UNCOVERING SUBCELLULAR REGULATORY NETWORKS

A Report submitted in partial fulfillment of Summer Internship for the degree

of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING
RGUKT RK VALLEY CAMPUS

By
S N Karishma
Entry No. R082546
(Summer Research Fellow)

To
Dr Sumeet Agarwal
(Assistant Professor, IIT-Delhi)

RGUKT RK Valley Campus,
Rajiv Gandhi university of Knowledge Technologies,
RK Valley, Kadapa, Andhra Pradesh.

July 2013

Certificate

Acknowledgements

I would like to thank **Dr. Sumeet Agarwal,** Asst. Prof, Department of Electrical Engineering, IIT-Delhi for giving me opportunity to undertake the summer internship. I wish to express my profound gratitude and indebtedness to him for introducing the present topic and for his inspiring guidance, constructive criticism and valuable suggestions throughout the course of internship.

I also thank Mr. Vamsi Krishna, Asst. Prof, Department of Electrical Engineering, GITAM University, Vishakapatnam for his invaluable suggestions. My sincere thanks to Mr. Sudhakar, Mr. Ramakant and Mr. Bhaskaraiah, Faculty of RGUKT for all the encouragement and support they have extended at the beginning of this undertaking.

S N Karishma
Entry No. R082546

Abstract

The understanding of complex real world systems has significantly increased with the advancements in the modern science of networks. Genetic regulatory networks are one of the first networked, real world dynamical systems. With the availability of gene expression and protein interaction data, studies on these networks have increased and large scale modeling attempts were made by relating structure to functionality. In this project, a data-driven approach is followed to study different empirical networks by employing a diverse range of diagnostics. A total of over two hundred real world networks obtained from different domains are analyzed. More than two hundred network diagnostics or summary statistics are computed and examined for these empirical networks. I demonstrate how this data-driven approach can be used to organize the networks, as well as to classify protein interaction networks from other real world networks.

Keywords: Networks, Dimensionality Reduction, Learning, Visualization, Classification.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Inspired by empirical studies of networked systems such as social networks, political networks and biological networks, researchers have in recent years developed a variety of techniques and models to help us understand or predict the behavior of the network systems. In this project, a data-driven approach to the study of complex networks is employed. I applied it to multiple domains in real world systems, focusing in particular on protein interaction networks. Since networks are intrinsically high dimensional objects, it is often hard to determine a suitable way of characterization for a given task. Due to dissimilarity and diversity in real world data, there is no systematic program for characterizing network structure. In addition, there are no particular subsets of diagnostics that are universally accepted. While studying a particular type of unfamiliar network, the observation that examining just one or a few network properties can be misleading motivated the researchers to attempt to develop a more holistic methodology for network investigation. They examined the network system from as many different perspectives as possible to get a handle on how they relate to other network types previously observed and studied. Here I simultaneously investigate many networks using many diagnostics in a data-driven fashion, and demonstrate how this approach serves to organize both networks and diagnostics.

Each network from real world data is taken as input, and network diagnostic is computed from the library of algorithms available. Thus design matrix is obtained, with networks as its rows and features as its columns. Each entry in the design matrix represents the value of one feature for one network. These empirical networks are modular and hierarchical, and have specific distribution of topological features that can be used to characterize them. Since our design matrix is hard dimension reduction problem, classification decision can't be guessed ahead of time. For classification, the goal is to map the input data into feature space in which the members from different classes are clearly separated. To visualize this classification, the high dimensional input data is mapped into a (2-3d) space that preserves the intrinsic structure as much as possible. Reduction of dimensionality helps in showing the clustered structure of network as well as in estimating a function of several features from the network-feature design matrix. The representation of network ensembles in a low dimensional space provides a tractable way of estimating the range of structural variants, or the region of structure space captured by a given network model with the given parameter settings.

I consider learning which actually refers to some form of algorithm for reducing error on the design matrix. In unsupervised learning, the system forms clusters of the input patterns. Different clustering algorithms lead to different clusters. Principal component analysis (PCA) is an unsupervised approach to finding the right features from the data. I seek to represent the high-dimensional data in a lower dimensional space of 2-3 dimensions. This will reduce degrees of freedom, the space and time complexities. PCA guarantees maximal retention of the variance when projecting data into a lower dimension. It finds a linear subspace and thus cannot deal properly with the real world data lying on the non linear manifolds. To overcome this problem, Isomap technique is employed. Compared with PCA, Isomap is characterized by two parts: it is manifold based and it has non linearity. Isomapping involves constructing neighbourhood graph, computing geodesic distances and constructing low-dimensional embedding. In brief, Isomap is a low-dimensional, neighbourhood-preserving embeddings of high-dimensional inputs. The most complex problem of Isomap is to explain the physical meaning of the

axes extracted. I have made attempts to determine what these axes can be. The resulting pattern of Isomap is affected by the scale of original dimensions and the definition of the neighbourhood. Moreover, Isomap can visualize the data well when the input data are well sampled and have little noise. Since real world data is noisy, Isomap often fails to nicely visualize them. In this situation, the class labels of the data, if known, can be used to retrieve the negative effect of noise. It is well known that points belonging to same class are often close to each other than those belonging to different classes.

In supervised learning, a category label is explicitly provided for the networks in the design matrix, and sum of the distances for the patterns is reduced. Supervised Isomap(S-Isomap), the improvised version of Isomap, utilizes the class information to guide the procedure of non linear dimensionality reduction. In S-Isomap, the neighbourhood graph of the input data is constructed according to a certain kind of dissimilarity between data points, which is specially designed to integrate class information. Hence, S-Isomap can be used to recover the true manifold of the noisy data.

Chapter 2: Related Work

192 real world networks (Appendix B) obtained from Onnela *et. al*. are analyzed. They include empirical networks drawn from different domains like biological (brain connectivity, protein interaction & metabolic growth), social networks, political networks (political voting, political cosponsorship, political committee) and others (financial correlation, word adjacency, fungal growth) [14]. These networks are modular and hierarchical, and have specific distribution of topological features that can be used to characterize them [12]. 70 network diagnostics that are taken from literature are used (Appendix A). These network algorithms include many a kind of structural properties like measures of degree freedom, clustering of links, different notions of node centralities, frequencies of small motifs, mesoscopic structure via partitioning into communities, spectral properties of adjacency matrix etc... Each of them takes network as input and computes some property of it. Some diagnostics return number and some return multiple features. A suite of 253 network metrics are used so as to comprehensively compare all 192 networks simultaneously, allowing for in depth evaluation of simultaneous models. Different metrics are put on a common scale to obtain meaningful comparison. Hence the design matrix, with the networks as its rows and metrics as its columns, is normalized. The normalization includes standardizing all the networks to have zero mean and unit standard deviation. These values are then mapped to unit standard interval via the logistic function $f(z) = (1+ \exp(-z))$[7].

Apart from computational constraints like time limit, some diagnostics are also undefined for certain networks, for instance those which are not connected. Hence, design matrix includes missing features for some networks. These missing values are handled either by removing the columns that are less than 80% full or by replacing missing values in a column with the average mean of the other entries in the respective column [7].

The presence of large number of features obstructs the interpretations of the useful patterns of the data. Many features are correlated with each other through linear combination or other functional dependence. Redundancy must be removed [7, 15]. Feature Selection aims to build a new feature space of reduced dimensionality, producing a compact representation of the network data that may be distributed across several of the original features. Using Feature Selection, we try to decide on a feature subset, discarding features that do not contribute towards predicting the response [16]. The high-dimensional feature-space network representation is mapped to a low-dimensional space and thus, I tried to find a few (2-4) particular dimensions that capture the bulk of variation between commonly studied network types. Feature Selection is not easily interpretable because the physical meaning of the response features can’t be directly retrieved. Principal Component Analysis (PCA), a Linear Dimensionality Reduction Technique is employed to identify the feature subset which associates strongly with various network characteristics to a surprisingly high degree. Thus Dimensionality Reduction determines a representation of that manifold that will allow the projection of data points on it. As shown in the figure 1, each data point on the scattered plot represents the network's position along the given two dimensions. Different symbols are used for different domains from which networks are taken. Design matrix is clustered to see the similarities [7].

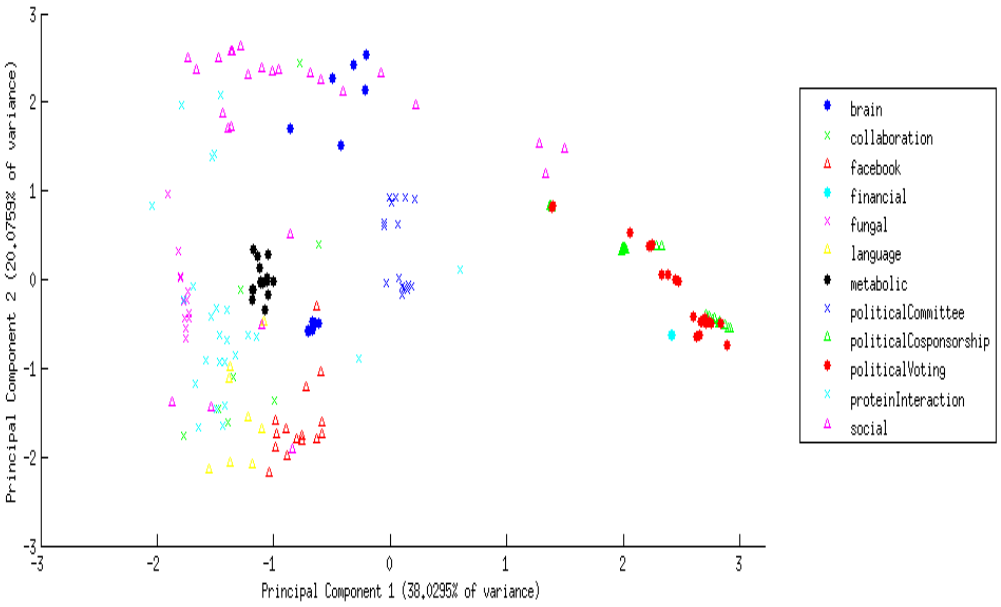


Figure 1: Network Clustering via PCA dimensionality reduction - I

The results reveal that some networks like financial correlation, fungal growth, and metabolic growth are highly cohesive and form tight clusters. Political voting, political cosponsorship are next to the above three, and are cohesive enough. Protein interaction and political committee networks are confined to a restricted space, but are less clear cut. This implies they include networks from wider range of sources, and are not well-defined. The first five dimensions captured 80% of variance with first two dimensions alone capturing 58% of it. PCA guarantees maximum retention of the variance when projecting data into a lower dimension.

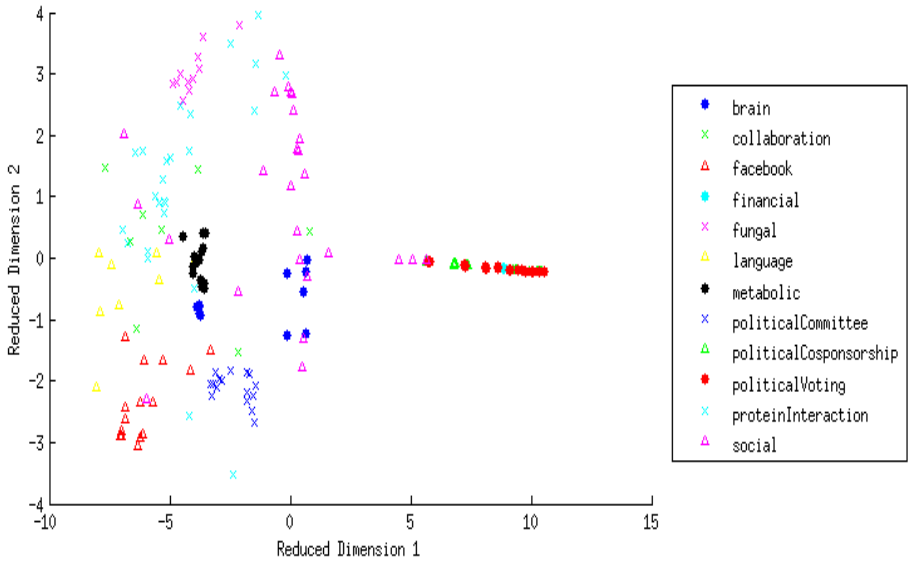


Figure 2: Network Clustering via Isomap Dimensionality Reduction – I

To examine in greater detail, non linear dimensionality reduction is carried out using Isomap. Isomap is an unsupervised learning algorithm that computes low-dimensional, neighbourhood-preserving embeddings of high-dimensional inputs. In contrast to previous algorithms, Isomap efficiently computes a globally optimal solution, and for an important class of data manifolds, it is guaranteed to converge asymptotically to the true structure. Each data point is connected to 22 nearest neighbours to obtain single largest connected component, thus having K = 22. Figure 2 shows the two dimensional embeddings of the network structure. It is observed from the computations that the first four dimensions alone capture 99% of total variance with first one alone accounting for over 96% of it. It is also observed that the two basic measures of network density and size are sufficient to capture the variability between different network types [7]. Since the overlaps between samples are not considerably high between different classes, these 253 network features can be used to classify the real world data.

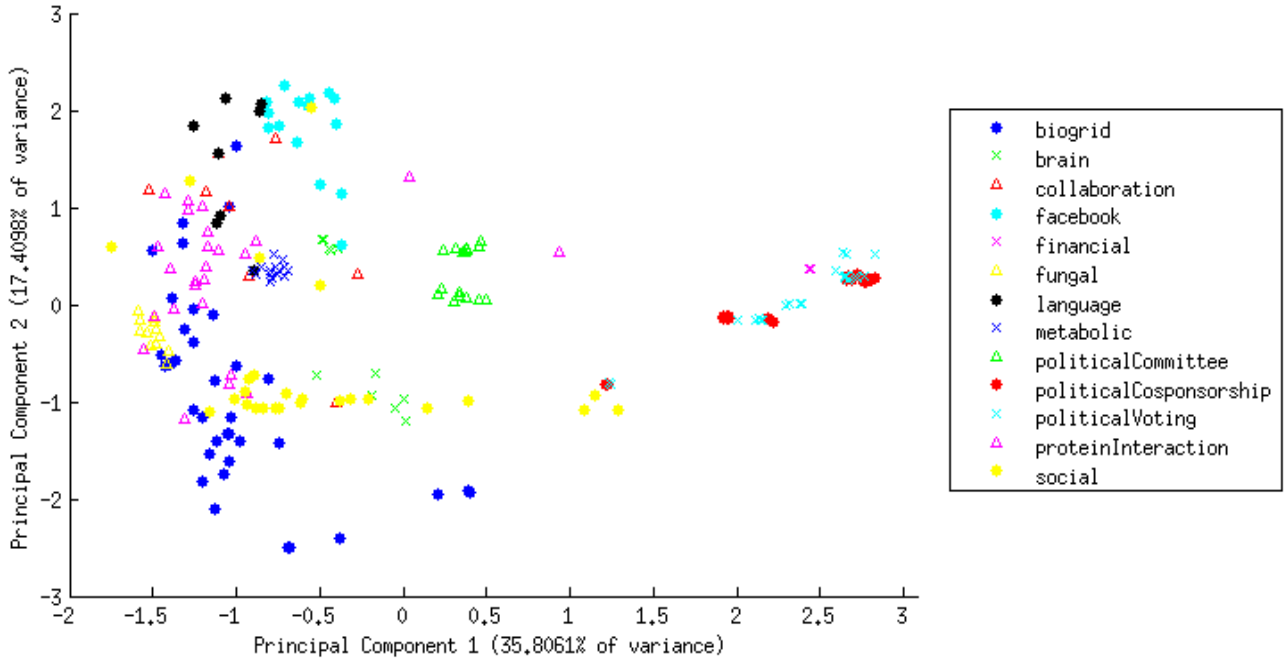
Chapter 3: Organization of Protein Interaction Networks

The above data-driven approach to organizing and using many different network diagnostics serves as a general purpose tool for further network investigation. I included the biogrid (Biological Genetic Repository for Interaction Database) data [6] and checked the classification. The data includes a curated biological database of protein-protein and genetic interactions for all major model organism species. Protein Interaction Network is a representation of proteins with directed edges joining them if a mechanistic physical interaction exists between the proteins. These Protein Interaction Networks are included to the dataset of 192 real world networks collected.

The protein interaction data is obtained using a C program to extract network ids from source file and map each unique id to an index. An adjacency matrix file is created in MATLAB that has a sparse matrix with 1's along the interaction cells and 0's at the rest. These files are loaded to SQL database which already consists of 192 real world networks, there by summing up to 234 networks. The previous network diagnostics are applied and 253 network features are computed for 234 networks. The network-metric matrix is normalized. There are more than 15% missing values, and they are handled by deleting the column if it less than 80% full or by replacing the missing values with average mean of the remaining values if the column is more than 80% full. After updating the missing values, the design matrix obtained was made up of 211 features for 234 networks.

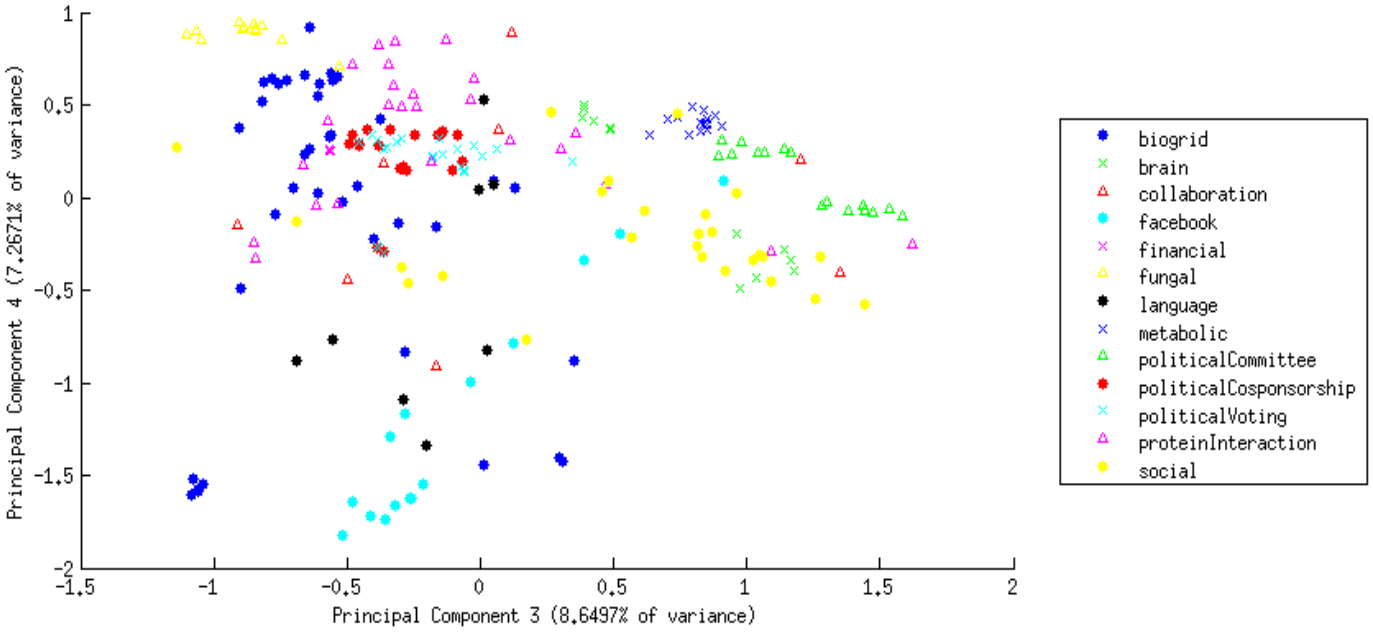
Chapter 4: Results

The Linear Dimensionality Reduction is done through PCA as shown in the figure 3. From the figure, it is observed that first two dimensions contributed for 53.2% of variance. Large energy loss is observed with reduction to low dimensionality by PCA and samples from various classes are mixed up in projection space. From figures 4 and 5, we can say that the misclassification occurs in the real world data with increase in the principal component number. PCA is only able to find a linear subspace and thus cannot deal properly with the data lying on non-linear manifolds.



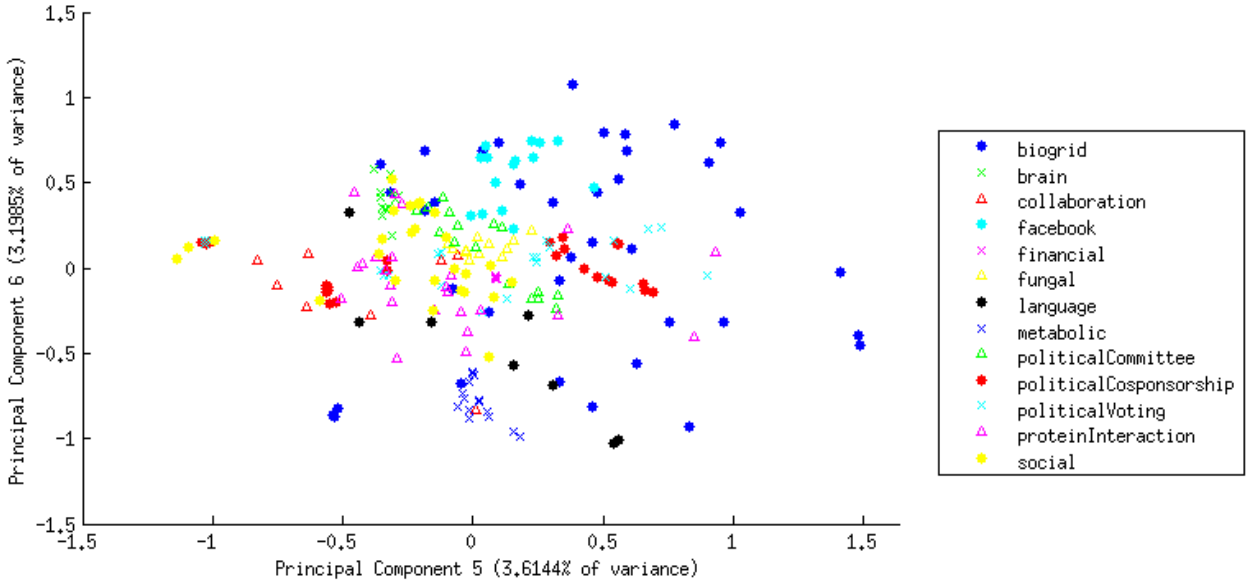
i . First two reduced dimensions

Figure 3: Network Clustering via PCA dimensionality Reduction - II



ii. Third and fourth reduced dimensions

Figure 4: Network Clustering via PCA dimensionality Reduction - II



iii. Fifth and sixth reduced dimensions

Figure 5: Network Clustering via PCA dimensionality Reduction – II

Isomap is used to reveal the structure of this dataset. The nearest neighbourhood factor K = 11 provides the largest connected component. Figure 7 shows curve of residual variance as the number of Isomap dimensions is increased. Intrinsic dimensionality can be as low as 4 to get variance of 99%.

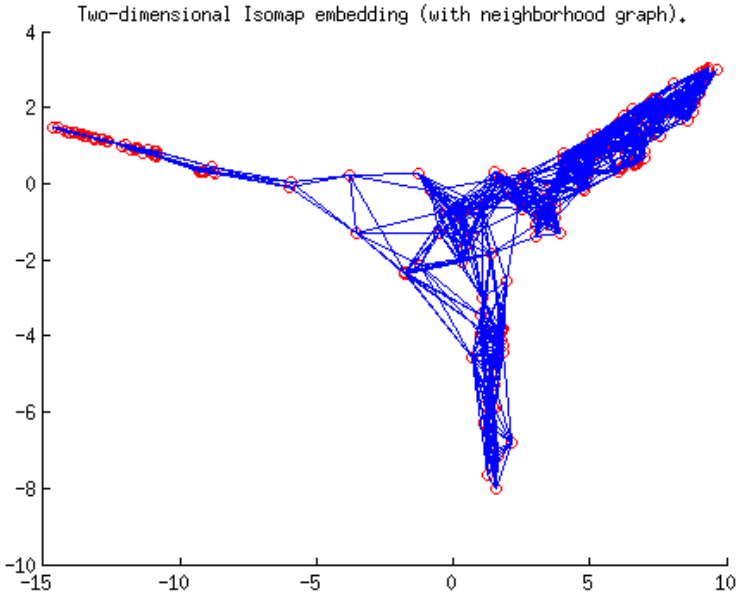


Figure 6: Two-dimensional Isomap embedding (with neighbourhood graph)

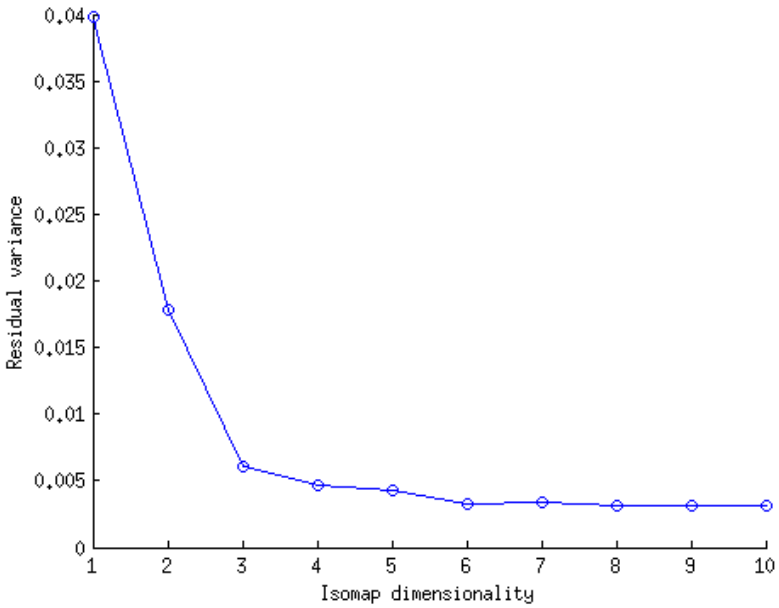
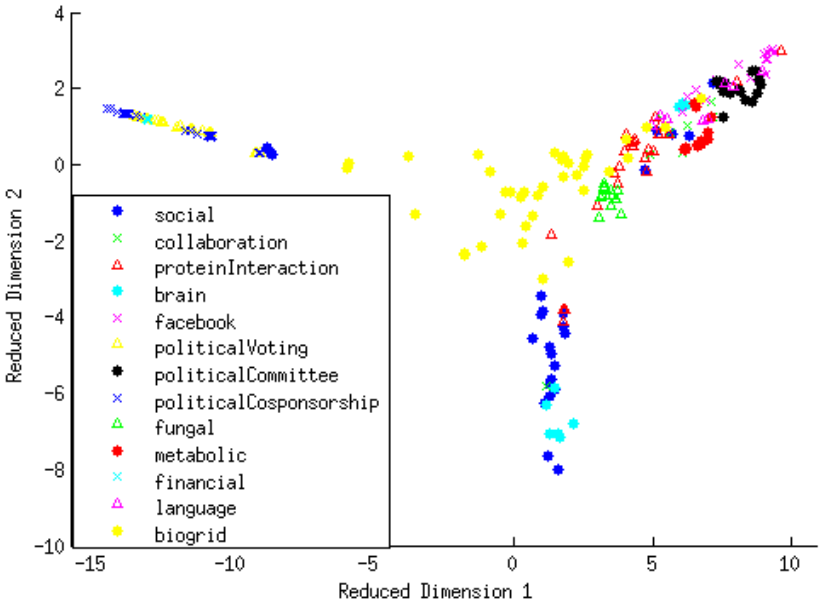


Figure 7: Residual variance as the number of Isomap dimensions is increased

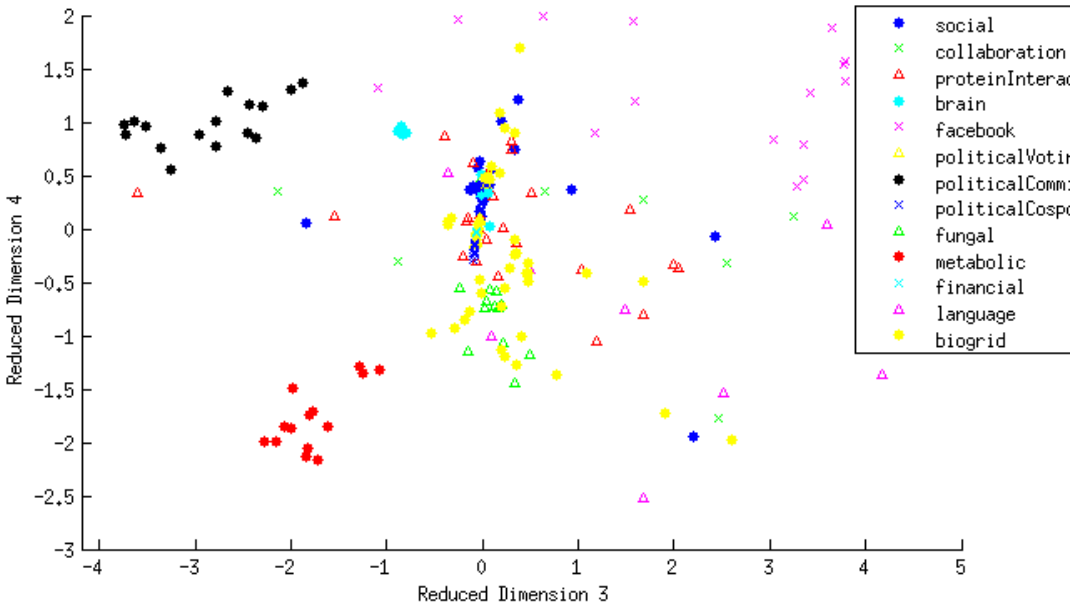
From figure 8, we observe that the biogrid samples formed compact clusters considerably, which indicates that those 211 features can be used to distinguish biogrid networks from other networks successfully. The shape of projection of social, political cosponsorship and political voting domains is

almost a straight line. This implies that the network data corresponds to a one dimensional manifold in high dimensional space.



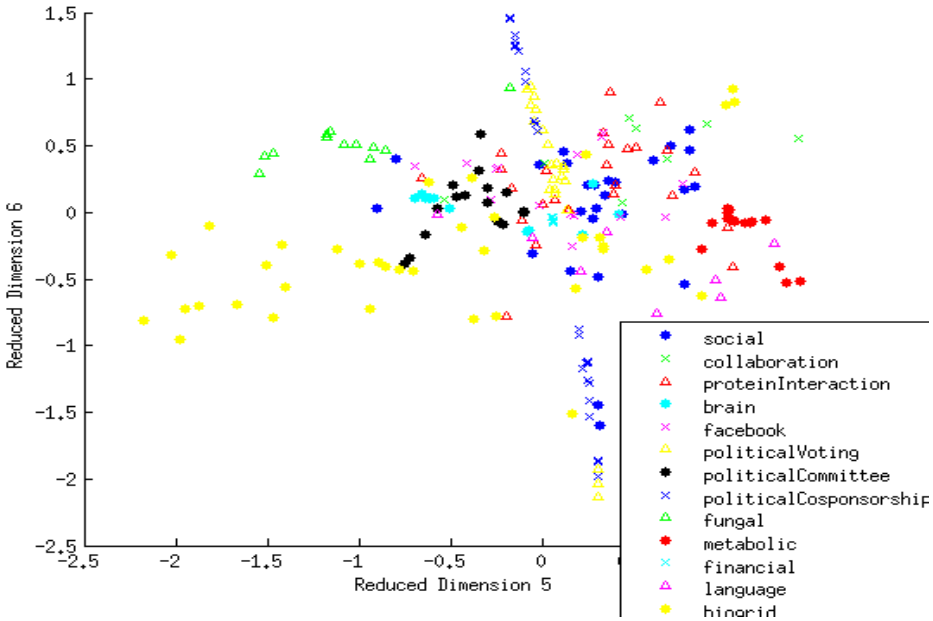
i. First two reduced Dimensions

Figure 8: Network Clustering via Isomap dimensionality Reduction - II



ii. Third and fourth reduced dimensions

Figure 9: Network Clustering via Isomap dimensionality Reduction - II



iii. Fifth and sixth reduced dimensions

Figure 10: Network Clustering via Isomap dimensionality Reduction – II

The correlation coefficients are computed for reduced dimensions with the feature-space design matrix to find the network-feature that the reduced dimension is capturing. The first and second maximally correlated values (r_1, r_2 respectively) are given in the table 1.

Dimension	**r_1**	**Feature-1**	**r_2**	**Feature-2**
1	-0.9266	degreeCentrality_harmmean	-0.9144	degreeCentrality_Geomean
2	0.7354	fraction2core_snowball100	0.7246	numNodes_snowball100
3	0.7604	evectorCentrality_fit_lognormal	0.7601	evectorCentrality_fit_wbl
4	0.6121	assortativeCoefficient_snowball100	0.5624	assortativeCoefficient
5	0.5025	clusteringCoeff_var	0.4973	clusteringCoeff_iqr

Table 1: Maximally correlated features for each reduced dimension in Isomap

The first dimension was found to have a very high correlation with the measurement of variability in node centrality. The second dimension is maximally correlated with the fraction of the network's nodes covered by the 2-core, and also with number of the nodes in the sampled network. Thus the two basic measures of network size and node centrality spread capture most of the variability in the network domains. The third dimension is found to have a maximum correlation with the model into which centrality fits, and the fourth dimension is correlated with the complexity of the network. The fifth dimension is highly correlated with the dispersion of clustering coefficient. This dimension gives a rough idea about how denser the network cluster is.

To obtain the proper classification of the protein interaction data from the real world database, supervised learning method is employed. A category label is explicitly provided for the networks in the design matrix, and the sum of the distances for these patterns is reduced. All the protein interaction and biogrid datasets that come under protein-protein interaction data are labeled as 'bio' and all the other domains are labeled as 'non bio'. The s-Isomap is computed. K=15 provides the largest connected component.

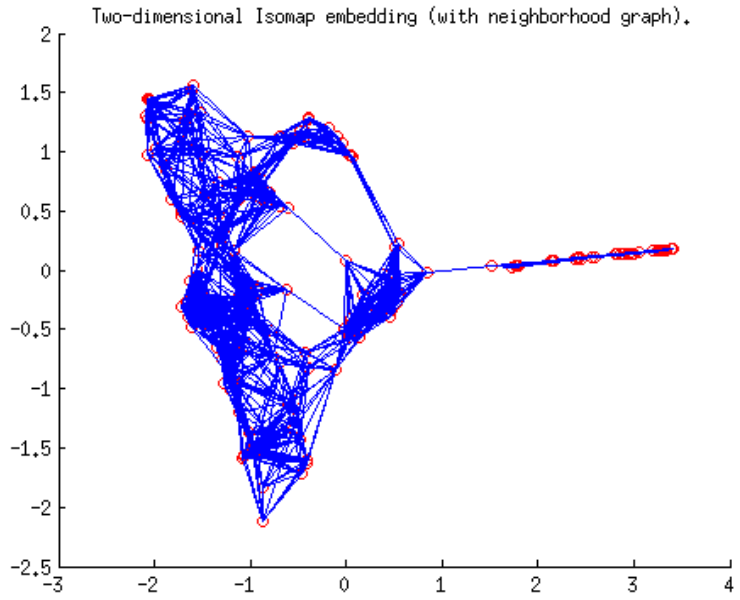


Figure 11: Two-dimensional S-Isomap embedding (with neighbourhood graph)

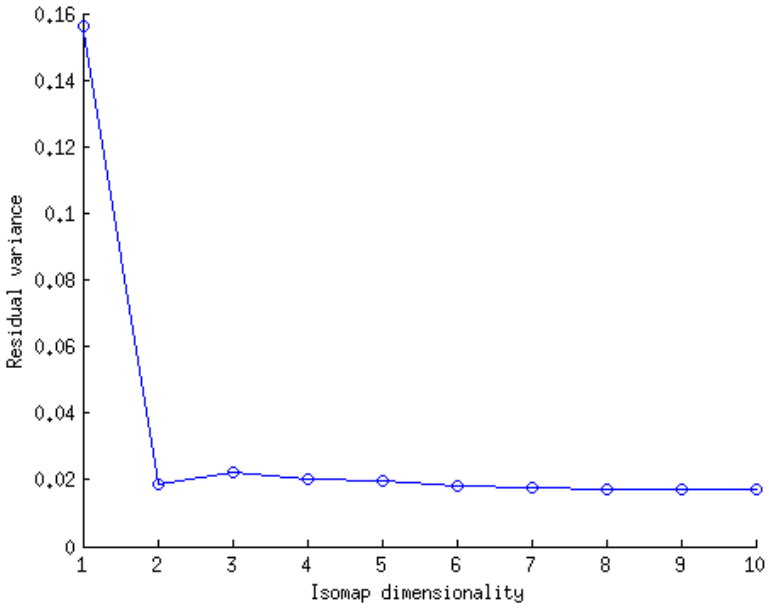
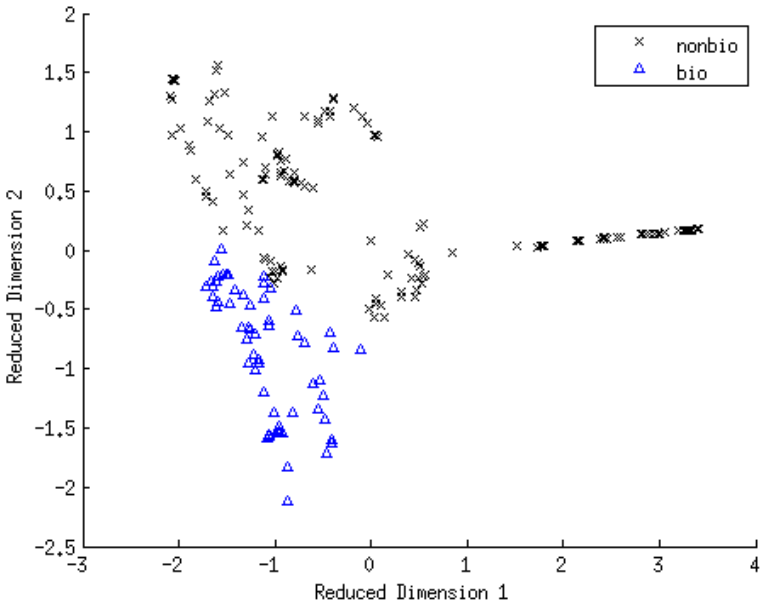


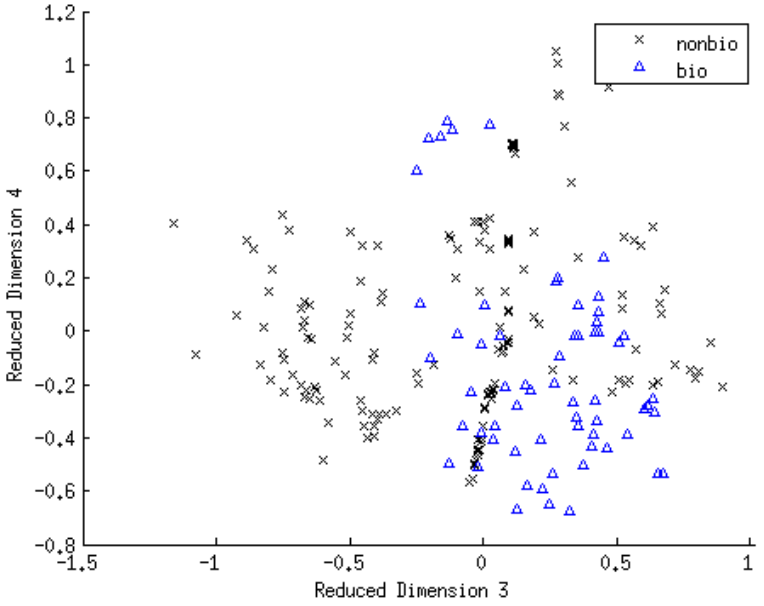
Figure 12: Residual variance as the number of S-Isomap dimensions is increased

The first two dimensions alone capture 98% of the variance. Residuality in the variance is seen to be improved to a little extent with third dimension, and the further increment in the dimensions has shown no considerable increase in the variance.



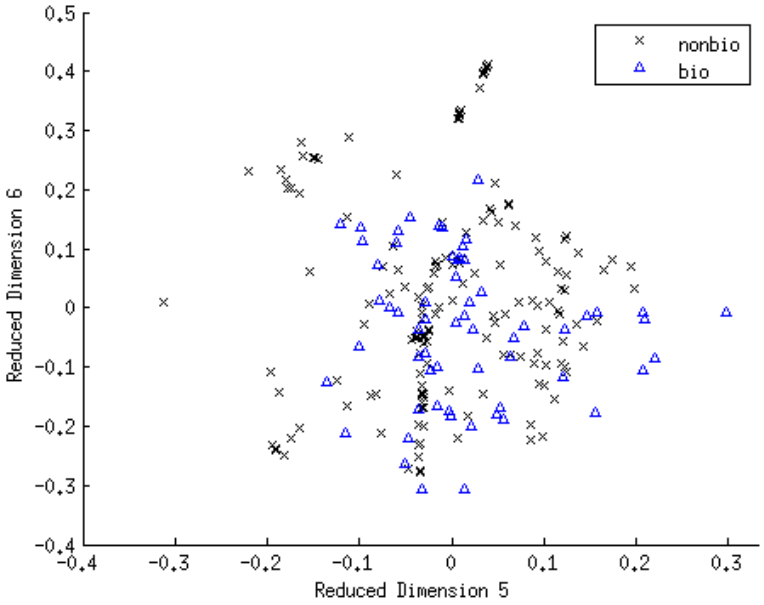
i. First two reduced dimensions

Figure 13: Network Clustering via S-Isomap dimensionality reduction



ii. Third and fourth reduced dimensions

Figure 14: Network Clustering via S-Isomap dimensionality reduction



iii. Fifth and sixth reduced dimensions

Figure 15: Network Clustering via S-Isomap dimensionality reduction

The correlation coefficients are computed as in isomap. Table 2 shows the first and second maximally correlated coefficients and the respective feature values.

Dimension	r_1	Feature-1	r_2	Feature-2
1	-0.9230	ClusteringCoeff_min	-0.8928	ClusteringCoeff_max
2	0.7515	fraction2core	-0.7302	eVectorCentrality_posrms
3	-0.5030	betweenCentrality_range	-0.5030	betweenCentrality_max
4	-0.5725	eVectorCentrality_trimean10	-0.5588	eVectorCentrality_mean
5	-0.4150	AssortativeCoefficient_snowball100	-0.4084	CyclomaticNumber

Table 2: Maximally correlated features for each reduced dimension in S-Isomap

The first reduced dimension is strongly correlated with the dispersion of the clustering coefficient. The second dimension is correlated with the fraction of networks nodes covered by the 2-core. Thus the first two dimensions indicate density of loops in the network and node degree respectively. The third dimension is highly correlated with the dispersion of the node centrality, and the fourth dimension is correlated with the importance of the node in the network. The above two dimensions give a measure of variability in the node centrality. The fifth dimension is correlated with the complexity measure of the network, and is also substantially correlated with the connectivity of the graph.

Chapter 5: Discussion

The visualization provides relative positions of different categories in the space, which increases our understanding of the intrinsic structure and distribution of real world data in different categories. From the figure 3, it is observed that biogrid and protein interaction samples overlap which is expected since both of them belong to same class. The principal components 3 and 4 also form a compact cluster of biogrid and protein interaction data, but misclassification may occur if classification is based on these features as political cosponsorship and political voting samples also overlap the protein data beyond a tolerable limit. As dimensionality is further increased, samples get mixed up in the space. From figure 8, it is seen that distance between biogrid and protein interaction samples considerably decreased by including non-linearity through Isomapping. As opposed to the previous work, the second dimension is found to be the spread of the centrality, rather than network density. Further increase in dimension in the non-linear manifolds misclassified the networks. When supervised learning is employed, the projection of samples from each category form their compact clusters. It means that differences in feature values correspond with the protein-protein interaction database. With increase in the dimensionality, residual variance increased indicating that the second dimension is the bottle neck for classification.

Chapter 6: Conclusion

I applied a wide-ranging set of diagnostics to protein interaction networks, and studied the most identified group of network metrics that leads to interesting aspects of network structure. The resulting functional outcome directs our attention to identify the signatures of complex network architecture. One limitation is that it scales poorly with the network size (for larger genomes like humans). Using sampling methods like snowball sampling helps in this regard. The comprehensive 'look-up' elucidated with the present set of protein-protein interaction networks may serve as a basis for further development in gene regulatory networks which is the main goal of this project.

References

[1] Code by Xin Geng. http://www.lamda.nju.edu.cn/code_S-Isomap.ashx/ (accessed July 2013).

[2] Code by S. Agarwal. http://www.comp-engine.org/

[3] See http://www.w3schools.com/

[4] See http://www.tutorialspoint.com/sql/

[5] See http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-094-introduction-to-matlab-january-iap-2010/index.htm/

[6] See http://www.thebiogrid.org/

[7] S. Agarwal, G. Villar, and N. S. Jones. Comparative network analysis. In preparation.

[8] M. E. J. Newman. The structure and function of complex network. *SIAM Review*, 45(2):167-256 (2003).

[9] S. Weng, C. Zhang, Z. Lin, X. Zhang. Mining the structural knowledge of high-dimensional medical data using Isomap. *Medical & Biological Engineering & Computing*, 43:1-3 (2005).

[10] V. Filkov, Z. M. Saul, S. Roy, R. M. D'Souza, and P. T. Devanbu. Modeling and verifying a broad array of network properties. *Europhysics Letters*, 86(2):28003 (2009).

[11] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75-174 (2010).

[12] S. Roy and V. Filkov. Strong associations between microbe phenotypes and their network architecture. *Physical Review* E, 80(4):040902 (2009).

[13] J. B. Tenenbaum, V. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319-2323 (2000).

[14] J.-P. Onnela, D. J. Fenn, S. Reid, M. A. Porter, P. J. Mucha, et al. Taxonomies of networks (2011). arxiv:1006.5731.

[15] S. Agarwal. Networks in Nature: Dynamics, Evolution and Modularity. Ph.D. thesis, University of Oxford (2012).

[16] A. Tsanas, M. A. Little, and P. E. McSharry. A simple filter benchmark for feature selection. *Journal of Machine Learning Research*. In review (2011).

[17] X. Geng, D. -C. Zhan, and Z. -H. Zhou. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, 35(6): 1098-1107 (2005).

[18] C. Bishop. *Pattern Recognition and Machine Learning*. Information science and statistics. Springer (2006).

[19] R. Bonneau. Learning biological networks: from modules to dynamics. *Nature Chemical Biology 4*:658-664 (2008).

[20] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662-1664 (2002).

Appendix A: List of Network Features

Here all of the diagnostics and summary statistics that were utilized in this report are listed. For each diagnostic, the short name given is that generally used to refer to it in the main text. For summary statistics, short names use a subscript to denote the summary (e.g., the maximum of the degree distribution is degreemax); shorthand summary names used in such subscripts (where applicable) are given in parentheses. The code to evaluate all these network diagnostics is obtained from Dr. Sumeet Agarwal, who has used it for his D. Phil thesis.

Short name	**Full name**
Connectivity	
degree	Degree distribution
avgNearestNeighbourDegree	Average of degrees of adjacent nodes
asortativeCoefficient	Assortative coefficient
density	Density
fractionArticulation	Fraction of articulation nodes
erosionTime	Erosion Time
dilationTime	Dilation Time
fraction2core	Fraction of vertices comprising 2-core
fraction3core	Fraction of vertices comprising 3-core
fraction4core	Fraction of vertices comprising 4-core
richClub	Rich-club index
richClubNormalised	Normalised rich-club index
Centrality	
degreeCentrality	Degree centrality
degreeCentralityGroup	Group degree centrality
betweenCentrality	Betweenness centrality
betweenCentralityGroup	Group betweenness centrality
closeness	Closeness
closenessGroup	Group closeness
evectorCentrality	Eigenvector centrality
subgraphCentrality	Subgraph centrality
subgraphCentralisation	Subgraph centralization
bipartivity	Estrada's measure of bipartivity
infoCentrality	Information centrality
infoCentralityGroup	Group information centrality
vulnerability	Vulnerability
Community	
modularity	Spectrally optimized modularity
modularityFast	Louvain optimized modularity
greedyPartitionEntropy	Entropy of Louvain partition
spectral	Newman's spectral community detection
greedyComm	Louvain community detection

Short name	**Full name**
pottsModel	Potts model community detection
infomap	Infomap community detection
Clustering	
transitivity	Transitivity
clusteringCoeff	Clustering coefficient
clustSofferGlobalMean	Global mean Soffer clustering coefficient
clustSofferLocalMean	Local mean Soffer clustering coefficient
Distance	
diameter	Graph diameter
radius	Graph radius
szegedIndex	Szeged index
cyclicCoefficient	Cyclic coefficient
geodesicDistanceMean	Mean geodesic distance
geodesicDistanceVar	Variance of geodesic distance
harmonicMeanGeoDist	Harmonic mean geodesic distance
Complexity	
cyclomaticNumber	Cyclomatic number
edgeFraction	Edge fraction
connectivity	Connectivity
logNumSpanningTrees	log(number of spanning trees)
graphIndexComplexity	Graph index complexity
mediumArticulation	Medium articulation
efficiency	Efficiency
efficiencyComplexity	Efficiency complexity
offDiagonalComplexity	Off-diagonal complexity
chromaticNumber	Chromatic number
tspl	TSP length from cross-entropy algorithm
$tspl_{ga}$	TSP length from genetic algorithm
$tspl_{sa}$	TSP length from simulated annealing
Spectral	
largestEigenvalue	Largest eigenvalue
spectralScalingDeviations	Deviations from ‘perfect spectral scaling’
algebraicConnectivity	Algebraic connectivity
algebraicConnectivityVector	Algebrain connectivity vector
fiedlerValue	Fiedler value
Statistical physics	
energy	energy
entropy	entropy
Motif	
fraction3motifs	Fraction of 3-motifs
fraction4motifs	Fraction of 4-motifs
Size	
numNodes	Number of nodes
numEdges	Number of edges
totStrength	Sum of all link weights
Model	

Short name	Full name
ergm_edges	Exponential random graph model for edges
fitPowerLawAlpha	Fitted power law exponent for degrees
fitPowerLawP	p-value of power law fit to degrees

Table 3: List of network diagnostics

Central tendency	Dispersion	Shape	Model fit log-likelihoods
Mean	Minimum (min)	Kurtosis	Normal
Geometric mean (geomean)	Maximum (max)	Skewness	Log-normal
Harmonic mean (harmmean)	Variance (var)		Exponential
Mean excluding 10% outliers(trimmean10)	Range		Extreme value
	Inter-quartile range (iqr)		Gamma
RMS of positive values (posrms)	Mean absolute deviation (meanad)		Weibull (wbl)
RMS of negative values (negrms)	Median absolute deviation (medad)		

Table 4: List of distribution summary statistics

Appendix B: Set of 192-Real World Networks

The set of 192 real world networks used in this project was obtained from Dr. Sumeet Agarwal, who has used it for his D. Phil thesis.

Name	Category
Human brain cortex: participant A1	Brain
Human brain cortex: participant A2	Brain
Human brain cortex: participant B	Brain
Human brain cortex: participant D	Brain
Human brain cortex: participant E	Brain
Human brain cortex: participant C	Brain
Cat brain: cortical	Brain
Cat brain: cortical/thalmic	Brain
Macaque brain: cortical	Brain
Macaque brain: visual/sensory cortex Brain	Brain
Macaque brain: visual cortex 1	Brain
Macaque brain: visual cortex 2	Brain
Co-authorship: astrophysics	Collaboration
Co-authorship: comp. geometry	Collaboration
Co-authorship: condensed matter	Collaboration
Co-authorship: Erdos	Collaboration
Co-authorship: high energy theory	Collaboration
Co-authorship: network science	Collaboration
Hollywood film music	Collaboration
Jazz collaboration	Collaboration
Facebook: Caltech	Facebook
Facebook: Cornell	Facebook
Facebook: Dartmouth	Facebook
Facebook: Georgetown	Facebook
Facebook: Harvard	Facebook
Facebook: Indiana	Facebook
Facebook: MIT	Facebook
Facebook: NYU	Facebook
Facebook: Oklahoma	Facebook
Facebook: Texas80	Facebook
Facebook: Trinity	Facebook
Facebook: UCSD	Facebook
Facebook: UNC	Facebook
Facebook: USF	Facebook
Facebook: Wesleyan	Facebook
NYSE: 1980-1999	Financial
NYSE: 1980-1983	Financial

Name	Category
NYSE: 1984-1987	Financial
NYSE: 1988-1991	Financial
NYSE: 1992-1995	Financial
NYSE: 1996-1999	Financial
Phanerochaete velutina control11-2	Fungal
Phanerochaete velutina control11-5	Fungal
Phanerochaete velutina control11-8	Fungal
Phanerochaete velutina control11-11	Fungal
Phanerochaete velutina control17-2	Fungal
Phanerochaete velutina control17-5	Fungal
Phanerochaete velutina control17-8	Fungal
Phanerochaete velutina control17-11	Fungal
Phanerochaete velutina control4-2	Fungal
Phanerochaete velutina control4-5	Fungal
Phanerochaete velutina control4-8	Fungal
Phanerochaete velutina control4-11	Fungal
Online Dictionary of Computing	Language
Online Dictionary Of Information Science	Language
Reuters 9/11 news	Language
Roget's thesaurus	Language
Word adjacency: English	Language
Word adjacency: French	Language
Word adjacency: Japanese	Language
Word adjacency: Spanish	Language
Metabolic: CE	Metabolic
Metabolic: CL	Metabolic
Metabolic: CQ	Metabolic
Metabolic: CT	Metabolic
Metabolic: DR	Metabolic
Metabolic: HI	Metabolic
Metabolic: NM	Metabolic
Metabolic: OS	Metabolic
Metabolic: PA	Metabolic
Metabolic: PG	Metabolic
Metabolic: PH	Metabolic
Metabolic: PN	Metabolic
Metabolic: SC	Metabolic
Metabolic: ST	Metabolic
Metabolic: TP	Metabolic
Bill cosponsorship: U.S. House 96	Political: cosponsorship
Bill cosponsorship: U.S. House 97	Political: cosponsorship
Bill cosponsorship: U.S. House 98	Political: cosponsorship
Bill cosponsorship: U.S. House 99	Political: cosponsorship
Bill cosponsorship: U.S. House 100	Political: cosponsorship
Bill cosponsorship: U.S. House 101	Political: cosponsorship
Bill cosponsorship: U.S. House 102	Political: cosponsorship

Name	Category
Bill cosponsorship: U.S. House 103	Political: cosponsorship
Bill cosponsorship: U.S. House 104	Political: cosponsorship
Bill cosponsorship: U.S. House 105	Political: cosponsorship
Bill cosponsorship: U.S. House 106	Political: cosponsorship
Bill cosponsorship: U.S. House 107	Political: cosponsorship
Bill cosponsorship: U.S. House 108	Political: cosponsorship
Bill cosponsorship: U.S. Senate 96	Political: cosponsorship
Bill cosponsorship: U.S. Senate 97	Political: cosponsorship
Bill cosponsorship: U.S. Senate 98	Political: cosponsorship
Bill cosponsorship: U.S. Senate 99	Political: cosponsorship
Bill cosponsorship: U.S. Senate 100	Political: cosponsorship
Bill cosponsorship: U.S. Senate 101	Political: cosponsorship
Bill cosponsorship: U.S. Senate 102	Political: cosponsorship
Bill cosponsorship: U.S. Senate 103	Political: cosponsorship
Bill cosponsorship: U.S. Senate 104	Political: cosponsorship
Bill cosponsorship: U.S. Senate 105	Political: cosponsorship
Bill cosponsorship: U.S. Senate 106	Political: cosponsorship
Bill cosponsorship: U.S. Senate 107	Political: cosponsorship
Bill cosponsorship: U.S. Senate 108	Political: cosponsorship
Committees: U.S. House 101, comms.	Political: committee
Committees: U.S. House 102, comms.	Political: committee
Committees: U.S. House 103, comms.	Political: committee
Committees: U.S. House 104, comms.	Political: committee
Committees: U.S. House 105, comms.	Political: committee
Committees: U.S. House 106, comms.	Political: committee
Committees: U.S. House 107, comms.	Political: committee
Committees: U.S. House 108, comms.	Political: committee
Committees: U.S. House 101, Reps.	Political: committee
Committees: U.S. House 102, Reps.	Political: committee
Committees: U.S. House 103, Reps.	Political: committee
Committees: U.S. House 104, Reps.	Political: committee
Committees: U.S. House 105, Reps.	Political: committee
Committees: U.S. House 106, Reps.	Political: committee
Committees: U.S. House 107, Reps.	Political: committee
Committees: U.S. House 108, Reps.	Political: committee
Roll call: U.S. House 101	Political: voting
Roll call: U.S. House 102	Political: voting
Roll call: U.S. House 103	Political: voting
Roll call: U.S. House 104	Political: voting
Roll call: U.S. House 105	Political: voting
Roll call: U.S. House 106	Political: voting
Roll call: U.S. House 107	Political: voting
Roll call: U.S. House 108	Political: voting
Roll call: U.S. Senate 101	Political: voting
Roll call: U.S. Senate 102	Political: voting
Roll call: U.S. Senate 103	Political: voting

Name	Category
Roll call: U.S. Senate 104	Political: voting
Roll call: U.S. Senate 105	Political: voting
Roll call: U.S. Senate 106	Political: voting
Roll call: U.S. Senate 107	Political: voting
Roll call: U.S. Senate 108	Political: voting
U.K. House of Commons voting: 1992-1997	Political: voting
U.K. House of Commons voting: 1997-2001	Political: voting
U.K. House of Commons voting: 2001-2005	Political: voting
U.N. resolutions 59	Political: voting
U.N. resolutions 60	Political: voting
U.N. resolutions 61	Political: voting
U.N. resolutions 62	Political: voting
Biogrid: A. thaliana	Protein interaction
Biogrid: C. elegans	Protein interaction
Biogrid: D. melanogaster	Protein interaction
Biogrid: H. sapiens	Protein interaction
Biogrid: M. musculus	Protein interaction
Biogrid: R. norvegicus	Protein interaction
Biogrid: S. cerevisiae	Protein interaction
Biogrid: S. pombe	Protein interaction
DIP: H. pylori	Protein interaction
DIP: H. sapiens	Protein interaction
DIP: M. musculus	Protein interaction
DIP: C. elegans	Protein interaction
Human: CCSB	Protein interaction
Human: OPHID	Protein interaction
Protein: serine protease inhibitor (1EAW)	Protein interaction
Protein: immunoglobulin (1A4J)	Protein interaction
Protein: oxidoreductase (1AOR)	Protein interaction
STRING: C. elegans	Protein interaction
STRING: S. cerevisiae	Protein interaction
Yeast: Oxford Statistics	Protein interaction
Yeast: DIP	Protein interaction
Yeast: DIPC	Protein interaction
Yeast: FHC	Protein interaction
Yeast: FYI	Protein interaction
Yeast: PCA	Protein interaction
Corporate directors in Scotland (1904-1905)	Social
Corporate ownership (EVA)	Social
Dolphins	Social
Family planning in Korea	Social
Unionization in a hi-tech firm	Social
Communication within a sawmill on strike	Social
Leadership course	Social
Les Miserables	Social
Marvel comics	Social

Name	Category
Mexican political elite	Social
Pretty-good-privacy algorithm users	Social
Prisoners	Social
Bernard and Killworth fraternity: observed	Social
Bernard and Killworth fraternity: recalled	Social
Bernard and Killworth HAM radio: observed	Social
Bernard and Killworth HAM radio: recalled	Social
Bernard and Killworth office: observed	Social
Bernard and Killworth office: recalled	Social
Bernard and Killworth technical: observed	Social
Bernard and Killworth technical: recalled	Social
Kapferer tailor shop: instrumental (t1)	Social
Kapferer tailor shop: instrumental (t2)	Social
Kapferer tailor shop: associational (t1)	Social
Kapferer tailor shop: associational (t2)	Social
University Rovira i Virgili (Tarragona) e-mail	Social
Zachary karate club	Social

Table 5: List of 192-real world networks

Appendix C: Set of 42-Biogrid Networks

The set of 42 protein interaction networks used in this project was obtained from open source [6].

Name	Category
Anopheles_gambiae	Protein interaction
Arabidopsis_thaliana	Protein interaction
Aspergillus_nidulans	Protein interaction
Bacillus_subtilis	Protein interaction
Bos_taurus	Protein interaction
Caenorhabditis_elegans	Protein interaction
Candida_albicans_SC5314	Protein interaction
Canis_familiaris	Protein interaction
Cavia_porcellus	Protein interaction
Chlamydomonas_reinhardtii	Protein interaction
Cricetulus_griseus	Protein interaction
Danio_rerio	Protein interaction
Dictyostelium_discoideum_AX4	Protein interaction
Drosophila_melanogaster	Protein interaction
Equus_caballus	Protein interaction
Escherichia_coli	Protein interaction
Gallus_gallus	Protein interaction
Hepatitus_C_Virus	Protein interaction
Homo_sapiens	Protein interaction
Human_Herpesvirus_1	Protein interaction
Human_Herpesvirus_2	Protein interaction
Human_Herpesvirus_3	Protein interaction
Human_Herpesvirus_4	Protein interaction
Human_Herpesvirus_5	Protein interaction
Human_Herpesvirus_6	Protein interaction
Human_Herpesvirus_8	Protein interaction
Human_Immunodeficiency_Virus_1	Protein interaction
Human_Immunodeficiency_Virus_2	Protein interaction
Leishmania_major	Protein interaction
Macaca_mulatta	Protein interaction
Mus_musculus	Protein interaction
Neurospora_crassa	Protein interaction
Oryctolagus_cuniculus	Protein interaction
Oryza_sativa	Protein interaction
Pan_troglodytes	Protein interaction
Plasmodium_falciparum_3D7	Protein interaction
Rattus_norvegicus	Protein interaction
Ricinus_communis	Protein interaction
Saccharomyces_cerevisiae	Protein interaction

Name	Category
Schizosaccharomyces_pombe	Protein interaction
Simian-Human_Immunodeficiency_Virus	Protein interaction
Sus_scrofa	Protein interaction
Ustilago_maydis	Protein interaction
Xenopus_laevis	Protein interaction
Zea_mays	Protein interaction

Table 6: List of 42-biogrid networks