Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Biomedical Informatics 39 (2006) 451–464

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Methodological Review

Graph theoretic modeling of large-scale semantic networks

Michael E. Bales *, Stephen B. Johnson

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

Received 14 April 2005
Available online 15 December 2005

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.

Keywords: Small world network; Scale-free network; Graph theory; Language; Linguistics; Semantics; Terminology; Review literature

1. 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 [2] divides large, **sparse**, real-world complex networks into four categories: social, information,

[*] Corresponding author. Fax: +1 208 694 4181.
E-mail address: michael.bales@dbmi.columbia.edu (M.E. Bales).

1532-0464/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2005.10.007

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).

1.1. 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.

1.2. 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 1960s [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.

1.3. 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, antonomy, 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.