression results are shown in figure 3 - figure 5 below.

```
> Regression4<-lm(Closeness.Centrality~ImpactFactor)
> summary(Regression4)

Call:
lm(formula = Closeness.Centrality ~ ImpactFactor)

Residuals:
       Min         1Q     Median         3Q        Max 
-7.215e-06 -2.945e-07  3.663e-07  6.360e-07  9.376e-07 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.692e-06  3.240e-08 237.412   <2e-16 ***
ImpactFactor 8.575e-10  8.245e-09   0.104    0.917    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.003e-06 on 1442 degrees of freedom
Multiple R-squared:  7.501e-06,	Adjusted R-squared:  -0.000686 
F-statistic: 0.01082 on 1 and 1442 DF,  p-value: 0.9172
```

Figure 5.3: Linear regression between impact factor of a journals and its betweenness centrality

```
> Regression3<-lm(Betweenness.Centrality~ImpactFactor)
> summary(Regression3)

Call:
lm(formula = Betweenness.Centrality ~ ImpactFactor)

Residuals:
    Min      1Q  Median      3Q     Max 
 -18758   -3860   -3586   -2028 1655839 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)    3430.1     1525.1   2.249   0.0247 *
ImpactFactor    212.8      388.2   0.548   0.5836  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47220 on 1442 degrees of freedom
Multiple R-squared:  0.0002084,	Adjusted R-squared:  -0.0004849 
F-statistic: 0.3006 on 1 and 1442 DF,  p-value: 0.5836
```

Figure 5.4 : Linear regression between impact factor of a journals and its closeness centrality

```
> Regression5<-lm(Eigenvalue~ImpactFactor)
> summary(Regression5)

Call:
lm(formula = Eigenvalue ~ ImpactFactor)

Residuals:
    Min      1Q  Median      3Q     Max 
-2028.4   -30.7    -0.5     5.4  3181.4 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.536      7.511  -0.471    0.638
ImpactFactor    1.551      1.912   0.811    0.417

Residual standard error: 232.6 on 1442 degrees of freedom
Multiple R-squared:  0.0004561,	Adjusted R-squared:  -0.0002371 
F-statistic: 0.658 on 1 and 1442 DF,  p-value: 0.4174
```

Figure 5.5: Linear regression between impact factor of a journals and its eigenvalue

As shown in Figure 5.3 to 5.5, p-values are 0.9172, 0.5836, and 0.4174 respectively, which are much larger than the cut-off significant value of 5%. We fail to reject the null hypotheses, meaning that the impact factor of a journal generally does not have impact on its betweenness centrality, closeness centrality, and eigenvalue.

6 Clustering Analysis

6.1 Methodology & Rationale

The Louvain Community Detection Method is applied to obtain meaningful clusters in the journal network. This algorithm is chosen because it is computationally fast compared to other known community detection methods, produces high-quality communities as measured by modularity and has verified accuracy on ad hoc networks (Blondel et al., 2008). More importantly, the method's validity has been tested on commonly used test-case networks that include a network of 9000 scientific papers and their citations, which is similar in nature to the journal network discussed in this paper.

6.2 Size Distribution

The community detection algorithm was configured to randomise the nodes to produce a higher-quality decomposition and to use edge weights that are directly related to a journal's impact factor. A resolution value of 2.5 is used as recommended by the method's guide, to obtain a suitable number of appropriately-sized communities that are conducive for meaningful analysis. As a result, the community detection algorithm has identified a total of 6 communities on the journal network. As shown in Figure 6.1, the size of these 6 communities ranges from having 21 nodes to having 479 nodes.

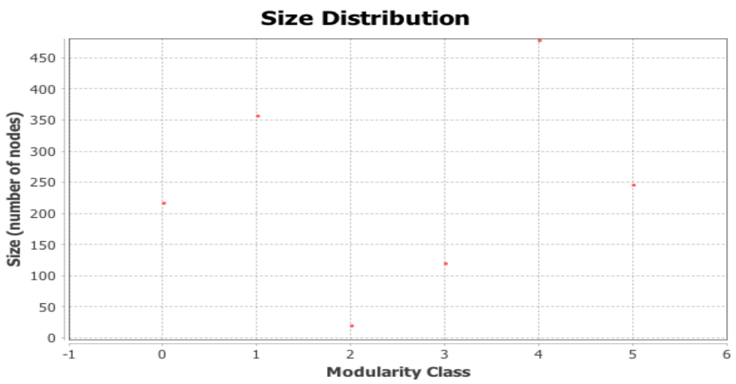

Figure 6.1: Size distribution of the detected communities, using randomisation, edge weights and a resolution of 2.5.

6.3 Clustering Results Discussion

Figure 6.2 shows these communities partitioned into 6 different colors. Each journal is indicated by a colored circle corresponding to the community that it belongs to. Orange, green, indigo, pink, purple and blue indicates community 0, 1, 2, 3, 4 and 5 respectively. The size of a circle is proportional to the number of journals that have referenced it before the year 2017. A journal is labeled using its name in a shortened form and the labels are sized proportional to the size of the circle, so that journals cited by the greatest number of publications, such as *Computers in Human Behavior*, are displayed in the most notable font. Journals are positioned relative to other journals based on repulsion strength. Distance between nodes are proportional to how strongly nodes reject each other. Considering only direct citations, those that are strongly related in terms of citations are placed closer to each other, and those weakly related tend to be located further apart. The curved edges with arrows indicate citation relations, with the arrow leaving the citing journal and pointing at the cited journal.

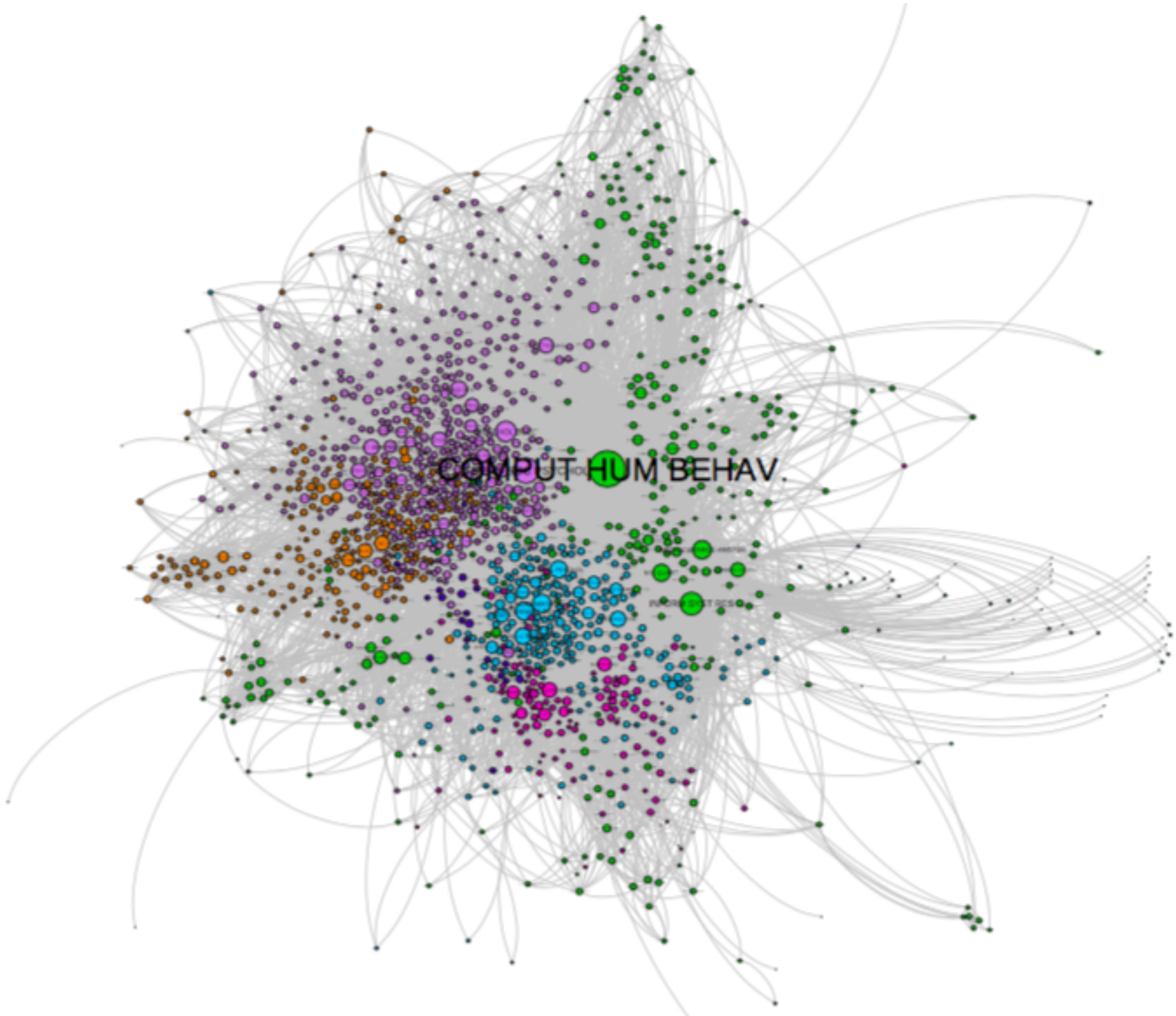

Figure 6.2: Graphical representation of the network of communities extracted from a network of journals and their citations

This network represents 1444 academic and scientific journals and their citations made to one another before 2017. The size of a node is proportional to the number of journals it has been cited by, and its color corresponds to the community that it belongs to. Only journals that have been cited by more than one journal, including itself, have been shown.

Figure 6.2 shows that the communities of journals are weakly segregated, with numerous node overlap between communities. Analysis of the underlying journal attributes reveals that journals in one community tend to be from fields of research that are deeply interlinked and center on a common theme. On a higher level, the position of a community relative to others reflects the connections between these fields of research.

To illustrate, community 0 (orange) is largely comprised of research publications on health and medicine. The topics of these publications range from public health, professional healthcare, health education to medical research and research methodologies. Under each topic, there is a variety of subtopics that a journal focuses more specifically on. For instance, within a collection of journals that can be classified under public health, there are journals dedicated to examining health policies, alcohol addiction and substance abuse individually; under professional healthcare, there are nursing education, patients education and rehabilitation; under health education, family health, sex health and mental health; under medical research, cancer and AIDS; and under research methodologies, assistive technologies and health informatics.

Community 4 (purple) chiefly consists of journals in psychology. A closer examination reveals that journals in this community covers a vast range of research and applied specialities in the psychology discipline, such as clinical psychology, developmental psychology, cognitive psychology and educational psychology. Psychology also has the largest number of publications, as community 4 has the most number of nodes.

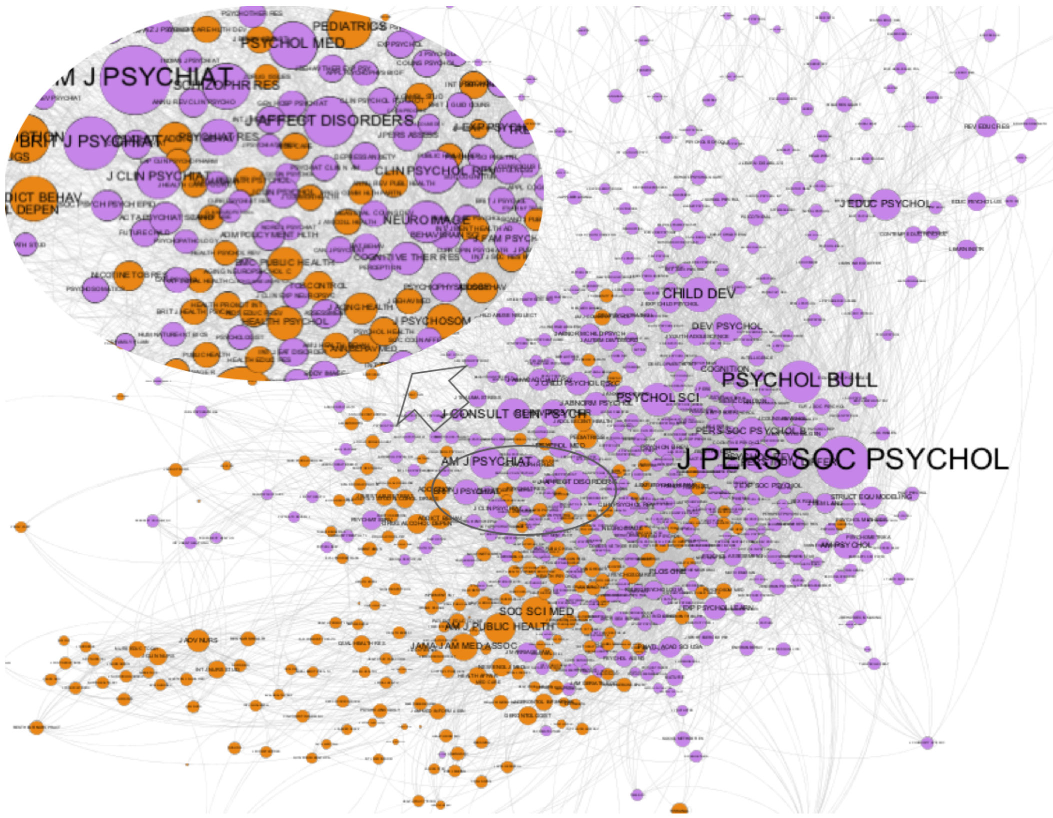

Figure 6.3: Network of journals in community 0 (orange) and community 4 (purple) (Size of a node and its label is proportional to its impact factor)

Figure 6.3 shows that research in the field of psychology and health is closely connected, as a significant overlap between community 0 and 4 is observed. A drilled-down view on the journals situated on this overlap reveals that they mostly focus on health issues where psychological factors play a significant role, such as addictions and mental health, or on the application of psychological studies in diagnosing and treating mental disorders, such as schizophrenia.

As shown in Figure 6.3, the presence of community 1 (green) is sparsely spread across the network, with the majority of nodes floating on the rim and away from other communities. Closer analysis reveals that this is due to the diverse set of scientific studies that journals in this community cover. By taking the relative position of nodes as an indicator, it can be seen that journals in this community span many fields of research. In Figure 6.3, nodes on the lower left corner cover sociology and further to the left, journals on criminology and justice stand out. On the bottom of the network, nodes mostly cover library science. On the right, nodes cover information systems and as they go up, they cover computer science. Nodes on further right cover communication and nodes further up cover education. Nodes at the top cover language. At the center, the few nodes from community 1 that are located inside other communities cover fields related to their neighbor journals, such as communication in community 4 and commerce in community 5. Community 1 aptly demonstrates how positioning journals based on their citations can reveal their relevance in content. Interestingly though, it is difficult to make an overarching connection among this diverse set of fields discovered in community 1.

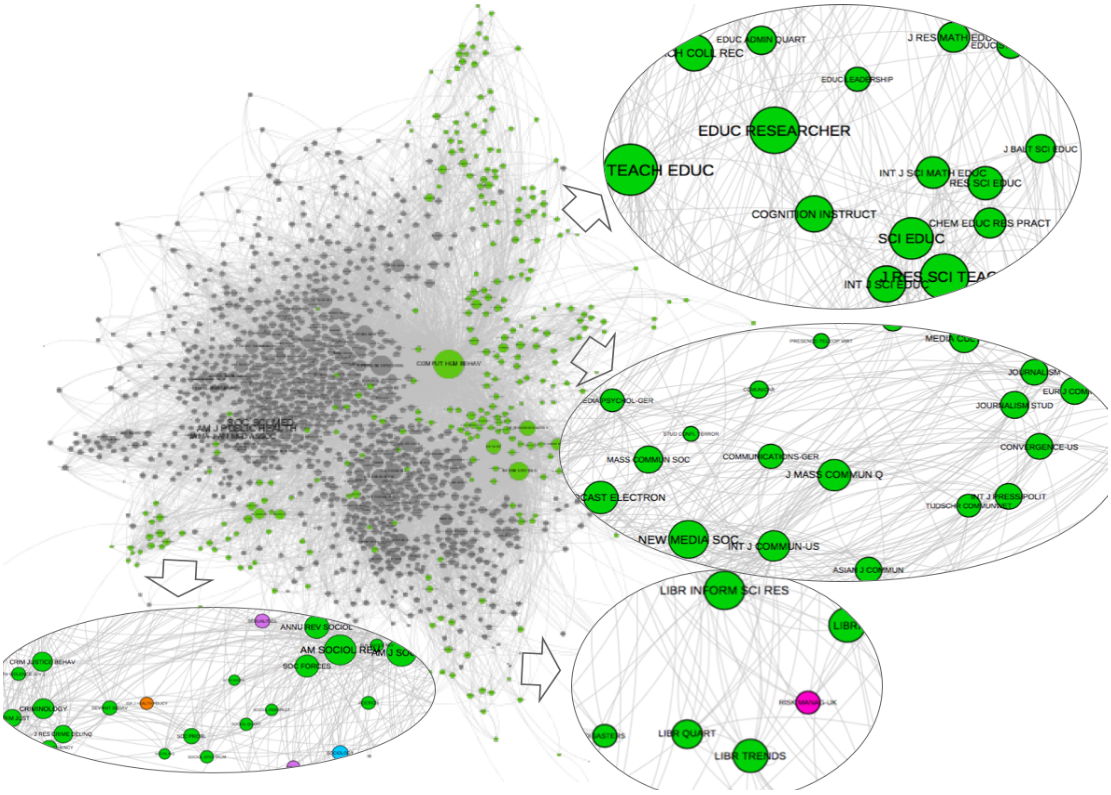

Figure 6.4: Network of journals in community 1 shown in green and journals in other communities shown in grey in the background (Size of a node and its label is proportional to its impact factor)

Figure 6.4 shows three more communities-community 2 (indigo), 3 (pink) and 5 (blue). Community 2 is the smallest community and covers work and transportation. This community reveals an interesting relationship between two ostensibly unrelated fields, as closer analysis shows that work and transportation consist of subtopics-safety, ergonometrics, and mobility, that share a theme of enhancing efficiency.

Community 3 centers on economics and business administration. Journals in this community generally highlights research and findings in econometrics, finance, accounting, management science, decision science, supply chain, logistics, operation and total quality. Readers of these publications are offered two perspectives to look at a business through the lens of its internal functions as well as its external market environment.

Journals in community 5, highlighted in bright blue and located at the heart of the network, publish insights on organisation management. This set of journals include works of applied psychology, organisational science, strategic management, marketing and hospitality.

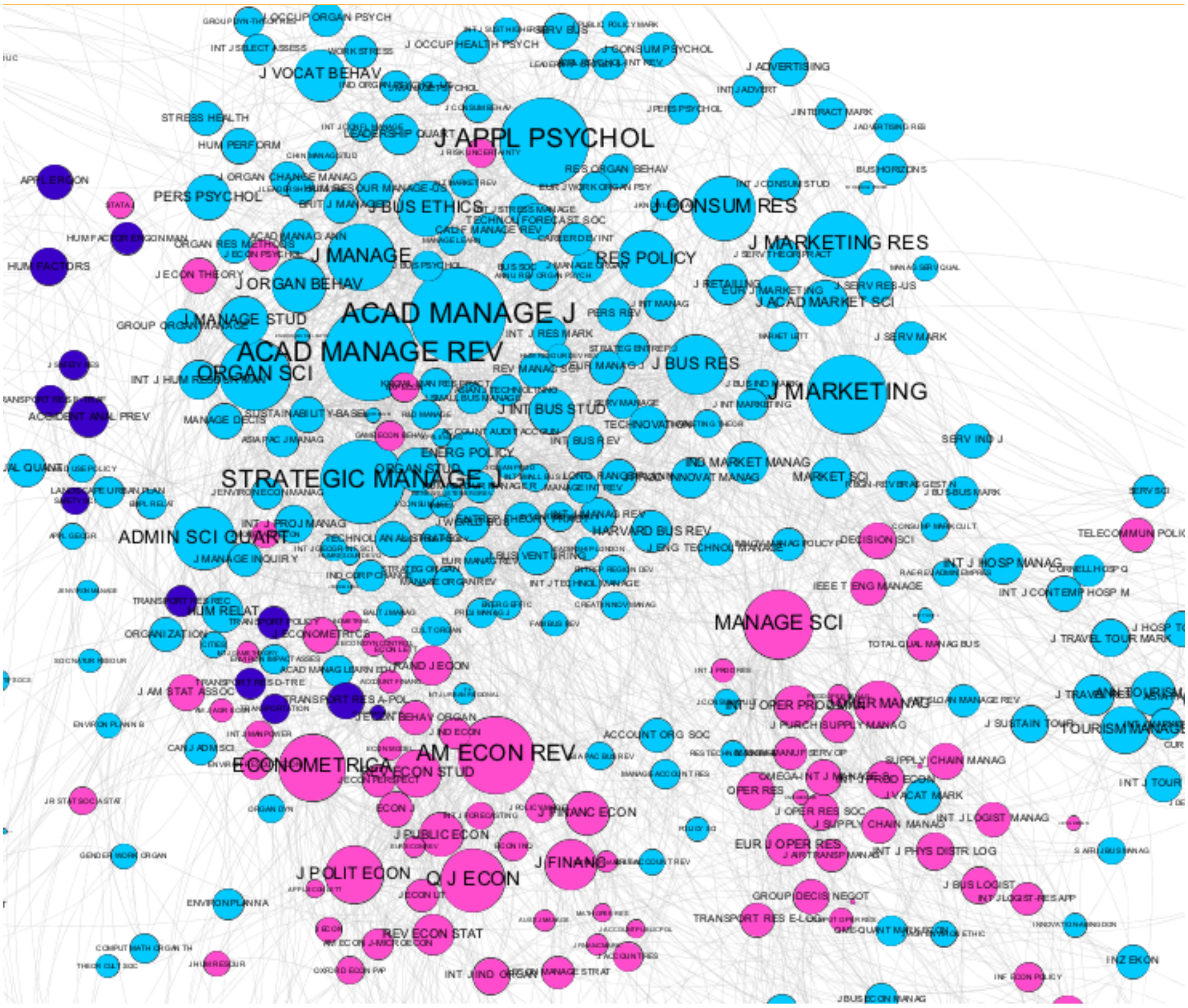

Figure 6.5: Network of journals in community 2 (indigo), 3 (pink) and 5 (blue) (Size of a node and its label is proportional to its impact factor)

An interesting observation is related to the presence of nodes from one community in the midst of a group of nodes from another community. They are relevant in published content to nearby journals and therefore often contain cited information from their neighbors, but are not substantially different in content from their own community or sufficiently close in content to the community they inhabit. These journals, many located at the boundary of its community, are often deemed as an interfacing link that bridges two or more fields of studies. Bringing together the distinctive components of several fields, these journals' interdisciplinarity quality binds the network. From a research point of view, the possibility to replicate, develop or test, evaluate and build up on the findings documented in these journals opens new perspectives for exploring uncharted fields of scientific research and for extending practical insights from theoretical bases.

7 Key Journal Analysis

Besides the impact factor of journals, network measurements discussed previously (degree centrality, closeness centrality, betweenness centrality and eigenvector centrality) are also used for analysing the level of importance of journals. Figures below includes all the top 5 journals using each measurement.

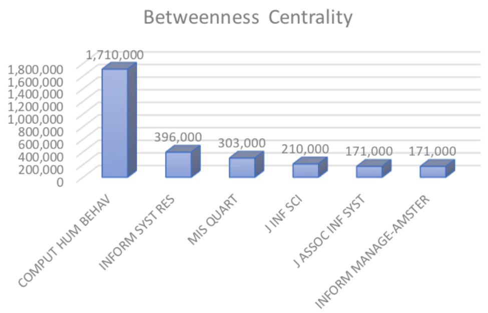

Figure 7.1 Top 5 Journals with Highest Betweenness Cetralities

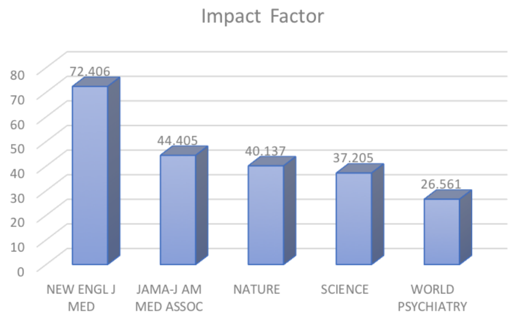

Figure 7.2 Top 5 Journals with Highest Impact Factors

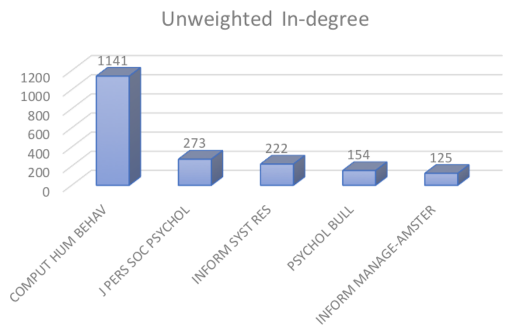

Figure 7.3 Top 5 Journals with Highest Unweighted In-degree Centralities

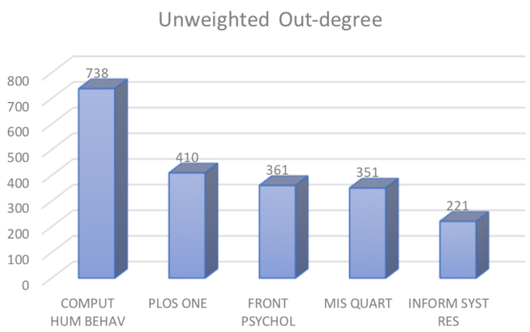

Figure 7.4 Top 5 Journals with Highest Unweighted Out-degree Centralities

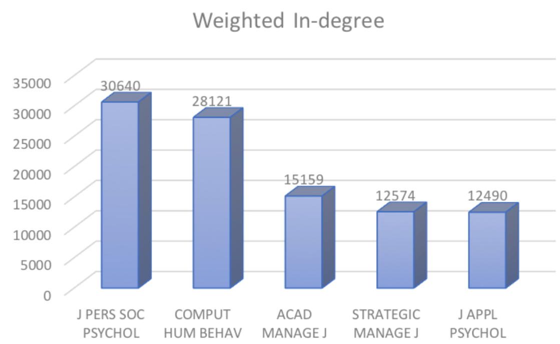

Figure 7.5 Top 5 Journals with Highest Weighted In-degree Centralities

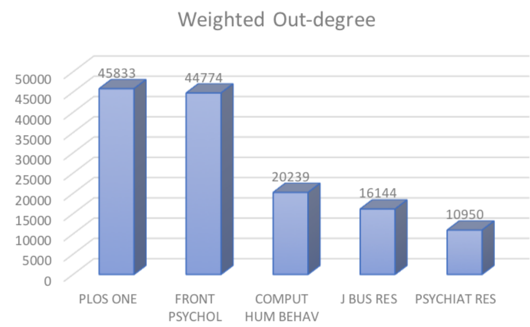

Figure 7.6 Top 5 Journals with Highest Weighted Out-degree Centralities

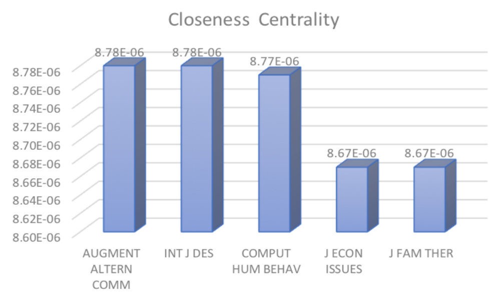

Figure 7.7 Top 5 Journals with Closeness Centralities

As discussed earlier in the measurement section, different measurement can be analyzed to determine a journal's characteristics and location in the network. From the figures above, we noticed that

"COMPUT HUM BEHAV" has high rankings in 5 of the 8 measurements. Therefore, the analysis starts with this journal.

- Firstly, "COMPUT HUM BEHAV" is ranked highest in betweenness centrality, and the value is much more than the second ranked journal. This means that it locates in the shortest paths between many pairs of journals. As betweenness centrality can be used to measure interdisciplinarity of journals (Leydesdorff, n.d.), with the fact that "COMPUT HUM BEHAV" has high betweenness centrality, the removal of it from the citation network may result in the network fall apart into less connected, in the way that some journals may take longer path to reach each other.

- Secondly, "COMPUT HUM BEHAV" has high degree centrality. High in-degree centrality indicates that it has been cited many times by other journals. This may because of its authoritativeness as well as comprehensive content it covers. Also, a broad coverage of other journals (as shown by the high out degree centrality) makes it more objective as introduced earlier. The fact that "COMPUT HUM BEHAV" does not have ranks in weighted degree centrality as high as that in unweighted degree centrality is due to the fact that there are extreme values in weights of links in the network and make the network unbalanced in a way. However, our groups think that unweighted degree centrality should be considered as more significant as it stands for a broader citation out from / in to the journal.

- In addition, "COMPUT HUM BEHAV" has high closeness centrality (almost the same as the first two journals). We can infer that from "COMPUT HUM BEHAV", readers can access to a large number of other journals with little effort. Also, its high multidisciplinarity makes it in a central location in the network that makes it easy to access to journals in different academic fields.

- Besides, its relatively low ranking in impact factor can also be explained. As introduced earlier, impact factor is calculated by total number of citations dividing by total publications in the journal. Therefore, a low impact factor of "COMPUT HUM BEHAV" should be due to its large amount of publications. This can also be indirectly proven by its high out-degree centrality: as it cites many other journals, its number of total publications must be huge.

Furthermore, although some other journals in the citation network are also important, they are not as significant as "COMPUT HUM BEHAV".

- "NEW ENGL J MED"(and other journals having high impact factor) has the highest impact factor, however, it has low rankings in all other measurements. This may because of its small number of publications.

- "J PERS SOC PSYCHOL" has higher weighted in-degree centrality than "COMPUT HUM BEHAV", its unweighted in-degree centrality is lower. This may due to its high weights in citations. In this project, we value more on the number of citations to and from different journals than times of citation to and from the same journal.

- "PLOS ONE" and "FRONT PSYCHOL" have higher weighted out-degree centrality than "COMPUT HUM BEHAV", their unweighted out-degree centralities are lower. This can be explained in a similar way to the previous point.

- "AUGMENT ALTERN COMM" and "INT J DES" has higher closeness centrality than "COMPUT HUM BEHAV", however, the difference is not significant. Besides, they are not as important as "COMPUT HUM BEHAV" with respect to other measurements.

To sum up, considering these reasons, our group chose "COMPUT HUM BEHAV" as the most important journal. The figure below is a visualization of the citation network with the highlight of "COMPUT HUM BEHAV". This illustrates the importance of "COMPUT HUM BEHAV" as shown in the below network diagram.

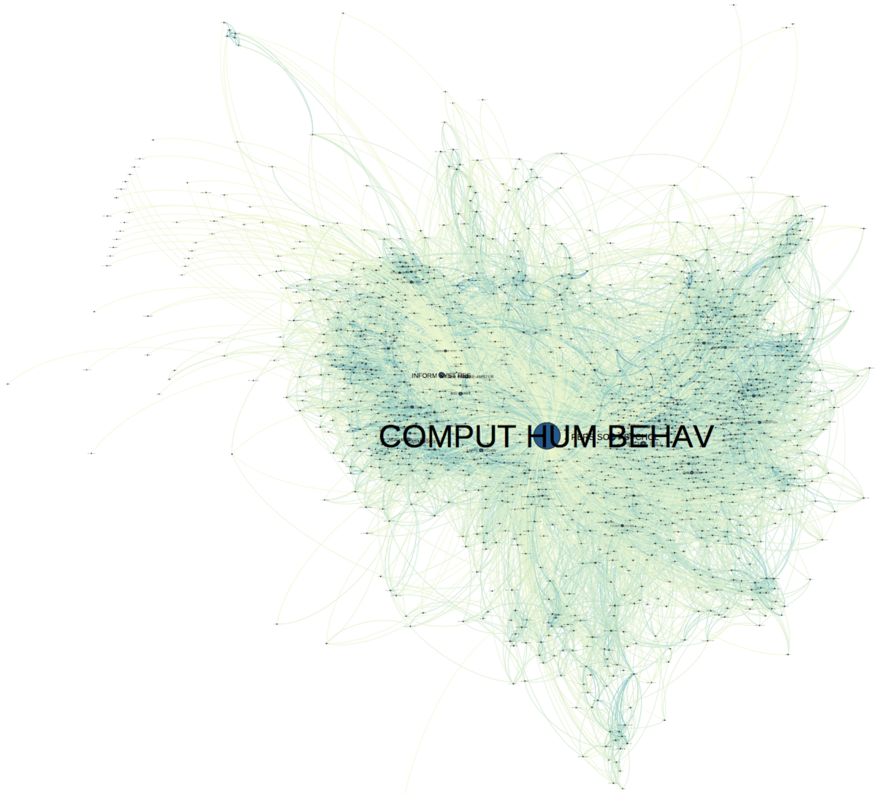

Figure 7.8 Network Diagram Overview

8 Limitation

Although the research and analysis aimed to be more systematic and comprehensive, it is undeniable that there exists a few limitations in the proposed network model. Overall speaking, the entire model would be a simplification of the actual journal citation network. For instance, as the network was plotted on the basis of five selected journals, possibly some important journals or connections would be missing on the network. Also, although web crawler was applied to find the connections among neighbouring journals, it would be difficult to guarantee that the crawler has digged all the available databases thoroughly. With more time and resources available, it would be ideal to construct a more all-rounded network to support more in-depth and accurate analysis.

In addition, in clustering analysis, the optimal clusters were generated via trial-and-error, i.e. identifying the most “appropriate” by trying several possible inputs (“resolution”). However, this method might not be applicable when the network size grows larger, since it would be extremely difficult to try out a bunch of inputs and find the optimal one. It would be worthwhile implementing some algorithms to find the optimal cluster size, so that solving the clustering problem could be more accurate and efficient.

9 Conclusion

As discussed and analyzed above, the project aims to plot and analyze the journal citation network on the basis of five well-selected journals in the information systems field. Through using a few analytic tools and applications, the network diagram could be plotted in a clear and succinct way, and various network metrics such as degree centralities and betweenness centralities could be calculated for further analysis. Regression and clustering have also been applied to analyze the journal network in a more detailed manner. To obtain more in-depth and all-rounded findings, it would be ideal to expand the research scope and represent a more general journal network.

References:

Du, D. (n.d.). Social Network Analysis: Centrality Measures. Retrieved from: http://www2.unb.ca/~ddu/6634/Lecture_notes/Lecture_4_centrality_measure.pdf

Garfield, E. (1999). Journal impact factor: A breif review. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1230709/pdf/cmaj_161_8_979.pdf

Leydesdorff, L. (n.d.). “Betweenness Centrality” as an Indicator of the “Interdisciplinarity” of Scientific Journals. Retrieved from: https://arxiv.org/pdf/0911.2719.pdf

Jengolbeck. (2013, July 22). Gephi Modularity Tutorial. Retrieved from https://www.youtube.com/watch?v=7LMnpM0p4cM

Quora. (n.d.). What are the limitations of graph centrality measures? Retrieved from: https://www.quora.com/What-are-the-limitations-of-graph-centrality-measures

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre. (2008). Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000. Retrieved from: https://arxiv.org/pdf/0803.0476.pdf

R. Lambiotte, J.-C. Delvenne, M. Barahona. (2009) Laplacian Dynamics and Multiscale Modular Structure in Networks. Retrieved from: https://arxiv.org/pdf/0812.1770.pdf

Appendix

Appendix files are complementary materials to support our analysis and arguments in the sections above, including:

- “Network Results (In-degree & Out-degree Centralities).xls”: Full list of all the journal connections and the corresponding in-degree & out-degree centralities
- “crawler.py”: Progrmaming file in Python used to do web crawler to get the journal connections among neighbouring journals
- “adjMatrix.csv”: Matrix file of all the journal connections
- “projectScript.R”: Programming file in R used to run the different network measurement data
- “journal-citation-visulisation.pdf”, “journal-citation-communities.pdf”: Network diagrams for the full network and the network with clustering analysis