BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net *Hypothesis*

Volume 11(2)

Application of centrality measures in the identification of critical genes in diabetes mellitus

Chintagunta Ambedkar[1], Kiran Kumar Reddi[2], Naresh Babu Muppalaneni[3] & Duggineni Kalyani[3]*

[1]S R K Institute of Technology, Vijayawada; [2]Krishna University, Machilipatnam; [3]C R Rao AIMSCS, Hyderabad; Duggineni Kalyani – Email: kalyani.duggineni@gmail.com; *Corresponding author

Received January 02, 2015; Accepted January 31, 2015; Published February 28, 2015

Abstract:
The connectivity of a protein and its structure is related to its functional properties. Many experimental approaches have been employed for the identification of Diabetes Mellitus (DM) associated candidate genes. Therefore, it is of interest to use various graph centrality measures integrated with the genes associated with the human Diabetes Mellitus network for the identification of potential targets. We used 2728 genes known to cause Diabetes Mellitus from Jensenlab (Novo Nordisk Foundation Center for Protein Research, Denmark) for this analysis. A protein-protein interaction network was further constructed using a tool Centralities in Biological Networks (CentiBiN) with 1020 nodes after eliminating the duplicates, parallel edges, self-loop edges and unknown Human Protein Reference Database (HPRD) IDS. We used fourteen centralities measures which are useful in identifying the structural characteristic of individuals in the network. The results of the centrality measures are highly correlated. Thus, we identified genes that are critically associated with DM. We further report the top ten genes of all fourteen centrality measures for further consideration as targets for DM.

Background:

People are facing major life threatening disease like diabetes, cancer, hyper tension, heart disease and stroke **[1].** We have chosen Diabetes Mellitus for our study. Diabetes Mellitus is a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both. The chronic hyperglycemia of diabetes is associated with long-term damage, dysfunction, and failure of various organs, especially the eyes, kidneys, nerves, heart, and blood vessels. As the risk of cardiovascular disease is much higher for a diabetic, it is crucial that blood pressure and cholesterol levels are monitored regularly **[2].**

Diabetes Mellitus is not a single disease but a group of disorders with glucose intolerance in common. Many online databases are available to research genes across species. Different databases available that allows access to information about phenotypes, pathways, and variations of many genes across species. Before the candidate-gene approach was fully developed, various other methods were used to identify genes linked to disease-states. However, these methods are not as beneficial when studying complex diseases for several reasons. In this scenario, candidate gene approaches were found in identifying the risk variants associated with various diseases of interest such as dementia, cancer, diabetes, asthma, and hypertension. The candidate gene approach to conducting genetic association studies focuses on associations between genetic variation within pre-specified genes of interest and phenotypes or disease states. **[3, 4]**

With the tremendous escalation of human protein interaction data, the entanglement of the techniques can be conquered through protein–protein interaction networks (PPINs). The function and activity of a protein are often modulated by other proteins with which it interacts **[5, 6].** Data might be represented as networks, in which the vertices (e.g. transcripts, proteins or metabolites) are linked by edges (correlations, interactions or reactions, respectively). Structural analysis of networks can lead to new insights into biological systems and is a helpful method for proposing new hypotheses **[7-10].**

Methodology:

Proteins are the representatives of the biological networks and they are realized only if the relationship between essentiality and topological properties such as the degree distribution, clustering coefficients, centrality measures, and community structures of the network are studied **[9].** Network centralities are used to rank elements of a network according to a given importance concept **[11].**

However, the use of centralities as a structural analysis method for biological networks is controversial and several centrality measures should be considered within an exploratory process **[16].** To support such analysis and due to the complexity of both biological networks and centrality calculations, a tool is needed to facilitate these investigations. Here we present CentiBin, an application for the calculation and visualization of centralities for biological networks.

The human protein interaction data was obtained from Human Protein Reference Database (HPRD). The main purpose of using HPRD dataset is it focuses on likely true Protein-Protein Interaction (PPI) set by generating sub networks around proteins of interest. HPRD represents a centralized platform to visually depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome **[17]**. We have followed the procedure mentioned in **Figure 1** for identifying the critical genes for diabetes mellitus.

Data Set:

We have extracted the human gene involving in Diabetes mellitus from the database developed by Jensen Group (Jensenlab) of Novo Nordisk Foundation Center for Protein Research, Denmark. Jensenlab is maintaining a DISEASES database. DISEASES database is a frequently updated web resource that integrates evidence on disease-gene associations from automatic text mining, manually curated literature, cancer mutation data, and genome-wide association studies. We have mined the Jensenlab DISEASES database for the genes causing Diabetes mellitus. We got that 2728 genes causing diabetes mellitus, after eliminating duplicate entries it reduced to 2017 genes.

Network Properties and Centrality Measures:

Here we have calculated fourteen different graph centrality measures such as degree, eccentricity, closeness, radiality, centroid values, Stress, shortest-path betweenness, current-flow closeness, current-flow betweenness, Katz status index, Eigen vector, hits-authority, hits-hubs and Page Rank using the tool CentiBin and are defined as follows**[12-13,15,16,18-32].**

Degree $C_{dev}(v) = \deg(v)$

Eccentricity $C_{ecc}(v) = \frac{1}{max\,\{dist\,(v,w): w \in v\}}$

Closeness $c_c(v) = \frac{1}{\sum_{u \in v} dist\,(u,v)}$

Radiality $C_{rad}(v) = \frac{\sum_{w \in v}(\Delta_G + 1 - dist\,(v,w))}{n-1}$

Stress C_{str}(v)= $\sum_{s \neq v \in V} \sum_{t \neq v \in V} \sigma_{st}(v)$

Shortest path Betweenness $c_B(v) = \sum_{s \neq t \neq v \neq V} \frac{\rho_{st}(v)}{\rho_{st}}$

Shortest path closeness $C_{cfc}(v) = \frac{n-1}{\sum_{t \neq v} p_{vt}(v) - p_{vt}(t)}$

$p_{vt}(t)$ equals the potential difference.

Katz status index $c_k = \sum_{k=1}^{\infty} \propto^k (A^t)\, \vec{1}$

Eigen Vector $\lambda\, C_{IV} = \mathrm{A} C_{IV}$

Centroid $C_{cen}(v) = \min\{f(v,w): v\{v\}\}$

Where f (v, w) = γ v (w) – γ w (v) and γ v (w) denotes the number of vertices that are closer to v than to w.

Page Rank $C_{pr} = dpC_{pr} + (1-d)\vec{1}$

Where P is the transition matrix and d is the damping factor.

Betweenness $C_{cfb}(v) = \frac{1}{(n-1)(n-2)} \sum_{s,t \in v} T_{st}(v)$

Where T_{st}(v) equals the fraction of electrical current running over vertex v in a network

HITS-Hubs $C_{hubs} = AC_{aut\,hs}$

Hits-authority $C_{aut\,hs} = A^T C_{hubs}$

Correlation analysis of centrality measures

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. The fourteen different centrality measures were calculated for each and every node in the interact and ranked based on their scores. Pair wise correlation between the various centrality measures was obtained through Spearman's rank correlation coefficient ρ which is defined as

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

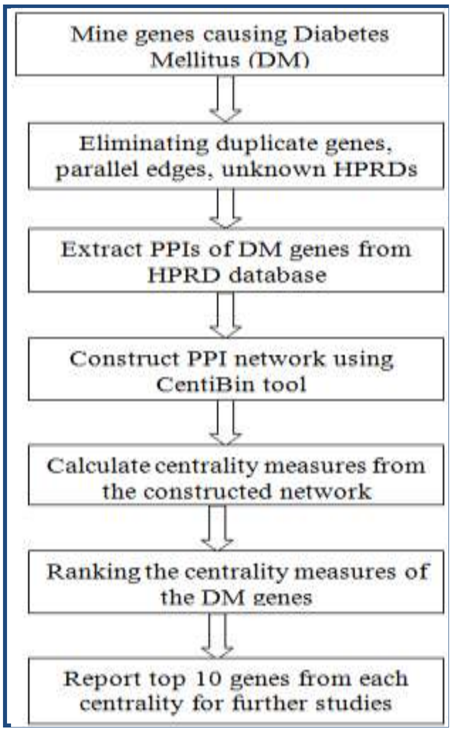
Mine genes causing Diabetes Mellitus (DM)
Eliminating duplicate genes, parallel edges, unknown HPRDs
Extract PPIs of DM genes from HPRD database
Construct PPI network using CentiBin tool
Calculate centrality measures from the constructed network
Ranking the centrality measures of the DM genes
Report top 10 genes from each centrality for further studies

Figure 1: Flow Chart

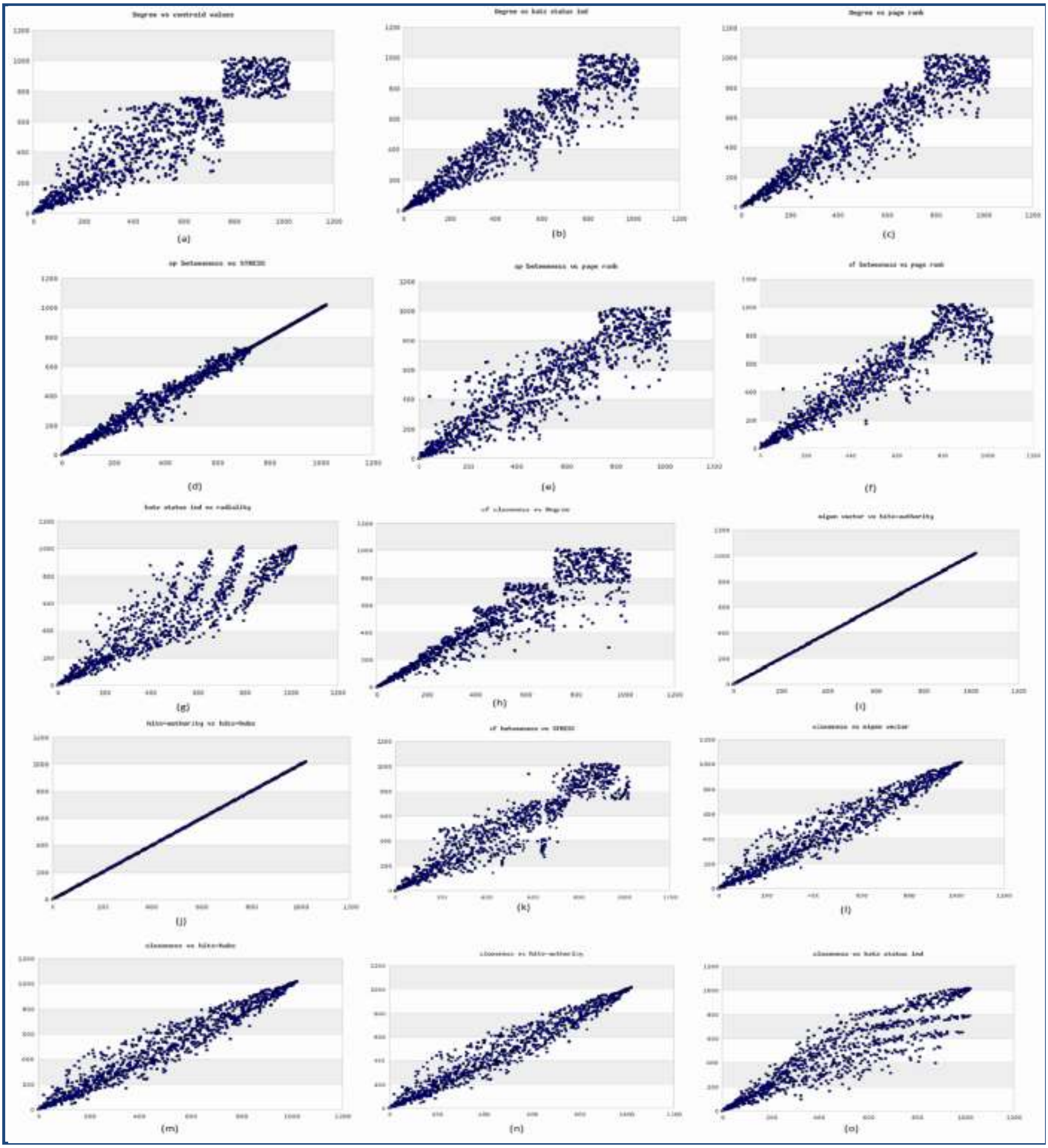

Figure 2: correlation among of different pairs of centrality measures for the diabetes mellitus genes whose correlation coefficient is above 0.9 **(a)** Degree vs Centroid **(b)** degree vs Katz status index **(c)** Degree vs page rank **(d)** sp betweeness vs Stress **(e)** sp betweeness vs page rank **(f)** cf betweeness vs page rank **(g)** Katz status index vs radiality **(h)** cf closeness vs degree **(i)** Eigen Vector vs hits authority (j) hits authority vs hits hubs **(k)** cf betweeness vs Stress **(l)** closeness vs eigen vector **(m)** closeness vs hits-hubs **(n)** closeness vs hits-authority **(o)** closeness vs katz status index.

Result & Discussion:

With the help of bioDBnet (http://biodbnet.abcc.ncifcrf.gov) we find the equivalent HPRD ids for the 2017 genes. Out of 2017 proteins we got HPRD IDs for 1834 proteins. We are unable to get the equivalent HPRD IDs for 183 proteins, so we find the HPRD IDs through their aliases. Still we couldn't find the HPRD ids for 39 proteins because these are the new entries in the database. After eliminating duplicates we got 1876 unique genes.

PPI Network

To construct the Protein-Protein Interaction network, we have downloaded and deployed the interactions database from HPRD website (http://hprd.org) in our local database. We have retrieved the PPIs for 1876 unique proteins where both source and sink proteins are in 1876 unique proteins. With that we have constructed a network using CentiBin with 1151 vertices and 3389 edges. Finally we got 1020 vertices with 2891 edges after eliminating the self edges, parallel edges from the network.

Using CentiBin we have calculated fourteen different graph centrality measures such as degree, eccentricity, closeness, radiality, centroid values, Stress, shortest-path betweenness, current-flow closeness, current-flow betweenness, Katz status index, Eigen vector, hits-authority, hits-hubs and Page Rank for the PPI network constructed. The top ten genes of each centrality measure are presented in **Table 1 (see supplementary material).**

Correlation analysis on centrality properties

The pair wise correlation coefficients of the fourteen centrality measures depicted for the Diabetes Mellitus elucidate that they all are positively correlated and their correlation value lies above 0.52 as represented in **Table 2 (see supplementary material), Figure 2**. Here, the difference d_i represents the difference in the ranks of each observation on the two variables which here represents the centrality scores.

Conclusion:

Many experimental approaches have been used to identify candidate genes in DM. We used various graph centrality measures integrated with the genes to identify potential drug targets. We calculated fourteen centralities measures for the constructed network with positive correlation having values greater than 0.52. This helped to identify genes that are highly critical in DM. We thus report the top 10 genes of all fourteen centralities for consideration as potential targets for DM.

Acknowledgment:

Dr M Naresh Babu and Ms Duggineni Kalyani would like to thank SERB, Department of Science and Technology, Government of India under FAST Track scheme order No. SB/FTP/ETA-436/2012

References:

[1] Sharma M *et al. Indian J Occup Environ Med.* 2009 **13:** 109 [PMID: 20442827]

[2] National Diabetes Data Group *Diabetes* 1979 **28:** 1039 [PMID: 510803]

[3] Chen Chen J *et al. Nucleic Acids Res.* 2009 **37**: W305 [PMID: 19465376]

[4] Miyata T, *Hypertension Res.* 2008 **31:** 173 [PMID: 18360034]

[5] Barabási A-L *et al. J Nat Rev Genet.* 2011 **12:** 56 [PMID: 21164525]

[6] Ackers GK, *Adv Protein Chem.* 1970 **24**: 343 [PMID: 4916268]

[7] Albert R, *Reviews of Modern Physics* 2002 **74:** 47

[8] Milo R *et al. Science* 2002 **298:** 824 [PMID: 12399590]

[9] Holme P *et al. Bioinformatics* 2003 **19:** 532 [PMID: 12611809]

[10] Wuchty S *et al. J Theor Biol* 2003 **223:** 45 [PMID: 12782116]

[11] Fell DA *et al. Nat Biotechnol* 2000 **18**: 1121 [PMID: 11062388]

[12] Zhang A. Chapter 4: Basic Properties and Measurements of protein interaction network. Protein Interaction Networks Computational Analysis. Cambridge University Press 2009 33-49.

[13] Jeong H *et al. Nature* 2001 **411:** 41 [PMID: 11333967]

[14] Caldarelli G, Scale-Free Networks: Complex webs in nature and technology. Oxford UK: Oxford University Press; 2007.

[15] Hegde SR *et al. PLoS Comput Biol* 2008 **4:** e1000237 [PMID: 19043542]

[16] Koschützki D, Schreiber F: Comparison of Centralities for Biological Networks. Proc German Conf Bioinformatics (GCB'04), Volume P-53 of LNI 2004:199-206

[17] http://hprd.org

[18] Kranthi T *et al. Mol BioSyst* 2013 **9:** 2163 [PMID: 23728082]

[19] Stelzl U *et al. Cell* 2005 **122:** 957 [PMID: 16169070]

[20] Chen J *et al. BMC Bioinformatics* 2009 **10:** 73 [PMID: 19245720]

[21] Newman, M.E.J. *Networks: An Introduction.* 2010 Oxford, UK: Oxford University Press.

[22] Borgatti *et al. Social Networks* (Elsevier) 2006 **28**: 466 doi: 10.1016 / j.socnet. 2005.11.005

[23] Kann MG, *Brief Bioinform* 2007 **8:** 333 [PMID: 17638813]

[24] Goh KI *et al. Proc Natl Acad Sci USA*. 2007 **104:** 8685 [PMID: 17502601]

[25] Junker HB *et al. BMC Bioinformatics* 2006 **7:** 219 [PMID: 16630347]

[26] Kleinberg JM, *Journal of the ACM* 1999 **46:** 604

[27] Ortutay C *et al. Nucleic Acids Res* 2009 **37:** 622 [PMCID: PMC2632920]

[28] Keshava Prasad TS *et al. Nucleic Acids Res* 2009 **37**: D767 [PMID: 18988627]

[29] Chatr-aryamontri A *et al. Nucleic Acids Res* 2007 **35:** D572 [PMID: 17135203]

[30] Xenarios I *et al. Nucleic Acids Res.* 2000 **28:** 289 [PMCID: PMC102387]

[31] Page L, Brin S, Motwani R, Winograd T: The Page Rank Citation Ranking: Bringing Order to the Web. Tech rep, Stanford Digital Library Technologies Project 1998.

[32] http://www.math.cornell.edu/~mec/Winter2009/Raluca Remus/Lecture4/lecture4.html

Edited by P Kangueane

Citation: **Ambedkar *et al.*** Bioinformation 11(2): 090-095 (2015)

Supplementary material:

Table 1: Top10 genes from each centrality measure for the network constructed with 1020 genes of Diabetes Mellitus

Degree	**eccentricity**	**closeness**	**radiality**	**centroid values**	**sp betweeness**	**cf closeness**	**cf betweeness**	**Katz status index**	**Eigen vector**	**page rank**	**hits-authority**	**hits-hubs**	**stress**
SRC	APP	SRC	SRC	SRC	SRC	SRC	SRC	SRC	SRC	SRC	SRC	SRC	SRC
ESR1	STAT3	ESR1	ESR1	ESR1	AKT1	MAPK1	AKT1	ESR1	ESR1	ESR1	ESR1	ESR1	AKT1
MAPK1	AKT1	MAPK1	MAPK1	MAPK1	ESR1	ESR1	TP53	MAPK1	MAPK1	AKT1	MAPK1	MAPK1	ESR1
AKT1	MAPK1	AR	AR	AR	TP53	AKT1	ESR1	AKT1	STAT3	MAPK1	STAT3	STAT3	MAPK1
GRB2	SRC	AKT1	AKT1	AKT1	HSP90AA1	GRB2	EGFR	EGFR	EGFR	TP53	EGFR	EGFR	TP53
EGFR	LCK	HSP90AA1	HSP90AA1	HSP90AA1	APP	EGFR	GRB2	GRB2	GRB2	EGFR	GRB2	GRB2	HSP90AA1
TP53	FYN	EGFR	EGFR	EGFR	EGFR	MAPK3	MAPK1	STAT3	PIK3R1	GRB2	PIK3R1	PIK3R1	EGFR
FYN	STAT1	STAT3	STAT3	STAT3	GRB2	STAT3	APP	PIK3R1	PTPN11	FYN	PTPN11	PTPN11	GRB2
MAPK3	MAPK8	GRB2	GRB2	MAPK3	MAPK1	PIK3R1	FYN	MAPK3	AKT1	MAPK3	AKT1	AKT1	APP
PIK3R1	HSP90AA1	CAV1	CAV1	GRB2	PIK3R1	TP53	CASP3	TP53	AR	PIK3R1	AR	AR	AR

Table 2: The correlation coefficients of different pairs of centrality measures shows that the ranking of the nodes differs based on their formalism

Centrality measures	**Degree**	**eccentricity**	**closeness**	**sp betweeness**	**cf betweeness**	**cf closeness**	**centroid**	**Eigen vector**	**hits-authority**	**hits-hubs**	**Katz status index**	**page rank**	**radiality**	**stress**
Degree	1	0.7023	0.7727	0.8687	0.9377	0.9365	0.9168	0.7099	0.7099	0.7099	0.9437	0.9423	0.7727	0.8867
eccentricity	0.7023	1	0.8444	0.6449	0.6668	0.7855	0.7742	0.7949	0.795	0.795	0.7982	0.5803	0.8444	0.6676
closeness	0.7727	0.8444	1	0.6444	0.7008	0.8892	0.84	0.9713	0.9713	0.9713	0.9136	0.611	1	0.6778
sp betweeness	0.8687	0.6449	0.6444	1	0.9292	0.7698	0.8087	0.5584	0.5584	0.5584	0.8099	0.913	0.6444	0.9941
cf betweeness	0.9377	0.6668	0.7008	0.9292	1	0.8828	0.8661	0.6122	0.6122	0.6122	0.8862	0.9389	0.7008	0.9296
cf closeness	0.9365	0.7855	0.8892	0.7698	0.8828	1	0.9451	0.8332	0.8332	0.8332	0.9762	0.8183	0.8892	0.7946
centroid	0.9168	0.7742	0.84	0.8087	0.8661	0.9451	1	0.762	0.762	0.762	0.9319	0.8337	0.84	0.8296
Eigen vector	0.7099	0.7949	0.9713	0.5584	0.6122	0.8332	0.762	1	0.9999	0.9999	0.8696	0.5342	0.9713	0.5996

hits-authority	0.7099	0.795	0.9713	0.5584	0.6122	0.8332	0.762	0.9999	1	0.9999	0.8696	0.5342	0.9713	0.5996
hits-hubs	0.7099	0.795	0.9713	0.5584	0.6122	0.8332	0.762	0.9999	0.9999	1	0.8696	0.5342	0.9713	0.5996
Katz status index	0.9437	0.7982	0.9136	0.8099	0.8862	0.9762	0.9319	0.8696	0.8696	0.8696	1	0.8443	0.9136	0.8363
page rank	0.9423	0.5803	0.611	0.913	0.9389	0.8183	0.8337	0.5342	0.5342	0.5342	0.8443	1	0.611	0.9202
Radiality	0.7727	0.8444	1	0.6444	0.7008	0.8892	0.84	0.9713	0.9713	0.9713	0.9136	0.611	1	0.6778
Stress	0.8867	0.6676	0.6778	0.9941	0.9296	0.7946	0.8296	0.5996	0.5996	0.5996	0.8363	0.9202	0.6778	1