

## **Graph theoretic modeling of large-scale semantic networks**

**Michael E. Bales, MPH**  
**Stephen B. Johnson, PhD**

Department of Biomedical Informatics, Columbia University, New York, NY, USA

**Journal of Biomedical Informatics (in press), 2005**

### **Correspondence and reprints:**

Michael E. Bales, MPH  
Department of Biomedical Informatics  
Columbia University  
Vanderbilt Clinic, 5<sup>th</sup> Floor  
622 West 168th Street  
New York, NY 10032  
[michael.bales@dbmi.columbia.edu](mailto:michael.bales@dbmi.columbia.edu)  
Fax 208-694-4181

## Abstract

During the past several years, social network analysis methods have been used to model many complex real-world phenomena, including social networks, transportation networks, and the Internet. Graph theoretic methods, based on an elegant representation of entities and relationships, have been used in computational biology to study biological networks; however they have not yet been adopted widely by the greater informatics community. The graphs produced are generally large, sparse, and complex, and share common global topological properties. In this review of research (1998-2005) on large-scale semantic networks, we used a tailored search strategy to identify articles involving both a graph theoretic perspective and semantic information. Thirty-one relevant articles were retrieved. The majority (28, 90.3%) involved an investigation of a real-world network. These included corpora, thesauri, dictionaries, large computer programs, biological neuronal networks, word association networks, and files on the Internet. Twenty-two of the 28 (78.6%) involved a graph comprised of words or phrases. Fifteen of the 28 (53.6%) mentioned evidence of small-world characteristics in the network investigated. Eleven (39.3%) reported a scale-free topology, which tends to have a similar appearance when examined at varying scales. The results of this review indicate that networks generated from natural language have topological properties common to other natural phenomena. It has not yet been determined whether artificial human-curated terminology systems in biomedicine share these properties. Large network analysis methods have potential application in a variety of areas of informatics, such as in development of controlled vocabularies and for characterizing a given domain.

## Introduction

This is a review of a knowledge representation approach known as **graph theoretic** modeling, and of how this approach has been applied to the study of **semantic networks**. The approach draws upon the mathematical formalisms of graph theory and upon analytic methods refined over decades of **social network** research. Networks consist of **nodes**, which represent entities, and lines, or **edges**, drawn between the nodes to indicate a connection between them. Advances in computer speed have provided an infrastructure for modeling of large and complex **network** models. These models allow for a study of relationships between entities both at the global and local level.

In the informatics community, the first wide-scale uses of this technique have been in bioinformatics and computational biology. Various biological systems such as protein-protein interaction and genetic regulatory networks have been studied, sometimes yielding new insights into cellular and molecular pathways and interdependencies. Network models are also used in informatics research in social[1] and cognitive science. Computational biology, social science, and cognitive science are all gaining prominence as areas of specialization in informatics, and all have adopted graph theoretic modeling; therefore it is possible that the approach will continue to permeate other fields of informatics.

Within a broader context, the method has been used across a variety of domains to examine a variety of real-world networks. In a thorough review summarizing recent research, Newman (2003)[2] divides large, **sparse**, real-world complex networks into four categories: social, information, technological, and biological. Some of the research has focused on words or other entities that carry semantic meaning. This article summarizes this collection of recent research involving graph theoretical depictions of various aspects of human language, such as networks derived from corpora and thesauri, and it also includes some studies involving biological neuronal networks. The results of these studies provide a backdrop for understanding the potential uses of the method in the broader informatics world.

Throughout the text, the words *network* and *graph* are used interchangeably. Words appearing in bold are defined in the glossary (Appendix 1.)

### *Knowledge representation is of central importance in biomedical informatics*

While this is a topic with potential applications in a variety of domains, the structure of semantic networks is of particular interest in biomedicine. In biomedicine, many human-curated semantic networks, such as controlled terminologies and ontologies, are used for organizing and communicating information. At the core of these vocabularies are discrete elements of knowledge, or entities, which carry meaning. The way in which these entities are arranged and encoded in electronic format is referred to as knowledge representation.

Knowledge representation is a central concern in biomedical informatics. Many of the core theories and methods of the field, ranging from bioinformatics databases to expert systems to disease surveillance approaches, depend on discrete representation of knowledge in a form that can be processed computationally. The output of any decision support tool, like the results of a given study, can only be interpreted in consideration of how information was modeled at the start of the process. In other words, the outputs depend upon the fundamental atomic units that constitute the inputs and how these units interrelate. As informatics continues to mature as a discipline, it is increasingly important to examine the knowledge representation approaches employed within various theories, methods, and systems.

### *An emerging knowledge representation approach: graph theoretic modeling*

Recently, graph theoretic modeling of information, propelled by decades of research in social network analysis, has become increasingly useful. Specifically, recent years have seen an increasing interest in the study of large, sparse, complex networks, in which graph-theoretic approaches are used to model the relationships between the entities in real-world systems. This body of research has prompted significant advances in the theory describing the form and function of complex networks.

The term “semantic network” has been used to refer to a family of knowledge representation techniques since the 1960's[3]. Classical semantic networks often represent defined relationships between entities, and the **topological structure** is typically defined by the designer. The networks in this review can be distinguished from earlier semantic networks in several ways: First, they are based on recent research (1998 or later); second, they are created from real-world data, and third, they are much larger and far more complex. The large-scale complexity of these networks could be considered surprising, since the networks are conceptually simple (generally having nodes of the same type and unweighted edges.) For example, in word association networks, a human subject is shown a particular word and is asked to name a related word[4-7]. An edge is assigned between two words if they are associated in this way. Networks can also be created by assigning an edge between two words if they co-occur in a large corpus[8-10]. The complexity of large semantic networks arises from a diversity of global and local features, which in turn emerge from the arrangements of links between the entities.

### *Artificial semantic networks have been modeled using graph theory*

Several existing semantic reference systems are amenable to graph theoretic modeling, since they include formalized lists of entities along with the connections between them. General purpose networks of this type include Roget's Thesaurus and the WordNet lexicon, a curated lexical reference system. In a network made from Roget's Thesaurus[11] two words were joined if one of the words was listed in the thesaurus entry of the other. The WordNet lexicon was

modeled as a network[12] in which the nodes were words, and an edge joined two words if they shared a given characteristic (hypernymy, antonymy, meronymy, or polysemy).

Among the most familiar semantic networks in biomedicine are artificial networks such as ontologies and other controlled vocabularies. These human-curated semantic networks are domain-specific; though useful in the domain for which they were developed, they may have little or no applicability in other domains. For example, a heart condition such as angina can have a significant impact on a person's daily activities. However, heart-related terminology in ICD[13] (the International Classification of Diseases), which was developed mainly for the purpose of representing patient diagnostic data for billing and reimbursement, is of limited use for encoding a patient's functional status information[14]. As a result of the domain-specific nature of such semantic networks, a variety of formal controlled vocabularies has been developed in parallel by various groups. Each of these vocabularies serves a different purpose and has its own global structure. There have been efforts to merge and unify a number of these vocabularies. The UMLS Metathesaurus[15] is the best known of these efforts.

### *Graph theory facilitates connection-oriented models*

For decades, rectangular data tables have been a dominant knowledge representation paradigm. In these tables, each record or row represents some entity and columns contain data that describe attributes of the entity. The rectangular shape of data tables imposes a restriction on the data they contain: Each entity in a given table is required to have an identical set of attributes, although the values of these attributes differ from one entity to the next. For example, the records in a data table might represent patients, and each patient's record includes the patient's first name, last name, date of birth, and other information. Data in this form are compatible with algorithmic processing and conventional statistical analysis.

Ontology has also become a dominant knowledge representation approach in biomedicine. An ontology is a formal model of the concepts in a given domain. Entities are assigned properties and relations between the entities are defined explicitly. When encoded using ontological principles, biomedical information can be used in a variety of computer applications that rely on discretely coded information for the execution of logical operations. The results of these logical operations are consistent and reliable.

Graph theoretic modeling is fundamentally different from rectangular database-oriented and ontological modeling approaches. Its focus is on entities and the relations between them, and it can be used to describe the global **topology**, or the community structure, of a system. A relational database is effective for managing and analyzing information about a set of entities of the same type (for example, patients' blood sugar levels taken over a period of time), but it does not include a convenient way to specify arbitrary connections between entities. For example, in a given population, suppose that a link were assigned among any two patients with similar diagnostic histories. In a data table, these links could be represented in a list. However, the graph theoretic model takes this list one step further by representing the global patterns occurring among this set of patients. Representation of these topological features, in turn, invites new analytical approaches. For example, one could examine whether the model conveys tightly clustered groups of patients, and if it does, whether these patients have something in common.

### *A connection-oriented model has its advantages*

The modeling fundamentals of **graph theory** are conceptually simple. The simplest type of model involves a set of nodes and the connections between them. With its built-in facility in representing connections between entities, graph theoretic modeling supplements relational database models and ontological approaches: A given system is viewed as a network rather than

as a collection of isolated entities, and the local and global topological structure can readily be examined. The models can be more complex; for example, edges can have weights, and nodes and edges can be of different types. But at its most basic level, the graph theoretic modeling paradigm is in fact more general than ontological modeling. It is concerned simply with how concepts relate to one another.

To be clear, it should be noted that databases and ontologies can also be represented as graphs. However, graphs made from databases and ontologies are constrained by an external set of rules, i.e., a schema or a set of axioms, which control how entities are connected. The graphs covered in this review are more flexible in that they are not subject to any such external control. As a result, the properties of the systems modeled reflect the systems themselves rather than the imposed organizational schema of the person who designed them.

### *Goals of this article*

This review article is a survey of recent research on large semantic networks. It is mainly directed towards people involved with the various domains of biomedical informatics, including medical informatics, bioinformatics, and public health informatics. However, given the interdisciplinary nature of the topic, the material may be of interest to individuals in other fields, such as artificial intelligence, library science, linguistics, and mathematics. The article concludes with a number of ways in which the techniques can potentially be applied, such as for biomedical vocabulary development or in electronic health records. It is hoped that a clearer understanding of the topological features of natural and artificial semantic networks will provide insight into the development of useful information systems in biomedicine.

### **Background**

For those who are less familiar with graph theoretic modeling, this section of the article covers the elementary principles of the method. Readers who have knowledge of graph theoretic modeling may wish to skip this section.

### *Graph theoretic modeling of networks*

Three recent review articles include concise and readable summaries of recent research on large, sparse, complex networks[2, 16, 17]. This interdisciplinary area has its roots in a number of diverse fields. Sociology and anthropology, given their longstanding focus on network analysis, have been particularly influential, as has discrete mathematics, with its history of investigations into graph theory[17]. A number of specific factors[16] have contributed to the explosive growth of this area, including increases in the availability of data and of computer power, interdisciplinary collaboration, and an increasing interest in moving beyond reductionist approaches to understand the behavior of entire systems.

### *Basic graph theory concepts*

Several articles summarize the elementary graph theoretic principles important in the study of complex networks[6, 18, 19]. Additional details can be found in the glossary in Appendix 1. To summarize, in graph-theoretic modeling, a **graph** is comprised of a set of **nodes**, also referred to as **vertices**, along with a set of either **edges** or **arcs** which connect pairs of nodes[6]. Edges are **undirected** connections, while arcs are directed. Although variations such as weighted edges are possible, simpler models are often favored because they are compatible with a wide array of algorithms and statistical measures.

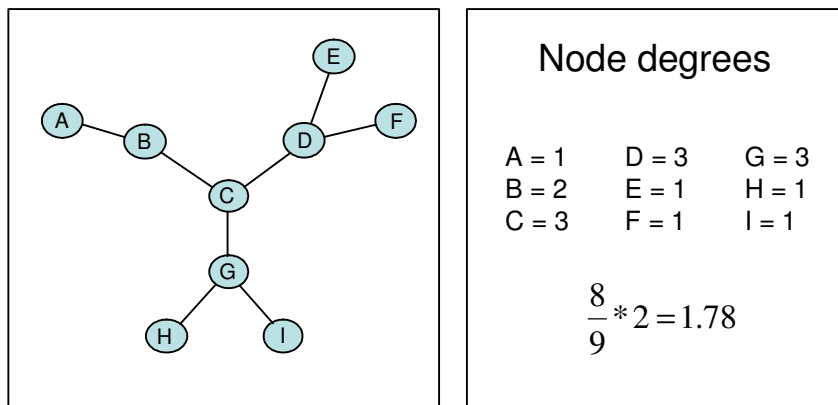
A graph containing only undirected edges is called an **undirected graph**; a graph with only arcs is a **directed graph**. Each directed graph can be converted to an undirected graph if

arcs are converted to edges by removing their **directionality**. A graph is considered **connected** if there is at least one **path** between any two nodes. The number of nodes to which a given node is immediately connected is its **degree**. A node and all of its adjacent nodes constitute a **neighborhood**.

*Signature measures of graph topology*

When nodes and edges are arranged into a large graph, what often emerges is a complex community structure. A single graph can have a variety of distinctive features, including highly clustered neighborhoods, treelike properties, islands, and **highly-connected hubs**. All of these emergent properties are part of the network’s topology. The topology of a connected, unweighted, sparse graph can be characterized using a variety of techniques and measures. A small handful of signature statistical properties have gained favor recently. These are the **average node degree**, the **average path length**, and the **clustering coefficient**. The average node degree, a measure of the density of a graph, is the average number of edges per node. It is calculated by dividing the number of edges by the number of nodes, and then multiplying by two (Figure 1.)

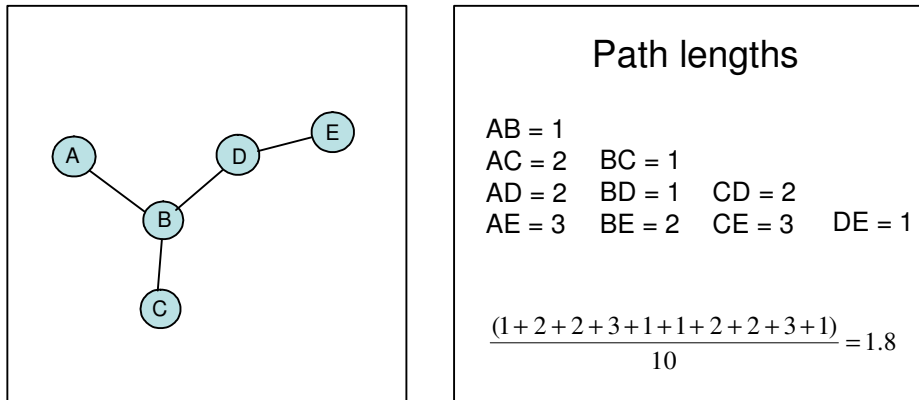
Figure 1. Average node degree



*The degree of Node B is 2 and the degree of Node C is 3. The average node degree for the entire graph is 1.78.*

The average path length, sometimes called the “average shortest path”, refers to the average **distance** between any two nodes. A simple algorithm determines the minimum distance between any node and any other node. An average is then calculated based on all of these values (Figure 2.)

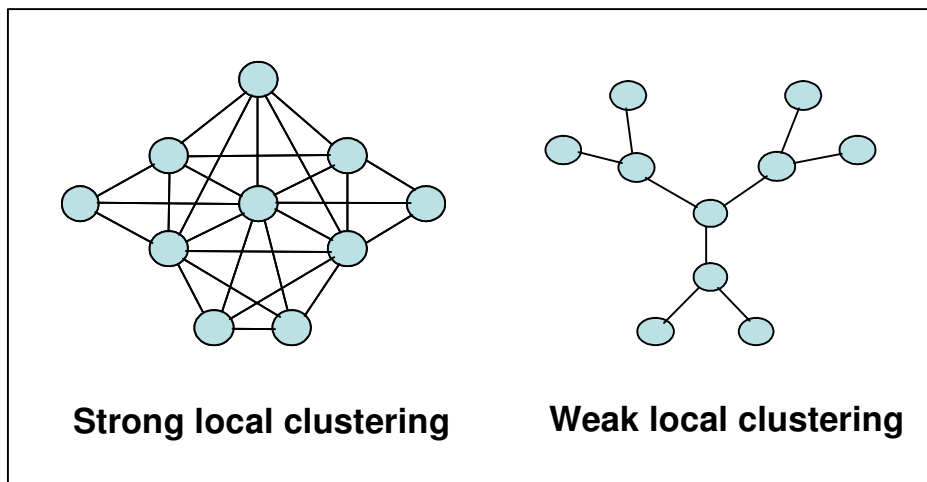
Figure 2. Average path length



The distance between nodes A and E is 3. The average path length for the entire graph is 1.8

There are several ways to calculate the amount of clustering in a graph at the local level[2]. A common approach is to calculate the clustering coefficient for a given node by counting the number of edges between the node's **neighbors**, and then dividing by all their possible edges. This results in a value between 0 and 1, which is then averaged over all nodes in a graph[19]. In Figure 3, the graph on the left, which exhibits strong clustering at the local level, has a high clustering coefficient; the graph on the right has a low clustering coefficient. Another way to think about the clustering coefficient is as the extent to which the neighborhoods of two neighboring nodes overlap[6]. In a **fully connected graph**, or a graph in which each node is connected to every other node, the clustering coefficient is 1.

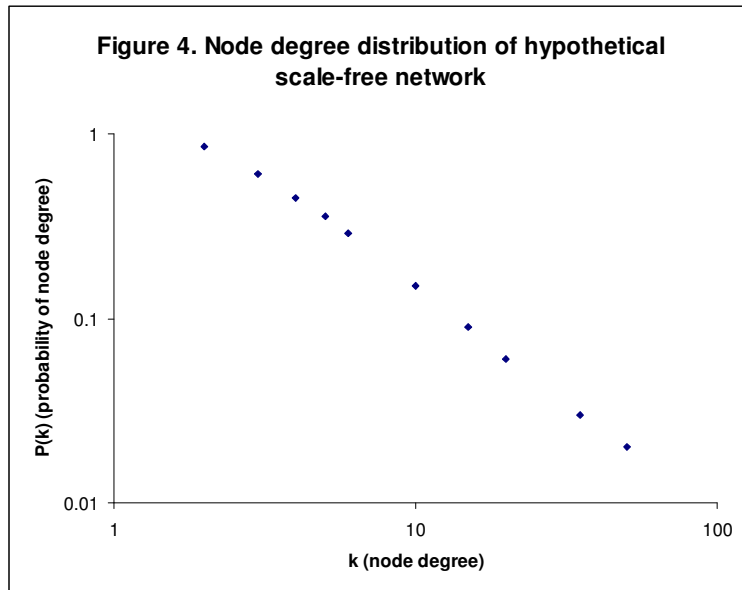
Figure 3. Components of hypothetical graphs with strong and weak local clustering



*Many real-world complex networks share common topological properties*

Although different graphs have varying topologies at the local level[20], networks constructed from real-world data often share common global properties. First, they are typically sparse: The vast majority of nodes are connected only to a small percentage of other nodes[6],

and the number of edges is closer to the number of nodes, than to the square of the number of nodes[21]. Second, they have a short average path length and **strong local clustering**: The neighbors of a given node are more likely to be connected to one another than would be expected through chance alone. **Random graphs**, by contrast, have a short average path length but a low clustering coefficient. Third, the distribution in node degree is characterized by a **power law**[6]. A **power law distribution** is a statistical distribution in which one variable is proportional to a power of the other[22]. When plotted on a log/log scale, individual points are distributed about a straight line (Figure 4). This means that there are a small number of nodes (the “hubs”) which have many neighbors and a large number of nodes that have only a few neighbors.



The second and third properties are of special note, because they are the signatures of *small-world* and *scale-free* characteristics, respectively. Several articles[6, 11] offer concise descriptions of these features. In graphs with small-world properties, there are highly clustered neighborhoods (see Figure 3) and it is possible to move from one node to another (see Figure 2) in a relatively small number of steps (often just two or three, on average.) As a small-world graph grows, the average path length increases slowly, as a function of the logarithm of the size of the graph[23]. This is in opposition to the longer path lengths of a regular **grid**-like network in which hundreds or thousands of steps may be required.

Scale-free networks have no characteristic scale of node degree; instead, they exhibit all scales of connectivity simultaneously[24]. As a result, they tend to have a similar appearance when examined at varying scales. They are also generally robust against random disruptions; if any node is removed at random, the statistical likelihood is low that the node will be a hub. It is far more likely instead that the node selected for removal will have a small number of neighbors. Therefore, a random change is not likely to have a significant effect on the network’s overall structural form. By contrast, if hubs are selected for removal, the topology of the network will change much more significantly, since the hubs are connected to a large number of neighbors. The scale-free property appears in many networks based on real data, and most scale-free networks also have small-world characteristics[11].

## **Methods**

To gather a collection of articles for this review, a search strategy was developed. The final search strategy evolved through an iterative process.

First, three seminal review articles[2] were identified. These articles were then used to identify related citations: articles cited within these articles, and articles which later cited the seminal articles. The indexing keywords appearing in these articles were then compiled into a list.

Next, a set of literature databases was identified. Because of the interdisciplinary nature of this research, it was not possible to confine the search to any one particular database. The authors consulted with seven professional reference librarians at six different libraries at our university. These librarians suggested more than 10 online databases possibly relevant to the topic. Each of these databases was then examined to assess the relevance of its content to the topic of large-scale semantic networks.

The list of author-assigned keywords was then used to search several of the suggested databases, including the Web of Science[25] and MEDLINE[26], to identify additional articles and their associated keywords. Most of these keywords fit naturally into one of two categories - they related either to graph theoretic approaches, or to semantic information. The terms were combined into two lists of search terms. The terms in each list were evaluated to determine their relevance and identifying power. Terms with low precision, i.e., terms that retrieved many irrelevant documents, were eliminated. The resulting search strategy (Table 1) was then executed to retrieve all articles that satisfied the selected criteria.

Table 1. Search strategy for retrieving documents pertaining to large semantic networks

**Network-related terms**

associat\* network\*  
 average shortest path\*  
 barabasi-albert  
 biological neural network  
 complex network\*  
 evolving network\*  
 growing network\*  
 interconnection network\*  
 local cluster\*  
 neuron\* network  
 path length  
 preferential attachment  
 real graph\*  
 real network\*  
 real-world network\*  
 scale-free network\*  
 scientific network\*  
 small world

**Semantics-related terms**

biomedical terminolog\*  
 biomedical vocabular\*  
 computer program\*  
 conceptual network\*  
 controlled terminolog\*  
 controlled vocabular\*  
 co-occurrence  
 cooccurrence  
 document collection\*  
 human language\*  
 informational cascade\*  
 lexical network\*  
 interaction of words  
 natural language\*  
 local knowledge  
 neuroanatomy  
 semantic network\*  
 semantic search  
 semantic structure\*  
 semantic web  
 SNOMED  
 UMLS  
 verb lexicon  
 verb semantics  
 word-adjacency  
 word association\*  
 word interaction\*  
 Wordnet

*The search strategy was to find all articles pertaining to at least one of the terms from each column. Five research databases were searched on July 25, 2005: MedLINE and PsycINFO; Compendex and Inspec; and Web of Science. The search and was limited to the years 1998-2005. An asterisk (\*) indicates that a wildcard character was employed.*

To classify these articles systematically, a coding form was developed. This form consisted of a series of data elements (Table 2). Each article was reviewed in turn, and the relevant information in each article was recorded on the coding form.

Table 2. Selected coding form data elements for tabulating data on articles pertaining to large semantic networks

<b>Data element</b>	<b>Description</b>	<b>Data type</b>
Reference	Complete bibliographic reference	Text
Pertains to semantic networks	Does the article pertain to networks of symbols that carry meaning?	Yes or no
Pertains to graph theory networks	Does the article pertain to networks based on graph theory?	Yes or no
Involves application of graph theory	Does the article describe research involving an application of graph theory?	Yes or no
Pertains to real-world data	In the article, is data from the real world used in the construction of a network?	Yes or no
Pertains to electronic networks	Does the article pertain to a network contained in electronic format, such as a thesaurus? (May include networks based on data collected by the authors.)	Yes or no
Pertains to brain networks	Does the article pertain to neuronal or corticocortical networks in the brain's physical structure? (Excludes artificial neural networks).	Yes or no
Pertains to associative networks	Does the article pertain to a conceptual network in a person's mind, such as a network of interrelated words?	Yes or no
Description	Description of network investigated	Text
Small world property	Did the authors report evidence of the small-world property?	Yes or no
Scale-free property	Did the authors report evidence of scale-free characteristics?	Yes or no
Average node degree	What was the average node degree, if reported?	Text
Path length	What was the average path length, if reported?	Text
Clustering coefficient	What was the clustering coefficient, if reported?	Text

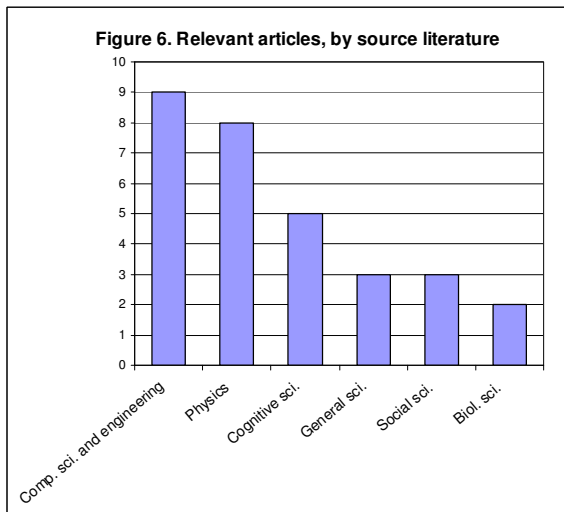
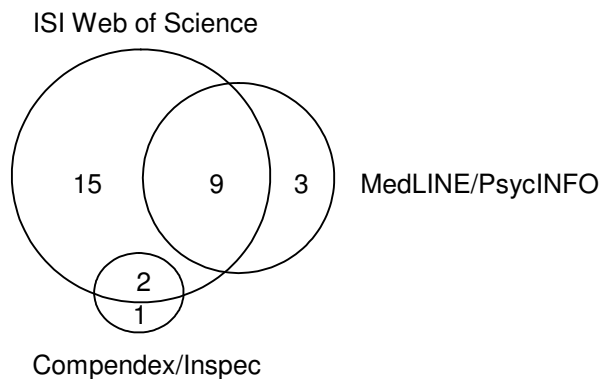
The articles were coded in two rounds. In the first round, all retrieved articles were coded, but only the first two data fields were entered. This allowed for the execution of a database query to identify the articles relevant to the topic. All articles which involved a graph theoretic approach to large networks, and which also pertained to semantic information, were retrieved, and were fully coded in the second round of coding.

## Results

### Summary of results

The search strategy retrieved a total of 116 documents. Of the 116 articles retrieved, 30 involved both *large-scale graph theoretic modeling* and *symbols that carry meaning* and were therefore relevant for this review. Twenty-six were identified through Web of Science, three through the combined search of Compendex and Inspec, and 12 through the combined search of MEDLINE and PsycInfo. Figure 5 shows the overlap between the databases in which the relevant articles were found. Figure 6 reflects the interdisciplinary nature of the topic; the articles appeared in a range of journals, books, conferences, and Ph.D. dissertations in cognitive sciences, physics, computer science and engineering, as well as general, biological, and social sciences.

**Figure 5. Databases in which articles appeared**



Additional articles were identified in other ways. For example, if a citation appearing in a given article appeared likely to be relevant, we obtained a copy of the article and assessed it for relevance using the same criteria. Ten documents were identified in this way or in similar ways, and these were also added to the document set, bringing the total to 40. Of these 40 articles, nine were review articles, resulting in a total of 31 articles upon which the remaining calculations are based.

Twenty-eight of these 31 articles (90.3%) involved an investigation of a real-world network. The sources of data for the networks investigated were varied and diverse, and included corpora[8, 10, 27-29], thesauri of words with similar meanings[6, 11, 12], dictionaries[21, 30], large computer programs[31], and files available on the Internet[9, 32]. In addition, twelve of the 28 (42.9%) pertained to networks in the human mind or brain. Seven of these pertained to conceptual networks, such as association networks of interrelated words[4, 6, 7, 21, 33-35]. The other five[18, 20, 36-38] pertained to the neuronal or corticocortical networks that constitute a part of the brain's physical structure.

Studies of words were common. Among the 28 articles involving a real-world network, 22 (78.6%) involved an experimental graph or graphs in which the nodes were comprised of words or phrases.

As for the topological properties common to many real-world networks, 15 of the 28 articles (53.6%) specifically mentioned finding evidence of the small-world characteristic in the network investigated. Eleven of the 28 (39.3%) reported evidence of scale-free properties. Five articles[9, 18, 36, 38, 39] identified the small-world phenomenon but did not mention scale-free properties in the network identified. One[30] reported scale-free properties but did not mention the small-world phenomenon.

As for the commonly-calculated measures of network topology, nine of the 28 articles (32.1%) measured the average node degree. Twelve (42.9%) reported average path length. Ten (35.7%) of the articles measured the clustering coefficient. Seven of the articles (25%) reported values for all three of these statistics. It is not possible to make direct comparisons of the values for these measures, for at least two reasons: First, the networks model different phenomena and were constructed in a variety of ways. Second, the values of the measures depend on factors such as the size of the network.

### *Topological features of networks derived from natural language*

General characteristics exhibited by large language-derived networks were sparsity, short average path lengths, a high degree of local clustering, and a power-law distribution in degrees of nodes. Graphs made from language have been shown to exhibit both scale-free[11, 40] and small-world[11, 29, 40] topological features. These features are not consistent with the features of arbitrarily structured networks and other conventional models of semantic organization often based on inheritance hierarchies[6]. These properties may be universal to all large semantic networks derived from language. In one recent paper, three semantic networks were examined (a word association database, WordNet, and Roget's thesaurus,) and all three exhibited small-world and scale-free properties[6]. Another study in which both of these properties were identified[11] involved a network composed of English language words with similar meanings. In this network, two words were assigned an edge if one word appeared in a thesaurus entry of the other. Small-world properties have also been found in Czech, Romanian, German[40], and Chinese[9], which suggests that this feature may be language-independent.

Small-world and power law features were also identified in a study of data of free-

association of ideas by human subjects[4]. In this study, a number of additional trends and patterns were also identified. One finding was that the number of new words input by the user diminished gradually, reaching an equilibrium state at which few new words were likely to be added. Another was what the authors termed asymmetry in edges. The presence of a directed edge, or arc, from one word to the next did not necessarily imply that there was an arc going the other direction. A third was context biasing: The words chosen by the research subjects tended to be influenced not only by the previous word, but by other recently-used words. These findings are distinct from many of the others reported in this review, since they were identified by examining the characteristics of a network of concepts from free-association by research subjects, rather than words used in language communication.

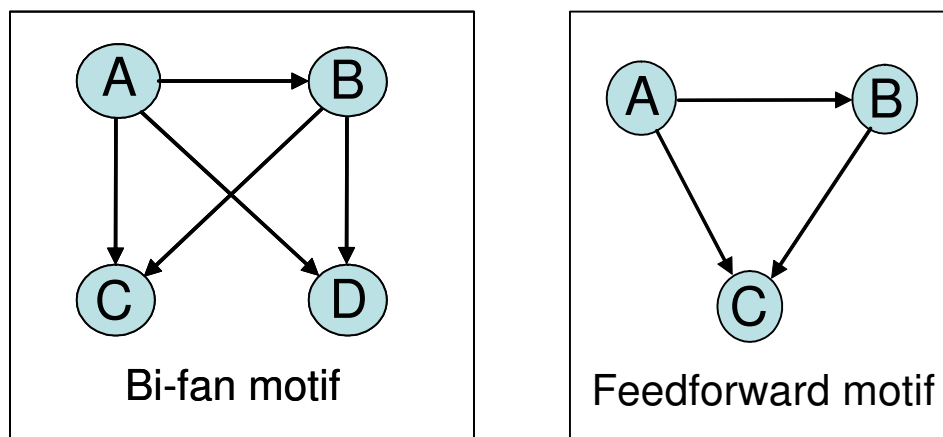
In a recent study of the global structure of a network derived from language[30], the level of clustering was not equivalent for all nodes, and appeared to be a function of the degree of a given node. While small nodes belonged to highly cohesive, densely interlinked clusters, hubs did not; their neighbors had a smaller chance of linking to each other. These results suggest that language has an organizational structure that repeats itself at various scales. At the local level, there are many small, densely interconnected clusters. These combine to form larger, less interconnected groups, and these groups again combine to form larger and even less cohesive groups. An analysis of networks made from Czech, Romanian, and German also identified a well-defined hierarchical organization[40] in each.

The dissertation of Old (2004)[39] demonstrated that the *implicit* (conceptual "hidden inner structure") of Roget's thesaurus, which is elicited using statistical approaches such as frequency counts and word connectivity patterns, differs from the *explicit* organizational structure (a hierarchy of concepts and sets of synonyms). Research on the thesaurus also identified a semantic core of highly-connected words related to agitation, motion, and survival. By contrast, the largest categories in the thesaurus related to concepts such as food, animals, clothing, and technology[39].

#### *Local structure of language networks*

Real-world networks from different fields share several global features, including small-world and scale-free properties. However, they may have significantly different local structure[37]. One way to analyze the local structure of graphs is to examine **network motifs**, which are described as simple building blocks of complex networks[20]. Certain local patterns of interconnections occur in real-world networks significantly more often than in randomized networks, and the most common patterns vary depending on the type of network. Networks involved in information processing give rise to significantly different types of motifs than networks of energy flow. Two motifs in particular (**feedforward loop** and **bi-fan**) (Figure 7), are common both to transcriptional gene regulation networks and to the neuronal connectivity network of the nematode *Caenorhabditis elegans*[20]. It has been surmised the feedforward loop may play a functional role in information processing[41].

**Figure 7. Feedforward loop and bi-fan motifs [20]**



The possible ways in which three nodes can share directed edges can be referred to as **triads**[37]. Networks from a particular field tend to have a similar distribution in the occurrence of selected triads (**triad significance profile**, or TSP.) Human languages share common topological properties at the local level: When texts of different sizes, and from different languages, were compared, they were found to have similar TSPs. This means that when a particular triad is found in one human language, it will also tend to be found in a high concentration in other human languages[37]. However, networks from other fields, such as biological networks, have different profiles. One distinctive feature of the local structure of human languages is that certain specific triangle-shaped triads are underrepresented in language; they occur far more frequently in other real-world systems. This may result from the way words are used in language. Words belong to categories, and a word from one category tends to be associated with a word from another category[42]. Words belonging to three different parts of speech, such as nouns, verbs, and adjectives, might preferentially be arranged into particular motifs.

Palla *et al* [43] have introduced a set of characteristic quantities to describe the statistics of communities in networks. These measures allow a node to participate in more than one community at a time. For example, the word *gold* can belong simultaneously to communities related to Olympic medals, metals, jewels, and prosperity. The topological structure of communities is shown to be scale-free for nodes with degrees above a given threshold; however, there is an exponential node degree distribution for words below the threshold. They present several illustrative graphs that show how a given word may be a member of multiple highly cohesive groups of nodes.

#### *Balance in real-world networks*

Networks of words are generally not considered as dynamic systems in which information flows from one point to another. However, the importance of balance in the structure of large semantic networks can perhaps best be understood in light of research on complex systems. Complex systems thrive at the border between order and chaos, and the number of connections is what maintains this balance [44]. With too few edges, a network may collapse. A large communication network may still function after some edges are removed, because

information can flow around the affected area. However, when additional edges are removed, and the demand on the network remains the same, the network is no longer able to accommodate all network traffic. On the other hand, while a system tends to gravitate towards maximum connectivity[44], a network with too many edges becomes unstable[45]. This is because when there is a high level of interconnectedness, a small change in a system can result in a cascade of evolutionary change[44]. The scale-free topology also helps maintain balance, as it helps control the rate of change and establishes order. Most nodes receive input from only a few other nodes, and change is limited to the local neighborhood[44].

#### *Balance in large semantic networks*

This delicate balance between order and chaos seems to apply to language as well. Language must be expressive and flexible, allowing for the expression of concepts in a virtually infinite cognitive space, but it must also be learnable. If there are not enough connections between concepts, the expressiveness of language is sacrificed. If there are too many connections, it will be more difficult for speakers to learn new words. Large language-derived networks show that the number of connections between words is at a point that maximizes its expressiveness without sacrificing its learnability[28].) Like the network motifs that describe the local structure of language, this tendency towards balance also appears to be language-independent[28].

#### *Applications of graph-theoretic modeling of large semantic networks*

Semantic networks have been used successfully for a number of practical applications, including automatic word sense disambiguation[32] and in formulating responses to natural language search queries[46]. Bordag *et al* (2003)[10] present a graph-theoretic approach to lexical disambiguation. In this approach, co-occurrence is used as a proxy for semantic similarity, allowing for the construction of graphs of related words. Possible applications of this theoretical framework include improvements in text classification methods, word sense disambiguation algorithms, and spell checking tools.

#### *Language acquisition and change*

The apparent scale-free and small-world characteristics of language may also have implications for psycholinguistic research on language acquisition[10]. The recent characterization of the topology of language-derived graphs has led to new ways of describing language evolution. Based on the data obtained by Ramon Ferrer i Cancho *et al*[29], language has been described as an evolving word web[47]. In this model, language is considered to be an evolving network of interacting words. The distribution of words has been shown to fall into two distinct regimes[48]. The first is referred to as the kernel lexicon, which varies slowly over time as a language changes. The second is a peripheral lexicon, used for more specific communication.

#### *Aspects of semantic networks that are not natural languages*

To complement this array of studies of large networks modeled after natural language, a handful of research has also examined the properties of other types of networks of semantic information. Large computer programs, for example, can be modeled as growing networks of related files; these files are analogous to natural language documents, rather than to words. Like natural language networks, these networks have been shown to exhibit small-world and scale-

free properties[31]. One of the implications of the scale-free property is that the information flow within the system is expected to be efficient; despite the size of the system, the average path length in the underlying network is small. Given the evidence for the existence of the small-world property, one practical implication is that in debugging and refining code, the highly-connected files could be checked first.

#### *Organization of cortical networks in the brain*

This article is not intended to address the structure and function of complex brain networks; this topic is well-covered in a recent review[19]. However, this area was addressed in the articles retrieved for this review, and the main ideas merit at least a brief mention.

In a review of research on the organization of long-range corticocortical connectivity in mammalian brains, Hilgetag and Kaiser (2004)[49] report that cortical projections are arranged in small-world networks. They form tight clusters which are highly interlinked with each other, but less frequently with other clusters. This distributed cluster structure achieves functional integration, while also allowing for different cortical areas to have individual specializations; it may therefore be an ideal design for achieving a high level of functional complexity[19, 49]. Small-world characteristics have been found across multiple scales of cortical organization; they have been found not only in corticocortical connection matrices, but also in large-scale cortical connection matrices[19]. The ubiquity of this topological feature points to its importance in the function of brains[50]. This architecture may relate to the need for global and local efficiency, in which local necessities can be addressed while facilitating wide-scope interactions[18].

#### *Organization of semantic information in the brain*

Despite the mounting research data describing the organization of complex brain networks, little is known about how the brain stores semantic information. However, it has been shown that persons with Alzheimer's disease experience a graceful degradation in their understanding of relations between concepts[51]. As the disease progresses, the organization of their semantic knowledge becomes increasingly abnormal. Scale-free networks degrade in a similar way; as individual nodes are removed, the integrity of the network remains intact. Removal of certain well-connected nodes can result in a cascade effect with more dramatic consequences; however, the destructive impact of such events still remains localized. Therefore, this research lends support to the hypothesis that concepts in the brain may also have a **scale-invariant** arrangement.

The small-world property of language may have arisen from the need for speed during the production of speech[29]. If speech can indeed be modeled as a path from word to word in the brain, it follows that an efficient organization of the words would improve the speed of speech production. Another important aspect of speech production is richness. The small-world nature of language suggests a cognitive model in which a speaker will normally choose words from commonly used words, but in which rare (yet perhaps more expressive) words are just a few degrees away[29].

## **Discussion**

#### *Graph theoretic modeling is flexible and has been used in many fields*

The wide range of source publications of the articles discussed in this review is a testament to the multidisciplinary nature of graph theoretic modeling. Modeling techniques for

large graphs, along with statistical measures to study them, have been shown to be useful in social network analysis, in studying the Internet, and in other fields. Given the increasing interest in similar modeling techniques in computational biology, graph theoretic modeling is also gaining importance in biomedical informatics.

*Graph modeling is intuitive for humans and is also machine compatible*

Network modeling is a flexible knowledge representation technique anchored in graph theory. The properties of graph theory make network models appealing for human conceptual understanding and also compatible with computational processing. For humans, the connectionist paradigm expressed in graphs is conceptually simple. Unlike other knowledge representation techniques, such as rectangular databases, graph models are composed of just two different building blocks. The nodes and edges of graphs are grounded in real data, and graph models can be built, drawn, and analyzed automatically using computer software. Curation of a semantic network can also be automatic: As new information becomes available, it can be incorporated into the network.

*Network modeling reveals the global and local structure of language*

When languages are represented using graph theory, the resulting network models reveal how individual semantic entities (often words) are related to one another. Historically, random graphs have been used to model real-world systems, under the apparent assumption that the networks underlying them are fundamentally random[16]. Given recent dramatic increases in the speed and ubiquity of computing resources, it is now possible to study complex networks in a way that assumes nothing in advance about the system's organization. Empirical, data-driven complex network models display actual relationships between a system's many entities.

Word networks, which have a distinctive structure at the global and local level, provide insights into how language serves as a framework for representing and communicating information. At the global or community level, many semantic network models exhibit the scale-free and small-world architectures common to many real-world phenomena[11, 29]. At the local level, the profiles of motifs [37] have been shown to be similar across several natural languages. Although the values of statistics used to measure these features differ from language to language, various authors have hypothesized that these global features are shared across all human languages.

*Graph modeling can aid in the development and maintenance of controlled vocabularies*

In biomedicine, rigidly arranged controlled vocabularies, often hierarchical, are one of the cornerstones of knowledge representation. Most controlled vocabularies contain data about the relationships between entities. The way these relationships are assigned is an important consideration in vocabulary development, because the structure of a vocabulary carries implications for its usability and adaptability. In a collaborative vocabulary development effort with multiple participants, developers could use a network modeling tool early in the development process to depict a vocabulary's proposed global and local structure. This structure could serve as a precursor for the development of a logical model or ontology, which could then be developed iteratively.

After the structure is formalized, graph theoretic modeling and visualization can also be useful for the maintenance of a vocabulary. Developers could use network modeling software to observe the vocabulary throughout the various phases of its evolution, and the network could be

examined periodically to detect errors and anomalies. For example, discovery of a sparse area in an otherwise dense network may prompt developers to consider adding additional concepts or links.

### *Topological features can influence operations on controlled vocabularies*

Cognitive science research on associative networks offers an analogous framework to illustrate how knowledge of topology can be used for controlled vocabulary maintenance. Just as word recognition, learning, and speech production are key cognitive operations on a semantic associative network, important operations in controlled vocabulary development and maintenance are finding, adding, editing, and deleting concepts. Features of network topology can help or hinder a user's ability to perform these operations.

Navigation through a semantic network to find a given term can be considered analogous to the task of finding a desired word during speech production. When navigating through a controlled vocabulary, a scale-free architecture can help a user identify a starting place to look for a concept. Highly connected hubs, small in number, can serve effectively as landmarks. As a user delves deeper into the network, there are smaller hubs at every scale which can also be used for orientation. A small-world architecture in a vocabulary implies that there are enough cross-links for efficient navigation between groups of concepts. Both of these properties, which are common in networks derived from natural language, could convey benefits if incorporated into the architecture of artificial networks.

The process of adding new terms can also be helped or hindered by various topological features. A scale-free architecture might facilitate the addition of terms because a growing network has a built-in way of limiting the semantic space to which new concepts are likely to be added. In **growing network models**[6], a network is constructed beginning with an initial node or set of nodes, and additional nodes are added sequentially. A process that commonly governs network growth in real-world networks is known as preferential attachment. New nodes attach to existing nodes according to a rule that can be summarized as "the rich get richer." Nodes that are already highly connected have a far higher probability of gaining new connections. If knowledge of these probabilities is used in vocabulary development, the space of concepts to which to join a new concept can be reduced to a limited number of likely choices. A small-world architecture could also confer an advantage for a vocabulary developer adding new concepts. If related words are already arranged into highly-connected clusters, it is easier to determine the concepts to which the new concepts should be linked.

The importance of topology can be illustrated further by considering the effect of architecture on the addition of new concepts. Suppose a given vocabulary is arranged into a **tree** in which the most important concepts are towards the top of the tree and only hierarchical relationships are permitted. If a new important concept is added near the top, it is impossible to assign links to other concepts more than one level away. Since more steps are now required to move between concepts, there is an increase in the number of operations when the vocabulary is used.

Operations associated with vocabulary editing and reorganization can also be influenced by topology. For example, examination of topological features can help a developer determine whether a given operation will adversely affect the network's structure. If a hub concept, or a concept in the neighborhood of a hub, is deleted, this will have a far greater effect on the network topology than if an isolated node is deleted. Likewise, if any one of the siblings of a fully-connected clique is moved to an entirely new part of the network, but its existing links to other

concepts are left in place, the topology may change significantly; there will be a large number of new links pointing to the concept's former neighborhood.

In preliminary research, we are comparing the global topologies of 16 biomedical vocabularies. The goals are to determine what can be learned by applying the most popular measures of network topology; to assess the extent to which artificial vocabularies share the topological properties common to natural semantic networks; and to determine whether it is possible to group vocabularies by topological structure. Initial results indicate that some controlled vocabularies share the scale-free and small-world topological features of networks made from natural language, and that network modeling can be used to visualize the global and local topologies of networks, leading to descriptions of the networks that are much more difficult to realize using conventional approaches.

#### *Graph modeling can convey a high-level summary of a given topic*

Network modeling offers some potential for use as a summarization tool for use across the biomedical informatics continuum, from the molecular level to public health. One way to define a given topic is to list the topics to which it is related. Semantic network models are sometimes built from co-occurrence information in which, for example, an edge is assigned between two words if they co-occur in a given sentence.

Suppose a given topic were represented as a large network derived from a corpus of text or speech on that topic. Individual words could then be represented as the portion of the semantic network pertaining to the word. A word in a given context might be given one representation, while that same word in a different context could adopt an altogether different representation. When represented as a network, language can serve as a window into knowledge, revealing how people organize information.

Another advantage of network modeling is that because real-world semantic networks are scale-free, they can be represented at varying levels of detail. Depending on how the network is constructed, it is often the case that the most important concepts are by nature the hubs of the network. Hubs are communicated even if the network is expressed at a low level of detail. Therefore, the broad structural configuration of the network is still preserved.

#### *Text in a patient's health record can be represented as a network*

In medicine, network modeling might be a useful adjunct to an electronic health record. Suppose a patient with a chronic kidney condition has a complex medical history that stretches back several decades. Imagine that this patient's record includes enough paper documents to fill an entire filing cabinet. When converted to electronic format, the records comprise many megabytes of data. If this patient's information were presented using network modeling techniques, it might reveal summary-level information that is not evident using conventional techniques. For example, since the patient has a history of a kidney disease, the word "renal" and related words would appear more frequently, and would therefore be more likely to be hubs in the graph model. Redundant information, such as multiple photocopies of a single discharge summary, would be absorbed into a single section of the graph. Graph visualization software with an intuitive interface could then be used to help new caregivers gain a basic understanding of the types of conditions that have affected the patient in the past.

#### *Areas of focus for future research*

Steyvers *et al* (2005)[6] have also proposed a number of areas for future research on large

semantic networks. They propose various types of inquiries involving linguistic constraints, in which words or connections are first categorized into semantic or syntactic classes, and statistical analyses are performed separately for each class. Another possible research direction would be to perform more subtle analyses in which qualitative or quantitative differences between connections are modeled. They also propose to compare the topologies of semantic networks of different languages and to develop search and retrieval algorithms that make use of the large-scale structure of semantic networks. Such work could eventually help describe the context sensitivity of meanings or the exact relationship between word meanings and concepts[6]. Other authors have proposed similar ways to use knowledge of the structure of semantic networks for practical applications, including lexical disambiguation, automatic summarization, spell checking, and document categorization.

Finally, characterizing the properties of large semantic networks may also have implications for the rapidly developing field of neuroinformatics[52]. Although it is tempting to draw parallels between representations of large semantic networks and the organization of semantic information in the brain, the relationship between the two is unclear. Another intriguing parallel is between the general processes that govern the way semantic networks function in computers, and human performance in semantic processing tasks[6]. Both of these questions are beginning to fall within the realm of scientific inquiry, given a confluence of theories and methods in information science, imaging, and neuroscience.

## **Conclusion**

Graph theoretic modeling of large networks has been influential in many areas of science, including sociology, physics, and computer science. Due in part to their prevalence in computational biology and bioinformatics, network modeling approaches are slowly becoming more recognized among informatics researchers.

Semantic networks have a key role in knowledge representation in health care and biomedicine. This article is meant to serve as a synthesis of research in this emerging field. The results confirm that large semantic networks derived from natural language share topological properties common to many real-world phenomena, including scale-free and small-world characteristics. A clearer understanding of the topological features of natural and artificial semantic networks will provide insight into the development of useful information systems in health care and biomedicine.

*Appendix 1: Glossary of key concepts from graph theory*

**arc:** In graph theory, a directed connection from one node to another. An arc, typically represented using an arrow, is distinguished from an edge, which is an undirected connection.

**average node degree:** The average node degree, a measure of the density of a graph, is the average number of edges per node. It is calculated by dividing the number of edges by the number of nodes, and then multiplying by two.

**average path length:** The average path length, also called the “average shortest path”, is the average distance between any two nodes in a graph.

**bi-fan motif:** A network motif involving four nodes in which there is an arc from one node to each of two adjacent nodes, as well as an arc from a second node to the same two adjacent nodes[20].

**bi-fan:** See *bi-fan motif*.

**clustering coefficient:** The probability that two neighbors of a randomly chosen node will themselves be neighbors, or alternatively, the extent to which the neighborhoods of neighboring nodes overlap[6]. The clustering coefficient is commonly calculated as the number of connections between a node’s neighbors divided by all their possible connections. It ranges between 0 and 1 and is typically averaged over all nodes of a graph, yielding the clustering coefficient value for the entire graph[10].

**connected:** A graph is considered connected if there is at least one path from each node to all other nodes.

**degree:** The number of edges connected to a node[2].

**diameter:** The diameter of a network is the length (in number of edges) of the longest path between any two nodes[2].

**directed graph:** A graph in which all connections between nodes are arcs rather than edges.

**directionality:** The direction of orientation of an arc connecting two nodes.

**distance:** The length of the shortest path between two nodes in a graph.

**edge:** The line connecting two nodes. Also called a bond (physics), a link (computer science), or a tie (sociology)[2].

**feedforward loop:** See *feedforward loop motif*.

**feedforward loop motif:** A network motif involving three nodes in which there is an arc from the first node to the second, from the second to the third, and from the first to the third[20].

**fully connected graph:** A graph in which each node is connected to every other node.

**graph theoretic:** See *graph theory*.

**graph theory:** The study of the properties of graphs.

**graph:** A set of nodes connected by either edges, which are undirected, or arcs, which are directed. A graph can be represented visually using dots to represent nodes, lines to represent edges, and arrows to represent arcs.

**grid:** A network in which nodes and edges are arranged into a repeating pattern of squares or cubes.

**growing network model:** A model in which a network is constructed beginning with an initial node or set of nodes, and additional nodes are added sequentially.

**hub:** A node with a disproportionately high number of connections to other nodes. In scale-free networks, hubs may have node degrees several orders of magnitude higher than the degrees of other nodes.

**motif:** See *network motifs*.

**neighbor:** In a network, two nodes that are connected by an edge.

**neighborhood:** A subset of nodes in a network consisting of a node and all of its neighbors[6].

**network motifs:** Patterns of interconnections occurring at the local level of a large network [20]. See also *bi-fan motif*; *feedforward loop motif*.

**network:** See *graph*.

**node:** An entity in a graph; may be connected to other nodes by either arcs or edges.

**path length:** The number of edges or arcs along a given path from one node to another. See also *average path length*.

**path:** A sequence of edges that connects one node to another[6].

**power law:** See *power law distribution*.

**power law distribution:** A statistical distribution in which the value of one variable is proportional to a power of the other[22].

**random graph:** A graph in which links between nodes are arranged randomly.

**scale-free:** Property of a graph that has no characteristic scale of node degree and instead exhibits all scales of connectivity simultaneously[24].

**scale-invariant:** See *scale-free*.

**semantic network:** An interconnected set of entities that carry meaning.

**small-world network:** A set of interconnected entities characterized by highly clustered neighborhoods and short average path lengths between the entities. In a small-world network, it is possible to move from one node to another in a relatively small number of steps.

**small-world:** See *small-world network*.

**social network:** A model describing a collection of people and the connections between them. In a social network model a connection is assigned between two people if they are connected in some way, such as through friendship or by way of a business relationship.

**sparse:** A characteristic of networks in which the vast majority of nodes are connected only to a small percentage of other nodes[6], and the number of edges is closer to the number of nodes, than to the square of the number of nodes[21].

**strong local clustering:** A characteristic of networks in which the neighbors of a given node are likely to be connected to one another more than would be expected through chance alone. Strong local clustering results in densely connected neighborhoods, one of the hallmark properties of small-world networks.

**topological structure:** See *topology*.

**topology:** The global configuration resulting from the arrangement of nodes in a graph and the connections between them.

**tree:** A model in which entities are arranged into hierarchies of parents and children.

**triad significance profile:** Distribution in the occurrences of various triad motifs in a network[37]. The triad significance profile is one way to characterize the local-level properties of a large network.

**triad:** A network motif consisting of three nodes and the edges connecting them.

**undirected graph:** A graph in which all connections between nodes are edges, which are undirected, rather than arcs, which are directed.

**undirected:** In a network model, a link between two entities is undirected if there is no inherent directionality in the relationship. In graph theory, an undirected link is represented using an edge.

**vertex (pl. vertices):** The fundamental unit of a network. In computer science, vertices are typically referred to as nodes, and in sociology, actors[2]. See also *node*.

**vertices:** See *vertex*.

## References

- [1] Anderson JG. Evaluation in health informatics: social network analysis. *Computers in Biology & Medicine* 2002;32(3):179-93.
- [2] Newman MEJ. The structure and function of complex networks. *Siam Rev* 2003;45(2):167-256.
- [3] Collins AM, Quillian MR. Retrieval time from semantic memory. *J Verbal Learn Verbal Behav* 1969;8:240-247.
- [4] Costa LD. What's in a name? *Int J Mod Phys C* 2004;15(3):371-379.
- [5] Pomi A, Mizraji E. Semantic graphs and associative memories. *Phys Rev E* 2004;70(6).
- [6] Steyvers M, Tenenbaum JB. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Sci* 2005;29(1):41-78.
- [7] Penz E, Meier-Pesti K, Kirchler E. "It's practical, but no more controllable": Social representations of the electronic purse in Austria. *J Econ Psychol* 2004;25(6):771-787.
- [8] Aizawa A, Kageura K. Calculating association between technical terms based on co-occurrences in keyword lists of academic papers. *Syst Comput Japan* 2003;34(3):85-95.
- [9] Zhu M, Cai Z, Cai Q. Automatic keywords extraction of Chinese document using small world structure. In: 2003 International Conference on Natural Language Processing and Knowledge Engineering; 2003 Oct. 26-29, 2003; Beijing, China: IEEE; 2003. p. 438-43.
- [10] Bordag S. Sentence co-occurrences as small-world graphs: A solution to automatic lexical disambiguation. In: Gelbukh A, editor. *Computational Linguistics and Intelligent Text Processing (Proceedings, Second Conference on Intelligent Text Processing and Computational Linguistics)*; 2003; Mexico City: Springer-Verlag; 2003. p. 329-332.
- [11] Motter AE, de Moura APS, Lai YC, Dasgupta P. Topology of the conceptual network of language. *Phys Rev E* 2002;65(6).
- [12] Sigman M, Cecchi GA. Global organization of the Wordnet lexicon. *P Natl Acad Sci USA* 2002;99(3):1742-1747.
- [13] National Center for Health Statistics (US). *International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)*. 2004 [cited 2004 Aug 5]; Available from: <http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm>
- [14] Bales ME, Kukafka R, Burkhardt A, Friedman C. Qualitative assessment of the International Classification of Functioning, Disability, and Health with respect to the desiderata for controlled medical vocabularies. *Int J Med Inf* (accepted) 2005.
- [15] McCray AT, Nelson SJ. The representation of meaning in the UMLS. *Method Inform Med* 1995;34(1-2):193-201.
- [16] Albert R, Barabasi AL. Statistical mechanics of complex networks. *Rev Mod Phys* 2002;74(1):47-97.
- [17] Watts DJ. The "new" science of networks. *Annu Rev Sociol* 2004;30:243-270.
- [18] Latora V, Marchiori M. Efficient behavior of small-world networks. *Phys Rev Lett* 2001;87:19(19).
- [19] Sporns O, Chialvo DR, Kaiser M, Hilgetag CC. Organization, development and function of complex brain networks. *Trends Cogn Sci* 2004;8(9):418-425.
- [20] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: Simple building blocks of complex networks. *Science* 2002;298(5594):824-827.
- [21] Gaume B, Duvignau K, Gasquet O, Gineste MD. Forms of meaning, meaning of forms. *J Exp Theor Artif In* 2002;14(1):61-74.

- [22] College of Geosciences, University of Oklahoma. An abbreviated glossary of system terminology. 2005 [cited 2005 Sep 21]; Available from: [http://www.esse.ou.edu/glossary\\_st.html](http://www.esse.ou.edu/glossary_st.html)
- [23] Gomez-Gardenes J, Moreno Y. Local versus global knowledge in the Barabasi-Albert scale-free network model. *Phys Rev E* 2004;69(3).
- [24] Barabasi AL, Albert R. Emergence of scaling in random networks. *Science* 1999;286(5439):509-512.
- [25] Institute for Scientific Information. Web of Science. Thompson Scientific, Philadelphia, PA, USA, 2005.
- [26] National Library of Medicine (US). MEDLINE. National Institutes of Health, Bethesda, MD, USA, 2005.
- [27] de Campos LM, Fernandez-Luna JM, Huete JF. The BNR model: foundations and performance of a Bayesian network-based retrieval model. *Int J Approx Reason* 2003;34(2-3):265-285.
- [28] Allegrini P, Grigolini P, Palatella L. Intermittency and scale-free networks: a dynamical model for human language complexity. *Chaos Soliton Fract* 2004;20(1):95-105.
- [29] Ferrer i Cancho R, Sole RV. The small world of human language. *P Roy Soc Lond B Bio* 2001;268(1482):2261-2265.
- [30] Ravasz E, Barabasi AL. Hierarchical organization in complex networks. *Phys Rev E* 2003;67(2).
- [31] de Moura APS, Lai YC, Motter AE. Signatures of small-world and scale-free properties in large computer programs. *Phys Rev E* 2003;68(1).
- [32] Veronis J. HyperLex: lexical cartography for information retrieval. *Comput Speech Lang* 2004;18(3):223-252.
- [33] Costa LD. The hierarchical backbone of complex networks. *Phys Rev Lett* 2004;93(9).
- [34] Diefenbach GJ. The role of trait and state anxiety in semantic network organization of information related to current concerns: The Louisiana State University And Agricultural And Mechanical College, US; 2000.
- [35] Capocci A, Servedio VDP, Caldarelli G, Colaiori F. Detecting communities in large networks. *Physica A-Statistical Mechanics And Its Applications* 2005;352(2-4):669-676.
- [36] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature* 1998;393(6684):440-442.
- [37] Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, et al. Superfamilies of evolved and designed networks. *Science* 2004;303(5663):1538-1542.
- [38] Stephan KE, Hilgetag CC, Burns G, O'Neill MA, Young MP, Kotter R. Computational analysis of functional connectivity between areas of primate cerebral cortex. *Philos Trans R Soc London, Ser B* 2000;355(1393):111-126.
- [39] Old LJ. The semantic structure of Roget's, a whole-language thesaurus: Indiana University, US; 2004.
- [40] Ferrer i Cancho R, Sole RV, Kohler R. Patterns in syntactic dependency networks. *Phys Rev E* 2004;69(5).
- [41] Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* 2002;31(1):64-68.
- [42] Allen J. *Natural Language Understanding*. Redwood City, CA: The Benjamin/Cummings Publishing Company, Inc.; 1995.

- [43] Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 2005;435(7043):814-818.
- [44] Thompson G. Is all the world a complex network? *Econ Soc* 2004;33(3):411-424.
- [45] Simon HA. *The Sciences of the Artificial*. 3rd ed. Cambridge, MA: The MIT Press; 1996.
- [46] Berger H, Dittenbach M, Merkl D. Activation on the move: Querying tourism information via spreading activation. In: Marik V, Retschitzegger W, Stepankova O, editors. *Proceedings of the 14th International Conference and Workshop on Database and Expert Systems Applications (DEXA '03)*; 2003 Sept. 1-5, 2003; Prague, Czech Republic: Springer-Verlag; 2003. p. 474-483.
- [47] Dorogovtsev SN, Mendes JFF. Language as an evolving word web. *P Roy Soc Lond B Bio* 2001;268(1485):2603-2606.
- [48] Ferrer i Cancho R. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's Law revisited. *J Quant Linguist* 2001;8(3):165-173.
- [49] Hilgetag CC, Kaiser M. Clustered organization of cortical connectivity. *Neuroinformatics* 2004;2(3):353-360.
- [50] Sporns O, Zwi JD. The small world of the cerebral cortex. *Neuroinformatics* 2004;2(2):145-162.
- [51] Chan AS, Salmon DP, Butters N. Semantic network abnormalities in patients with Alzheimer's disease. In: Parks R, Levine D, Long D, editors. *Fundamentals of neural network modeling: Neuropsychology and cognitive neuroscience*. Cambridge, MA: The MIT Press; 1998. p. 381-393.
- [52] Koslow SH, Subramaniam S. *Databasing the Brain: From Data to Knowledge (Neuroinformatics)*. Hoboken, NJ.: John Wiley & Sons, Inc.; 2005.