# Analysis of ResearchGate, A Community Detection Approach

Mohammad Heydari, MSc in Network Science
Department of Information Technology Engineering
School of Industrial and Systems Engineering
Tarbiat Modares University
Tehran, Iran
m_heydari@modares.ac.ir

Babak Teimourpour, Assistant Professor
Department of Information Technology Engineering
School of Industrial and Systems Engineering
Tarbiat Modares University
Tehran, Iran
b.teimourpour@modares.ac.ir

*Abstract*— **We are living in the data age. Communications over scientific networks creates new opportunities for researchers who aim to discover the hidden pattern in these huge repositories. This study utilizes network science to create collaboration network of Iranian Scientific Institutions. A modularity-based approach applied to find network communities. To reach a big picture of science production flow, analysis of the collaboration network is crucial. Our results demonstrated that geographic location closeness and ethnic attributes has important roles in academic collaboration network establishment. Besides, it shows that famous scientific centers in the capital city of Iran, Tehran has strong influence on the production flow of scientific activities. These academic papers are mostly viewed and downloaded from the United State of America, China, India, and Iran. The motivation of this research is that by discovering hidden communities in the network and finding the structure of intuitions communications, we can identify each scientific center research potential separately and clear mutual scientific fields. Therefore, an efficient strategic program can be design, develop and test to keep scientific institutions in progress path and navigate their research goals into a straight useful roadmap to identify and fill the unknown gaps.**

Keywords—: *Social Network Analysis, Scientific Network, Louvain Algorithm, Network Centrality*

## I. INTRODUCTION

In the age of information, data known as the most important asset to the various types of large scale, medium and small enterprises around the world. Text data is ubiquitous and growing quickly. The web, blogs, emails, social networks provide millions of data in seconds and all of these resources can be a beneficial and powerful base for knowledge extraction and useful results. It's called the power of text analysis. Currently, social network platforms are under the scrutiny of many academic researchers and developers. The topic is widely utilized in various scientific, social, commercial and industrial aspects[1]. Information systems can be mapped In Complex network form that contains linked nodes and relation[2]. Social Network Analysis has a deep potential to acts a tremendous role in human activities around the world. By analysis of huge range of extracted data from various kinds of social networks, a number of useful hidden patterns emerges. Social Network Analysis is a type of structure analysis way expanding in many research fields which focuses on the related research and is mainly used to describe and measure the relationship and information individually.[3][4]

Social Network Analysis has been proved to be successful in studies of scientific collaboration network[5][6]. Interaction with other authors in social networks will increase the co-authorship, besides co-authorship increase citation[7]. According to the [8] ResearchGate scored the highest, 61.1 percent followed by Acadmia.Edu with 48.0 percent score ranked "above average" and Mendeley with 43.9 percent ranked "average". The success of RG has already enabled researchers to announce their ideas and share their publication free of charge to facilitate the remote access for the researchers all around the world[9]. Based on Nature survey[10], RG has 2nd rank in Science and engineering topic and 4th in social science, arts, and humanities. It should be noted RG network became bigger compares to the Nature survey in 2014. To prove the mentioned claim, we looked up in Google Trend Service. The visualization of searched data based on keywords proves that the RG has the most interest to be visited among other similar networks. RG growth is increasing every day and now it has more than +15 million members all around the world[11]and ranked 162 among all web pages in the world[12]. Recently RG Score has become a major source of academic papers[13].

The RG Online Scientific Network has a significant part of scholar's communication around the world. Based on a survey of Nature Journal, 48% of science and engineering researchers and 35% of social science, arts and humanities scholars visit RG regularly. The percent this social network is five more than Academia.Edu which is known as its nearest rival[10][14]. In RG participants can utilize their own profile to share and represent their ultimate publications and projects. It must be noted that Graphical User Interface of RG is better than rivals. The search section of the site has been implemented intelligently and impressively cause by a single phrase search you can achieve various results about Researchers,

Projects, Publications, Questions, Jobs, Institutions, and Departments that may search item relates to the mentioned filed. Since number of institutions haven't joined to, and they don't have official profiles, the score that system assigns to the scholars and institutions is still an argumentative problem. Based on RG Network, at the time of writing this paper, it's introducing a new trial metric to measure researchers score called Research Interest. Research Interest is how to measure scholarly interests in any research. This score is focused on research items and scientists' interactions with them, using concepts that are familiar to RG members. To provide an overview.

This score is focused on research items and researcher interaction with them, using concepts that are familiar to RG members. To provide an overview of a researcher's body of work, RG also added a Total Research Interest score, which simply adds up the Research Interest scores from all of an author's research items. When researchers read, recommend or cite a research item, its Research Interest goes up. RG decided on a system for weighting the different forms of interaction based on a reader has a weighting of one, a full-text read has a weighting of three, a recommendation has a weighting of five and a citation has a weighting of ten[15]. A comparison of growth rate of RG against another famous academic portal is shown in **Fig. 1**.



Fig. 1 Growth Rate of Interest in ResearchGate Network Compares to other Scientific Networks based on Google Trends Service[16]

## II.    RELATED WORKS

Abbasi et al,[7] extended a theory model based Social Network Analysis methods to investigate on collaboration network of Researchers by usage of famous Social Network Analysis methods (i.e., normalized degree, closeness, betweenness, eigenvector centralities and average ties strength and efficiency) for investigating consequence of social network on the Researchers efficiency on a certain topic (i.e., Complex networks). Abbasi also mention that Researchers should work with lots of students as an alternative for the rest of well-known Researchers.

Manca et al,[8] proposed a multi-level framework towards analysis of RG and Academia.Edu. His aim is two exemplify how these two online social networks are socio-technical systems that support researcher's knowledge activity sharing and professional learning. His proposed framework includes three layers: A. The "socio-economic layer", B. the techno-cultural layer, C. the networked-scholar layer. Then he adopted described social networks to these three layers.

Naderbeigi et al,[11] made an investigation on mapping profile of research activities of faculty members of Sharif University of Technology in RG. They intend to test the correlation h-index between the RG and Web of Science and Scopus and Google Scholar. Then investigate Sharif University of Technology faculty members' top h cited research RG in WoS, Scopus, and GS. After all, investigate the Altimetric score of SUT faculty members' top h cited research RG with Altimetric Explorer. They already mention that more than %75 of Sharif University of Technology faculty member has already set their profile on RG. Besides that, the highest correlation of RG Score is with h-index and citation.

Muscanell et al,[17] examined usage and utility of RG by employing an online survey approach to target scientist who has an active RG profile. They find that most researchers who have an RG profile did not utilize usually. In the following, members did not perceive many benefits from RG profile and RG use was not considered to members work satisfaction or informational benefits but was considered to productivity and stress.

Newman[18] has already investigated in the scientific collaboration network by analyzing papers in computer science and physics in a 5-year period from 1995 to 1999. He constructed a collaboration graph based on a data extracted from for databases. The idea of collaboration patterns by using data extracted from scientific networks is not a new topic at all. However, to our knowledge, no similar collaboration network of all various types of Iranian scientific institutions has previously been attempted.

Haiyan et al,[19] worked on the structure of scientific collaboration network in Scientometric at the level of individuals by utilizing bibliographic data of published papers from 1978 to 2004. His novelty is the construction of a combined Social Network Analysis, Co-occurrence analysis, cluster analysis and frequency analysis of words to expose the collaborative center of the network, major collaborative fields of the network and various

collaborative sub-networks and microstructure of the collaboration network.

Šubelj et al,[20] researched on convexity in the scientific collaboration network. They analyzed a particular dataset Slovenian researcher in computer science, physics, and other fields and identified convex skeletons in the collaboration network by eliminating weak links of them and provided frequency distributions of several data parameters of the skeleton and of remained graphs. Finally demonstrated that convex skeleton is a good abstraction of the original co-authorship network. To be clearer about convexity in a graph, it refers to an attribute of its subgraphs to comprise all shortest path between nodes of that particular subgraph. The scale values can be in [0, 1]. The biggest part of a graph that appears after the removal of the minimum number of an edge is called a convex skeleton which is fundamentally a tree of cliques. Perianes-Rodriguez et al,[21] suggested a new fractional counting method to build bibliometric network which already compared with a traditional full counting method. In examination of university co-authorship network and journal bibliometric network, each one of two methods reveal distinct results. The first mentioned approach is known superior in various causes.

Bihari et al,[22] point out a key problem in research communities which is every author complete reputation of citation count on a particular paper, but in many cases their contribution is not the same at all. To eliminate this specific issue, they decide to utilize Poisson distribution to spread the contribution reputation in multi-authored paper and for examine research impacts of an independent, the eigenvector centrality has been utilized.

## III. RESEARCH METHODOLOGY

The research steps are clarified in **Fig. 2** - Research Methodology Phases. In the first step, ResearchGate selected as the target scientific network to study. necessary data from objective scientific institutions gathered straight completely for the main goal of the study. After the data preprocessing phase, the database prepared to be investigated. By the help of network science, we created the collaboration network graph. At the final step, Louvain algorithm applied to the network to identify hidden communities and the structure transparently.
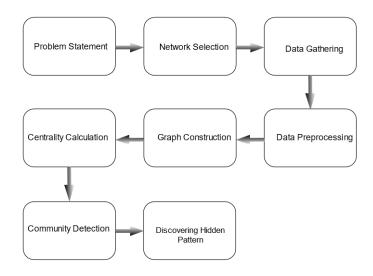


Fig. 2 - Research Methodology Phases

### A. Data Gathering

Data is the heart of every data science project. The data gathered and collected from the profile page of each institution which has an active profile on RG until. Our research target, covers all types of scientific institutions in Iran such as Public, Medical, Technical and fundamental institutions. The main reason that medical, technical and other types of institutions are considered in our study is covering various kind of theoretical communication among all major and academic fields at a comprehensive level. We don't aim to focus just on a particular research field. That's because different institutions type can be influential on results.

### B. Data Preprocessing

Data preprocessing is a fundamental method that initiate reshaping raw data into a readable style. After the data-gathering phase, the data cleaning task initiated to prepare them for graph construction from scratch. Many datasets are usually imperfect, incompatible with absence in certain behaviors, and is possibly to contain many missing values. Data preprocessing is a standard technique of eliminating such issues and preparing raw data for better processing.

## C. Graph Construction

Since our approach is towards social network analysis field, in our study, nodes refer to a particular university and edges defines scientific collaboration among them. The relation model is shown in **Fig. 3** - Sample Relations among Nodes in the Graph. The main attributes of the network are also mentioned in **TABLE 1.**



Fig. 3 - Sample Relations among Nodes in the Graph

TABLE 1 - ATTRIBUTES OF THE NETWORK

| Attribute | Value |
|---|---|
| Nodes | 183 |
| Edges | 320 |
| Average degree | 3.4973 |
| Average Path Length | 3.016 |
| Average Clustering Coefficient | 0.516 |
| Diameter | 7 |
| Central Degree | University of Tehran |
| Connectedness | True |
| Modularity | 0.535 |

## D. Centrality Measures Calculation

Centrality Metrics role is significant in network science. In our study, we calculate four main centralities among them are Degree, Betweenness, Closeness and PageRank Centrality.

- *Degree Centrality*

Degree centrality explained as the number of links that exist on a node (i.e., the number of connections that a node has). The degree centrality of a sample vertex $V$ for a given graph $G := (V, E)$ with $|V|$ vertices and $|E|$ edges is explained as described in **Equation (1)**

$$C_D(v) = \deg(v) \quad (1$$

(1)

- *Betweenness Centrality*

Betweenness Centrality is a measure of centrality in a network. It represents the degree that stands between each other nodes. a node with higher betweenness

centrality would have more control over the network, because more information will pass through that node[23] as described in **Equation 2**.

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st\,(v)}}{\sigma_{st}}$$

(2)

Where $\sigma_{st}$ is the total number of shortest paths from node $s$ to node $t$ and $\sigma_{st\,(v)}$ is the number of those paths that pass-through $v$.

- *Closeness Centrality*

In a connected graph, Closeness centrality is a process of discovering nodes that are capable to spread info very effectively within a graph. The calculation is based on its average farness to the rest of other nodes. vertices with a high closeness result have the shortest distances to all other nodes. The concept developed by Bavelas which described in **Equation 3**.

$$C(x) = \frac{1}{\sum_y d(y,x)}$$

(3)

The $d(y,x)$ is the distance between node $x$ and $y$.

- *PageRank Centrality*

PageRank is a link analysis algorithm, originally used to rank web pages. Google defines PageRank as "PageRank operates by counting the number and quality of links determine a rough evaluate of how significant the webpage is. Also based on Google Developer Team, they already updated the PageRank vector."[24] PageRank is defined in the original Google paper as described in **Equation 4**.

$$PR(A) = (1 - d) + d\left(\frac{PR(T1)}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)}\right)$$

(4)

we assume that a page ***A*** has pages ***T1*** to ***Tn*** which point to it (i.e., are citations). ***d*** is a damping factor which can be set between 0 and 1. It is usually set to 0.85. ***C(A)*** is defined as the number of links going out of page ***A***.

## E. Community Detection

The *Louvain* is a greedy optimization technique for community detection. It maximizes a modularity score for each community, where the modularity measures the quality of an allocation of nodes to communities by evaluating how much more densely connected the nodes

in a community are, compared to how connected they would be in a random network. It is one of the fastest modularity-based algorithms. It also demonstrates a hierarchy of communities at different scales, which can be functional for understanding the global functioning of a network. The inspiration for this method of community detection is the optimization of modularity as the algorithm advances. Modularity is a scale value in the middle of $[-1, 1]$. It measures the density of edges inside communities to edges outside communities. Optimizing this value theoretically results in the best possible clustering of the nodes of a sample network, however going through all possible iterations of the nodes into groups is infeasible, so heuristic algorithms are utilized. In the Louvain algorithm, first small groups are found by optimizing modularity locally on all nodes, then each small group is organized into one node and the first step is repeated. The algorithm process as described in **Fig. 4** is a heuristic method based on modularity optimization. The quality of the communities detected by this algorithm is better than previous methods as measured by modularity metric.
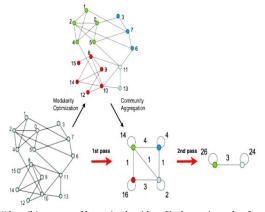


Fig. 4 – "describing steps of Louvain algorithm. Each pass is made of two steps: first where modularity is optimized by allowing only local changes of groups; second where the found groups are aggregated in order to create a new network of group"

As it is shown in the **Fig. 5**, public universities located in the capital city of Iran, Tehran has significant role in scientific production flow. By employing Louvain algorithm, nine communities identified on the network which discussed in **RESULT**.

## IV. **RESULT**

In the final section, we present the results of our study. with Average Degree 3.4973 it clears that each institute is at least connected to three other institution which can be conclude that our graph is not dense and we encounter with a sparse space. Density value with 0.019 can be as

proof for the claim. Based on degree centrality measure, the graph central node is University of Tehran.

NetworkX graph library[25] and Gephi visualization software[26] employed to visualize the network in an explicit way. An interesting point is that institutions with the highest RG score are also ranked in highest position in Islamic World Citation and Webometrics ranking. The reason that Shahid Beheshti University of Medical (SBMU) and Iran University of Medical Science (IUMS) are ranked among top medical institutions is their strong connection with Tehran University of Medical Science (TUMS) which indicates the PageRank Phenomenon. It means SBMU and IUMS both recommended by a Strong Node which is (TUMS) that leading them both in Medical Section.
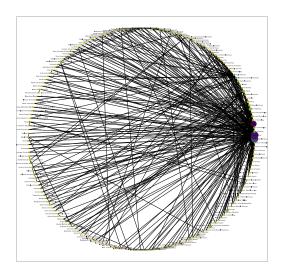


Fig. 5 Circular Visualization of the Network

Community structure is a common characteristic of social networks. nine communities identified by employing Louvain algorithm. The number of communities has not been fixed and obtained by the process of community detection. The method is similar to the earlier method[27] that connect communities whose merges produces the largest community in modularity. To evaluate the quality of detected communities, it has been tested by Girvan-Newman clustering algorithm. the process indicates Maximum found modularity is 0.48. Visualization of the network in Correlation between Centrality Measures and Communities demonstrated in **Figure 6**, **Figure 7** and **Figure 8**.
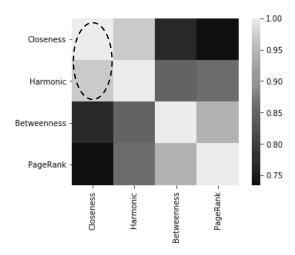
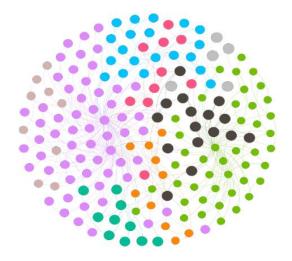Figure 6 – Heat Map of Correlation between Centrality Measures



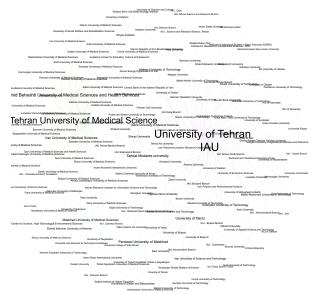Figure 7 - Modularity-based Community Detection of the Network



Figure 8 - Modularity-based Community Detection of the Network

To demonstrate geographical presence of identified communities we already initiated them to the country map to have a big picture of the network structure as it is shown in the **Figure 9**. Each color refers to a particular community.
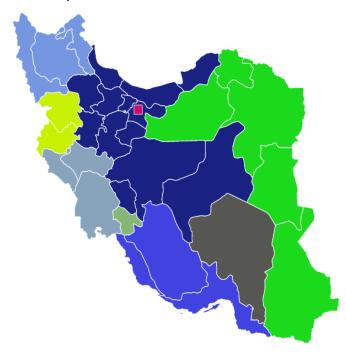


Figure 9 - Community Detection Visualization on Geographical Map

## V. CONCLUSION

In this study we utilized network science techniques to create, analyze and visualize collaboration network of Iranian scientific institutions for the first time. The contribution of our research is discovering structure of communities within the network. Our results demonstrated that geographic location closeness and ethnic attributes has important roles in academic interactions establishment. Besides, it shows that famous scientific centers in the capital city of Iran, Tehran, has strong influence on the production flow of scientific activities. The output of this study can be beneficial for Science and Technology Policy Organization who aims to design and initiates comprehensive strategies to gain more profit of conducted research and development in various knowledge-based fields. The most important benefit of community structure awareness of our network is identifying each particular scientific institute's research potential separately and recognize hot topic fields.

TABLE 2 - CENTRALITY MEASUREMENT CALCULATION RESULTS TO IDENTIFY KEY NODES ON THE NETWORK

| Nodes | Degree | Nodes | Page Rank | Nodes | Betweenness | Nodes | Closeness | Nodes | Harmonic |
|---|---|---|---|---|---|---|---|---|---|
| IAU | 0.36263 | IAU | 0.109805 | UT | 0.483103 | UT | 0.628019 | UT | 0.703846 |
| UT | 0.346153 | UT | 0.095966 | TUMS | 0.473596 | TUMS | 0.543933 | IAU | 0.678205 |
| TUMS | 0.280219 | TUMS | 0.084518 | IAU | 0.321358 | IAU | 0.539419 | TUMS | 0.644872 |
| SBMU | 0.104395 | SBMU | 0.029387 | FUM | 0.032998 | TMU | 0.460993 | TMU | 0.489744 |
| TMU | 0.065934 | IUMS | 0.018656 | Tabriz | 0.031145 | IUT | 0.431894 | IUT | 0.462821 |

## REFERENCES

[1] M. Basiri, A. Nilchi, and N. Ghassem-Aghaee, "A Framework for Sentiment Analysis in Persian," *Open Trans. Inf. Process.*, pp. 1–14, 2014, doi: 10.15764/OTIP.2014.03001.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.

[3] Liu J, "Introduction to School Network Analysis. Beijing: Social Science Documentation Publishing House," 2004.

[4] M. S.-Q. C. Min-Qiang, "Application of Social Network Analysis in Psychology," *J. Pyschology*, no. 19(5), pp. 755-764., 2011.

[5] R. R. Otte E, "Social network analysis: a powerful strategy, also for the information sciences," *J. Inf. Sci.*, vol. 28(6), pp. 443–455, 2002.

[6] A. I. Kretschmer H, "Visibility of collaboration on the Web. Scientometrics," vol. 61(3), pp. 405–426., 2004.

[7] and L. H. Abbasi, Alireza, Jorn Altmann, ""Identifying the Effects of Co- Authorship Networks on the Performance of Scholars: A Correlation and Regression Analysis of Performance Measures and Social Network Analysis Measures," *J. Informetr.*, vol. 5, no. 4, pp. 594–607, 2011.

[8] S. Manca, "An analysis of ResearchGate and Academia. edu as socio-technical systems for scholars' networked learning: a multilevel framework proposal," *Ital. J. Educ. Technol.*, vol. 25, no. 3, pp. 20–34, 2017, doi: 10.17471/2499-4324/985.

[9] M.-C. Yu, Y.-C. J. Wu, W. Alhalabi, H.-Y. Kao, and W.-H. Wu, "ResearchGate: An effective altmetric indicator for active researchers?," *Comput. Human Behav.*, vol. 55, pp. 1001–1006, Feb. 2016, doi: 10.1016/J.CHB.2015.11.007.

[10] R. Van Noorden, "Online collaboration: Scientists and the social network," *Nature*, vol. 512, no. 7513, pp. 126–129, Aug. 2014, doi: 10.1038/512126a.

[11] F. Naderbeigi, "Researchers' Scientific performance in ResearchGate: The Case of a Technology University," *Libr. Philos. Pract.*, 2018.

[12] W. Team, "ResearchGate," *Wikipedia*, 2018. [Online]. Available: https://en.wikipedia.org/wiki/ResearchGate.

[13] M. Thelwall and K. Kousha, "ResearchGate articles: Age, discipline, audience size, and impact," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 2, pp. 468–479, Feb. 2017, doi: 10.1002/asi.23675.

[14] B. Lepori, M. Thelwall, and B. H. Hoorani, "Which US and European Higher Education Institutions are visible in ResearchGate and what affects their RG score?," *J. Informetr.*, vol. 12, no. 3, pp. 806–818, Aug. 2018, doi: 10.1016/J.JOI.2018.07.001.

[15] R. Team, "Metric Score Calculation on Researchgate Social Network," 2019. [Online]. Available: https://researchgate.org.

[16] "Google Trend Service which Shows Exploration of Keyworkd Searched by the World," 2019. [Online]. Available: https://trends.google.com.

[17] S. U. Nicole Muscanell, "Social networking for scientists: An analysis on how and why academics use ResearchGate," *Online Inf. Rev.*, vol. 41, no. 5, pp. 744–759, 2017.

[18] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Natl. Acad. Sci.*, vol. 98, no. 2, pp. 404–409, Jan. 2001, doi: 10.1073/pnas.98.2.404.

[19] and Z. L. Hou, Haiyan, Hildrun Kretschmer, "The structure of scientific collaboration networks in Scientometrics.," *Scientometrics*, vol. 75, no. 2, pp. 189–202, 2007.

[20] L. Šubelj, D. Fiala, T. Ciglarič, and L. Kronegger, "Convexity in scientific collaboration networks," *J. Informetr.*, vol. 13, no. 1, pp. 10–31, Nov. 2018, doi: 10.1016/j.joi.2018.11.005.

[21] A. Perianes-Rodriguez, L. Waltman, and N. J. van Eck, "Constructing bibliometric networks: A comparison between full and fractional counting," *J. Informetr.*, vol. 10, no. 4, pp. 1178–1195, Jul. 2016.

[22] A. Bihari, S. Tripathi, and A. Deepak, "Collaboration Network Analysis Based on Normalized Citation Count and Eigenvector Centrality," *Int. J. Rough Sets Data Anal.*, vol. 6, no. 1, pp. 61–72, Dec. 2018, doi:

10.4018/ijrsda.2019010104.

[23]     K.-K. R. C. c Yan Hu a , b , Shanshan Wang a , Yizhi Ren b, "User influence analysis for Github developer social networks," *Expert Syst. With Appl. Elsevier*, vol. 108, pp. 108–118, 2018.

[24]     and C. D. M. Langville, Amy N., *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

[25]     N. Develpers, "Software for Complex Network," 2019.

[Online]. Available: https://networkx.github.io/.

[26]     M. J. Mathieu Bastian, Sebastien Heymann, "Gephi: An Open Source Software for Exploring and Manipulating Networks," in *Third International AAAI Conference on Weblogs and Social Media*, 2009.

[27]     C. Clauset, Aaron; Newman, M. E. J.; Moore, "Finding community structure in very large networks," *Phys. Rev.*, 2004.