1. Introduction and Background

The field of social network analysis has gained higher importance with the emergence of companies like Facebook and other social applications like Flicker, Instagram etc. As data is being generated at an unprecedented rate and volume (petabytes of data) from social media interactions majority of research work today is focused at developing efficient algorithms to identify community structures and patterns of entity interactions [11]. Extracting data from online communities and then constructing a research network for further study can help identify influential people of the community and important links of communication if an information is to be made viral. For instance, the Newman's network data comprises of data from three bibliographic databases – biology, physics and mathematics. If all the profiles of network scientists are extracted correctly and quality of data is ensured, then we will have a large collection of well structure data of network scientists [1].The profiles extracted from such networks can help us identify subject-specific experts, potential researchers and potential contributors for conferences, workshops etc. [1]. It can also tell us how active a research community is, how co-operative and collaborative the researchers are within their community, how likely the researchers are to co-author a paper together, collaboration patterns and many other characteristics of a research community [1].

In this paper, we explore and understand the **co-authorship network** of scientists working on Network theory as collected by Scientist Mark J. Newman. This dataset is a subset of the original dataset that comprises of data from three bibliographic databases namely Physics, Mathematics and Biology. The resulting collaboration visualization is a graph where nodes represent the scientists and edge represent the relation between two scientists. Mark Newman had assigned a weight to each of the edge. These edge weights represent the strength of relationship between two authors. A thicker edge means a stronger edge weight and a stronger relationship between two authors. We use node, authors, scientists and actors interchangeably in this paper.

2. Literature Review

Generally, such collaboration network of scientists is constructed by considering different aspects of social scientific relations such as papers co-authored together and similar research interests' areas [1]. According to Tomobe H. and his other co-researchers, the co-authorship relationship is the most important measure of collaboration among the research scientists [2].

Various studies have been conducted from the network extracted from different sources such as email network, search engines, purchase history etc. One such study conducted combining data from social networks and using collaborative filtering procedure is the *Referral Web*. The Referral web uses the co-occurrence of names in close proximity in any of the documents publicly available in the World Wide Web as evidence of the direct relationship [3]. It uses the data from personal homepages of scientists, list of co-authors in papers, citations etc. The resulting network from the Referral Web is the Ego-Centric network, in that an author or a scientist is at the center of obtained network [1]. The Referral web network can be used to find a subject/topic expert by making a simple query such as, "which friend of mine knows about Gephi?"

Similarly, another collaboration network named *Flink* was developed by Peter Mika, a Research scientist at Yahoo for the purpose of visualization of the social connectivity of the Semantic Web

Researchers. Flink uses the data gathered from Researchers' emails, homepage, affiliation, research interests, participation at a semantic web conferences etc. Similar to the Referral Web, Flink also employs the co-occurrence analysis of semantic web researchers to establish a relation between the two researchers. Flink has influenced many other semantic web researches such as European-On-To-Knowledge Project and SEKT project. Both of these research projects use the technological architecture of Flink.

Another example of a collaborative network is *Rexa* developed by Andrew McCallum and his team. Rexa extracts a de-duplicated cross-referenced database of not just papers (and references), but also people and grants, and so also publication venues and institutions [4].

3. Data Collection

This dataset is available for use at https://networkdata.ics.uci.edu/data.php?id=11. The dataset was originally compiled by Network Scientist Mark Newman and contains data from three bibliographic databases i.e Physics, Mathematics and Biology. However, the data used for this research analysis was compiled from the bibliographies of two review articles on networks, M.E.J. Newman, SIAM Review 45, 167-256 (2003) and S. Boccaletti et al., Physics Reports 424, 175-308 (2006), with a few additional references added by hand [6]. The data was compiled using a parser written in C programming language. The parser used was developed by the scientist Mark Newman himself.

4. Network Analysis and Research Questions

In this research paper, we are simply doing an analysis of a co-authorship network as prepared by Dr. Newman. We are especially interested in identifying the connectivity amongst the network scientists as well as identifying the characteristics of scientific communities as identified in the Gephi analysis. Using the network analysis metrics such as betweenness centrality, Eigen vector centrality, degree centrality etc. we focus on answering the following research questions:

- Who is the most influential scientist in the entire network as well as the most influential scientist from each of the communities?
- Which of the scientists frequently collaborate with each other?
- Do scientists from cross background co-author a paper together?
- How many scientists belong to the largest connected component of the network and define community characteristics as detected from Gephi analysis.
- Tie strength identified by the number of papers they have co-authored together.

5. Results

After loading the file in Gephi, I applied the "Force Atlas" and "Yifan Hu Proportional" algorithm before applying any filters to the network. Following is the statistical observation made before applying any filter:

A. This is not a single connected network and scientists from cross communities do co-author a paper together.

Next to abstract a more manageable network, I chose to filter the network using the "Giant Component" filter, filtering out nodes less than degree 1. Then I ran the modularity statistics with

the resolution value of 4.0 which yielded 9 communities. The nodes are sized by betweenness centrality and are color coded by modularity class. The results thus obtained are as follows:

B. The filtered network consists of 379 nodes which represents only 23.85 percentage of the entire network. The extracted network, comprises of 9 co-authorship communities.

C. Following table shows the top 5 most influential scientists per betweenness centrality, closeness centrality, and degree centrality measures.

Degree Centrality	Closeness Centrality	Betweenness Centrality	EigenVector Centrality
Barabasi, A	Dickman, R	Newman, M	Barabasi, R
Jeong, H	Rothman, D	Pastorsatorras, R	Kurths, S
Newman, M	Kalapala, V	CaidareLli, G	Jeong,H
Kurths, S	Sanwalani, V	Moreno, Y	Bocalletti, S
Bocaletti, S	Ye, N	Stauffer, D	Newman, M

Table 1.0 Top 5 Influential authors per centrality measures

D. Communities identified. Scientists from the green and red community are most collaborative. Frequent collaborators from each community identified (Refer Table 3.0)

E. Collaboration patterns identified and discussed.

6. Discussion

The results obtained above are explained as follows:

A. This is not a single connected network. This is not a single connected network scientists from cross communities do co-author a paper together.

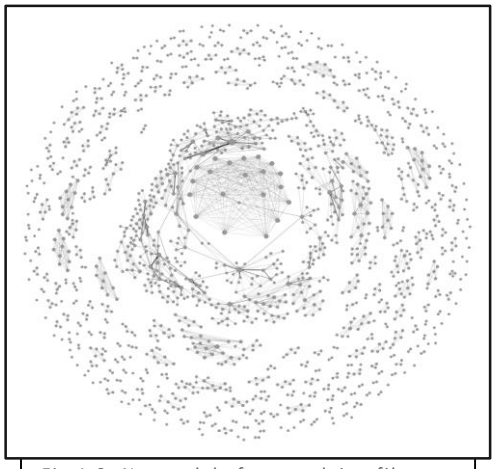

Fig 1.0: Network before applying filters

Fig 1.0 displays the output obtained from running the layout algorithms. As we can see, the NetScience network consists of a numerous clusters, closed network cliques and few isolates. Hence, this is not a single connected network. As a result, we can conclude that the network consists of number of sub-networks, which is evident by the number of cliques and clusteras shown by Fig 1.0. Also at the center of the graph, we see that a few nodes from cluster-1 are linked to a distant node of a different cluster-2. This indicates the collaboration of network scientists from two different clusters or different subject specialization. For example, Scientists Mark Newman (sky blue) and Stauffer (orange) have together co-authored a paper *"Dynamics of a simple evolutionary process"*.

B. The filtered network consists of 379 nodes which represents only 23.85 percentage of the entire network. The extracted network as seen, comprises of 9 co-authorship communities.

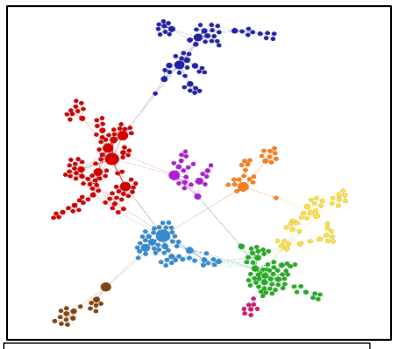

Fig 2.0: Communities in Network

The extracted network as seen, comprises of 9 co-authorship communities that are color coded by blue, red, purple, orange, yellow, green, pink, sky-blue and brown. Fig 2.0 shows the communities identified in the given network.

The red community consists of mostly Spanish and French scientists whose works are mostly focused in socio-physics and epidemics spreading in scale free networks. The scientists belonging to this community shares the second most collaboration strength (after green

The purple community consists of computer scientists who are mostly based on London and whose work is mostly focused on scale free networks such as World Wide Web, Internet, and Emails etc. Scientists who share stronger collaboration tie from this community are Caldarelli. G and Capocci.A. They have frequently collaborated paper regarding evolution and structure of the internet.

The sky blue comprises of computer scientists cum mathematicians whose work is more focused in developing fast algorithms for network epidemiology. It consists of scientists like Strogatz and Watz who together formulated the "Small world Network Model". Scientists Mark Newman and Watts.D share greater stronger collaboration tie in this community. Together they have co-authored several papers relating to structure and dynamics of a network.

The blue community consists of physicists cum mathematicians whose work is focused on synchronization of complex networks, and introducing selection criteria for the wiring topology that enhances synchronized dynamics in weighted networks. In this community, scientists Kurths.J, Rosenblum, and Pikovsky.A share strongest collaboration strength.

The orange represents scientists who are physicists and whose works are noted for application of statistical physics and computational physics in the areas of econophysics and sociophysics. This is a weakly connected community and members sharing stronger tie have co-authored only 4-5 papers together.

The yellow comprises of scientists who are both mathematician and computer scientists. Their works are mostly focused on devising centrality measures in spatial networks of urban streets. This is also a weakly connected community. However, scientists Marchiori, V. Latora and Crucitti have co-authored several papers together. One such paper is ""Error and attack tolerance of complex networks".

The green consists of scientists who share the background of biology as well as physics. The works of these scientists are mostly focused in metabolic networks and measuring preferential attachment in evolving networks. The scientists belong to this community share the strongest collaboration relationship. This is supported by the edge width as shown in the Net Viz. graph and from the fact that the scientist belonging to this community have most frequently collaborated paper with one another than in any other community. Also, scientists Jeong H, Albert R and Barabase form a closed triad. This indicates that they share strongest tie (in this community) which is evident from the fact that they have published several papers together.

The pink community includes scientists who have worked on percolation critical exponents in scale free networks. This is a weakly connected community. However, Halvin.S, Cohen.R and Benavraham.D have frequently co-authored paper relating to internet breakdowns. One such paper is "Resilience of the internet to random breakdowns". The last

brown community consists of scientists who are mathematicians as well as computer scientists whose work is focused on inferring web communities from link topology and mining the link structure from World Wide Web. The scientists from this community share the weakest collaboration strength. Scientists from this community who have frequently co-authored paper together are C.L Giles and Lawrence.S.

C. Most influential scientists /Hub leaders

The most important people in the network can be identified using the centrality measures. The nodes in our graph are characterized by betweenness centrality which measures the node's role as a connecter/bridge between other nodes. It tells us how critical the actor is to the network in its functioning as a bridging point between other actors. Here, the size of each node correspond to its critical role to serve as a connector in the network. Table 1 shows top 5 individuals from the above network who scored the highest with respect to betweenness centrality, closeness centrality, Eigenvector Centrality, and degree centrality measures. From table 1, we can note that the eigenvector and degree centrality measures list almost the same authors. This means that in the given network, authors with most node connections are also the authors connected to other authors with higher node connections.

Also, hub leaders (based on betweenness centrality) in each of the communities identified are shown in the table 2.0. This means that Bocaletti, S and other listed scientists act as a central gatekeeper in their respective scientists. It also means that information flow is most efficient through these nodes i.e. these nodes connect with most other nodes via the shortest path in their respective networks. However, when the whole network is taken into consideration, scientist Newman, M is the author with the highest betweenness centrality which means he is at the center of information flow of in the entire network and disseminates information most efficiently across the network.

Communities	Hub Leaders
Blue	Bocaletti, S
Red	Pastorrastorras, R
Sky Blue	Newman, M
Purple	Caldarelli, G
Orange	Stauffer, D
Brown	Klienberg, G
Green	Jeong, H
Pink	Crucitti, P
Yellow	Halvin, S

Table 2.0 Network Leaders

D. Collaboration / Relationship Strength and Frequent Collaborators

The collaboration tie strength in the given network can be identified using the edge width measure. Since the given network came with the weight assigned, no calculation was performed to obtain the edge weight. Higher the edge width, higher the collaboration strength. From Fig 3.0 (Refer to Appendix), it is evident that the authors in the green share a greater collaboration strength (followed by Red) than the other communities. This means that the authors in the green community are more collaborative. This is also supported by the fact that the nodes in green community are tightly clustered in cliques and traids. For example, authors Jeong.H, Barabasi. A, Albert.R (green nodes) form a close triad, which means that they form a close trusted loop of collaboration network of their own. Upon a quick research, I found that they have been frequently publishing papers together (http://barabasilab.com/pubs.php). Additionally, most of the green nodes have the highest clustering coefficient (of 0.5) in the entire network followed by the red community. Scientists with stronger collaboration from each of the communities have been discussed in section B. The table 3.0 shows frequent collaborators from each communities.

Communities	Frequently Collaborating
Blue	Kurths.J, Rosenblum, and Pikovsky.A
Red	Pastorsatorras and Vespiagini
Sky Blue	Mark Newman and Watts.D
Purple	Caldarelli. G and Capocci.A
Orange	Stauffer. D and Aharony. A
Brown	C.L Giles and Lawrence.S.
Green	Jeong H, Albert R and Barabase
Pink	Halvin.S, Cohen.R and Benavraham.D
Yellow	Marchiori, V. Latora and Crucitti

Table 3.0 Frequent Collaborators

E. Collaboration Patterns

An interesting observation in this network was that the scientists sharing the same nationality or heritage shared greater collaboration tie. For instance, scientists such as Pastorsatorras, Vespiagini,

Vazquez, and Moreno (from Red community) are Spanish (by heritage). They have collaborated many papers together which means they share relatively greater bond than with any others
Likewise, scientists such as Balthemy and Alian Barrat are French and shared greater relationship strength.
Similarly, scientists from same institutions or institutions with similar ranking also seem to have co-authored many papers together. For instance, the Alma matter of scientists Duncan J Watts and Steven Strogatz (both sky blue) are Cornell University and Harvard University respectively. Together they have co-authored several papers together.

The final observed pattern was that the scientists with same academic/ research background co-authored more papers together such as mathematicians co-authored more papers with mathematicians. For example, scientists Marichiori and V. Latora share the mathematics background and have co-authored many papers together than with any other scientists.

7. Conclusions

In this paper, we analyzed the collaboration network of scientists working on network theory.
We identified communities in the given network and its respective leaders who act as central network leader to disseminate information most efficiently across their respective network. We also identified which of the scientists from each community share stronger collaboration strength based on the edge weight and papers they have co-authored together and collaboration patterns of the network scientists.

References:

1. TaslimArif, Rashid Ali, M. Asger. "Scientific Co-authorship Social Networks: A Case Study of Computer Science Scenario in India" Available at: https://www.academia.edu/5287441/Scientific_Co-authorship_Social_Networks_A_Case_Study_of_Computer_Science_Scenario_in_India
2. Tomobe, H., Matsuo, Y. and Hasida, K. "Social Network Extraction of Conference Participants." Available at: http://ymatsuo.com/papers/www2003/p92-tomobe.html
3. Henry Kautz, Bart Selman and Mehul Shah. "Referral Web: Combining Social Networks and Collaborative Filtering". Available at: http://www.cs.cornell.edu/selman/papers/pdf/97.cacm.refweb.pdf
4. Andrew McCallum. "Extraction, Integration and Mining of Bibliographic Data" Available at: http://people.cs.umass.edu/~mccallum/research.html
5. Andrew Papachristos. Available at: http://nnscommunities.org/our-work/innovation/social-network-analysis
6. Available at: http://vlado.fmf.uni-lj.si/pub/networks/data/collab/netscience.htm
7. Available at: http://webwhompers.com/graph-theory.html
8. Available at: https://en.wikipedia.org/wiki/Clustering_coefficient
9. Available at: http://nnscommunities.org/our-work/innovation/social-network-analysis
10. Available at: http://www.researchgate.net/profile/Scott_Decker2/publication/248529612_Gangs_gang_homicides_and_gang_loyalty/links/547cb25e0cf285ad5b088462.pdf
11. Available at: https://www3.nd.edu/~dial/papers/PRL10.pdf

Appendix:

1. Fig 3.0 –Final Vizualization Graph

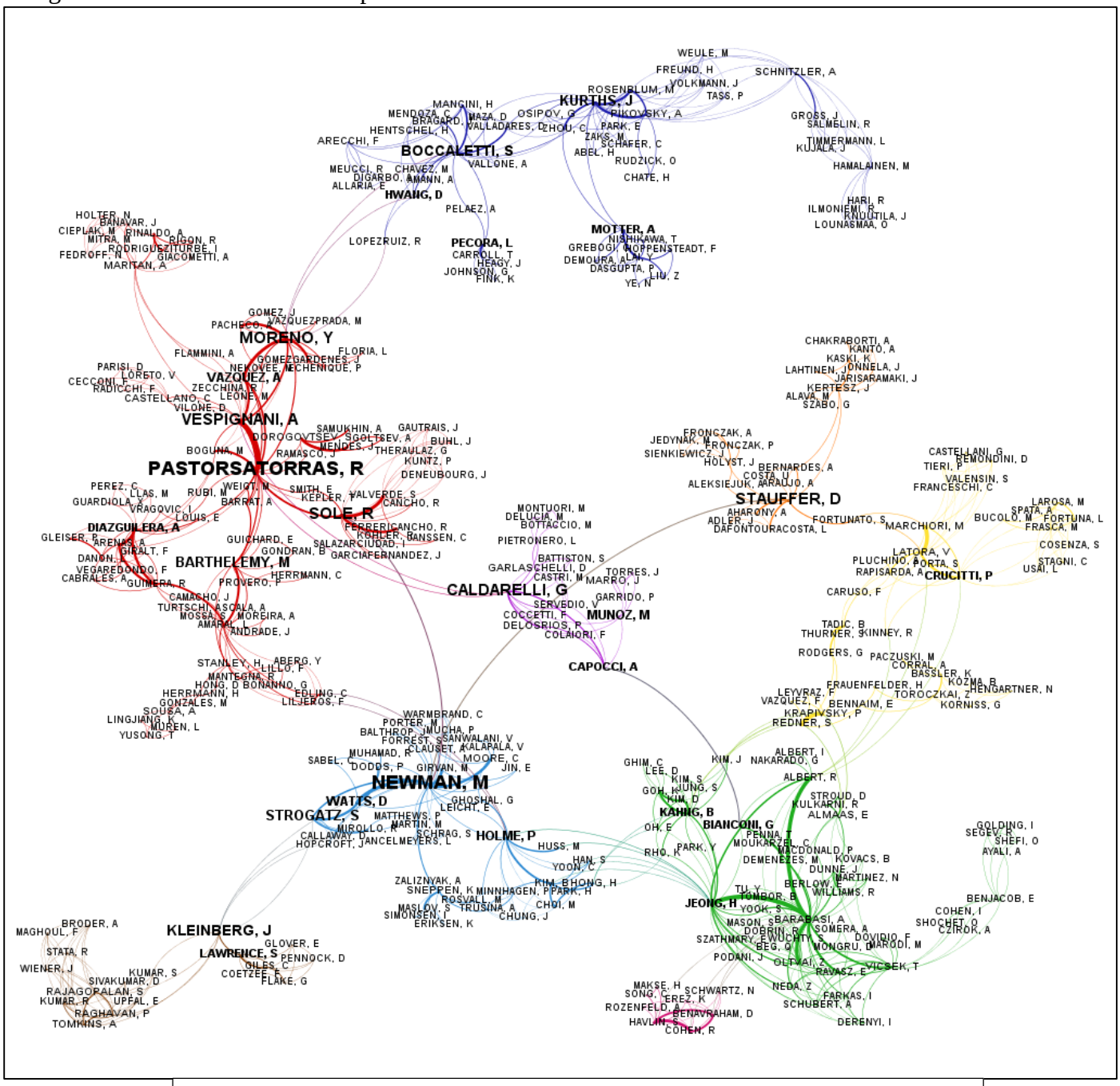

Fig 4.0 Network Vizualization of the most connected network of the Netscience