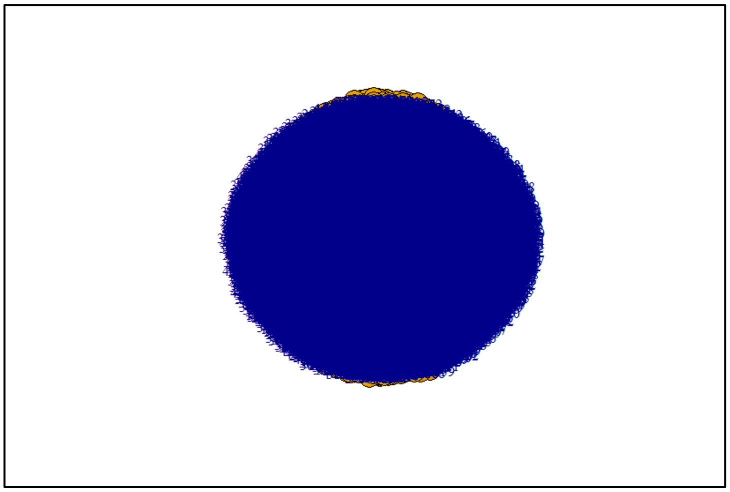
Devesh Sharma
UIN: 664049903

# ADVANCED LAB 3

## Community structure of the SAP Online Knowledge Community Platform

## Original SAP Community Structure:

### Visualization:



### Analysis and Description:

*The original graph:*

- # Edges: 57,102
- # Vertices/ Nodes: 98,288
- Weighted Graph: True
- Simple Graph: False
- Strong Connected Graph: False
- Weak Connected Graph: False
- Directed Graph: True
- Transitivity: 0.005
- Diameter: 29

*After simplifying the graph:*

- # Edges: 52,132
- # Vertices/ Nodes: 98,288
- Weighted Graph: True
- Simple Graph: True
- Strong Connected Graph: False
- Weak Connected Graph: False

- Directed Graph: True
- Transitivity: 0.005
- Diameter: 33

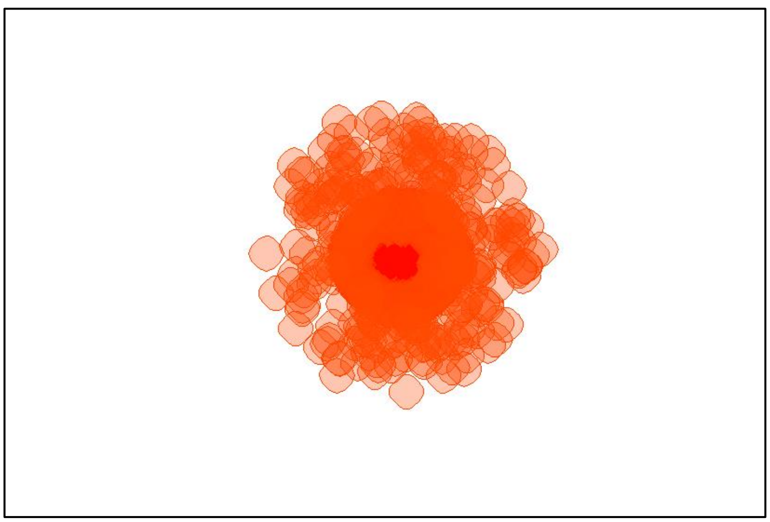
This showcases that the simplified graph reduces the number of edges (by combining edges to form directed ones) and increases the diameter as it removed the self-loops and multiple edges. Low transitivity suggests very low interconnectedness in the network. Since, the size of the network is very large, therefore it is neither strongly nor weakly connected.

# Community Detection Algorithms:

## Fast Greedy:

This algorithm will not work on this graph because it works for undirected graphs only whereas the SAP community network which we have is a directed graph. Therefore, I converted the graph to be undirected and then ran the Fast Greedy algorithm.

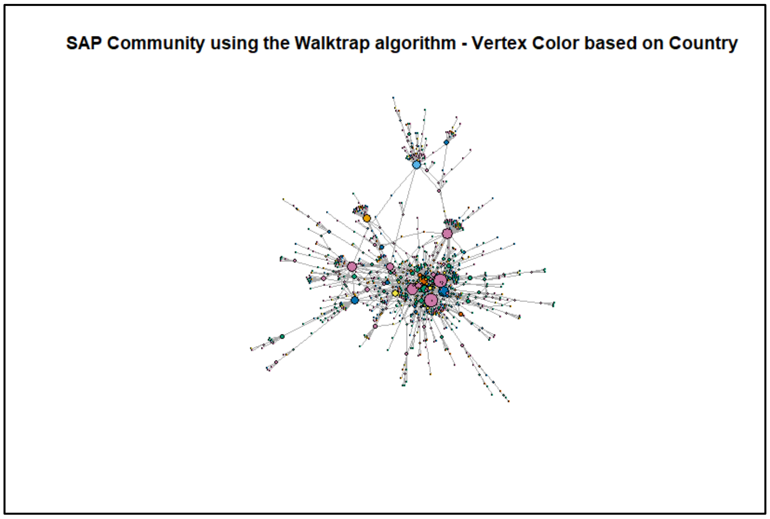
- Modularity = 0.795
- Length = 64,066



## Walktrap Algorithm:

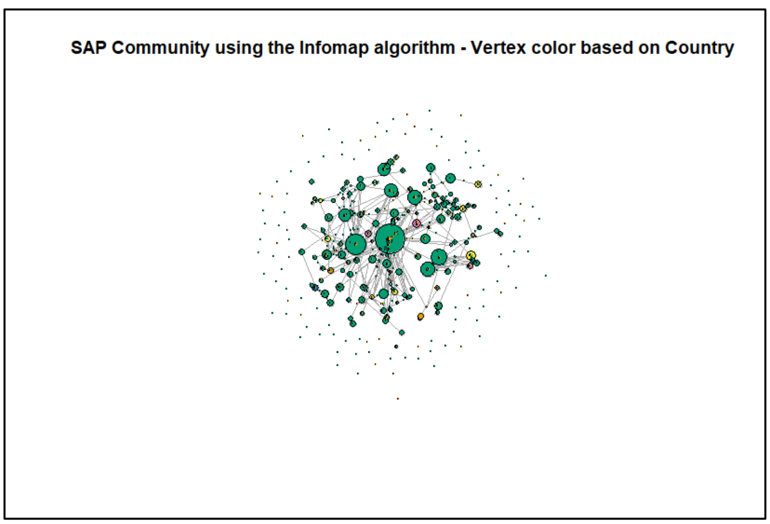
- Modularity = 0.698
- Length = 67,068

Devesh Sharma
UIN: 664049903




## Infomap Algorithm:

- Modularity = 0.529
- Length = 68,767



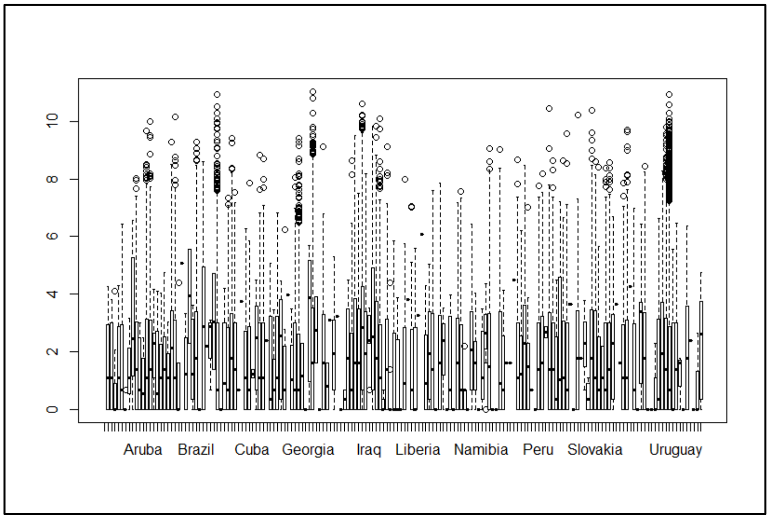

## Spinglass Algorithm:

This algorithm will not work on this graph because it does not work with unconnected graph and our SAP community graph is unconnected.

*Despite the slight differences between algorithms, they all identified around 60,000 communities which are homogenous (strongly connected) within themselves but heterogenous (weak connection)* *among each other. Number of edges are only half the number of nodes, which substantiates the observations from plots where high number of inactive users/ lurkers are present.*

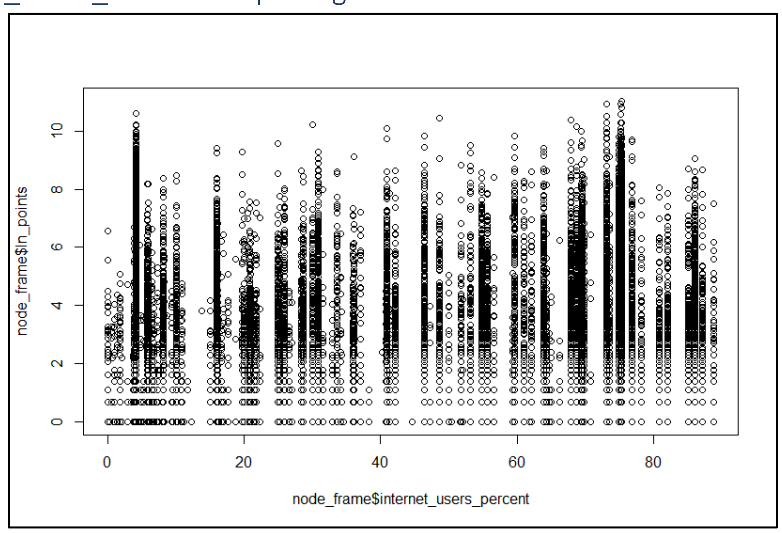
# Nationality and Other Observable Attributes:

## Impact of Country on placing into communities:



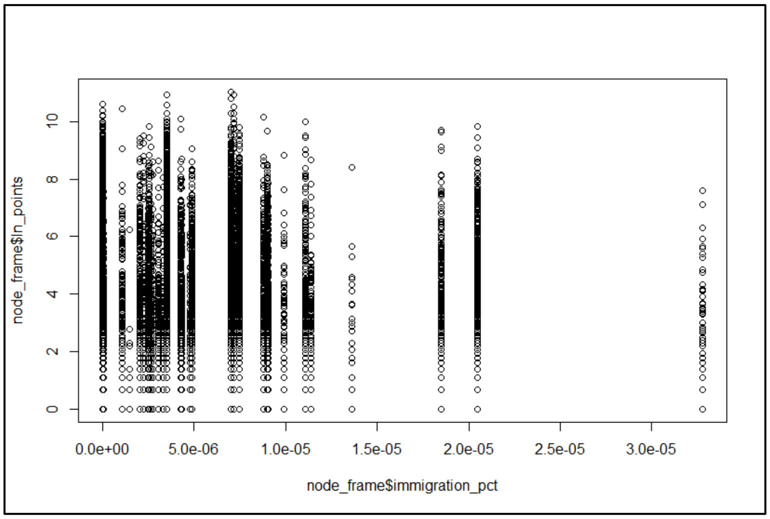

On further analysis of the data, top 10 countries with descending number of users: India, the United States, Germany, the United Kingdom, Canada, Australia, Spain, China, Netherlands, and France. These 10 countries represent 85% of users in the network, out of 169 total countries.

## Impact of Internet_Users_Percent on placing into communities:



Devesh Sharma
UIN: 664049903

### Impact of Immigration_Pct on placing into communities:




We can observe from the above plots that communities are not showing any preferential bias towards user nationality. This makes sense also because belonging to a certain nationality should not give any higher probability to be assigned in certain communities. If this would have been a social media platform, then we could have expected nationality to play a role in communities.

## Community Structure and Points Generated:

ln_points are award points earned by each user. When aggregated at the community level, this measure could represent the productivity level of each community in producing new knowledge in the system.

When applied to our SAP network, we can say that the communities with a high ln_points value would be generating more knowledge compared to the ones with low ln_points. We can assume that communities with high ln_points are communicating better with their community members. However, there could be a catch in this. Users can earn reward points by writing blogs or contents which are completely irrelevant to knowledge in theory and there is no way, with the current data, to segregate such cases