General Information and Overall Structure of the Network

The SAP online knowledge community network is a directed graph forming a discussion thread structure, where a directed edge represents an answer provided by a user to a question posted by another user. The network consists of 3415 nodes and 6090 edges. However, since a user can answer multiple questions posted by another user on one or more threads, there exists duplicate edge-pairs. Thus, the network is simplified to produce directed edges of varying weights. Now there are about 4120 edges in the simplified graph.

We can get some useful information just by looking at the direction of the edges from a node. An outgoing edge from a node representing a user answering to another user's question usually implies that the replier has a superior proficiency on the topic than the asker. Thus, the outdegree weights of the network would represent how often a user helps another. On the other hand, nodes receiving a lot of incoming edges, i.e., users receiving many answers from other users can represent users that are prestigious in the online community. Thus, such a network structure can be used to rank people's expertise in SAP online community. In order to analyze the network, several network metrics and network cohesion measures are used which are discussed below.

Degree Distribution

User with ID 592540 has the maximum overall degree in the network having a total of 454 connections in the SAP community. This user also happens to have maximum outdegree in the network with 452 outgoing connections to other nodes. This implies user with ID 592540 provides answers to 452 questions in the SAP community, but it receives answers from only two other users. Further, user with ID 3510478 has maximum indegree with 25 incoming connections, which means this user receives maximum answers to his/her questions in the SAP community. This user has outdegree of 6 which means it provides answers to 6 other questions in the community.

The users' relative connectedness in this network can be described by degree distribution. The vertex strength distributions, i.e. weighted degree distributions, for in-paths and out-paths are shown in Fig. 1. We see that there is a substantial fraction of nodes of quite low degree in both the histograms. In particular, there is more frequency of users which are not answering questions posted by other users compared to users who are not posting any questions. In other words, there are more users asking questions in the forum than there are users who answer questions posted by others. Further, we see that both the histograms generally decay as the vertex strength increases.

Given the nature of the decay in the distributions of Fig. 1, a log–log scale could be more effective in summarizing the degree information as shown in Fig. 2. We see that there is a fairly linear decay in the log-frequency (cumulative probability) as a function of log-degree in both indegree and outdegree distributions. This again reflects the highly uneven distribution of participation. Instead of everybody helping each other equally, in the SAP community, there are some extremely active users who answer a lot of questions while a majority of users answer only a few. Likewise, many users ask only a single question, but some ask plenty of questions.

Next, to analyze how vertices of different degrees are linked with each other, a plot of average neighbor degree versus vertex degree in the network is shown in Fig. 3. The plot suggests that there is a tendency for vertices of lower degree to link with vertices of both lower and higher degrees, while vertices of higher degree tend to link with vertices of relatively lower degree.

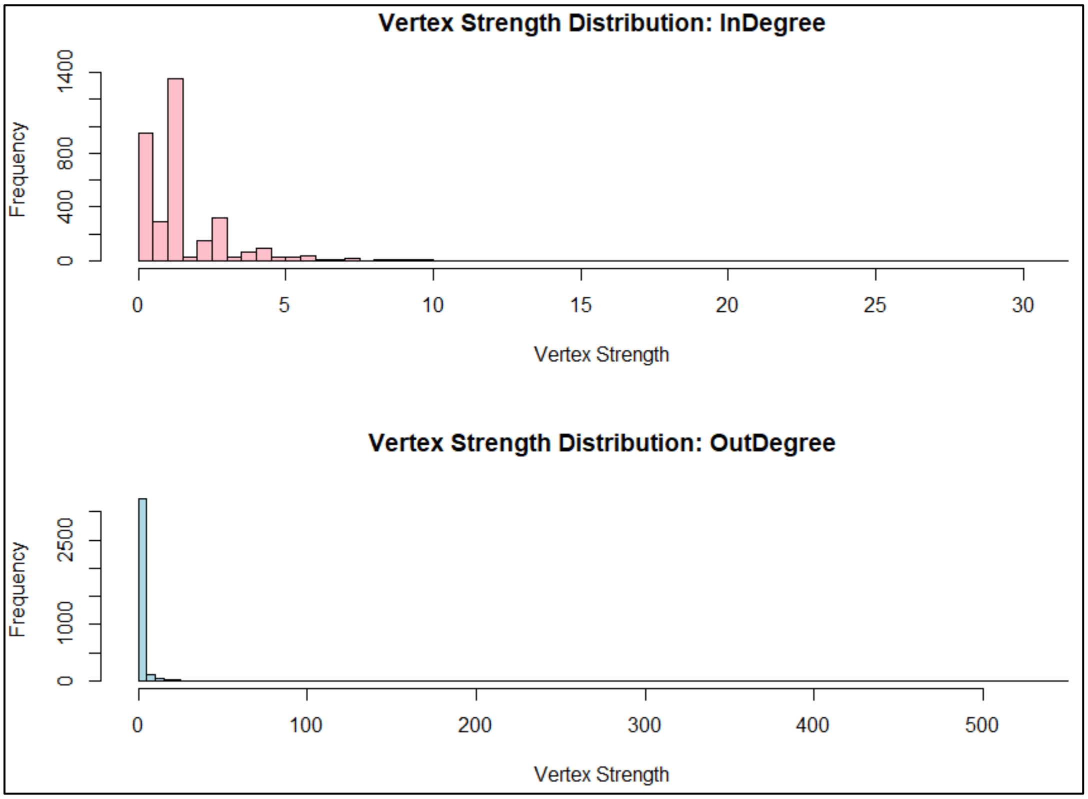


Fig. 1. Distribution of vertex strength: InDegree (top), OutDegree (bottom)

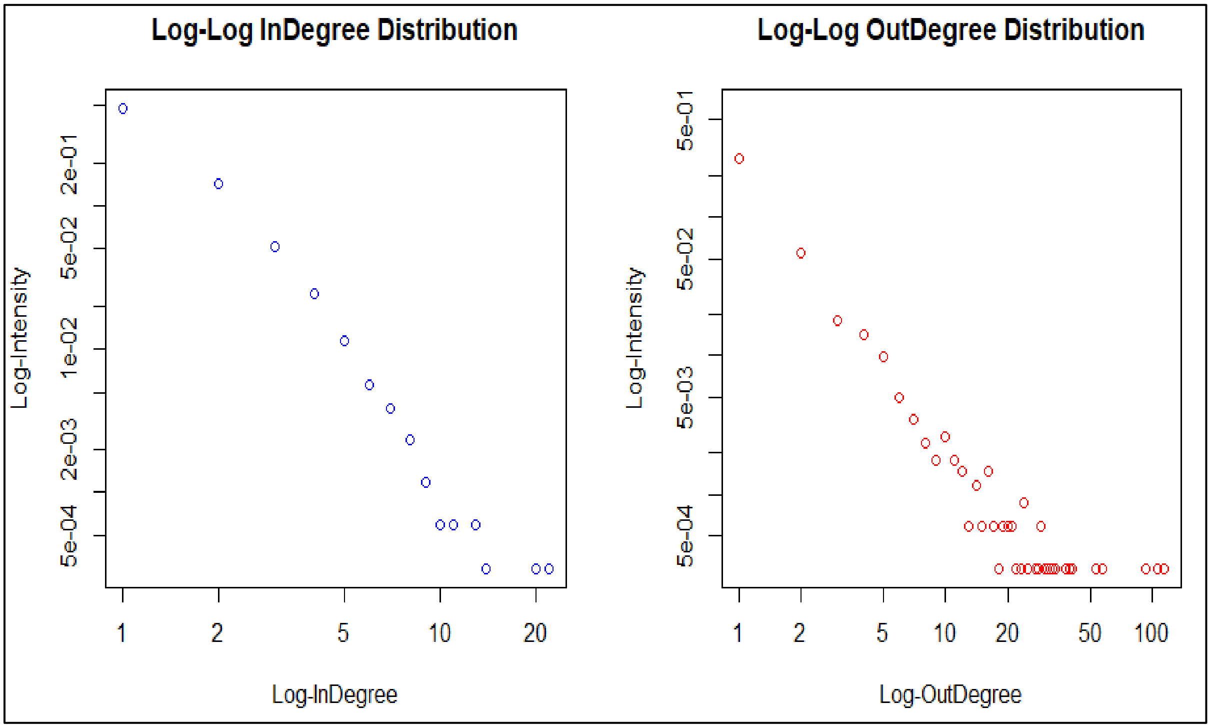


Fig. 2. Distribution of degree in log-log scale: InDegree (left), OutDegree (right)

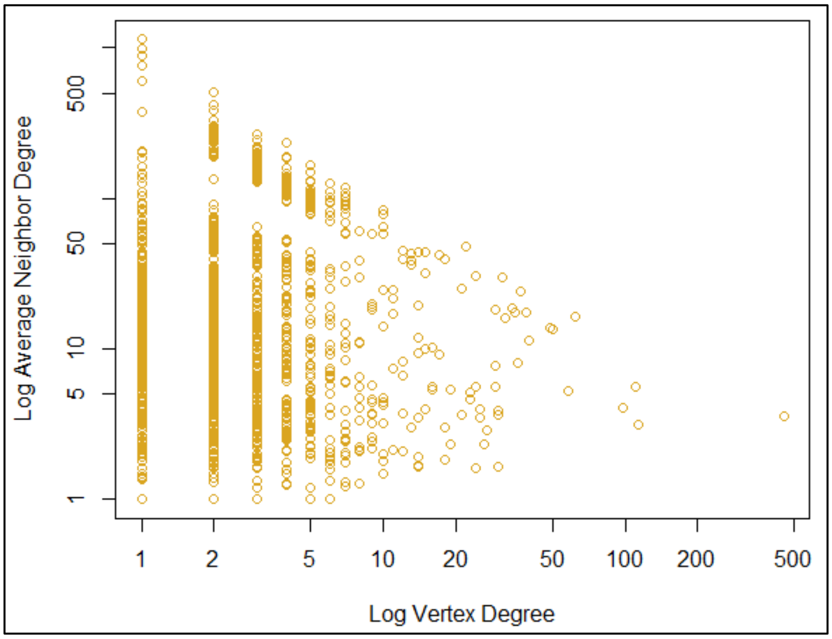
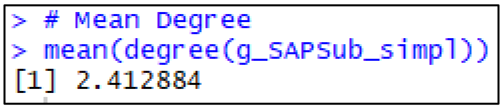


Fig. 3. Average neighbor degree versus vertex degree (log–log scale)

Average Degree

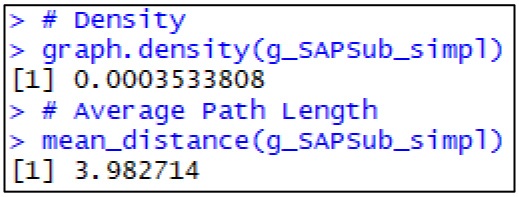
The degree refers to the amount of ties a node (user) has in the network and could help to reveal the most powerful individuals in a network. On average a user has ties to around 3 individuals.



```
> # Mean Degree
> mean(degree(g_SAPSub_simpl))
[1] 2.412884
```

Network Density and Average Path Length

Network density represents the proportion of connected ties over the total possible connections in the network. This gives an idea about how connected the network is. We see that the density value is very low indicating that the SAP online community network doesn't maintain good ties and thus is not well connected. The average path length refers to average number of steps in the shortest path to navigate the network. It gives us a sense of how efficient the flow of information is through the SAP network. We see that generally it takes about 4 ties on average to get all the way around in the network. Average path length can be correlated with density. A larger average path distance often implies that we have a less dense network.



```
> # Density
> graph.density(g_SAPSub_simpl)
[1] 0.0003533808
> # Average Path Length
> mean_distance(g_SAPSub_simpl)
[1] 3.982714
```

Connectivity

From the analysis thus far, we have seen that some people reply almost exclusively, and others ask almost exclusively, which suggests that there might be several components in the network isolated from each other. Thus, the SAP online community network may not contain giant strongly connected components (SCCs). SCCs represent those sets of users, such that one user can be reached from any other, following directed edges from asker to replier. A large SCC indicates the presence of a community where many users interact, directly or indirectly. Table I gives the sizes of the SCCs. We see that the network decomposes into a total of 3397 SCCs, with the largest cluster only having 9 users in it. Most of the SCCs contain singleton vertices.

A network graph is weakly connected if when considering it as an undirected graph it is connected. It happens that this network is also not weakly connected. Table II gives the sizes of the weakly connected components (WCCs). The largest WCC consists of 2696 users.

Table I. Strongly connected components (SCCs)

# SCC Clusters	3387	8	1	1
# Nodes	1	2	3	9

Table II. Weakly connected components (WCCs)

# WCC Clusters	239	43	11	3	5	2	1	1
# Nodes	2	3	4	5	6	7	9	2696

Reciprocity

Reciprocity refers to the ratio of reciprocal relations in a network. An example to describe a reciprocal relation would be a situation where user A answers user B, and user B answers back to user A. The reciprocity of the given SAP network is only 0.005825243, i.e., only about 0.6% reciprocal relations exists between users out of the total number of edge relations in the network.

There are only 12 mutual edges (relations where users are answering each other) in this network. The vertexes (users) on these mutual edges are listed in the Table III. The occurrence represents the number of times a vertex is a part of a mutual edge.

Table III. Table of vertexes in mutual edges

User ID	Occurrence	User ID	Occurrence
1119	1	3526200	1
1195854	2	3554509	1
131143	2	3609218	1
131313	1	3616618	1
2269373	1	3621372	1
2470693	1	3685584	1
3081842	1	3982689	1
3454297	1	4095126	1
3461918	1	4183082	1
3477685	1	44034	1
3483752	1	653026	1

Clustering and Transitivity

Clustering gives an idea of potential underlying mechanisms in the network. A network with high clustering has a higher proportion of closed triads to all triads. In other words, when there is mutuality there will be high transitivity. In the given SAP online community network, since the density of ties is lower overall, we would expect transitivity to be low.

The global transitivity or the global clustering coefficient of the network is 0.009985725 i.e. about 1.0%, which means that not more than 1.0% of the connected triples are close to form a triangle. This means that it barely happens that if two users answer the same questions asked by another user, then these two users might tend to answer the questions asked by the each other. Thus, there does not exist much triadic closure in the SAP network and it is sparse.

Diameter of Graph

The diameter of a graph is the length of the longest geodesic. Using the regular weights for edges, the diameter of the SAP network graph is 26.0. When the inverse logarithm weight for edges are used, the diameter of graph reduces to 14.27228. With the inverse logarithm weights, the length of the longest geodesic is shorter.

Clique Census

In addition to clusters and communities that we saw before, one another approach to define network cohesion is through clique census. Cliques are complete subgraphs and hence are subsets of vertices that are fully cohesive, in the sense that all vertices within the subset are connected by edges.

Table IV summarizes maximal clique census which represents cliques that are not a subset of a larger clique. We see that in the SAP network there are 3320 edges (cliques of size 2), followed by 335 triangles (cliques of size three), followed by 39 cliques of size 4, and 5 cliques of size 5. The largest clique of size 5, of which there are 5, includes User with ID 592540 in common. Here, we do not see larger cliques because the SAP network is very sparse.

Table IV. Maximal Clique Census

Clique Size	2	3	4	5
# Cliques	3320	335	39	5

Hub and Authority Scores

A vertex is considered a hub if it points to many other nodes and a vertex is considered an authority if it has many nodes linking to it. Table V and VI show few users with high hub and authority scores, respectively. Previously we had seen that user with ID 3510478 had maximum indegree, however, this user doesn’t seem to have a high authority score.

Table V. Users having high hub scores

User ID	592540	131143	22328	701187	3552437	1666938
Hub Score	1.0000	0.0954	0.0705	0.0490	0.0454	0.04121

Table VI. Users having high authority scores

User ID	3462364	3497637	3469006	182703	1636378	3720625
Authority Score	1.0000	0.9953	0.9727	0.9688	0.9533	0.9464

Positional Features and Network Centralities

Various network centralities are evaluated, to be used as local measures, for the SAP online community network. Some of these measures include degree, node betweenness, edge betweenness, closeness centralities, etc. These centralities are compared through correlation plots and statistically through correlation matrix.

From the graph on left in Fig. 4, we observe there is a positive correlation to some extent between average betweenness of edges and edge weights. On the right, we see a significant negative correlation between average betweenness of nodes and local clustering coefficients. Thus, when individuals' clustering in the SAP network is very low, the betweenness of those individuals is quite high, and high betweenness leads to formation of structural holes and local bridges. This leads to an important conclusion about the SAP network structure- *networks low in cohesion and structural equivalence are rich in structural holes. The strength of weak ties.*

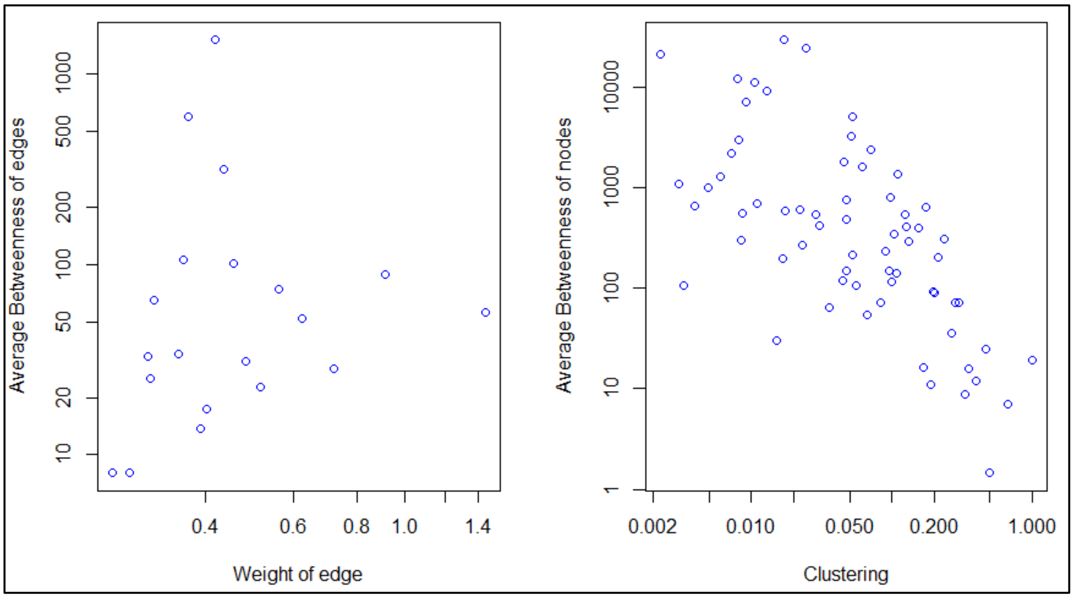


Fig. 4. Correlation plots of various centralities: Set 1

Local bridges play a very important part in an online knowledge community like the given SAP network. Local bridges or structural holes lead to connections among users who have no other common connection. This in fact leads to better flow of knowledge and there is less redundancy compared to that of a strong network. Through a local bridge, critical information can be shared which may not be quite possible in a strong network. Local bridges become important part of knowledge sharing, and information can be useful for the whole community as a whole. Because of high betweenness, there is low overlapping of information. The local bridges help the SAP community to gain important information. Strong cohesive ties sometimes lead to data holding and redundancy and so structural holes play a very important role.

Further, more correlation plots are shown in Fig. 5. In bottom-left plot there is a negative correlation between individuals' average clustering coefficient and degree which implies that for the given SAP network the global clustering is lower than the average clustering. In top-left plot, we see that average betweenness of nodes increases as their degree increases. Further, the embeddedness decreases with increase in degree in bottom-right plot.

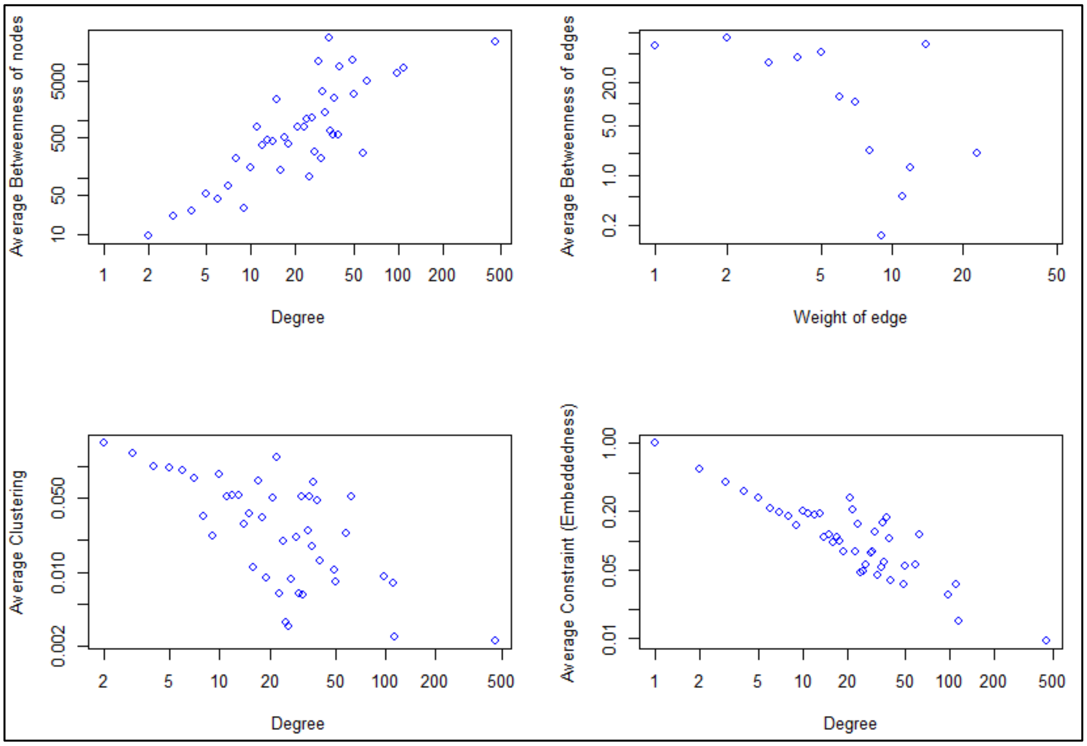


Fig. 5. Correlation plots of various network centralities: Set 2

Table VII shows a correlation matrix with correlation values for each pair of centralities. We can see, for example, that degree centrality is positively correlated with hub score at a 0.89 level, while it has almost negligible correlation with the authority score.

Table VII. Correlation matrix for various network centralities

	Degree	**Node Betweenness**	**Edge Betweenness**	**Closeness**	**Eigen**	**Hub Score**	**Authority Score**
Degree	1	0.582	-0.0024	0.2693	0.7569	0.8984	0.0358
Node Betweenness	0.582	1	0.0089	0.3929	0.4384	0.4893	0.0214
Edge Betweenness	-0.0024	0.0089	1	0.0184	- 0.0057	- 0.0019	-0.0055
Closeness	0.2693	0.3929	0.0184	1	0.1732	0.1928	-0.0179
Eigen	0.7569	0.4384	-0.0057	0.1732	1	0.7484	0.2351
Hub Score	0.8984	0.4893	-0.0019	0.1928	0.7484	1	0.006
Authority Score	0.0358	0.0214	-0.0055	-0.0179	0.2351	0.006	1

Network Visualization

The SAP online community network graph shown in Fig. 6 is obtained by applying Fruchterman and Reingold layout algorithm, which is one of the most used force-directed layout algorithms. The node sizes are proportional to their vertex strength i.e. weighted vertex degree. The nodes are colored based on their indegree as shown in Fig. 6. Additionally, the width of the edges is proportional to the logarithm of edge weights. If a node appears to be large and has a lighter shade of blue, that would indicate a user with higher outdegree and lower indegree. Similarly, if a node appears to be small but has a darker shade of blue, that would indicate a user with higher indegree and lower outdegree.

Another popular force-directed algorithm that produces nice results for connected graphs is Kamada Kawai which is shown in Fig. 7. Like Fruchterman Reingold, it attempts to minimize the energy in a spring system.

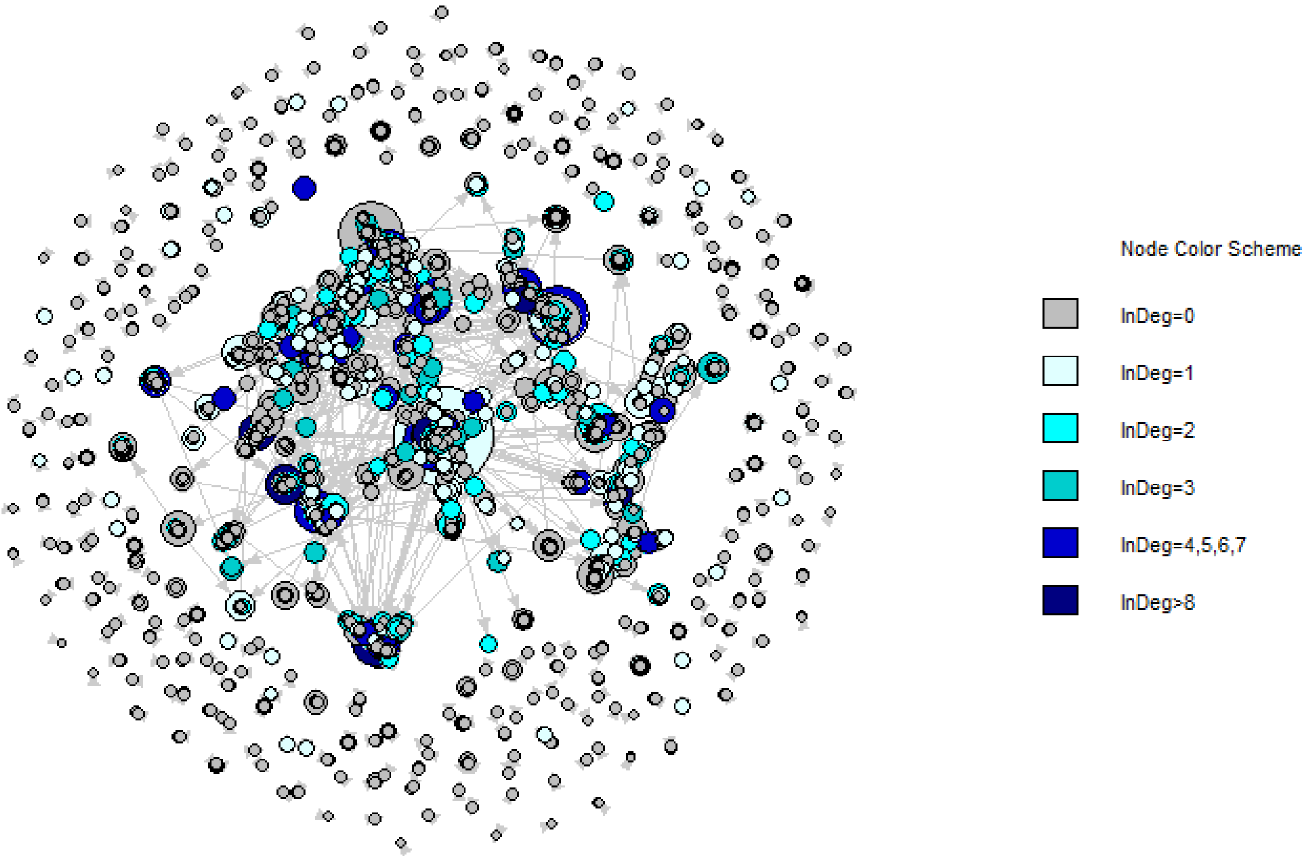


Fig. 6. SAP network layout using the method of Fruchterman and Reingold

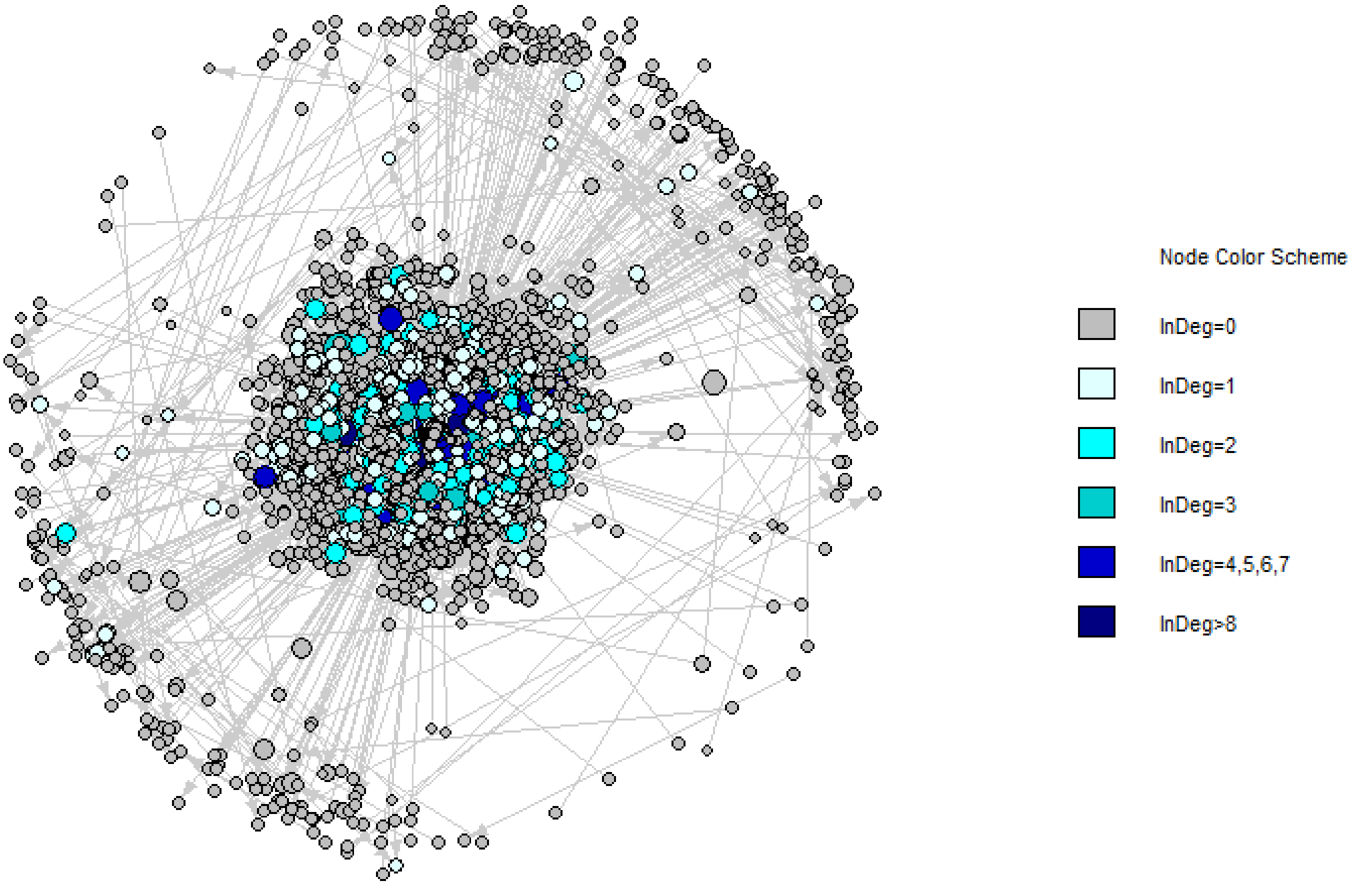


Fig. 7. SAP network layout using the method of Kamada and Kawai

References

Kolaczyk, E. D., & Csárdi, G. (2014). *Statistical analysis of network data with R* (Vol. 65). New York: Springer.

Aral, S., & Van Alstyne, M. (2011). The diversity-bandwidth trade-off. *American Journal of Sociology*, *117*(1), 90-171.

Zhang, J., Ackerman, M. S., & Adamic, L. (2007, May). Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web* (pp. 221-230). ACM.

Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008, April). Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web* (pp. 665-674). ACM.