INTERACTIONS IN A DRUG-TARGET NETWORK

Varsha S. Kulkarni

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
□Doctor of Philosophy□
in the School of Informatics and Computing,□
Indiana University□
December 2016

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

David J. Wild, PhD

_____

Gerardo Ortiz, PhD

_____

John Beggs, PhD

_____

Y.Y. Ahn, PhD

April 18, 2016

To

my father

Suresh N. Kulkarni

for introducing me to the world of learning

Thank you *to my loved ones to have nurtured the ethos*

that academic pursuits promote intellect and excellence in a humanistic way.

Acknowledgments

I am very thankful to Prof. David Wild for his encouragement of my work and efforts. Completion of a doctoral dissertation is interesting and often a challenging process. Prof. David Wild's advice, kindness and support helped me reach the finish line.

I am grateful to Profs. Gerardo Ortiz, John Beggs, Y.Y. Ahn for helpful discussions and comments on this dissertation. I thank Prof. Ying Ding for discussions and all those staff, students, and faculty at Informatics and Computing who have helped me in this program. I gratefully acknowledge the financial support of my research through assistantships, fellowships, and instructorships at Indiana University.

I also thank Prof. Anatole Beck of the University of Wisconsin-Madison for his encouragement, and for the confidence he showed in my capability.

In addition, I thank the professors and scholars at all the institutions I have studied. Collaborations and discussions with them have significantly contributed to my academic advancement.

Varsha S. Kulkarni

INTERACTIONS IN A DRUG-TARGET NETWORK

Highly chemically similar drugs usually possess similar biological activities but small changes in chemistry result in large differences in biological effects. Chemically similar drug pairs showing extreme deviations in activity represent distinctive drug interactions. The presence of these interactions adversely affects prediction of structure and activity associations. Their identification has crucial implications on drug development and innovations. Given the multitude of drugs in an ensemble, pairs possess multilevel distinctiveness in terms of their attributes of structural and activity similarity or variation. The cliff characterization for describing drops in similar activity has received considerable attention, however, it remains quantitatively less refined. In this dissertation, I investigate distinctiveness of drug interactions using a large drug-target network and provide a quantitative rationale for characterization of the pharmacological topography. I consider rises in pairwise similarity and variation in activity of drugs on proteins with chemical similarity (*c*) to assess levels of distinctiveness. These activity measures are affected by the presence of few drugs (targets) having multiple targets (drugs). I quantify interactions between drugs by considering similarity and variation jointly with *c*. The probability of distinctiveness is predicted by employing joint probability of structure and activity measures. Intermittent spikes in variation along the axis of *c* represent canyons in the activity landscape. This new representation accounts for distinctiveness through relative rises in activity measures and offers an enhanced perspective. It provides a mathematical basis for predicting the probability of occurrence of distinctiveness. It identifies the drug pairs at varying levels of distinctiveness and non-distinctiveness. Prediction is validated even if data approximately satisfy the conditions of the formulation. The difference in distinctive interactions emphasizes the importance of studying both measures, and reveals that the choice of measurement can affect the interpretation.

Further, I find that minor changes in methods or perturbations of measures can crucially alter the classification of interactions as distinctive. Identification and interpretation of distinctiveness, therefore, gain relevance through methodological specifications. The present analysis of structure and activity provides an in depth modeling and assessment of distinctiveness and the probability of its occurrence. It could potentially influence decision-making in research and development.

_____

David J. Wild, PhD

_____

Gerardo Ortiz, PhD

_____

John Beggs, PhD

_____

Y.Y. Ahn, PhD

Contents

Chapters

Chapter 1

INTRODUCTION

I. Context

The field of drug discovery primarily constitutes creation of new medicinal compounds having targeted therapeutic impacts for improvement of human health. It entails identification of optimum combination of physicochemical properties that can cure diseases and ailments of diverse nature observed. The drugs are prepared using the physicochemical properties specified according to [1,2] a procedure that emphasizes the importance of select properties required for the medicinal compounds. However, with the development of an increasing number of drugs, the process seems to have favored few properties. A study conducted on a group of chemists showed that majority of chemists based their strategy on two or three properties if not less [1]. This bias that magnifies with time may in part be attributed to the cumulative discretionary decisions of the drug manufacturers, scientists and chemists involved in the process.

This complexity in drug discovery gives rise to a variety of interactions or associations between pairs of drugs. In a system of drugs, an interaction between two drugs is a *connection,* which usually refers to the similarity or dissimilarity of their *activities* on protein targets. In other words, two drugs are connected depending on whether they have a common or an uncommon target [3,4]. However, there is a dimension to an interaction between drugs in addition to the one based on their activities. Pairs of drugs can also be connected through the level of similarity (or dissimilarity) in their *chemical* or structural properties [5]. These two dimensions of pairwise interactions of drugs together account for a various structure-activity associations observed between drugs. These associations are crucial for understanding drug development and studied extensively by researchers [1,2,5,6]. Consider that a person is looking to cure a common ailment and A is the standard medicine s/he is aware of. Someone

claims that medicine B is as good as A. [7-12] Now the person has two alternatives[1]. Further inspection of the chemical constituents of A and B shows little variation. Hence, A and B are chemically analogous and it is not surprising that they are substitutes having similar therapeutic actions. S/he finds another medicine C which is chemically analogous to A and B. It should be obvious then that there are three alternatives for the cure[2]. However, it turns out that C has a therapeutic action different from A and B. While the connection between A and B is a usual or expected, the one between A and C or B and C is unusual.

These kinds of connections between pairs of drugs that possess similar chemical compositions but highly variant physiological activities often occur in drug development. Extreme deviations in activity of pairs of chemically similar drugs describe *distinctive* connections or interactions between pairs of drugs. Distinctive interactions occur between pairs of drugs with high chemical similarity and activity variation, or, high chemical similarity and low activity similarity. The presence of these kinds of interactions is known to adversely affect the prediction of activity changes of drugs caused by changes in their chemical structures. Identification of such interactions can significantly improve drug development. However, given the multitude of drugs in the ensemble, pairwise associations potentially have multilevel attributes (of structural and activity similarity or variation). Thus *distinctiveness* can be found at varying levels, which makes it challenging to identify the distinctive interactions. This dissertation focuses on selective identification of distinctive interactions by employing rigorous quantitative techniques. It attempts to understand the extent to which this classification of distinctive versus non-distinctive depends on the choice of measurement. It would help in knowing the

[1] The price of drugs can be a determinant of the drugs chosen by individuals [7,8] and there are socio-economic considerations on the distribution of health care, medicines that would affect their usage [9-11]. Further, an increased availability of some chemical combinations in the market decreases the effectiveness of drugs prepared for limited, advanced treatments. Thus, financial incentive of the drug industry can potentially affect the health outcomes.

[2] This decision is further dependent on the growth of medical innovations in the society [10]. The individuals' choice of new health products is often influenced by and dependent on certain aspects of the social network and the properties of the innovation [12].

significance of the classification in the context of the methodology and govern the quantitative assessments, and the health outcomes.

II. Previous research

The kinds of connections between drugs described above as usual, are those interactions that conform with the similarity principle in medicinal chemistry [13]. According to this principle, a pair of structurally similar drug compounds tends to behave similarly on proteins. This, however, is not indicative of their behaviors on proteins in general. The proteins targeted by a pair of structurally similar drugs may not necessarily be similar in their chemical composition. Distinctive interactions refer to those pairs of drugs that do *not* conform with the similarity principle. This principle is of vital significance because pharmacologists are constantly in search of variants of lead substances to make drugs that do not vary in behavior, cure the same kinds of ailments or are active on the same proteins [1-6, 13, 14]. This achieves substitution with little cost particularly for substitutes not very different from the original compound. Further, it affects the pricing of the variants or substitutes in the market. If a small structural change in the drug achieves a similar therapeutic impact, it comes at low cost, and that would allow a substitute to be introduced at a similar price. On the contrary, if one has to widen the search in the chemical space to find a variant in order to achieve a similar functionality, the typical cost incurred is high.

This can also work in another way. If diversity between chemical compounds is limited, then according to similarity principle, many substitutes so generated would add neither to the knowledge base nor to the range of therapeutic outcomes. Thus, it may be a difficult decision because a small and simple modification tends to churn out a vast number of substances that can be possibly derived. A small change in chemical space resulting in substantially different functional behavior is also helpful, from the perspectives of both drug discovery and marketability. For instance, if a structural analog

produced has many targets, it can be helpful for increasing drug applicability[3] [1,3,4]. Thus, structure activity analysis is crucial for understanding pharmacological outcomes. It is an important tool for predicting how the non-synthesized drugs would act on proteins based on the knowledge of functional behaviors of a given ensemble of drugs. In fact, the decisions surrounding the identification of the magnitude of change in chemical structure required to produce a desired change in functionalities, are still made informally by pharmacologists using their experience and knowledge [1, 15]. A *quantitative* analysis of structure and activity is meant to assist in making rigorous and accurate decisions and experimental design.

The functionality of a drug indicates its physiological activity (on a protein, for instance), if it is active or inactive. Structural properties of compounds are governed by the physicochemical properties that determine that they would to be active on certain proteins and inactive on others. The activity may be measured as success in a drug's ability to bind with a protein or cure a disease or affect genes, among others. Structure activity relationship [1,2,14] is formalized as the modification in physiological activity of a drug resulting from a modification in its chemical constitution or chemical structure. It can be written as $\Delta(physiological\ activity) = f(\Delta(chemical\ structure))$.

The quantitative structure activity analysis uses a training data sample from which one can extract information on changes in activity caused by changes in structure. The patterns obtained from the analysis are then applied to predict the activity outcomes of a test sample of data on drugs and proteins [1,14,16]. As this involves analysis of the data, a suitable application of statistical and mathematical techniques becomes critical. The development of sound and effective quantitative structure activity models for the analysis and prediction in diverse data samples is a major area of research in the fields of cheminformatics, computational chemistry, medicinal chemistry, drug discovery. These models are designed for investigating the space of interactions of drugs through changes in structure and

[3] This implies that existing drugs can be profiled for curing diseases in addition to those they are intended to treat. The marginal effect of a chemical innovation on the biological activity can vary in a wide range, and the large effects can lead to inventions in medicine as the entire chemical space is explored.

corresponding activities. The space of interactions defined by the structures and activities of drugs is an *activity landscape*. The activity landscape is a two dimensional space on which drug interactions may be mapped as the similarity in physiological activity between two drugs with the corresponding similarity in their chemical structure or composition. Thus the interactions between a pair of drugs may be quantified in terms of two properties or quantities- chemical or structural similarity and similarity (or variation) in their functionalities. In the activity landscape, researchers have identified regions along the chemical space where there is a huge drop in similar activity. The extreme deviations in activity observed for pairs of drugs that are highly chemically similar are represented as discontinuities in the landscape. These interactions have been quantified as activity cliffs [7,8] in previous research. Their presence hinders the prediction of quantitative models and therefore it is important to identify them and the regions of the landscape where they are observed.

III. Toward a new characterization of the activity landscape

Are pairs of structurally analogous drugs similar in their activities or not? Which drug interactions are distinctive and which of them conform with the similarity principle? How probable is their occurrence in a given data sample? This forms the basis for studying multilevel interaction effects in an ensemble of drugs, which I focus on in this dissertation. Structure activity analysis requires two properties for all pairwise interactions. It may either condition the functionality based on a constant structural similarity or condition the structural similarity based on constant functionalities. Therefore, all the pairwise interactions can be quantified at multiple levels of each of the two attributes (i) chemical similarity (c) and (ii) similarity or variation in activity. Similarity (s) and variation (d) in activity of a pair of drugs are measured respectively as the number of common and uncommon targets of theirs. Here, I use *data* on biological *activity* of 1354 drugs on 1596 proteins. These are drugs listed in drugbank database. The information is stored as a bipartite graph. In this graph, connections are between two types of vertices. A link from a drug to a protein indicates that the protein is an active target of that drug. The information on activity is entered in a 1354×1596 adjacency matrix. The existence of a connection

implies activity and the corresponding element of the matrix is 1 whereas absence of a connection implies inactivity with corresponding entry of 0. Targets of a drug are the proteins that it is active on. The binary information of a drug's activity on all proteins stored in the form of an array constitutes the activity profile of the drug. Next, I obtain data on pairwise *chemical similarity* of all the drugs in the ensemble. The chemical structure of a drug compound is based on molecular combinations having physicochemical properties. The structural representations of compounds are encoded by descriptors of these properties. The descriptors are in the form of binary strings, wherein the presence and absence of a feature are given by 1 and 0 respectively. The degree of structural resemblance of a pair of drugs is quantified by a coefficient of chemical similarity (c) of their strings. Chemical similarity is computed using the widely applied Jaccard coefficient as this has been shown to be effective.

Activity cliffs are formed by pairs having both high c *and* low similarity in their activity values [17,18]. On the landscape it is expected that as c increases, similarity in physiological activity would also increase. Interestingly, some pairwise drug interactions deviate from this behavior. Activity cliffs are commonly observed in almost all data collections on structure and activity of drugs. Researchers have focused particularly on using quantitative models to account for such aberrations [14-18]. However, they consider s and d as complementary measures of activity. Cliffs are quantitatively characterized using differences in activity values and the chemical similarity of drugs. In this dissertation, I use two measures of activity, s and d (number of common and uncommon targets) and consider each of these jointly with c for every interaction. These measures quantify the interacting system approach. I find intermittent rises in both measures, particularly, d in the landscape. This leads to a *canyon* characterization of the landscape. The distinctive rises in d and s along the chemical space represent distinctive and non-distinctive interactions at varying levels. Further, in this new representation the level of distinctiveness exhibited by a pair can vary according to multiple levels of c in combination with those of d or s. I compute the probability of distinctiveness using joint probability of finding a pair with particular levels of attributes c, s, d. This can be used to predict the occurrence of distinctive interactions in the landscape.

IV. Main Contributions

This analysis verifies the presence of multilevel distinctiveness using a large collection of drugs through the distribution of pairs in various levels of (combinations of) *c* and *s*, *c* and *d*. I quantify aggregate (non) distinctiveness with the distribution of pairs in various levels of *c* and *s* or *d*. Quantitative modeling has advanced the structure activity analysis significantly in drug discovery but there has not been a mathematical treatment for accurate interpretation of regularities and irregularities in the landscape. Studies have relied on less precisely constructed measures for quantification of cliffs in the landscape. In fact, the existing quantitative structure activity analysis in this field is insufficient for analyzing cliffs and some of the methods applied are not well defined. The effort to increase the level of sophistication and technicality in quantitative assessment of the landscape brings to fore the lack of a precise specification of the abrupt deviations in activity or the distinctiveness. In this dissertation, I give a rationale for defining distinctive interactions and predicting the probability of their occurrence. The formulation here consists of predicting the probability of distinctiveness of a pair of drugs using the probability distributions of *c*, *s*, *d*. This approach facilitates identification of distinctiveness using both measures *s* and *d* and prediction of how it varies in the landscape. An aim of the dissertation is to introduce the subjectivity inherent in the quantitative interpretation of distinctiveness. It is intriguing to find little consensus on the interpretation yielded by similarity and variation, and different scales of their measurements. The canyon characterization offers an accurate classification of distinctiveness in relation to the technique applied. This would facilitate efficacious use of drugs for specified treatments and achieve improved health of the society.

A challenge to the analysis of structure and activity in drug discovery is that the magnitude of the data available [19] affects the inferences. Differences in sizes of the data samples are known to affect estimation and hence the interpretations. This can significantly distort the inferences about activity cliffs and distinctiveness in the activity landscape. For instance, the power law pattern of probability distributions of number of targets per drug and drugs per target has been verified for drug-target data

[3, 20]. Power law behavior indicates heterogeneous nature, that some drugs (proteins) possess (are targets of) many more targets (drugs) than average. The drug target data studied in this dissertation also conform with this behavior. However, different data samples may display variations in this behavior and the extent of heterogeneity. This can affect the pairwise measurements *s* and *d* and their distributions. It is, therefore, worthwhile to find that the predicted probability formulation derived in this thesis is valid for data that may not be exactly identical in nature to the sample used for its construction. This is the main merit of the approach. It is based on multilevel interaction effects of *c*, *s*, *d* and gives the probability of distinctiveness for every pair of drugs.

Hence, I show that an accurate interpretation of the landscape for distinctive interactions is in relation to the methodology used to analyze it. While structure activity relationships have been quantified using advanced statistical techniques, this aspect of the subjectivity of landscape interpretation and its implications for identifying distinctive interactions has largely been ignored in literature. Studying both measures *s* and *d* enhances our perspective on this problem while contributing to the adaptive meanings of distinctiveness. Few drug interactions, however, would remain consistently distinctive but a lot of them may be susceptible to a methodological modification, and this should be considered. This study recommends a targeted choice of methods that is consistent with the nature of distinctiveness being addressed by a researcher in a particular context.

V. Outline

The rest of the dissertation is organized in four chapters. **Chapter 2** provides a detailed mathematical analysis of the bipartite graph of drugs and protein-targets. It shows distributions of drugs per target and targets per drug. I also calculate the probability distribution of number of drugs that a drug has at least one common target with. These findings help to assess how a few multi-target drugs can possibly govern the measures of interaction that I subsequently analyze in this work. I begin the study of distinctive interactions using the network of drugs. **Chapter 3** extends the network analysis to understand distinctive drug-drug interactions using weighted measures of *c*, *s*, *d*. I study how

connections between pairs of drugs are distributed in multiple levels of distinctiveness, while pointing out the effect of employing *s* and *d* measures on the interpretation. I conclude this chapter with inferences on distinctiveness based on the measures used as compared to previously established research and a discussion of the limitations of the previous analyses. In **chapter 4**, I introduce the formulation of predicted probability of distinctiveness in the landscape using a canyon characterization. I enunciate the merits of this characterization (which implements distinctiveness through both activity measures) and the probabilistic analysis which accurately predicts the occurrence of distinctive interactions at various levels. I identify the distinctive interactions using both measures and discuss the differences and the central contributions. I end the dissertation in **chapter 5** giving an auxiliary analysis of the data using some information theoretic measures. This is another form of comparing the effectiveness of activity measures for classification of interactions as distinctive, the inherent uncertainty. It shows a possible ranking scheme of rareness of distinctive interactions based on probabilities of occurrence of the levels of interaction attributes *c*, *s*, *d*. Finally, in the conclusion of this thesis, I reflect on the main themes emerging from this analysis. I discuss how this analysis can enhance the understanding on distinctiveness in the activity landscape, thereby contributing to pharmacological practice and research.

Chapter 2

CONNECTIONS IN A DRUG NETWORK

A network is a graph of vertices and links connecting them. This thesis focuses on the evaluation of connections between drugs. A *network* of drugs as vertices forms a useful tool for this study of *interactions* between drugs. This provides the perspective of an interacting system. In a bipartite graph, there are links connecting vertices of two types. The bipartite graph of drugs and protein targets has links from drugs to proteins signifying that the drugs are active on those proteins. This information on activity can be used to construct a network of drugs, so that a pair of drugs is connected if both drugs have one or more common targets. Alternatively, a network can be constructed where pairs of drugs are connected because they have uncommon targets. Thus, two drugs can interact because of their common or uncommon activity. It is also known that all pairs of drugs are connected through their structural similarities. Pairwise associations of activity and structure of drugs reflect the structure activity relationship. In this chapter, I give a preliminary drug network analysis and study how it contributes to our understanding of distinctive interactions.

2.1 Data

I use the information on activity of N =1354 drugs on N' =1596 protein targets. The activity information is binary, $A_{i(k)}$=1 if a drug i is active on protein k and 0 if it is inactive (i = 1,2…,1354; k = 1,2,….1596). This constitutes the activity profile of drug i. I use chemical similarity of all pairs of drugs. As explained above, structural representation of a drug is in the form of physicochemical descriptors encoded in a binary string. The descriptors chosen here are in accordance with extended connectivity fingerprints (ECFP) [1,2,5,6,21]. This is because ECFP descriptors include a wide range of features. The similarity between the strings of two drugs is computed by employing the Jaccard coefficient. The chemical similarity of all pairs of drugs specifying the weight of the link between two

drugs is provided in C as $N \times N$ weighted adjacency matrix. The elements of the matrix, C_{ij} give the chemical similarity of the pair of drugs i, j and $0 \leq C_{ij} \leq 1$.

2.2 Network of Drugs and Protein Targets

The connections between pairs of drugs in a network are formed according to whether or not they have any common or uncommon targets in their profiles. Therefore, it is useful to observe the distributions of number of drugs per target and number of targets per drug. This is because if there are drugs with many targets, they may form connections with many drugs through commonality of a target. On the other hand, if there is a target on which many drugs are active, it reflects the presence of many pairwise connections. Therefore, drugs having many (few) connections in the network indicate increase (decrease) in similar activity, and if their connections are highly chemically similar, these interactions (do not) conform to the similarity principle. If connections between drugs are formed through the non-commonality of their targets, then drugs having many (few) connections in the network indicate decrease (increase) in similar activity, and if their connections are highly chemically similar, these interactions are (non) distinctive. The probability patterns of drugs per target and targets per drug are plotted in Figs.2.1, 2.2 on a logarithmic scale. Pareto power law behavior is evident, $P(k) \sim k^{-\gamma}$, k being the number of targets per drug or drugs per target[4]. $\gamma \approx 2$ implies possibility of a very high variance in sizes of activity profiles [22, 23], however, the exponent may be different for different samples of data analyzed. The presence of power law implies heterogeneity in terms of [3,19,20,22] how drugs bind to proteins and proteins react to drugs. This means that some drugs tend to be active on many more proteins than average and few targets are acted on by many more drugs than average. Multi-target drugs are beneficial as they can treat ailments of diverse nature [1].

[4] Power law behavior is expressed as $P(k) \sim k^{-\gamma}$. This can be approximated by the equation, $\log P \sim -\gamma \log(k)$. The coefficient of regression line on the log-log plot of $\log P$ versus $\log(k)$ gives the power law exponent γ.

Mathematical research on networks has proved some useful quantities for understanding network structure in the context of both unipartite and bipartite random graphs with arbitrary degree distribution [22, 23]. These graphs are considered random in the sense that the degrees of all vertices are independent and identically distributed. In the present drug-target network of N drugs and N' proteins, if a drug is active on an average of μ proteins and a protein has on an average ν drugs functionally active on it, then

$$\frac{\mu}{N'} = \frac{\nu}{N} \qquad (2.1).$$

It is observed in the data that $\mu \approx 3.77, \nu \approx 3.2$. From Figs.2.1, 2.2, the overall behavior of drug activity and protein response may be formalized as the normalized probability of a drug being active on (a protein responding to) k_p proteins (k_d drugs) which is

$$P(k_p) \approx \frac{k_p^{-\gamma_1} e^{-\frac{k_p}{\kappa_1}}}{Li_{\gamma_1}\left(e^{-\frac{1}{\kappa_1}}\right)} \qquad (2.2a)$$

$$P(k_d) \approx \frac{k_d^{-\gamma_2} e^{-\frac{k_d}{\kappa}}}{Li_{\gamma_2}\left(e^{-\frac{1}{\kappa_2}}\right)} \qquad (2.2b).$$

In both equations (2.2a,b), the constants κ_1, κ_2 denote the exponential cut-off parameters[5]. $Li_\gamma(x)$ is the polylogarithm function and γ_1, γ_2 are the power law exponents computed from the data.

I analyze the characteristics of the drug network wherein two drugs are connected or linked if they are jointly active on a protein. Theoretically, this is done using the formulations of generating functions established in previous research. The results derived using these are analyzed for certain specific kinds

[5] The Eqs. 2.2 a,b are variants of power law behavior representing heavy tails of power law. In situations like these, or when the data is limited, the upper range of the tail is truncated using an exponential cutoff. It implies that the probability of finding very high values decreases sharply.

of networks and distributions [22]. Both the distributions (of k_p, k_d) are governed by power laws in the bipartite drug-target network. Hence, in the drug network, the first and second neighbors of a randomly chosen drug- z_1, z_2 respectively, are given by (see derivation in Appendix 1.1)

$$z_1 = \frac{Li_{\gamma_1-1}\left(e^{-\frac{1}{\kappa_1}}\right)}{Li_{\gamma_1}\left(e^{-\frac{1}{\kappa_1}}\right)}\left\{\frac{Li_{\gamma_2-2}\left(e^{-\frac{1}{\kappa_2}}\right) - Li_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)}{Li_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)}\right\} \tag{2.3}$$

$$z_2 = \frac{Li_{\gamma_1-1}\left(e^{-\frac{1}{\kappa_1}}\right)}{Li_{\gamma_1}\left(e^{-\frac{1}{\kappa_1}}\right)}\left(\frac{Li_{\gamma_1-2}\left(e^{-\frac{1}{\kappa_1}}\right) - Li_{\gamma_1-1}\left(e^{-\frac{1}{\kappa_1}}\right)}{Li_{\gamma_1-1}\left(e^{-\frac{1}{\kappa_1}}\right)}\right)\left(\frac{Li_{\gamma_2-2}\left(e^{-\frac{1}{\kappa_2}}\right) - Li_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)}{Li_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)}\right)^2 \tag{2.4}.$$

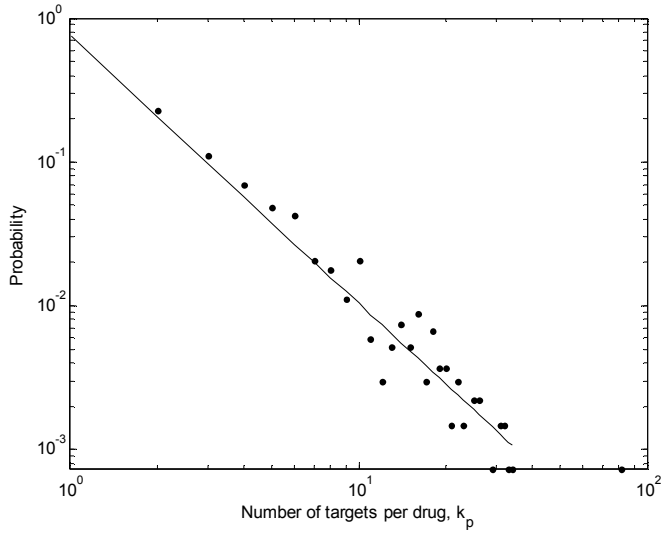


Fig. 2.1 Probability distribution of targets per drug (k_p) on log-log plot. The straight line approximates power-law exponent $\gamma_1 \approx 1.75 \pm 0.09$. γ_1 and scale parameter κ_1 are obtained by regression Eq. (2.2a) on a logarithmic scale, the exponential cutoff $e^{-1/\kappa_1} \approx 0.97$.

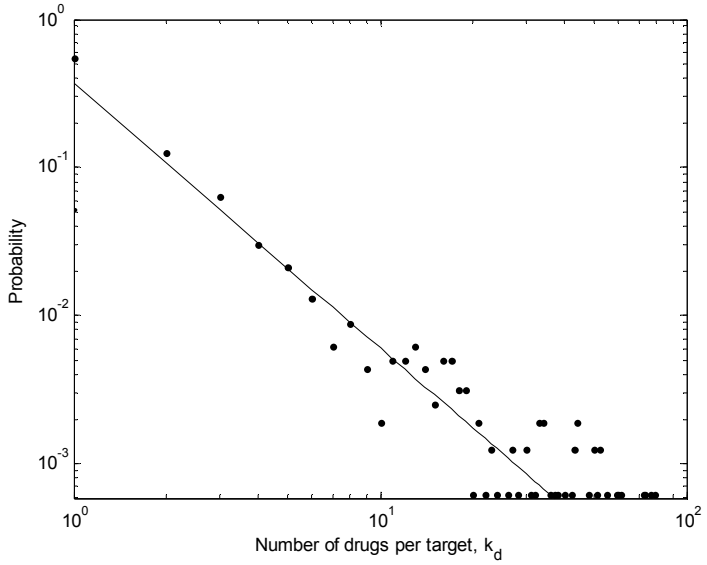


Fig. 2.2 Probability distribution of drugs per target (k_d) on logarithmic plot. The straight line approximates power-law exponent $\gamma_2 \approx 1.9 \pm 0.1$. γ_2 and scale parameter κ_2 are obtained by regression of Eq.(2.2b), the exponential cutoff $e^{-1/\kappa_2} \approx 0.98$.

I find by substituting the parameters, $z_1 \approx 34$, $z_2 \approx 3148.5$. Therefore, the path length [22] for the drug network is $l = \log\left(\frac{N}{z_1}\right)/\log\left(\frac{z_2}{z_1}\right) = 1.76$. The low path length is not surprising given the presence of drugs possessing many targets and it could point to a lower diversity in the activity profiles of many drugs. Note that the computed values are highly sensitive to decimal place considerations of the cutoff parameter and power law exponents. However, the computed values match with those observed within a range.

Further, I approximate the probability distribution of the numbers of co-drugs, that is, the number of drugs that are connected through sharing a common protein as an active target. It is given in Eq.(2.5) as the probability that a drug has *k* links in the (co)drug network,

$$p_k \approx \frac{Li_{\gamma_1-1}\left(\frac{1}{Li_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)}\right)\left(Li_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)\right)(k+1)^{1-\gamma_2}}{Li_{\gamma_1}\left(e^{-\frac{1}{\kappa_1}}\right)} \qquad (2.5).$$

This is based on some approximations considered for mathematical convenience (a sketch of the derivation is provided in Appendix 1.2). In order to see how it deviates from the observed degree distribution of the network of drugs constructed from the data, we plot the two distributions in Fig.2.3. This calculation is helpful because the actual distribution does not show a perfect power law behavior. The adjacency matrix of this graph of co-drugs is $A'_{ij} = 1$ if drugs i, j are commonly active on at least 1 protein and $A'_{ij} = 0$ if the activity profiles of i, j have no commonalities. We impose the range of observed connectivity to get a normalized predicted probability. Fig.2.4 compares the cumulative distribution functions of distributions shown in Fig.2.3. The figures and Kolmogorov-Smirnov test statistic confirm the correspondence between observation and the prediction in Eq.(2.5). It is interesting to see a power law connectivity distribution of drugs as it suggests the greater diversity shown by some drugs (as compared to others) that they are active on many different proteins.

2.3 Chemical Similarity of Co-drugs

According to the similarity principle, chemically similar drugs tend to act similarly on the proteins. The connections between drugs can be categorized in two dimensions. These correspond to the structure and activity attributes of interactions. One of them is the pairwise chemical similarity indicated by the weighted adjacency matrix *C* and the other is the activity similarity. In this chapter, I consider activity similarity as a binary attribute, implying whether (1) or not (0) a pair of drugs has a target in common. It determines the connections in co-drugs network. In line with the similarity principle, a drug sharing high chemical similarity with a high number of drugs should have a high

number of co-drugs. However, it is known that the presence of activity cliffs in the pharmacological space contradicts this behavior. It is therefore important to investigate the extent to which the similarity principle is valid in the given data. When chemically similar drugs do not share any targets, the interactions would be highly *distinctive*.

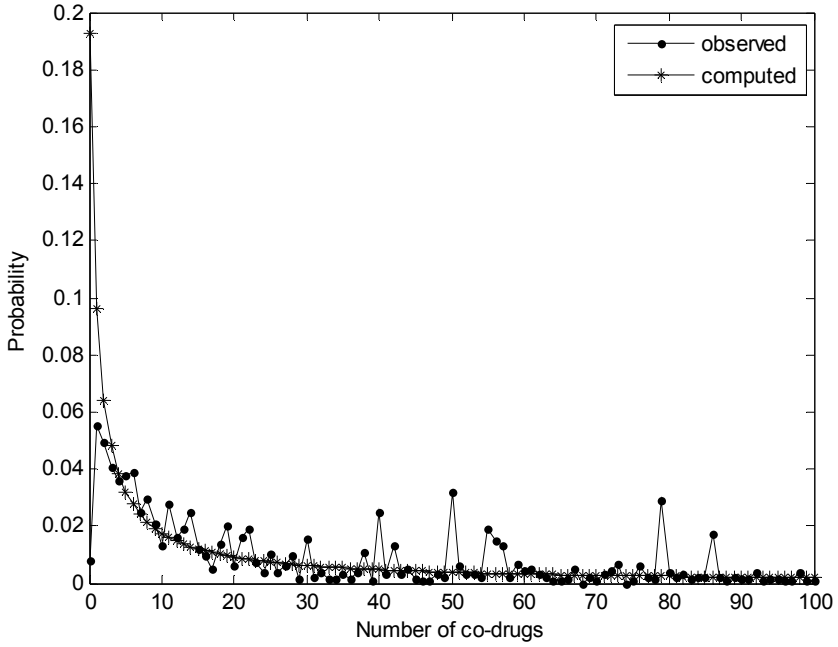


Fig. 2.3 Comparison of observed and computed probability distributions of connectivity or the number of co-drugs.

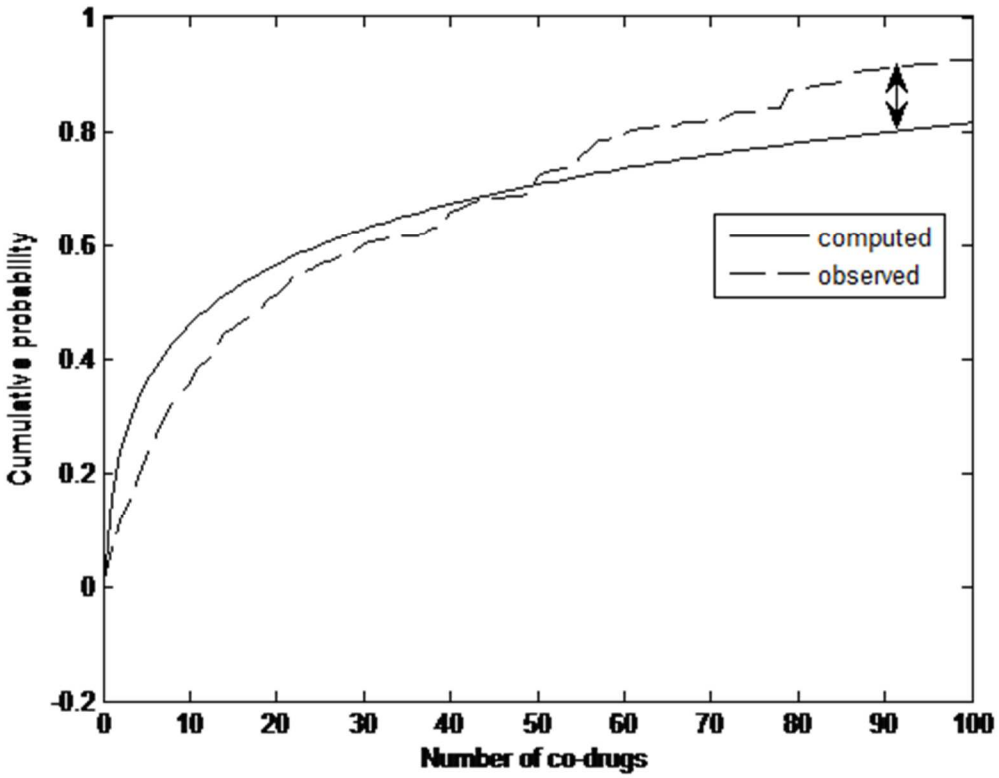


Fig.2.4 Comparison of observed and computed cumulative probabilities of connectivity or the number of co-drugs. Cumulative probability is plotted with number of co-drugs. Kolmogorov Smirnov test statistic using cumulative distribution functions Fn, F is D=sup|Fn-F|=0.17 which is significant at 0.05 level. Hence the correspondence between observed and computed probabilities is statistically significant.

In this co-drug network, the average chemical similarity observed between drug neighbors is 0.12. This is not high as expected from the similarity principle. Fig.2.5 indicates this, and how majority of neighbors share lower chemical similarity than higher. It means that the links between co-drugs (possessing a common target) may not possess a very high chemical similarity.

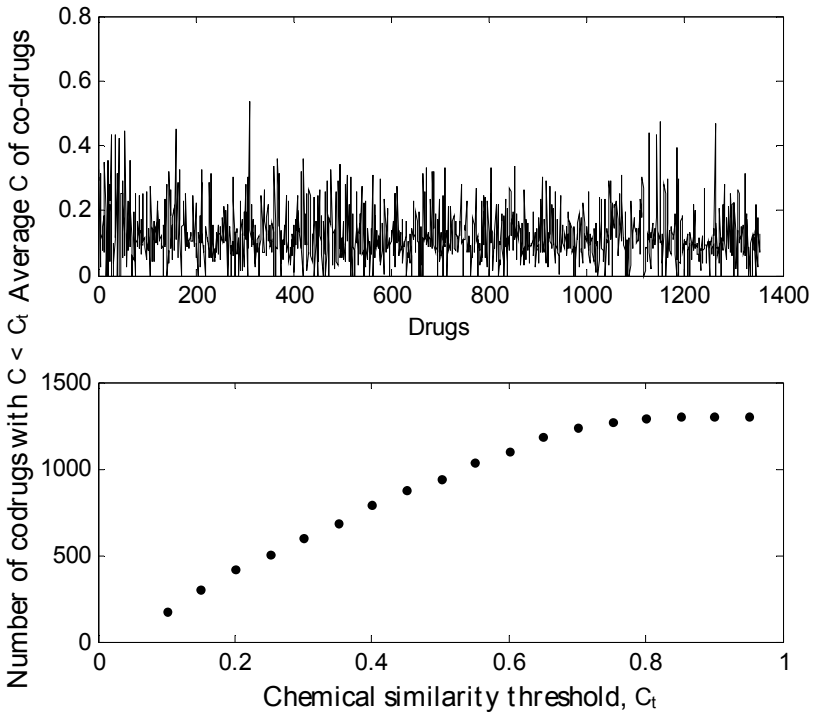


Fig. 2.5 (Top) Average chemical similarity of the links in the neighborhood of a drug. (Bottom) Variation of the number of neighbors or co-drugs with chemical similarity below a threshold C_t, with C_t.

Why is this and how does the deviation become important for identifying distinctiveness? In particular, what is the probability that a drug has *chemically similar* co-drugs? Chemically similar co-drugs are those pairs of connected drugs that share chemical similarity (c) at least as high as c =0.3. For studying distinctiveness, it is helpful to investigate the typical distribution of chemical similarities in the neighborhood of drugs. For instance, according to the similarity principle, two chemically similar drugs should have many targets in common. If a drug shares a target with many others, is it necessarily similar in its structure to theirs? The number of drugs in the neighborhood of a drug and the distribution of those drug links in levels of chemical similarity becomes important for distinguishing distinctive drug interactions from non-distinctive ones.

It is known that the probability that a link points to a vertex with degree k is proportional to kp_k and [22] the average connectivity of the drug network is $\langle k \rangle = \sum_k kp_k$. Further, every link must possess c

in one of the n categories in the range (chemical similarity $c_i = c\,\epsilon[0,1]$ for all i =1,2..n). The probability of finding links in category $c = c_i$ is given by $p(c_i)$ and is obtained from the data. If x_i is the number of links in this category, then the probability that a drug having k co-drugs will have its links distributed into the bins of chemical similarity would be proportional to

$$p_k \frac{k!}{x_1!\,x_2!\,\ldots.\,x_n!} p(c_1)^{x_1} p(c_2)^{x_2}\,\ldots.\,p(c_n)^{x_n} \qquad (2.6)$$

where $\sum_i p(c_i) = 1, \sum_{i=1}^{n} x_i = k$, considering k links to be independent trials. This way, the links around a drug would be distributed in the various categories of chemical similarity in accordance with the multinomial distribution. In this network, $\langle k \rangle$ is the average connectivity, so the expected frequency of links in i^{th} category of chemical similarity is given by $\langle k \rangle p(c_i)$. The variance is $\langle k \rangle p(c_i)(1 - p(c_i))$, and the average chemical similarity in the neighborhood of a drug is $\sum_{i=1}^{n} c_i x_i / \langle k \rangle$. Here, for the drug network, $\langle k \rangle = 38.24$, the variation of $\langle k \rangle p(c_i)$ is shown in Fig. 2.6 for n=19 categories $c_1\epsilon$ [0, 0.05], $c_2\epsilon$ [0.05, 0.1], $c_3\epsilon$ [0.1, 0.15]…. $c_{19}\epsilon$ [0.95, 1]. The sharp drop in the expected frequency as c_i increases is attributed to the disproportionately high number of links having c < 0.3. These links are not of pharmacological significance yet they form 99.7-99.8% of the links. These are considered redundant in pharmacology and we ignore these in future.

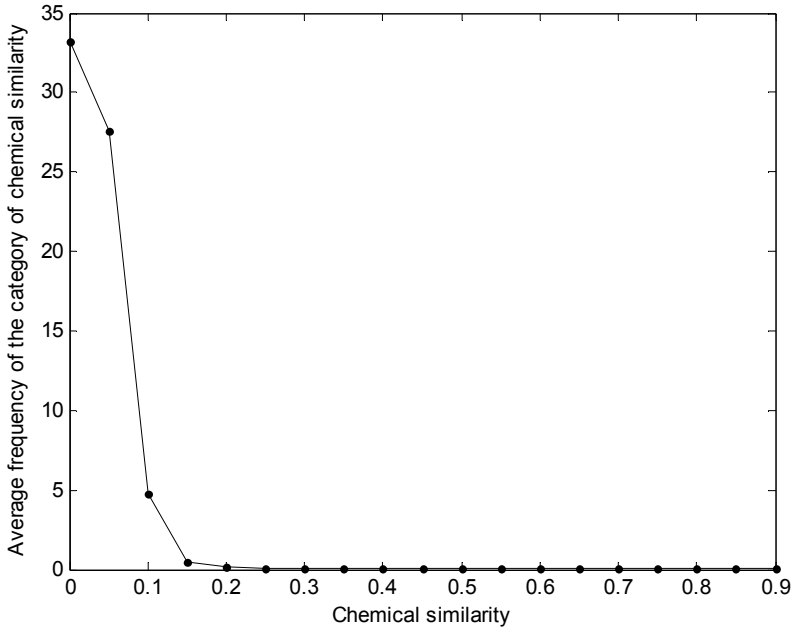


Fig.2.6 Variation of average frequency of links in the categories of chemical similarity in the interval [0,1] divided in subintervals of 0.05 width.

Considering that $\langle k \rangle$ links with co-drugs are independent, these links may be distributed in different categories of *c*, ranging from low to high. The resulting distribution can be approximated as a multinomial. I assign the number of links in these intervals of *c* and as before, compute the probabilities $x_i/\langle k \rangle$. The probability mass function (pmf) of the multinomial distribution can be computed. It would be relevant for distinctiveness to study whether or not the combinations favoring the categories of higher chemical similarities are probable. The most probable combination of number of links assigned to all *n* categories corresponds to the one with maximum entropy [24,25]. In other words, the most probable combination is the one with maximum multiplicity $M = \frac{\langle k \rangle!}{x_1!x_2!....x_n!}$.

I perform two experiments. First, I randomly assign each of the $\langle k \rangle = 39$ links to one of the *n*=3 categories of chemical similarities in the ranges [0-0.3], [0.3-0.6], [0.6-1] with equal probability of 1/3 for each category. In 100 random samples, I find the combination of x_i which maximizes *M*. This happens for $x_i \epsilon$ [12, 13, 14] and any of the permutations. Considering the middle values of all subintervals or categories as their representatives of *c*, the average chemical similarity in the

neighborhood of a drug is around 0.46. The probabilities *observed* for finding links in these categories are given as $p(c_i)$=[0.998, 0.0018, 0.0002], respectively. This means that $p(c_1)$=0.998, $p(c_2)$=0.0018, $p(c_3)$=0.0002. Again, considering the midpoint representatives of *c* in the categories, the expected chemical similarity = $0.15 \times 0.998 + 0.45 \times 0.0018 + 0.75 \times 0.0002 \approx 0.1507$. This is much lower than in the previous case. The difference is because of the large proportion of chemically redundant links that may dominate the neighborhood of a drug. Therefore, the high *c* categories are highly under-represented in the drug neighborhoods.

In the next experiment, I consider the links with chemical similarity at least 0.25 and ignore all the redundant links that the co-drugs may have. In doing so, a new drug network is constructed in which the interactions between two drugs are two dimensional as the linkages are formed by considering two independent properties. These linkages between two medicines are characterized as the pairs of medicines with *c* at least as high as 0.25 *and* one or more common targets. I refer to the network formed by these special interactions as the co-star network of drugs. The average connectivity of the network is $\langle k \rangle \approx 4$. The proportion of links retained is only 0.3%. The categorization of c_i considered is in intervals [0.25-0.5], [0.5-0.75], [0.75-1], midpoints of the intervals being representatives [25]. The observed probabilities of finding co-star links in these intervals are given as $p(c_i)$=[0.8489, 0.132, 0.0191]. The average chemical similarity of the configuration expected around a vertex or in the network is $0.375 \times 0.8489 + 0.625 \times 0.132 + 0.775 \times 0.0191 \approx 0.47$. Now assigning $k = 4$ links randomly 200 times to the categories with the *observed probability* scheme, the configuration of maximum *M* is x_i = (1, 2, 1) or (2, 1, 1) or (1, 1, 2). Table 2 gives the results. The average chemical similarity around a vertex in the co-star network corresponding to maximum entropy configuration is around 0.625.

Table 1 Taking equal probabilities of the categories (1/3) and <k>=39

(x_1, x_2, x_3)	M	**Average chemical similarity** $\sum_{i=1}^{3} c_i \times x_i/39$
(12,13,14) and its permutations	7.84×10^{16} (maximum)	0.46
(15, 12, 12) and its permutations	6.8×10^{16}	0.458

Table 2 Probabilities of the categories [0.8489, 0.132, 0.019] and k links, for maximum M

k	(x_1, x_2, x_3)	**Average chemical similarity** $\sum_{i=1}^{3} c_i \times x_i/k$
4	(1, 2, 1) and the permutations	0.625
10	(5, 4, 1) or (6, 2, 2)	0.51
15	(10, 3, 2) or (7, 7, 1)	0.49
20	(9, 9, 2) or (13, 5, 2) or (12, 7, 1)	0.5

The co-star network predicts average chemical similarity around a vertex that is closer (than the previous case) to that of the most probable configuration obtained by simulation, more so, as k is increased. This is interesting because the observed probability for interval of lowest c is still very high as compared to intervals of high c. Thus co-star network can reveal distinctive interactions between medicines as it is constructed using two criteria for formation of links- activity similarity and chemical similarity. Further, after removal of redundant links, the linkages of co-star network become perceptible for comparison across various classes or intervals of chemical similarities. The connections of individual vertices with higher connectivity than average would be dominated by relatively lower than higher chemical similarities. The multinomial distribution puts greater mass (pmf) on combinations favoring lower c. From Eq.(2.5) the probability of finding drugs with high connectivity diminishes as power law. So, from Eq.(2.6), as the connectivity (or functional similarity) of a drug increases, its links tend to be less restricted to the low c domain. Equivalently, the preponderance increases in intervals of higher c. This is a plausible explanation for the results in table 2. That being said, it is interesting that the most probable average c in the neighborhoods of highly connected drugs decreases and is closer to the expected value.

This network can help in identification of distinctiveness at various levels of non-redundancy. This means changing the minimum *c* criterion to construct the network can reveal the various ways in which chemical similarity may be distributed among the links around a vertex. Increasing minimum *c* reveals the extent to which the compliance with similarity principle would increase but this is not always feasible because there is a limit on the range of *c* that can be considered as redundant. The above analysis is clearly not in accordance with the similarity principle and distinctiveness emerges from this re-analysis of the information.

Chapter 3

DISTINCTIVE INTERACTIONS: A QUANTITATIVELY REASONED PERCEPTION

The drug network constructed from the drug target data in Chapter 2 showed that drugs sharing activities were quite chemically similar to each other but the average *c* in the neighborhoods of drugs was not high. This means that there is a large number of drugs sharing activities but the chemical similarity of these drug pairs is not high. Therefore, one can find a lot of pairs with low or intermediate *c* having at least one common target. As we have seen in the last chapter, as the number of connections of vertices in the drug network increases, the links may not be restricted to the low *c* range, but the average *c* (of the most probable configuration) in the drug neighborhoods would decrease. This indicates a tradeoff between increase in similar activity and the average *c*. For assessing the compliance with the similarity principle, however, it is more useful to enumerate the activity similarity by number of common targets. This method would give the level of activity similarity of a pair and help in discerning the pairs that are highly similar in activity at any level of $c>0.3$. Such pairs at high levels of *c* can be accurately classified as non-distinctive in nature.

Activity cliffs in the pharmacological landscape are indicated by discontinuities in structure activity association. These type of cliff interactions are formed by pairs of drugs having high *c* and highly divergent activity profiles, meaning, the drugs have vastly different activities on the same proteins. They are represented by distinctive drops in similar activity along the chemical space in the landscape. The extent of their non-compliance with the similarity principle (or as I refer to these aberrations as their distinctiveness) is quantified by measuring the dissimilar activity of the drugs in these pairs [13,17]. Previous studies have focused on dissimilarity and similarity in activity profiles of drugs as complementary measures. Measures like sali are constructed using pairwise differences in activity values and the chemical similarity of drugs [17]. These are studied as pairwise weights in a weighted

adjacency matrix of all the drugs from which cliff-like pairs are identified as the ones showing high *c* and high dissimilarity in activity. However, these kinds of measures are threshold dependent and not well defined as I discuss later in this chapter.

In this chapter, I specify an interaction between two drugs quantitatively using two measures of activity (i) dissimilarity or variation (*d*) and (ii) similarity (*s*) in activity profiles. For drugs, the measures do not add to 1596 necessarily, in fact $d \neq 1596 - s$. The measurements are made using the activity information of 1354 drugs on 1596 proteins in *A* (described in section 2.1). Distinctive interactions are defined in two ways in terms of these two measures. Interactions are *distinctive* if the pairs of drugs either have high *c* and high *d*, or have high *c* and low *s*. I use chemical similarity data as mentioned above. Pairwise measures of *c*, *s*, *d* are weights of the interactions and can be represented as elements of a weighted adjacency matrix. Using this representation, I try to compare the network characteristics in terms of *c*, *s*, *d*. This is relevant for understanding distinctiveness because it helps us to find correspondence between clusters of chemically similar drugs and those of drugs having similar or dissimilar activity. In particular, which drugs in the network have high *c* and which have high *s*, high *d*? Is there a difference?

Fig. 3.1 shows the difference in the spectral properties of *C* and the weighted network obtained by random assignment of weights in the adjacency matrix, C_r, averaged a 100 times. The weights of C_r are randomly chosen in the range [0,1]. The eigenvalue distribution of C_r is quite different from that of *C*. This quantified by the spectral distance [26, 27], which computes the difference between the eigenvalues as in Eq. (3.1). If $\lambda_i^C, \lambda_i^{C_r}$, *i*=1,2,..*N* are eigenvalues of *C*, C_r respectively, then spectral distance is

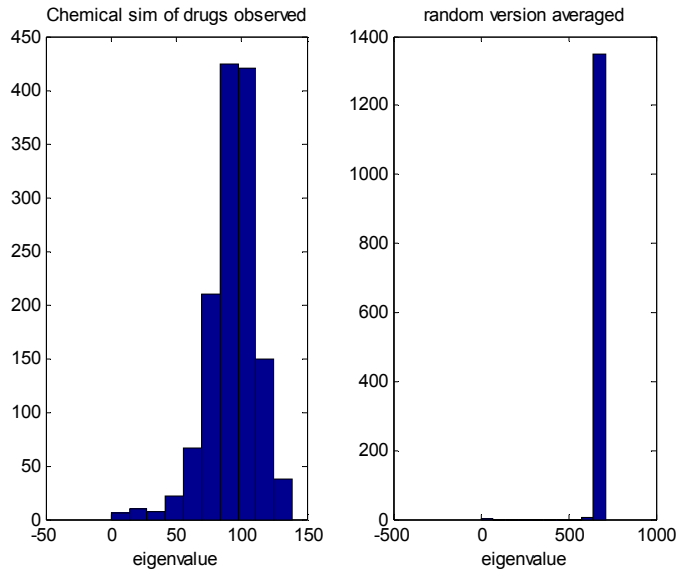


Fig. 3.1 Histogram of eigenvalues of C (left) and of the random version of C (right).

$$\partial_s = \sqrt{\sum_{i=1}^{N} (\lambda_i^C - \lambda_i^{C_r})^2} \qquad (3.1).$$

It is as high as 21454. Drug Lithium shares least chemical similarity with others in the network.

3.1 Network Modifications and Distinctiveness

I first compare the networks using the unweighted versions of networks C, Ψ, ζ for N=1354 drugs. The unweighted versions are given by $N \times N$ adjacency matrices as (i) $C_{ij}^u = 1$ if the chemical similarity of medicines *i,j* is at least as high as a constant (*const*), and 0 otherwise (ii) $\Psi_{ij}^u = 1$ if the medicines *i,j* are commonly active on at least 1 protein, and 0 otherwise (iii) $\zeta_{ij}^u = 1$ if the activities of medicines *i,j* are variant on at least 1 protein, and 0 otherwise. Ψ_{ij}^u is the co-drugs network of Chapter 2 [27]. We also refer to Ψ, ζ as co-similar and co-variant networks respectively. Clearly, ζ is a denser network than Ψ because it is more likely to find drugs that act dissimilarly on at least one protein. We will see this affects the interpretation of distinctiveness. Note that *const* specifies the threshold of chemical similarity for a link to be considered as viable. This specification affects the number of connections in all the networks (if we consider the similarities and dissimilarities in functionalities of

pairs of drugs (*ij)* for which $C_{ij}^{u} = 1$). A quantity studied is the sparsity of the network, which is defined as the ratio of number of links observed (n_{links}) to the total number of links. Sparsity= n_{links}/N_{C_2}. Fig. 3.2 depicts the decrease in the number of links as *const* increases, the networks becomes less dense.

Weighted versions of these network measures consider the magnitudes of similar and variant activity between medicines. *C* represents the weighted adjacency matrix of chemical similarity. If A_i and A_j are activity profiles of drugs *i,j*, that is the proteins *k* for which $A_{i(k)}$ =1 and $A_{j(k)}$ =1, then $\Psi_{ij} = |A_{i(k)} \cap A_{j(k)}|$. Here Ψ has weighted adjacency matrix for *N*=1354 drugs. We denote the similarity weights or magnitudes as *s*. The commonly unresponsive proteins are ignored. Every element of the matrix denotes the weight of the link or the interaction between pairs of drugs.

Variation between the activity profiles of two drugs measures the number of differences or the dissimilarity between their binary attributes, that is the number of proteins on which they differ in activity. The weighted network of pairwise variation of drugs are given by $N \times N$ weighted adjacency matrix $\zeta_{ij} = |(A_{i(k)} \cap \overline{A_{J(k)}}) \cup (\overline{A_{\iota(k)}} \cap A_{j(k)})|$. This is *not* the complement of Ψ as we ignore the commonly unresponsive proteins on all pairs of drugs while computing Ψ. Both these measures [27] are constructed along the lines of Jaccard coefficient without normalization. The constructions of Ψ and ζ are crucial for understanding distinctiveness because if their magnitudes and variances are distinct, one can compare the actual numbers of similar and variant activities.

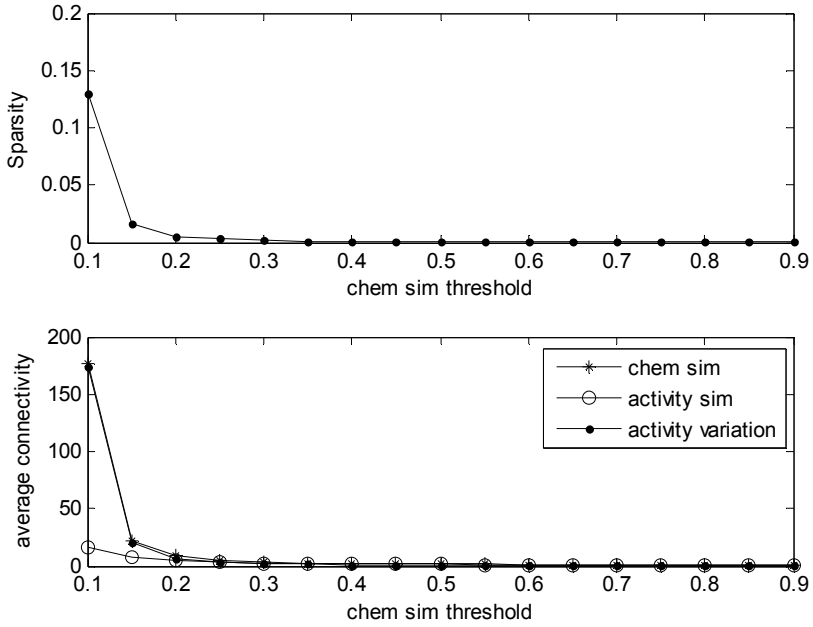


Fig. 3.2 (Top) Sparsity plotted with the constant of chemical similarity. (Bottom) The decrease in average connectivity of the unweighted networks C^u (chem sim), Ψ^u (activity sim), ζ^u (activity variation). The horizontal axis in both figures shows const as chem sim threshold.

3.2 Spectral properties of C, Ψ, ζ

This section analyzes the spectral properties of the unweighted and weighted adjacency matrices of chemical similarity, activity similarity and activity variation of the medicines. Particularly, we focus on the eigenvalues of the $N \times N$ Laplacian matrix (L).

$$L = D - X \qquad (3.2).$$

Here D is the $N \times N$ diagonal matrix with entries $D_{ii} = \sum_{j=1}^{N} X_{ij}$. The matrix X is considered as the observed (weighted and unweighted) adjacency matrices for C, Ψ, ζ. L is defined for each of the measures as L_c, L_Ψ, L_ζ respectively and for the unweighted measures as $L_c^u, L_\Psi^u, L_\zeta^u$ respectively. The eigenvalues of L are $0 = \lambda_1 \leq \lambda_2 \leq \lambda_3 \ldots\ldots \leq \lambda_N$. The smallest eigenvalue of L is 0 and the second smallest non-zero eigenvalue of L is called the Fiedler value, which gives the algebraic connectivity of the network. Its lower bound is inversely proportional to the diameter of the network and the eigenvector corresponding to this eigenvalue, the Fiedler vector, gives the spectral bisection of the network, as a community structure. It is the structure obtained such that the vertices in a group are

more connected (or tightly connected for weighted networks) to each other than to vertices belonging to another group. Figs. 3.3-3.5 show the community structure given by Fiedler vector of L_c^u constructed using *const*=0.3, 0.45, 0.65. In all these bisections, there is little difference between weights of links within groups and those across groups. The nature of drugs constituting the communities changes drastically as *const* increases. This may happen because the relative chemical similarity of drugs in the network changes on an average as some interactions of *c*<*const* are ignored. In Fig.3.3, the bisection is not perfect, the structure of the pair (Teniposide, Doxorubicin) is maximally similar (=1) and becomes prominent in the absence of a large number of links of intermediate *c*. Both these drugs affect DNA synthesis, structure but have different therapeutic functionalities.

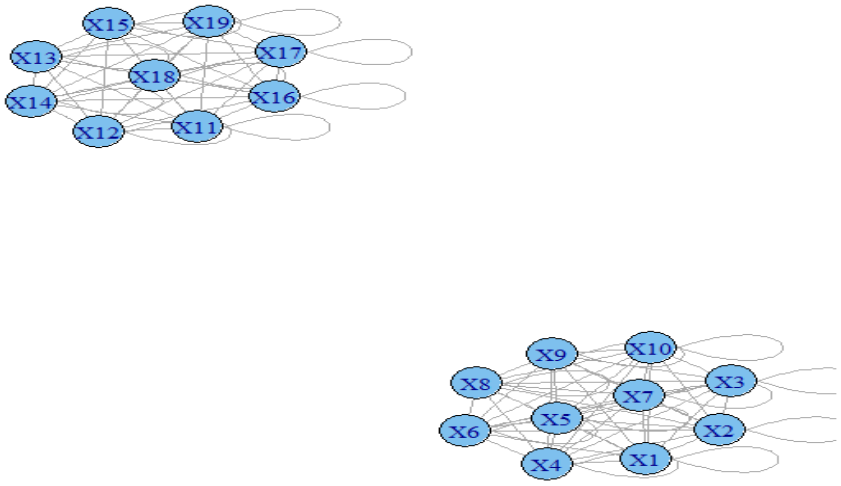


Fig. 3.3 Community structure of C^u *with* $\mathrm{const} = 0.3$. *The links gives the weight or chemical similarity between medicines within the two groups. For clarity the chemical similarities between medicines across the groups are not indicated. These drugs are: group 1 (top panel) –(Trospium, Oxyphenonium, Oxyphencyclimine, Clidinium, Glycopyrrolate, Oxybutynin, Tiotropium, Mepenzolate, Aclidinium) and group 2 (bottom panel)-(Benzatropin, Cetrizin, Buclizine, Hydroxyzine, Cinnarizine, Meclizine, Diphenylpyraline, Cyclizine, Almitrine, Flunarizine).*

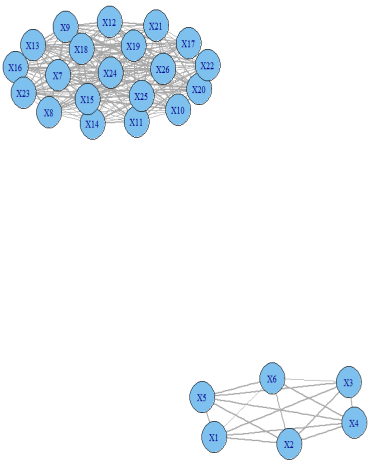


Fig. 3.4 Community structure of C^u with $const = 0.45$. The links indicate their weights or chemical similarity within the two groups. For clarity, the pairwise chemical similarities across the groups are not indicated. These drugs are: group 1 (top panel) –(Ceftazidime, Cefalotin, Cefotaximine, Cefdinir, Cephalexin, Cefixime, Cephaloglycin, Cefaclor, Cefditorin, Cefuroxime, Cefapirine, Cefadroxil, Cefaprozil, Ceftriaxone, Ceftizoxime, Cefradine, Cefepime, Cefacetrile, Ceftibuten, Cefpodoxime) and group 2 (bottom panel)- (Cefotiam, Cefamandole, Cefomicid, Cefoperazone, Ceforanide, Cefpiramide).

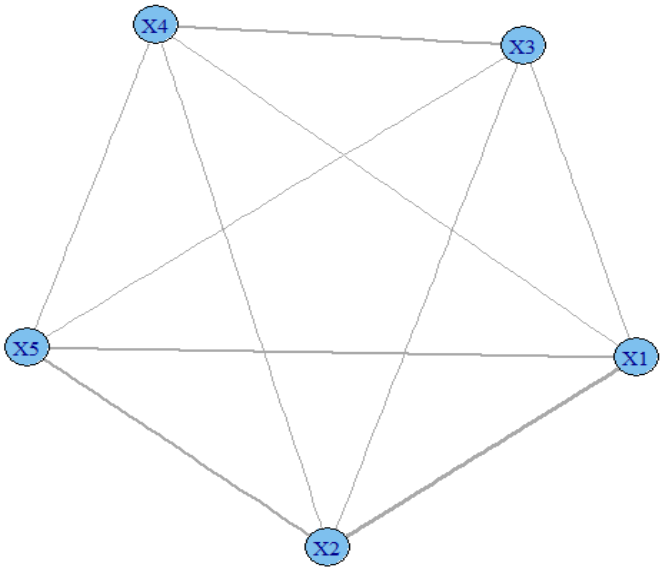


Fig. 3.5 Community structure of C^u with $const = 0.65$. The links are shown with their chemical similarity weights within the two groups. High weight interactions within group are (Teniposide, Doxorubicin) and those across the groups are (Doxorubicin, Idarubicin), (Teniposide, Idarubicin), (Rescinnamine, Deserpidine).

In all cases shown in this chapter, the Fiedler vector contributions of majority of vertices are negligible. Hence, for the identification of community structure, we consider the vertices within a group as those whose components have the same sign and magnitudes at least as large as the standard deviation of the Fiedler vector components. Figs. 3.6-3.7 show the interactions filtered through a spectral bisection of Ψ^u, ζ^u and the variation of community structure. The medicinal groups obtained for C^u in Fig.3.4 and for ψ^u in Fig.3.7 overlap to some extent. It is not surprising that the co-variant medicines appear in different groups of co-similarity.

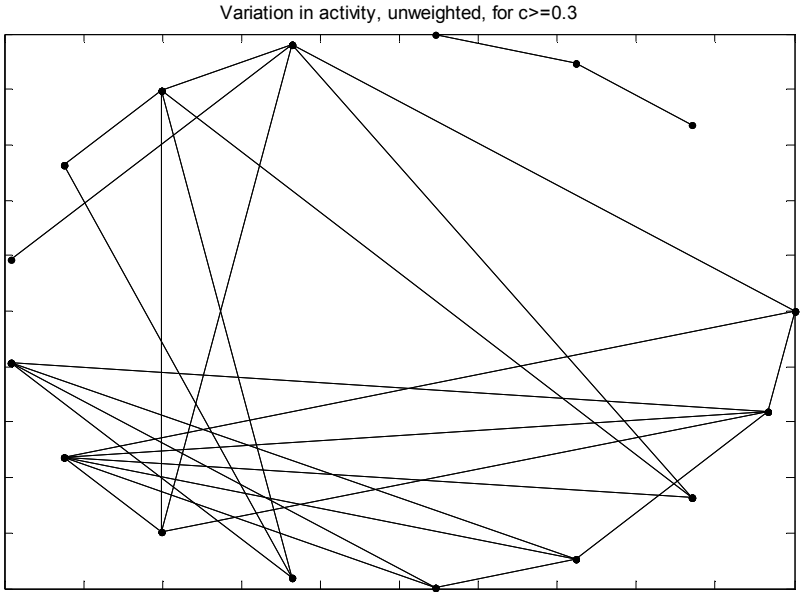


Fig. 3.6 Network of interactions of ζ^u obtained by spectral bisection for $const = 0.3$. Connections between all medicines in the groups are shown. The drugs are: group1- (Lorazepam, Clobazm, Diazepam, Estazolam) and group2- (Temazepam, Alprazolam, Adinazolam, Clorazepate, Midazolam, Flurazepam, Halazepam, Triazolam, Flunitrazepam, Fludiazepam, Prazepam, Quazepam, Cinolazepam).

It is important to observe that groups shown in Fig.3.7 consist of covariant medicines affecting ailments of central nervous system that are structurally quite similar. The difference in their activities is at a finer level of specification within the class of nervous system disorders. Also, not all medicinal pairs across the two groups are highly co-similar in activity. Therefore the groups represent medicines of nervous system with a common benzodiazepine structure but having fine distinctions in their

activity, which explains they are covariant on at least one protein. Weighted network analysis is required to infer how co-variant they may be.

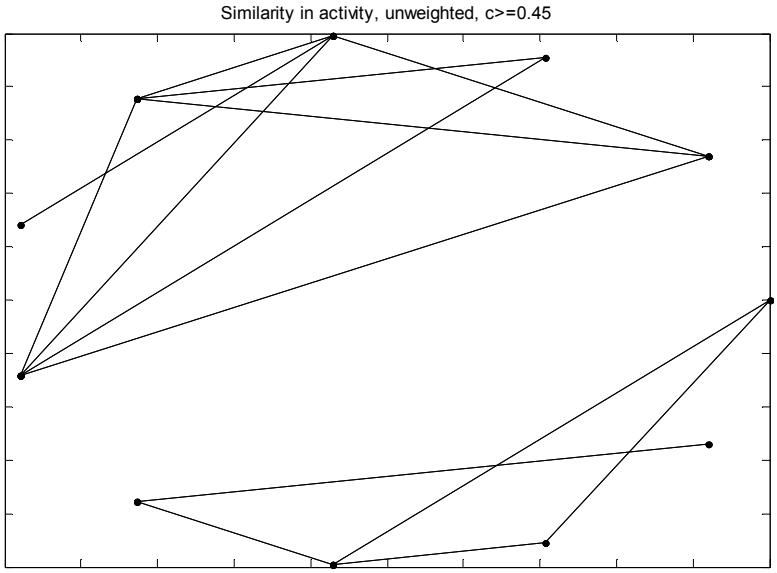


Fig. 3.7 Network of interactions of Ψ^u obtained by spectral bisection for $const = 0.45$. Connections between all medicines in the groups are shown. The drugs are: group1-(Cefmenoxime, Ceftazidime, Ceftizoxime, Cefepime, Cefecetrile, Ceftibuten, Cefpodoxime) and group2-(Cefalotin, Cefotaximine, Cefditorin, Cefapirin, Ceftriaxone).

The correspondence between *d* and *c* of medicinal pairs is further confirmed by computing the spectral distance in Eq. (3.1) between the distributions of eigenvalues of *L* for all 3 pairs of measures. Fig.3.8 gives the impact of changing chemical similarity threshold *const* on spectral distance. It is not surprising that ∂_s is high for (C^u, Ψ^u) given the difference in the construction of these measures. It decreases with increase in *const* because the average connectivity in Ψ^u decreases. It is interesting that while ∂_s decreases for all 3 pairs (Fig. 3.8), it is the least for (C^u, ζ^u) at all levels of *const*. This may imply that the imposition of the threshold affects the overall connectivity (and structure) in both adjacency matrices equally. However, the change in *const* may make Ψ^u a lot sparser. The magnitude of the largest eigenvalue decreases due to this perturbation and would be proportional to the probability of links having $c \geq const$. This is shown in Theorem in Appendix 2. It highlights the difference between the correspondences of behaviors of structural similarity with those of activity similarity and activity variation. Such a quantitative reasoning is crucial for identification of what can be referred to as *distinctive interactions*, which we also see in detail later. The algebraic connectivity

of the graphs increases with *const* indicating a decrease in diameter. It must be mentioned that the Fiedler value in this case may not be accurately observed because the gap between eigenvalues is very low and smallest non-zero eigenvalue of the Laplacian may be chosen quite arbitrarily.

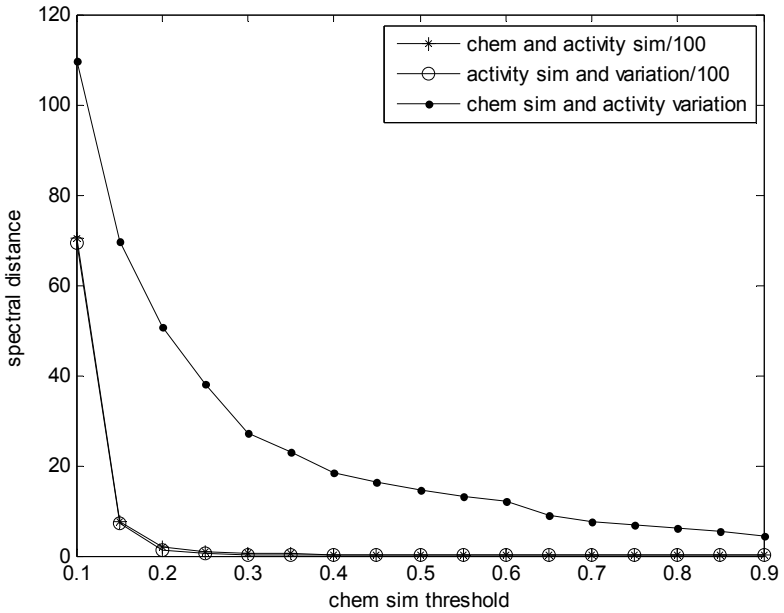


Fig. 3.8 Variation of spectral distance with const for $(C^u, \Psi^u), (\Psi^u, \zeta^u), (C^u, \zeta^u)$ *indicated by -*-, -o-, -.- respectively. The spectral distances of* (C^u, Ψ^u) *and* (ζ^u, Ψ^u) *are divided by 100.*

3.3 How Distinctive and How Probable?

In this section, I extend the analysis of distribution of links in the co-star drug network of the previous chapter to study the distribution of *distinctive* links of a drug. Distinctiveness of an interaction is quantified in terms of the (i) magnitudes of chemical similarity *and* similar activity (*s*) exhibited by the pair of drugs, and (ii) magnitudes of chemical similarity *and* dissimilar or variant activity (*d*) of the pair (*d* is also referred to as variation sometimes). Intuitively, these two measures appear to be complementary but they give rise to very different conclusions on identification of distinctive interactions, as shown later. These measures allow a quantitative specification for identification of distinctive interactions by imposing (i) minimum level of *c* and maximum level of *s* for (*c, s*) specification, and (ii) minimum levels of *c* and *d* for (*c, d*) specification. A distinctive interaction is

typically one characterized by high *c* but low *s*. It can also be the one characterized by high *c* and high *d*. Therefore, the levels of *c*, *s* and *d* specify *how* distinctive an interaction is.

Further, the probability of finding distinctiveness interactions would depend on the probability distributions of the levels of *c*, *s*, *d* observed. If we only consider the value of *const* used for the construction of the unweighted networks, we would have no information on the *level* of distinctiveness that is present in the drug network. The co-similarity network Ψ contains information on the magnitude of similar activities between drugs and the co-variant network ζ contains information on the magnitude of dissimilar or variant activities between drugs. If a drug has *k* co-similar drugs, then these *k* links may be distributed in categories of *c* and *s* simultaneously. The magnitude of distinctiveness around such a drug (in accordance with the multinomial pmf as used in section 2.3) is proportional to

$$p_k \frac{k!}{x_1!x_2!\ldots.x_{nm}!} p(c_1,s_1)^{x_1} p(c_1,s_2)^{x_2} \ldots p(c_1,s_m)^{x_m} p(c_2,s_1)^{x_{m+1}} \ldots\ldots p(c_n s_m)^{x_{nm}} \qquad (3.3)$$

where $\sum_{i=1}^{n}\sum_{j=1}^{m} p(c_i,s_j) = 1, \sum_{l=1}^{nm} x_l = k$. The structural and functional similarities are divided into *n* and *m* categories respectively and the links are distributed in a total of *nm* categories of *c* and *s* jointly. If a drug has *k* co-variant drugs, then these *k* links may be distributed in categories of *c* and *d* simultaneously. The magnitude of distinctiveness around such a drug is proportional to

$$p'_k \frac{k!}{x_1!x_2!\ldots.x_{nm}!} p(c_1,d_1)^{x_1} p(c_1,d_2)^{x_2} \ldots p(c_1,d_m)^{x_m} p(c_2,d_1)^{x_{m+1}} \ldots\ldots p(c_n d_m)^{x_{nm}} \qquad (3.4)$$

where $\sum_{i=1}^{n}\sum_{j=1}^{m} p(c_i,d_j) = 1, \sum_{l=1}^{nm} x_l = k$. p'_k is the probability of having *k* covariant drugs. The structural similarity and functional dissimilarity are divided into *n* and *m* categories respectively and the links are distributed in a total of *nm* categories of *c* and *d* jointly.

Considering n=m=3 results in nm=9 categories of joint (c, s) with c_i in the ranges [0.25-0.5], [0.5-0.75], [0.75-1] and s_j in the ranges [0-8], [9-17], [18-28]. The 3 representative midpoints of these ranges are $c_1 = 0.375, c_2 = 0.625, c_3 = 0.875$ and $s_1 = 4, s_2 = 13, s_3 = 23$.

The joint probabilities of finding links in 9 categories $(c_i \cap s_j)$ of $[c_1, s_1], [c_1, s_2], [c_1, s_3], [c_2, s_1], [c_2, s_2], [c_2, s_3], [c_3, s_1], [c_3, s_2], [c_3, s_3]$ are observed as [0.8169, 0.0292, 0.0028, 0.1264, 0.0053, 0.0004, 0.019, 0, 0] respectively. The average distinctiveness is $\sum_{i,j} c_i s_j p(c_i, s_j)$. Assuming structural and functional similarity as independent attributes of a drug interaction, $p(c_i, s_j) = p(c_i)p(s_j)$. I repeat the experiment of assigning k links to the 9 categories. The average distinctiveness calculated is 1.823. The probability of link in the first category of lowest c and s is the highest, and for any level of c, the probability decreases as s increases. While the multinomial distribution puts greatest mass on lowest category, Fig. 3.9 shows a relative reduction in its contribution to average (non) distinctiveness.

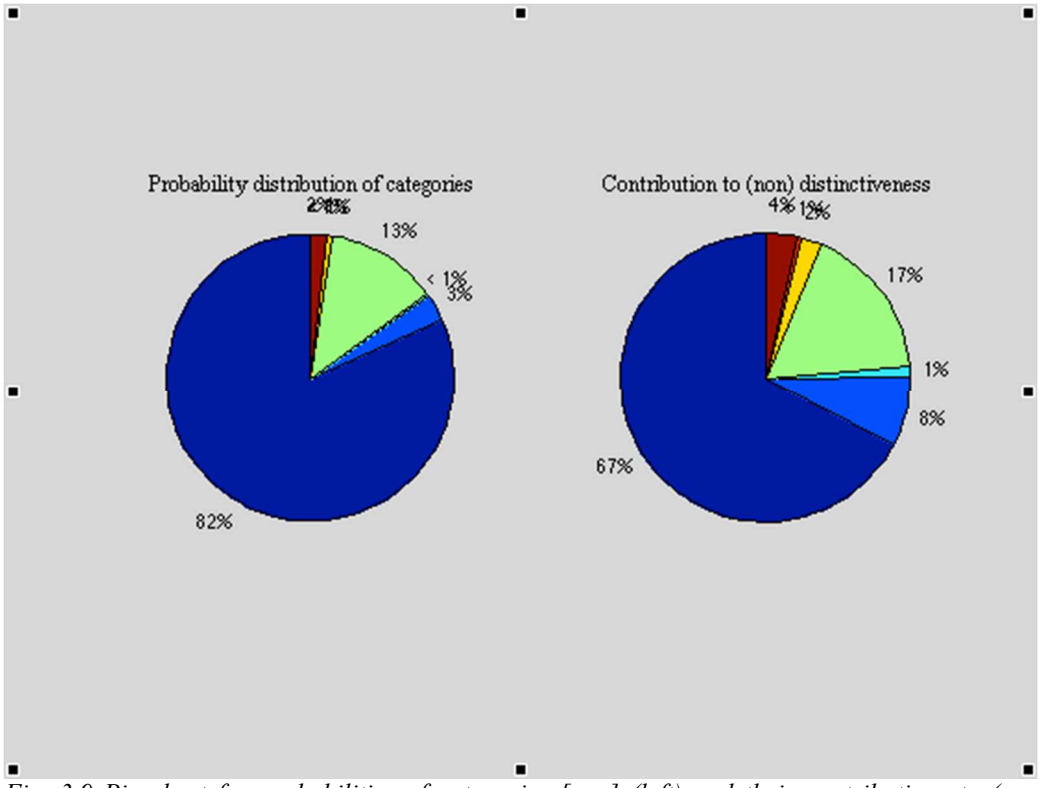


Fig. 3.9 Pie chart for probabilities of categories [c, s] (left) and their contributions to (non) distinctiveness (right). The blue region represents the lowest [c,s] category.

Therefore, even after removing the redundant links (of c<0.25), the similarity principle is contradicted, as the probability is 0 in the last two categories. This is confirmed by simulations as most links are assigned to $[c_1, s_1]$ or $[c_2, s_1]$. Also, if the connectivity of a drug increases, the average distinctiveness of the most probable configurations remains close to the expected average (non) distinctiveness. Table 3. (Note they are still different). It implies that when the number of co-similar links increases, the links tend to be distributed in more categories than just the first, spanning the categories of high c and high s. This indicates increase in non-distinctiveness. Fig. 3.10 shows the changes in observed distribution of links from expected. In the experiment of random assignment of links using the expected probabilities, the difference between contribution of $[c_1, s_1]$ and the other categories is highly reduced. Note that expected probabilities refer to the observed probability scheme used for the simulation.

Table 3 Configurations and distinctiveness of the most probable configurations using c, s

k	$(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9)$	Average (non) distinctiveness $\sum_{i,j} c_i s_j p(c_i, s_j)$
4	(2,0,0,1,0,0,1,0,0) or (2,1,0,0,1,0,0,0,0) or (2,1,0,1,0,0,0,0,0)	1.596
10	(5,1,0,3,0,0,1,0,0) or (5,1,0,2,1,0,1,0,0)	2.5
15	(8,1,0,5,0,0,1,0,0) or (8,0,0,5,1,0,1,0,0)	2.41

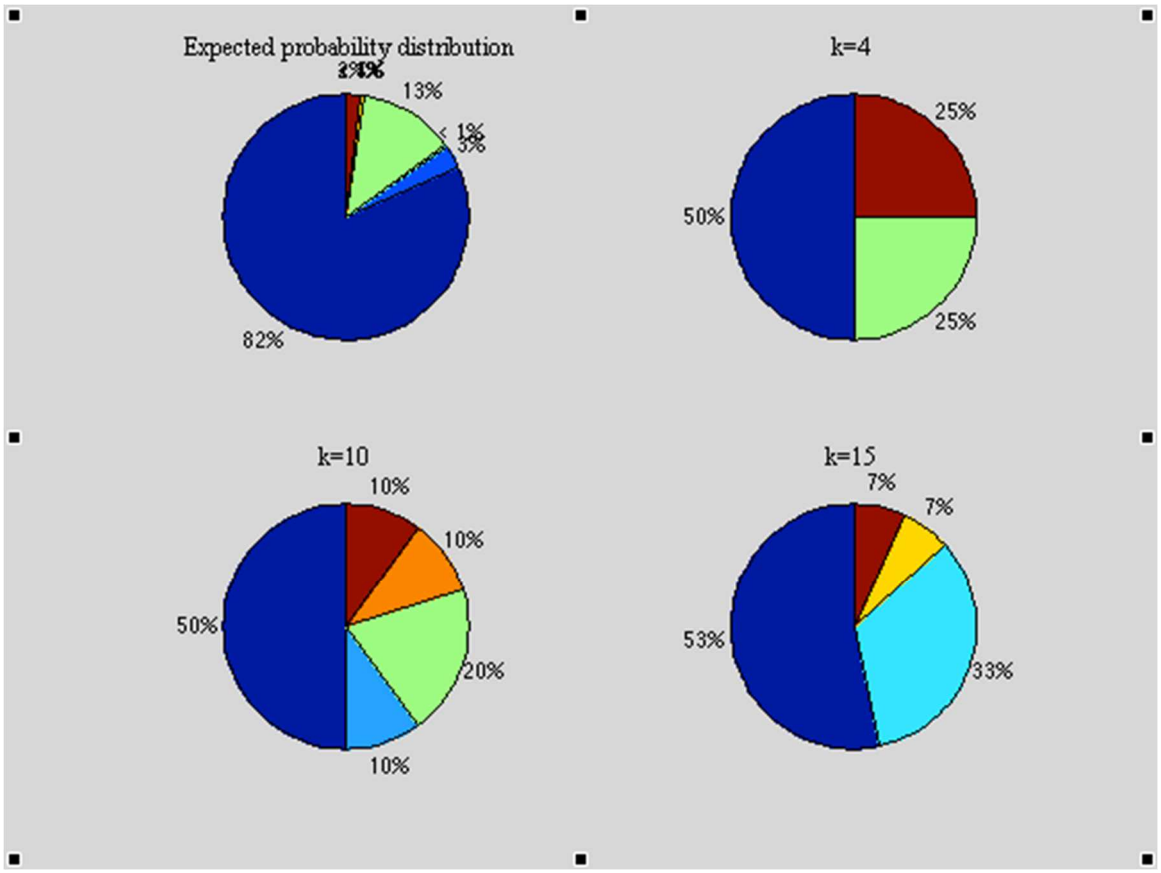


Fig.3.10 (Top left) The expected probabilities of categories [c, s] based on observations. The rest of the pie charts give the probabilities of categories obtained by observations in table 3 for the experiment k=4, 10, 15.

If in the neighborhood of a drug, the probability of finding connections of high *c* and high *s* is large, then neighborhood is non distinctive. If probability of high *c* and low *s* is high, then the interactions are distinctive. Intermediate ranges are inconclusive and can be classified as either distinctive or non-distinctive.

Similarly, the analysis can be repeated by considering *d* instead of *s*. For the same ranges of *c*, the 3 ranges d_j of interactions indicating the magnitude of covariance (*d*) between activities of drugs (from the covariant network, ζ).

The ranges are [0-37], [38-75], [76-114], and the midpoint values representative of these ranges are $d_1 = 18.5, d_2 = 56.5, d_3 = 95$ respectively. The probabilities of the 9 categories $(c_i \cap d_j)$ of $[c_1, d_1], [c_1, d_2], [c_1, d_3], [c_2, d_1], [c_2, d_2], [c_2, d_3], [c_3, d_1], [c_3, d_2], [c_3, d_3]$ are observed as

[0.843, 0.0035, 0.0025, 0.1317, 0.0003, 0, 0.019, 0, 0] respectively. The average distinctiveness is $\sum_{i,j} c_i d_j p(c_i, d_j)$. I repeat the experiment of assigning k links to the 9 categories. The average distinctiveness calculated is as high as 7.85. It is higher than obtained by using *s* because the magnitudes of *d* are higher. From the observed probability scheme, we can say that highly distinctive interactions are rarer (upper categories), but distinctive interactions are prevalent as indicated by the high probability concentrated in the first two categories and $[c_3, d_1]$. It is important to investigate their nature. It is also striking from results in Table 4, that the average distinctiveness of the most probable configurations obtained for various levels of connectivity, k in ζ is somewhat higher than the expected. Again, expected probabilities are those computed using observations. This may mean a bias toward the first categories or that observed numbers in first categories must be higher. While the multinomial probability puts greatest mass on lowest category, Fig.3.11 shows a relative reduction in its contribution to average distinctiveness. It implies that when the number of co-variant links increases, the links tend to be distributed in more categories than just the first, spanning the space including categories of high *c* and high *d* (increase in overall distinctiveness). This can be seen from Fig. 3.12 showing the changes in observed probability distribution of links from expected.

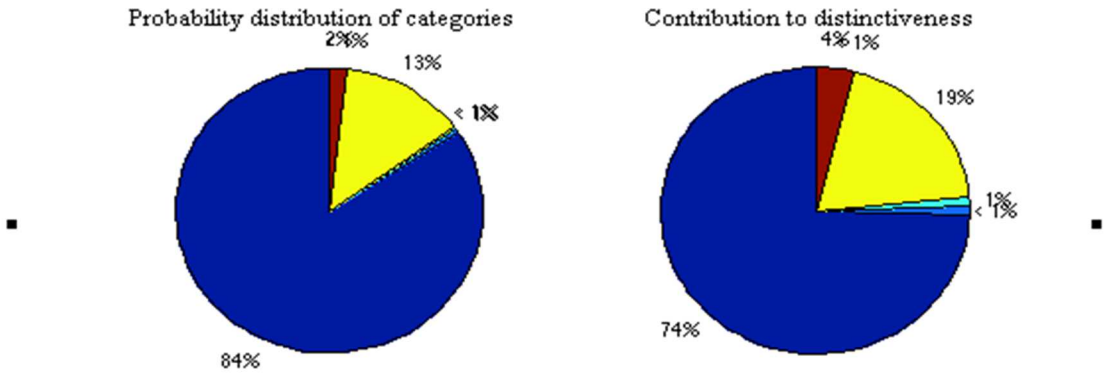


Fig.3.11 Pie chart for probabilities of categories [c,d] (left) and their contributions to distinctiveness (right). The blue region represents the lowest [c,d] category.

Table 4 Configurations and distinctiveness of the most probable configurations using *c*, *d*

k	$(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9)$	**Average distinctiveness** $\sum_{i,j} c_i d_j p(c_i, d_j)$
7	(3,0,0,3,0,0,1,0,0) or (4,0,1,1,0,0,1,0,0)or (3,1,0,2,0,0,1,0,0)	11.63
10	(5,1,0,3,0,0,1,0,0) or (6,0,0,3,0,0,1,0,0) or (5,0,0,3,0,0,2,0,0)	10.03
15	(9,0,0,4,0,0,2,0,0) or (9,0,0,5,0,0,1,0,0) or(10,0,2,3,0,0,0,0,0)	10.06

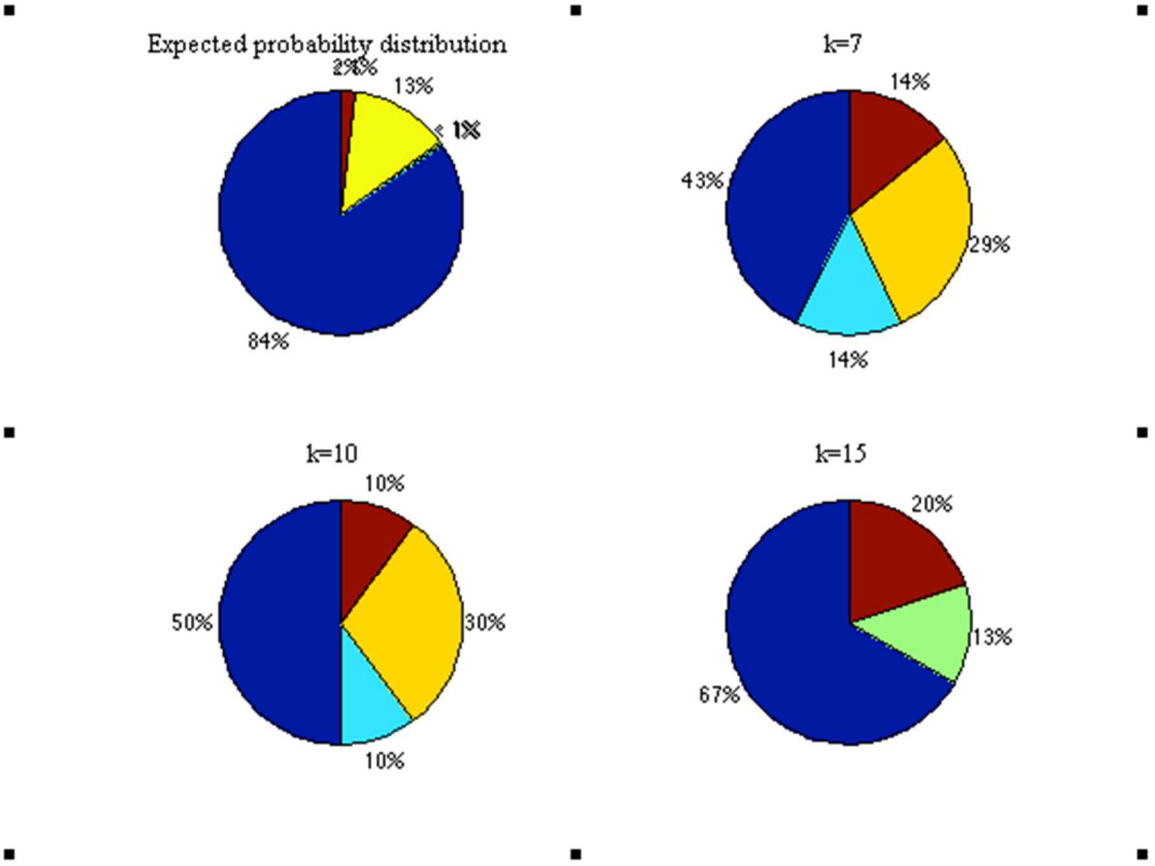


Fig.3.12 (Top left) The expected probabilities of categories [c,d]. The rest of the pie charts give the probabilities of categories obtained by observations in table 4 for the experiment with k=7, 10, 15.

It is interesting to note that if the probability of high *c* and high *d* is large, then the neighborhood[6] is highly distinctive, but if probability of high *c* and low *d* is high, then the interactions are considered distinctive. This is the major difference that the application of measures *s* and *d* can make on the interpretation of distinctiveness. It arises from the fact that while computation of *s* from weighted network Ψ is proportional to Jaccard index (as is *d* from weighted network ζ), there is a number of proteins that pairs of drugs are inactive on in the data. (Indeed there are drugs that are not active on any of the proteins). These are ignored in the calculation of *s*, and so the drop in similar activity is not necessarily indicative of the rise in *d*. However, a huge rise in *s* does guarantee a drop in *d*. Therefore,

[6] These experiments help to establish the inherent (increase in) uncertainty (or high information content) of the link attributes (*c*, *s*) and (*c*, *d*) actually found in the neighborhood of a drug (most probable configuration). It shows the possibility of high levels of distinctiveness and non-distinctiveness of links.

this analysis suggests that for distinctiveness, activity variation may be a more suitable measure to study than activity similarity. It indicates distinctiveness at different levels in the data and is a more selective filter of the interactions.

3.4 Identification of Distinctive and Non-distinctive Interactions

3.41 Measures of cs and cd

As seen from the ranges specified above for *c*, *s*, *d* that the weights of the links are in the ranges of $c_1 s_1 < cs < c_n s_n$ and $c_1 d_1 < cd < c_n d_n$. I construct the weighted networks of these link weights *cs* and *cd* for all links having $c \geq 0.3$. Fig. 3.13 gives the probability density of cs and cd on logarithmic and semi-logarithmic scales. For cs, a power law probability is indicated while cd shows fast exponential decrease initially with the very high values are equally probable.

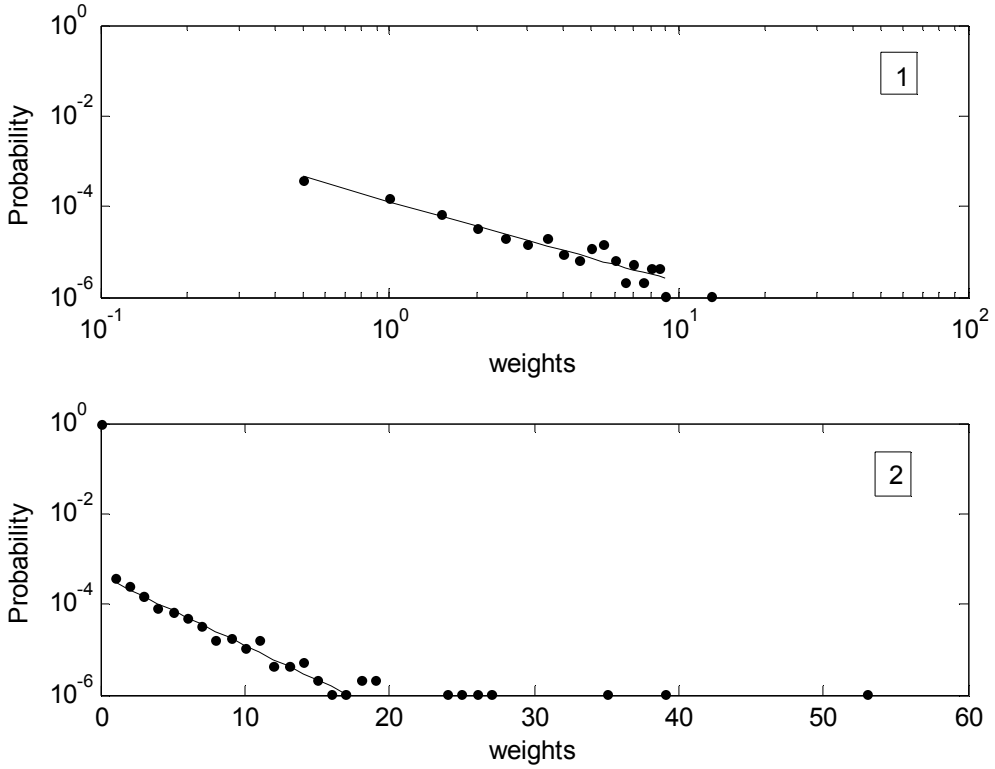


Fig. 3.13 Probability distribution of weights as (1) cs on logarithmic plot with straight line regression coefficient -1.78 ± 0.13 and (2) cd on semi-log plot with straight line regression coefficient -0.35 ± 0.01.

The observed ranges of these weights are $0 \leq cs \leq 13.3$ and $0 \leq cd \leq 53.5$. The mean and standard deviation of *cs* are given by $\mu_{cs} = 0.0068, \sigma_{cs} = 0.0083$ respectively and mean and standard deviation of *cd* are given by $\mu_{cd} = 0.0015, \sigma_{cd} = 0.0645$ respectively. While high *cs* weights would correspond to high non-distinctiveness, high *cd* weights correspond to high distinctiveness. It must be clarified that here 'high' is meant as being above average and how high is indicated by the number of standard deviations above average. It is interesting that the probability of finding links of weight *cs* decreases as we go above the mean that is $cs \geq \mu_{cs} + n\sigma_{cs}$ (for *n*=0,1,2,..) as a power law (Fig.3.14, power law exponent =1.41). It falls off faster as *n* is increased beyond 49-50. The 19 highly non-distinctive interactions observed at *n*=100 are given in Fig. 3.15. These represent links between medicines and their analogs (sharing high *c*) having similar functionalities on same diseases. However, the probability of finding links of weight *cd* decreases very slowly initially (for low *n*) but as a power law for higher *n*, (Fig.3.16, exponent =2.87), indicating very high variance in *cd*). The 14 distinctive interactions filtered by considering *n*=85 are between medicines and their analogs (highly chemically similar) which cure ailments of a particular class or kind but which are variant in their functionalities (Fig.3.17). These may be studied further to know the nature of distinctiveness.

Distinctive interactions occur between drugs of the same class, treating the same kinds of diseases but have finer distinctions in their therapeutics. The highly non-distinctive drug interactions are observed between drugs that are used commonly such as nutrients thus having same functionalities on diseases, proteins. Thus a hierarchical nature of link distribution across various levels of distinctiveness and non-distinctiveness is indicated by power law, but the interpretations of links in the uppermost categories of (largest *n*) both measures *cs* and *cd* are totally different. Finding distinctive interactions with *cs* measure (low *cs*) is not precise because we observe $cs \leq \mu_{cs}$ for 97% of the links. The networks in Figs. 3.15, 3.17 show clustering, which means that structurally similar compounds participate in many interactions. The power law dependence of probability is interesting as it indicates a hierarchical emergence of distinctiveness by relaxation of the evaluation criteria. It is also crucial to

note that the medicines constituting the interactions in Fig. 3.15 are also observed in Fig.3.6 which seems surprising because the metric applied in that analysis was *d* and not *s*.

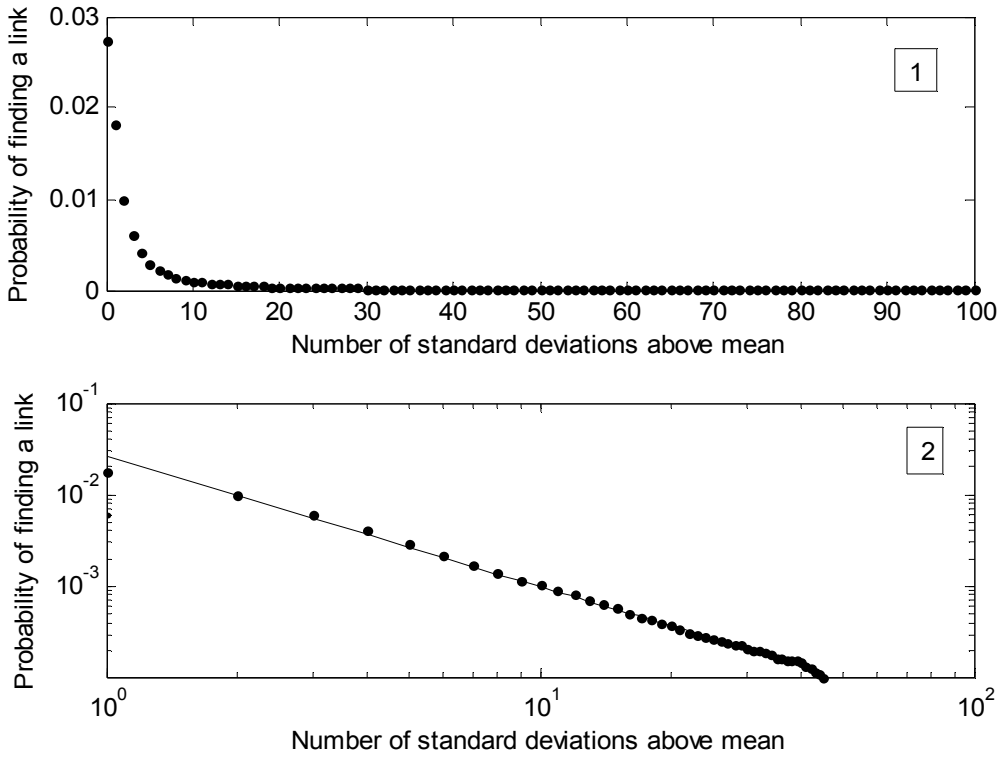


Fig. 3.14 (Top) Probability of finding links of weight $cs \geq \mu_{cs} + n\sigma_{cs}$ *with* n *(Bottom) Logarithmic plot of probability with the number of standard deviations, straight line in the tail has regression coefficient or power law exponent* -1.41 ± 0.027.

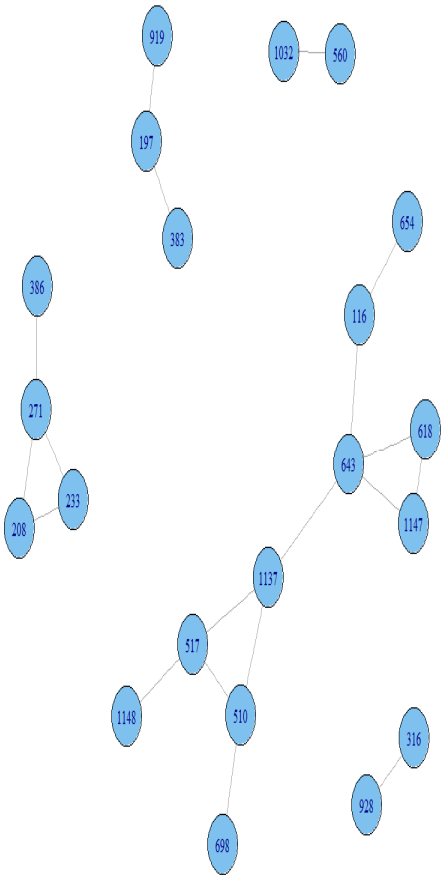


Fig.3.15 Non distinctive interactions filtered using $cs \geq \mu_{cs} + 100\sigma_{cs}$*. These are (Temazepam, Diazepam), (Temazepam, Oxazepam), (Amitriptyline, Notriptyline), (Amitriptyline, Doxepin), (Olanzapine, Clozapine), (Olanzapine, Loxapine), (Clozapine, Loxapine), (Loxapine, Amoxapine), (Imipramine, Desipramine), (Midazolam, Flurazepam), (Midazolam, Triazolam) Midazolam, Fludiazepam), (Flurazepam, Fludaazepam), (Flurazepam, Quazepam), (Risperidione, Fencamfamine), (Halazepam, Diazepam), Halazepam, Prazepam), (Diazepam, Fludiazepam), (Diazepam, Prazepam).*

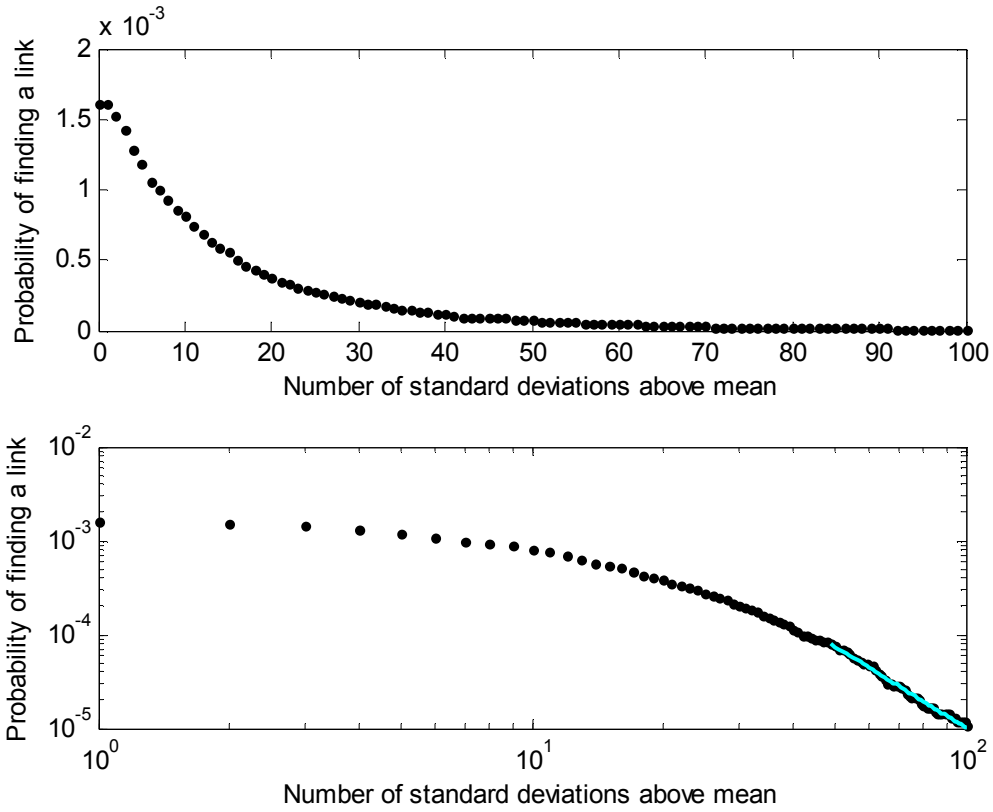


Fig.3.16 (Top) Probability of finding links of weight $cd \geq \mu_{cd} + n\sigma_{cd}$ *with n (Bottom) Logarithmic plot of probability with the number of standard deviations, straight line in the tail has regression coefficient or power law exponent* -2.87 ± 0.04.

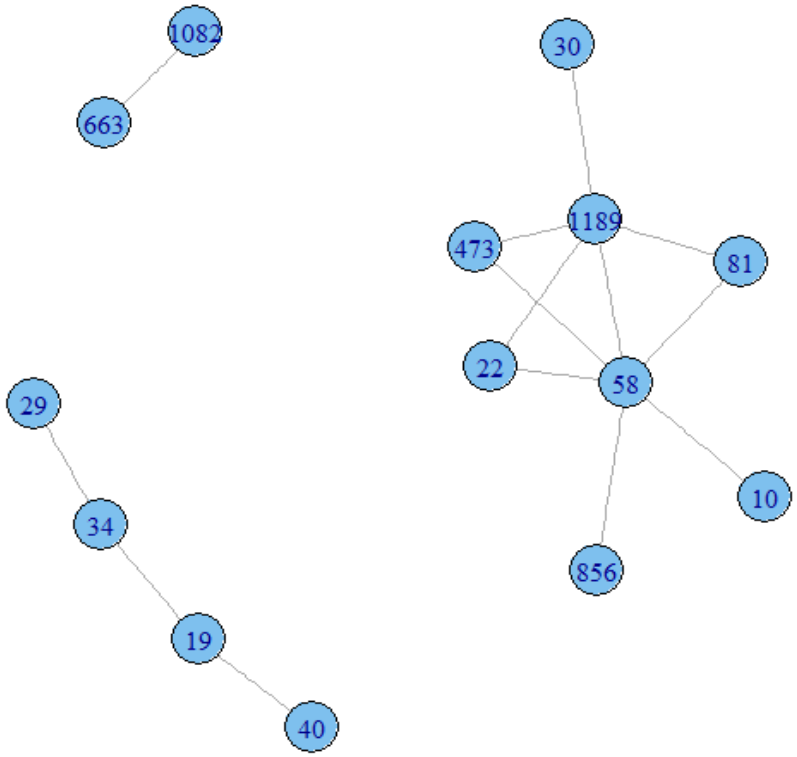


Fig.3.17 Distinctive interactions filtered using $cd \geq \mu_{cd} + 85\sigma_{cd}$ *. These are (S-Adenosinmethionine, Adenosine Triphosphate), (L-Aspartic acid, Glycine), (L-Aspartic acid, L-Cysteine), (Adenosine monophosphate, Adenosine triphosphate), (Adenosine monophosphate, Flavin Adenine Dinucleotide), (Succinic acid, Glycine), (Riboflavin, Flavin Adenine Dinucleotide), (Adenosine triphosphate, Vidrabine), (Adenosine triphosphate, Adenosine), (Adenosine triphosphate, Fludrabine), (Adenosine triphosphate, Flavin Adenine Dinucleotide), (Vidrabine, Flavin Adenine Dinucleotide), (Adenosine, Flavin Adenine Dinucleotide), (Pseudoephedrine, Ephedrine).*

3.42 Structure Activity Landscape Index and *cd*

Researchers have struggled with appropriate quantification of activity cliffs. This is because qsar models are less effective when small structural changes in drugs result in large variation in their activity. They are known to require additional information from experiments to offer predictions of cliffs. Also, the quantitative methods developed cannot be used to filter out most distinctive links. The

most commonly used measure for activity cliffs is the -structure activity landscape index (sali) [17] as it combines the chemical similarity and activity variation of drugs. For two drugs *i, j* sali is given by $sali_{ij} = \frac{\zeta_{ij}}{1-C_{ij}}$. This measure is used to characterize the drug network with an adjacency matrix by placing the elements of the matrix as 1 (or connected) if sali is greater than a threshold or sali constant, and 0 otherwise. While sali puts more weight on activity variation between pairs that possess high chemical similarity, it is rather contentious. This is because the sali constant is not well defined [17]. And the mathematical basis is insufficient because in this integrated measurement, high scores may arise from very low *c* and high *d*. This does not help in analysis of distinctiveness, which strictly requires both high *c* and high *d*. Since this method of network analysis is dependent on choice of sali constant, it is important to know the probability distributions of weights of drug interactions. I apply this approach later for a better characterization of the pharmacological distinctiveness.

The measure cd considered above is an integrated method that accounts for *c* and *d* jointly. However, as seen earlier, unweighted network analysis is not useful because of the increasing sparseness with threshold. In the Laplacian constructed using sali adjacency matrices, a decrease in the algebraic connectivity is indicated as sali constant is increased (Fig.3.18).

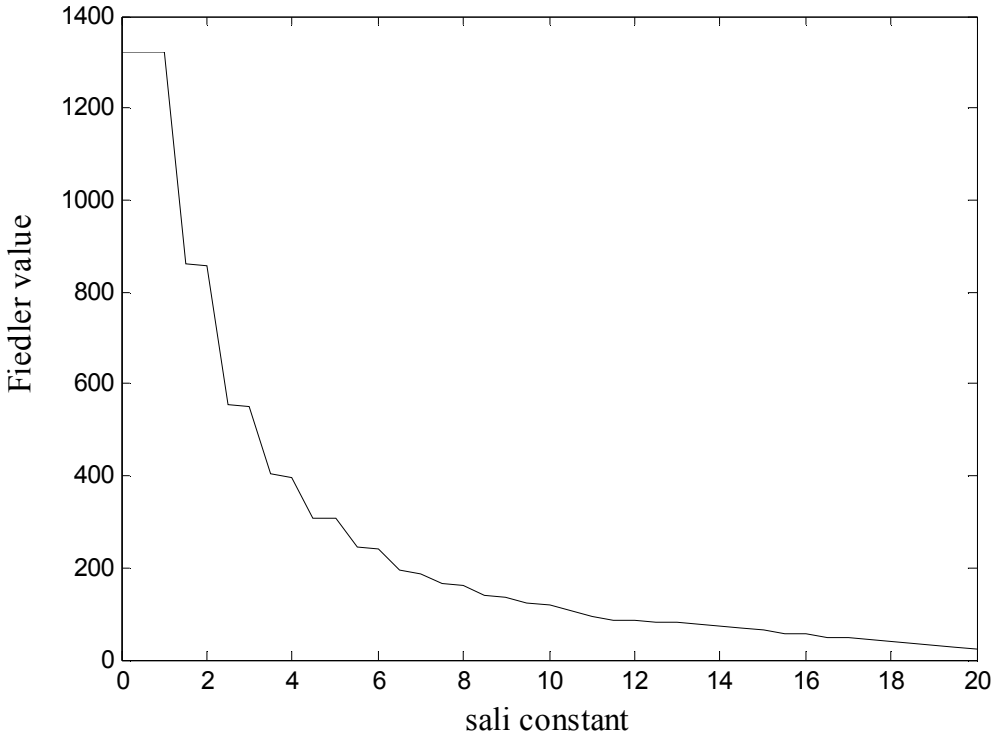


Fig. 3.18 Variation of algebraic connectivity or the Fiedler value of sali Laplacian with sali constant.

The inconclusiveness of this method for distinctiveness is indicated by the shift in the components of eigenvector corresponding to the Fiedler value as the sali constant is decreased. This makes it difficult to divide drugs into groups, as majority of links are either positive or negative in a very small range. Further, the maximum spectral gap is very large which rules out the existence of community structure. This leads us to classification of links as distinctive or non-distinctive based on whether or not their sali is greater than sali constant. Since there is a difference in the constructions of sali and $c \times d = cd$, the interpretations of distinctive and non-distinctive will differ and in fact there is no specification for non-distinctiveness given by sali. Thus sali is useful for filtering out the pairs of interactions that possess low structural similarity and very low variation in activity. The pairs scoring intermediately in these two dimensions of c and d are not properly classified.

3.43 Correlation Matrix Analysis

This subsection checks the results of analysis of correlation matrix of drug activity with the similarity measure s introduced previously. The matrix of correlations between drug activity profiles is computed as the Pearson coefficient and $-1 \leq Cor_{ij} \leq 1$. A positive correlation indicates similar behavior of two drugs on same proteins but negative or anti-correlation means they act oppositely on the same proteins. Note that the coinciding 0 activity values in two binary activity profiles also contribute to correlation whereas these are ignored in calculating s. I compare the eigenvalues of the observed *Cor* with those of a random correlation matrix generated by randomly shuffling the activities of drugs on proteins (random shuffling of activity profiles of all drugs) and averaging over 100 realizations. The comparison is made including all links in Fig. 3.19, keeping only the links whose $c \geq 0.3$ in Fig. 3.20 and only the links whose $c \geq 0.45$ in Fig. 3.21. In the last two cases, the eigenvalues of the random matrix are smaller than the corresponding eigenvalues of observed *Cor*. Fig.3.22 shows a drop in largest eigenvalue λ_{max} with *const*.

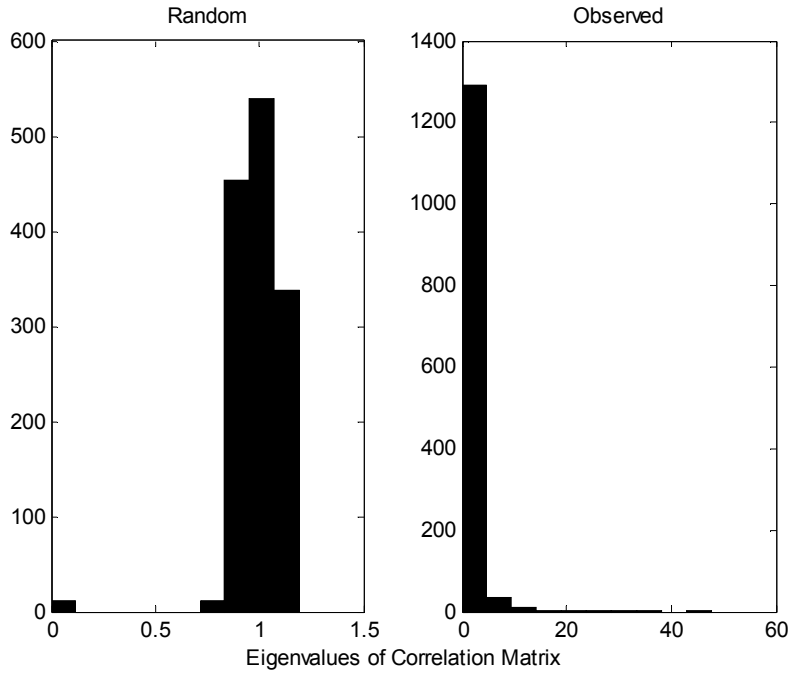


Fig. 3.19 Histogram of eigenvalues of correlation matrix: random (left) observed (right).

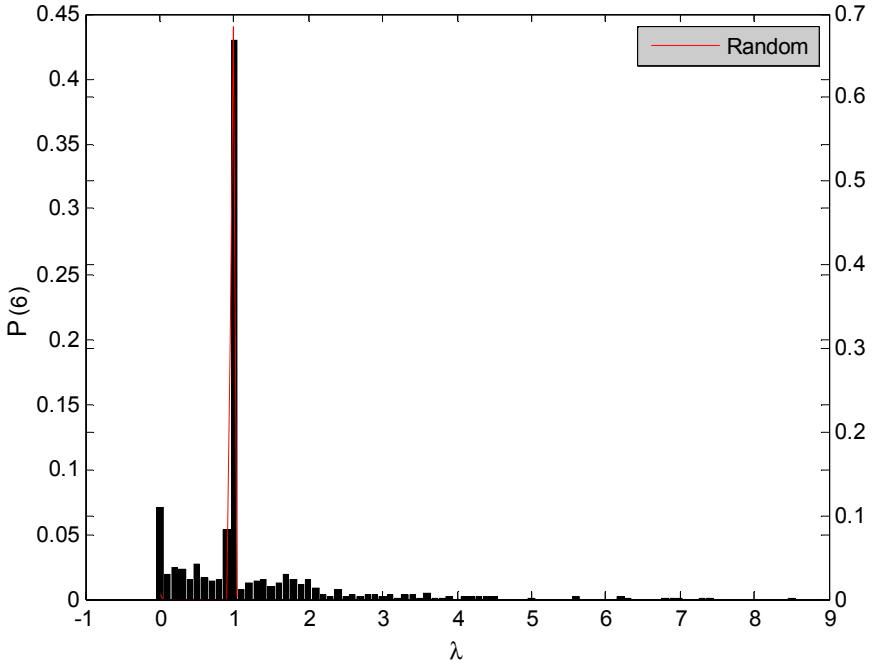


Fig.3.20 The eigenvalues λ are plotted with probabilities $P(\lambda)$. Comparing of eigenvalues of the random correlation matrix with those of the observed correlation matrix. Only the links wits $c \geq 0.3$ are retained.

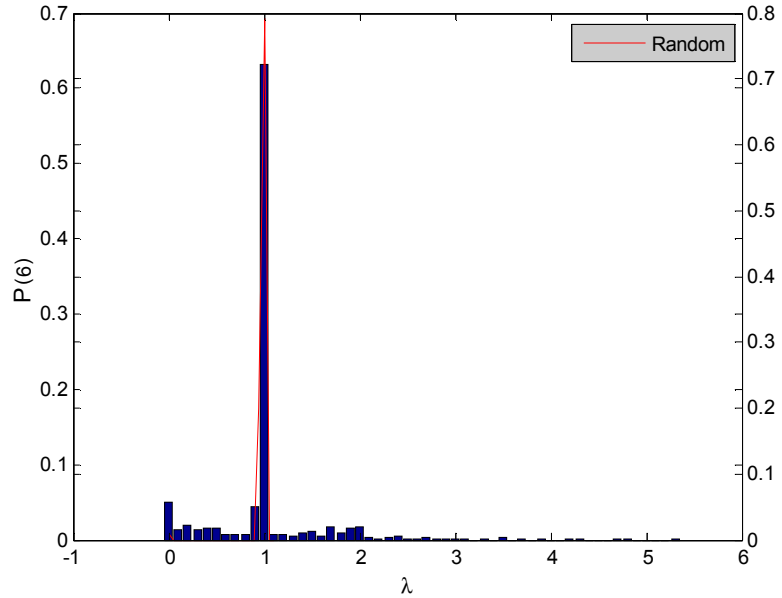


Fig.3.21 The eigenvalues λ are plotted with probabilities $P(\lambda)$. Comparing of eigenvalues of the random correlation matrix with those of the observed correlation matrix. Only the links with $c \geq 0.45$ are retained.

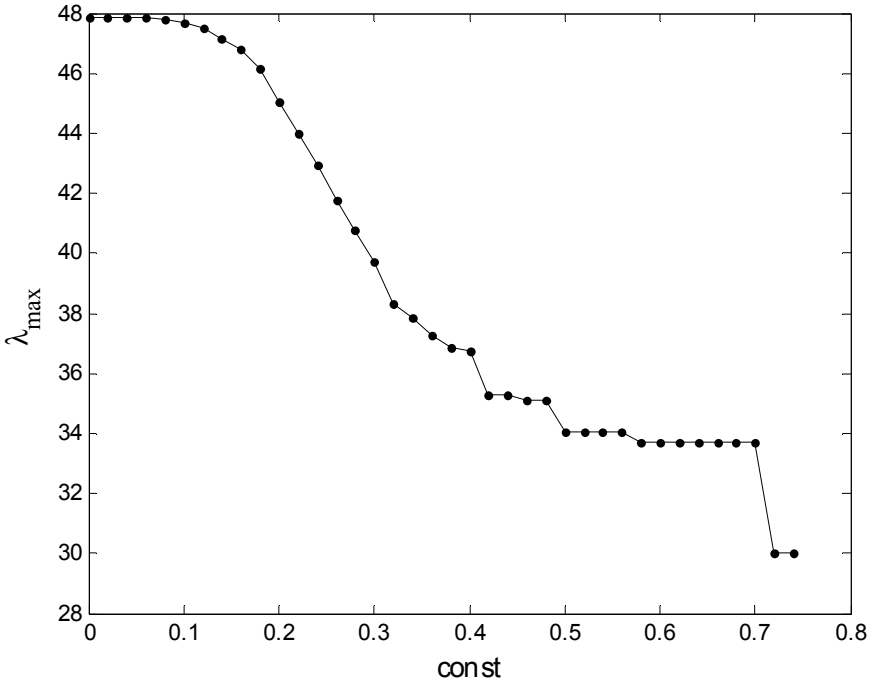


Fig.3.22 Largest eigenvalue of observed Cor retaining links with $c \geq const$ plotted with const.

These analyses reveal that *Cor* has a non random character but there is no significant information on distinctive or non-distinctive interactions.

In this chapter, I have tried to identify distinctive pairs of drugs using two measures of activity, similarity and variation. Network analysis shows that drugs contributing highly to chemical similarity in the network are not the same as those contributing to *s* or *d*. This means that there may be *ranges* of these interaction attributes –*c*, *s*, *d* where distinctiveness and non-distinctiveness become perceptible. The distribution of links in various (joint) categories of (*c*, *s*) and (*c*, *d*) confirms the preponderance decreases in the upper ranges, however, for links at any level of *s* or *d*, in the range *c*>0.3, interactions can be classified as distinctive and non-distinctive. Not differentiating between these kinds of interactions can lead to errors in prescriptions and adverse health outcomes. Therefore, as this chapter has shown, the difference between *s* and *d* can have implications for accurate identification of distinctiveness at multiple *levels*. Further, the standard methods of sali and correlation analysis are

insufficient for investigating distinctiveness. Measuring distinctiveness using a combination of attributes, *cs* or *cd* yields distinctiveness at different levels. These findings motivate a probabilistic analysis of distinctiveness at various levels in different ranges of the chemical space. This is introduced in the next chapter.

Chapter 4

TOWARD A CANYON CHARACTERIZATION OF THE ACTIVITY LANDSCAPE

The structural features of a drug compound describe its physicochemical properties that determine its biological activity on protein targets [21]. A vast number of combinations or analogs that result from slight changes in chemical structure may not contribute to the diversity in biological activity or functionality of drugs. We saw in chapter 2 that the drug target behavior of this data shown in Figs. 2.1, 2.2 resembled that previously established for different data. Few drugs are active on many more targets than average while a large number of drugs are active on much fewer targets [3,20]. This points to the discovery of certain molecular combinations having highly versatile functionalities. If targets are sufficiently distinct in their chemical nature, then the high druggability of few targets points to low specificity of a large number of medicines. Dissimilarity in activities of drugs when measured pairwise, often reveals pairs of drugs with highly dissimilar activity. Is variation in activity of drugs necessarily based on their chemical structures? In this chapter, I analyze in detail the structure activity association for the present data sample using the features of the *activity landscape*. The aim is to devise a quantitative characterization of the landscape for a precise identification and interpretation of the distinctive interactions.

Medicinal chemistry has generally relied on the similarity principle [1,13,14]. However, many exceptions are known. Development of medicines and analogs with many targets facilitates the production of multipurpose medicines. It also leads to the emergence of activity cliffs [17,18], specifying pairs of structurally similar drugs having highly variant biological activities. These cliff-like interactions between drugs correspond to extreme behaviors of pairs of chemically similar drugs showing unusually large deviations in activities. They are quantified by drops in similar activity of two

drugs. However, one may use the rise in pairwise variation of their activities instead. Distinctiveness can be measured as the either the drop in similar activity or rise in variant activity of two drugs.

Activity landscape of drugs (or the pharmacological topography) is a two-dimensional space of chemical structure and functionality to which a pair of drugs can be mapped. Researchers have focused greatly on quantitative structure activity models to study the activity landscape. Nonetheless, the effectiveness of these models has been found to be inadequate when studying activity cliffs, that is when small structural variations result in large changes in biological activity [1,17,18]. In previous work, structure activity analysis has relied mainly on measures such as sali (discussed in detail in 3.42) and some algorithmic or statistical analysis [17,18]. We saw in the last chapter, these methods and specifications of distinctive drug interactions have limitations as they are threshold driven and descriptive. Moreover, increased activity variation may not indicate decrease in activity similarity particularly when the magnitude of similarity is determined by number of commonly active targets. Previous studies have not differentiated between these two measures of quantifying the aberrations. Further, the magnitude of data considered for the analysis can significantly change the patterns of drug activity on targets and the inferences on similar activity [19].

This chapter provides a mathematical basis of the *extreme deviations* observed. It introduces a probabilistic analysis of the pharmacological topography and uses two measures of activity (similarity, s and variation, d) jointly with the corresponding chemical similarity c to investigate distinctiveness. It would facilitate identification of distinctive interactions on different scales of measurements. Probability distributions of s, d, c are used to identify medicinal categories involved in the highly distinctive interactions. I compute the predicted probabilities of s and d jointly with c. This gives the probability of distinctiveness at varying levels. The level of distinctiveness of a pair of drugs (or drug interaction) is specified according to the magnitudes of the attributes s, d, c for that pair. This chapter shows that the choice of the measurement (low s or high d) affects the interpretation of distinctiveness and the landscape. I use statistical logic based on established methods to argue that d is more efficient estimate than s for obtaining an optimal forecast (though neither of them can be used to construct a

statistically consistent estimate). In particular, the probabilistic analysis of distinctiveness using both measures induces a transition from cliffs to a new characterization of the landscape called activity *canyons*.

4.1 Non Random ψ^u, ζ^u

First, the spectral properties of drug network ψ^u are compared with those of a corresponding random network R. In this network ψ^u, two drugs are connected if they have at least one target in common. A network generated by random shuffling of binary activities of drugs in the bipartite network results in a random network of drugs R. The random shuffling preserves the sizes of the activity profiles of drugs. Since we are interested in the analysis of drugs, we do not shuffle the protein response profiles. The adjacency and Laplacian matrix for the drug network is the same as described in section 3.1. ψ^u shows non-random behavior as indicated in Fig. 4.1. In the exact same manner, Fig. 4.2 compares the eigenvalue distribution of the Laplacian of ζ^u with that of a corresponding random network R. ζ^u is also found to have a non-random behavior.

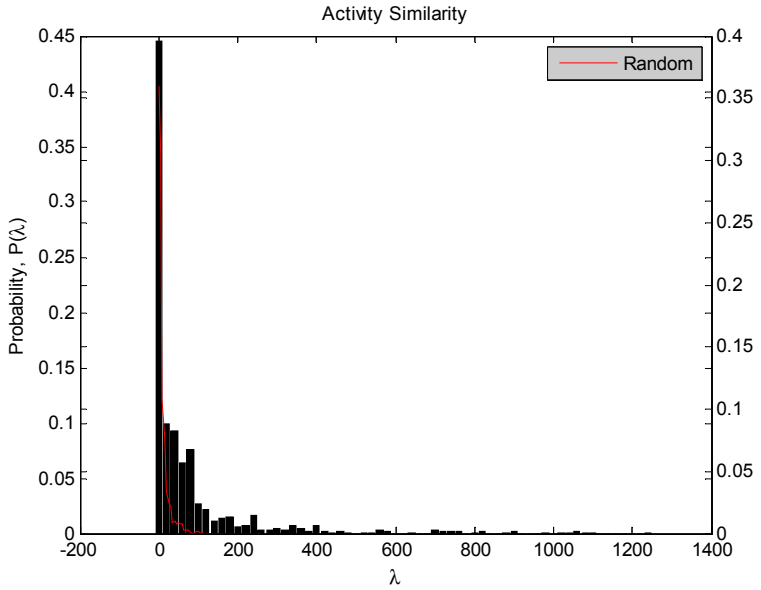


Fig.4.1 Comparison of probability distribution of eigenvalues λ of Laplacian of Ψ^u (L_ψ^u) defined in the previous chapter Eq.(3.2) with that of a randomly generated adjacency matrix R, averaged on 25 realizations. The eigenvalues of R are much smaller than those of L_ψ^u.

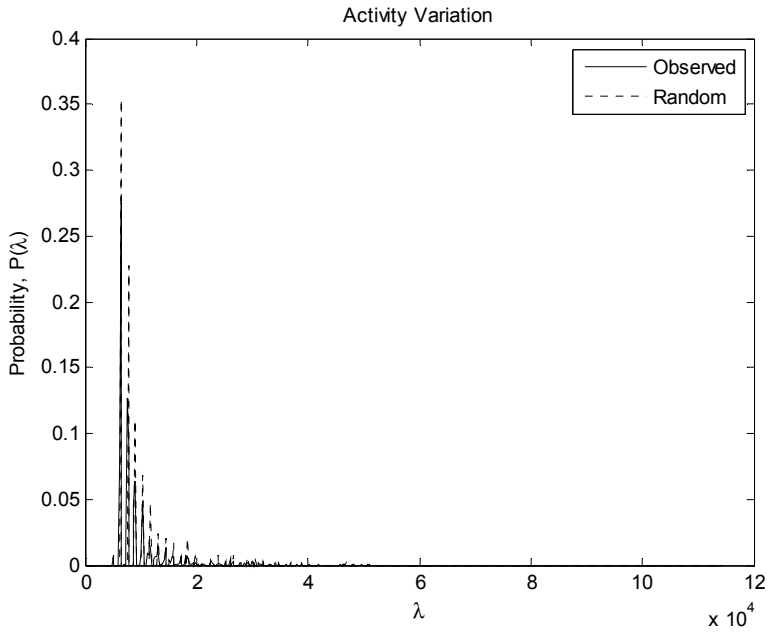


Fig. 4.2 Comparison of probability distribution of eigenvalues λ of Laplacian of ζ^u (L_ζ^u) defined in the previous chapter Eq.(3.2) with that of a randomly generated adjacency matrix R, averaged on 25 realizations. The eigenvalues of R are much smaller than those of L_ζ^u.

4.2 Distinctiveness in the Landscape : The Canyon Representation

Distinctiveness is the property of drugs having high structural similarity but varying activity profiles. An interaction between two drugs refers to the difference or similarity of activity of the drugs on the same protein. The discontinuities arising in the activity landscape due to sharp drops in similar activity along the chemical space represent drug interactions that are distinctive in nature [14,17,18]. They are not regular as they do not conform to the similarity principle. An interaction is *distinctive* when structurally similar drugs tend to behave oppositely on the same proteins. There may be two alternative specifications of a distinctive interaction in any pharmacological space. One, a pair of structurally similar drugs could have low similarity in activity. Two, the activity profiles of two structurally similar drugs may be highly dissimilar or variant. These are used alternatively for characterizing the well known phenomenon of activity cliff [17,18]. However, as shown here, this choice of measurement becomes crucial if *s* is measured by only the commonly active targets of drugs and ignores the commonly unresponsive proteins. Thus, the activity variation may not totally indicate

the magnitude of similarity of drugs. I determine which of these behaviors is dominant in the present interaction space using the measures of *s*, *d*.

4.21 Similarity and Dissimilarity in Activity

The weighted adjacency matrices Ψ and ζ of drug-drug interactions measure the magnitude of similarity and difference or variation in activity of drug pairs respectively. In the given pharmacological space, $0 \leq \Psi_{ij} \leq 28;\ 0 \leq \zeta_{ij} \leq 114$. As before, these interaction weights representing elements of Ψ and ζ are denoted by *s* and *d* respectively.

4.22 Comparison in the Chemical Space

This subsection examines how the measures vary on all pairs of drugs in different ranges of chemical similarity, *c*. Fig. 4.3 illustrates the structure-activity associations for *s* versus *d* in progressing windows of *c*. Detailed inspection suggests that although the averages of both *s* and *d* are more or less constant over the chemical space, in every window, the fractions of the weights of *d* found above average remain higher than fractions of weights of *s* found above average. These *d* weights tend to be higher in magnitude than *s* weights owing to constructions of *s*, *d*. While both *s* and *d* decline with *c* overall, it must be noted that the decrease in *s* occurs rather gradually and hence the cliff representation may be deficient. Further, the density of magnitudes (or number of points) found above average for *s* and *d* across each row of plots in Fig.4.3 is different. In all the subfigures, and particularly the ones in the intermediate region of the chemical space $0.3 \leq c \leq 0.65$, it is clear that the preponderance is more in the region marked by above average *d* than the corresponding region for *s*. This implies that a pair of structurally similar drugs is more likely to possess highly variant than highly similar activity.

Intermittent spikes in activity variation occur in the intermediate range of chemical similarity $c \in [0.3, 0.65]$ or $c \in [0.4, 0.65]$ with a dense layer of points beneath. This feature can be used to reasonably delineate the structure versus activity graphs from the perspective of activity variation.

Thus in terms of variation, the landscape maybe characterized as activity *canyons* or *gorges* depending on the steepness. This also applies to *s* but less distinctly.

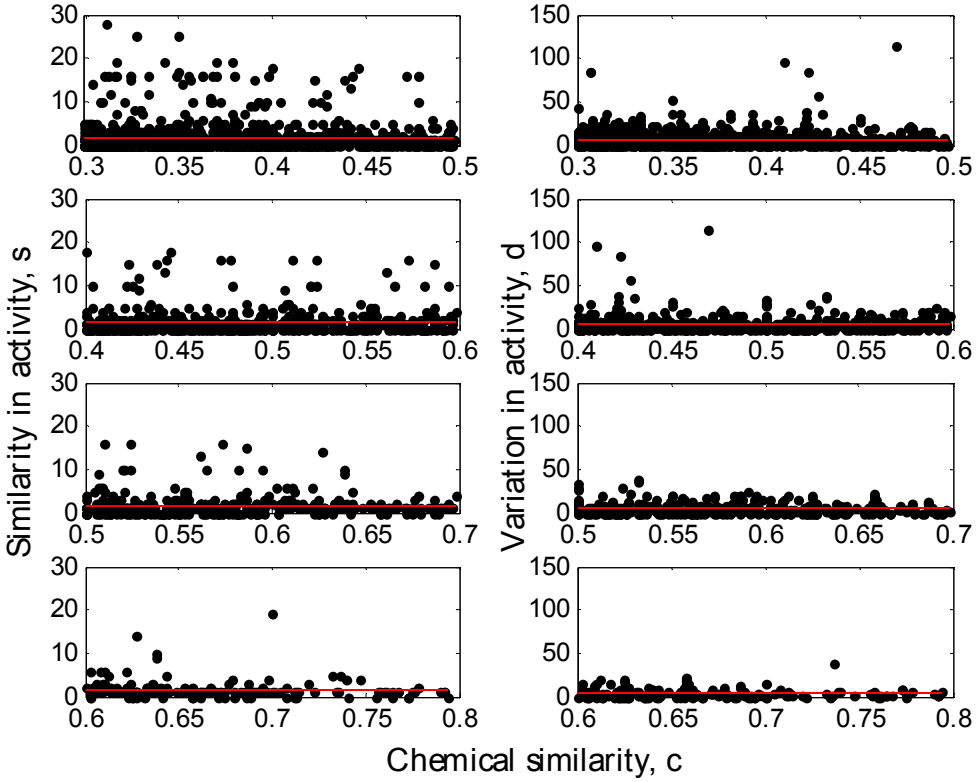


Fig. 4.3 (Left) Plotting s with c in four chemical similarity intervals (topmost to bottom in order) of [0.3, 0.5], [0.4, 0.6], [0.5, 0.7], [0.6, 0.8]. (Right) Plotting d with c in four chemical similarity ranges (topmost to bottom in order) of of [0.3, 0.5], [0.4, 0.6], [0.5, 0.7], [0.6, 0.8]. The horizontal solid red lines correspond to the averages of s and d in the respective intervals of c.

4.3 Structure Activity Quantification

In general medicinal chemistry has relied on quantification of structure activity to obtain a function *f* such that, change in physiological activity= *f* (change in structure) [1]. Here the nature of the function and hence the landscape characterization would depend on whether similarity or variation was considered for quantifying biological activity. We are interested in the two dimensional interactions relating structure (*c*) and activity (*s* or *d*). Measures like sali have been applied to quantify activity cliffs. However, as discussed earlier, high sali could also arise from high *c* (low *d*) or low *c* (high *d*). The overall score can be misleading for identification of distinctive links with high *c and* high *d*.

Therefore, in this chapter, I present a probabilistic analysis of the each of the measures *s* and *d* jointly with *c*. This informs us of not only the significance of cliffs but also where (if at all) they are found in the chemical space. The main purpose of this representation is for finding the probability of occurrence of various levels of (extreme) deviations. In other words, for every pair of drugs, it quantifies: *how distinctive and how probable?*. This was addressed in chapter 3 experimentally by calculating the empirically observed probabilities of different (joint) categories of *c*,*s* and *c*,*d*. Here, I approach it with a predicted probability model.

4.31 Predicted Probability

The joint probabilities of structural or chemical similarity with activity variation and similarity are approximated using the probability distributions of these individual properties Eqs.(4.1)-(4.3).

Assuming continuous distributions for the link weights as random variables for similarity (*s*), variation (*d*), chemical similarity (*c*), we can write the mathematical forms of the normalized probability distributions as

$$P(s) = \frac{0.045}{s_{min}} \left(\frac{s}{s_{min}}\right)^{-\alpha} \qquad (4.1),$$

$$P(d) = 0.042\, e^{\lambda d_{min}} e^{-\lambda d} \qquad (4.2),$$

$$P(c) = \frac{\gamma - 1}{c_{min}} \left(\frac{c}{c_{min}}\right)^{-\gamma} \qquad (4.3).$$

The behaviors are shown in Fig.4.4. Here s_{min}, c_{min}, d_{min} correspond to the minimum values of s, c, d respectively. We have $d_{min} = 0$ and for a well defined power-law distribution[7], we consider $s_{min} = 1$, $c_{min} = 0.004$.

This analysis is based on the distributions of s, d and c weights of all the links. Fig.4.4 shows the different kinds of probability distributions of drug-drug interactions in terms of $P(\psi = s)$, $P(\zeta = d)$, $P(C = c)$. While s and c conform to a power law probability distribution, d follows an exponential distribution[8]. Note that the distributions found are approximations used to provide a generalized prediction of the propensity of distinctiveness.

[7] The power law approximations in Eqs. 4.1 and 4.3 are normalized and the constants are determined by regressing probability and activity measures (s and d) on a logarithmic scale. This is shown as an activity canyon characterization of the pharmacological topography [21].

[8] The approximation in Eq. (4.2) is an exponential distribution. The constant λ is obtained as the coefficient of regression $\log(P(d)) \approx \lambda d$. The semi log plot of $\log(P(d))$ versus d is shown in Fig. 4.4.

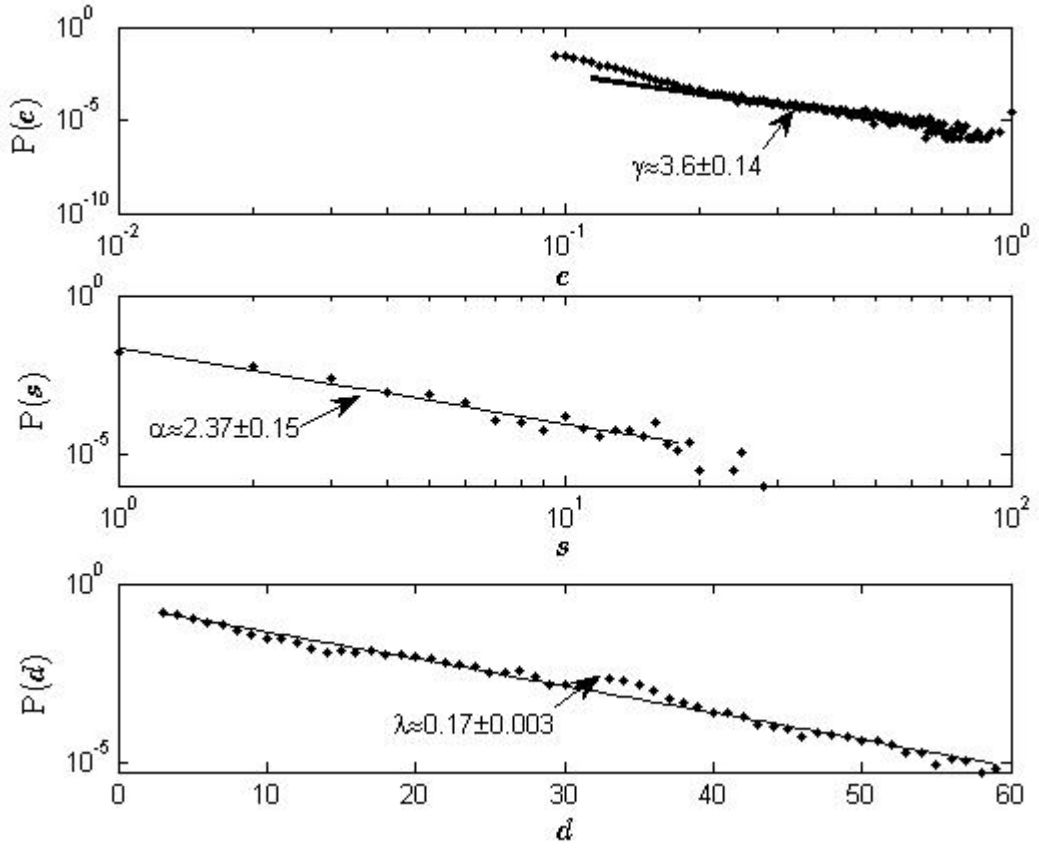


Fig. 4.4 (Top) Probability of pairwise chemical or structural similarities P(C=c) on a log-log plot. The straight line approximates power law behavior on a range of c with exponent 3.6±0.14.
(Middle) Probability of number of similar activities $P(\psi = s)$ *on log-log plot. The straight line approximates power law behavior on a range of s with exponent 2.37±0.15.*
(Bottom) Probability of number of dissimilar activities or variation $P(\zeta = d)$ *on a semilog plot. The straight line approximates exponential behavior on a range of d with exponent 0.17±0.003.*

This indicates that probability of finding extremely large *d* diminishes much faster than corresponding to *s*, however, the magnitudes of *d* are much higher than those of *s* when $c \geq 0.3$. Further, the power law exponents of *c*, *s* indicate that the second moment (indicative of the variance) of *c* is much higher than that of *s*, implying that *s* is more or less homogeneously distributed in the interaction space.

In this interaction based predicted probability model, the joint probability $P(s = s_0 \cap c = c_0)$ or $P(d = d_0 \cap c = c_0)$ for each pair of drugs specifies the chance of finding the level of distinctiveness shown by the interaction between that pair of drugs. s_0 , c_0 , d_0 are parameters denoting magnitudes of *s*, *c*, *d* which we vary. In the same way, one can specify ranges of study $s_1 \leq s \leq s_2$, $c_1 \leq c \leq c_2$, $d_1 \leq d \leq d_2$ and find the corresponding probabilities (of finding links with weights in those ranges).

The calculations below give the probability of finding links exhibiting a given level of distinctiveness (or non-distinctiveness).

If structure and activity are independently generated, then predicted joint probabilities are given by

$$P\big((c_1 \le c \le c_2) \cap (s_1 \le s \le s_2)\big) = c_{min}^{\gamma-1} s_{min}^{\alpha-1} (s_1^{1-\alpha} - s_2^{1-\alpha})(c_1^{1-\gamma} - c_2^{1-\gamma}) \quad (4.4)$$

$$P\big((c_1 \le c \le c_2) \cap (d_1 \le d \le d_2)\big) = c_{min}^{\gamma-1} \frac{0.042}{\lambda} \big(e^{-\lambda d_1} - e^{-\lambda d_2}\big)(c_1^{1-\gamma} - c_2^{1-\gamma}) \quad (4.5)$$

I provide the proof of the above statement in Appendix 3.1.

The two measures are compared by computing the predicted probabilities of *s* and *d* above their respective averages μ_s and μ_d. This is done by comparing Eqs. (4.4) and (4.5) taking s_{min}=1, c_{min}=0.004, d_{min}=0, $s_1 = \mu_s$, $s_2 = \infty$, $d_1 = \mu_d$, $d_2 = \infty$, in different ranges of [c1, c2] as [0.3,) [0.3, 0.65], [0.65,) for the purpose of integration. The standard deviations of the link weights of *s* and *d* are respectively σ_s and σ_d. In the range of $c \ge 0.3$, we have $\mu_s = 1.62$, $\sigma_s = 2.95$, $\mu_d = 4.85$, $\sigma_d = 7.53$. The values of these averages and standard deviations are computed for all progressing windows along the chemical space.

In order to compare *s* with *d* above their respective averages, we measure the predicted probability of finding $s \ge \mu_s + j\sigma_s$ and $d \ge \mu_d + j\sigma_d$ for *j*=0,1,2.. in the same chemical space. These are considered as measurements for the categories $P\big((s \ge \mu_s) \cap (c_1 \le c \le c_2)\big)$ and $P\big((d \ge \mu_d) \cap (c_1 \le c \le c_2)\big)$ for similar and variant activities respectively. Also calculated are the probabilities for standardized data where all the values of *s* and *d* are subtracted from their averages and divided by their standard deviations for a certain chemical space. Then the comparison is between the probability of finding *s* and *d* above their average=0 in all intervals of *c*. The probabilities calculated with the model for the categories *j*=0,1..8 for *d* are almost consistently and significantly higher than those for *s*. When they are not higher, the difference is insignificant. This is important and interesting to note and

is given in Fig. 4.5. It must also be noted that c is considered at least equal to 0.3, links below which have little consequence for a meaningful analysis, as in chapter 2.

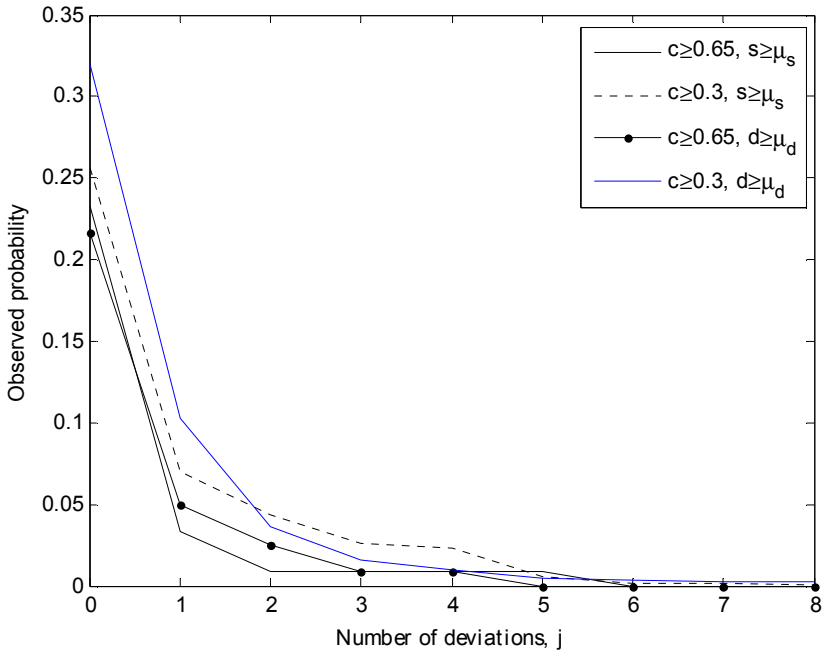


Fig. 4.5 Variation of observed joint probabilities of $c \cap s$ and $c \cap d$ with the number of standard deviations, j.

4.32 Validation

First, the predicted probability formulas (derived in Eqs. 4.4,4.5) are validated for the entire data of 1354 drugs by comparing the predicted probability with the observed probability for categories $s \geq \mu_s + n\sigma_s$ and $d \geq \mu_d + n\sigma_d$ taking $c \geq 0.3$. For this purpose, I take observations and compute predictions at n= 0, 0.2, 0.4,…9. I use the averages and standard deviations for the entire (original) data and compute RMSE for judging the agreement between the predicted probability (p_{pred}) and observed probability (p_{obs}) for s and d. $RMSE = \sqrt{\sum_j \frac{(p_{pred} - p_{obs})^2}{j-1}}$, j = 1,2… Note that the observed values are computed by considering the number of links sharing both properties of being within the given range of c and s (or d). The bivariate condition is expressed as $P(s \cap c) = P(s|c)P(c)$. If we find the joint probability of the two properties in a particular range of c then $P(c)$=1.

RMSE is as low as 0.0001 and 0.0006 for calculations of *d* and *s* respectively. It is remarkable that while generalized predictions are made with the assumption of independence of *s* and *c*, and, *d* and *c* on all pairs of drugs, they confirm the observed dominance of above average *d* values. Fig. 4.6 shows the agreement between observed and predicted values for *d*.

Figs. 4.7, 4.8 show the steady decrease in predicted and observed probabilities with increasing levels of screening (or *n*), the patterns are close to each other.

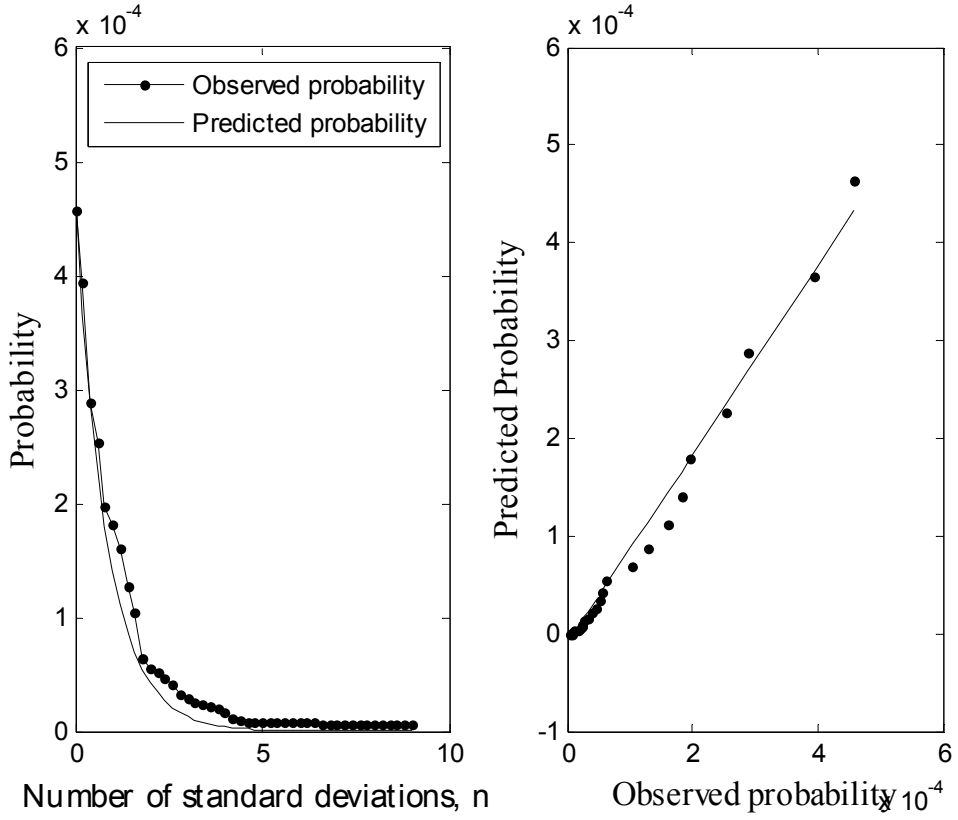


Fig. 4.6 Observed and predicted probabilities (scaled) for $d \geq \mu_d + n\sigma_d$ *and* $c \geq 0.3$ *plotted with n using the original data (left). Observed versus predicted probabilities showing a good agreement (right).*

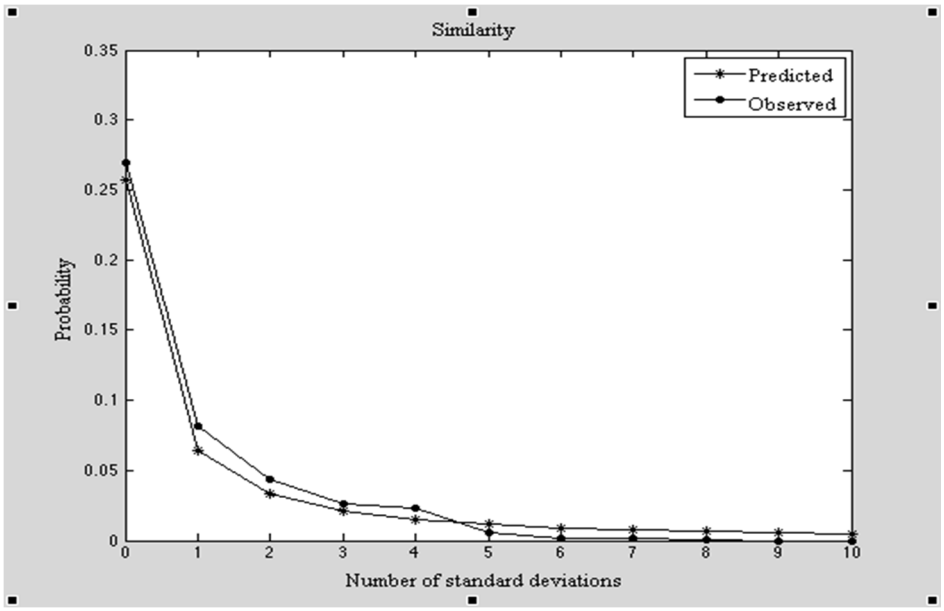


Fig. 4.7 Predicted and observed joint probabilities of $s \geq \mu_s$ *for* $c \geq 0.3$ *plotted with n. The observations are scaled by 40000.*

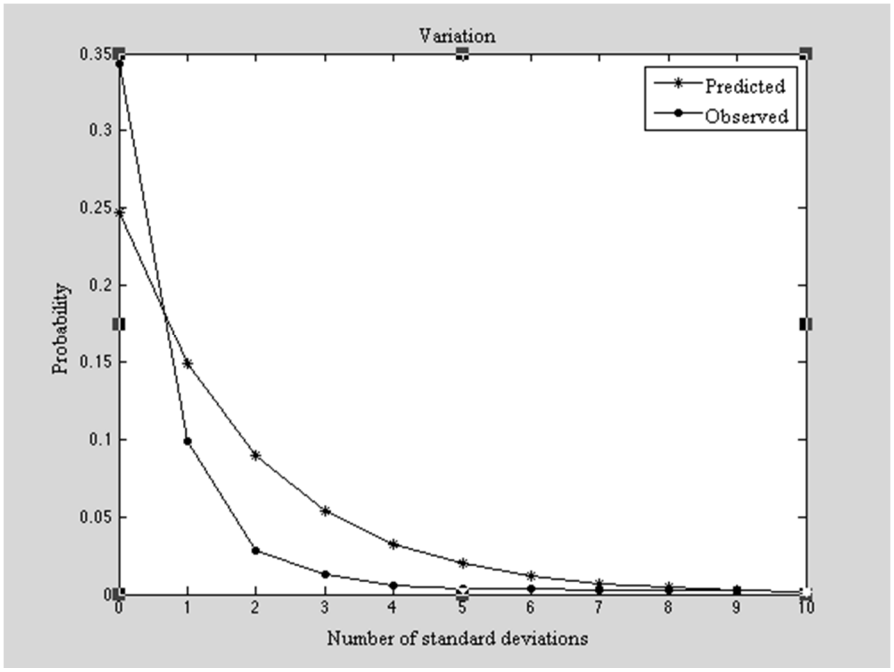


Fig. 4.8 Predicted and observed joint probabilities of $d \geq \mu_d$ *for* $c \geq 0.3$ *plotted with n. The observations are scaled by 100000.*

Next, I validate the predictions of the probability of distinctiveness using the system of drugs of the nervous system. This system consists of data on activity of 146 drugs on 219 protein targets. The probability distributions of interaction weights measured using pairwise *s*, *d*, *c* values for this system

are similar to those shown in Fig. 4.4 but the patterns are not identical. The probability distributions of s and d are indicated as power law and exponential respectively (as before) in Figs. 4.9, 4.10, there may be some difference in the patterns, exponents.

While the probability distribution of c is the same as that obtained above, the distributions of s and d show approximately similar patterns as there may be some deviations. As before, I compare the observed probabilities with the predicted probabilities using the formulas, for categories $s \geq \mu_s + n\sigma_s$, $d \geq \mu_d + n\sigma_d$ taking $c \geq 0.3$. Using the information for this drug ensemble c_{min}=0.0089,$s_{min} = 1$, $d_{min} = 0$, $\mu_s = 1.18, \sigma_s = 3.63, \mu_d = 16.67, \sigma_d = 13.39$, I show in Fig. 4.11 the agreement between observations and predictions for d. RMSE is 0.002 and 0.0006 for s and d calculations respectively. Notably, there is good agreement between observations and predictions despite the deviations[9] from the assumptions and conditions involved in the construction of the formulas.

[9] I might digress a little at this point to expand on the significance of the quantitative assessment of this information. The data put as a network has information on the drug generation carried on for a long time and we analyze the state to which the discovery has led. Thus, it is more about using this evidence to understand the evolutionary process. The network structure constructed out of an ensemble of drugs performing different functions may represent an integrative system- of individuals in the society or neural ensemble. The structure and biology of drugs in the larger ensemble are more diverse than in the one consisting of the drugs of the nervous system. In the networks constructed with both ensembles, drugs perform different functions and are connected by their structural and functional similarities or variations. Therefore, the integrative action of these topographies may be indicative of the evolution and the structural functionalism, and that the principles of this collectivism are the same for different or larger ensembles that may be created with time, similar to the cells in the democratic nervous system [29, 30]. Quantitative analysis of the given system is known to yield crucial information on evolution, for instance, the evolved regulatory processes in a cell or protoplasm established through "a harmonic relation between progressive events of the physiological dynamics of the cell". Researchers [30] have focused on subjecting "the physiological properties of the cell to the analytical mode of quantitative studies with their time-progressions in velocities, as against a mass estimation verified through experiments". Moreover, "the plant protoplasm has fundamental features of elasticities of various kinds with which it adjusts a harmonic progression of vital phenomena".
In the networks constructed, there may be drugs whose presence governs the collection significantly. However, with the constant addition and changes of drugs, the topographical changes would remain insignificant for a duration, due to a balance but the decisions in clinical preparation and marketing may become perceivably biased, gradually. These changes may not affect the conditional distributions to the extent of affecting the predictions on distinctiveness, and the samples considered maybe representative. Identification of distinctiveness in this case becomes vital for effective drug usage.

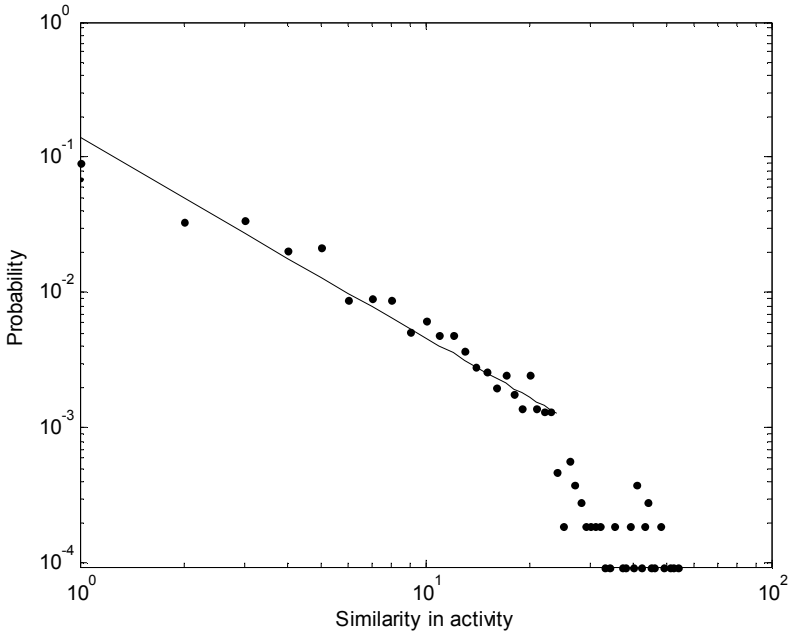


Fig. 4.9 Logarithmic plot of similarity in activity, s and the probability P(s) for the drug network of the Central Nervous System. The straight line has a regression coefficient~ 1.47.

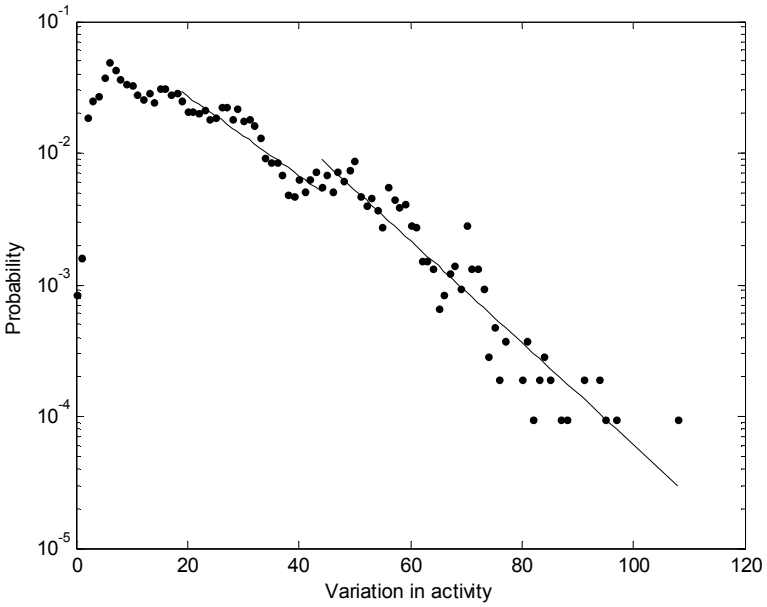


Fig. 4.10 Semilog plot of similarity in activity, d and the probability P(d) for the drug network of the Central Nervous System. The two straight lines indicate double scaling in the region.

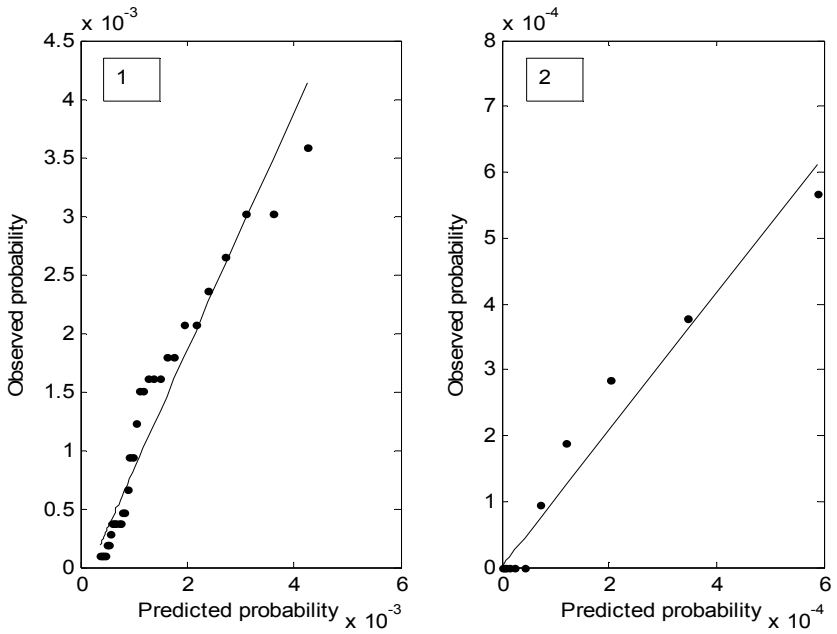


Fig. 4.11. Predicted versus observed probabilities for (1) s and (2) d for n=0, 0.2, 0.4,..9. The observations are scaled by a factor and the slopes of the straight lines are 1.06 (1) and 1.22 (2). RMSE is 0.0021 in (1) and 0.0006 in (2).

4.33 Hypotheses Tests

This subsection uses hypothesis testing to establish whether or not *d* dominates *s* by comparing observed probabilities (p_1, p_2) in regions spanning the chemical space $0.3 \leq c \leq 1$. The tests are listed below. Tests 1-3 show this comparison for $0.3 \leq c \leq 1$. Comparisons between different regions are given in tests 4-5. Test 6 gives the significance of variation as a refined filter of rare distinctive interactions. I report the test results along with significance level ρ,

1. $p_1 = P(c \geq 0.3 \cap s \geq \mu_s),\ p_2 = P(c \geq 0.3 \cap d \geq \mu_d)$

$p_2 > p_1$, $\rho < 0.01$

2. $p_1 = P(0.3 \leq c \leq 0.65 \cap s \geq \mu_s),\ p_2 = P(0.3 \leq c \leq 0.65 \cap d \geq \mu_d)$

$p_2 > p_1$, $\rho < 0.01$

3. $p_1 = P(c \geq 0.65 \cap s \geq \mu_s),\ p_2 = P(c \geq 0.65 \cap d \geq \mu_d)$

$p_2 > p_1$, ρ ~0.1

4. $p_1 = P(0.3 \leq c \leq 0.65 \cap s \geq \mu_s),\ p_2 = P(c \geq 0.65 \cap s \geq \mu_s)$

$p_1 > p_2$, ρ <0.05

5. $p_1 = P(0.3 \leq c \leq 0.65 \cap d \geq \mu_d),\ p_2 = P(c \geq 0.65 \cap d \geq \mu_d)$

$p_1 > p_2$, ρ < 0.05

6. $p_1 = P(c \geq 0.65 \cap s \leq \mu_s),\ p_2 = P(c \geq 0.65 \cap d \geq \mu_d)$

$p_1 > p_2$, ρ < 0.01.

Testing Methodology

Two statistical tests are employed in order to test the significance of the difference between the observed probabilities associated with activity variation and similarity. They are described below.

4.331 McNemar test for Comparing Proportions of *s* and *d* in a Chemical Region

I apply McNemar's test [31] to compare the relative frequencies (or probabilities) of *s* and *d* above their averages in the same chemical region $c \in [c_1, c_2]$. This test is suitable for comparisons of the two measurements made on the same links. The contingency table for a given chemical region is

	$d \geq \mu_d$	$d < \mu_d$
$s \geq \mu_s$	n_1	n_2
$s < \mu_s$	n_3	n_4

If $n_1 + n_2 + n_3 + n_4 = n$, we test whether or not marginal probabilities are equal : $\frac{n_1+n_2}{n} \_ \frac{n_1+n_3}{n}$ and $\frac{n_3+n_4}{n} \_ \frac{n_2+n_4}{n}$ The null and alternative hypotheses can be written as

H_0: $n_2 = n_3$

H_1: $n_2 < n_3$ (frequency of $d \geq \mu_d$ is greater than that of $s \geq \mu_s$ in the given chemical region)

H_0 is rejected in favor of H_1 at the level of significance ρ if $\frac{(n_2-n_3)^2}{n_2+n_3} > \chi_1^2$ or if the binomial probability $P(X \geq n_3 | n = n_2 + n_3, p = 0.5) \leq \rho$. The comparisons made here suit the design of the analysis, hence no corrections are required.

4.332 Paired t test

This test is used to compare the average of the difference between probabilities $p_1 = P((c_1 \leq c \leq c_2) \cap (s \geq \mu_s + j\sigma_s))$ and $p_2 = P((c_1 \leq c \leq c_2) \cap (d \geq \mu_d + j\sigma_d))$ observed (pairwise) at *j*=0,1,...8. All the probabilities so computed can be generalized as $p_1 = P((c_1 \leq c \leq c_2) \cap (s \geq \mu_s))$ and $p_2 = P((c_1 \leq c \leq c_2) \cap (d \geq \mu_d))$. These probabilities represent the results obtained by two different measures of interaction distinctiveness- *s*, *d*. In this way, we consider the variation of probabilities in the space above the average sampled at increasing number of standard deviations.

We can test the null hypothesis H_0 of equality of probabilities on average against one-sided alternative H_1

H_0: $\langle \Delta \rangle = 0$

H_1: $\langle \Delta \rangle < 0$ or H_1: $\langle \Delta \rangle > 0$

with the test statistic, $t = \frac{\Delta}{\sigma_\Delta / n}$ at significance level $\rho = 0.05$ for *n*-1 d.f. Here Δ represents the vector of differences, p_1 - p_2, σ_Δ is the standard deviation of Δ and sample size *n*=9 here. The null hypothesis of no difference is rejected in favor of alternate H_1 if the observed *t* is greater than the critical value of test statistic i.e. $t > t_c$ (0.05,8). This is applied for comparing probabilities of *s* and *d* in different chemical regions.

Results of Hypothesis testing

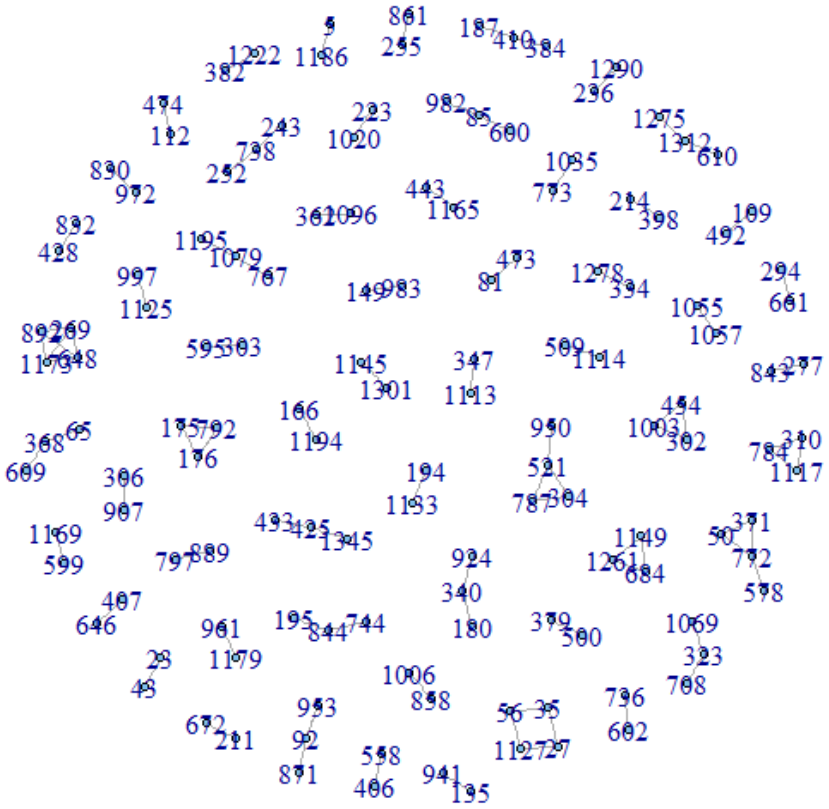
Results of the first three tests confirm that variant activity behavior dominates the similar activity behavior of the drug pairs consistently in the chemical space. The probability of finding above average

d magnitudes is more than that of finding above average *s* magnitudes in the region $c \geq 0.65$, but the difference is less significant. The intermediate region of the chemical space $0.3 \leq c \leq 0.65$ marks the transitional regime when structural similarity starts being positively associated with *d* more than with *s*. Distinctiveness emerges here. Tests 4, 5 indicate that *d* and *s* decrease with increase in *c*. However, *d* significantly dominates *s* in both proportion and magnitude particularly in the intermediate chemical region. The paired t test is applied to compare these probabilities as they specify the effect of changing the level or range of *c* of interactions on their *s* and *d* weights. Test 6 establishes the advantage of *d* as a potential filter of distinctive interactions. Few above average *d* weights (32.5%) are higher in magnitude than 75% of below average *s* weights (Fig.4.12(i)-(iii)). This helps to identify all the rarely occurring distinctive interactions, most of which are not revealed with *s* but are crucial for characterization of activity landscape. Further, *d* yields distinctiveness at multiple levels of hierarchy set by criterion of the number of standard deviations above the average.

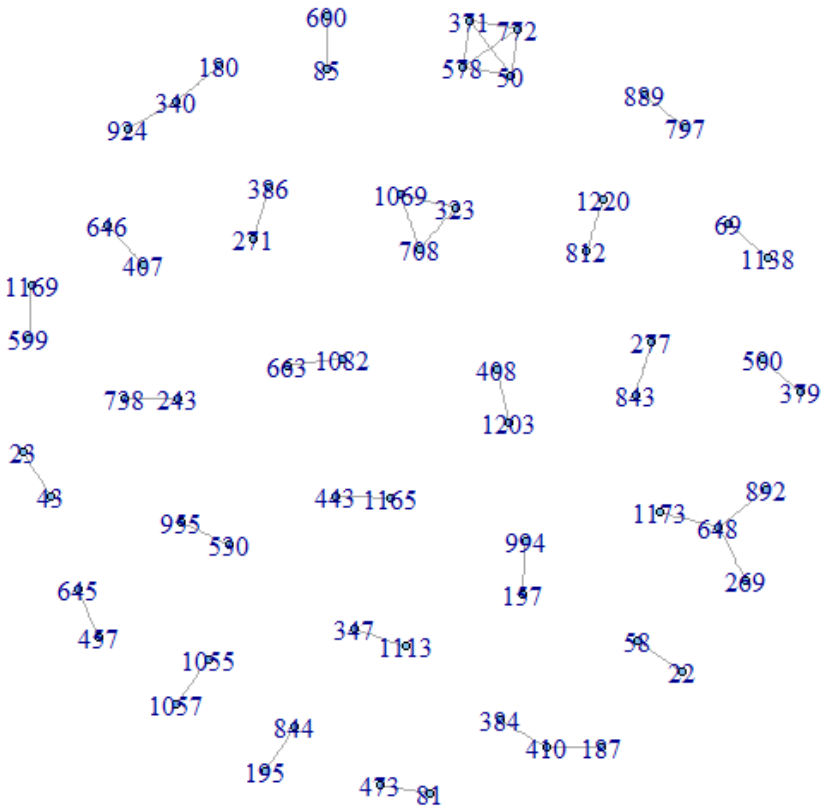
Most distinctive interactions occur between vitamins and medicines curing ailments of classes such as central nervous system or dermatological in nature. Several less similar interactions filtered as distinctive, include compound pairs from distinct classes. Fig.4.13((i),(ii)) shows structural similarity of Sitagliptin and Nefazodane that affect insulin release and the nervous system, respectively. Interestingly, highly variant interactions are fewer and more distinctive as they include drugs and their structural analogs belonging to the same class but performing intricately different functions. Fig.4.13((iii)-(vi)) shows structurally similar Loxapine and Amoxapine affect the nervous system but Amoxapine is more versatile. Felodipine and Clevidipine are analogs for curing hypertension but Clevidipine treatment is more advanced. Thus, activity variation calculations are advantageous for identifying medicinal analogs that provide more specialized treatments for same kinds of ailments. The activities of such medicinal analogs typically vary a lot. This measure is particularly helpful in making finer distinctions between medicines having apparently similar treatments. It can therefore identify selective interactions, which would ensure a better matching of drugs to desirable treatments.

Tests1-3 and 6 use McNemar test. Test 6 compares *d* as a filter for highly distinctive drug interactions with *s*. We compare the probabilities $p_1 = P(c \geq 0.65 \cap s \leq \mu_s - n\sigma_s)$ and $p_2 = P(c \geq 0.65 \cap d \geq \mu_d + n\sigma_d)$, *n*=0,1,2.. For *n*=1, $p_1 = 0$ and $p_2 = \frac{11}{120} = 0.09$. The *d* measure yields maximally distinctive interactions as *n* is increased, implying distinctiveness at multiple levels. Fig.4.12(iii) indicates this, and for *n*=1, I report the pairs of drugs that are most distinctive. The proportion and nature of interactions filtered are significantly different from that obtained with *s*.

(i).



(ii).


600
180
85
340
924
889
797
386
1069
323
1220
646
271
708
812
69
1158
407
1169
599
297
663
1082
408
500
843
379
738
243
1203
28
48
443
1165
892
1173
955
648
590
994
645
269
157
457
347
1113
58
1055
22
1057
844
384
410
187
195
473
81

(iii).

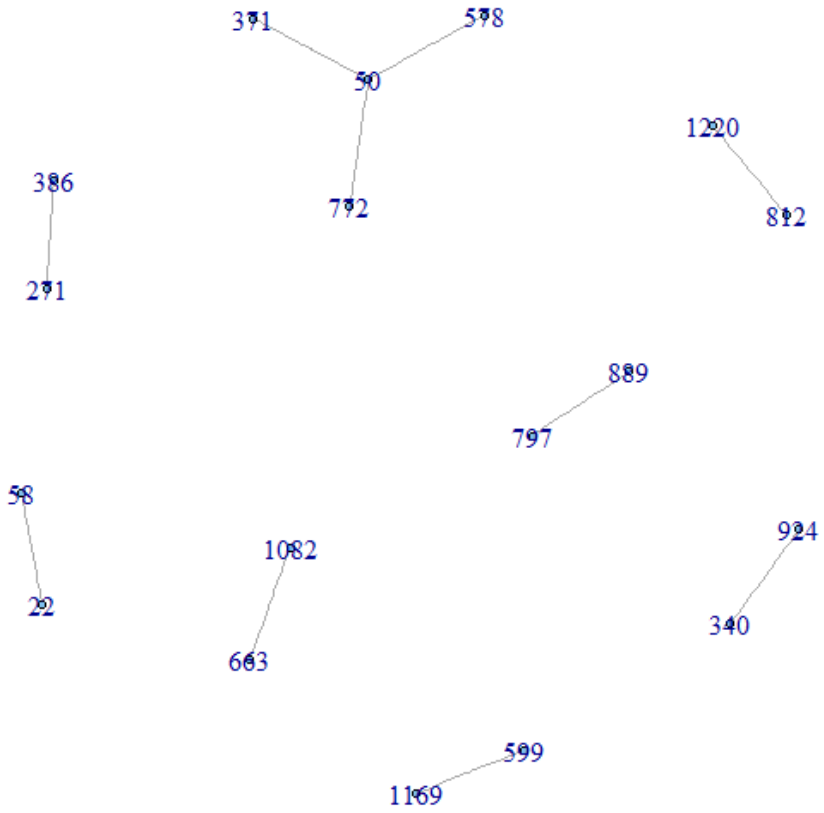
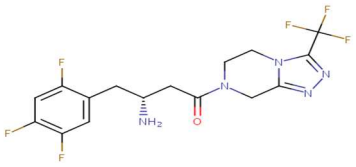


Fig. 4.12 (i) Network of 91 distinctive interactions determined by s below average

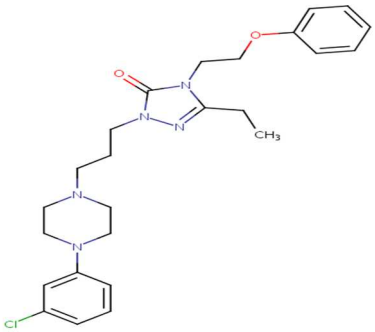
(ii). Network of 39 distinctive interactions determined by d above average.

(iii) Network of 11 distinctive interactions determined by d by calculation for $d \geq \mu_d + \sigma_d$. *The medicine links are (Lysine-Ornithone), (Adenosine monophosphate-Adenosine triphosphate), (Vitamin A- Alitretinoin), (Vitamin A- Tretinoin), (Vitamin A- Isotretinoin), (Propoiomazine-Aceprometazine), (Loxapine-Amoxapine), (Dicloxacillin-Cloxacillin), (Pseudoephidrine-Ephedrine), (Tioconazole-Miconazole), (Felodipine-Clevidipine).*

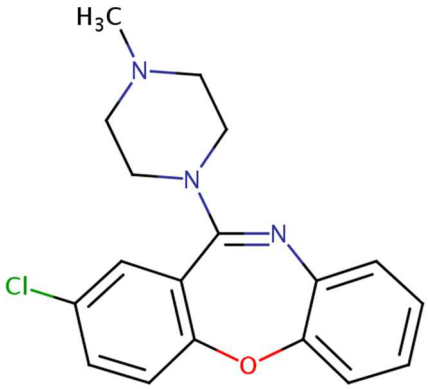
(i).



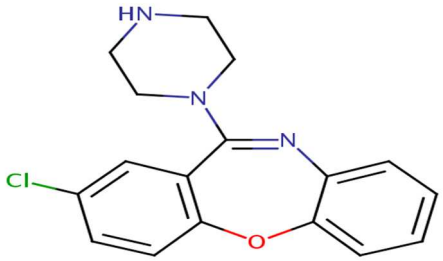
(ii).



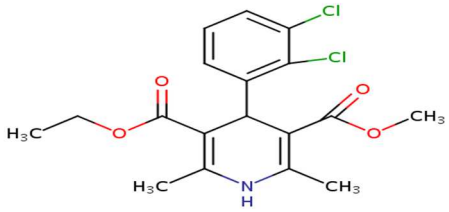
(iii).



(iv).



(v).



(vi).

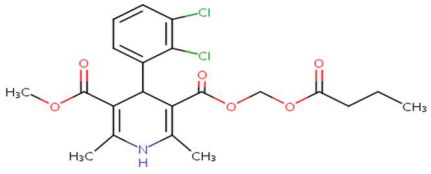


Fig. 4.13 (i). Sitagliptin- used for control of type 2 diabetes mellitus, increased release of insulin.

(ii). Nefazodone-analogous to Sitagliptin but affects central nervous system, has a palliative action.

(iii). Loxapine-antipsychotic drug.

(iv). Amoxapine- analogous to Loxapine but used for many other neurotic disorders and sedation.

(v). Feldopine- calcium channel blocker for moderate hypertension.

(vi). Clevidipine- analogous to Feldopine, is calcium channel blocker, but for advanced treatment of blood pressure.

4.4 Implications of the Characterization

There are two contributions of the analysis in this chapter. One is a methodological contribution as it provides a mathematical basis for identifying distinctive interactions. The predicted probability analysis introduced here is based on simple assumptions but predicts the overall distinctiveness in various chemical regions to a reasonable accuracy. The second contribution is the canyon representation of the activity landscape of drugs or as we refer to it, the pharmacological topography. This representation is appropriate for the probabilistic analysis of distinctiveness because the two measures of similarity and variation are on highly different scales in the data. When similarity of activity is considered as given by the commonly active targets, some medicinal pairs may show both distinctiveness and non-distinctiveness and cliff representation is not sufficient.

In the present pharmacological space, distinctive interactions are more probable than non-distinctive interactions. Non-distinctive interactions are regular as they conform to the similarity principle. Distinguishing among the distinctive interactions requires precision. The range of *d* is much higher than that of *s*. Quantitative comparison of the landscape reveals that structural and biological similarities, two a priori independent manifestations of drug interactions, are unevenly associated and maybe complementary. The presence of distinctiveness is highlighted by both *s* and *d*. As chemical similarity increases, both measures decrease. However, the decrease in *s* is more significant than that in *d*, as *d* maintains its above average value. Activity variation is a more suitable measure for characterization of the pharmacological interaction space than activity similarity[10]. Intermittent rises in *d* with chemical similarity maybe interpreted as activity *canyons* or *gorges* of varying levels. In this representation, the rises in both measures can be assessed relative to their respective ranges of magnitudes and compared across all pairs. The rises in activity similarity and variation are interpreted as increase in non-distinctiveness and distinctiveness respectively, characterized as activity canyons instead of cliffs. Therefore, distinctiveness is a general term that may represent abrupt deviations in the landscape as cliffs or canyons depending on the method of measurement. Moreover, the canyon

[10] Appendix sections 3.1-3.3.

representation allows for the quantitative analysis of 'how distinctive and how probable?' [11], demarcating the pharmacological subspace for finding distinctiveness. These rises in activity variation are probable[12] all along the chemical space and more distinct than drops in *s*. The predicted probability model introduced in this chapter helps in identifying distinctive interactions and the probability of their occurrence in the landscape. The crucial merit of this approach is that whatever the assumptions and considerations or probability distributions in the construction of the formulas maybe, the predictions would also apply to data that are different in nature. The model helps in identifying distinctive drug associations that are not only significant and rare but also less apparent.

[11] Multilevel distinctiveness is captured in this model by allowing a specification of the levels of *c*, *s*, *d* of the interactions. According to the Eqs. 4.4, 4.5, the probability of an interaction being distinctive (or non-distinctive) can be found in the range $[s, \infty]$, $[d, \infty]$, $[c, \infty]$.

[12] The probability at the level *s*, *c* (or *s*, *d*) is given by the predicted estimate $\theta = k_1 s^{1-\alpha} c^{1-\gamma}$ (or $= k_2 e^{-\lambda d} c^{1-\gamma}$) as α, γ are greater than 1, and k_1, k_2 are constants in Eqs. 4.4, 4.5. Further, if $c = c_0$, then the estimates using both measures can be compared with mean square error of $\theta(s|c)$ and $\theta(d|c)$ as the optimum prediction criterion: $M_\theta = Var_\theta + bias_\theta^2$. Thus the optimal prediction (optimum M_θ) by the measures depends on the variance and bias generated with their estimates $\theta(d), \theta(s)$. Now $Var\big(\theta(s)\big) = k_1^2 c_0^{2(1-\gamma)} Var(s^{1-\alpha})$ and $Var\big(\theta(d)\big) = k_2^2 c_0^{2(1-\gamma)} Var\big(e^{-\lambda d}\big)$.

Using Taylor's expansion, $Var(s^{1-\alpha}) = \mu_s^{1-\alpha} - 0.5\alpha(1-\alpha)\mu_s{}^{-\alpha-1} Var(s)$ and $Var\big(e^{-\lambda d}\big) = e^{-\lambda \mu_d} + 0.5\lambda^2 e^{-\lambda \mu_d} Var(d)$. As we know, *s* is distributed as Pareto power law with $2 < \alpha < 3$ for which the variance diverges in the limit of infinite size of the network and would be large for finite size networks. And as we know that *d* has an exponential distribution. Its variance ~ λ^{-2}, $\lambda > 0$, is finite. Therefore, for any limited or even large bias, using the efficiency of the two measures for an estimate of the predicted probability defined as $Var\big(\theta(s)\big)/Var\big(\theta(d)\big)$, it turns out that *d* would be better and more reliable measure for an optimal prediction.

Chapter 5

INFORMATION, IMPLEMENTATION, AND MEASUREMENT

The predicted probability model for distinctiveness developed in the previous chapter was shown to rely on the probability distributions of the two properties considered for distinctiveness (*c*, *s*) or (*c*, *d*). It was able to account for the observed distinctiveness despite the assumption of independence of the probabilities of the two interaction attributes involved, and the approximations applied. If the interaction of a pair of drugs is distinctive, then it possesses both high chemical similarity *and* highly variant (or less similar) activity. Is the knowledge of the distribution of chemical similarity of links between drugs indicative of the nature of their similar or variant activity (*s* or *d*) distributions? It would be interesting to investigate this. I apply information theoretic measures [32, 33] to find the mutual dependence of the pair of variables or properties- (*c*, *s*) or (*c*, *d*). Using mutual information, one may infer whether and how much the joint distribution of probabilities of these variables is similar to the products of their marginal distributions. Further, it gives the relative entropy or the amount of information associated with a property. Here, it can be used to infer whether or how much knowing chemical similarity of a link reduces the uncertainty of *s* and *d* attributed to the link. This can provide a justification for the using both attributes simultaneously for classifying the interactions as distinctive[13].

[13] Conditional probability $P(d \geq \mu_d | c \geq c_0) = \frac{P(d \geq \mu_d \cap c \geq c_0)}{P(c \geq c_0)}$. If *d* and *c* are independent features of a connection, $P(d \geq \mu_d | c \geq c_0) = P(d \geq \mu_d)$. Distinctiveness involves *d* and *c*, however, some connections may have both high *d* and high *s*. Using Bayes theorem with all the features of a connection, probability of distinctiveness is $P(d \geq \mu_d \cap c \geq c_0) = \sum_i P(s_i) P((d \geq \mu_d \cap c \geq c_0) | s_i)$, summing on all discrete levels of *s* observed. It should be obvious that if a connection has a high level of *s*, then the measure of distinctiveness is not independent of non-distinctiveness. Therefore, how good a measure *d* is for distinctiveness would depend on *P(s)*. If higher levels of *s* occur with very low probability, then the non-distinctiveness in the sum can be ignored and *d* is an adequate measure. We can think similarly for *s*. This emphasizes the importance of knowing the probability distributions of all features of a connection.

5.1 Mutual Information, Uncertainty, and Relative entropy

Mutual Information of two random variables *X* and *Y* is given by

$$I(X;Y) = \sum_y \sum_x p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad (5.1).$$

p(x,y) is the joint probability distribution. Considering *X=c*, *Y=s*, their joint probability distribution can be computed.

The range for analysis of this data is $c_1 s_1 \leq cs \leq c_2 s_2$ and $c_1 d_1 \leq cd \leq c_2 d_2$ and all the links are distributed in the discrete categories $0.3 \leq c \leq 0.9$, $0 \leq s \leq 28$, $0 \leq d \leq 114$. Using Eq.(5.1), I compute the empirical mutual information by summing over all values which gives $I(c,s) = 2.16 \times 10^{-4}$ and $I(c,d) = 6.56 \times 10^{-5}$. These values are very close to 0, the value expected when the variables are independent. Fig. 5.1 shows the variation of $I(c = 0.3, s \geq \mu_s + n\sigma_s)$ and $I(c = 0.3, d \geq \mu_d + n\sigma_d)$ with *n.* Increasing *n* increases the level at which mutual information for distinctiveness and non-distinctiveness is measured by $I(c,d)$ and $I(c,s)$ respectively.

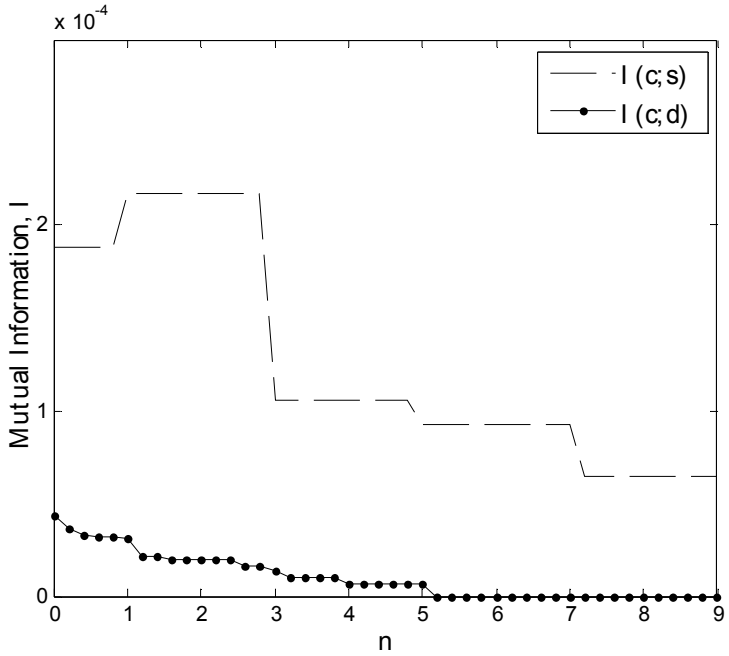


Fig. 5.1 Mutual information $I(c \geq 0.3, s \geq \mu_s + n\sigma_s)$ *and* $I(c \geq 0.3, d \geq \mu_d + n\sigma_d)$ *plotted with n.*

The mutual information decreases as the level of identification of distinctiveness becomes more selective, i.e. as *n* increases. This would imply that specific information about the levels of each of the attributes *c*, *d* (or distinctiveness of) a link becomes important. In other words, for finding links at high levels of distinctiveness, the relevance of both *c and d* increases. It is interesting that the mutual information of non-distinctive interactions also decreases as the level of screening for non-distinctiveness increases but it is still higher than *I*(*c*, *d*).

I use a related measure, the Uncertainty coefficient *U* to determine the predictability of one variable when the other is given [32]. It is normalized mutual information. In general, for variables *X*, *Y*,

$$U(X|Y) = \frac{I(X,Y)}{H(X)} \qquad (5.2)$$

where $H(X) = -\sum_x p_X(x)\log(p_X(x))$ is the entropy associated with *X*.

The computed *U*(*c*|*s*), *U*(*s*|*c*), *U*(*c*|*d*), *U*(*d*|*c*) plotted with *n* show that apart from *U*(*c*|*s*) (and *U*(*c*|*d*) to some extent), all other quantities are small. Thus there is some (at least a small, finite) predictability that can be achieved for levels of *s* and *d* when *c* is specified but it is not true the other way around. If *s* or *d* levels are given, we cannot be sure which *c* level the links belong to.

This can be further investigated by finding the relative entropy of probability distribution of chemical similarity of links with respect to that of activity similarity (and activity variation). Kullback-Leibler divergence (KLD) gives the increase of information [33] when *c* instead of *s* (or *d*) is used to characterize distinctiveness. Symmetrized KLD is useful for analyzing the differences in the contribution of links to the divergence according to the levels of their features *c*, *s*, *d*. In the context of distinctiveness, this involves comparison of probabilities of links by categorization of their weights *w* in terms of two properties - w_c (the weight of chemical similarity) and w_s (or w_d), weights of similarity (or variation) in activity. This procedure is also used in text categorization research. All the links are analyzed individually and the weights of all attributes given by w_c, w_s, w_d are considered for

every link. It is examined how probable it is to find weights such as these. These probabilities are then compared. The assignment is done by sampling the data into broad categories of $c \geq const = 0.3, 0.5, 0.7$. In each sample,

let $p_c = P(w_c = c)$, $p_s = P(w_s = s)$ and $p_d = P(w_d = d)$, then

$$KLD(c,s) = \sum_{all\ links}(p_c - p_s)\log\left(\frac{p_c}{p_s}\right) \qquad (5.3a)$$

$$KLD(c,d) = \sum_{all\ links}(p_c - p_d)\log\left(\frac{p_c}{p_d}\right) \qquad (5.3b).$$

By assigning the links to the categories and computing the probabilities, I compute *KLD*(*c*, *s*) and *KLD*(*c*, *d*) using Eqs. 5.3a, 5.3b. It must be noted that the categories of *s* and *d* are further divided as $s \geq \mu_s + n\sigma_s$ and $d \geq \mu_d + n\sigma_d$ for *n*=0,1,2... Also, since the links are dealt with individually, the categories of *c* are continuous as they appear in the data. It turns out that *KLD*(*c*, *s*) decreases much slower than *KLD*(*c*, *d*) and *KLD*(*c, d*) is very small.

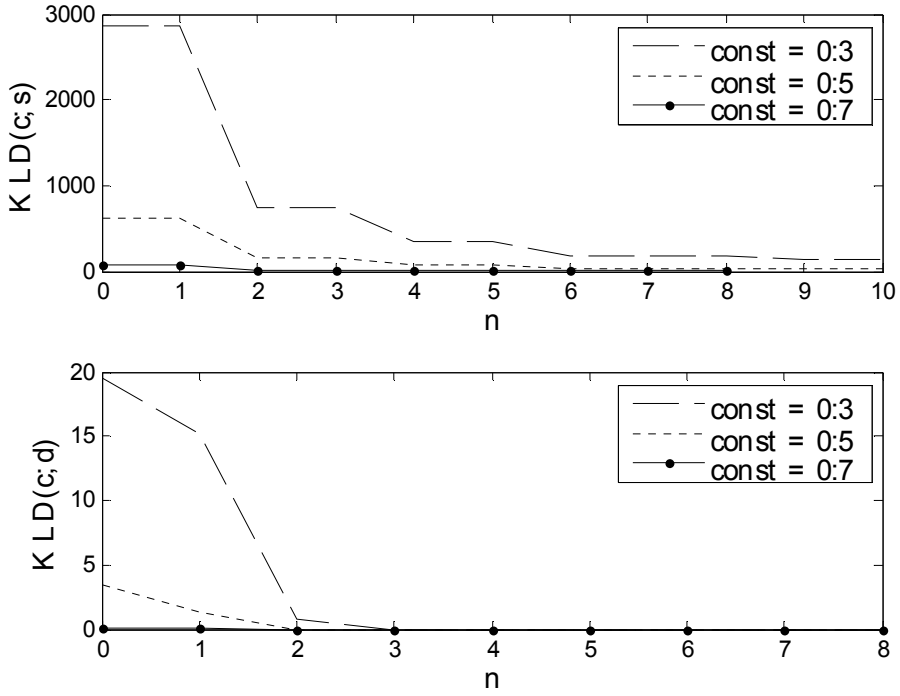


Fig. 5.2 Variation with n of KLD(*c,s*) *(Top) and KLD*(*c*, *d*) (*Bottom*) *for data samples obtained by removing redundant links having* $c \geq const = 0.3, 0.5, 0.7$.

KLD decreases in all cases shown in Fig.5.2. The decrease in KLD implies decrease in relative entropy. It is not surprising that it is the highest when *const* = 0.3 because it is the most dense category with most of the links having chemical similarity in this range. However, it is interesting that in this and all other ranges specified by *const* = 0.5, 0.7 the *KLD*(*c, d*) is always much lower than *KLD*(*c*, *s*) and vanishes quickly with increase in *n*. If $p_c = 0$ and, or $p_d = 0$ then the interactions (or links) are not distinctive. If $p_c = 0$ and, or $p_s = 0$ then the similarity principle is contradicted. Moreover these links contribute nothing to *KLD.* A decrease in *KLD* would mainly occur with increase in such cases because it becomes rarer to find links of high *s* and high *d* as *n* increases. It may be interesting to note that few links that contribute to these categories – high *c*, *s* and high *c*, *d* are rare, and more importantly, it is *rarer* to find *highly distinctive* links (high *c*, *d*) than to find highly non-distinctive links (high *c*, *s*).

For quantifying the rareness of such distinctive links, I define a *rareness coefficient*, $r \geq (p_c p_d)^{-1}$, for $p_c \neq 0,\ p_d \neq 0$[14]. This means that the links that correspond to the least probable chemical space as well as the least probable *d*, score the highest on rareness or are the rarest distinctive. Rareness is also a measure of the information content in the link and the uncertainty of its attributes (*c*, *d*)[15]. By replacing p_d with p_s in calculation of *r*, we can obtain the rareness of the non-distinctive interactions. The rareness of the distinctive interactions can be ranked using this scheme (fixed by *r*). Figs. 5.3-5.5 below show the number of distinctive interactions divided into quadrants of rareness, *r* depending on p_c, p_d.

[14] This requirement is to avoid a singularity. If either $p_c = 0$ or $p_d = 0$ or both then such a link is infinitely rare, meaning it would be impossible to find this level of distinctiveness (specified by *c*, *d*) in the present data.
[15] This rareness coefficient can be considered proportional to the information content in the interactions. Taking logarithm of *r*, we can write $\log r = -\log p_c - \log p_d$. Thus for an interaction with highly probable attributes [32], the uncertainty is very low, but it is very high when at least one of the attributes of the interaction is highly improbable.

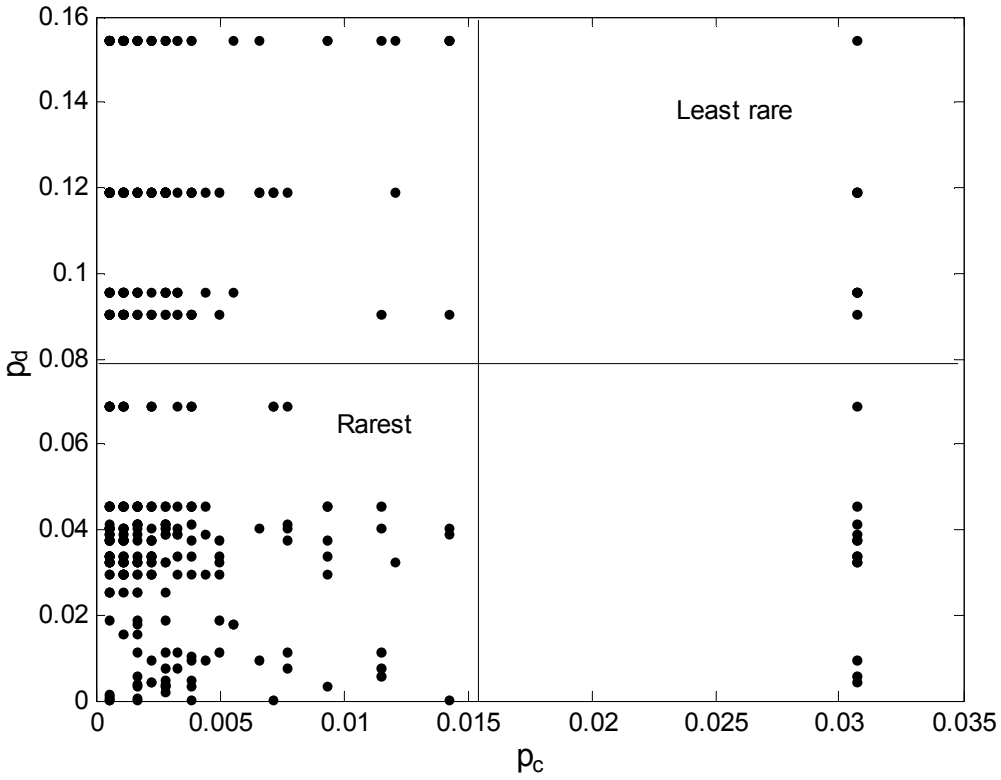


Fig. 5.3 Plot of $p_d = P(w_d = d \geq \mu_d)$ *on the vertical axis with* $p_c = P(w_c \geq 0.3)$ *on the horizontal axis for all links in the data sampled at const=0.3. The quadrants divide the 419 distinctive interactions on the basis of their rareness, indicating least and most rare quadrants.*

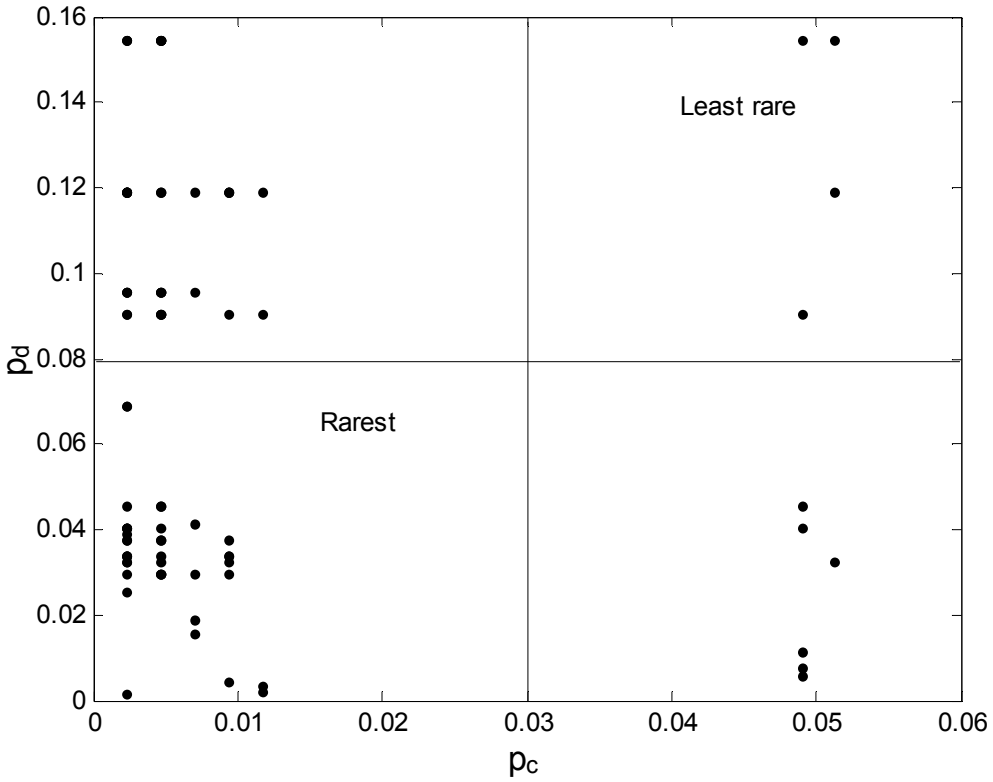


Fig. 5.4 Plot of $p_d = P(w_d = d \geq \mu_d)$ *on the vertical axis with* $p_c = P(w_c \geq 0.5)$ *on the horizontal axis for all links in the data sampled at const=0.5. The 86 distinctive interactions are represented and the least and most rare regions of the plot are indicated.*

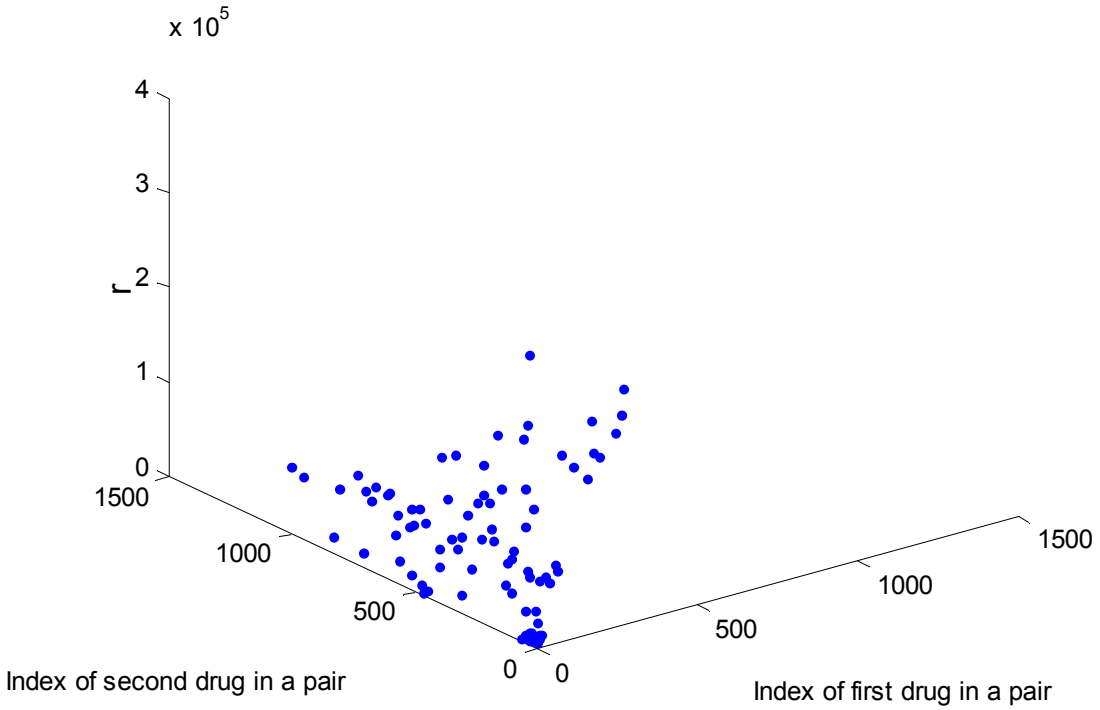


Fig. 5.5 Drug index plot of drugs involved in the distinctive interactions shown in Fig.5.4 for $p_d = P(w_d = d \geq \mu_d)$ *and* $p_c = P(w_c \geq 0.5)$ *in the data sampled at const=0.5.*

For *const*=0.3, most distinctive interactions score highly on rareness but both the number of most distinctive interactions and *r* reduces for *const*=0.5. In the category of most highly distinctive interactions, which is specified by *const*=0.7 and $d \geq \mu_d$, the 6 rare (not the rarest of all) distinctive interactions are

(*Adenosine monophosphate- Adenosine triphosphate*), (*Flucloxacillin- Dicloxacillin*), (*Adreparin-Amoxapine*), (*Alitretinoin- Tretenoin*), (*Tretenoin-Isotretenoin*), (*Pseudoephedrine-Ephedrine*).
The medicines constituting these pairs are structurally similar and their functionalities are not identical but quite similar, as they perform intricately different functions.

In the category of most highly non-distinctive interactions (conforming most to the similarity principle) which is specified by *const*=0.7 and $s \geq \mu_s$, the 4 rare (not the rarest of all) non-distinctive interactions are

(*Adenosine monophosphate- Adenosine triphosphate*), (*Nisoldipine-Nifedipine*), (*Eszopiclone-Zopiclone*), (*Adreparin-Amoxapine*).

The medicines constituting these pairs are identical in their functioning as per the description in the drug bank. Note that the pairs (*Adenosine monophosphate- Adenosine triphosphate*) and (*Adreparin-Amoxapine*) are both distinctive and non distinctive. This is different from the observation in chapter 3. This points to the importance of method of analysis as it may render the interpretation of distinctiveness as subjective.

5.2 Perturbation and its Implications for Identifying Distinctive Interactions

Drug discovery often relies on optimization algorithms such as minimum, maximum spanning trees for finding structurally diverse compounds and in other applications [33]. Such medicinal compounds may or may not be functionally similar and it is important to investigate that for distinctiveness. I will close this thesis with a discussion of how and whether distinctiveness is affected by *perturbations*. In quantum mechanics, a system is perturbed slightly so that the deviations in the energy and eigenstates represent corrections to states of the original system. In chapter 3, we saw the effect of retaining links possessing a minimum level of *c* in the network using un-weighted adjacency matrices of C^u, Ψ^u, ζ^u. Different kinds of distinctive interactions were identified by modifying the networks. Distinctiveness was also studied using weighted matrices of *cd*. High value of *cd* on links indicated high level of distinctiveness but the level of distinctiveness was tuned by the parameter: number of standard deviations above mean. Those interactions that survived in distinctive categories of large *n* were not only the ones possessing greatest distinctiveness but also the most robust. In this last section, I apply a perturbative approach to investigate the robustness of distinctive interactions, and to thereby assess the sensitivity of the system (drug network) to the perturbations.

Specifically, I iteratively perturb the weight matrix *W=cd* by reducing the elements of the matrix each time by a small fraction. I wish to analyze the impact this has on our perception of distinctiveness. If the elements of the weighted adjacency matrix *W* are perturbed uniformly[16] as $W_{ij} \rightarrow W_{ij} - pW_{ij}\ \forall i,j$

[16] This is simple. It would be useful to consider other perturbations.

for $0 < p < 1$, for a total of n' iterations, then the corresponding change in eigenvalues of *W* at the end of n' iterations is proportional to $\Delta_{n'} = 1 - (1 - p)^{n'}$. This is given by the Theorem in Appendix 4. It must be noted that the value *p* that would cause an optimum (or minimum) deviation depends on the number of iterations. I apply different values of p (Corollary 1 in Appendix 4 gives the optimum perturbation criterion and the optimum relation between p and n') for n' iterations to look for any modifications in classification of distinctive interactions at different levels of perturbation. The minimum level of perturbation represents the weakest disturbance to the system and any modifications in distinctiveness would represent corrections in identification.

I compare the principal eigenvector of *W* before and after perturbation. The maximum eigenvalue (λ_{max}) of the matrix decreases steadily as *p* increases. See Fig. 5.6. As for the perturbation applied previously, that is removing links having *c<const*, the decrease in λ_{max} with increase in restriction on the retainment of links is expected in accordance with the Theorem (Appendix 2) and shown in Fig.5.7. In Fig. 5.8, I plot the drop in λ_{max} due to perturbation of individual vertices, keeping rest of the network fixed. The vertex whose perturbation causes maximum drop would satisfy Corollary 2 in Appendix 4. When we apply the minimum perturbation using *p*=0.2, $n' = 4$, the decrease is shown to be the greatest for the vertex representing drug Adenosine triphosphate[17]. In the decreasing order of their impacts on $\lambda_{max}(W)$, the drugs representing the vertices are Adenosine monophosphate, Adenosine, Vidrabine, Riboflavin, S-Adenosinemethionine. This minimum perturbation applied to the elements of any other drugs does not result in any change in $\lambda_{max}(W)$. It must be noted that most of these are constituents of highly distinctive interactions indicated Fig. 3.17 except Riboflavin.

[17] This result would remain the same for *p*=0.2, $n' = 1$ applied in accordance with the corollary.

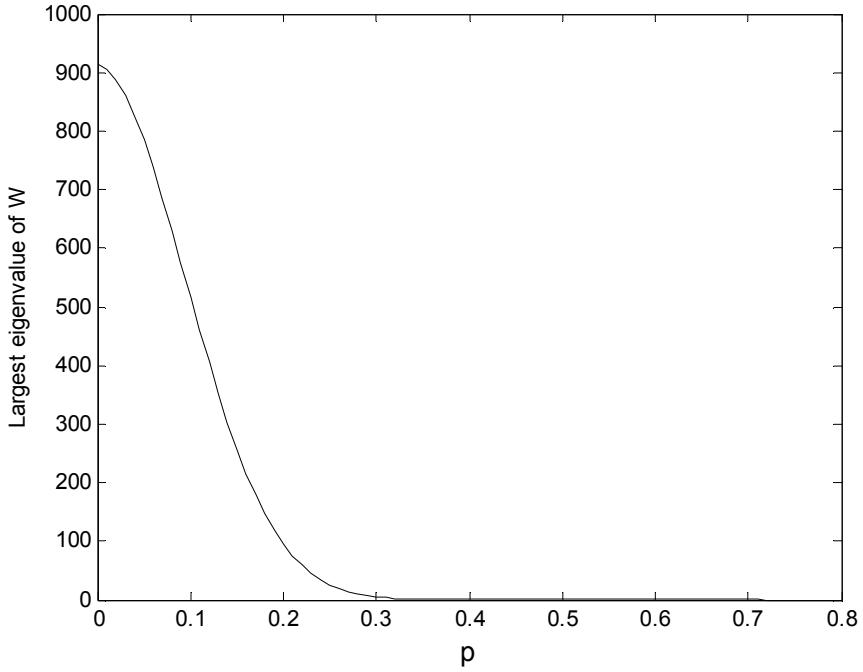


Fig. 5.6 Variation of λ_{max} of W with the perturbation parameter p.

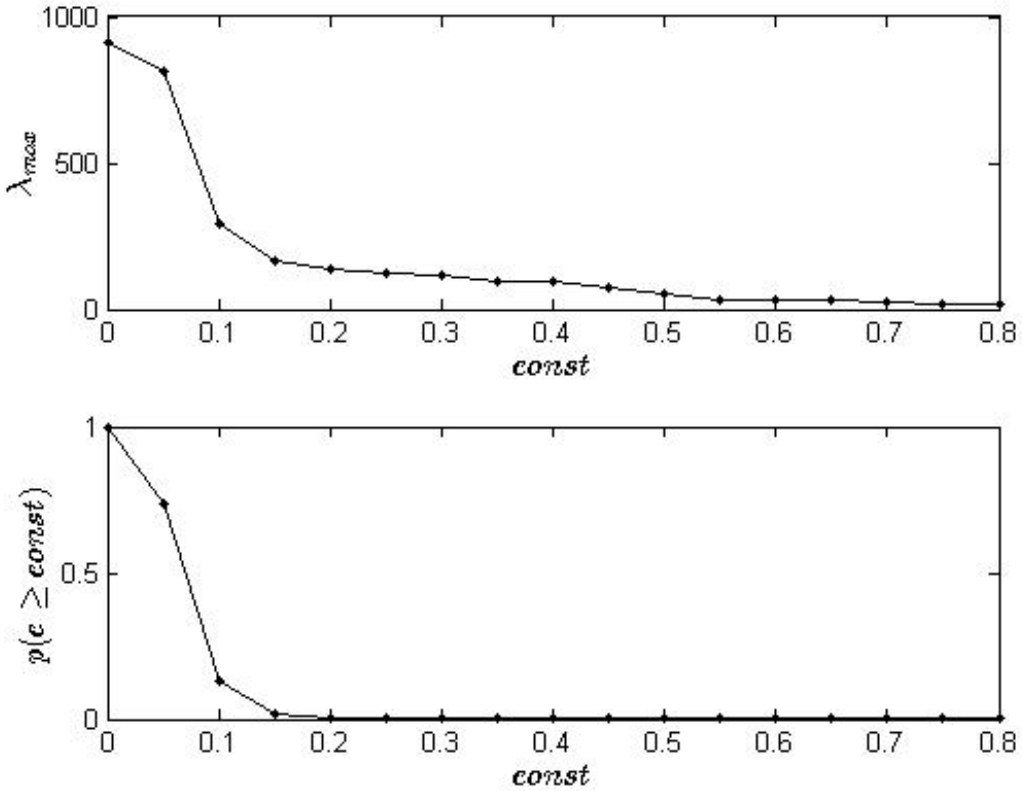


Fig. 5.7 Variation of λ_{max} of W with const (Top) and of the probability p(c≥const) with const (Bottom).

Now, I remove the redundant links for which c<0.3 from the network and analyze the effect of perturbing $W \to W - pW$ once and $n' > 1$ times, on the apparent distinctiveness that emerges. I (minimally) perturb the matrix using p=0.2 once and for $n' = 4$ times, and p=0.3 for $n' = 2$ times. To compare the distinctiveness before and after perturbation, I filter the links in the perturbed network whose weights are n standard deviations higher than the average of the weights in the unperturbed matrix, W. Fig. 5.9 compares the number of distinctive interactions in the perturbed and unperturbed network as n increases. The perturbation reduces λ_{max} and shifts the components of the corresponding eigenvector from negative to positive. Evidently, the decrease in number of distinctive interactions filtered is sharper for the perturbed network and this allows us to analyze the difference. For instance, for n=34, we are able to retrieve the 15 most highly distinctive interactions using the perturbed W. This is much lower as compared to 90 distinctive interactions predicted by the unperturbed W using the same criterion for filtering, which is $cd \geq \mu_{cd} + n\sigma_{cd}$. The identification of distinctiveness is highly sensitive to such perturbations.

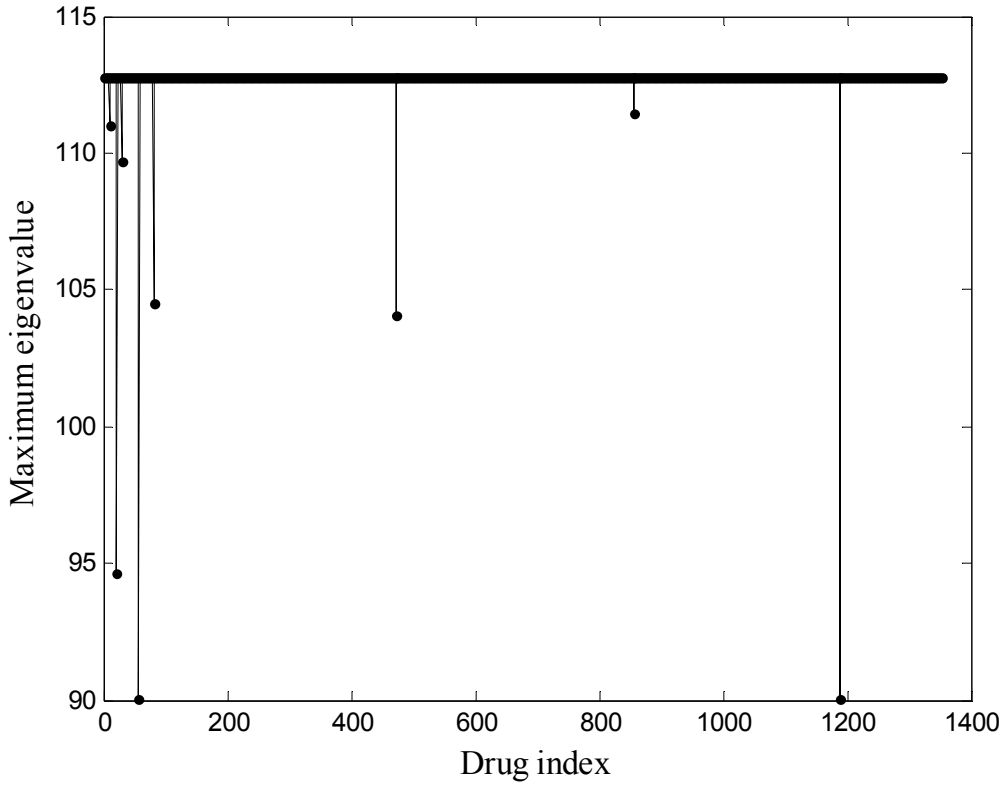


Fig. 5.8 The largest eigenvalue of the W plotted with the index of the individual drug whose elements are perturbed using p=0.2, $n' = 4$. *The solid line represents the largest eigenvalue of the unperturbed weight matrix.*

Moreover, the 15 most highly distinctive interactions are not only the same as those identified in Fig.3.17, but also highly robust, as confirmed by this analysis. The drugs constituting these are the also ones causing significant drops in largest eigenvalue discussed above. Fig. 5.10 shows that the impact of the minimum perturbation applied p=0.3, $n' = 2$ is smaller. Hence the number of distinctive interactions decreases slowly with increase in n which makes many other distinctive interactions perceptible (as *distinctive*). For instance, for n=34, we obtain 22 distinctive interactions and for n=30, there are 43 distinctive interactions including those 15 interactions observed above. This shows how sensitive the system is to change in perturbation and therefore points to the *adaptive* nature of the *interpretation* of what is distinctive. The network in Fig. 5.11 shows all of these, and the 43-15=28 interactions are also indicated. Here, an additional interaction identified as (L-Aspartic acid, L-Serine) was not considered distinctive in Fig.3.17. It is interesting to see interactions between drugs affecting the same classes such as the nervous system but having functional diversity. There are also few interactions affecting ailments of different classes. One can continue this analysis for various levels of optimal perturbation and observe the number and kinds of distinctive interactions that can remain classified as distinctive for any kind of perturbation.

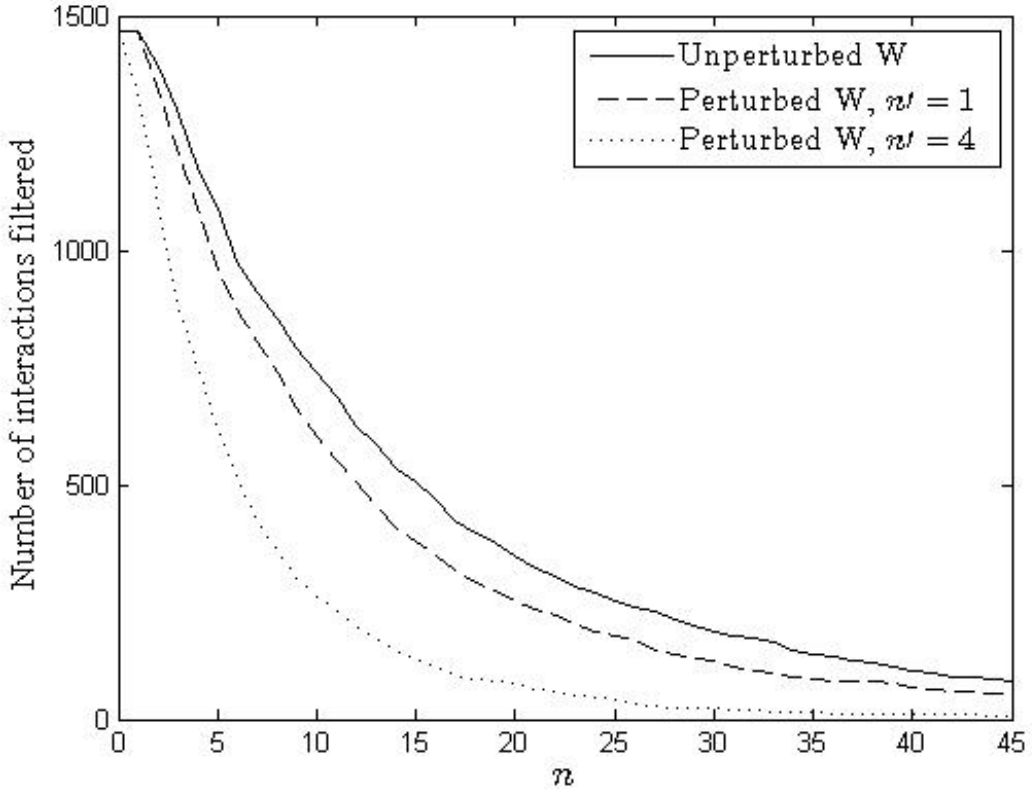


Fig. 5.9 Number of distinctive interactions filtered as links whose weights obey the criterion $cd \geq \mu_{cd} + n\sigma_{cd}$ *plotted with n. p=0.2.*

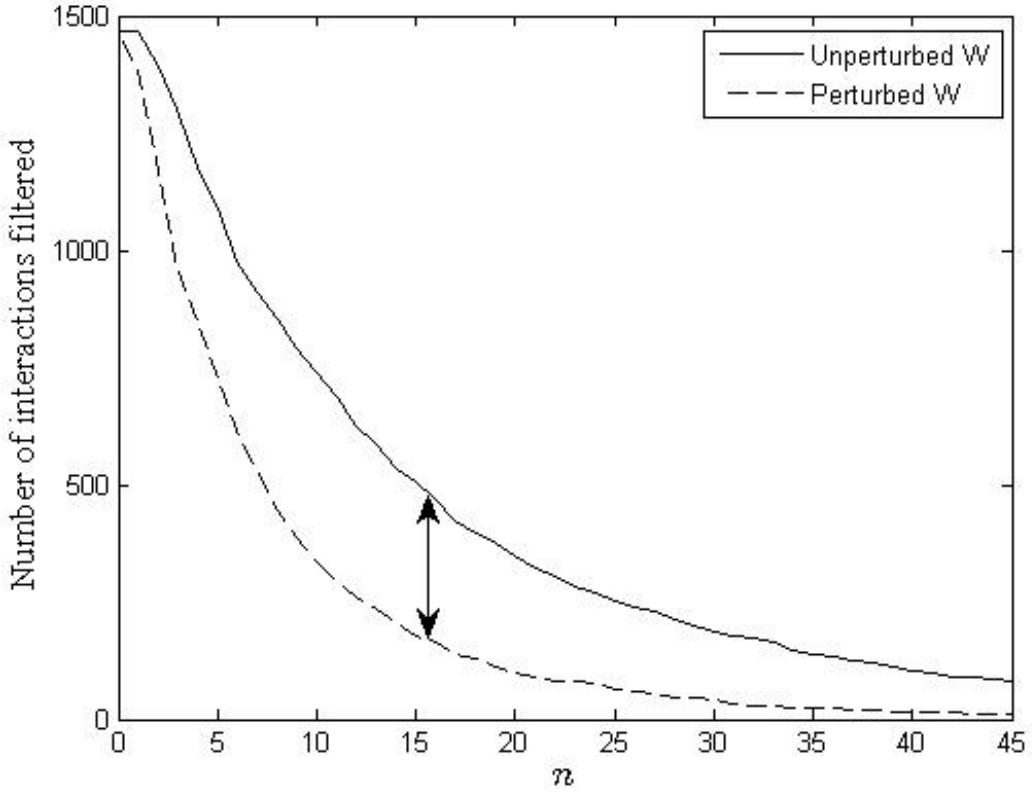


Fig. 5.10 Number of distinctive interactions filtered as links whose weights obey the criterion $cd \geq \mu_{cd} + n\sigma_{cd}$ *plotted with n. p=0.3,* $n' = 2$.

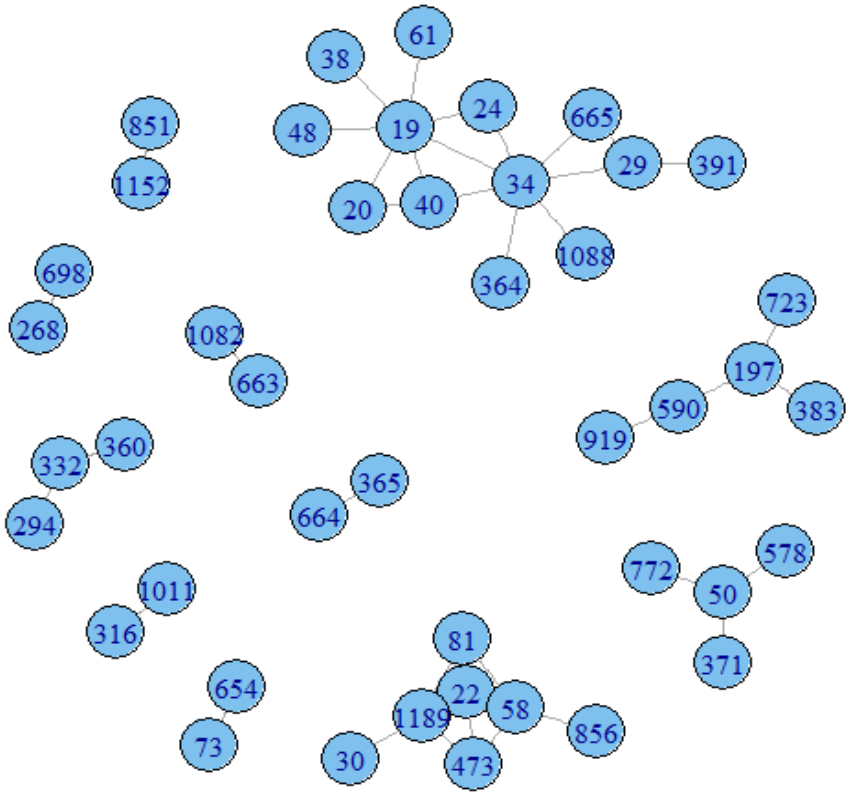


Fig. 5.11 Network of distinctive interactions filtered using the criterion $cd \geq \mu_{cd} + n\sigma_{cd}$ *for n=30. The distinctive interactions in addition to those indicated in Fig.3.17 are: (L-Aspartic acid, L-Ornithine), (L-Aspartic acid, L-Serine), (L-Aspartic acid, L-Alanine), (Adenosine monophosphate, Vidrabine), (Adenosine monophosphate, Adenosine), (L-Ornithine, L-Cysteine), (L-Serine, Glycine), (Succinic acid, Azelaic acid), (Succinic acid, Aminolevulinic acid), (Glycine, L-cysteine), (Glycine, Aminocaproic acid), (Glycine, Aminolevulinic acid), (Glycine, Dihydroxyaluminium), (Vitamin A, Alitretenoin), (Vitamin A, Tretenoin), (Vitamin A, Isotretenoin), (Lorazepam, Oxazepam), (Amitriptyline, Notriptyline), (Amitriptyline, Olapatidine), (Amitriptyline, Cyclobenzaprine), (Alprazolam, Trizolam), (Prochlorperazine, Chlorpromazine), (Chlorpromazine, Triflupromazine), (Dextromethorphan, Levorphanl), (Olapatadine, Doxepin), (Clonazepam, Nitrazepam).*

It is fascinating that high number of distinctive interactions obtained with the unperturbed matrix in this case reveals a lot of hidden distinctiveness. For instance, there are interactions such as

(Adenosylmethionine, Adenosine monophosphate) that are highly distinctive but sensitive to perturbations of the system. This is crucial for any analysis of distinctiveness in drug interactions.

CONCLUSIONS

This thesis quantitatively analyzes the structure activity associations in a large collection of drugs and their targets. In particular, I study *distinctive* interactions between pairs of drugs occurring in the pharmacological activity landscape. I consider interactions between pairs of drugs as connections and assess them using three attributes of their chemical similarity (*c*), magnitudes of activity similarity (*s*) and activity variation (*d*). A preliminary analysis of the drug network constructed through commonality of targets between drugs gives limited evidence for the prevalence of similarity principle. Few drugs with many targets share similar activity with many drugs but few of these drug connections have high chemical similarity.

Distinctive pairs are identified as those with high *c* and *d* or with high *c* and low *s*. This means that there are medicines having selective treatments and distinctiveness distinguishes these *selective* interactions. While previous studies on activity cliffs have considered similarity and variation in activity as alternate specifications for quantifying distinctiveness, their methods do not define the *level* of distinctiveness well. Furthermore, they consider activity similarity as determined exactly by variation. I point out here that they may not necessarily be indicative of each other. In the given drug ensemble, I observe that the pairwise interactions are distributed across multiple levels or categories of *c* and *s* or *c* and *d* jointly. The probability of finding links in upper ranges of both of the combined categories decreases, yet, it is nonzero. This shows the presence of multilevel distinctiveness and non distinctiveness in the pharmacological landscape.

The present dissertation focuses on a generalized prediction or identification of multilevel distinctiveness based on the data. In this respect, the analyses make two important contributions. First, is a methodological contribution in the form of a predicted probability model for identifying distinctive interactions, commonly studied as activity cliffs in literature. It is also able to predict the distinctiveness observed in a pharmacological space that may not be identical to the one considered for its construction. Further, the method is general as it can be applied in evaluation using different kinds

of information available, e.g. if the information on activity is both binary and continuous. The second contribution is toward a new characterization of the pharmacological topography in the form of activity canyons. It is associated with the probabilistic analysis of distinctiveness and helps to filter drug interactions at varying levels of distinctiveness and non-distinctiveness. It emphasizes the relevance of both measures for enhancing our perspective on distinctive interactions. It facilitates identification of distinctive interactions and the region of the chemical space to find them. The methodology aims to rigorously quantify the distinctiveness that can occur in diverse forms. This is helpful as the presence of these deviations is known to hinder the quantitative structure activity modeling. Accurate identification of distinctive interactions would decrease the risk of indiscriminate prescription and use of medicines by health care providers and consumers respectively.

Activity cliffs have been studied in detail in past research. These aberrations make it harder to coherently generalize the quantitative association between changes in physiological activity and structure, which is vital for medicinal chemists. Methods applied can influence the decision process of preparing fresh drugs and drug development through identification of distinctive pairs. The deviations in activity due to small chemical innovations have the potential of producing inventions in medicine. Thus, distinctiveness could be an externality for the drug industry.

The interpretation of the topography of pharmacological interaction space can vary according to the choice of measurement of c, s, d. I illustrate this with a probabilistic analysis of the pharmacological space which considers the probability of each of the measures s and d jointly with c. The consistency between observations and predictions corroborates that structural (c) and biological (s or d) properties can be considered as independently generated for studying the discontinuities in the landscape. A distinctive interaction can be measured either as high c, d or high c and low s. The almost contrasting construction of both measures results in a much higher range and distribution of magnitudes of d than of s. This affects the quantitative assessment of the interaction landscape, particularly for distinguishing amongst structure-activity functions displayed by drugs. As I show here, it is less significantly likely for a pair of drugs with high structural similarity to act similarly than oppositely on

a target. Further, the decrease in similar activity along the chemical space lacks the steepness of a cliff. In the given pharmacological space, high *d* dominates the chemical space in terms of not only magnitudes but also the probabilities. This behavior is mostly observed in the intermediate range of *c*. This is a transitional regime as it marks the increasing significance of *d*. The formulation shows the estimation of the prediction using *d* is more efficient.

As *c* increases, interactions corresponding to high *d* are relatively rare and more distinctive than those of low *s*. This facilitates preparation of analogs for intricately specified treatments. Low *s* interactions can occur between drugs affecting different classes. However, distinctive interactions of high *d* are rarer and occur between structural analogs aimed to treat same kinds of ailments with increasingly intricate specifications. Such analogs providing advanced or versatile treatments of same kinds of ailments vary highly with respect to their activity profiles. The differentiation between these kinds of drugs requires precision, and is highly crucial because a medicine may be rendered less effective in improving health if its treatment is not restricted to the particular treatment it is designed for. It is important to note that such associations are identified with high *d* instead of low *s*.

Based on the complex nature of structure-activity relationship study, this analysis aims to suggest an alternative characterization resulting from mathematical and statistical formulations. A crowded valley of magnitudes with intermittent spikes of variation in the chemical subspaces, may be characterized as activity canyons or gorges. This can be an encouraging sign for diversity in drug development. For instance, if a minor change in structure gives rise to an analog, then the two may be active on different kinds of targets more often than not. The sizes of their activity profiles may also be quite different. The methods provide a basis for practitioners and pharmacologists to identify distinctive interactions and the region of chemical space that they are probable to occur in.

Finally, the interpretation of distinctiveness and non-distinctiveness in the activity landscape is sensitive to perturbations and the ranges of measures used. At a given level of *c*, the distinctive interactions obtained with low *s* are different from those obtained with high *d*. This dissertation is an attempt to highlight the role of the techniques and criteria in evaluating distinctiveness in drug

interactions. They can render this concept of distinctiveness open to multiple interpretations and often adapting to the choice of measurement. I hope that this analysis would contribute to better health by meting out drugs to the desirable therapeutic targets. It is a crucial consideration for making quantitative assessments that would have the potential to affect contemporary and future research, drug discovery and development.

References:

1. Davis A. and Ward S. *The Handbook of Medicinal Chemistry : Principles and Practice*. The Royal Society of Chemistry, Cambridge, UK, 2014, 1-183.
2. King, F.D. *Medicinal Chemistry* : *Principles and Practice.* The Royal Society of Chemistry, Cambridge, UK, 2002.
3. Yildirim M.A., Goh K.L., Cusick M.E., Barabasi A.L., Vidal M. Drug Target Network. *Nat. Biotechnol.*, 2007, 25, 1119-1126.
4. Drews J. and Ryser S. Classic Drug targets. *Nat. Biotechnol.*, 1997, 15, 1297-1350.
5. Kubiniyi H. Similarity and Dissimilarity : A Medicinal Chemist's view. *Perspectives in Drug Discovery and Design* 1998, 9, 225-252.
6. Gillet, V.J., Wild, D.J., Willett, P., Bradshaw, J. Similarity and Dissimilarity Methods for Processing Chemical Structure Databases, *The Computer Journal*, 1998, 41, 547-558.
7. DiMasi, J.A., Hansen, R.W., Grabowski, H.G. Cost of innovation in the pharmaceutical industry, *J. of Health Economics*, 1991, 10, 107-142.
8. Danzon, P. Economics of the Pharmaceutical Industry, *National Bureau of Economic Research*, 2006.
9. Kulkarni, S. *Health for Peace*, Northern Book Center, Delhi, India, 1992.
10. Coleman, J.S., Social Capital in the creation of Human Capital, *Am. J. Sociol*., 1988, 94, S95-S120.
11. Arrow, K.J. Uncertainty and the Welfare Economics of Medical Care, *Am. Econ. Rev*. 1963, 5, 941-973.
12. Kulkarni, V.S. Temporal Evolution of Social Innovation : What Matters?, *SIAM J. Appl. Dyn. Syst.*, 2016, 15, 1485-1500.

13. Martin, Y.C., Kofron, J.L., Traphagen, L.M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.*, 2002, 45, 4350-4358.
14. Wassermann, A., Wawer, M., Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis, *J. Med. Chem. Perspective*, 2010, 53, 8209-8223.
15. Bickerton, G.R., Paolini, G.V., Besnard, J. Mureasan, S., Hopkins, A.L. Quantifying the chemical beauty of drugs, *Nature Chemistry*, 2012, 4, 90-98.
16. Wood, D.J., Carlson, L., Eklund, M., Norinder, U., Stalring, J. QSAR with experimental and predictive distributions: an information theoretic approach for assessing model quality, *J. Comput. Aided Mol. Des.*, 2013, 27, 203-219.
17. Stumpfe, D. and Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.*, 2012, 55, 2932-2942.
18. Cruz-Monteagudo, M., Medina-Franco, J.L., Perez-Castillo, Y., Nicolotti, O., Cordeiro, M.N., Borges, F. Activity Cliffs in Drug Discovery : Dr. Jekyll or Mr. Hyde. *Drug Discov. Today*, 2014, 19, 1069-1080.
19. Mestres, J., Gregori-Puegjane, E., Valverde, S., Sole, R.V. Data Completeness: Achilles heel of drug-target networks. *Nat. Biotechnol,* 2008, 26, 983-984.
20. Hu, T.M. and Hayton, W.L. Architecture of the drug-drug interaction network, *J. Clin. Pharm. Ther.*, 2011, 36, 135-143.
21. Kulkarni, V.S. and Wild, D.J. An Activity Canyon Characterization of the Pharmacological Topography, *J. of Cheminformatics*, 2016, 8:41, 1-12.
22. Newman, M.E.J., Strogatz, S.H., Watts, D.J. Random graphs with arbitrary degree distributions and their applications, *Phys. Rev. E.*, 2001, 64, 026118-1-17.
23. Albert, R., Barabasi, A-L., Statistical mechanics of complex networks, *Rev. Mod. Phys.*, 74, 47-97, 2002.
24. Jaynes, E.T. Information Theory and Statistical Mechanics, *Phys. Rev.*, 1957, 106, 620-630.
25. Freund, J.E. *Mathematical Statistics*, Prentice Hall, London, 1987.

26. Wilson, R.C., Zhu, P. A study of graph spectra for comparing graphs and trees, *Pattern Recognition*, 2008, 41, 2833-2841.
27. Chung, F. Lu, L, Vu, V. Spectra of random graphs with given expected degrees, *Proc. Natl. Acad. Sci.,* 2003, 11, 6313-6318.
28. Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D., Wienberger, L.E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors, *J. Med. Chem.*, 1996, 39, 3049-3059.
29. Sherrington, C.S. *Man on his Nature*, Macmillan, Cambridge, UK, 1940, 172-201, 277.
30. Inamdar, R.S. Auto-regulation of Physiological Processes in Plants in the Wake of Cosmic Phenomena, *Special Lectures*, Invited, India, 1956.

    Inamdar, R.S. The Auto-regulation of Physiological Processes in Plants, *Presidential address*, Indian Science Congress, 1926.
31. McNemar, Q. Note on sampling error of the difference between correlated proportions or percentages, *Psychometrika*, 1947, 12, 153-157.
32. Theil, H. *Economics and Information Theory*, North Holland, Amsterdam, 1967.
33. Kullback, S. *Information Theory and Statistics*, Wiley, London, 1959, 1-31.

Appendix

Mathematical Proofs

1.1 Derivation of z_1 and z_2 in Eqs. (2.3), (2.4).

Sketch of the derivation: The two quantities z_1, z_2 can be derived using the procedure [22, 23] of random graphs, for the bipartite network with both kinds of vertices possessing power law distribution. The generating functions of k_p, k_d are

$f_0(x) = \frac{Li_{\gamma_1}\left(xe^{-\frac{1}{\kappa_1}}\right)}{Li_{\gamma_1}\left(e^{-\frac{1}{\kappa_1}}\right)}$ and $g_0(x) = \frac{Li_{\gamma_2}\left(xe^{-\frac{1}{\kappa_2}}\right)}{Li_{\gamma_2}\left(e^{-\frac{1}{\kappa_2}}\right)}$ and

$f_1(x) = \frac{Li_{\gamma_1-1}\left(xe^{-\frac{1}{\kappa_1}}\right)}{xLi_{\gamma_1-1}\left(e^{-\frac{1}{\kappa_1}}\right)}$ and $g_1(x) = \frac{Li_{\gamma_2-1}\left(xe^{-\frac{1}{\kappa_2}}\right)}{xLi_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)}$ respectively.

$z_1 = f_0'(1)g_1'(1)$ and $z_2 = f_0'(1)f_1'(1)[g_1'(1)]^2$

Substituting the above expressions in those of z_1 and z_2, and considering x=1, we get the required expressions for z_1, z_2 in Eqs. (2.3), (2.4).

1.2 Derivation of p_k in Eq. (2.5).

Sketch of the derivation:

I derive the probability distribution for the network [22] with the generating functions,

$$G_0(x) = f_0\big(g_1(x)\big)$$

which implies $$G_0(x) = \frac{Li_{\gamma_1}}{Li_{\gamma_1}\left(e^{-\frac{1}{\kappa_1}}\right)}\left(\frac{Li_{\gamma_2-1}\left(xe^{-\frac{1}{\kappa_2}}\right)}{xLi_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)}\right)$$

Since γ_1>1, $Li_{\gamma_1}\left(e^{-\frac{1}{\kappa_1}}\right) \approx \varsigma(\gamma_1)$, ς is called Riemann zeta function, but this is highly sensitive to small shifts or variations in $e^{-\frac{1}{\kappa_1}}$ [which is not =1 as shown in Fig. 2.1], and similarly for γ_2.

Probability of k links can be found by the derivatives $p_k = \frac{1}{k!}\frac{d^k G_0}{dx^k}$ at x=0.

Let $\frac{Li_{\gamma_2-1}\left(xe^{-\frac{1}{\kappa_2}}\right)}{xLi_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)} = \vartheta$, $p_1 = \frac{dG_0}{dx} = \frac{1}{Li_{\gamma_1}\left(e^{-\frac{1}{\kappa_1}}\right)}\frac{Li_{\gamma_1-1}(\vartheta)}{\vartheta}d\vartheta/dx$ at $x = 0$.

$\frac{d^2G_0}{dx^2} = \frac{1}{Li_{\gamma_1}\left(e^{-\frac{1}{\kappa_1}}\right)}\left(\frac{Li_{\gamma_1-1}(\vartheta)}{\vartheta}\frac{d^2\vartheta}{dx^2} + \frac{Li_{\gamma_1-2}(\vartheta)}{\vartheta^2}(d\vartheta/dx)^2 - \frac{Li_{\gamma_1-1}(\vartheta)}{\vartheta^2}(d\vartheta/dx)^2\right)$ at x=0.

$\frac{d^2G_0}{dx^2} = \frac{1}{Li_{\gamma_1}\left(e^{-\frac{1}{\kappa_1}}\right)}\left(Li_{\gamma_1-1}\left(\frac{1}{Li_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)}\right)Li_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)\frac{2}{3^{\gamma_2-1}} + \frac{Li_{\gamma_1-2}(e^{-\frac{1}{\kappa_1}}) - Li_{\gamma_1-1}(e^{-\frac{1}{\kappa_1}})}{(2^{\gamma_2-1})^2}\right)$ for x=0

Now, the computation of $e^{-\frac{1}{\kappa_1}}$ is subject to errors, and I consider the least estimate in the interval around the standard error, so that $\frac{Li_{\gamma_1-2}\left(e^{-\frac{1}{\kappa_1}}\right) - Li_{\gamma_1-1}\left(e^{-\frac{1}{\kappa_1}}\right)}{(2^{\gamma_2-1})^2} \ll 1$. This is also feasible owing to the susceptibility of the function to small changes in inputs. Hence, I retain only the first terms involving Li_{γ_1-1} in all the derivatives and ignore all higher order terms. This is done for convenience of solving.

$$\frac{d^3G_0}{dx^3} = \frac{1}{Li_{\gamma_1}\left(e^{-\frac{1}{\kappa_1}}\right)}\left(\frac{Li_{\gamma_1-1}(\vartheta)}{\vartheta}\frac{d^3\vartheta}{dx^3} + \frac{Li_{\gamma_1-3}(\vartheta)}{\vartheta^3}(d\vartheta/dx)^2 + \frac{Li_{\gamma_1-2}(\vartheta)}{\vartheta^2}\frac{d}{dx}(d\vartheta/dx)^2 + \frac{Li_{\gamma_1-2}(\vartheta)}{\vartheta^2}\frac{d\vartheta}{dx}\frac{d^2\vartheta}{dx^2} + h.o.\right)$$

Again, the higher order terms h.o. are ignored and only first term involving $Li_{\gamma_1-1}$is retained and then approximated. The reason is the same, because $\gamma_1 \approx \gamma_2 \approx 2$, $Li_{\gamma_1-3}\left(\frac{1}{Li_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)}\right) \ll 1$ and similarly for other higher order terms.

This inspection of terms for *k*=1,2,3 can be used for *k*>3, and thus retaining only the first terms, the approximate connectivity distribution is as given in Eq.(2.5),

$$p_k = \frac{1}{k!}Li_{\gamma_1-1}\left(\frac{1}{Li_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)}\right)Li_{\gamma_2-1}\left(e^{-\frac{1}{\kappa_2}}\right)k!\,(k+1)^{-\gamma_2+1}.$$

2. Link retaining procedure

Theorem: A weighted adjacency matrix *W* is perturbed by retaining only the links with chemical similarity $\mathrm{c} \geq \mathrm{const}$, then the maximum eigenvalue λ_{max} of the perturbed matrix W' satisfies $\lambda_{max}(W') \leq p\lambda_{max}(W)$, where $p = p(\mathrm{c} \geq \mathrm{const})$ is the probability of finding links at the level of chemical similarity *const* or greater.

Proof: It is known that for any real symmetric matrix such as *W* having eigenvalues λ and eigenvectors *x*, the quantity

$$\lambda = \frac{x^T W x}{x^T x}$$

is maximized for the maximum eigenvalue $\lambda_{max}(W)$ and the corresponding eigenvector, that is when $\lambda = \lambda_{max} = \frac{x^T W x}{x^T x}$.

Imposing the condition of retaining only those links in *W* that possess chemical similarity $\mathrm{c} \geq \mathrm{const}$, the number of links is reduced and so is the average connectivity <*k*> of *W*. The average connectivity

is calculated for the unweighted adjacency matrix derived from *W* in which the non-zero elements of *W* are considered as 1 (or connected). Upon doing so, the new perturbed matrix W' has average connectivity $< k >'$ and the maximum eigenvalue $\lambda_{max}(W')$. From the distribution of chemical similarity of links, we can find the required probability, *p* using information from the chemical similarity adjacency matrix *C*. Assuming that the links of every node in the network (of size *N*) are formed independently, we can say that all nodes have approximately (1-*p*)*N* nodes with whom they do not have connections and approximately *pN* nodes with whom they retain the connections after perturbation. A node in *W* having *k* links is reduced to having $k' = pk$ links in W' after perturbation, so the average connectivity of W' is $< k >' \approx p < k >$.

Define an indicator function $\mathrm{I}_{ij} = 1\ if\ \mathrm{C}_{ij} \geq const$; $\mathrm{I}_{ij} = 0\ if\ \mathrm{C}_{ij} < const$. This specifies the un-weighted adjacency matrix for the constraint applied on retaining of links before after perturbation as I and I′ respectively. Thus $W' = W \circ \mathrm{I}'$ is obtained by taking the entrywise product (or Hadamard product) of the two matrices. Considering λ_{max} as the spectral radius of the matrices in this case, it is established that $\lambda_{max}(W \circ \mathrm{I}') \leq \lambda_{max}(W)\, \lambda_{max}(\mathrm{I}')$. Considering the maximum eigenvalue as roughly equal to the average connectivity of the network (for any graph, $< k > \leq \lambda_{max} \leq$maximum connectivity),

$\lambda_{max}(\mathrm{I}') \approx p\, \lambda_{max}(\mathrm{I})$

which proves that

$$\lambda_{max}(W') \leq p\lambda_{max}(W).$$

3.1 Joint probability computation

If structure and activity are independently generated, then predicted joint probabilities are given by

$$P\big((c_1 \leq c \leq c_2) \cap (s_1 \leq s \leq s_2)\big) = c_{min}^{\gamma-1} s_{min}^{\alpha-1} \big(s_1^{1-\alpha} - s_2^{1-\alpha}\big)\big(c_1^{1-\gamma} - c_2^{1-\gamma}\big)$$

$$P\big((c_1 \leq c \leq c_2) \cap (d_1 \leq d \leq d_2)\big) = c_{min}^{\gamma-1} \frac{0.042}{\lambda} \big(e^{-\lambda d_1} - e^{-\lambda d_2}\big)\big(c_1^{1-\gamma} - c_2^{1-\gamma}\big)$$

Proof: Since the structure and activity are independently generated, using probability distributions of $P(c)$ and $P(s)$ from above, the joint probabilities approximate to

$$P\big((c_1 \le c \le c_2) \cap (s_1 \le s \le s_2)\big) = P(c_1 \le c \le c_2)P(s_1 \le s \le s_2)$$

$$= \int_{s_1}^{s_2} \frac{0.045}{s_{min}} \left(\frac{s'}{s_{min}}\right)^{-\alpha} ds' \int_{c_1}^{c_2} \frac{\gamma-1}{c_{min}} \left(\frac{c'}{c_{min}}\right)^{-\gamma} dc' \qquad (1).$$

Similarly, using probability distributions of $P(c)$ and $P(d)$, we obtain

$$P\big((c_1 \le c \le c_2) \cap (d_1 \le d \le d_2)\big) = P(c_1 \le c \le c_2)P(d_1 \le d \le d_2)$$

$$= \int_{d_1}^{d_2} 0.042\, e^{\lambda d_{min}} e^{-\lambda d'} dd' \int_{c_1}^{c_2} \frac{\gamma-1}{c_{min}} \left(\frac{c'}{c_{min}}\right)^{-\gamma} dc' \qquad (2).$$

Solving equations (1) and (2) we obtain

$$P\big((c_1 \le c \le c_2) \cap (s_1 \le s \le s_2)\big) = c_{min}^{\gamma-1} s_{min}^{\alpha-1} (s_1^{1-\alpha} - s_2^{1-\alpha})\big(c_1^{1-\gamma} - c_2^{1-\gamma}\big) \qquad (3).$$

$$P\big((c_1 \le c \le c_2) \cap (d_1 \le d \le d_2)\big) = c_{min}^{\gamma-1} \frac{0.042}{\lambda} \big(e^{-\lambda d_1} - e^{-\lambda d_2}\big)\big(c_1^{1-\gamma} - c_2^{1-\gamma}\big) \qquad (4).$$

Eqs. (3) and (4) give the predicted probabilities of simultaneously finding structural and biological similarity, and structural similarity and distance in a range, respectively.

3.2 Predicted probability and range

The lower bound on the predicted probability calculated above using two variables (s and c) or (s and d) works differently for both measures s and d.

Using Chebyshev inequality, $P\big((l_{11} \le A_1 \le l_{12}) \cap (l_{21} \le A_2 \le l_{22})\big) \ge 1 - \sum_{i=1,2} T_i$,

$T_i = \frac{4\sigma_i^2 + (2\mu_i - (l_{i1} + l_{i2}))^2}{(l_{i2} - l_{i1})^2}$, μ_i, σ_i are the mean and variance of variable A_i respectively.

Taking $A_1 = c$, $A_2 = s$, we find the probability in any range of these variables is uncertain because s has a power law distribution with very high variance, and by the inequality it would imply that the probability of finding c and s in any range can be any value between 0 and 1. However, when $A_1 = c$ and $A_2 = d$, the variances of c and d are finite and we can adjust the ranges of the variables to control for the uncertainty of the estimate.

However, this technique is of little use if the domain is restricted to high or extreme values of *c*, *s*, *d* for distinctiveness. The explicit calculation in 3.1 helps in quantifying the extremes of both variables jointly.

3.3. The effect of choice of technique for discerning distinctiveness and non-distinctiveness

In the predicted probability model introduced above, the levels of distinctiveness or non-distinctiveness are specified by the levels or ranges of *c*, *s*, *d*, assuming independence of structure and activity. The two estimates of predicted probability of distinctiveness are $\pi_s = P\big((c \geq c_0) \cap (s \geq \mu_s + n\sigma_s)\big) = p_{ch}P(s \geq \mu_s + n\sigma_s)$, taking $p_{ch} = P(c \geq c_0)$.

And $\pi_d = P\big((c \geq c_0) \cap (d \geq \mu_d + n\sigma_d)\big) = p_{ch}P(d \geq \mu_d + n\sigma_d)$,

According to Chebyshev inequality, $P(|s - \mu_s| \geq n\sigma_s) \leq 1/n^2$ and $P(|d - \mu_d| \geq n\sigma_d) \leq 1/n^2$. This gives the bounds on finding the probability of distinctiveness and non-distinctiveness at the extremes using both measures *s* and *d* in the same chemical space. Now assuming a constant $\upsilon = n\sigma_i$, with *i* as a symbol for *s* and *d*, the probability of finding large values are bounded above as $P(|s - \mu_s| \geq \upsilon) \leq \sigma_s^2/\upsilon^2$ for *s,* and, $P(|d - \mu_d| \geq \upsilon) \leq \sigma_d^2/\upsilon^2$ for *d*. As the variance of *s* if very high and would become infinite as the system size becomes infinitely large, the prediction in any interval around the mean or away from it or in the tails of the distribution, is highly arbitrary for all values of υ, p_{ch}.

In contrast, σ_d^2 is measurable and finite. Thus it is possible to find υ so that $|\upsilon| \geq |\sigma_d|$ ensures a finite predicted probability (for *n*>1) of finding highly distinctive or non-distinctive interactions. Now to consider both distinctiveness and non-distinctiveness that can be discerned by using *d,* the predicted probability of finding either kind of interaction between drugs can be written as $\pi_d = p_{ch}P(|d - \mu_d| \geq \upsilon) \leq p_{ch}\sigma_d^2/\upsilon^2$.

This facilitates the specification of an interval around the mean: $|\upsilon| \geq \sqrt{p_{ch}}|\sigma_d|$. The equality corresponds to a greater maximum of the probability that can be estimated. The values on the right outside of this interval would correspond to distinctive interactions and those to the left would

correspond to the non-distinctive interactions. The probability of finding them is lower, meaning they are rarer. Moreover, higher chemical similarity values are less probable and the interval around the mean that distinguishes highly distinctive from non-distinctive diminishes. When c_0 is relatively lower, it is the other way around. This indicates a tradeoff between chemical similarity and σ_d for discernibility of distinctiveness. Interestingly, high distinctiveness (and high non-distinctiveness) will be more discernible and rarer at very high levels of both attributes.

4.

Theorem: If the weighted adjacency matrix W, with eigenvalues Λ is iteratively perturbed as $W'_{ij} = W_{ij} - pW_{ij} \ \forall i,j$ for $0< p < 1$, then at the end of n iterations, the change in eigenvalues $\|\Delta\Lambda\| \propto |\Delta_n|$ where $\Delta_n = 1 - (1-p)^n$.

Proof sketch:

The perturbations grow successively as follows.

For the first step $m = 1,\ W'_{ij} = W_{ij}(1-p)$,

for $m = 2,\ W''_{ij} = W'_{ij} - pW'_{ij} = W_{ij}(1-p)^2 = W_{ij} - 2pW_{ij} + p^2W_{ij}$,

for $m = 3,\ W'''_{ij} = W''_{ij} - pW''_{ij} = W_{ij}(1-p)^3 = W_{ij} - 3pW_{ij} + 3p^2W_{ij} - p^3W_{ij}$,

.

.

.

and therefore inductively generalizing for $m = n, W_{ij} \rightarrow W_{ij}\{\Delta_n\}$ where $\Delta_n = 1 - (1-p)^n$.

It is well established that the perturbation of W to $W + \Delta\mathrm{W}$ results in change in eigenvalues as $\|\Delta\Lambda\| = \|X^{-1}\|\|X\|\|\Delta\mathrm{W}\|$ (where X is the set of eigenvectors of the matrix).

After m=n iterations, as shown above, $\|\Delta\mathrm{W}\| = |\Delta_n|\|\mathrm{W}\|$ and therefore $\|\Delta\Lambda\| \propto |\Delta_n|$.

Corollary 1: (Optimum perturbation).

For the weighted adjacency matrix W to be perturbed at least once by proportion p, the corresponding change would be

$$\Delta_n \leq p$$

Proof: It is known from theorem above that when the elements of W are perturbed by a fraction p, such that $0<p<1$ then the change in the maximum eigenvalue (after n perturbations) is proportional to $\Delta_n = 1-(1-p)^n$. Since this is a monotonic function in p and n, we can approximate the change to be as low (or as high) as ε so that $\varepsilon > 0$. This means that we require

$$1-(1-p)^n = \varepsilon$$

which implies that $n = \frac{\log(1-\varepsilon)}{\log(1-p)}$. For $n \geq 1$, solving the inequality

$$\frac{\log(1-\varepsilon)}{\log(1-p)} \geq 1$$

we get $\varepsilon \leq p$. This means if the perturbation is applied $n \geq 1$ times, then $\Delta_n \leq p$.

Corollary 2: When the link weights of individual vertices in W are perturbed, the perturbation of the link weights of the vertex i maximizes the decrease $\Delta\lambda_{max}(W)$ if there is an indicator function $I(i)$ defined as $I_{i'j} = 1 \ \ \forall j$ if $i' = i$ and $I_{i'j} = 0 \ \ \forall j$ if $i' \neq i$, so that $\lambda_{max}(I(i)) \geq \lambda_{max}(I(i')) \ \ \forall i' = 1,2 \ldots N$.

Proof:

The weighted adjacency matrix W is perturbed using an indicator function $I = I(i)$ which specifies the unweighted adjacency matrix such that the elements of i^{th} row and column corresponding to i^{th} vertex are equal to 1 and the other elements in the matrix are 0. It is symmetric. Therefore, the perturbation is $W_{ij} \rightarrow W_{ij} - p(W^{\circ}I)_{ij}$ and its magnitude is $\|\Delta W\| = |p|\|W^{\circ}I\| \leq$

$|p|\sqrt{\|I^T I\|\|W^T W\|}$. Since the norm of the indicator function is the maximum eigenvalue, $\|I^T I\| = \lambda_{max}(I)$. For any vertex $i' \neq i$, the indicator function $I' = I(i')$ satisfies $\|I'^T I'\| = \lambda_{max}(I')$. Now, the magnitude of perturbation to W is greatest due to a perturbation of link weights of vertex i when $\lambda_{max}(I) \geq \lambda_{max}(I')$. Thus from the theorem the drop in maximum eigenvalue $\Delta\lambda_{max}$ is greatest when the links of vertex *i* are perturbed for same level of perturbation $|\Delta_n|$.

Curriculum Vitae

Varsha S. Kulkarni

Education

Ph.D. December 2016, Indiana University, Bloomington, USA

Minor in Statistics, Indiana University, Bloomington, USA

Graduate level Coursework, Department of Statistics, University of Wisconsin-Madison, USA

M.Sc. Physics, Department of Physics & Astrophysics, University of Delhi, India

B.Sc. (Honors) Physics, Miranda House, University of Delhi, India

Awards, Honors, Distinctions

2015 Travel award, Kellogg School of Management, Northwestern University, USA

2014-15 NSF IGERT fellowship, Indiana University, Bloomington

2012 Fellow, Graduate Workshop in Computational Social Science, Santa Fe Institute, USA

2011-12 Travel award, Indiana University, Bloomington

2011 Grant for research in summer, Australian National University, Australia

2009 Fellow, Complex Systems Summer School, Santa Fe Institute, USA

2008-9 Research-project fellowship, Indian Institute of Technology, Kanpur, India

2003-4 Paper competition prize, Department of Physics, St. Stephen's College, University of Delhi, India: research on a dynamic model in social networks.

Merit Scholarship for M.Sc., University of Delhi, Delhi, India.

Merit Scholarship for M.Sc., Government of Karnataka, India.

2002 Ranked third, B.Sc. (Honors), I-III, Miranda House, University of Delhi, India.

Ranked top ten in B.Sc. (Honors), I-III, University of Delhi, India

Publications

Kulkarni, Varsha S. 2016. Temporal Evolution of Social Innovation : What Matters?, *SIAM J. on Appl. Dyn. Syst*., 15 (3), 1485-1500.

Kulkarni, Varsha S. and D. Wild. 2016. An Activity Canyon Characterization of the Pharmacological Topography, *J. of Cheminformatics*, 8:41.

Kulkarni, Varsha S. 2014. 'Complexity, Chaos, and the Duffing-Oscillator Model: An Analysis of Inventory Fluctuations in Markets', *J. of Applied Nonlinear Dynamics*, 3 (2), 147-158.

R. Jha and **Varsha S. Kulkarni**. 2015. 'Inflation, its Volatility, and the Inflation-Growth Tradeoff in India', *International J. of Emerging Markets*, 10 (3), 350-361.

R. Gaiha, K. Hill, G. Thapa, **Varsha S. Kulkarni**. 2014. 'Have Natural Disasters become Deadlier?', in *Sustainable Economic Development: Resources, Environment, and Institutions*, A. Balisacan, U. Chakravorty, M.L. Ravago eds., Academic Press Oxford.

R. Gaiha, K. Hill, G. Thapa, **Varsha S. Kulkarni**. 2013. 'Have Natural Disasters become Deadlier?', *Brooks World Poverty Institute publications.*

G. Thapa, R. Gaiha, K. Imai and **Varsha S. Kulkarni**. 2009. 'Soaring food prices: A threat or opportunity in Asia?', *Brooks World Poverty Institute publications.*

Kulkarni, Varsha and N. Deo. 2007. 'Correlation & Volatility of an Indian stock market: A Random Matrix Approach', *The European Physics Journal B,* 60 (1), 101-109.

Kulkarni, Varsha and N. Deo. 2006. 'A Random Matrix Approach to Volatility of an Indian Financial Market', in *Econophysics of Stock and Other Markets*, eds. B. Chakrabarti, A. Chatterjee, Springer-Verlag Milan.

Working paper

Kulkarni, Varsha S. 'Spreading of Infection through social networks'.
Previous version- Complex Systems Summer School, Santa Fe Institute, 2009.

Opinion Editions in Newspapers

'Is it demand or costs that are driving food inflation?' with R. Gaiha, *Business Standard*, 2008.

'Pay more to pay less' with R. Gaiha, *Indian Express*, 2008.

'Costlier food, reformed farms' with R. Gaiha, *Indian Express*, 2008.

'Need to shift focus from GDP per capita to other indicators, including poverty and inequality' with R. Gaiha, A. Vikram, *Economic Times*, 2012.

Presentations

2015 Computational Social Science Summit, Kellogg School of Management, Northwestern University, USA

2011 Conference on Dynamical Systems, Society for Industrial and Applied Mathematics, Utah, USA

2010 NetSci conference, Massachusetts Institute of Technology, Cambridge, MA, USA.

2009 Conference on Disorder, Complexity & Biology II, Banaras Hindu University, Varanasi, India

2006 Econophysics of Stock Markets and Minority Games, Saha Institute of Nuclear Physics, Kolkata, India

2006 School on Complex systems, Institute of Mathematical Sciences, Chennai, India.

Work

2010-2016 Associate Instructor, School of Informatics and Computing, Indiana University, Bloomington, USA.

2009-2010 Research Assistant, Center for Complex Networks and Systems Research, Indiana University, Bloomington, USA

2006-2007 Graduate Student, Department of Statistics, University of Wisconsin-Madison, USA.

Teaching Assistant, Department of Statistics, University of Wisconsin-Madison, U.S.A

2004-05 Research scholar, University of Delhi for National Science Foundation program on Human and Social Dynamics, University of Chicago and Santa Fe Institute, U.S.A.

2004-06 Research Scholar, Department of Physics and Astrophysics, University of Delhi, India.

Teaching

School of Informatics and Computing, Indiana University, associate instructor

Info 590- *Data Science for Drug Discovery*

Info 201- *Mathematical Foundations of Informatics*

H201- *Mathematical Foundations of Informatics, Honors*

Info 400- *Informatics in Disasters and Emergency Response*

Info 400- *Introduction to Networks* (Analysis of complex networks: theory and applications)

Department of Statistics, University of Wisconsin-Madison, teaching assistant

Stat 301 *Introduction to Statistics*

Stat 441 *Introduction to Biostatistics*

Professional service

Reviewer for scientific research journals and programs.