Advanced Lab 3: Community Structure of the SAP Online Knowledge Community Platform

Part 1). Community Structure: Visualization and analysis of the SAP online knowledge community.

Overview of the dataset:

The given data set consists of two .csv files with infile_edges.csv indicating the edge connections and infile_nodes.csv has the nodes information and their corresponding attribute characteristics. The community network is composed of 98,288 vertices and 57,102 edges. The nodes have total of 9 attributes in total out of which country is a categorical variable and the rest are continuous variables.

Analysis:

To understand the underlying community structure, walk- trap community detection algorithm was run on the network which resulted in about 96000 communities. Also observing the vast difference between the number of vertices and edges suggests the presence of isolated and disconnected components, which explains the unusual number of communities in the above case. So, to understanding the network, a more densely closely connected network is needed, which can be extracted by the below 2 methods:

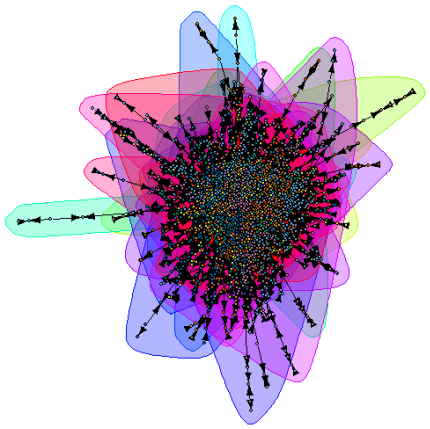
1. Removing the isolated, triples, paired components and then combining the decomposed network back to "Recomposed component"
2. Extracting giant component

Various community detection algorithms are applied to the above networks. The results are as below,

Type of component	Community detection algorithm	Number of communities observed
Recomposed	Louvain Cluster	235
Recomposed	Fast greedy	399
Recomposed	Walk-trap	3,234
Recomposed	Springless	Doesn't run on unconnected graph
Recomposed	Label Propagation	1,750
Giant	**Louvain Cluster**	**92**
Giant	Fast greedy	253
Giant	Walk-trap	3,064
Giant	Springless	Takes too long to complete
Giant	Label Propagation	1,569

By looking at the above numbers, it can be observed that Louvain Cluster algorithm when applied to the giant component results in least number of communities, which is 92. Hence this combination is taken as a base for the further analysis. The network is simplified, multiple edges removed and converted to weighted edges instead after initialising the edge weights to 1.

The giant component was plotted using layout.fruchterman.reingold layout:



Evaluating Assortativity & Clustering coefficients of SAP online knowledge platform

For the complete network:

1. **Degree Assortativity:** The degree Assortativity of SAP online knowledge platform is negative, having a value of -0.01096796. This means that two connected nodes do not have same degree. To put it in other words, there are some people with vast number of followers while others have comparatively lesser number of followers. Assortativity is also applied to ln_points and the value is found to be positive (0.02234436) in the ove rall network. This means that a there is some level of homophily in the overall networ k with respect to intelligence points gathered by individuals in the SAP network.

2. **Clustering Coefficient**: A vast difference in overall (0.005203974) and average local clustering coefficient (0.03404859) represents that there are certain localised commun ities/components in the network which are cut-off from the rest of the network. This o bservation seems reasonable as the total network showed about 96000 communities w ith just 57,102 edges.

Assortivity and Clustering coefficients of Giant community:

1. **Degree Assortativity**: The value of degree assortativity for the giant component is -0 .01697888. Negative value represents a star-like network structure wherein the degree of a node is negatively correlated with the degree of other nodes connected to it. negat ive values mean that large-degree nodes tend to attach to low degree nodes. Hence It c an be inferred that there are certain people in the network with multiple followers who themselves have lower degree. It can also be inferred that a hub and spoke structure is present. When assortativity is applied to ln_points , the value obtained is -0.03240836, which implies the presence of negative homophily. This means that a there is a negati ve correlation in the level of intelligence among the nodes which are connected. A si milar pattern is followed in the communities within the network with negative assortat ivity of the In Points. However, out of 92 communities observed, there are 17 commu

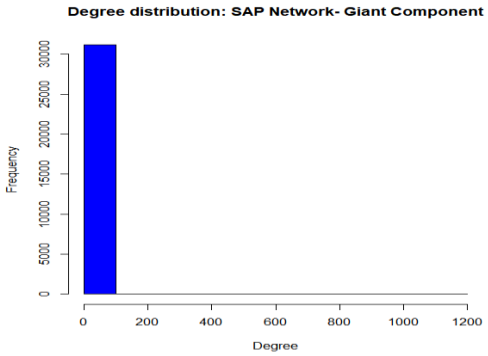
nities which follow a different pattern, that is people of similar level of intelligence po ints gained are connected in the network.

2. **Clustering Coefficient**: Similar to the entire network, the giant component also has a vast difference in overall (0.005204738) and average local clustering coefficient (0.03 52601). It indicates that there might be modularity present in the network which might be a reason of information now flowing within the network. In the communities prese nt within the giant component, a different trend is observed with most of the communi ties having almost no difference between the average and overall clustering coefficien ts. This is indicative of a well-connected nature in the sub communities.

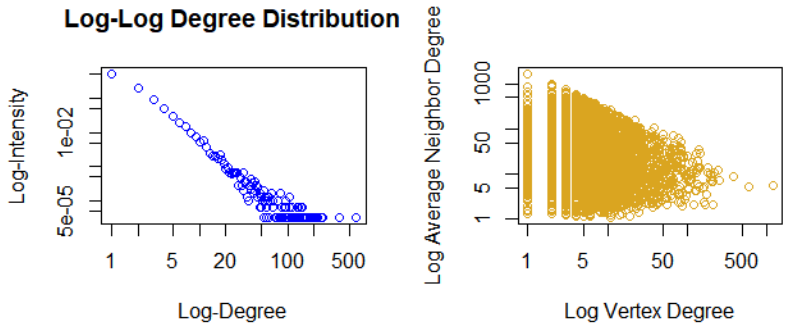
3. **Modularity:** modularity coefficient of the community turns out to be 0.7858535. This confirms with the above conclusion made based on the clustering coefficient.

Interpretation of the giant component:

1. **Degree Distribution:** The degree distribution of the giant suggests "preferential attachment" as the distribution doesn't have a normal shape. This is a case of a scale free network where people with higher number of links attract more people.



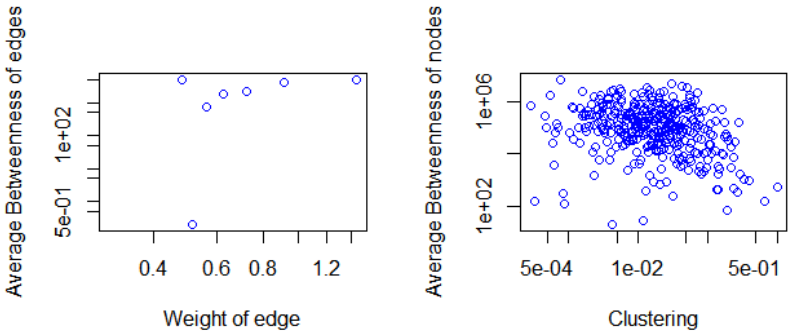
2. a) **Log-Log plot of degree distribution and log of the frequency** which shows a power law, a presence of popular individuals in the network. Noise can be noticed in the tail of the distribution.
   b) **Plot of average neighbour degree versus vertex degree** in the SAP Network, suggests that while there is a tendency for vertices of higher degrees to link with similar vertices, vertices of lower degree tend to link with vertices of both lower and higher degrees



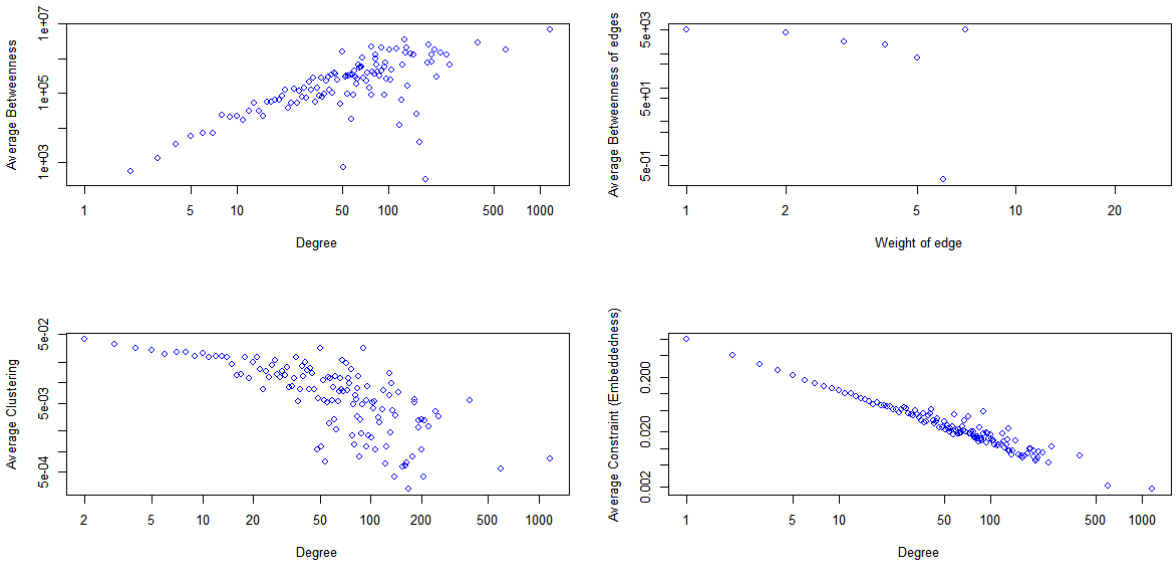
3. a) **Negative relationship between average betweenness of nodes and clustering**: Strengthens the finding that there are certain popular individuals (having high

betweenness centrality) to which answer many questions on the SAP community forum and serve a link between other individuals who are not connected. This is similar to a hub and spoke structure

b) **Negative relationship between inverse weight and average betweenness of edges**: Indicates preferential attachment or scale free network described above



4. The charts shown below are summarized below:
   a) **Positive relationship between degree and betweenness centrality:** Reflects that individuals who have higher links or the people who answer many questions, serve as a link between many other people who don't answer many questions. This shows the presence of a "Star-Like" relationship
   b) **Negative relationship between edge weight and average betweenness of edges**: Indicates preferential attachment or scale free network described above
   c) **Negative relationship between degree and local clustering & degree and average constraints**: Shows presence of individuals in the network which are not connected. This shows a presence of structural holes (popular individuals) surrounded by constraints (not connected individuals)



Part 2) <u>Analysis of effect of nationality and other attributes on community structure:</u>

This part of the report deals with the analysis of effect of nationality and other attributes on the structure of the community. We want to know how confidence we are that the attributes play a role in the formation of communities. We use statistical methods,

especially Analysis of Variance (ANOVA) for this. Our null hypothesis would be that the attributes have no effect on the community structure and the distribution is uniform across the communities. Alternative hypothesis is that attributes in fact have an effect and effect the distribution. We find the F statistic value for this hypothesis which in turn helps in finding the P-value. If P- value is significantly low, we reject null hypothesis and conclude that attributes influence the community structure. Below are the F statistic and P values for all the attributes.

Attributes	F value	P Value
Country	15.11	<2e-16
ln_points	7.311	1.04e-05
ln_fdi	16.36	1.6e-12
Fdi_pergdp	17.17	5.25e-05
ICT_goods_import_percent	11.72	4.61e-09
Internet_users_percent	9.376	2.75e-07
Immigration_pct	15.2	1.16e-11

Looking at the above table, the P values of all the attributes are significantly low, which allows us to reject the null hypothesis. So, it can be concluded that all the attributes have effect on the community structure.

Part 3) <u>Analysis of correlation between network attributes and knowledge contributions</u>:

At community level, aggregated ln_points represent the productivity of each community in producing new knowledge in the system. So, to analyse the effect of network attributes on new knowledge production, a linear regression is formed with ln_points as the dependent variable and network attributes as the independent ones. Network attributes considered are overall clustering, average clustering, average degree, average betweenness, average hub score and average authority score. The result of the regression is as below,

```
lm(formula = ln_points ~ overall_clustering + avg_clustering +
    avg_degree + avg_betweens + avg_hub_scr + avg_auth_scr, data = a
ds)


Residuals:
    Min      1Q  Median      3Q     Max
-2815.2  -415.3   -83.2   283.2  3724.3

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2701.904   1345.608  -2.008  0.04786 *
overall_clustering -49610.226  23724.467  -2.091  0.03954 *
avg_clustering        404.998   6305.602   0.064  0.94894
avg_degree           1988.289    645.141   3.082  0.00278 **
avg_betweens           11.807      1.412   8.362 1.14e-12 ***
avg_hub_scr          -476.666    707.595  -0.674  0.50239
```

```
avg_auth_scr        -1261.398    678.005  -1.860  0.06632 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 879 on 84 degrees of freedom
Multiple R-squared:  0.8578,       Adjusted R-squared:  0.8477
F-statistic: 84.47 on 6 and 84 DF,  p-value: < 2.2e-16
```

The above results can be interpreted by checking two factors from the above results. P value of average clustering, average hub scores are high suggesting that they are not statistically significant. So nothing can be said about their influence on knowledge productivity.

Overall clustering has a P value or 0.03, which makes it significant for standard alpha = 0.05%. So, it does have effect on ln_points. The coefficient (-49610.226) suggests that a negative relation exists between these two factors, which means higher overall clustering results in lower productivity of each community in producing new knowledge in the system.

Similiarly, average degree and average betweenness have P values 0.00278 and 1.14e-12 making both statistically significant. Both have positive coefficients indicating positive relation with ln_points . This implies higher degree centrality and higher betweenness would drive for higher productivity of knowledge in communities.

As hypothesized we now see a positive relation between the knowledge contained within a community and the degree and betweenness centralities within the network.