Community Structure of the SAP Online Knowledge Community Platform

ANALYSIS AND COMPARISON OF THE COMMUNITY DETECTION ALGORITHMS ON THE SAP NETWORK:

Running community detection algorithms on the entire network structure yields the following results:

Algorithm	**Time taken**	**No. of communities**	**Size of largest Community**
Fast Greedy Algorithm	13.32	63994	5125
Walktrap Algorithm	64.13	67128	6618
Label Propagation Algorithm	1.78	65648	5171
Eigenvector Algorithm	7.11	63821	19406

There were many nodes in this network with a zero degree. ***A total of 61237 nodes had zero degree***. Eliminating all such nodes, the new results by running the community detection algorithms were as follows:

Algorithm	**Time taken**	**No. of communities**	**Size of largest Community**
Fast Greedy Algorithm	11.42	2757	5125
Walktrap Algorithm	43.72	5891	6618
Label Propagation Algorithm	0.93	4601	4957
Eigenvector Algorithm	5.73	2584	19406

The above results imply that the only the communities that had a size of one have been eliminated from the first table to the second.

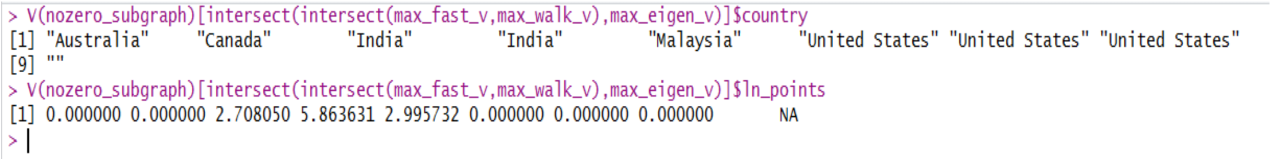
EACH ALGORITHM HAS DIFFERENT NODES IN ITS LARGEST COMMUNITY

An interesting observation is that each algorithm forms communities with nodes not present in the largest community formed by the other algorithms.

By following the below steps to check the common nodes in the largest community generated by each algorithm:

1) Consider the largest community in each algorithm
2) Check which nodes are common in all the 4 largest communities using intersect function in R.

To summarize, taking the largest communities from all four algorithms together, there are no common nodes. However, when the label propagation algorithm was eliminated, there were 9 common nodes. Further checking the attributes of these nodes, it could be seen that they were not the ones having highest points or from a specific country. From the R result below we can see that the common nodes in the largest community generated by 3 of the algorithms seem to be extremely random.



```
> V(nozero_subgraph)[intersect(intersect(max_fast_v,max_walk_v),max_eigen_v)]$country
[1] "Australia"     "Canada"        "India"         "India"         "Malaysia"      "United States" "United States" "United States"
[9] ""
> V(nozero_subgraph)[intersect(intersect(max_fast_v,max_walk_v),max_eigen_v)]$ln_points
[1] 0.000000 0.000000 2.708050 5.863631 2.995732 0.000000 0.000000 0.000000       NA
> |
```

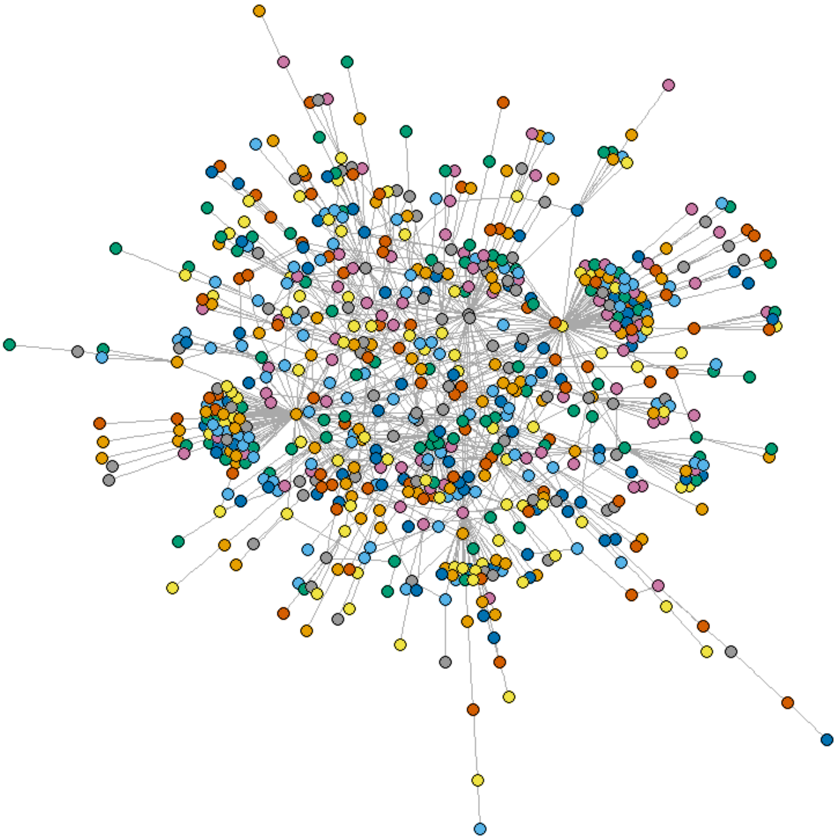
SIGNIFICANCE OF COMMUNITIES:

On running the Wilcox test for significance on the communities obtained by all four algorithms, the conclusion was that fast greedy algorithm and walktrap algorithm returned the maximum number of significant communities.

ROLE OF COUNTRIES IN FORMING COMMUNITIES:

If we colour the nodes of the largest communities generated by the fast greedy algorithm as per the country, we find that ***countries have absolutely no part to play in formation of communities***. The procedure followed was as follows:

1) Obtain the largest community from the fast greedy algorithm. It has 5125 nodes.
2) Run the fast greedy algorithm again on this community and get the largest community again. It has 660 nodes. The visualization obtained is as below. Nodes are coloured as per countries.



3) The next step was to check if this trend holds across communities by plotting the various communities with varying sizes and coloring the nodes as per countries:

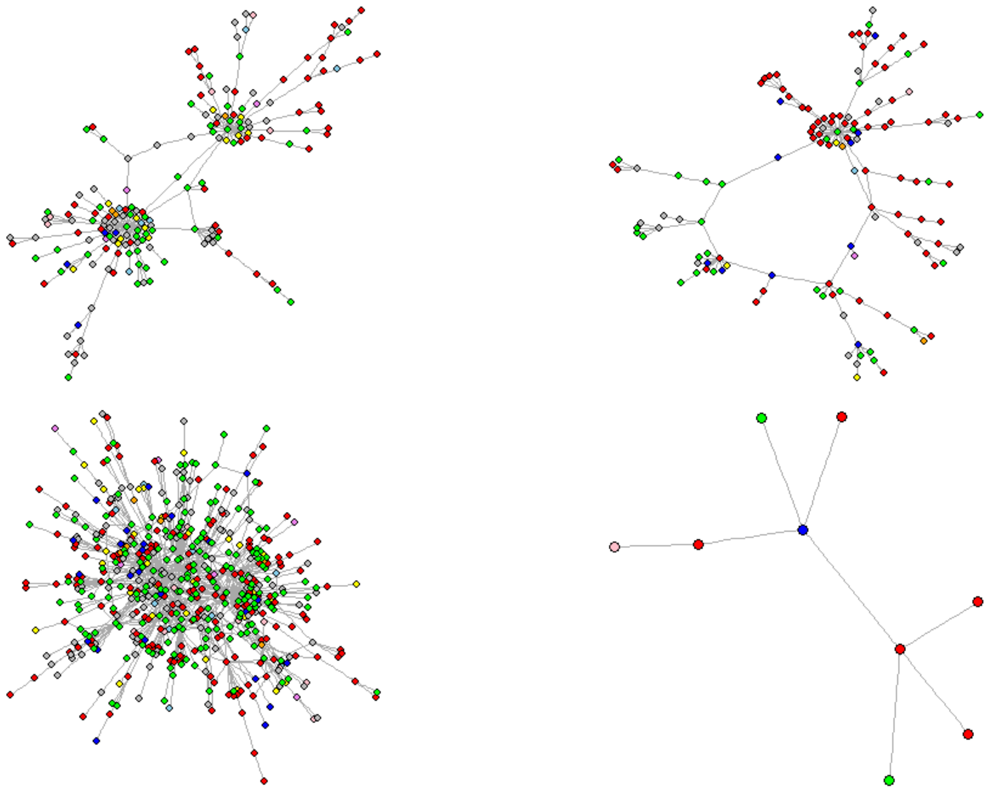
Colors indicate the top 8 countries to which the vertices belong. The following is the color code:

India : red ***United States : green*** ***Germany : yellow*** ***United Kingdom : blue***

Canada : pink ***Australia : skyblue*** ***Spain : orange*** ***China : violet***

All others: Gray

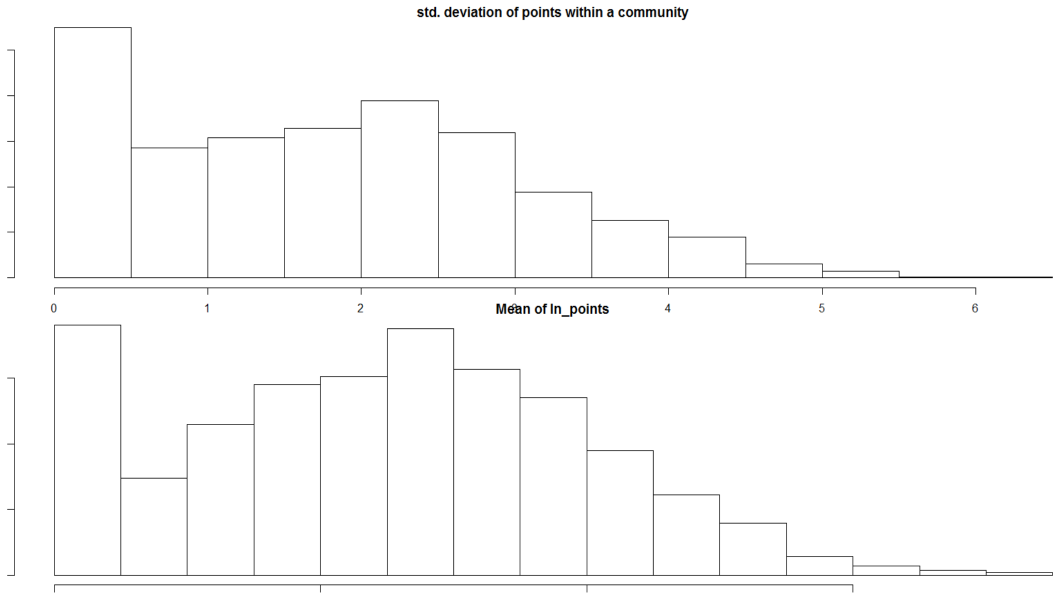
As is evident from the plot below of four different communities of varying sizes, the trend of countries not contributing to community membership hold in large communities as well as small communities. Even “bridges” in the network are most often not formed between nodes belonging to the same countries.



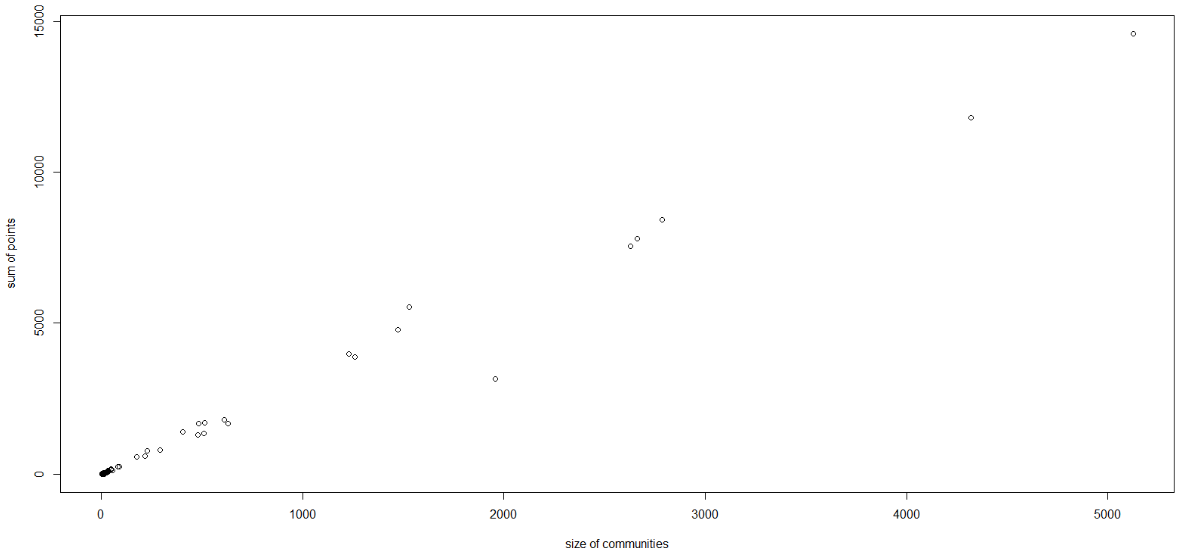
From the above visualizations, it is clear that a community is not comprised of any “dominant country”

The other attributes fdi by gdp, ICT imports, percentage of internet users, latitude and longitude are specific to a country and constant for a country. Hence these attributes will not matter in determination of communities if the countries do not matter.

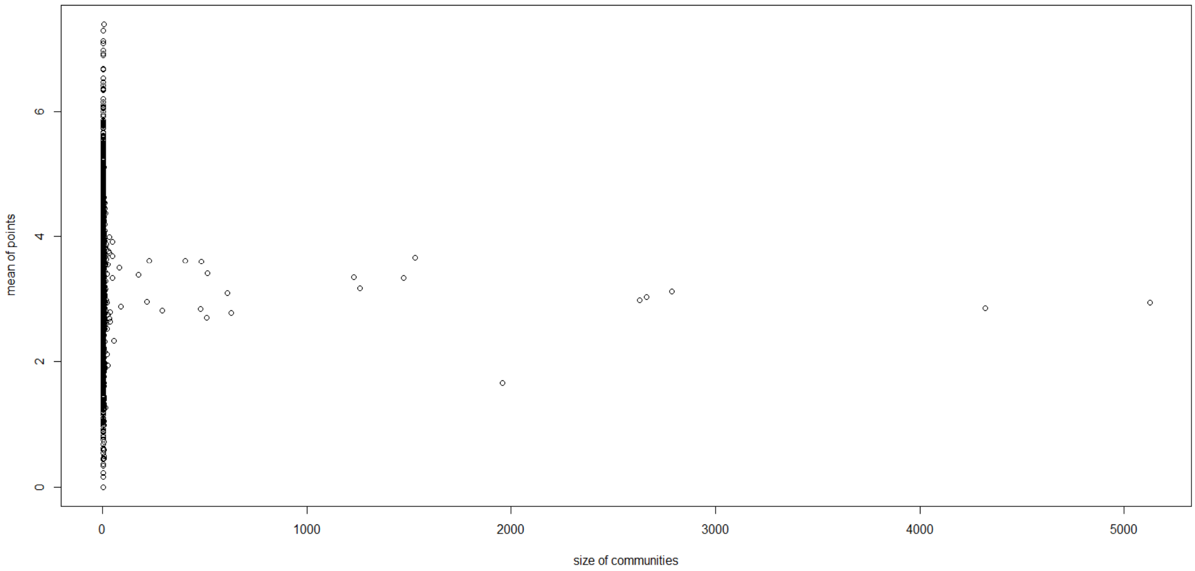
AGGREGATION OF LN_POINTS BY COMMUNITIES FORMED BY FAST GREEDY ALGORITHM:
Aggregating and summarizing the points by communities for all the communities obtained by the fast greedy algorithm:
The below plot of the mean and standard deviation shows that most communities have a mean which is on the left tail or around the median of the distribution. The same is true of the standard deviation of the distribution of means.



If we plot the ***sum of points in a community against the size of the community***, the correlation is strong as seen from the below plot.



However, if the ***mean of points was plotted against the size of community***, the largest community does not have a higher mean as evident from the below plot:



This shows that the largest communities may not be formed by the users who have won the most number of points.

From the plots of the Sum of points against size of communities and the plot of mean against the size of communities, it can be inferred that even though the largest communities generate the most knowledge in the SAP network, the largest communities may not have the "largest contributing nodes of the network in the dominant community" .

CONCLUSION:

- The communities within the SAP online network are not formed due to external factors related to a country or the GDP, FDI or the information technology import volume of a country. The communities are formed purely on the basis of knowledge transfer. The intuition is that these communities are formed due to the "various areas of technical expertise" within the SAP network. Thus these are not limited by countries or other such factors.

- The observation regarding the award points in a community and for nodes, show that a community may not have the nodes that contribute the "most" knowledge. This is also intuitive as moderate or less amount of knowledge means that members reach out to other members outside the community thus expanding the community.

- The high number of nodes with zero degree indicate that there is a potential in future of new communities being formed when these nodes share their knowledge with existing communities or nodes. The overall community structure of the SAP network will be highly sensitive to time due to this factor.