# Abstract:

As the internet becomes more accessible all around the world, the different networks around the internet such as p2p, email etc. are also growing in size. Therefore, analyzing large networks for the purpose of knowledge discovery to be able to learn different governing patterns that dictate these large networks is quite necessary. With that goal in mind, in this study we studied a plethora of large networks that varied from more active human driven networks such as email communications, social networks to more passive networks such as p2p. We employed unsupervised learning techniques to visualize the underlying structures of these networks by representing them as 2 dimensional manifolds. We hypothesize that network structures are more affected by the behavior of the nodes/users of the networks instead of having a predefined shape. It is our assumption that same type networks (p2p network) will have a similar structure within an error interval, while networks of different types (social network vs p2p network) will be distinctly different in terms of network structure. Based on our findings, we also propose a hierarchical categorization of networks in a broader sense, such as communication networks, have hierarchies within their structures, where we can observe the structures changing in a certain pattern or trend. This kind of behavior based unsupervised knowledge discovery methods can help us find further meaningful patterns in large random human networks which than can be used to identify and generalize different networks such as migration networks, criminal networks or terrorist networks.

Keywords: Unsupervised Learning, Network Visualization, Knowledge Discovery

# Introduction:

In recent years, network analysis techniques have evolved (Chen et al., 2019; Williamson & Tec, 2019) in proportion to the rapid growth of real world networks. Much research has been done on networks such as social networks (Bródka et al., 2011; Culotta & Cutler, 2016), communication networks (Mccallum, 2007), citation networks etc. As time progresses, these networks become bigger and more complex, consequently holding vast amounts of interesting information which can be used for various purposes. Network analysis techniques include, but are not limited to, using random graph models to capture or derive the properties of real world networks (Williamson & Tec, 2019), subgraph isomorphism (Cordella et al., 2004), graph simulation etc. A significant portion of the research conducted on networks has been on centralities (Newman, 2010) to deduce, “which” are the more important nodes in the graph i.e.. central nodes or nodes that propagate most information, and how traffic flows through them. In turn, calculated centralities of various network graphs have been used in many other research fields as well. For example, analyzing social networks, such as tweet classification (Hussain & Islam, 2016), detecting political discussion practices (Miller et al., 2015) and many other such applications (Cohn et al., 2019; Rossman et al., 2010; Yang & Liu, 2008); has involved the generation of multiple centralities. Centralities have also been used in analyzing road traffic networks, such as finding road network patterns (Zhang et al., 2011), tourism management (Lee et al., 2013) etc.. Another important application of centrality measurements is in analyzing biological networks, as can be seen in (Bell et al., 1999; Joyce et al., 2010; Narayanan, 2005; Park & Kim, 2009).

Centralities such as Degree, Closeness, Betweenness, Crossclique, Pagerank, etc. have all been proposed and developed (Crucitti et al., 2006) over the years to answer “Which node is the most influential?” in various different applications (Landherr et al., 2010). “Significance” or “importance” of nodes varies from one context to another. For example, in some cases, it is essential to identify which node(s) propagate more information locally or globally in a graph(Newman, 2010), whereas in other cases, detecting the central node(s) might be of more value (Chiu et al., 2016). Centralities can, therefore, be seen as features/characteristics of a graph; thus, it is possible to discriminate between networks based on their centrality values (Wang & Krim, 2012).

In this research, centrality measures have been used to visually analyze the structure of real world networks. The hypothesis is that, the structure of a network is not random, but is rather dictated or largely affected by the behavior of the nodes in the network. As different centrality measures express different characteristics of the nodes in the network, the information given by the combination of centrality metrics should give similar visual representations for similar networks (such as email networks of two institutions), and different representations for different networks (such an email network and a product co-purchase network). Using dimensionality reduction methods, analysis of the higher dimensional representations of the networks is possible.

# Literature review/relevant works/ background study

*// needs to be redone with more visual analysis papers*

# Centralities:

This study largely revolves around centralities as they are useful metrics to define the role of a node in a network. They can be used to explain the node’s importance both globally, or in their local communities/cliques. Centrality measures essentially describe the connectivity of a node within a network, based on factors such as number of connections, geodesic distances with other nodes, placement in between different cliques etc. (Borgatti, 2005; Brandes & Pich, 2007; Dwyer et al., 2003). The centralities used in this research were chosen based on whether they expressed the global or local presence of the node, resource intensiveness and computability. The selected centralities are: degree, eigenvector, Pagerank, authority, Hubscore, Betweenness, closeness, eccentricity, Density of Maximum Neighborhood Component (DMNC), lobby index, leverage, local bridging.

* Degree: Total number of edges going into and out of a node. Denoted by: *//FORMULA NEEDS CHECKING*
* Eigenvector: Ranking of nodes based on how many well-connected nodes they are connected to. Denoted by: //*DEFINITION NEEDS CHECKING*

Where CIV is an eigenvector and λ is an eigenvalue.

* Pagerank: Ranking of nodes based on the frequency of their appearance in a random traversal of the network. Denoted by: //*DEFINITION NEEDS CHECKING*

Where ti , i=1,..n are the nodes that have a directed edge to node v, C(v) is the number of edges going out from v, and d is the damping factor.

* Authority: Ranking of nodes based on the principal eigenvector of t(A)\*A, where A is the adjacency matrix of the graph.*// DEFINITION COPIED, NO FORMULA FOUND*
* Hubscore: Ranking of nodes based on the principal eigenvector of A\*t(A), where A is the adjacency matrix of the graph. //*DEFINITION COPIED, NO FORMULA FOUND*
* Betweenness: Ratio to define how many shortest paths pass through a node amongst total shortest paths between all the node pairs of the network. Denoted by: //*DEFINITION AND FORMULA NEEDS CHECKING*

is the number of shortest paths between node s and t and (v) is the number of shortest paths passing on a node v out

* Closeness: Reciprocal of total distance from node v to all other nodes in the network. Denoted by:

Where dist(v,t) denotes the distance between node v and t.

* Eccentricity: Maximum distance between a node v and all other nodes. Denoted by:
* DMNC: Density of the Maximum Neighborhood Component of a node. Denoted by: //*DEFINITION NEEDS CHECKING*

Where

* Lobby index: Defines the biggest integer k such that the node has at least k neighbors which have degree of at least k. //*DEFINITION NEEDS CHECKING, NO FORMULA FOUND*
* Leverage: Relationship between the degree of a node and the degree of all of its neighbors, averaged over the total number of neighbors. Denoted by: //*DEFINITION NEEDS CHECKING*
* Local bridging: <https://www.centiserver.org/centrality/Local_Bridging_Centrality/> *// COPY PASTING THE LINK CAUSE I CANNOT WRITE THIS*

Table Node centralities VS Community/Clique Centralities

|  |  |  |
| --- | --- | --- |
| Node Centrality | Community Centrality | Both Node and Community |
| Degree | DMNC | Leverage |
| Eigenvector | Lobby index | Eccentricity |
| Pagerank | Local bridging |  |
| Authority |  |  |
| Hubscore |  |  |
| Betweenness |  |  |
| Closeness |  |  |

# Datasets

The datasets used in this study belong to the following categories: Social Networks, Citation Networks, Collaboration Networks, Product Co-Purchase Networks, Communication networks, Internet P2P networks, Ground Truth Networks and Signed Networks. The description of each network are given below.

## Social Network

### Ego Networks (Facebook and Twitter)

The Ego networks are essentially networks centered around one particular node. Such networks are formed by taking one node, and finding all the vertices that are directly connected to it while also finding the connections between those vertices. Two ego networks have been used in this research, namely the “Ego-Facebook” dataset which contains 4039 nodes and 88233 edges as it combines 10 ego networks, and “Ego-Twitter” dataset which contains 81306 nodes and 2420765 edges. The Facebook ego networks were collected by Facebook users willingly using an app and twitter information was collected using web crawlers.

### Page to Page Networks (Facebook)

Another type of network is formed between pages instead of account-holders. The “Gemsec-Facebook” datasets consist of networks of 8 different categories- Artists (50515 nodes and 819306 edges), Athletes (13866 nodes and 86858 edges), Company (14113 nodes and 52310 edges), Government (7057 nodes and 89455 edges), New sites (27917 nodes and 206259 edges), Politicians (5908 nodes and 41729 edges), Public figures (11565 nodes and 67114 edges), and Tv shows (3892 nodes and 17262 edges). Each individual dataset consists of verified Facebook pages of that category, and edges exist between pages that like each other.

### Twitch

Twitch is a website on which people can livestream, and it is used mostly by gamers. The “Musae-twitch” datasets consists of 6 networks – DE (9498 nodes and 153138 edges), ENGB (7126 nodes and 35324 edges), ES (4648 nodes and 59382 edges), FR (6549 nodes and 112666 edges), PTBR (1912 nodes and 31299 edges), RU (4385 nodes and 37304 edges) - divided by the languages used by the streamers. The edges represent friendships between the streamers. The data was collected on May 2018.

### Last.fm

Last.fm is a website that has been providing music services since 2003. The “Feather-lastfm-social” dataset contains a social network of LastFM users which was collected in March 2020. The nodes represent users residing in Asian countries and edges represent mutual follow relationships. It has 7624 nodes and 27806 edges.

## Communication Networks

Communication networks can be constructed by collecting email data from institutes. Such networks are the “Email-EuAll” and “Email-enron” networks. “Email-EuAll” has 265214 nodes and 420045 edges, and was collected from a European institute from October 2003 to May 2005. The nodes represent email address and a directed edge signifies that source node sent at least one email to target node. “Email-enron” has 36692 nodes and 367662 edges and was collected, and made public, from the Enron Corporation when it was being investigated. This network is undirected and contains an edge between nodes if any email was exchanged between them.

## Collaboration Networks

The Arxiv collaboration networks were collected from the e-print arXiv websites. The edges represent collaborations between authors, so if one author collaborated with another an undirected edge exists between them. The data consists of papers published between January 1993 to April 2003. The “Ca-AstroPH” dataset has 18772 nodes and 396160 edges, and represents collaborations in papers of the Astro Physics category. The “Ca-HepPH” dataset has 12008 nodes and 237010 edges, and represents collaborations in papers of the High Energy Physics - Phenomenology category.

## Citation Networks

The Arxiv citation networks were also collected from the e-print arXiv database. These datasets represent which papers cite each other within this database, and if one paper cites another, a directed edge is drawn from the former to the latter. Information regarding cited papers that do no exist within the database is not present in the networks. The data is collected from papers published between January 1993 and April 2003. The “cit-HepPH” dataset contains 34546 nodes and 421577 edges and consists of papers submitted to the High Energy Physics- Phenomenology category. The “cit-HepTH” dataset contains 27770 nodes and 352807 edges and consists of papers submitted to the High Energy Physics- Theory category.

## Co-purchase Networks

The Amazon product co-purchasing networks were collected by crawling the website and collecting information from the “Customers Who Bought This Item Also Bought” feature. A directed edge exists from one product to another if the former is frequently bought companying the latter. The “Amazon0302” dataset was collected on 2nd March 2003 and has 262111 nodes and 1234876 edges. The “Amazon0601” dataset was collected on 1st June 2003 and has 403394 nodes and 3387388 edges.

## Internet P2P Networks

The hosts in a peep-to-peer sharing system also form networks. Such networks were constructed from the Gnutella file sharing network by taking 9 snapshots across a few days in August 2002. The ones used in this research are: “P2p-Gnutella04” (10876 nodes and 39994 edges) which was collected on 4th August, “p2p-Gnutella30” (36682 nodes and 88328 edges) which was collected on 30th August, “p2p-Gnutella31” (62586 nodes and 147892 edges) which was taken on 31st August.

The following table shows the details of each dataset, sorted in order of increasing nodes:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset name | Nodes | Edges | Average Degree | Graph Density | Graph Transitivity |
| musae-Twitch-PTBR | 1912 | 31299 | 32.73953975 | 0.008561595 | 0.130980962 |
| Gemsec-Facebook-Tv shows | 3892 | 17262 | 8.870503597 | 0.001139582 | 0.590643566 |
| Ego-Facebook | 4039 | 88233 | 43.69051745 | 0.005408581 | 0.519189302 |
| musae-Twitch-RU | 4385 | 37304 | 17.01436716 | 0.001940065 | 0.048648019 |
| musae-Twitch-ES | 4648 | 59382 | 25.55163511 | 0.00274867 | 0.084234875 |
| Gemsec-Facebook-Politician | 5908 | 41729 | 14.12626947 | 0.00119552 | 0.301073736 |
| musae-Twitch-FR | 6549 | 112666 | 34.40708505 | 0.002626896 | 0.054128273 |
| Gemsec-Facebook-Government | 7057 | 89455 | 25.35213263 | 0.00179624 | 0.223768822 |
| musae-Twitch-ENGB | 7126 | 35324 | 9.914117317 | 0.00069563 | 0.042433249 |
| Feather-lastfm | 7624 | 27806 | 7.294333683 | 0.00047838 | 0.178622548 |
| musae-Twitch-DE | 9498 | 153138 | 32.24636766 | 0.001697535 | 0.046470889 |
| p2p-Gnutella04 | 10876 | 39994 | 7.354542111 | 0.000338109 | 0.005402029 |
| Gemsec-Facebook-Public figure | 11565 | 67114 | 11.60639862 | 0.00050179 | 0.166564488 |
| CA-HepPh | 12008 | 237010 | 39.47534977 | 0.00164371 | 0.659477009 |
| Gemsec-Facebook-Athletes edges | 13866 | 86858 | 12.52819847 | 0.00045176 | 0.129227029 |
| Gemsec-Facebook-Company | 14113 | 52310 | 7.413023454 | 0.000262631 | 0.153198695 |
| CA-AstroPh | 18772 | 396160 | 42.20754315 | 0.001124215 | 0.318001581 |
| Musae-Facebook | 22470 | 171002 | 15.22047174 | 0.000338684 | 0.232321437 |
| Cit-HepTh | 27770 | 352807 | 25.40921858 | 0.000457494 | 0.119569073 |
| Gemsec-Facebook-New sites | 27917 | 206259 | 14.77658774 | 0.000264652 | 0.113984782 |
| Cit-HepPh | 34546 | 421577 | 24.4067041 | 0.000353249 | 0.145656674 |
| p2p-Gnutella30 | 36682 | 88328 | 4.815876997 | 6.56E-05 | 0.005163701 |
| Email-Enron | 36692 | 367662 | 20.04044478 | 0.00027309 | 0.085310796 |
| Gemsec-Facebook-Artist edges | 50515 | 819306 | 32.43812729 | 0.000321074 | 0.05349711 |
| p2p-Gnutella31 | 62586 | 147892 | 4.726040968 | 3.78E-05 | 0.003872019 |
| Ego-Twitter | 81306 | 2420765 | 59.5470199 | 0.000366191 | 0.170572209 |
| Amazon0302 | 262111 | 1234876 | 9.422542358 | 1.80E-05 | 0.236082716 |
| Email-EUAll | 265214 | 420045 | 3.167592963 | 5.97E-06 | 0.004106431 |
| Amazon0312 | 400727 | 3200440 | 15.97316877 | 1.99E-05 | 0.16048945 |
| Amazon0601 | 403394 | 3387388 | 16.79443918 | 2.08E-05 | 0.165622114 |

# Ablation Study

An ablation study was conducted to see the effects of the different centralities on the structure, and hence the visualizations, of the networks. The ablation models were designed based on what the centralities express and which category (node centrality or community centrality) they belong to. An ablation matrix is given below:



**Model 1**

In model 1, all centralities were kept to see the impact of both community and central node driven centralities.

**Model 2**

In model 2, we removed only degree. Degree centrality is the most influential centrality in the sense that it is used to calculate many other centralities. Therefore, we wanted to see how removing affects the model

**Model 3**

In model 3, centralities that calculated community were removed. The goal was to see how well the model performs when only centralities that are calculated globally (WITHOUT taking neighborhood/ community into account) are used.

**Model 4**

In model 4, centralities that used global information for central node calculation were removed. The idea was to see how the model performed when global centralities were taken out of the equation and centralities that depend on neighborhood/ community were used.

**Model 5**

In model 5, eigenvector, pagerank, hubscore and authority centrality were removed. These centralities calculate importance of a node in similar manner. Therefore, the goal was to see how the impact of “importance” amongst nodes differs from the impact of “distance” between nodes.

**Model 6**

In model 6, betweenness, eccentricity, closeness and information centrality were removed to see how the impact of “distance” between nodes differs from the impact of “importance” amongst nodes.

**Model 7**

Model 7 is mix of both central nodes based centralities and community oriented centralities. In this model leverage, lobby, dmnc, local bridging, information centrality, eigenvector and closeness were not used. The goal was to see how a model with more global information about “central” nodes with some community information behaves.

**Model 8**

This is same as model 7. However, here degree, hubscore, authority, betweenness, information, eccentricity and leverage were not used. The purpose was to see the results of equal proportions of community and central node information.

**Model 9**

Same as prior 2 models. In model 9, closeness, eigenvector, authority, hubscore, betweenness, leverage and information centrality were unused. The goal was to see how more information about community and some information about central nodes affects the result.

# Results and Analysis

# Discussion

# Conclusion

# References:

Bell, D. C., Atkinson, J. S., & Carlson, J. W. (1999). Centrality measures for disease transmission networks. *Social Networks*, *21*(1), 1–21. https://doi.org/10.1016/S0378-8733(98)00010-0

Bródka, P., Skibicki, K., Kazienko, P., & Musiał, K. (2011). A degree centrality in multi-layered social network. *Proceedings of the 2011 International Conference on Computational Aspects of Social Networks, CASoN’11*, 237–242. https://doi.org/10.1109/CASON.2011.6085951

Chen, H., Yin, H., Chen, T., Viet, Q., Nguyen, H., & Xue, W. P. (2019). Exploiting Centrality Information with Graph Convolutions for Network Representation Learning. *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 590–601. https://doi.org/10.1109/ICDE.2019.00059

Chiu, C. C., Balkunid, P., & Weinberg, F. (2016). When managers become leaders : The role of manager network centralities , social power , and followers ’ perception of leadership. *The Leadership Quarterly*. https://doi.org/10.1016/j.leaqua.2016.05.004

Cohn, A. M., Amato, M. S., Zhao, K., Wang, X., Cha, S., Pearson, J. L., Papandonatos, G. D., & Graham, A. L. (2019). Discussions of Alcohol Use in an Online Social Network for Smoking Cessation: Analysis of Topics, Sentiment, and Social Network Centrality. *Alcoholism: Clinical and Experimental Research*, *43*(1), 108–114. https://doi.org/10.1111/acer.13906

Cordella, L. P., Foggia, P., Sansone, C., & Vento, M. (2004). A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(10), 1367–1372. https://doi.org/10.1109/TPAMI.2004.75

Crucitti, P., Latora, V., & Porta, S. (2006). Centrality in networks of urban streets. *Chaos*, *16*(1). https://doi.org/10.1063/1.2150162

Culotta, A., & Cutler, J. (2016). *Predicting Twitter User Demographics using Distant Supervision from Website Traffic Data*. *55*, 389–408.

Hussain, J., & Islam, M. A. (2016). Evaluation of graph centrality measures for tweet classification. *2016 International Conference on Computing, Electronic and Electrical Engineering, ICE Cube 2016 - Proceedings*, 126–137. https://doi.org/10.1109/ICECUBE.2016.7495209

Joyce, K. E., Laurienti, P. J., Burdette, J. H., & Hayasaka, S. (2010). A new measure of centrality for brain networks. *PLoS ONE*, *5*(8). https://doi.org/10.1371/journal.pone.0012200

Landherr, A., Friedl, B., & Heidemann, J. (2010). A Critical Review of Centrality Measures in Social Networks. *Business & Information Systems Engineering*, *2*(6), 371–385. https://doi.org/10.1007/s12599-010-0127-3

Lee, S. H., Choi, J. Y., Yoo, S. H., & Oh, Y. G. (2013). Evaluating spatial centrality for integrated tourism management in rural areas using GIS and network analysis. *Tourism Management*, *34*, 14–24. https://doi.org/10.1016/j.tourman.2012.03.005

Mccallum, A. (2007). *Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email*. *30*, 249–272.

Miller, P. R., Bobkowski, P. S., Maliniak, D., & Rapoport, R. B. (2015). *Talking Politics on Facebook : Network Centrality and Political Discussion Practices in Social Media*. https://doi.org/10.1177/1065912915580135

Narayanan, S. (2005). *The Betweenness Centrality Of Biological Networks A Study of Betweenness Centrality*.

Newman, M. (2010). Networks: An Introduction. In *Networks: An Introduction*. https://doi.org/10.1093/acprof:oso/9780199206650.001.0001

Park, K., & Kim, D. (2009). *Localized network centrality and essentiality in the yeast – protein interaction network*. 5143–5154. https://doi.org/10.1002/pmic.200900357

Rossman, G., Esparza, N., & Bonacich, P. (2010). I’d like to thank the academy, team spillovers, and network centrality. *American Sociological Review*, *75*(1), 31–51. https://doi.org/10.1177/0003122409359164

Williamson, S. A., & Tec, M. (2019). Random clique covers for graphs with local density and global sparsity. *35th Conference on Uncertainty in Artificial Intelligence, UAI 2019*.

Yang, B., & Liu, J. (2008). Discovering global network communities based on local centralities. *ACM Transactions on the Web*, *2*(1). https://doi.org/10.1145/1326561.1326570

Zhang, Y., Wang, X., Zeng, P., & Chen, X. (2011). Centrality characteristics of road network patterns of traffic analysis zones. *Transportation Research Record*, *2256*, 16–24. https://doi.org/10.3141/2256-03

Borgatti, S. P. (2005). *Centrality and network flow ଝ*. *27*(November 2003), 55–71. https://doi.org/10.1016/j.socnet.2004.11.008

Brandes, U., & Pich, C. (2007). Centrality estimation in large networks. *International Journal of Bifurcation and Chaos*, *17*(7), 2303–2318. https://doi.org/10.1142/S0218127407018403

Dwyer, T., Kosch, D., & Xu, K. (2003). *Visual Analysis of Network Centralities*. *Wuchty 2002*.