ADVANCED LAB 3

Siddhesh Tiwari

UNIVERSITY OF ILLINOIS AT CHICAGO

UIN:657796780

Contents

Visualize and Analyze Network Community

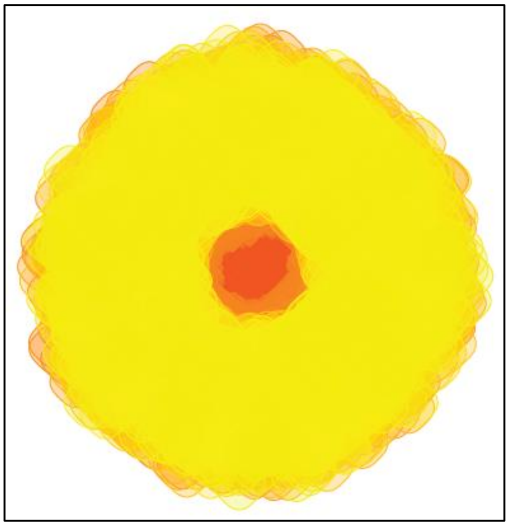


Fig 1.1 Community graph for whole data

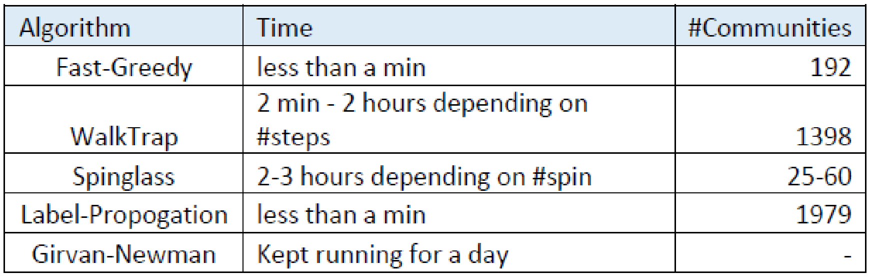
Above graph is the result of first run of fast-greedy algorithm on unprocessed graph i.e graph with all nodes and edges. The problem with this graph is that it's not at all discernible which is due to the fact that there were 63994 communities. This in turn was because of imbalance in number of edges (52067) and nodes (98288) in graph indicating that there are nodes which are not connected to any other node. Moreover, there were nodes which had only one or maximum of 2 edges. These disconnected nodes led to generation of such large number of community.

```
>length(sap_fast_raw)

[1] 63994

> vcount(g_simple)

[1] 98288

> ecount(g_simple)

[1] 52067
```

In order to get some understanding of community from network I decided to preprocess the graph for which I did the following 2 steps:

- Removing nodes which are single i.e not connected to anyone in network.
  ```
  > g_sap_sub <- subgraph.edges(graph = g_sap_simple,eids =
  E(g_sap_simple),delete.vertices = TRUE)
  ```
- Decompose the graph and get only connected part of the network which in our case had the largest number of nodes.
  ```
  > g_sap_decompose <- decompose.graph(g_sap_sub)
  > g_sap_final <- g_sap_decompose[[1]] #every other connected network had only 2-3 nodes
  ```

After getting the connected portion of network I ran Fast-Greedy, WalkTrap, Spinglass, Label-Propogation & Girvan-Newman algorithm on network giving the following result:



Algorithm	Time	#Communities
Fast-Greedy	less than a min	192
WalkTrap	2 min - 2 hours depending on #steps	1398
Spinglass	2-3 hours depending on #spin	25-60
Label-Propogation	less than a min	1979
Girvan-Newman	Kept running for a day	-

Table 1.1 Algorithm Runs

From the above table it was evident that Fast-Greedy was faster and efficient in detecting community as compared to others. Therefore, I decided to do my further analysis using Fast-Greedy algorithm.

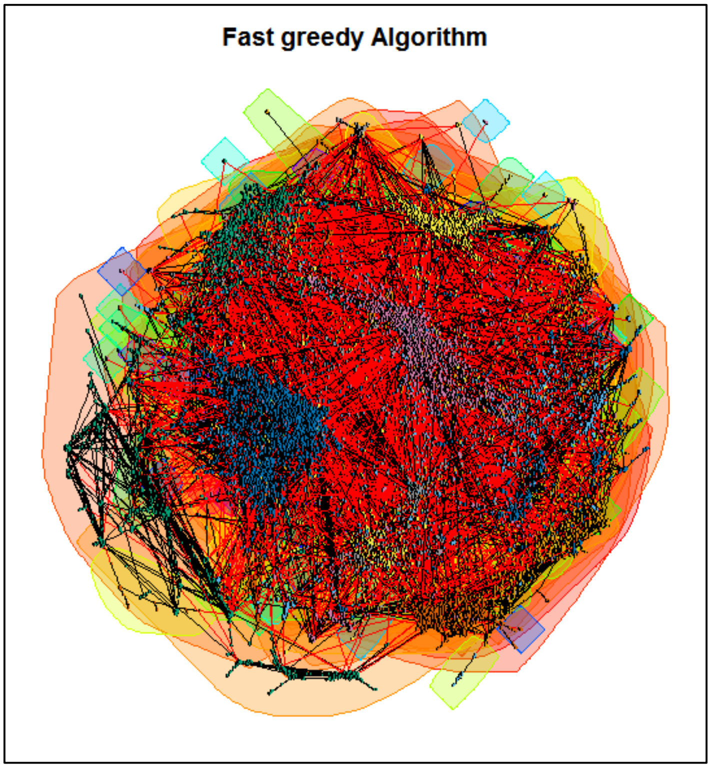


Fig 1.2 Community graph after preprocessing

The preprocessed graph consists of 31176 nodes, 48771 edges & 192 communities.

```
> vcount(g_sap_undir)
[1] 31176
> ecount(g_sap_undir)
[1] 48771
> length(sap_fast)
[1] 192
```

The graph obtained after preprocessing is much comprehensive from what we had seen before. In *Fig 1.2* I can distinguish among different communities by looking at the colors. Inspite of being more comprehensive than before it's difficult to see a clear separation of communities in this massive network which is limiting my understanding of communities here. Maybe further analysis of communities in next section based on their node attributes will give us more insight.

Community Analysis Based on Attributes

Attributes of nodes like Country, ln_fdi, fdi_pergdp, ICT_goods_import_percent, internet_users_percent, immigration_pct, lat, & lng are purely correlated to Country. Therefore, it's not useful to keep all these attribute when they can be represented by Country. My final node attributes had Country & Ln_points.

For purpose of easing the analysis of community based on nationality I did following preprocessing of data:

- Obtain a data frame giving Community, Country, Frequency
  ```
  > country_final <- get.vertex.attribute(g_sap_undir, "country")
  > fast_table <- data.frame(table(c.m.fast,country_final, useNA = c("no")))
  ```
- Remove tuples which had missing value for country.
- Remove those countries from analysis which had less than 20 nodes.

Using the data frame obtained from preprocessing I obtained following graph giving the distribution of countries in 192 communities

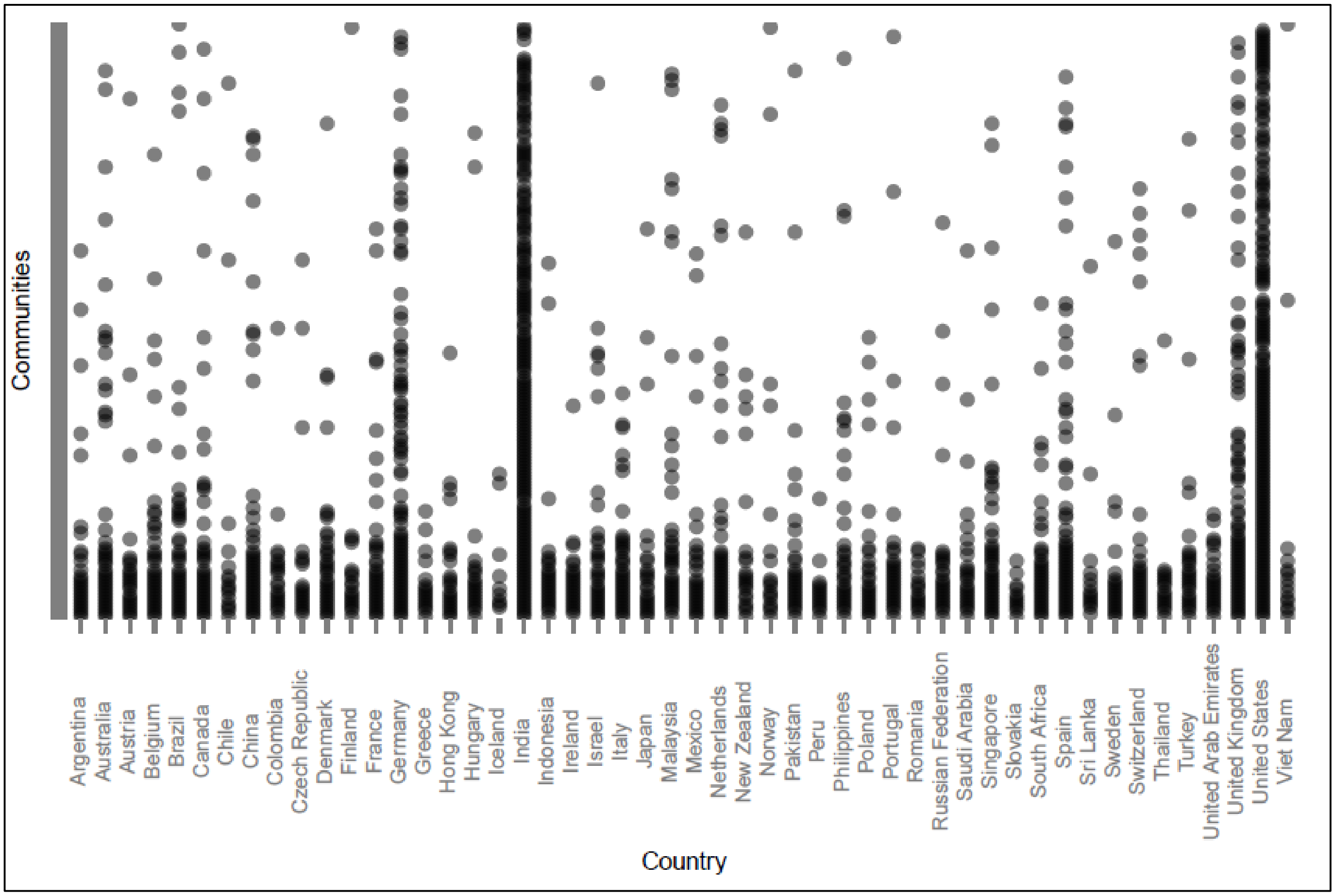


Fig 1.3 Country distribution in Communities

Analysis of the above graph tells us that nationality of individuals is not responsible for placing them in distinct communities. Most of the communities at the bottom of community spectrum have nodes from almost every country, the dark portion at the bottom of the graph signifies this. Countries like India, United States, United Kingdom & Germany with large number of nodes are prevalent in most of the communities this is evident from the continuous data points plotted against the community by these countries in *Fig 1.3*.

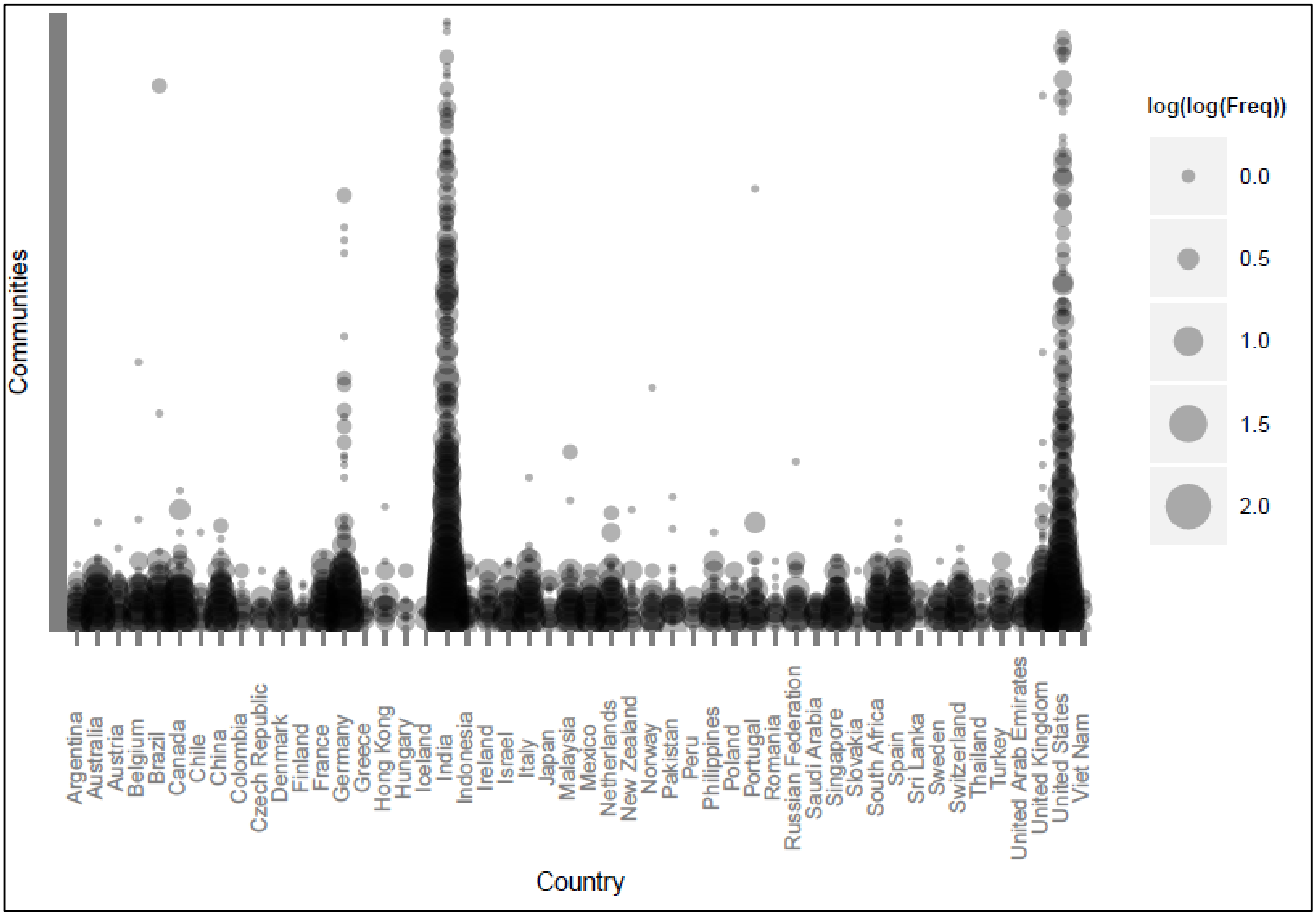
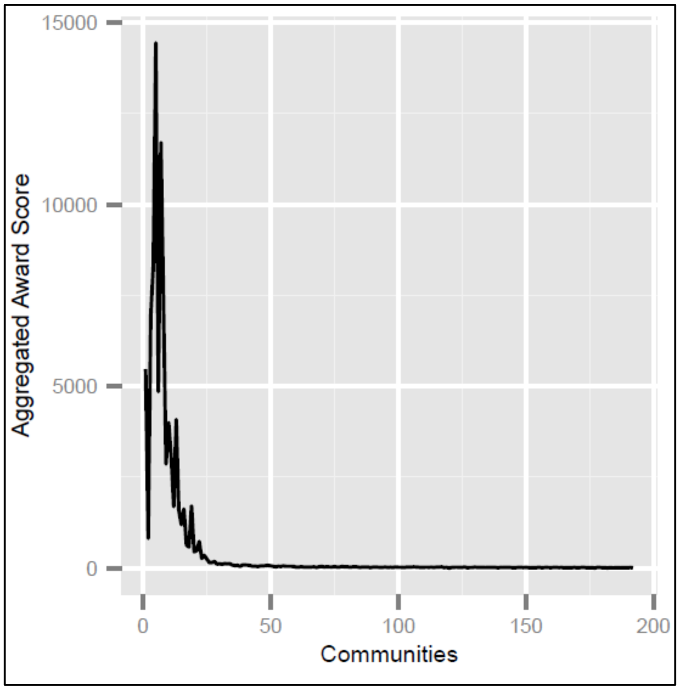


Fig 1.4 Bubble Chart of Country distribution in Communities

To strengthen my point about larger countries(node wise) are prevalent among different communities we can check the *Fig 1.4* which clearly shows how India and United States are prevalent in most on the communities. Not only they are prevalent in all communities but they are dominant within the communities in which they exist, it is evident from the bigger circles representing their frequencies in different communities.

So, communities in our case are not segregated based on nationality but contains a good distribution of nodes from different countries at bottom of community spectrum and most of the community in community spectrum consists of countries which have larger number of nodes.

Knowledge Contribution at Community Level



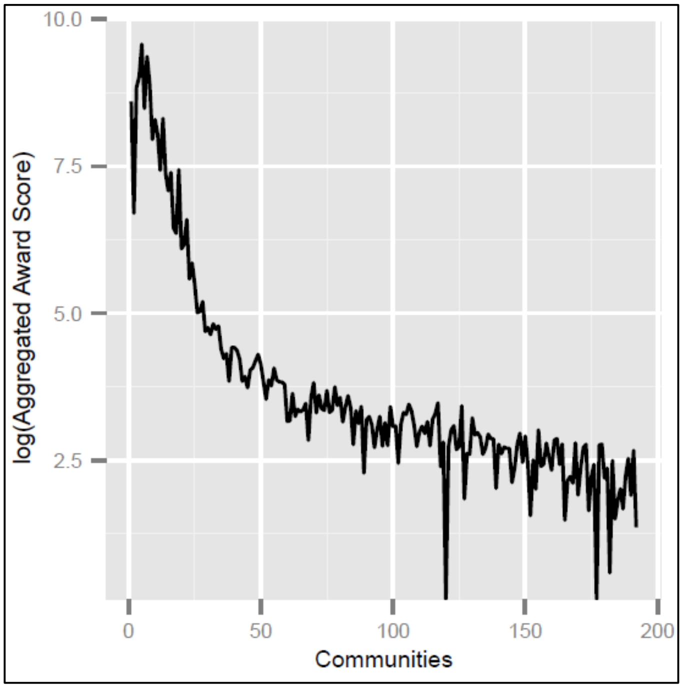


Fig 1.5 Aggregate Award Score vs Communities

Fig 1.6 log(Aggregate Award Score) vs Communities

By aggregating ln_points attribute at community level I obtained the above graph which clearly shows productivity of different communities. Communities till level 25 on the left have award points above average, it is in these communities where most of the knowledge contribution is made. As we go more towards left the award point increases indicating highly productive communities. We can say that these communities are highly productive as they have members from almost every country as visible from Fig 1.4 in previous section. The availability of nodes from various countries in these communities may be a cause of making them more productive.

As we go towards right the aggregated award point in community decreases along with decrease in number of nodes in the community. These communities are less productive and it's because of the fact that there are relatively less number of nodes which contribute to the knowledge of society.

The award point variable significantly brings out the knowledge contribution within a community.