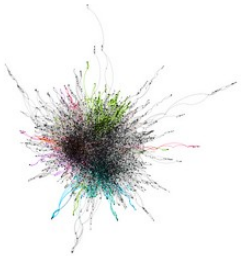
Online Social Networks and Media

Assignment 2

Full Name – Student ID: Terizi Chrysoula
Date: 04 – 12 – 2019

Question 1

There is a variety of scientific branches which have studied and analyzed the issue of clustering [1, 2, 3]. Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group are more similar, in some sense, to each other than to those in other clusters. The most discussed algorithms for community detection based on modularity measure are the Newman's spectral method using a fine-tuning stage and the method of Clauset, Newman, and Moore (CNM) [4]. The hierarchical agglomeration CNM algorithm [5] was implemented in the context of this assignment.



The given DBLP10 dataset consists of ~ 31k users and ~ 106k undirected connections between authors. The average degree and the density of the graph is ~ 6.77 and ~ 0.000215 respectively. The link express that a publication have been written by $author_i$ and $author_j$. Snap library implements the CNM algorithm **CommunityCNM(Graph, CmtyV)**[1] where at every step two communities that contribute maximum positive value to global modularity are merged. The total number of clusters is 1953 and the modularity of the network is 0.728547. The modularity of each cluster was calculated using the function **GetModularity(Graph, NidV, Gedges=-1)**[2].

Observing Figure 1(a), there are only two big clusters which consist of ~ 5k to ~ 6k individuals, five clusters from 1000 to 2000 users and the rest of the groups (about 98 percent, see Figure 1(b)) contain a few dozen authors. The modularity distribution follows the same form as this community size. About modularity measure, this is about a system property which

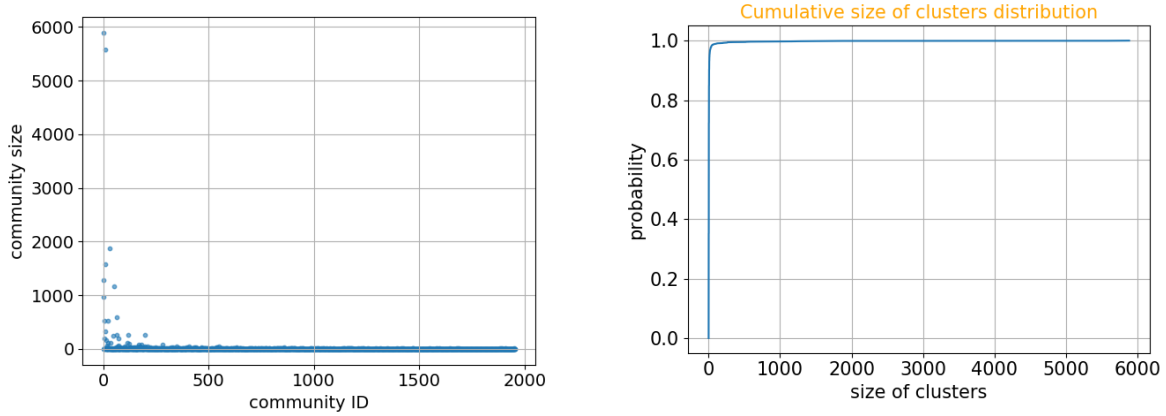


Figure 1: (a) Plot for each community, the total number of authors (cluster's size) it consists of, (b) Plot the cumulative distribution of the size of clusters.

measures the degree to which densely connected compartments within a system can be decoupled into separate communities or clusters which interact more among themselves rather than other communities [6]. The value of the modularity for unweighted and undirected graphs lies in the range $[-\frac{1}{2},1]$ [7]. It is positive if the number of edges within groups exceeds the number expected on the basis of chance. In Figure 2(b) (similar to Figure 2(a)), it is observed that for this

1 https://snap.stanford.edu/snappy/doc/reference/CommunityCNM.html

2 https://snap.stanford.edu/snappy/doc/reference/GetModularity.html

particular implementation, there are only positive modularity values and the modularity value for the two largest groups is close to 0.15. Similar to the size distribution (see Figure 1(b)), the 98 percent of the modularity values is too low (see in Figure 2(b)).

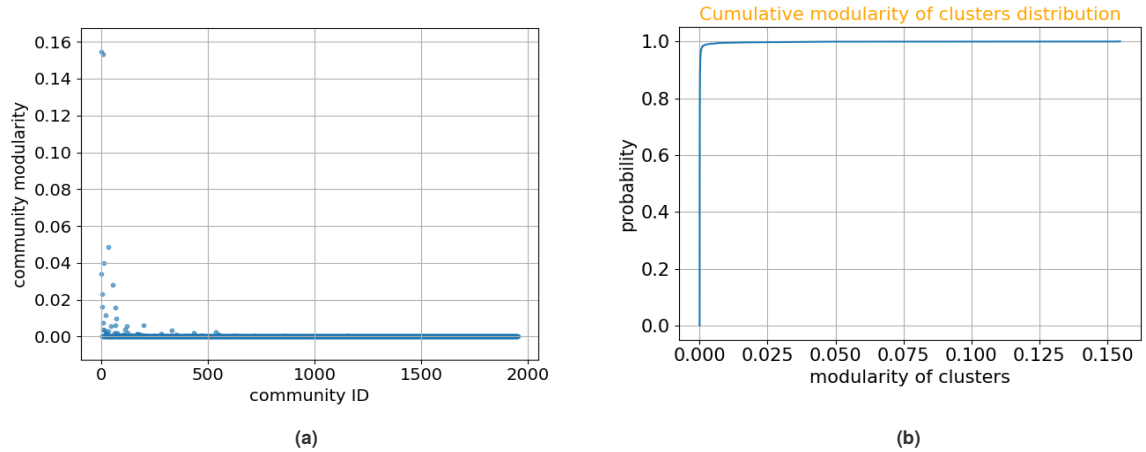


Figure 2: (a) Plot for each community its modularity value, (b) Plot the cumulative distribution of modularity values of clusters.

Homogeneity

In statistics, homogeneity arise in describing the properties of a dataset, or several datasets. They relate to the validity of the often convenient assumption that the statistical properties of any one part of an overall dataset are the same as any other part. In networks, homogeneity of the clusters means that all of the observations with the same class label are in the same cluster.

The labels of the users were needed to evaluate the homogeneity of the clusters which has emerged from the CNM algorithm. Each author is characterized by a set of conferences (let it be $conf_i$ for author i) in which he has published an article. There are 12 different names of conferences and the process of calculating homogeneity of the clusters is the following,

1. For each pair of clusters C_i and C_j :
2. For each combination of authors (let it be $a_{cluster\ ID}$) in C_i and C_j :
3. Calculate JaccardSimilarity($a_i \in C_i$, $a_j \in C_j$) = $\frac{|conf_i \cap conf_j|}{|conf_i \cup conf_j|}$
4. Calculate the average similarity = $\frac{\sum Jaccard\ similarities(step3)}{total\ number\ of\ combinations(step2)}$

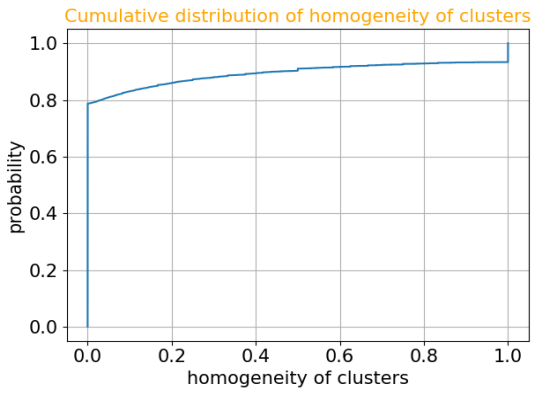


Figure 3: CDF plot for homogeneity of the clusters.

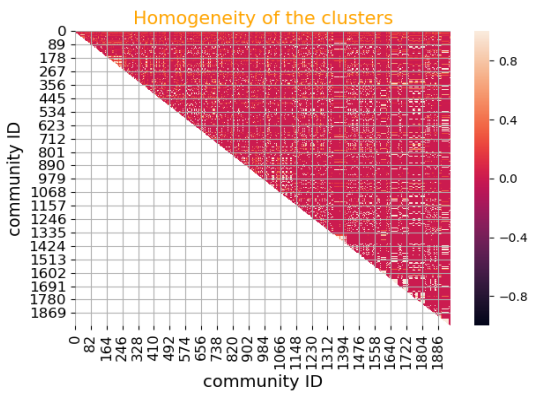
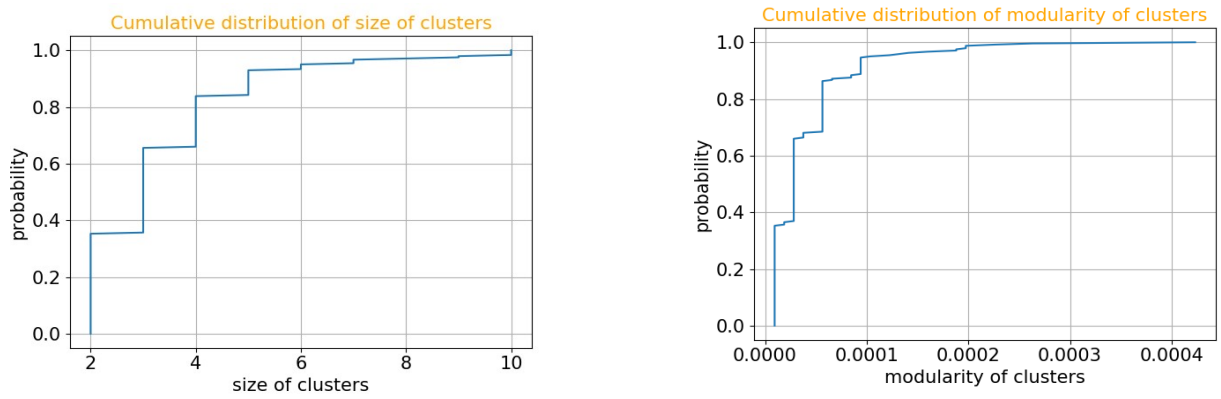


Figure 4: Heatmap plot of the homogeneity of the clusters for all cluster combinations.

In Figure 3, notes that the ~ 10 percentage of clusters has homogeneity over 0.50 which means that only a small percentage of the total number of clusters' combinations consists of authors with similar labels. The total number of clusters' combinations is ~ 1.5m and the percentage of pairs with average similarity over 0.85 is ~ 6.87%. All average jaccard similarities for all clusters' combinations are shown in Figure 4. Does not exist a group of authors or a set of groups of authors which achieve high similarity with the rest of all clusters. For the top clusters, clusters which achieve high similarity with more than ~ 200 clusters, have not a large number of authors in their groups and their modularity is low. The ~ 65% of the top clusters contain from 2 to 4 individuals and the ~ 2 percentage of the top clusters has size equals to 10 (see Figure 5(a)). Similar, the modularity of the top clusters is not high, around ~ 69 percentage of the clusters have modularity less than 0.0001 value (see Figure 5(b)).



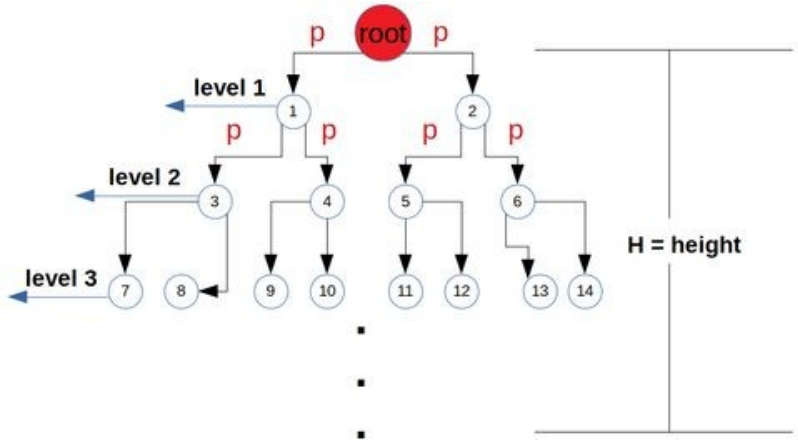
(a) (b)

Figure 5: (a) Cumulative distribution plot for the size of the top clusters in the network, (b) Cumulative distribution plot for the modularity of the top clusters that exist in the network.

1. Driver and Kroeber, "Quantitative Expression of Cultural Relationships", University of California Publications in American Archaeology and Ethnology, pages: 211–256, 1932
2. Zubin, Joseph, "A technique for measuring like-mindedness", The Journal of Abnormal and Social Psychology, vol. 33 (4), pages 508–516, 1938. doi:10.1037/h0055441
3. Cattell, R. B., "The description of personality: Basic traits resolved into clusters", Journal of Abnormal and Social Psychology, vol. 38 (4): 476–506, 1943. doi:10.1037/h0054116
4. Vinícius da Fonseca Vieira, Carolina Ribeiro Xavier, Nelson Francisco Favilla Ebecken, and Alexandre Gonçalves Evsukoff, "Performance Evaluation of Modularity Based Community Detection Algorithms in Large Scale Networks," Mathematical Problems in Engineering, vol. 2014, Article ID 502809, 15 pages, 2014. https://doi.org/10.1155/2014/502809
5. Clauset, Aaron, M. E. J. Newman, and Cristopher Moore. "Finding Community Structure in Very Large Networks." Physical Review E 70.6 (2004): n. pag. Crossref. Web
6. Ali Kharrazi, Brian Fath, Encyclopedia of Ecology (Second Edition), Elsevier, 2019, Pages 414-418, ISBN 9780444641304, https://doi.org/10.1016/B978-0-12-409548-9.10751-1
7. Brandes, U. Delling, D. Gaertler, M. Gorke, R. Hoefer, M. Nikoloski, Z. Wagner, D., "On Modularity Clustering", IEEE Transactions on Knowledge and Data Engineering, 20 (2): 172–188, February 2008, doi:10.1109/TKDE.2007.190689

Question 3

Consider a graph that is a binary tree of depth (height) H . Based on Independent Cascade model, each edge has probability P to transmit the disease/rumor/news on a non-active neighbor. At time t ,



- **t = 0** → The root is the only one active node (user) in the initial step

- **t = 1** → – The probability to fail (not transmit the news) at level 1 and and stop transmitting it is $(1-P)^{2^{level}}$

  – The probability to active one of each neighbors is $P(1-P)$

  – The probability to activate all of his neighbors is $P^{2^{level}}$

- **t = 2** → ** Look nodes 1 and 2. Each one of the nodes has two states active and non-active. If node 1 is non-active then he can not transmit the news to his neighbors and he stops. Hence, only the second node can transmit the news, if he had became active at time 1. E.g. node 2 is active therefore, at level 2:

  – The probability to fail is $[P(1-P)](1-P)^2$

  – The probability to fail one of its neighbors is $[P(1-P)][P(1-P)]=P^2(1-P)^2$

  – The probability to activate all his neighbors is $[P(1-P)]P^2$

  ** Both nodes are active:

  – The probability to fail is $P^2(1-P)^{2^{level}}$

  – The probability to active one of each neighbors is $P^{2^{level}}(1-P)^2$

  – The probability to activate all of his neighbors is $P^{2^{level}}P^2$

  .
  .
  .

- **t = H** → – At each time step only one node is active: $P^H(1-P)^H$, >> expected spread = $H+1$

  – At each time step only a node is active except the last step where two nodes are active: $P^{(H+1)}(1-P)^{(H-1)}$, >> expected spread = $H+2$

  – At each time step both two nodes are active: $\prod_{i=1}^{H} P^{2^i}$, >> expected spread = 2^H-1-2^{H-1}

  – At each time step both two nodes are active except the last step where only one is active: $P^2\prod_{i=2}^{H} P^{2^{i-1}}(1-P)^{2^{i-1}}$, >> expected spread = 2^H-1