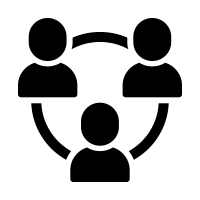
Gautam BHOWMICK

gautam.bhowmick@capgemini.com



nEtwork analysis

Facebook Network of Douglas Luke in 2013

**Introduction**

The data set selected for the network analysis project is part of the network dataset ‘UserNetR’ package in R. This igraph network object contains the Facebook friends of Douglas Luke and their friendship connections, as of 2013. The network contains a list of the 93 Facebook friends of Douglas Luke, as of 2013. The data were all obtained from Facebook using the Facebook application NameGenWeb. Due to changes in the Facebook privacy policy and data API, it is now no longer possible to download the data in the same way.

**Scope**

* Visualize and explain key characteristics of Network.
* Use multiple community detection algorithms and explain findings.
* Use network modeling using ERGM.

**Key Questions**

* Which community structure is most interesting?
* Which is best model fit for ERGM modeling?

|  |
| --- |
| **Visualization and Network Summary** |

**Dataset **

|  |  |
| --- | --- |
| The data was picked up from the data sets in package ‘UserNetR’. The whole data set is being used for this project.   * A network object with 93 vertices and 323 edges * Here we see some of the rows contained in the edges   and vertices csv files.   * The network is undirected   Node characteristics:   * name - Label for Facebook member * group - Code to indicate type of friend:   + F - Family   + W - Work   + S - Spiel (German for game or boardgame)   + B - Book Club   + M - Music   + H - High School   + C - College   + G - Graduate School | **Edges** |

|  |  |
| --- | --- |
| * sex - Gender * relationship\_status - Facebook relationship status * friend\_count - How many friends each member has * mutual\_friend\_count - How many friends each member has who were also friends with Douglas Luke | **Vertices** |

**Exploratory Data Analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| **Network Statistics** | **Value** | **Network Statistics** | **Value** |
| Node Count | 93 | Density | 0.07550257 |
| Edge Count | 323 | Connectivity - Strong | False |
| Directed | False | Clustering Coefficient | 0.6662791 |
| Size | 93 | Average degree | 6.946237 |
| Diameter | 4 | Modularity | 0.6145798 |
| Hyper | False | Loops | False |
| Multiple | False | Bipartite | False |
| Clusters | 10 | Diameter | 3 |
| Modularity | 0.6145798 | Missing edges | 0 |

**Network Summary**

1. Graph Density = 0.075;

From the statistic, as the value is closer to zero the graph is sparse as observed and therefore a majority of available connections have not been completed as expected, which means a weak network.

1. Average Degree: 6.946237;

Average degree of the graph indicating that this network has fair number of relations (or edges).

1. Modularity = 0.615;

which is a measure of the strength of division of a network into groups or clusters. A high modularity score indicates sophisticated internal structure. From the network statistic, it can be interpreted that the network has intermediate internal structure in complexity.

1. Diameter = 4;

A diameter of 4 suggests that this network is compact.

1. Clustering Coefficient = 0.666;

The transitivity for the network suggesting a high level of clustering. Based on the higher value Clustering the tendency to formed closed triangles is high in this network.

**Network Visualization – Figure I**

|  |  |
| --- | --- |
| Normal Network | Normal Network (With label) |
|  |  |

**Advance Network Visualization – Figure II**

|  |  |
| --- | --- |
| Circle Layout | Sphere Layout |
|  |  |

**Network Group Visualization – Figure III**

|  |  |  |
| --- | --- | --- |
| Generate node colors based on Group and and Legand | | |
|  | | |
| **Subgroup/Community Detection** |

The social systems contained in networks often exhibit complex structures. Many social networks are made up of relatively densely connected subgroups. Following techniques to examine subgroups that may be contained in larger social network.

|  |  |
| --- | --- |
| **Edge-betweenness:** Modularity is fairly high (0.631). The membership function reveals that 14 different subgroups have been identified. | **Fast-greedy:** Modularity is fairly high (0.632). The membership function reveals that 12 different subgroups have been identified. |
|  |  |
| **Label Propagation:** Modularity is fairly high (0.619). The membership function reveals that 14 different subgroups have been identified. | **Leading eigenvector:** Modularity is fairly high (0.635). The membership function reveals that 10 different subgroups have been identified. |
|  |  |

|  |  |
| --- | --- |
| **Louvain:** Modularity is fairly high (0.636). The membership function reveals that 13 different subgroups have been identified. | **Cluster Optimal:** Modularity is fairly high (0.638). The membership function reveals that 12 different subgroups have been identified. |
|  |  |

**Summary:** These results show that all the detection algorithms identify subgroups between 10 to 14. Modularity ranges from 0.619 to 0.638.

Louvain community structure, which I find most interesting and it holds second highest modularity and subgroups. The method consists of repeated application of two steps. The first step is a "greedy" assignment of nodes to communities, favoring local optimizations of modularity. The second step is the definition of a new coarse-grained network in terms of the communities found in the first step.

|  |
| --- |
| **ERGM Models (Exponential-family Random Graph Models)** |

|  |  |
| --- | --- |
| **Null Model:** A baseline model to judge how much subsequent models are improving. This model only has one term, edges which produces random graph model. The coefficient of edges is negative (-2.50508) indicating the density of the network is less than 25%. |  |

|  |  |
| --- | --- |
|  | **Null Model characteristics:** Simulating 100 networks based on the NULL model (analysis done with “edges” only). The ‘x’ shows there are close to 750 triangles in the Facebook network, but the simulations based on the NULL model shows very few triangles. |

|  |  |
| --- | --- |
| **Model (I) – Adding Sex attribute:** The large p value after ‘sex’ attribute was added shows that gender is not a significant variable in predicting connection between two nodes. AIC value is slightly larger than the null model indicating that the null model is a better fit. |  |

|  |  |
| --- | --- |
|  | **Model (II) – Adding Sex and Group:** Low p values for Group C, F, G and H shows that group is a significant variable in predicting the connection between two “nodes”.  The AIC value of 2007 is smaller than the one obtained for the null model, and Model (I) indicating better fit. |

**Goodness of Fit 1**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | edges | degree0 | degree1 | degree2 | degree3 | degree4 | degree5 | Triangle |
| Facebook | 323 | 3 | 5 | 11 | 11 | 7 | 15 | 746 |
| Null | 330 | 0 | 1 | 0 | 1 | 14 | 10 | 53 |
| Model(I) | 297 | 8 | 2 | 11 | 9 | 11 | 11 | 128 |
| Model (II) | 317 | 0 | 0 | 3 | 2 | 9 | 13 | 47 |

Simulation statistics shows that **Model (II)** model performed much better than other two models so far. Overall edges, and number of nodes close to original network.

**Goodness of Fit 2 -> [Model (II)]**

|  |  |
| --- | --- |
|  | High p value for “goodness-of-fit for degree” shows that the there are no significant differences when it comes to number of isolates (0 degree) between the simulated networks and the original network. And same can be said for all degree nodes from 0 to 32 (except for nodes with degrees 1,2,3,5,8,15 and 32 as they all p values less than .05). |
|  |  |

**Goodness of Fit Evaluation-> [Model (II)]**

|  |  |  |
| --- | --- | --- |
| The fact that dark line is within the confidence interval  bands (grey lines) indicates good fit  for degree. However dark line 50 % above grey line in case of dyads and edgewise shared partners. |  |  |
|  |  |

**Conclusion: Model (II)** matches the original model better, with fair goodness of fit.