Graphs in Libraries: A Primer

James E. Powell, Daniel Alcazar, Matthew Hopkins, Robert Olendorf, Tamara M. McMahon, Amber Wu, and Linn Collins

Whenever librarians use Semantic Web services and standards for representing data, they also generate graphs, whether they intend to or not. Graphs are a new data model for libraries and librarians, and they present new opportunities for library services. In this paper we introduce graph theory and explore its real and potential applications in the context of digital libraries. Part 1 describes basic concepts in graph theory and how graph theory has been applied by information retrieval systems such as Google. Part 2 discusses practical applications of graph theory in digital library environments. Some of the applications have been prototyped at the Los Alamos National Laboratory Research Library, others have been described in peer-reviewed journals, and still others are speculative in nature. The paper is intended to serve as a high-level tutorial to graphs in libraries.

Part 1. Introduction to Graph Theory

Complexity surrounds us, and in the twenty-first century, our attempts at organization and structure sometimes lead to more complexity. In layman's terms, complexity refers to problems and objects that have many distinct but interrelated issues or components. There also is an interdisciplinary field referred to as "complex systems," which investigates emergent properties, such as collective intelligence.[1] Emergent properties are an embodiment of the old adage "the whole is greater than the sum of its parts." These are behaviors or characteristics of a system "where the parts don't give a real sense of the whole."[2] Libraries reside at the nexus of these two definitions: they are creators and caretakers of complex data sets (metadata), and they are the source of explicit records of the complex and evolving intellectual and social relationships underlying the evolution of knowledge.

Digital libraries are complex systems. Patrons visit libraries hoping to find some order in complexity or to discover a path to new knowledge. Instead, they become the integration point for a complex set of systems as they juggle resource discovery by interacting with multiple systems, either overtly or via federated search, and by contending with multiple vendor sites to retrieve articles of interest.

Contrast this with Google's simple approach to content discovery: a user enters a few terms in a single box, and Google returns a large list of results spanning the Internet, placing the most relevant results at the top of this list. No one would suggest using Google for all research needs, but its simplicity and recognized ability to answer routine searches is compelling. How, we wonder, can we bring a bit of Google to the library world?

Google harvests vast quantities of data from the web. This data aggregation is obviously complex. How does Google make sense of it all so that it can offer searchers the most relevant results? Answering this question requires understanding what Google is doing, which requires a working knowledge of graph theory. We can then apply these lessons to library systems, make sense of voluminous bibliometric data, and give researchers tools that are as effective for them as Google is for web surfers. Just as web surfers want to know which sites are most relevant, researchers want to know which of the relevant results are the most reliable, the most influential, and of the highest quality. Can quantitative metrics help answer these qualitative questions?

The more deeply libraries and librarians can mine relationships between articles and authors and between subjects and institutions, the more reliable are their metrics. Suppose some librarians want to compare the relative influence of two authors. They might first look at the authors' respective number of publications. But are those papers of equally high quality? They might next count all citations to those papers. But are the citing articles of high quality? Deeper still, they might assign different weights to each citing article using its own number of citations. At each step, whether realizing it or not, they are applying graph theory. With deeper knowledge of this subject, librarians can embrace complexity and harness it for research tools of powerful simplicity.

PageRank and the Global Giant Graph

Indexing the web is a massive challenge. The Internet is a network of computer hardware resources so complex that no one really knows exactly how it is structured. In fact, researchers have resorted to conducting experiments to discern the structure and size of the Internet and its potential vulnerability to attacks. Representations of the data collected by these experiments are based on network

James E. Powell (jepowell@lanl.gov) is Research Technologist, **Daniel A. Alcazar** (dalcazar@lanl.gov) is Professional Librarian, **Matthew Hopkins** (mfhop@lanl.gov) is Library Professional, **Tamara M. McMahon** (tmcmahon@lanl.gov) is Library Technology Professional, **Amber Wu** (amber.ponichtera@gmail.com) is Graduate Research Assistant, and **Linn Collins** (linn@lanl.gov) is Technical Project Manager, Los Alamos National Laboratory, Los Alamos, New Mexico. **Robert Olendorf** (olendorf@unm.edu) is Data Librarian for Science and Engineering, University of New Mexico Libraries, Albuquerque, New Mexico.

science, also known as graph theory. This is not the same network that ties all the computers on the Internet together, though at first glance it is a similar idea. Network science is a technique for representing the relationships between components of a complex system.[3] It uses graphs, which consist of *nodes* and *edges*, to represent these sets of relationships.

Generally speaking, a *node* is an actor or object of some sort, and an *edge* is a relationship or property. In the case of the web, universal resource locators (URLs) can be thought of as nodes, and connections between pages can be thought of as links or edges. This may sound familiar because the Semantic Web is largely built around the idea of graphs, where each pair of nodes with a connecting edge is referred to as a triple. In fact, Tim Berners-Lee refers to the Semantic Web as the Global Giant Graph—a place where statements of facts about things are published online and distinctly addressable, just as webpages are today.[4]

The Semantic Web differs from the traditional web in its use of ontologies that place meaning on the links and in the expectation that nodes are represented by universal resource identifiers (URIs) or by literal (string, integer, etc.) values, as shown in figure 1, where the links in a web graph have meaning in the Semantic Web.

Semantic Web data are a form of graph, so graph analysis techniques can be applied to semantic graphs, just as they are applied to representations of other complex systems, such as social networks, cellular metabolic networks, and ecological food webs. Herein lies the secret behind Google's success: Google builds a graph representation of the data it collects. These graphs play a large role in determining what users see in response to any given query.

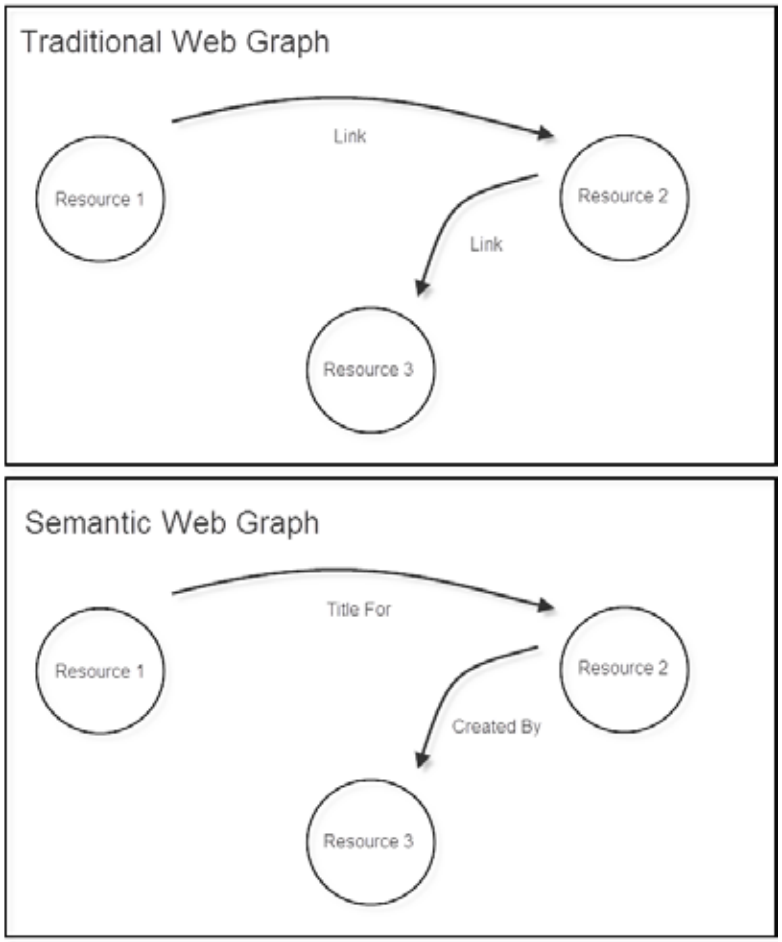

Figure 1. A traditional Web graph is compared to a corresponding Semantic Web graph. Notice that replacing traditional Web links with semantic links facilitates a deeper understanding of how the resources are related.

Google uses a graph analysis technique called Eigenvector centrality.[5] In essence, Google calculates the relative importance of a given webpage as a function of the importance of the pages that point to it. A simpler centrality measure is called degree centrality. Degree centrality is simply a count of the number of edges a given node has. In a social network, degree centrality might tell you how many friends a given person has. If a person has edges representing friendship that connect him to seventeen other nodes, representing other people in the network, then his degree value is seventeen (see figure 2). If a person with seventeen friends has more friendship edges than any other person in the network, then he has the highest degree centrality for that network.

Eigenvector centrality expands on degree centrality. Consider a social network that represents the amount of influence a person has in a business context. If we want to analyze this aspect of the network, then it makes sense to consider the fact that some relationships are more influential than others. For example, a relationship with the president of the company is more significant than a relationship with a coworker, since it is a safe assumption that a direct relationship with the company leader will increase influence. So we assign weights to the edges based on who the edge connects to.

Google does something similar. All the webpages they track have centrality values, but Google's weighting algorithm takes into account the relative importance of the pages that connect to a given resource. The weighting algorithm bases importance on the number of links pointing to a page, not the page's internal content, which makes it difficult for website authors to manipulate the system and climb the results ladder. So if a given webpage

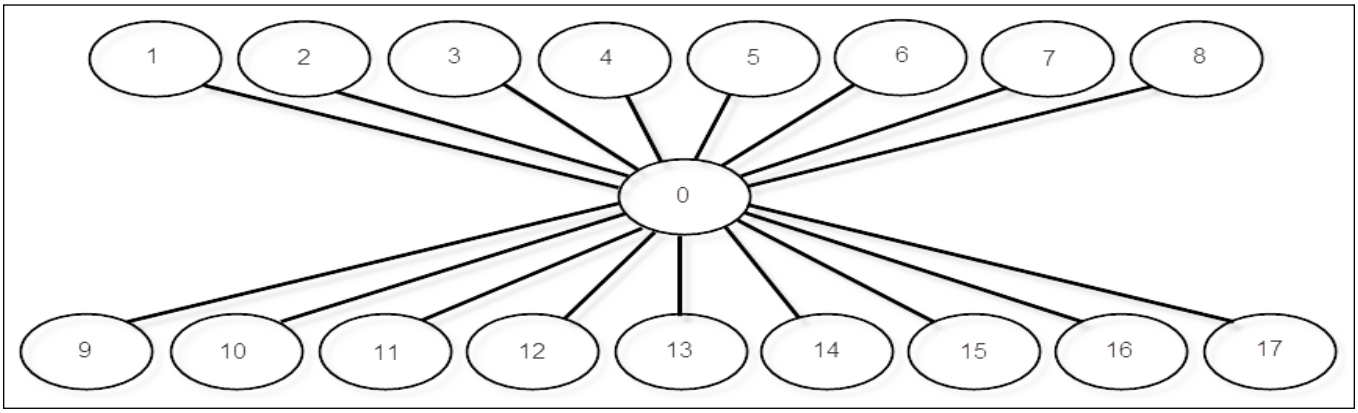

Figure 2. Friendship network

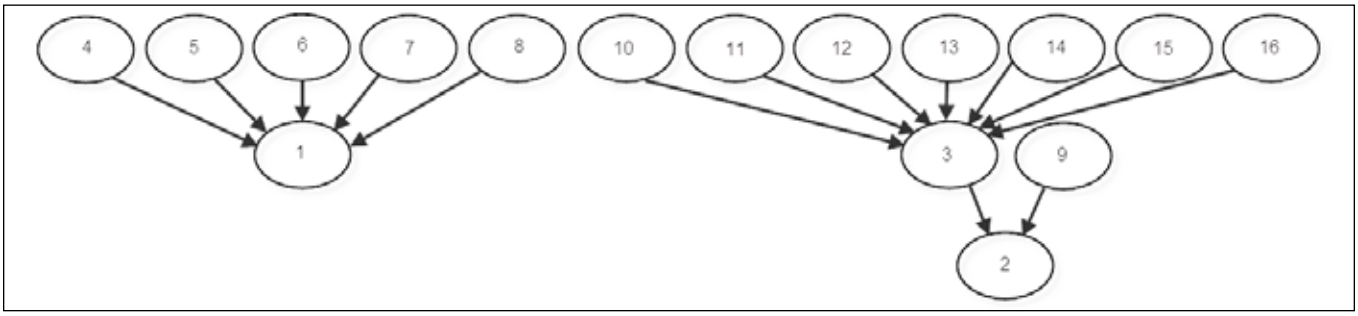

Figure 3. Node 2 ranks higher than node 1 because node 3, which connects to node 2, has more incoming links than node 1. Node 3 is deemed more important than node 9, which has no incoming links.

has only two edges, it may still rank higher than a more connected page if one of the pages that links to it has a large number of pages pointing to it (see figure 3).

This weighted degree centrality measure is Eigenvector centrality, and a higher Eigenvector centrality score causes a page to show up closer to the top of a Google results set. The user never sees a graph, but this graph-based approach to exploring a complex system (the web), works quite well for routine web searches.

Graph Theory

Graph theory, also known as network science, has evolved tremendously in the last decade. For example, information scientists have discovered hubs in the web that connect large numbers of pages, and if removed, disconnect large portions of the network.[6] Biologists have begun to explore cellular processes, such as metabolism, by modeling these processes as networks and have even found in these networks evidence for the evolution of metabolic processes.[7] Chemists have used networks to model reactions in a step-wise fashion by "editing" graphs representing models of molecules and their reactivity,[8] and they also have used graphs to better comprehend phase transition states, such as the freezing of water or the emergence of superconductivity when a material is cooled.[9] Economists have used graphs to model market trades and the effects of globalization.[10] Infectious disease specialists have used networks to model the spread of disease and to evaluate prospective vaccination plans.[11] Sociologists have modeled the complex interactions of people in communities.[12] And in libraries, computer scientists have explored citation networks and coauthorship networks,[13] and they have developed maps of science that integrate scientific papers, their topics, the journals in which they appear, and comsumers' usage patterns to provide a new view of the pursuit of science.[14]

Network science can make complexity more comprehensible by representing a subset of actors and relationships in a complex system as a graph. These

graphs can then be explored visually and mathematically. Graphs can be used to represent systems as they are, to extract subsets of these systems, or to discover wholly artificial collections of relationships between components of a speculative system. Data also can be represented as graphs when they consist of "measurements that are either of or from a system conceptualized as a network."[15] In short, graphs offer a continuum of techniques for comprehending complexity and are suitable either for a layman with casual interest in a topic or a serious researcher ferreting out discrete details.

At the core of network science is the graph. As stated earlier, a graph is a collection of nodes and the edges that connect some of those nodes, together representing a set of actors and relationships in a type of system. Relationships can be unidirectional (e.g., in a social network, when the information flows from one person to another) or bidirectional (e.g., when the information flows back and forth between two individuals). Relationships also can vary in significance and can be assigned a weight—for example, a person's relationship to his or her supervisor might be weighted more heavily than a person's relationship to his or her peers. A graph can consist of a single type of node (for subjects) and a single type of edge connecting those nodes (for predicates). These are called *unipartite* graphs. From the standpoint of graph theory, these are the easiest types of graphs to work with. Graphs that represent two relationships (*bipartite*) or more are typically reduced to unipartite graphs in the process of exploring them because the vast majority of techniques for evaluating graphs were developed for graphs that address a single relationship between a set of nodes.

Global Properties of Graphs

There are other aspects of graphs to consider, sometimes referred to as "global graph properties."[16] There are two basic classes of networks: homogeneous networks and inhomogeneous networks.[17] These graphs exhibit characteristics that may not be comprehensible by close examination (e.g., by examining degree centrality, node clustering, or paths within a graph)[18] but may be apparent, depending on the size and the way in which the graph is rendered, merely by looking at a visualization of the graph. In homogeneous graphs, nodes have no significant difference between their number of connections. Examples include *random graphs*, *complete graphs*, and *small world networks*. In *random graphs* there is an equal probability that any two nodes will be connected (see figure 4), while in *complete graphs* (see figure 5) all nodes are connected with one another. Random graphs are often used as tools to evaluate networks that describe real systems. Complete graphs might occur in social networks as subgraphs, e.g., in the case where a person has two friends who are also mutual friends.

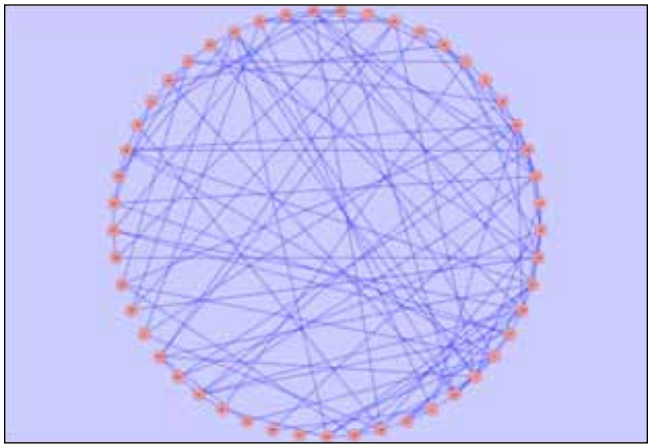

Figure 4. A Random Graph. Any given node has an equal probability of being linked to any other node

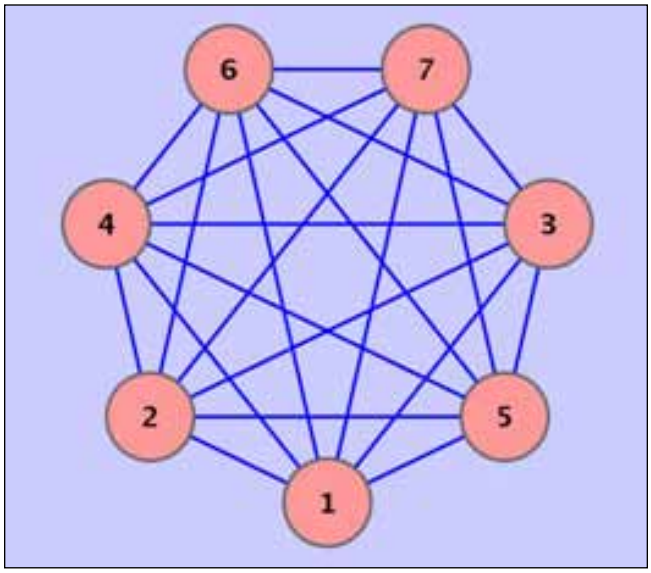

Figure 5. A Complete Graph. All nodes are connected to all other nodes

Small world networks have numerous highly interconnected subgroups called clusters. These may be distributed throughout the network in a regular fashion, with a few random connections that connect the otherwise disconnected clusters. These random links have the effect of greatly reducing the path length between any two nodes and explain the oft-cited six degrees of separation that connect all people to one another. In social networks, agency is often described as the mechanism by which

these random links get established. Agency refers to the idea that multiple, often unpredictable actions on the part of individuals in a network result in unanticipated connections between people. Examples of such actions are hobbies, past work experience, meeting someone new while on a trip to another country—pretty much anything that takes a person outside his or her normal social circles.

In the case of inhomogeneous graphs, not all nodes are created equal. One type, *scale-free networks*, is common in a variety of systems ranging from biological to technological (see figure 6. These exhibit a structure in which a few nodes play a central role in connecting many others. These hubs form as a result of preferential attachment, known colloquially as "the rich get richer." Researchers became aware of scale-free networks as a result of analysis of the web when it was in its infancy. Scale-free networks have been documented in biology, social networks, and technological networks. As a result, they are quite important in the field of information science. Small world and scale-free networks are typical of complex systems that occur in nature or evolve because of emergent dynamic processes, in which a system self-organizes over time. Small world networks provide fast, reliable communication between nodes, while scale-free networks are more fault tolerant, making them ideal for systems such as living cells, which are frequently challenged by the external environment.[19]

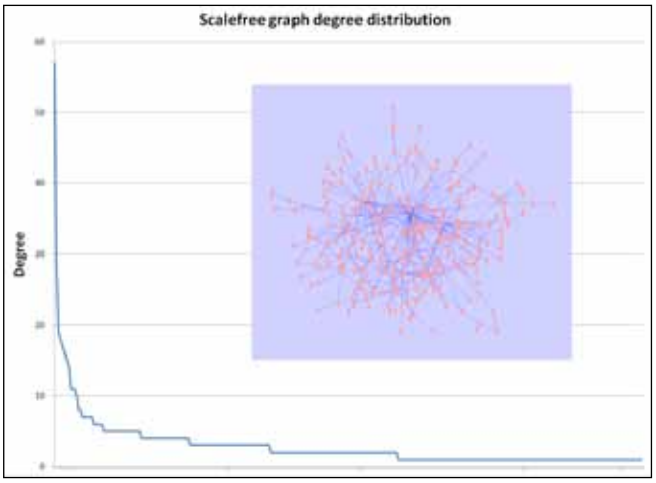

Figure 6. Example of a Scale-Free Coauthorship Network. A few nodes have many links, and most nodes have few or a single link to another node

Local Properties of Graphs

Below the ten-thousand-foot system-level view of networks, graphs can be scrutinized more closely using many other techniques. We will now consider four broad categories of local characteristics that describe networks and how they are, or could be, applied in digital libraries: node centrality measures, paths between nodes, motifs, and clustering.

Centrality measures make it possible to determine the importance of a given node in a network. Degree centrality, in its simplest form, is simply a count of the number of edges connected to any given node in a network: a node with high-degree centrality has many connections to other nodes compared to a typical node in the graph.

Paths make it possible to explore the connections between nodes. An author who is two degrees removed from another author—in other words, the friend of a friend of a friend—has a path length of 2. Researchers are often interested specifically in the shortest path between a given pair of nodes. Many other types of paths can be explored depending on the type of network, but in libraries, paths that describe the flow of ideas or communication between people are most likely to be useful.

Motifs are the fundamental recurring structures that make up the larger graph, and they often are called the building blocks of networks.[20] A three-node feedback motif is a set of nodes where the edges between them form a triangle and the edges are directional. In other words, node A is connected to (and might convey some information to) node B; node B, in turn, has the same relationship with node C; and node C is connected to (and conveys information back to) node A. In digital libraries, for example, if similar papers exhibit the same pattern of connectivity to a group of subject or keyword categories, motifs will make it possible to readily identify the topical overlap between them.

Collections of nodes that have a high degree of connectivity with each other are called *clusters*.[21] In many complex systems, clusters are formed by preferential attachment. A group of highly clustered nodes that have low connectivity to the larger graph is known as a *clique*.

While there are other aspects of graphs that can be explored, these four—node centrality measures, paths between nodes, motifs, and clustering—are accessible to most users and are significant in graphs representing bibliographic metadata and textual content. This will become clearer in the examples that follow.

Quantitative Evaluation of Graphs

Returning now to centrality measures, two of particular interest in digital libraries are *degree centrality* and *betweenness centrality* (or flow centrality). An interesting aspect of graphs is that, regardless of the data being represented, centrality measures and clustering characteristics often reveal important clues about the system that the data

describes, whether it's coauthorship relationships or protein interactions in the cell of a living organism. Often the clusters or nodes that exhibit a higher score in some centrality calculation are significant in some meaningful way compared to other nodes.

Recall that *degree centrality* refers to how many edges a given node has. Degree centrality can vary significantly in strength depending on the relationships that are represented in the graph. Consider a graph of citations between papers. While it may be obvious to humans that the mostly highly cited papers will have the highest-degree centrality, computers have no idea what this means. It is still up to humans to lend a degree of comprehensibility to the raw calculation: in other words, to understand that a paper with high-degree centrality is an important paper, at least among the papers the graph represents.

Betweenness centrality exposes how integral a given node is to a network. Basically, without getting into the mathematics, it measures how often a node falls on the shortest path between other nodes. Thus, nodes with high betweenness centrality do not necessarily have a lot of edges, but they bridge disparate clusters. In an informational network, the nodes with high betweenness centrality are crucial to information flow, social connections, or collaborations. Hubs are examples of nodes with high betweenness centrality. The removal of a hub causes large portions of a network to become detached. In figure 7, the node labeled "Folkner, W.M." exhibits high betweenness centrality, since it connects two clusters together.

A cluster coefficient expresses whether a given node in a network is a member of a tightly interlinked collection of nodes, or clique. The cluster coefficient of an entire graph reveals the overall tendency for clustering in a graph, with higher cluster coefficients typical of small world graphs. In other types of graphs, clusters sometimes manifest as *homophily*; that is, nodes of a given type are highly interconnected with one another and have few connections with nodes of other types. In social networks, this is sometimes referred to as the "birds of a feather" effect. In a more current reference, the effect was explored as a function of the likelihood that someone would "unfriend" an acquaintance on the social networking site Facebook.[22] In some networks (such as the Internet), clusters are connected by hubs, while in others, the hub is the primary connecting node of other nodes.

Paths refer to the edges that connect nodes. The simplest case of a path is an edge that connects two nodes directly. Path analysis addresses the set of edges that connect two nodes that are not, themselves, directly connected. The shortest path, as its name implies, refers to the route that uses the least number of edges to connect from node A to node B and measures the number of edges, not the linear distance. Walks and paths refer to a list of nodes between two nodes, with walks allowing repeat visits to nodes, and paths not allowing them. Cycles refer to a path that connects a node through other nodes back to itself. Within graph visualization tools, the placement of nodes can vary from one layout to another. What matters is not the pictorial representation (though this can be useful), but the underlying relationships between nodes (the topology). Along with clustering, paths help differentiate motifs, which are considered to be building blocks of some types of complex networks.

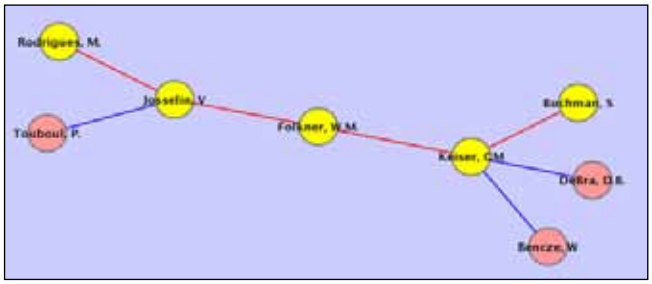

Figure 7. Paths in a Coauthorship Network

Since bibliographic metadata represents communication in one form or another, it is often most common to apply social network theory to graphs. But it is also possible to apply various centrality measures to graphs that are not social and to use these to discover significant nodes within those graphs. In part 2 we consider various unipartite and bipartite graphs that might be especially useful for examining digital library metadata.

Part 2. Graph Theory Applications in Digital Libraries

Library systems, by virtue of the content they contain, are complex systems. Fielded searches, faceted searches, and full-text searches all allow users to access aspects of the complex system. Fielded searches leverage the explicit structure that has been encoded into the metadata describing the resources that users are ultimately trying to find (articles, books, etc). Full-text searches enable users to explore in a more free-form manner, subject of course to the availability of searchable text. Often, full-text search means the user is searching titles, abstracts, and other content that summarizes a resource, rather than the actual full text of articles and books. Even if the user is searching the full content, there are relationships and aspects of the content that are not readily discernible through a full-text search. Furthermore, there is not one single, comprehensive digital library—many library systems live in the Deep Web, that is, they are databases that are not indexed by search engines like Google, and so users must

search each individually. But if more of these systems adopted Semantic Web standards, they could be explored as graphs, and relationships between different databases would be easier to discern and represent to the user.

Many libraries have tried to emulate Google by incorporating federated search engines with a single search box as an interface. This copies the form of Google's search engine but not its underlying power. To do that, libraries must enhance full-text searches by drawing on relationships. A full-text search will (hopefully) find relevant papers on a given topic, but a researcher often wants to find the *best* papers on that topic. To meet that need, libraries must harness the information contained in relationships; otherwise each paper is stuck in a vacuum.

Cited references are one way to connect papers. For researchers and librarians alike, this is a familiar metric for assessing a paper's relative importance. The Web of Science and Scopus are two databases that perform this function. Looked at another way, citation counts are nothing more than degree centrality applied to a simple graph in which papers are nodes and references are edges. Thus, in the framework of graph theory, citation analysis is just a small sliver of a world of possible relationships, many of which are unexplored.

The following examples outline use case scenarios in which graph techniques are or could be applied to library data, such as bibliographic metadata, to help users find relationships and conduct research.

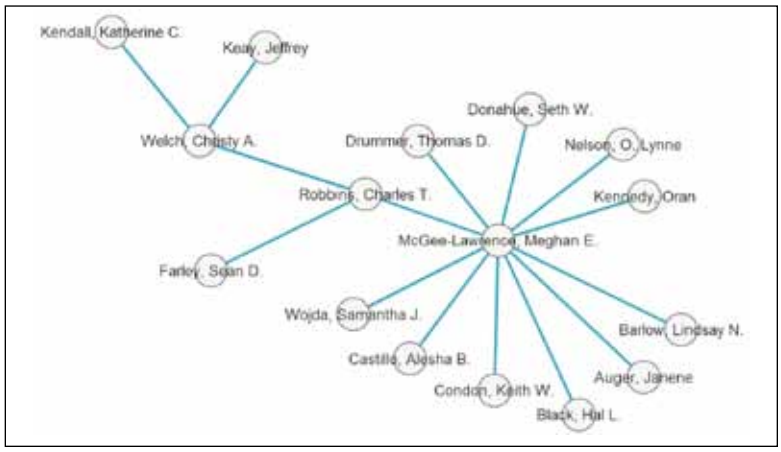

Figure 8. A Coauthorship Network

Informational Graphs Intrinsic to Digital Library Systems

There are multiple relationships represented within and between metadata contained in library systems that can be represented as graphs and explored using graph techniques. Some of these, such as citation networks, are among the most well-studied informational networks. Citation networks are valued because the data describing them is readily accessible and because scientists studying classes of networks have used them as surrogates for exploring scale-free networks. They are often evaluated as static networks (i.e., a snapshot in time) but some also have dynamic characteristics (e.g., they change and grow over time or they allow information-flow analysis). Techniques such as PageRank can be used to evaluate information when the importance of a linking resource is as important as the number of links to a resource. Multirelational networks can be developed to explore dynamic processes in research fields by using library data to provide the basic topological framework for some of the explorations.

Coauthorship (Collaboration) Networks

Coauthorship (collaboration) networks are typically small world networks in which cross- and interdisciplinary work provides the random links that connect various clusters (see figure 8). These graphs can be explored to determine which researchers are having the most influence in a given field; influence is a function of frequency of authorship. A prime example is the collaboration network graph for Paul Erdős, a highly productive mathematician. The popularity of his influence in academia has lead to the creation of the Erdős Number, which is "defined as indicating the topological distance in the graph depicting the co-authorship relations."[23] Liu et al. proposed a node analysis measure that they called AuthorRank, which establishes weighted directed edges between authors. The author's AuthorRank value is a sum of the weighted edges connected to that author.[24] These networks also can be used to explore how an idea spreads and what opportunities may exist for future collaborations, as well as many other existing and potential relationships.

Citation Graphs

Citation graphs more strongly resemble scale-free networks, in which early papers in a given field tend to accumulate more links. Such hub papers can be cited hundreds or even thousands of times while most papers are cited far less often or not at all. Many researchers have explored citation graphs, though the person often credited with first noting the network characteristics of citation patterns was Dereck J. de Solla Price in 1965.[25] More recently, Mark Newman introduced the concept of what he calls "first mover advantage" to describe the preferential attachment observed in citation networks.[26]

Subject–Author (Expertise) Graphs

Graphs that connect authors by subject areas can vary because of the granularity of subject headings (see figure 9). High-level subject headings tend to function as hubs, but more useful relationships are revealed by specific subject headings and author-provided keywords. The map of science merges publications and citations with actual end user usage patterns captured in library systems and deals, in part, with categories of scientific research.[27] It clusters publications and visualizes them "as a journal network that outlines the relationships between various scientific domains." Implicit in this a model is the relationship of authors to subject areas.

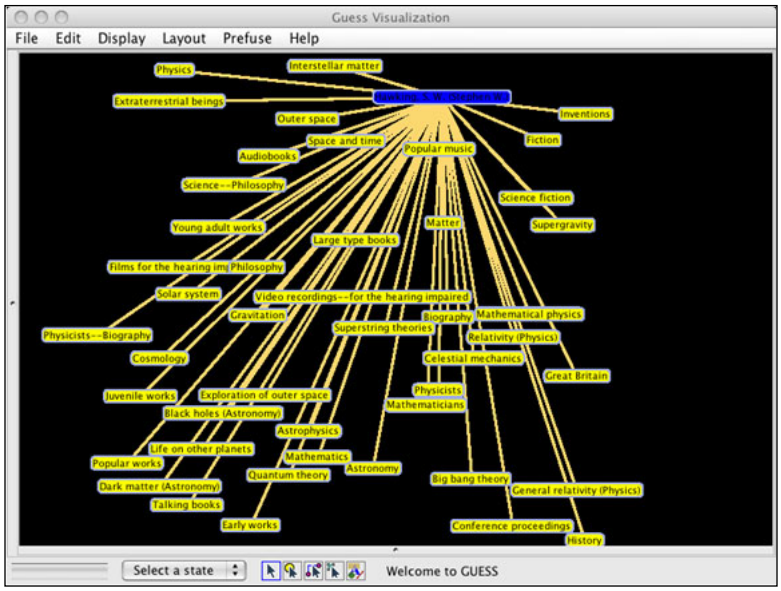

Figure 9. A Subject–Author Graph for Stephen Hawking

Institution–Topic and Nation–Topic (Expertise) Graphs

From a commercial or geopolitical perspective, graphs that represent institutional or national expertise can reveal valuable information for scientists, elected officials, and investors, particularly in networks that represent the change in a given organization or region's contributions to a field over time. Metadata for scientific papers typically includes enough information to generate nodes and edges describing this. The resulting graph can reveal unexpected details, such as national or institutional efforts to nurture expertise in a given field, and the results of those efforts. The visualization of this data may take the form of icons that vary in shape and size depending on various aspects of nodes in the institution-topic network. These visual representations can then be overlaid onto a map, with various visual aspects of the icons also affected by centrality measures applied to a given institution's contributions.[28]

Graphs as Tools

Graph representations can be used as tools to explore a variety of complex systems. Even systems that do not initially appear to manifest networks of relationships can often be better understood when some aspect of the system is represented as a graph. This approach requires thinking about what aspects of information needs, discovery, or consumption might be represented or evaluated using networks. Two interesting examples from other fields will illustrate the point.

A 2009 paper in *Acta Astronautica* proposed that techniques to reduce the amount of space junk in orbit around the earth could be evaluated using graph theory techniques.[29] The authors propose a dynamic multirelational network with three types of nodes: one to represent individual pieces of debris, a second to represent collections of debris that are the original object that the debris is a fragment of, and a third to represent conjunction events (near misses) between objects.

Another example of graphs being used as tools is the case of developing vaccination strategies to curtail the spread of an infectious disease.[30] In this case, networks have been used to determine that one of the best strategies for curtailing the transmission of a disease is to identify and vaccinate hubs, rather than to conduct mass vaccination campaigns.

In libraries, graphs as tools could be used to help researchers identify collaboration opportunities, to disambiguate author identities and aggregate related materials, to allow library staff to evaluate the academic contribution of a group of researchers (bibliometrics), and to explore geospatial and temporal aspects of information, including changes in research focus over time.

Graphs for Author Name Disambiguation

Author name disambiguation is a long-standing problem in libraries. Many resources have been devoted to manual and automatic name authority control, yet the problem remains unsolved. Projects such as OCLC VIAF and efforts to establish unique author identifiers will no doubt improve the situation, but many problems remain.[31] Meanwhile, we have experimented with an approach to author name matching by generating multirelational graphs. Authors

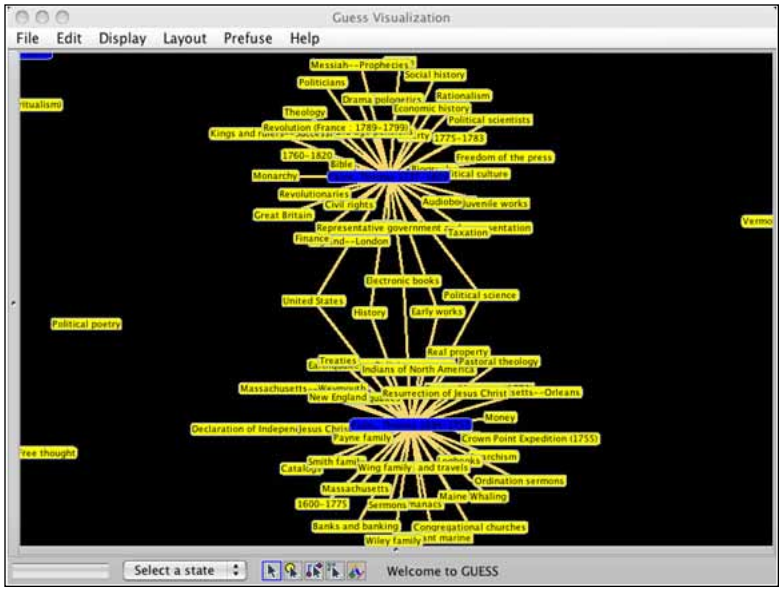

Figure 10. Two Authors with Similar Names Linked by Subject Nodes

are the primary nodes of interest, but relationships such as topic areas, titles, dates, and even soundex representations of names also are represented. As one would expect, phonetically similar names cluster around particular soundex representations. Shared coauthorship patterns and shared topic areas can reveal that two different names are for the same author as, for example, when a person's name changes after marriage (see figure 10).

Graphs for Title or Citation Deduplication

String edit distance involves counting the number of changes that would need to be made to one string to convert it to another, and it is one of the most common approaches to deduplicating titles, citations, and author names. Multirelational graphs, in which titles are linked to authors, publication dates, and subjects, result in subgraphs that appear virtually identical when two title variants are represented. Centrality measures can be applied to unipartite subgraphs of these networks to home in on areas where data duplication may exist.

Temporal-Topic Graphs for Analyzing the Evolution of Knowledge Over Time

A particularly active area of research in graph theory is the representation of dynamical systems as networks. A dynamical system is described as a complex system that changes over time.[32] Computer scientists have developed various strategies and technologies to cope with computation over time because it is typically so important to understanding data. Allen's temporal intervals address machine reasoning over disparate means of recording the temporal aspects of events.[33] Another temporal computing concept that has applicability to graphs is from the Memento project, which makes it possible for users to view prior versions of webpages.[34] Entities in the Memento ontology can become predicates in triples, which in turn can become edges in graphs. Using graphs, time can be represented as a relationship between objects or as a distinct object within a graph. Nodes that connect through a temporal node may overlap, coincide, or co-occur. Nodes that cluster around time represent something important about the objects.

Genomic-Document and Protein-Document Networks

Many people hoped that mapping the human genome would result in countless medical advances, but the process whereby genes manifest themselves in living organisms turned out to be much more complex—there wasn't just a simple mapping between genes and organism traits, there were other processes controlled by genes representing additional layers of complexity scientists had not anticipated. Today biologists apply network science to these processes to reveal the missing pieces of this puzzle.[35] Just as the process itself is complex, the information needs of these researchers benefit from a more sophisticated approach. Biologists often need to find papers that reference a given gene or protein sequence. And so, representing these relationships (e.g., article–gene) as graphs has the added benefit of making the digital library research data compatible with the methods that biologists already use to document what they know about these processes. Although this is a specialized type of graph, a similar approach might be valuable to researchers in a number of scientific disciplines, including materials science, astrophysics, and environmental sciences.

Graphs of Omission

One of the less obvious capabilities of network science is to make predictions about complex systems by looking for missing nodes in graphs.[36] This has many applications: for example, identifying a hub in the metabolic processes of bacteria can yield new targets for antibiotics, but it is vital to know that interrupting the enzyme that serves as that hub will effectively kill the organism. Making predictions about the evolution of research by identifying areas for cross-disciplinary collaboration or areas where little work has been done—enabling a researcher to leverage

the first-mover advantage and thus advance his or her career—is a valuable service that libraries are well positioned to provide (see figure 11).

Machine-supplied suggestions offer another type of prediction. For example, providing the prompt "Did you mean John Smith and climate change?" can leverage real or predicted relationships between author and subject (see figure 12). Graphs, in turn, can be used to create tools that will simplify an author–subject search.

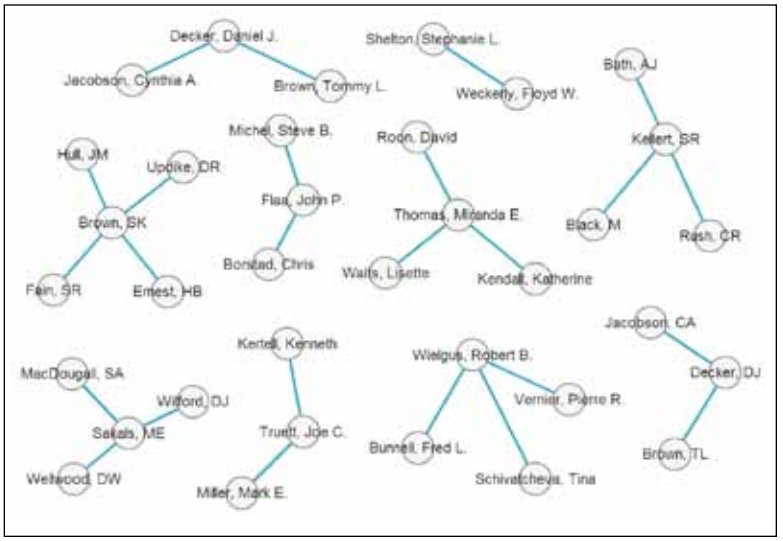

Figure 11. Identifying Areas for Collaboration: A co-author graph with many simple motifs and few clusters might indicate a field ripe for collaboration

Viral Concept Detection

Phase transition typically refers to a process in thermodynamics that describes the point at which a material changes from one state of matter to another (e.g., liquid to solid). Phase transition also applies to the dispersal of a new idea. Interestingly enough, graphs representing matter at the point of phase transition, and graphs representing the spread of a fad in a social network, exhibit the same recognizable pattern of change: suddenly there are links between many more nodes, there's a dramatic increase in clustering, and something called a *giant component* emerges.[37] In a giant component, all of the nodes in that portion of the graph are interlinked, resulting in a complete graph like figure 5. This is not so different from what one observes when something "goes viral" on the Internet. In a library, a dynamic graph showing the usage of new keywords for emerging subject areas would likely reflect a similar pattern.

Linked Data Graph Examples

Cross-collection graphs, or graphs that link data under your control to data published online, can be constructed by building links into the Web of Linked Data.[38] Linked data refers to semantic graphs of statements that various organizations publish on the web. For example, Geonames.org publishes millions of statements about geographic locations on the Linked Data Web.[39] As these graphs grow and evolve, opportunities emerge for using this data in combination with your own data in various ways. For example, it would be quite interesting to develop a network representation of library subject headings and their relationships to concepts in the encyclopedic linked data collection known as DBpedia.[40] The resulting graph could be used in a variety of ways: for example, to evaluate the consistency of statements made about concepts, to establish semantic links between user-provided tags and concepts,[41] or to function as the basis for an on-the-fly search expansion tool. A query-suggestion tool might look at user-entered terms and determine that some are hubs, then suggest related terms from nodes that connect to those hub nodes. Remember, graphs need not be visible to be useful!

Global Subject Resolution using Dbpedia

Although Dbpedia appears to lag behind Wikipedia in terms of completeness and scrutiny by domain experts, it offers one mechanism for unifying user-provided tags, author keywords, and library-assigned subject headings with a graph of known facts about a topic. Links into and out of Dbpedia's graphs on a given topic would enable serendipitous knowledge discovery through browsing these semantic graphs.

VIAF Linked Author Data

OCLC's effort to convert tens of millions of identity records into graphs describing various attributes of authors promises to enhance exploration of digital library content on the author dimension.[42] These authority records contain a wealth of information, linking name variations, basic genealogical data such as birth and death dates, associations with institutions, subject areas, and titles published by authors. Although some rough edges need to be smoothed (one of the authors of this paper discovered that his own authorship data was linked with another author of the same name), iterative refinement of this data as it is actually used may enable crowd-sourced

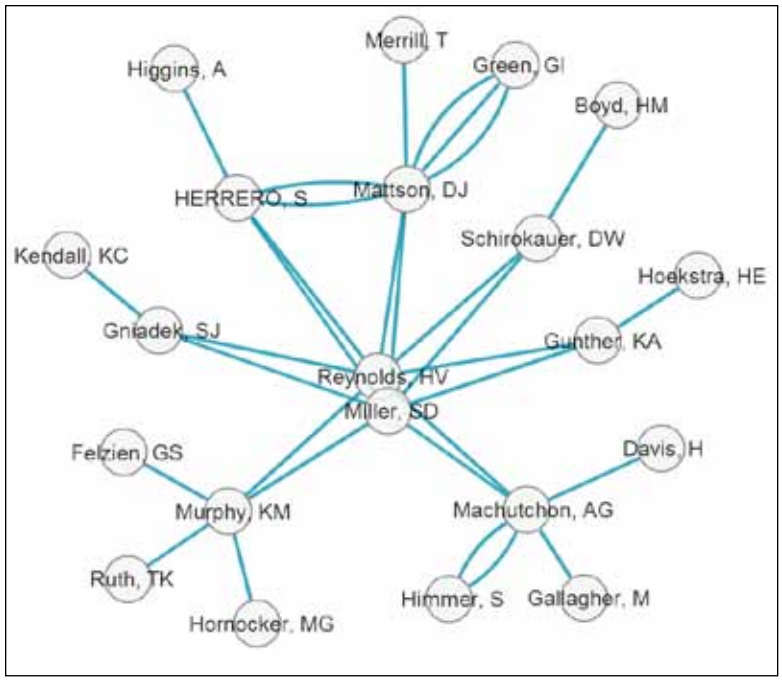

Figure 12. Find Similar Articles: A search for HV Reynolds might prompt a suggestion for SD Miller, who has a similar authorship pattern

quality control that will more rapidly identify and resolve these problems.

Linked Geographic Data Using Geonames

It is ironic that the use of networks to describe geographic aspects of the world is in its infancy, considering that many consider Leonhard Euler's attempt to find a mathematical solution to the Seven Bridges of Königsberg problem in 1735 to be the birth of the field.[43] As some authors have pointed out, geometric evaluation of geographic relationships is actually a poor way to explore geographic relationships.[44] Graphs can be used to express arbitrary relationships between geographically separated objects, and it is perhaps no accident that our road and railway systems are in fact among the most familiar graphs that people encounter in the real word. A subway map is a graph where subway stations are nodes linked by railway. Graphs can represent the relationships between topological features, the visibility of buildings in a city to one another, or the land, sea, and air transportation that links one country to another. Geonames supplies a rich collection of geographic information that includes descriptions of geopolitical entities (cities, states, countries), geophysical features, and various names that have been ascribed to these objects. The geographic relationships in intellectual content could be represented and explored as a graph, and some research has already shown that geographic networks—especially those representing human-constructed entities such as cities and transportation networks—exhibit small world characteristics.[45]

Another way graphs can express geographic relationships in useful ways would be in representing the concept of nearness. Waldo Tobler's first law of geography states that "everything is related to everything else, but near things are more related than distant things."[46] In practice, human beings define nearness in different ways, so a graph representing a shared concept of nearness would be very valuable, particularly in exploring works associated with biological, ecological, geological, or evolutionary sciences. Graph representations of nearness could be developed by librarians working with scientists and could be the geographic equivalent to subject guides and finding aids. They also might be useful across disciplines and would enable machine inferencing across data that include geographic relationships.

Still Other Kinds of Graphs

What might a digital library tool based on graph theory look like? What could it do? It wouldn't necessarily depict visualizations of graphs (though in some cases visual graphs are the most efficient way to impart concepts). After all, citation databases utilize graph theory, but the user only sees a number (cite count) and lists of articles (citing or cited). In many cases, then, the tool would perform graph evaluation techniques behind the scenes, translating these metrics into simple descriptive queries for the user. For example, a user interested in the most influential papers in his field would enter his subject, and then on the backend, the tool would apply Eigenvector centrality to that subject's citation graph. If the same user finds an especially relevant article, clicking a "find similar articles" button will produce a list of papers in that graph with the shortest path length to the paper in question.

Researchers also could use this tool to evaluate authors and institutions in various ways:

- *Is my output diverse or specialized compared to my colleagues?* The tool assigns a score for each author based on degree centrality in a subject-author graph.
- *I want to find potential collaborators.* The tool returns authors connected to researcher by the shortest path length in a coauthorship graph.
- *I want to collaborate with colleagues from other departments at my institution.* High betweenness centrality

in a subject–author graph for that institution may locate potential "bridge" subjects to collaborate in.

- *I'm leaving my current job. What other institutions are doing similar work?* In an institution–subject graph, the shorter the path length between two institutions, the more comparable they may be.

Graphs also enable libraries to reach outside their own data to build connections with other data sets. Heterogeneity, which makes relational database representations of arbitrary relationships difficult or impossible, becomes a trivial matter of adding additional nodes and edges to bridge collections. The Linked Data Web defines simple requirements for establishing just such representations, and libraries are well-positioned to build these bridges.

Conclusion

For many centuries, libraries have served as repositories for the accumulated knowledge and creative products of civilization, and they contain mankind's best efforts at comprehending complexity. This knowledge includes scientific works that strive to understand various aspects of the physical world, many of which are complex and require the efforts of numerous researchers over time. Since the advent of the Dewey Decimal System, librarians have worked on many fronts to make this knowledge discoverable and to assist in its evaluation. Qualitative evaluation increasingly requires understanding a resource in a larger context. We suggest that this context is itself a complex system, which would benefit from the modeling and quantitative evaluation techniques that network science has to offer. We believe librarians are well positioned to leverage network science to explore and comprehend emergent properties of complex information environments. As motivation for this pursuit, we offer in closing this prescient quote from Carl Woese, which though focused on the discipline of biology, is equally applicable to the myriad complexities of modern life: "A society that permits biology to become an engineering discipline, that allows that science to slip into the role of changing the living world without trying to understand it, is a danger to itself."[47]

References

1. Melanie Mitchell, *Complexity: A Guided Tour* (Oxford, England; New York: Oxford Univ. Pr., 2009).

2. Carl Woese, "A New Biology for a New Century," *Microbiology and Molecular Biology Reviews* (June 2004): 173–86, DOI: 10.1128/MMBR. 68.2.173–186.2004.

3. National Research Council (U.S.), *Network Science* (Washington, D.C.: National Academies Pr., 2005).

4. Tim Berners-Lee, "Giant Global Graph," online posting, Nov. 21, 2007, timbl's blog, http://dig.csail.mit.edu/breadcrumbs/node/215.

5. Lawrence Page et al., *The PageRank Citation Ranking: Bringing Order to the Web (1999),* http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1768.

6. Duncan S. Callaway et al., "Network Robustness and Fragility: Percolation on Random Graphs," *Physical Review Letters* 85, no. 25 (2000): 5468–71.

7. Adreas Wagner and David A. Fell, "The Small World Inside Large Metabolic Networks," *Proceedings of the Royal Society B: Biological Sciences* 268, no. 1478 (2001): 1803–10.

8. Gil Benko, Christopher Flamm, and Peter F. Stadler, "A Graph-Based Toy Model of Chemistry," *Journal of Chemical Information and Modeling* 43, no. 4 (2003): 1085–93.

9. Tad Hogg, Bernardo A. Huberman, and Colin P. Williams, "Phase Transition and the Search Problem," *Artificial Intelligence* 81 (1996): 1–15.

10. Vladimir Boginski, Sergiy Butenko, and Panos M. Pardalos, "Mining Market Data: A Network Approach," *Computers & Operations Research* 33, no. 11 (2006): 3171–84.

11. Zoltán Dezső and Albert-László Barabási, "Halting Viruses in Scale-Free Networks," *Physical Review E* 65, no. 5 (2002), DOI: 10.1103/PhysRevE.65.055103.

12. Hans Noel and Brendan Nyhan, "The 'Unfriending' Problem: The Consequences of Homophily in Friendship Retention for Causal Estimates of Social Influence," Sept. 2010, http://arxiv.org/abs/1009.3243.

13. Johan Bollen et al., "Toward Alternative Metrics of Journal Impact: A Comparison of Download and Citation Data," *Information Processing & Management* 41, no. 6 (2005): 1419–40; Xiaoming Liu et al., "Co-authorship Networks in the Digital Library Research Community," *Information Processing & Management* 41, no. 6 (2005): 1462–80.

14. Johan Bollen et al., "Clickstream Data Yields High-Resolution Maps of Science," ed. Alan Ruttenberg, *PLoS ONE* 4, no. 3 (3, 2009): e4803.

15. Eric Kolaczyk, *Statistical Analysis of Network Data* (New York; London: Springer, 2009).

16. Alejandro Cornejo and Nancy Lynch, "Reliably Detecting Connectivity using Local Graph Traits," *CSAIL Technical Reports* MIT-CSAIL-TR-2010–043, 2010, http://hdl.handle.net/1721.1/58484 (accessed Feb. 17, 2011).

17. Réka Albert, Hawoong Jeong, and Albert-László Barabási, "Error and Attack Tolerance of Complex Networks," *Nature* 406, no. 6794 (2000): 378–82.

18. M. E. J. Newman, "Scientific Collaboration Networks. II. Shortest Paths, Weighted Networks, and Centrality," *Physical Review E* 64, no. 1 (2001), DOI: 10.1103/PhysRevE.64.016132.

19. Albert, Jeong, and Barabási, "Error and Attack Tolerance."

20. R. Milo, "Network Motifs: Simple Building Blocks of Complex Networks," *Science* 298, no. 5594 (2002): 824–27.

21. Lawrence J. Hubert, "Some Applications of Graph Theory to Clustering," *Psychometrika* 39, no. 3 (1974): 283–309.

22. Noel and Nyhan, "The 'Unfriending' Problem."

23. Alexandru Balaban and Douglas Klein, "Co-authorship, Rational Erdős Numbers, and Resistance Distances in Graphs," *Scientometrics* 55, no. 1 (2002): 59–70.

24. Liu et al., "Co-authorship networks in the digital library research community."

25. Derek J. de Solla Price, "Networks of Scientific Papers,"

Science 149, no. 3683 (July 30, 1965): 510–15.

26. M. E. J. Newman, "The First-Mover Advantage in Scientific Publication," *EPL (Europhysics Letters)* 86, no. 6 (2009): 68001.

27. Bollen et al., "Clickstream Data Yields High-Resolution Maps of Science."

28. Chaomei Chen, Jasna Kuljis, and Ray J. Paul, "Visualizing Latent Domain Knowledge," *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)* 31, no. 4 (Nov. 2001): 518–29.

29. Hugh G. Lewis et al., "A New Analysis of Debris Mitigation and Removal Using Networks," *Acta Astronautica* 66, no. 1–2 (2010): 257–68.

30. Dezso and Barabási, "Halting Viruses in Scale-Free Networks."

31. OCLC, "VIAF (The Virtual International Authority File) [OCLC—Activities]," http://www.oclc.org/research/activities/viaf/ (accessed Feb. 17, 2011).

32. Mitchell, *Complexity: A Guided Tour*.

33. James F. Allen, "Toward a General Theory of Action and Time," *Artificial Intelligence* 23, no. 2 (1984): 123–54.

34. Herbert Van de Sompel et al., "Memento: TimeMap APO for Web Archives," http://www.mementoweb.org/events/IA201002/slides/memento_201002_TimeMap.pdf (accessed Feb. 17, 2011).

35. Hawoong Jeong et al., "Lethality and Centrality in Protein Networks," *Nature* 411 (May 3, 2001): 41–42.

36. Aaron Clauset, Cristopher Moore, and M. E. J. Newman, "Hierarchical Structure and the Prediction of Missing Links in Networks," *Nature* 453, no. 7191 (2008): 98–101.

37. M. E. J. Newman, "The Structure of Scientific Collaboration Networks," *Proceedings of the National Academy of Sciences of the United States of America* 98, no. 2 (Jan. 16, 2001): 404–9.

38. Chris Bizer, Richard Cyganiak, and Tom Heath, *How to Publish Linked Data on the Web?* http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/ (accessed Feb. 17, 2011).

39. GeoNames, http://www.geonames.org/ (accessed Feb. 17, 2011).

40. DBpedia, http://dbpedia.org/ (accessed February 17, 2011).

41. Alexandre Passant and Phillippe Laublet, "Meaning of a Tag: A Collaborative Approach to Bridge the Gap between Tagging and Linked Data," *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008)*, Bejing, Apr. 2008, doi: 10.1.1.142.6915.

42. OCLC, "VIAF"; OCLC homepage, http://www.oclc.org/us/en/default.htm (accessed Feb. 17, 2011).

43. Norman Biggs, *Graph Theory, 1736–1936* (Oxford, England; New York: Clarendon, 1986).

44. Bin Jiang, "Small World Modeling for Complex Geographic Environments," in *Complex Artificial Environments* (Springer Berlin Heidelberg, 2006): 259–71, http://dx.doi.org/10.1007/3-540-29710-3_17.

45. Gillian Byrne and Lisa Goddard, "The Strongest Link: Libraries and Linked Data," *D-Lib Magazine* 16, no. 11/12 (2010), http://www.dlib.org/dlib/november10/byrne/11byrne.html (accessed Feb. 17, 2011).

46. Daniel Sui, "Tobler's First Law of Geography: A Big Idea for a Small World?" *Annals of the Association of American Geographers* 94, no. 2 (2004): 269–77.

47. Woese, "A New Biology for a New Century."