Comparing a web graph with an Erdős–Rényi random model

Sameen Islam, 140869323

s.islam@se14.qmul.ac.uk

05.03.2018

I. Introduction

The 'web-edu' graph used in this analysis [1] have 3037 nodes which represent web pages. There are 6480 edges which represent a hyperlink between two web pages. 'web-edu' is an undirected sparse network, so we expect it to have a smaller number of links than the maximum possible. Sparse networks are scale-free, therefore the degree distribution should be as-per the power-law (long-tail):

$$P_{deg}(k)\alpha k^{-\gamma} \tag{1}$$

where γ is some exponent and $P_{deg}(k)$ decreases slowly as degree k increases. This results in an increased probability of finding a small number of nodes with a high degree (hubs), with the majority of the nodes having a smaller degree.

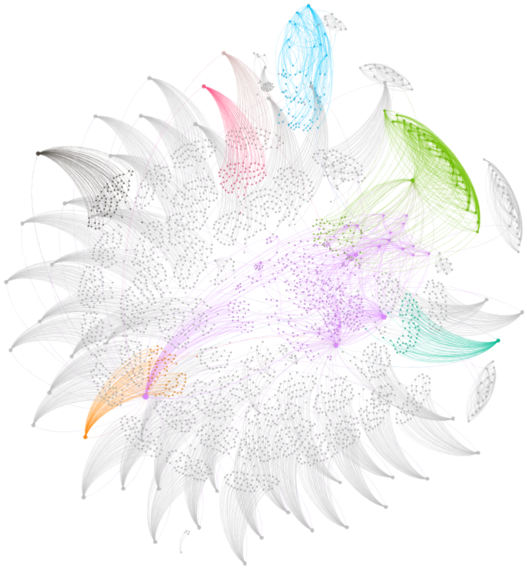

Figure 1: 'web-edu' network

The networks were visualised and analysed using using a combination of Gephi [2] and iGraph [3] based on R.

II. Generating Random Graph Model

The Erdős–Rényi (E-R) random graph model have two variants. In the first form $G(n, p)$, there are n vertices and an edge is inserted between any pair of vertices independently with probability p. In the second form, $G(n, M)$ there are n vertices and the number of edges, m is specified. A graph with m edges is then chosen uniformly, at random.

We chose to generate our random model using the $G(n, M)$ variant as we want the random network to be connected - just as our 'web-edu' dataset is. Our rationale was that since the $G(n, p)$ model inserts edges at random between nodes, it goes to show that if the wiring probability, p is too low then not all nodes will be connected and if it is high, the network will become a mesh which does not resemble a scale free network.

Figure 2: E-R network

However, it should be noted that the E-R model due to it's probabilistic nature does not generate graphs that can be considered to be scale free, and therefore follow the power law distribution (1).

III. Degree Distribution

Degree distribution is the fraction of nodes in the network with the degree k. In Figure 3, we can see the network adhering to the long-tail distribution as there are many nodes with a small degree. From the figure, we see that over 2000 nodes have a degree of 3. We also find that most nodes in the graph have a degree between 5 - 25, while a rare few nodes have a degree over 100. This network has an average degree distribution of 4.267.

Comparing this with the E-R graph (Figure 4), we see that the average degree distribution is 4.338 which is quite close to our web graph. However, as expected with a random graph, we find that most of the nodes in the network have a similar number of degree (approx. 3 to 12 degrees).

IV. Clustering Coefficient

This is a measure of the extent to which nodes in a graph tent to cluster together. From the clustering coefficient value, we can find out how likely it is for two connected nodes to be part of a larger, highly connected group of nodes (clique).

Our analysis shows that the web graph has a network clustering coefficient of 0.652 (Figure 5). The network clustering coefficient is computed by finding the mean from individual nodes in the network. A coefficient value of 1.00 would show that the network is highly connected (mesh). The E-R model we generated gave a clustering coefficient value of 0.002 (Figure 6), which is to say that it's neighbours are not connected at all.

V. Modularity

Modularity tells us how densely nodes are connected in a network. This metric is commonly used for detecting communities. A community can be thought of as a densely connected group of nodes within the network.

In our web graph, we detected 60 communities using [2] with an overall modularity value of 0.952 (Figure 7). In our E-R graph, used the same community detection technique and found 30 communities with a modularity of 0.485 (Figure 8). This would suggest that the communities in the web graph are more tightly connected than the random model. If we compare this with the betweeness centrality of the web graph (Figure 9), we can see that a very small fraction of nodes exist which have approximately 1.25 million shortest paths passing through them.

VI. Centralities

Betweeness centrality help us identify nodes which connect the most different communities together. These nodes are the most 'influential' in the network and can affect large amounts of nodes if removed. This metric is computed by considering two connected nodes in a network. Between these nodes, there exists at least one shortest path that pass through both of these nodes. The betweeness centrality is the number of such shortest paths passing through the pair of nodes.

The diameter of a graph can also give us a frame of reference when considering shortest paths through a network. Diameter is defined to be the shortest distance between two most distant nodes in the network. We find that the $diameter = 11$ in our web graph compared to our E-R model which shows $diameter = 12$ which would suggest the size of the two networks are similar.

Our E-R graph's betweeness centrality distribution (Figure 10), shows that there are no influential nodes in the network as most nodes have a betweeness centrality value of 0. However, as previously mentioned, the web graph does contain hubs which connect different communities because there are a number of nodes which have a betweeness centrality $\geq$ 250k (Figure 9).

References

[1] Ryan A. Rossi and Nesreen K. Ahmed. "The Network Data Repository with Interactive Graph Analytics and Visualization". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015. URL: http://networkrepository.com.

[2] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. "Gephi: an open source software for exploring and manipulating networks." In: *Icwsm* 8 (2009), pp. 361–362.

[3] Gabor Csardi and Tamas Nepusz. “The igraph software package for complex network research”. In: *InterJournal, Complex Systems* 1695.5 (2006), pp. 1–9.

VII. Appendix

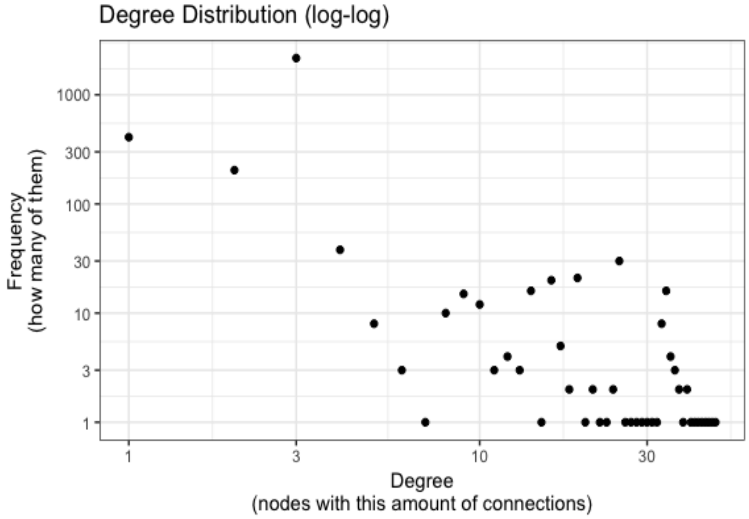

Figure 3: Degree Distribution of ’web-edu’ network (log-log plot)

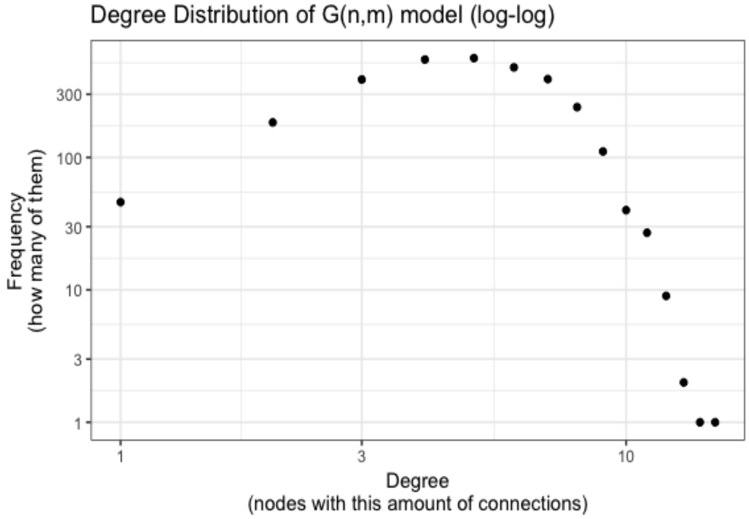

Figure 4: Degree Distribution of random network (log-log plot)

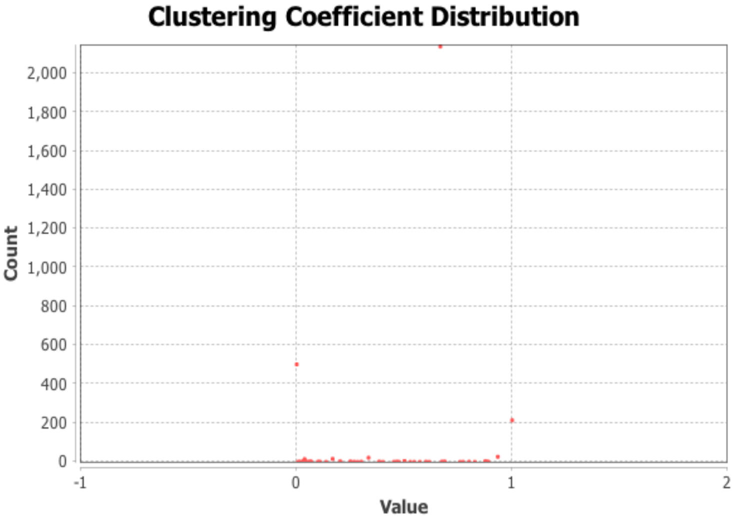

Figure 5: Clustering coefficient distribution of 'web-edu' network

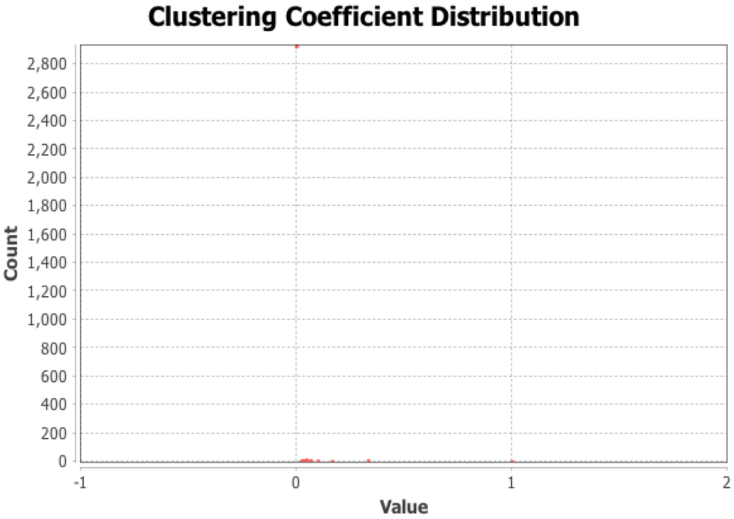

Figure 6: Clustering coefficient distribution of random network

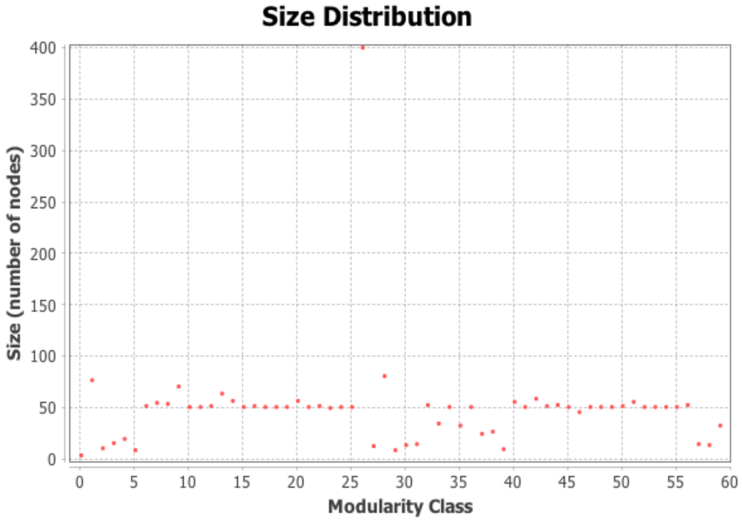

Figure 7: Size distribution of communities in ’web-edu’ network

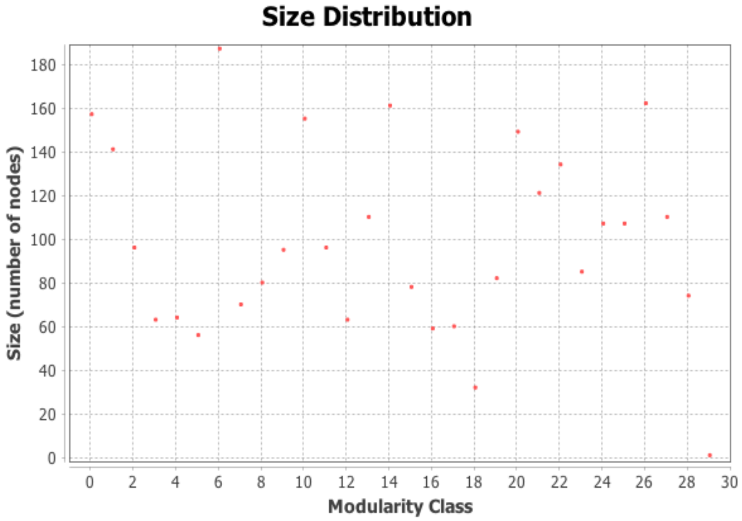

Figure 8: Size distribution of communities in random network

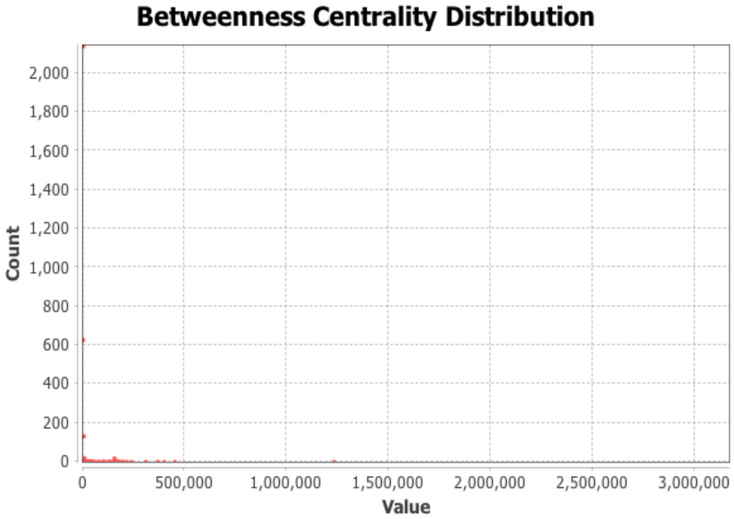

Figure 9: Betweeness centrality distribution of 'web-edu' network

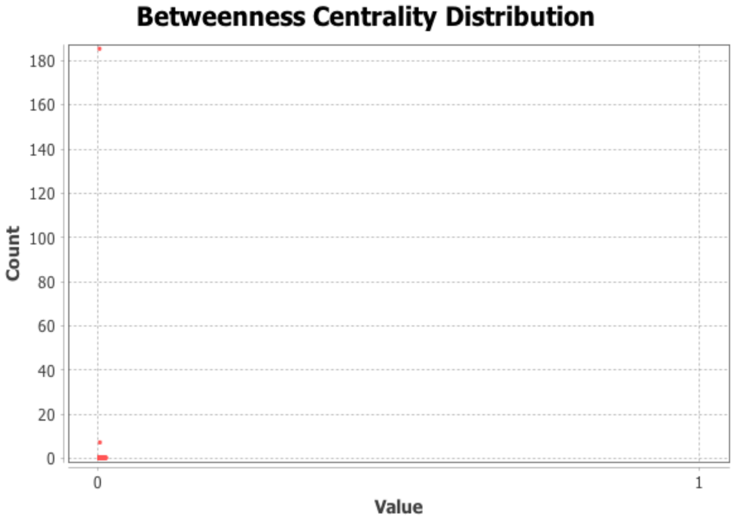

Figure 10: Betweeness centrality distribution of random network