Chapter 14
Visual Analytics: Leveraging Cognitive Principles to Accelerate Biomedical Discoveries

Suresh K. Bhavnani

14.1 Introduction

The *Open Science* movement (e.g., data from NIH-funded studies being made publicly available), combined with digital access to patient clinical records, in addition to rapid advances in the development of inexpensive high throughput technologies (e.g., multiplex assays for measuring whole genome data across many patients) has resulted in vast digital resources accessible by both scientists and the lay public (Molloy 2011). However, the sheer magnitude of such resources far exceeds our cognitive abilities to exploit them for the prevention, diagnosis, and treatment of diseases. For example, translational teams consisting of biologists, clinicians, and epidemiologists increasingly need to integrate and comprehend the relationships among large and disparate types of information including molecular, biochemical, and environmental variables, with the goal of comprehending complex phenomena such as heterogeneities and corresponding pathways underlying different diseases.

Portions of this chapter in sections "VISUAL ANALYTICS: THEORETICAL FOUNDATIONS", and "STRENGTHS AND LIMITATIONS OF NETWORK ANALYSIS" appeared in Bhavnani, Drake, and Divekar, 2014. With kind permission from Springer Science + Business Media: Bhavnani et al. (2014b). Portions of this chapter in the section "NETWORK ANALYSIS: MAKING DISCOVERIES IN COMPLEX BIOMEDICAL DATA", appeared in an article in the *Proceedings of the AMIA Summit on Translational Bioinformatics* (2014), Bhavnani, S.K., Dang, B, Caro, M., Bellala, G., Visweswaran, S., Heterogeneity within and across Pediatric Pulmonary Infections: From Bipartite Networks to At-Risk Subphenotypes".

S.K. Bhavnani, Ph.D. (✉)
Institute for Translational Sciences, University of Texas Medical Branch, 301 University Blvd,
Galveston, TX 77555, USA
e-mail: skbhavnani@gmail.com

V.L. Patel et al. (eds.), *Cognitive Informatics for Biomedicine*, Health Informatics,
DOI 10.1007/978-3-319-17272-9_14

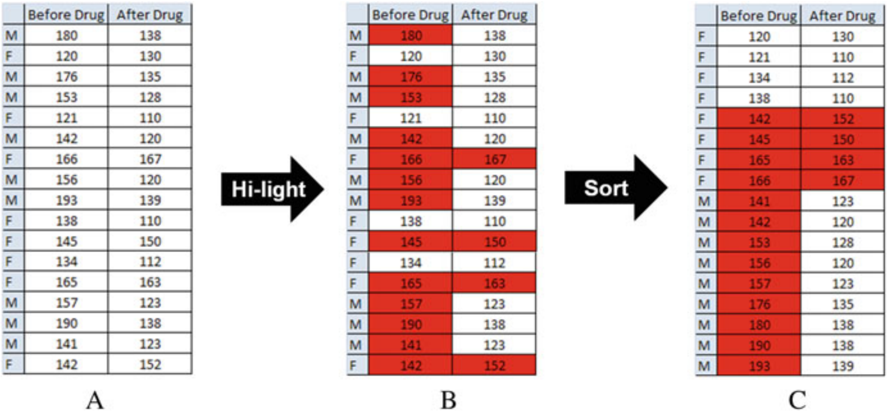
One approach to integrate and comprehend such vast and disparate information is through methods being developed in the new field of visual analytics. This chapter begins by presenting an overview of the evolving theoretical foundations for visual analytics, and the cognitive and task-based motivations to use methods from this field to help comprehend complex biomedical data. Next, the chapter provides a brief overview of visual analytical applications in the biomedical domain, with a demonstration of how to use one of the most advanced forms of visual analytics called networks, which are particularly useful for analyzing complex molecular and clinical data. These analyses reveal the strengths and limitations of network analysis, which are critical for its practical use to analyze ever increasing and complex biomedical data. The chapter concludes with theoretical, applied, and pedagogical hurdles that need to be addressed through future, research which will enable visual analytics to fully realize its potential in accelerating biomedical discoveries.

14.2 Visual Analytics: Theoretical Foundations

Visual analytics is defined as the science of analytical reasoning, facilitated by interactive visual interfaces (Thomas and Cook 2005). The primary goal of visual analytics is to augment cognitive reasoning by translating symbolic data (e.g., numbers in a spreadsheet) into *visualizations* (e.g., a scatter plot) which can be manipulated through *interaction* (e.g., highlight only some data points in the scatter plot). As discussed below, visualizations, and interaction with those visualizations, are powerful for helping analysts comprehend complex relationships in biomedical data because of the nature of human cognition, and the nature of tasks performed by analysts.

14.2.1 Why Do Visualizations Matter?

Visualizations of data are often powerful because they leverage the massively parallel architecture of the human visual system consisting of the eye and the visual cortex of the brain (Card et al. 1999). This parallel cognitive architecture enables the rapid comprehension of multiple graphical relationships simultaneously, which often leads to insights about relationships in complex data such as similarities, trends, and anomalies (Thomas and Cook 2005). For example, Fig. 14.1a shows a spreadsheet representing the systolic blood pressure of patients before and after taking a drug. The task of determining which of the two conditions have more patients with systolic >140 is time consuming and error prone because the analyst has to compare the number in each cell with 140, remember the result of each comparison, and then make a final count to determine which column has a higher number of patients with systolic >140. Such symbolic processing is serial in nature,


Hi-light
Sort
A
B
C

	Before Drug	After Drug
M	180	138
F	120	130
M	176	135
M	153	128
F	121	110
M	142	120
F	166	167
M	156	120
M	193	139
F	138	110
F	145	150
F	134	112
F	165	163
M	157	123
M	190	138
M	141	123
F	142	152

	Before Drug	After Drug
M	180	138
F	120	130
M	176	135
M	153	128
F	121	110
M	142	120
F	166	167
M	156	120
M	193	139
F	138	110
F	145	150
F	134	112
F	165	163
M	157	123
M	190	138
M	141	123
F	142	152

	Before Drug	After Drug
F	120	130
F	121	110
F	134	112
F	138	110
F	142	152
F	145	150
F	165	163
F	166	167
M	141	123
M	142	120
M	153	128
M	156	120
M	157	123
M	176	135
M	180	138
M	190	138
M	193	139

Fig. 14.1 An example of how symbolic data in a spreadsheet (**a**) when converted into a visual representation (**b**) leverages the parallel processing abilities of the visual cortex which enables faster comprehension of patterns in the data. Because visual processing is parallel in nature, it scales to handle large amounts of data. When the same data is sorted by gender (**c**), the visual representation reveals yet another pattern demonstrating how interaction with the data is a critical aspect of visual analytics, and can guide the verification of the patterns using the appropriate quantitative measures

and therefore highly dependent on the number of data points, which when large can quickly overwhelm an analyst.

In contrast, as shown in Fig. 14.1b, if all cells in the spreadsheet with values >140 are colored red, the resulting visual representation enables processing of red cells in each column to be conducted in parallel, resulting in a more rapid determination that the left column has more red cells compared to the right column. Such parallel processing is independent of the number of cells, and therefore scales up well to large amounts of data. Data visualizations therefore help to shift processing from the slower symbolic processing areas of the human brain, to the faster graphical parallel processing of the visual cortex enabling detection of patterns in large and complex biomedical data sets. Furthermore, by externalizing key aspects of the task, the representation in Fig. 14.1b shifts information from an internal to an external representation, making other tasks such as counting the number of patients with systolic >140 in each column much easier (Zhang and Norman 1994).

Unfortunately, not all data visualizations are effective in augmenting cognition. For example, a road map pointing south is not effective for a driver who is facing north because it requires a mental rotation of the map before it can be useful for navigation. Similarly, an organizational chart of employee names and their locations laid out in a hierarchy based on seniority is not very useful if the task is to determine patterns related to the geographical distribution of the employees. Finally, if a chart has an incorrect or missing legend and axes labels, the visualization is difficult to comprehend because it cannot be mapped to concepts in the data. Therefore visualizations need to be aligned with mental representations of the user (Tversky et al. 2002), tasks (Norman 1993), and data, before those visualizations can be effective in augmenting cognition.

14.2.2 Why Does Interactivity Matter?

While static visualizations of data can be powerful if they are aligned with mental representations, tasks, and data, they are often insufficient for comprehending complex data. This is because data analysis typically requires many different tasks performed on the same data such as discovery, inspection, confirmation, and explanation (Bhavnani et al. 2012), each requiring different transformations of the data. For example, if the task in Fig. 14.1b is to understand the relationship of the drug to gender, then the data can be sorted based on gender. As shown, interaction with the data through such sorting reveals that the drug has no effect on females (low values remain low, and high values remain high), whereas it has a dramatic effect on lowering systolic values in males (all high values become low). Therefore, while it is well accepted that interactivity is crucial for the use of most computer systems, interaction with data visualizations can help to reveal relationships that are otherwise hidden when using a single representation of the data.

Interactivity is also critical when analysis is done in teams consisting of different disciplines, where each member often requires a different representation of the same data. For example, a molecular biologist might be interested in which genes are co-expressed across patients, whereas a clinician might be interested in the clinical characteristics of patients with similar gene profiles, and later how they integrate with the molecular information. To address these changes in task and mental representation, visualizations require interactivity or the ability to transform parts, or the entire visual representation.

14.2.3 Theories Related to Visual Analytics

Although the field of visual analytics has drawn on theories and heuristics from different disciplines such as cognitive psychology, computer science, and graphic design, the development of theories and taxonomies for visual analytics are still in early stages of development (Thomas and Cook 2005). For example, there are a number of attempts to classify visual analytical representations (Heer et al. 2010; Shneiderman 1996), and interaction intents at different levels of granularities (Yi et al. 2007; Amar et al. 2005).

One attempt to classify visual analytical representations groups them into (1) time series (e.g., line graphs showing how the expression of different genes change over time), (2) statistical distributions (e.g., box-and-whisker plots), (3) maps (e.g., pie charts showing percentages of different races at different city locations on the US map), (4) hierarchies (e.g., top-down tree showing the management structure of an organization), and networks (e.g., a social network of how friends connect to other friends such as on Facebook). Once these visualizations are generated, they are considered visual analytical if they enable interaction directly or indirectly with part, or all of the information being represented. Examples for such interactivity include transforming a top-down tree into a circular tree, coloring nodes in the tree based on specific properties such as gender, or dragging a node in the tree to swap its location with another sibling node.

Similarly, there have been several attempts to classify interactions with visualizations at different levels of granularity. For example, Amar et al. (2005) proposed 8 low-level interaction intents: retrieve value, filter, compute derived value, find extremum, sort, determine range, characterize distribution, find anomalies, and cluster and correlate. In contrast, Yi et al. (2007) proposed 6 higher level interaction intents typically used: select, explore, reconfigure, encode, abstract/elaborate, filter and connect.

While the above classifications of visual analytical representations and interaction with them are useful as check lists for building effective visual analytical systems, they do not provide an integrated understanding of how they work together to enable analytical reasoning, a primary goal of visual analytics. To address this gap, Liu and Stasko (2010) proposed a framework which integrates visual representation, interaction, and analytical reasoning. The framework specifies that central to reasoning with an external visual analytical representation (e.g., the table in Fig. 14.1b) is a *mental model* which is an analog of the external representation stored in working memory, and which is "runnable" to enable reasoning of the data and relationships. This is achieved by creating a mental model in working memory which is a "collage" of some or all of the structural, semantic, and elemental details present in the visual representation, in addition to other information from long term memory relevant to the task. For example as shown in Fig. 14.1b, an analyst conducting the task of determining which of the two columns have more patients with systolic >140 might construct a mental model in working memory consisting of two columns with cells colored red and white, but excluding elements such as the numbers in the cells. Similar to the speed of accessing information stored in the memory of a computer versus from disk, a mental model stored in the brain's working memory can be used to rapidly achieve tasks such as determining which of the two columns have more red cells, or even determining that the first column has approximately three times more red cells compared to the second column.

The framework further specifies that because working memory has size constraints, a mental model can typically contain only some of the information present in the external visualization at any given time. Therefore, when the task changes, it motivates a tight interactive coupling between the internal mental model and the external visual representation, through which new information is extracted from the existing state of the visualization or from long term memory, irrelevant information in the mental model is discarded to make room for new information, the external visual representation itself is transformed to reveal new relationships, or the conceptual information is externalized onto the visual representation to enable future tasks. For example, when the task described in Fig. 14.1 involves exploring or determining the relationship of systolic blood pressure to gender, then a tight coupling between the internal and external representations is triggered enabling the extraction of gender-related information and its relationship to systolic blood pressure. This can be done either by extracting the information from the current representation (requiring often costly mental manipulations) to identify patterns, or by transforming the external representation through manipulations such as sorting (requiring relatively cheaper physical actions) to reveal new relationships, which are then immediately available for internal reasoning tasks such as determining inequalities between the columns. Furthermore, information about the current or previous task such as a discovered pattern can be externalized onto the visual representation through annotations, and therefore freeing up working memory for subsequent tasks.

The framework proposes that the coupling of internal and external representations can be characterized by three interacting goals: (1) *External anchoring* or the process of connecting conceptual structures (e.g., systolic blood pressure >140) to material elements of the visualization (red colored cells), (2) *Information foraging* or the process of exploring the external visual representation through extraction (e.g., counting the red cells related to female patients) or through transformation (e.g., sorting) of the representation, and (3) *Cognitive offloading* or the process of transferring a conceptual structure onto the visual representation to reduce working memory demands (e.g., encircling or annotating in Fig. 14.1c all female patients who have systolic >140 before and after taking the drug).

While the above integrated framework of visual representation, interaction, and analytical reasoning still needs to be elaborated into a theory and tested through predictive models, it provides a first step into how the critical concepts of visual analytics could be working together to enable analytical reasoning, leading to implications for the design and evaluation of effective visual analytical systems.

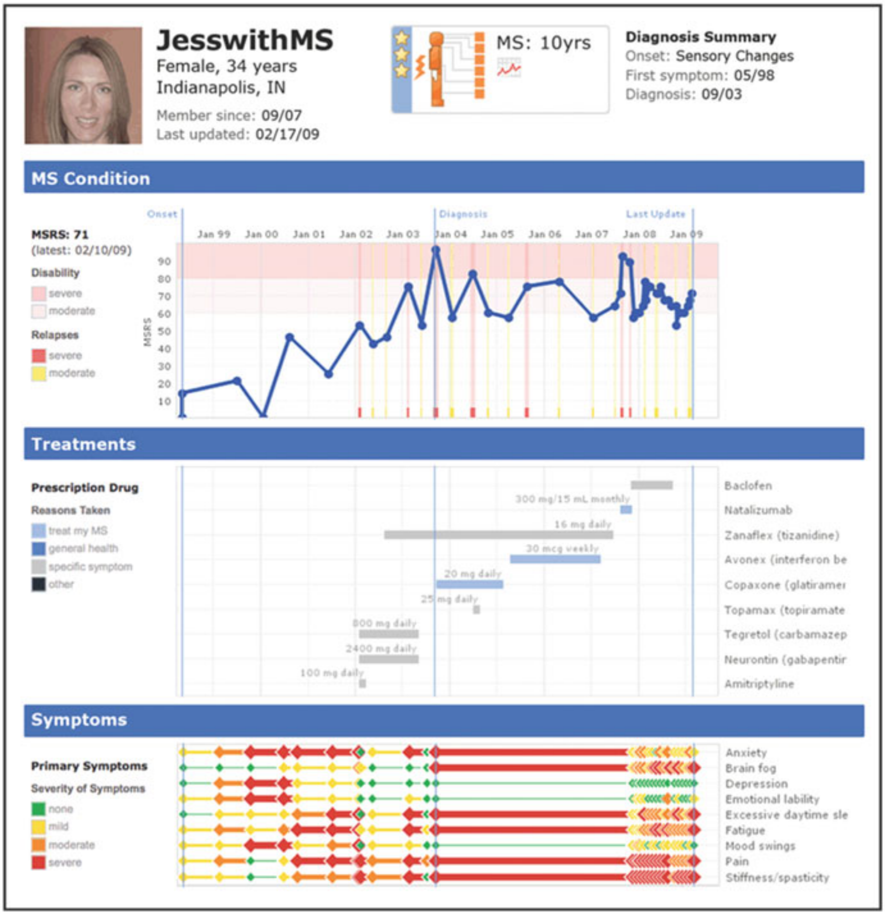
Finally, it is important to note that visual analytics has considerable overlap with the fields of scientific visualization (focused on modeling real-world geometric structures such as earthquakes), and information visualization (focused on modeling abstract data structures such as relationships). However, as described above, visual analytics places a large emphasis on approaches that facilitate reasoning and making sense of complex information individually and in groups (Thomas and Cook 2005).

14.3 Visual Analytics: Biomedical Applications

The use of visual analytical representations is increasingly becoming pervasive in the biomedical domain. The selection of visual analytical representations is highly dependent on the users of the information and their goals, which can be classified in the following two broad categories:

14.3.1 Information Consumers

The primary goal of information consumers is to make biomedical information actionable in terms of directly affecting change in health-related behaviors. An important class of information consumers is patients and care providers whose primary goal is to track and modify personal health and life style behaviors through the use of biomedical and social data. For example, the website *PatientsLikeMe* (2014) enables users to input health and lifestyle variables of specific individuals. As shown in Fig. 14.2, this information is displayed using visual analytical representations such as longitudinal charts and graphs which can be modified to display



JesswithMS
Female, 34 years
Indianapolis, IN
Member since: 09/07
Last updated: 02/17/09

MS: 10yrs

Diagnosis Summary
Onset: Sensory Changes
First symptom: 05/98
Diagnosis: 09/03

Fig. 14.2 A visual analytical display of patient information provided by PatientsLikeMe, a website that enables patients and caregivers to upload information about individuals, and search for other patients with a similar condition (Reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology (Brownstein et al. 2009), copyright 2009)

different granularities of data. Users can also find patients who are similar to their profile, and learn about their real-world experiences of dealing with their diseases, with the goal of improving the quality of life for themselves or for those they provide care. Similarly, personal and wearable activity monitors (e.g., fitbit) have been developed to motivate behavior change such as weight loss by monitoring how many steps a user has taken on a particular day, and displaying that information on a smart phone using visualizations such as a progress bar and the recommended target. Such information can be shared with other users in a social network to provide additional motivation through competition.

Another important class of information consumers consists of healthcare providers such as physicians and first-responders whose primary goal is to make healthcare decisions relevant to specific patients and situations by extracting relevant information from databases such as electronic health records. For example, the Twinlist system (Plaisant et al. 2013) was developed to reconcile multiple lists of drugs (e.g., from the hospital records versus what the patient reports taking) associated with a patient by graphically displaying what is similar and different among the different lists. The goal of this prototype was to enable caregivers to rapidly reconcile contradictory information with the goal of reducing errors in treatment.

A third class of information consumers consists of policy makers from federal and state agencies whose primary goal is to make policy decisions based on public health information. For example, the Centers of Disease Control provides interactive maps showing the incidence of different disease outbreaks across the US (CDC 2014), with the goal of enabling faster response.

Given that the primary goal of information consumers is to make specific forms of biomedical information actionable, an active area of research is to determine which visual analytical representations are appropriate for which classes of users and goals, and to design and evaluate systems which are easy to learn, and intuitive to use (Shneiderman et al. 2013). For example, while interactive time series, maps, and hierarchies when designed carefully are considered easy to comprehend and to interact with, other representations such as networks with more than a few dozen nodes are considered more difficult to comprehend and tend to be avoided as representations for information consumers.

14.3.2 Information Analysts

In contrast to information consumers, the primary goal of information analysts in academic and industrial settings is to make contributions to biomedical scientific knowledge. While the goal of all biomedical information users is to ultimately improve health outcomes, the process of reaching that long-term goal is achieved by information analysts through progressive contributions to scientific knowledge. An important class of information analysts consists of biologists and bioinformaticians whose primary goal is to decipher the biological mechanisms involved

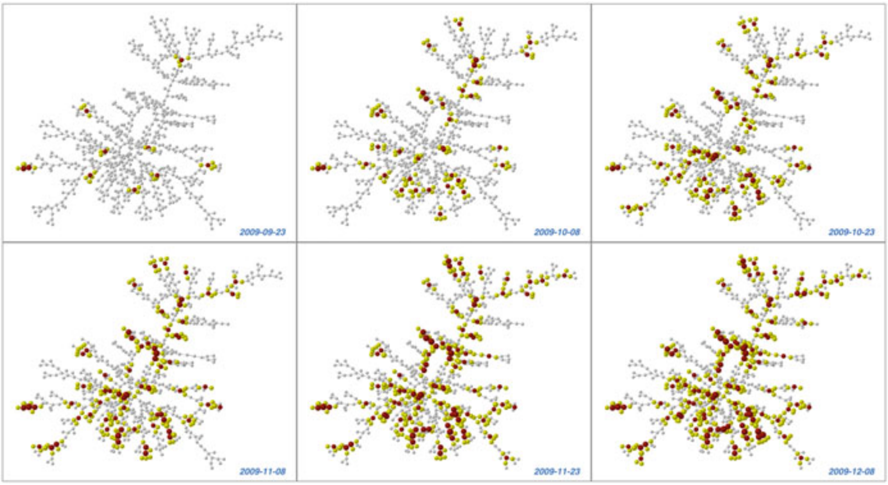


Fig. 14.3 Progression of the flu infection through a social network of students from Harvard University (Christakis and Fowler 2010). The *red nodes* represent infected students, the *yellow nodes* represent friends of infected students, and the edges connecting the nodes represent self-reported friendship links (Reprinted under the Creative Commons Attribution license)

in different diseases. For example, biologists often use network visualization and analysis tools like Cytoscape (2014) to comprehend complex disease-protein associations (Ideker and Sharan 2008) with the goal of deciphering the functions and pathways related to proteins of interest.

A second class of information analysts consists of clinical researchers and medical informaticians whose primary goal is to develop new methods to improve patient treatment by analyzing the relationship between clinical variables and outcomes. For example, networks visualizations have been used to analyze Medicare claims from more than 30 million patients, which enabled researchers to infer patterns in the progression of different diseases (Hidalgo et al. 2009). One of the their observations was that that highly connected nodes in the network had high lethality implying that patients with such diseases are more likely to have an advanced stage of disease.

A third class of information analysis consists of epidemiologists whose primary goal is to analyze public health information. For example as shown in Fig. 14.3, Christakis and Fowler (2010) found that the flu infection in a social network consisting of Harvard students peaked two weeks earlier compared to a random set of students from the same population. Such advanced warning could be effective for planning immunizations during outbreaks of infectious diseases.

An active area of visual analytics research is to develop new approaches that integrate molecular, clinical, and epidemiological information, in a single representation. For example, translational scientists working in teams have used network visualization and analyses to integrate molecular and clinical information with the goal of inferring heterogeneity in asthma, and the respective biological mechanisms (e.g., Bhavnani et al. 2014a, b).

Given the importance of networks for the analysis and presentation of complex relationships in a wide range of data types, and because it is one of the most advanced form of visual analytics, the rest of this chapter focuses on providing a concrete understanding of this approach as applied to the integrative analysis of molecular and clinical information.

14.4 Network Analysis: Making Discoveries in Complex Biomedical Data

Networks (Newman 2010) are an effective representation for analyzing biomedical data because they enable an interactive visualization of complex associations. Furthermore, because they are based on a graph representation, they also enable the quantitative analysis and validation of the patterns that become salient through the visualization. Networks are increasingly being used to analyze a wide range of molecular measurements related to gene regulation (Albert 2004), disease-gene associations (Goh et al. 2007), and disease-protein associations (Ideker and Sharan 2008). A network (also called a graph) consists of a set of nodes, connected in pairs by edges; nodes represent one or more types of entities (e.g., patients or genes). Edges between nodes represent a specific relationship between the entities (e.g., a patient has a particular gene expression[1] value). Figure 14.4 shows a sample bipartite network where edges exist only between different types of entities (Newman 2010), in this case between patients and genes.[2]

Network analysis of biomedical data typically consists of three steps: (1) **exploratory visual analysis** to identify emergent bipartite relationships such as between patients and genes; (2) **quantitative analysis** through the use of methods suggested by the emergent visual patterns; (3) **inference** of the biological mechanisms involved across different emergent phenotypes. This three-step method used across several studies (Bhavnani et al. 2010, 2011b, 2012) have revealed complex but comprehensible visual patterns, each prompting the use of quantitative methods that make the appropriate assumptions about the underlying data, which in turn led to inferences about the biomarkers and underlying mechanisms involved. Each of the three steps of this method is described below, followed by its application to analyze a data set of subjects and gene expressions.

[1] Gene expression is the process by which the information in a gene is translated into a gene product such as a protein which can be involved in biological processes like inflammation during an infection.

[2] Researchers have explored a wide range of network types including unipartite, directed, dynamic, and networks laid out in three dimensions to analyze complex data. As this wide range is beyond the scope of this chapter, we suggest other excellent sources (Newman 2010) for such information.

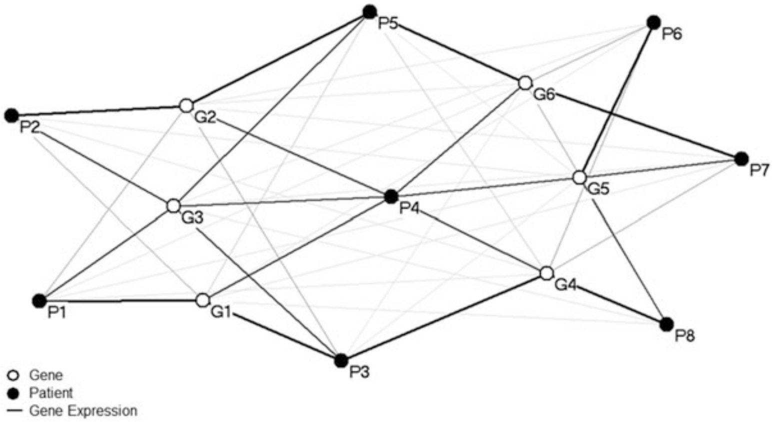


Fig. 14.4 A sample bipartite network where edges exist only between two different types of nodes. In this case, nodes represent either patients (*black*) or genes (*white*), and edges connecting the two represent gene expression

14.4.1 Exploratory Visual Analysis

Network analysis typically begins by transforming symbolic data into graphical elements in a network. To achieve this, the analyst needs to decide which *entities* in the data represent the nodes in the network, in addition to how other useful information can be mapped onto the node's shape, color, and size. Similarly, the analyst needs to decide which *relationships* between the entities in the data are represented by the edges in the network, in addition to how to map other useful information to the edge's thickness, color, and style. These selections are made based on an understanding of the kinds of relationships that need to be explored, and is often an iterative process based on an understanding of the domain and the nature of the data being processed.

Once the symbolic data has been mapped to graphical elements, the resulting network is laid out so the nodes and edges can be visualized. The layout of nodes in a network can be done where either the distances between nodes has no meaning (e.g., nodes laid out randomly or along a geometric shape such as a line or circle), or where the distance between nodes represents a relationship such as similarity (e.g., similar cytokine expression profiles). Layouts where distance has meaning are typically generated through force-directed layout algorithms. For example, the application of the *Kamada-Kawai* (1989) layout algorithm to a network results in nodes with a similar pattern of connecting edge weights to be pulled together, and those with different patterns to be pushed apart.

Figures 14.5, 14.6, 14.7 and 14.8 show the steps that were used to generate a bipartite network of 101 subjects and 18 genes, data which is described in more detail in the original study (Ioannidis et al. 2012). The 101 subjects consisted of 28 influenza (flu), and 51 respiratory syncytial virus (RSV) cases, and 22 age,

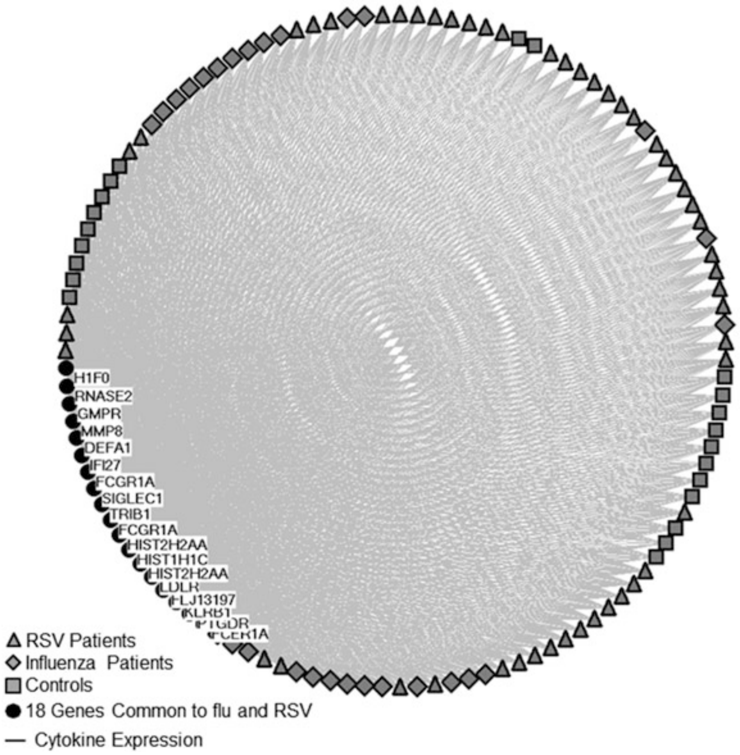


Fig. 14.5 A bipartite network showing subject nodes (RSV patients = *triangles*, flu patients = *diamonds*, and controls = *squares*) and gene nodes (*black circles*) connected in pairs by edges, which represent normalized gene expression. Patient and gene nodes were separately grouped and randomly laid out equidistantly around a circle

gender, and race matched healthy controls. The 18 genes were highly significant, differentially-expressed genes that were common to both infections. The goal of this analysis was to identify subgroups of cases that had different molecular profiles and therefore could suggest sub-phenotypes that require different treatments. Figure 14.5 shows how the three types of subjects were represented as RSV (gray triangles), flu (gray diamonds), and controls (gray squares), and the genes were represented as circular black nodes. Furthermore, normalized gene expression values were represented as edges connecting each subject to each gene. These nodes were laid out equidistantly around a circle. Figure 14.6 shows the same network but where the edge thicknesses are proportional to the normalized gene expression values. Therefore, thicker edges represent higher gene expression values as compared to the thinner edges. Furthermore, the size of the node was made proportional to the total expression value of the connecting edges. Therefore, larger patient nodes have overall higher aggregate gene expression values compared to smaller patient nodes.

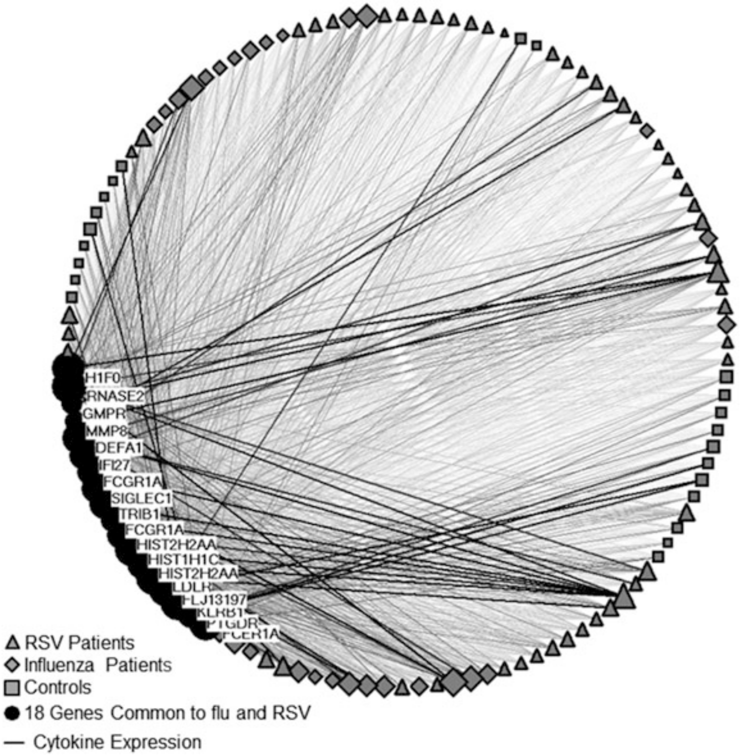


Fig. 14.6 The same network as in Fig. 14.5 but where edge thickness is proportional to the normalized gene expression value and the size of each node is proportional to the total expression values of the connecting edges. Thick edges represent higher gene expression values compared to thin edges. Similarly, larger subject nodes have higher aggregate gene expression values compared to smaller patient nodes

Although the patients, genes, and the gene expression have been visually represented, the distances between the nodes have no meaning. To better comprehend the data, the subjects that have higher expression value for a particular gene should be spatially closer to that gene compared to those that have lower gene expressions. This approach of using short distances between entities to show similarity, and long distances between entities to show dissimilarity is typical across clustering algorithms. As shown in Fig. 14.7 and previously reported (Bhavnani et al. 2014a, b), application of the forced-directed algorithm Kamada-Kawai to the circular layout results in nodes that have a similar pattern of gene expression to be pulled together, and those that are not similar to be pushed apart.

The resulting layout suggests that there exist distinct clusters of subjects and genes. As shown in Fig. 14.7, the subjects had a complex but understandable topology consisting of a majority of the cases (triangles and diamonds) on the top cluster which had a preferential expression of the top 14 genes, and a majority of the

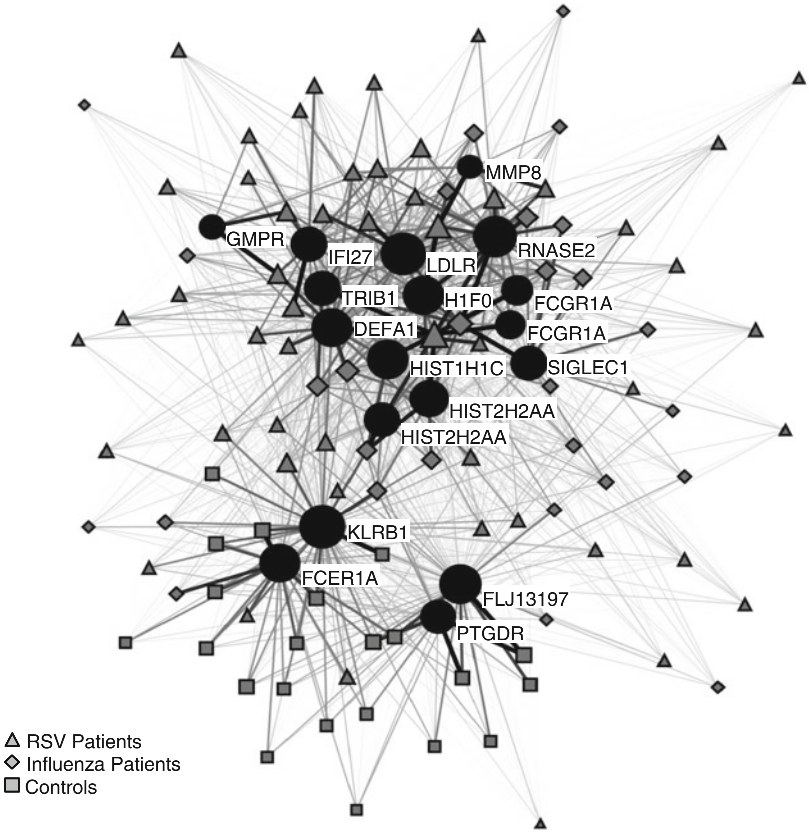


Fig. 14.7 Application of *Kamada-Kawai*, a force-directed algorithm, to the circular layout. The algorithm pulls nodes with similar gene expression patterns closer together while pushing apart those with dissimilar expression patterns. The layout of the network suggested the existence of distinct subject and gene clusters, and revealed inter-cluster relationships such as how the subject clusters express particular gene clusters. However, quantitative methods must be used to identify cluster boundaries

controls (squares) at the bottom of the network which had preferential expression of the bottom 4 genes. In addition, the cases on the top had a core-periphery topology, where there were some cases with high overall gene expression in the center, and many patients with low overall gene expression in the periphery. Finally, there were four cases (triangles and diamonds) that were clustered with the controls at the bottom of the network.

While the network layout suggests the existence of distinct clusters, it is not designed to reveal the members of each cluster. We therefore need to use quantitative methods that are explicitly designed to identify the boundaries of clusters based on a multivariate analysis of the data.

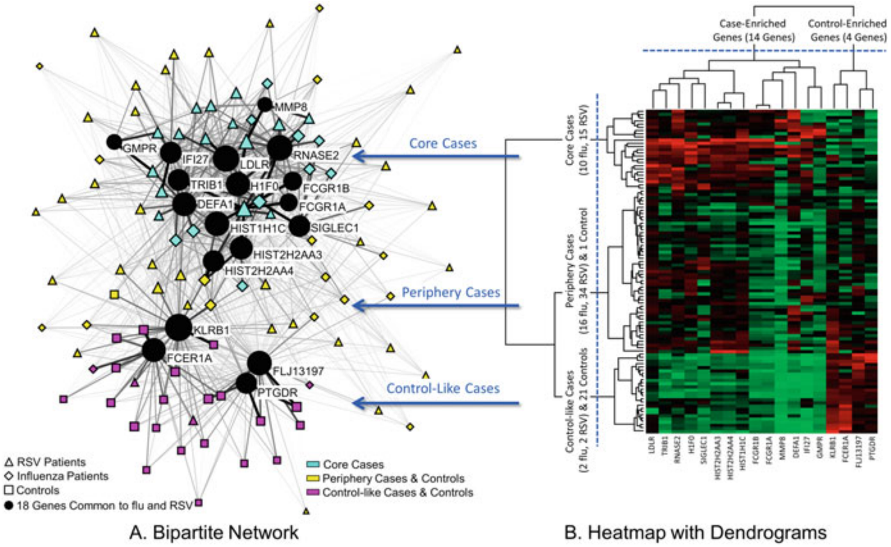


Fig. 14.8 A heatmap with dendrogram generated through hierarchical clustering helped to identify the boundaries of three subject clusters, which were superimposed onto the network shown in Fig. 14.4 using colored nodes to denote cluster membership. The network also shows the relationship of the subject clusters to the top gene cluster consisting of 11 genes, and bottom gene cluster consisting of 4 genes (Bhavnani et al. 2014a)

14.4.2 Quantitative Verification and Validation

There exist a wide range of quantitative methods to verify and validate patterns discovered through network visualization methods. While in principle any statistical method can be used to quantitatively analyze a pattern observed in a network, many patterns are often analyzed using graph-based methods (Newman 2010) that specialize in analyzing complex relationships. For example, *degree assortativity* measures whether one type of nodes in a network which have high weighted degree (e.g., subjects that have large nodes in Fig. 14.7), are preferentially connected to another type of nodes that have high degree (e.g., genes that have large nodes in Fig. 14.7), or vice versa.

Another approach that can be used to verify patterns in a network is hierarchical clustering (Johnson and Wichern 1998). This unsupervised learning method attempts to identify the number and boundary of clusters in the data. For example, hierarchical clustering can be used to identify clusters of patients based on their relationship to genes, or clusters of genes based on their relationship to patients. The method begins by putting each node in a separate cluster, and then progressively joins nodes that are most similar based on their relationship to connected nodes. This progressive grouping generates a tree structure called a *dendrogram*, where distances between subsequent layers of the tree represent the strength of dissimilarity between the respective clusters; the larger the distance between two subsequent layers, the stronger the clustering. Analysts therefore determine the number and membership of the clusters by identifying relatively large breaks between the layers in the dendrogram.

Given the wide range of quantitative methods available, the patterns in the network are used to guide the selection of the appropriate method. For example, if distinct clusters do not exist in a network, then it is not appropriate to apply a clustering algorithm to the network. This approach of selecting methods based on the inspection of the data is similar to how statisticians determine whether to use parametric or non-parametric inferential methods based on the underlying distribution of the data.

Because the network in Fig. 14.7 suggested the existence of distinct clusters, hierarchical clustering was used to identify the boundary and members of the clusters. As shown in Fig. 14.8b, the horizontal dendrogram represents the gene clusters, the vertical dendrogram represents the patient clusters, and the colored cells represent normalized gene expression ranging from green (0) to red (1). The dendrograms shows a clear break at two clusters for the genes, and three clusters for subjects (as shown by the corresponding blue dotted lines across each dendrogram).

While there may be clear breaks in the dendrograms, the overall pattern could have occurred by random chance. Patterns discovered in networks, and subsequently the dendrograms, are therefore, validated by determining their significance. One approach to do this is to compare the patterns in the data to random permutations of the network.

To test whether there were significant breaks in the dendrogram (denoting the existence of distinct clusters), the variance, skewness, and kurtosis of the dissimilarities (generated by the hierarchical clustering algorithm) in the flu/RSV network were compared to 1,000 random permutations of the data. For each network permutation, the number of nodes and the number of edges connected to each node, in addition to the edge weight distribution of subjects were preserved when analyzing the gene dendrogram, and vice versa. Significant breaks in the subject or gene dendrograms would result in a significantly larger variance, skewness, and kurtosis of the dissimilarity measures, compared to the same measures generated from the random networks. As previously reported (Bhavnani et al. 2014a, b) the results showed the clusteredness of the subjects in the network was significant as measured by the variance of the dissimilarities (flu/RSV = 2.75, Random-Mean = 0.88, $p < .001$ two-tailed test), skewness of the distribution of dissimilarities (flu/RSV = 5.55, Random-Mean = 3.94, $p < .001$ two-tailed test), and kurtosis of the distribution of dissimilarities (flu/RSV = 38.69, Random-Mean = 25.03, $p < .001$ two-tailed test).

The same approach was used to test the clusteredness of the gene clusters. The results showed that the gene clustering was also significant when compared to 1,000 random networks based on variance of the dissimilarities (flu/RSV = 2.91, Random-Mean = 0.24, $p < .001$ two-tailed test), skewness of the distribution of dissimilarities (flu/RSV = 2.01, Random-Mean = 0.80, $p < .001$ two-tailed test), and kurtosis of the distribution of dissimilarities (flu/RSV = 7.81, Random-Mean = 3.16, $p < .001$ two-tailed test).

To understand why the subjects and genes were clustered, and how they related to each other, the cluster memberships were superimposed onto the network. As shown in Fig. 14.8a, the subject nodes were colored (blue, yellow, and pink) to denote their membership in three separate clusters referred to as core cases, periphery cases, and control-like cases. Furthermore, the 14 genes on the top, and the 4 genes at the bottom also formed distinct clusters, but because they were easy to distinguish by their spatial separation, they were kept black to reduce visual complexity.

As shown in Fig. 14.7, in addition to the above clustering, the core cases appeared to have higher overall gene expression (based on their size which is proportional to the sum of their edge weights) compared to the periphery cases. This pattern was quantitatively verified by comparing the weighted degree centrality (sum of edge weights) of the core cases to those of the periphery cases. This can be done with well-known statistical tests such as the Mann Whitney *U* test, a non-parametric test, which can be used to determine if the median of a variable is significantly different across two groups.

The results showed that the core cases (Median = 4.55) was significantly different ($U = 49.00$, $p < .001$, two-tailed test) compared to the periphery cases (Median = 2.52) verifying that the overall gene expression of the patients in the core was higher compared to those in the periphery. Furthermore, the median gene expression of the 14 genes across the 25 core cases (Median = 4.22) was significantly higher ($U = 16$, $p < .001$, two-tailed test) compared to the 50 periphery cases (Median = 1.95). This pattern can also be seen in the high expression values (shown in mostly red cells) in the upper left-hand corner of the heatmap in Fig. 14.8b. Finally, there was no significant difference ($\chi^2(2, \mathrm{N} = 79) = 0.86$, $p = 0.652$) in the proportion of flu vs. RSV patients across the three case clusters, suggesting that the gene-based clustering was common across both types of infection.

The above results of the cluster analysis superimposed over the network, in addition to quantitative analysis of gene expression across the clusters enabled the identification of three potential sub-phenotypes: (1) **core-cases** who had a significantly higher gene expression of the top cluster of 14 genes, (2) **periphery cases** who had a medium expression of the top 14 genes, and (3) **control-like cases** whose profiles were similar to the controls with high expression of the bottom cluster 4 genes. These three sub-phenotypes were common across both infections.

14.4.3 Inference of Sub-phenotypes and Biological Mechanisms

While the visual and quantitative analysis helped to reveal patterns in the data, the ultimate goal of the network analysis is to infer the biological mechanisms involved, and the emergent sub-phenotypes in the data. This inferential step requires an integrated understanding of the molecular and clinical variables.

One approach to conduct such an integrated analysis, is to analyze how the patients in each emergent cluster (based on molecular profiles), differ in their clinical variables. As the primary data included disease severity of each patient (Ioannidis et al. 2012), we used the Mann Whitney *U* test to analyze if the core and periphery cases were significantly different in their disease severity. The test revealed that the disease severity of core cases (Median = 7) was significantly higher ($U = 261.50$, $p < .001$, two-tailed test) compared to periphery cases (Median = 2). This result suggested a significant association between the high gene expression of the 14 top genes in the core-cases, and higher disease severity.

The bipartite visualization and quantitative verifications therefore revealed not only sub-phenotypes based on the molecular profiles, but also how they related to clinical variables, which enabled the domain experts to infer three possible sub-phenotypes and their potential pathways (Bhavnani et al. 2014a, b).

1. The **core cases** have significantly higher expression of 14 up-regulated genes, which included 4 histone genes, 4 genes with to date have unknown function in antiviral response, and 6 immune-related genes each of which has a well-known non-overlapping antiviral function. An Ingenuity Pathway Analysis (Ingenuity 2014) of the 14 genes suggested an indirect but strong interferon signature including TNFα and IL-6 cytokines involved in antiviral and innate inflammatory responses. Because the core cases also had a significantly higher disease severity score, they represent a distinct at-risk sub-phenotype that are hyper responsive to pathways targeted to viral clearance, and possibly carry a risk for long-term epithelial cell damage.
2. The **periphery cases** have a medium expression of all 18 genes and therefore suggest a second subphenotype with a subdued anti-viral response relative to the above hyperresponders.
3. The **control-like cases** have a high expression of 4 down-regulated genes, and low expression of the 14 up-regulated genes, and therefore mirror the expression patterns in uninfected controls. The results therefore suggest that the down-regulation of these 4 genes indicates a "protective" phenotype making them similar to the uninfected controls. Existing literature on these genes provide some confirmatory evidence. While the exact role of the high-affinity receptor which binds to the constant portion of IgE (FcER1) is unknown in viral pathogenesis, SNPs included on this gene have been shown to be associated with severe RSV disease (Janssen et al. 2007). Additionally, KLRB1, which has been shown to have inhibitory functions on natural killer (NK) cells (Pozo et al. 2006) was downregulated, suggesting an enhanced antiviral response in patients resembling the immune response of controls. Finally, PTGDR a receptor important in mast cell function was downregulated, but the exact role of this receptor in viral infection is still unknown. Overall, control-like cases suggests a third subphenotype which have a "just enough" response to the virus, without overt

stimulation of virally induced genes, and therefore potentially with reduced bystander damage.

One might argue that the above result could also be the result of the progression of infection over time. For example, the core cases could be at the peak of infection, the periphery cases could be later in the infection, and the control-like cases could be recovering from the infection. However, an additional analysis revealed that the 3 case clusters were not significantly different ($H(2, \mathrm{N} = 79) = 2.56$, $\mathrm{p} = 0.278$) in time of sample collection after hospitalization. There is of course the possibility that the children were infected at very different times before hospitalization, but controlling such a variable is practically impossible in the analysis of naturally infected humans. Therefore, we provide two explanations for why sample collection time is probably not an adequate explanation for the results: (1) Because all case samples were collected from patients that were hospitalized indicating severe illness, a resolution of such severity in the short time window of 42–72 h is unlikely to occur. (2) The gene expression changes in the PBMCs of the patients suggest a specific induced innate immune response (e.g., Toll-like receptor) to viruses. Such signaling pathways (which induce interferon secretion and contribute to anti-viral immunity) last several days which exceeds the sample collection time window in this study. We therefore propose that the three case clusters are more likely the result of inherent host differences in anti-viral responses, and therefore represent distinct sub-phenotypes.

Informed by these underlying molecular processes, the network analysis of subjects and genes therefore helped to infer not only the sub-phenotypes, but also the possible mechanisms involved, and which sub-phenotypes had a high risk of developing severe complications. The results therefore provided data-driven hypotheses of sub-phenotypes and their mechanisms which can be validated in future research with other datasets. Such analysis therefore could lead to future treatments that are targeted to specific sub-phenotypes, and is therefore an important step towards precision medicine.

14.5 Strengths and Limitations of Network Analysis

Network analysis has several strengths and limitations, whose understanding can lead to informed uses of the method, appropriate interpretation of the results, and insights for future enhancements and complementary methods.

14.5.1 Strengths

Network visualization and analysis provide four distinct strengths for enabling rapid discovery of patterns in complex biomedical data.

1. **Provides Integrative Visualizations.** Because networks are based on graph theory, they provide a tight integration between visual and quantitative analysis. For example as shown in the Fig. 14.8a, networks enable the integrative visualization of multiple raw values (e.g., subject-gene associations, gene expression values, subject phenotype), aggregated values (e.g., sum of gene values), and emergent global patterns (e.g., clusters) in a single representation. This uniform visual representation leverages the parallel processing power of the visual cortex enabling the comprehension of complex multivariate, quantitative relationships.
2. **Guides Quantitative Analysis.** Networks do not require *a priori* assumptions about the relationship of nodes within the data, in contrast to hierarchical clustering or k-means which assume the data is hierarchically organized or contain disjoint clusters, respectively. Instead, by using a simple pairwise representation of nodes and edges, network layouts enable the identification of multiple structures (e.g., hierarchical, disjoint, overlapping, nested) in a single representation (Nooy et al. 2005). Therefore, while layout algorithms such as Kamada-Kawai depend on the force-directed assumption and its implementation, such algorithms are viewed as less biased for data exploration because they do not impose a particular cluster structure on the data, often leading to the identification of more complex structures in the data (Bhavnani et al. 2010). The overall approach therefore enables a more informed selection of quantitative methods to verify the patterns in the data.
3. **Enables Pathway Inference through Co-occurrence.** Network layouts such as the one shown in Fig. 14.8a, preserve highly-correlated variables (such as genes) and display them through clustering. Furthermore, the bipartite network representation enables the comprehension of inter-cluster relationships such as between variable (e.g., genes) clusters and subject clusters. These features provide important clues to domain experts about the pathways that involve those variables. This is in contrast to many supervised learning methods which drop highly correlated variables in an attempt to identify a small number of variables that together can explain the maximum amount of variance in the data. While this approach is powerful for developing predictive models, the reduction in variables could limit the inference of biological pathways involved in the disease.
4. **Accelerates Discovery through Interactivity.** Networks enable high interactivity enabling the rapid modification of the visual representation to match the changing task and representation needs of analysts during the analysis process. For example, nodes that represent patients in a network can be interactively colored or reshaped to represent different variables such as gender and race, enabling the discovery of how they relate to the rest of the network.

14.5.2 Limitations

Networks have three important limitations that are important to understand for their current use, and need to be addressed in future research.

1. **Constrains Number of Node Properties.** While node shape, color and size can represent different variables, there is a limit on the number of variables that can be simultaneously represented. Furthermore, a visual representation can get overloaded with too many colors and shapes, which can mask rather than reveal important patterns in the data. Therefore, while networks can reveal complex multivariate patterns in the data based on a few variables, they often require complimentary visual analytical representations such as Circos ideograms (Krzywinski et al. 2009; Bhavnani et al. 2011a) to explore data that is high-dimensional (e.g., large number of attributes related to entities such as subjects in the network).
2. **Requires Advanced Computational Skills.** While networks provide a rich vocabulary of graphical elements to represent data, their design and use requires iterative refinement based on an understanding of the domain, knowledge of graphic design and cognitive heuristics, and the use of complex interfaces that are designed for those facile in computation. This combination of knowledge required to conduct network analyses makes domain experts dependent on network analysts to generate and refine the representations, which can limit the rapid exploration and interpretation of complex data.
3. **Lacks Systematic Approaches for Finding Structure in Hairballs.** While network layout algorithms are designed to reveal complex and unbiased patterns in multivariate data, they often fail to show any patterns in the data resulting in what is colloquially called a "hairball". In such cases, the nodes appear to be randomly laid out providing little guidance for how to proceed with the analysis. While network applications offer many interactive methods to filter data such as by dropping edges and nodes based on different thresholds, many of these methods are arbitrary and therefore unjustifiable to use when searching for patterns especially in important domains such as biomedicine. There is therefore a need to develop more systematic and defensible methods to find hidden patterns in network hairballs.

14.6 Future Directions in Network Analysis of Biomedical Data

The limitations of networks discussed above motivate future research with the goal of overcoming theoretical, practical, and pedagogical hurdles. **Theoretically**, we need better frameworks that tightly integrate existing theories from cognition, mathematics, and graphic design. Such theories can help predict for example which combination of visual representations can together help researchers to best comprehend patterns in different types of data such as genes versus cytokines. Furthermore, given that many network layouts show no structure, future algorithms should attempt to integrate different methods from machine learning to enable the discovery of hidden patterns. These research directions could enable the rapid discovery of patterns in the age of big data and translational medicine. **Practically**, visual analytical tools tend to be designed for analysts, often requiring substantial programming to make a dataset ready for visualization, and therefore limiting the use of the methods to only a few biologists and physicians. This hurdle motivates the need for tools that enable biologists and physicians to explore data on their own so that they can better leverage their domain knowledge in interpreting the patterns in the data. Of course such patterns need to be statistically validated by subsequent analyses, but currently the exploration and validation is done mostly by analysts, who could miss important associations due to the lack of domain knowledge. **Pedagogically** there needs to be a concerted effort to train the next generation of biomedical informaticians for developing and using novel visual analytical approaches, and to train biologists and physicians on how to make important biomedical discoveries in visual analytical representations of their data. Such advances should enable visual analytics to fully realize its potential to accelerate discoveries in increasingly complex and big biomedical data.

Discussion Questions

1. Why are visualizations and interactivity critical in making discoveries in complex biomedical data?
2. What are the strengths and limitations of networks, and how can future research fully exploit the strengths, and overcome the limitations?

Acknowledgements I thank Shyam Visweswaran, Rohit Divekar, and Bryant Dang for their contributions to this chapter. This research was supported in part by NIH CTSA #UL1TR000071, the Institute for Human Infections and Immunity at UTMB, the Rising Star Award from University of Texas Systems, and CDC/NIOSH #R21OH009441-01A2.

Additional Readings

Card, S., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization: Using vision to think*. San Francisco: Morgan Kaufmann Publishers.

Newman, M. E. J. (2010). *Networks: An introduction*. Oxford: Oxford University Press.

Thomas, J. J., & Cook, K. A. (2005). *Illuminating the Path: The R&D agenda for visual analytics national visualization and analytics center*.

Tufte, E. R. (1983). *The visual display of quantitative information*. Chesire: Graphics Press.

References

Albert, R. K. (2004). Boolean modeling of genetic regulatory networks. *Complex Networks, 21*, 459–481.

Amar, R., Eagan, J., & Stasko, J. (2005, October). Low-level components of analytic activity in information visualizations. In *Proceedings of IEEE InfoVis'05*, Minneapolis, MN, USA (pp. 111–117).

Bhavnani, S. K., Bellala, G., Ganesan, A., et al. (2010). The nested structure of cancer symptoms: Implications for analyzing co-occurrence and managing symptoms. *Methods of Information in Medicine, 49*, 581–591.

Bhavnani, S. K., Pillai, R., Calhoun, W. J., et al. (2011a). How circos ideograms complement networks: A case study in asthma. In *Proceedings of AMIA summit on translational bioinformatics*, Bethesda, MD.

Bhavnani, S. K., Victor, S., Calhoun, W. J., et al. (2011b). How cytokines co-occur across asthma patients: From bipartite network analysis to a molecular-based classification. *Journal of Biomedical Informatics, 44*, S24–S30.

Bhavnani, S. K., Bellala, G., Victor, S., et al. (2012). The role of complementary bipartite visual analytical representations in the analysis of SNPs: A case study in ancestral informative markers. *Journal of the American Medical Informatics Association, 19*, e5–e12.

Bhavnani, S. K., Dang, B., Caro, M., Bellala, G., & Visweswaran, S. (2014a). Heterogeneity within and across pediatric pulmonary infections: From bipartite networks to at-risk subphenotypes. In *Proceedings of AMIA summit on translational bioinformatics*, Bethesda, MD.

Bhavnani, S. K., Drake, J. A., & Divekar, R. (2014b). The role of visual analytics in asthma phenotyping and biomarker discovery. In A. Brasier (Ed.), *Heterogeneity in asthma* (pp. 289–305). New York: Springer.

Brownstein, C. A., Brownstein, J. S., Williams, D. S., III, Wicks, P., & Heywood, J. A. (2009). The power of social networking in medicine. *Nature Biotechnology, 27*, 888–890.

Card, S., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization: Using vision to think*. San Francisco: Morgan Kaufmann Publishers.

Centers for Disease Control and Prevention. (2014, April 28). Retrieved from the website http://nccd.cdc.gov/DHDSPAtlas/#

Christakis, N. A., & Fowler, J. H. (2010). Social network sensors for early detection of contagious outbreaks. *PLoS ONE, 5*(9), e12948.

Cytoscape. (2014, April 28). Retrieved from the website http://www.cytoscape.org/

Goh, K., Cusick, M., Valle, D., et al. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 8685.

Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Communications of the ACM, 53*, 59–67.

Hidalgo, C. A., Blumm, N., Barabási, A.-L., & Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology, 5*(4), e1000353.

Ideker, T., & Sharan, R. (2008). Protein networks in disease. *Genome Research, 18*, 644.

Ingenuity. (2014, April 28). Retrieved from the website http://www.ingenuity.com/products/ipa

Ioannidis, I., McNally, B., Willette, M., et al. (2012). Plasticity and virus specificity of the airway epithelial cell immune response during respiratory virus infection. *Journal of Virology, 86*(10), 5422–5436.

Janssen, R., Bont, L., Siezen, C. L., et al. (2007). Genetic susceptibility to respiratory syncytial virus bronchiolitis is predominantly associated with innate immune genes. *Journal of Infectious Diseases, 196*(6), 826–834.

Johnson, R. A., & Wichern, D. W. (1998). *Applied multivariate statistical analysis*. Upper Saddle River: Prentice-Hall.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters, 31*, 7–15.

Krzywinski, M., Schein, J., Birol, I., et al. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research, 19*, 1639–1645.

Liu, Z., & Stasko, J. T. (2010). Mental models, visual reasoning and interaction in information visualization: A top-down perspective. *IEEE Transactions on Visualization and Computer Graphics, 16*(6), 999–1008.

Molloy, J. C. (2011). The open knowledge foundation: Open data means better science. *PLoS Biology, 9*, e1001195.

Newman, M. E. J. (2010). *Networks: An introduction*. Oxford: Oxford University Press.

Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. Cambridge: Cambridge University Press.

Norman, D. (1993). *Things that make us smart*. New York: Doubleday/Currency.

PatientsLikeMe. (2014, April 28). *PatientsLikeMe*. Retrieved from the website http://www.patientslikeme.com/

Plaisant, C., Chao, T., Wu, J., Hettinger, A. Z., Herskovic, J. R., Johnson, T. R., Bernstam, E. V., Markowitz, E., Powsner, S., & Shneiderman, B. (2013, November 16). Twinlist: Novel user interface designs for medication reconciliation. In *Proceedings of AMIA annual symposium* (pp. 1150–1159).

Pozo, D., Valés-Gómez, M., Mavaddat, N., Williamson, S. C., Chisholm, S. E., & Reyburn, H. (2006). CD161 (human NKR-P1A) signaling in NK cells involves the activation of acid sphingomyelinase. *Journal of Immunology, 176*(4), 2397–2406.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualization. *Visual Languages, 93*, 336–343.

Shneiderman, B., Plaisant, C., & Hesse, B. (2013). Improving health and healthcare with interactive visualization tools. *IEEE Computer, 46*(5), 58–66.

Thomas, J. J., & Cook, K. A. (2005). *Illuminating the path: The R&D agenda for visual analytics national visualization and analytics center*.

Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: Can it facilitate? *International Journal of Human-Computer Studies, 57*, 247–262.

Yi, J. S., Kang, Y. A., Stasko, J., et al. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics, 13*, 357–369.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science, 18*, 87–122.