Börner, Katy. In Press. Network Science: Theory, Tools, and Practice.
In William Sims Bainbridge, Ed. *Leadership in Science and Technology: A Reference Handbook.*
Thousand Oaks, CA: Sage Publications

040. Network Science: Theory, Tools, and Practice

Overview

Today’s world is flooded with data, information, knowledge, and expertise. Science and technology (S&T) leaders and other decision makers need effective approaches and qualitatively novel tools to help them identify what projects, patents, technology claims, or other developments are most important or profitable and hence deserve their full attention. They need approaches and tools that let them analyze and mine terabytes of relevant data and communicate results in an easy to understand way, enabling them to increase the quality of their decisions.

This chapter reviews and exemplifies approaches applied in the study of scholarly networks such as co-author networks and co-inventor networks, paper citation networks and patent citation networks, or topic maps. The goal of these studies is the formation of accurate representations of the world of science and technology, its structure, evolution, inputs, outputs, and flows. The work builds on social network analysis, physics, information science, bibliometrics, scientometrics, econometrics, informetrics, webometrics, communication theory, sociology of science, and several other disciplines.

Types of Analysis

A well designed and executed network analysis satisfies a specific insight need, i.e., it provides an answer to a concrete question. At the most general level, there exist five general types of questions: when, where, what, with whom, and why, as detailed below. Most real world decisions benefit from answers to multiple question types. Note that different question types require the application of rather different algorithms and approaches developed in specific domains of science.

Temporal Analysis (When): Science and technology evolve over time. Companies as well as research areas are born; they grow, stabilize, or shrink; potentially merge and split; and might cease to exist. Employees get hired, papers and patents have publication dates, scientists might get cited and funded at some point in time. External events such as changes in the stock market or government policies impact the S&T system. Different entities have different latency rates.

Temporal analysis takes a sequence of time-stamped observations as input, e.g., hires, citation counts, to identify patterns, trends, seasonality, outliers, and bursts of activity. The sequence of events or observations ordered in time can be continuous—i.e., there is an observation at every instant of time; or discrete—i.e., observations exist for regularly or irregularly spaced intervals. Temporal aggregations, e.g., over days, years, or decades—are common. Filtering can be applied to reduce noise and make patterns more salient. Smoothing, i.e., averaging using a time window of a certain duration, or curve approximation might be applied.

In practice, the number of entities per time point is often plotted to get a first idea of the temporal distribution of a dataset, e.g., first and last time point covered; completeness of data; maximum and minimum data value; or seasonality. In addition, it is interesting to know the data value at a certain point; what growth, latency to peak, or decay rate exists; if there are correlations with other time series; or what trends are observable. Data models such as the least squares model—available in most statistical software packages—are applied to best fit a selected function to a data set and to determine if the trend is significant.

Kleinburg's burst detection algorithm (Kleinberg 2002) is commonly applied to identify words that have experienced a sudden change in frequency of occurrence. These words may be extracted from documents or email text, could be names of authors, stocks, companies, or countries. Rather than using simple counts of word occurrences, the algorithm employs a probabilistic automaton whose states correspond to the frequencies of individual words. State transitions correspond to points in time around which the frequency of the word changes significantly. The algorithm generates a ranked list of the word bursts in the document stream, together with the intervals of time in which they occurred. This can serve as a means of identifying topics, terms, or concepts important to the events being studied that increased in usage, were more active for a period of time, and then faded away.

It might also be of interest to see the evolution of a social network, customer-relations network, or industry-academia partnership network over time. Here relevant data, e.g., joint publications or funding awards, executed contracts, or emails sent are 'sliced' in time and each time slide is analyzed and visualized separately and then all slices are combined to an animation. Time slices can be cumulative, i.e., later data includes information from all previous intervals, or disjoint, i.e., each slice only includes data from its own time interval. Cumulative slices are used to show growth over time while disjoint slices show changes in network structure more effectively.

Geospatial Analysis (Where): Geospatial analysis has a long history in geography and cartography. It aims to answer the question of where something happens and with what impact on neighboring areas. Just like temporal analysis requires time stamps, geospatial analysis requires spatial attribute values or geolocations for each entity to be mapped. Geospatial data can be continuous, i.e., each person has a latitude and longitude position per time unit, or discrete, i.e., each country occupies an area often represented by a so called shape file. Spatial aggregations, e.g., merging via ZIP codes, congressional districts, counties, states, countries, and continents, are common.

In practice, a geospatial analysis might identify key competitors or potential collaborators in a given spatial area; or analyze the impact of space on the flow of expertise via hiring and firing of employees or via collaborations. Collaboration patterns but also information flow pathways are commonly represented and studied as networks, see Figure 1.

Topical Analysis (What): Topical analysis aims to identify the semantic coverage of an entity, e.g., the expertise profile of a person or the topics captured by a set of papers. It uses keywords or words extracted from text and might apply linguistic and other techniques to extract meaning from free text. Topical aggregations, e.g., grouping papers by journal volume, scientific discipline, or institution, are common.

In practice, a set of unique words or word profiles and their frequency are extracted from a text corpus. Stop words, such as 'the' and 'of' are removed. Stemming can be applied so that word like 'education', 'educator', 'educated' are all matched to the word stem 'educat' and are treated as the same word. Salton's term frequency inverse document frequency (TFIDF) is a statistical measure used to evaluate the importance of a word in a corpus (Salton and Yang 1973). The importance increases proportionally to the number of times a word appears in the paper but is offset by the frequency of the word in the corpus. Dimensionality reduction techniques are commonly used to project high-dimensional information spaces, i.e., the matrix of all unique patents multiplied by their unique terms, into a low, typically two-dimensional space.

Co-word analysis identifies the number of times two words are used in the title, keyword set, abstract and/or full text of a document (Callon, Courtial, Turner et al. 1983). The space of co-occurring words can be mapped providing a unique view of the topic coverage of a dataset, see Figure 2. S&T entities can be clustered according to the number of words they have in common.

Network Analysis (With Whom?): The study of networks aims to increase our understanding of what entities interact with each other in what ways. Datasets are represented as nodes and edges. Nodes might denote authors, institutions, companies, and countries or words, papers, patents, or funding awards. Edges represent social, scholarly, financial, or other interlinkages. Base maps, e.g., of scientific papers based on bibliographic coupling, can be created and used to overlap other data, e.g., funding awards. Diverse algorithms exist to extract, preprocess, analyze, or visualize networks, see section *Network Science Theory and Practice*.

Modeling (Why?): S&T decision makers are often asked to do the impossible: to predict the future outcomes of current decisions and to measure the returns on recent investments while much of the work is still in progress. Process models aim to simulate, statistically describe, or formally reproduce statistical and dynamic characteristics of interest. Different network models exist for generating, e.g., random networks, small world networks, or scale free networks. *Random networks* are a theoretical construct that is well understood and their properties can be exactly solved. They are commonly used as a reference, e.g., in tests of network robustness and epidemic spreading. A *small-world network* is one whose majority of nodes are not directly connected to one another, but still can reach any other node via very few edges. Many real world networks, e.g., social networks, the connectivity of the Internet, or gene networks, exhibit small-world network characteristics. *Scale-free networks* have a degree distribution that follows a power law, at least asymptotically. Examples are collaboration network among scientists, paper citation networks, or the movie actor network. See (Weingart, Guo, Börner et al. 2010) for details on corresponding models.

Level of Analysis

Data analysis can be performed at different levels such as micro (individual), meso (local), and macro (global). The different levels employ different approaches, algorithms, and tools and support different types of insight. The combination of insights from all levels is considerably larger than their sum.

Micro/Individual Level studies typically deal with 100 or less data records, e.g., all employees in a small firm, all patents on a given narrow topic. They are often qualitative, i.e., data is acquired via interviews, questionnaires, or ethnographic studies. Each record might have many different attributes. The data can often be analyzed by hand. Tabular listings and interactive visualizations are used to communicate results.

Meso/Local Level studies typically deal with more than 100 and up to 10,000 records. They increase our understanding of teams, groups of firms, sets of countries, etc. Most analyses are quantitative and involve large numbers of records that might be downloaded from databases, listserv archives, or are extracted from RSS feeds. Algorithmic means are needed to make sense of the data. Different types of visualizations—often involving a stable reference system, e.g., a map of a country, and multiple data overlays, e.g., a set of companies, their product flows and profits, help understand and communicate results.

Macro/Global Level studies often deal with more than 10,000 records. These global studies might examine world trade flows, emerging research areas in the landscape of science, or the diffusion of knowledge in large populations. The quantitative analysis of large scale datasets, requires extensive data storage and computing resources commonly available via cyberinfrastructures (Atkins, Droegemeier, Feldman et al. 2003). Parallelization of software code is common to run algorithms with high algorithmic complexity or to study datasets with more than 1 million nodes. Analysis results cannot be communicated at once. Instead, focus and context techniques are applied to show details within global patterns. Shneiderman's information seeking mantra: "Overview first, zoom and filter, then details on demand" guides the design of interactive interfaces (Shneiderman 1996). Stable reference systems with multiple data overlays are also used.

Sample Studies

Major analysis types and levels of analysis are shown in Table 1 together with sample analyses, see (Weingart et al. 2010) for details. Most real world decision making requires answers to a subset of when, where, what, with whom, and why questions. Four sample studies (bold in Table 1) are discussed below.

Table 1: Major analysis types and levels of analysis.

Analysis Types and Sample Studies	**Micro/Individual (1-100 records)**	**Meso/Local (101–10,000 records)**	**Macro/Global (10,000 < records)**
Statistical Analysis/Profiling	Individual persons and their expertise profiles	Larger labs, centers, universities, research domains, or states	All scientists in the US, NSF funding, English speaking publications.
Temporal Analysis (When)	Funding portfolio of one individual	**Mapping topic bursts in 20-years of PNAS**	113 years of physics research
Geospatial Analysis (Where)	Career trajectory of one individual	**Mapping a state's intellectual landscape**	PNAS publications
Topical Analysis (What)	Base knowledge from which one grant draws	Knowledge flows in Chemistry research	**Topic maps of NIH funding**
Network Analysis (With Whom?)	NSF Co-PI network of one individual	Co-author network	**NIH's core competency**

Recent progress in data analysis and visualization and the mapping of science (Börner 2010; Shiffrin and Katy Börner eds 2004) makes it possible to study and communicate the structure of science at a local and global scale. Maps of science might show the intellectual landscape of a geospatial region (see Fig. 1), bursts of activity in biomedical research (Mane and Börner 2004) (see Fig. 2), or a base map of science with overlays of funding by the National Institutes of

Health (NIH) (Boyack, Börner, and Klavans 2009).

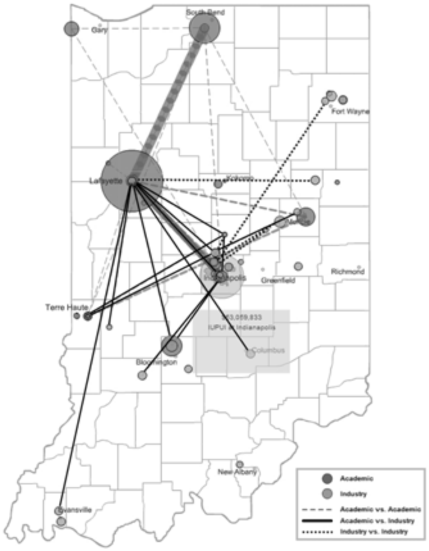

Figure 1: Biomedical grant proposals and awards from an Indiana funding agency covering 2001-2006 were analyzed to identify pockets of biomedical innovation, pathways from ideas to products, and the interplay of industry and academia. Nodes represent researchers and practitioners in industry (yellow) and academia (red). Links represent collaborations in grant proposals and awards of different types: Within industry (yellow), within academia (red), industry-academia (orange). Hovering over a node brings up the total award amount for a node. Clicking on a node leads to details formatted as sortable tables. Collaboration patterns are mapped in geospatial space.

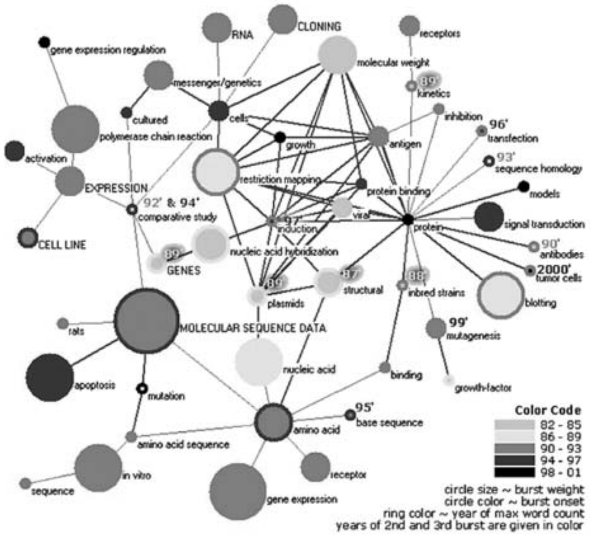

Figure 2: Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982–2001. Each node represents a single word. Nodes are size coded by burst strength, i.e., the amount of change of the word’s usage frequency. They are color coded by the year in which the first burst appears (subsequent bursts are given as year numbers next to the node). Node ring color indicates the year of the maximum word count. As

can be seen, many words burst earlier (ring color is darker than filling) or concurrently (ring and filling have same color) with the year of max word usage. Hence, busts serve as an indicator of future word usage activity. A dynamic abstract space of topics is given a two-dimensional shape.

Online maps of neuroscience abstracts (http://scimaps.org/maps/neurovis) or NIH projects funded in 2007 (http://scimaps.org/maps/nih/2007) support interactive data exploration. The NIH map renders each project funded in 2007 as a dot that is color coded by the institute which funded it. Successive zooms are provided together with expert assigned area labels. A set of projects can be selected to examine their distribution over the different NIH institutes and to access details such as investigators, title, abstract, etc. Keyword search can be used to see what projects target what scientific topics and what institutes fund that research. The international Mapping Science exhibit, online at http://scimaps.org, provides more examples of the state of the art in mapping science.

Workflow Design

The design of an analysis should be driven by an explicit set of prioritized user needs, see (Börner 2010), particularly pages 50-51. A typical workflow comprises about five or more data acquisition, preprocessing, analysis, modeling, and visualization steps and respective parameter values, see next section for details.

Results should be validated and interpreted in close collaboration with domain experts and other stakeholders. Insights gained might generate additional user needs or inspire changes to the workflow. The workflow design process is highly incremental, often demanding many cycles of revision and refinement to ensure the best datasets are used, optimal algorithm parameters are applied, and clearest insight are achieved. All datasets, algorithms, and parameter values used in a study should be documented at a detail that supports replication and interpretation.

Network Science Theory and Practice

Introduction

There are a number of recent popular science books (Barabási 2002; Barabási 2010; Börner 2010; Watts 1999; Watts 2003) but also reviews and textbooks that introduce network science approaches and techniques to physics (Barrat, Barthélemy, and Vespignani 2008; Bornholdt 2003; Newman 2010; Newman, Barabási, and Watts 2006; Pastor-Satorras and Vespignani 2004), computer science (Borgman and Furner 2002; Börner, Sanyal, and Vespignani 2007; Easley and Kleinberg 2010; Thelwall, Vaughan, and Björneborn 2005), communication studies (Monge and Contractor 2003), scientometrics and bibliometrics (Merton 1973; Nicolaisen 2007; White and McCain 1989; Wilson 1999), and social science perspective (Burt 2010; Carrington, Scott, and Wasserman 2005; Christakis 2009; Freeman 2004; Nooy, Mrvar, and Batagelj 2005; Scott 2000; Wasserman and Faust 1994). Here we focus on a rather small yet commonly used set of practical approaches and techniques relevant for S&T leaders.

Network Types

Different types of networks answer different questions. Co-author, co-inventor, or co-director networks help understand collaboration patterns on papers, patents, or S&T boards respectively. Networks that link a scholar or company via citation or reading patterns to terms, papers, or journals help us understand knowledge *import* or consumption. Citation, download, and product purchasing records capture the *export* or usage of papers and patents, online services, or product offerings. Plus, there are co-occurrence linkages, e.g., two terms are frequently used together or two patents are jointly cited by a third patent; authors, institutions, or countries appearing on a paper or patent can be co-cited by a third paper.

Some of these networks are unimodal networks, i.e., they have one node type. An example is co-author networks. Other networks are bi-modal, e.g., author-paper networks. The edges in a network can be weighted or not, directed or not. Co-author networks are weighted (e.g., by number of papers or citations) and undirected (i.e., the direction of knowledge transfer is unknown). In contrast, author-paper networks are unweighted (as every unique author is listed only once on a paper) and directed (the author actively publishes a paper).

Historically, bibliometric and scientometric studies have analyzed unimodal networks of one node type. Examples are co-author networks, paper citation networks, journal co-citation, or term co-occurrence networks. In these networks, nodes of one type, e.g., authors, terms, papers, journals, are connected by one edge type such as direct citation, co-citation, or co-occurrence of, e.g., references (bibliometric coupling), authors (co-authorship), or terms (co-term analysis).

Note also that networks can be extracted at different levels of aggregation. For example, there are paper-citation and journal-citation networks. Each type of network answers a very specific question.

Data Acquisition and Preparation

Typically, about 80 percent of the total project effort is spent on data acquisition and preprocessing; yet well prepared data is mandatory to arrive at high-quality results. Datasets might be acquired via questionnaires, crawled from the Web, downloaded from a database, or accessed as continuous data stream. Datasets differ by their coverage and resolution of time (hours, days, months, years), geography (industrial areas, languages and/or countries considered), and topics (keywords, product areas). Their size ranges from several bytes to terabytes (trillions of bytes) of data. Some datasets are curated by domain experts others are retrieved from the Web. Datasets might have to be augmented, e.g., a geospatial mapping requires geocoding where address information is used to determine latitude and longitude for a record; topical analysis benefits from stemming and the removal of stop words. Based on a detailed needs analysis and deep knowledge about existing databases, the best suited yet affordable datasets have to be selected, filtered, and interlinked.

Network Extraction

Before network analysis and visualization techniques can be applied, datasets that come in tabular or other formats, must be re-represented as networks. For example, authors, institutions, and countries, as well as words, papers, journals, patents, and funding can be represented as nodes and their complex interrelations as edges. Nodes and edges can have (time-stamped) attributes. Commonly studied networks comprise:

- Paper/Journal citation networks as a direct trace of scientific production and consumption.
- Patent citation networks as a direct trace of technological progress and intellectual property claims.
- Co-author/institution/country networks – “invisible communities” of social interaction that produce scientific products.
- Keyword or semantic maps that represent topic and product spaces.

To extract a citation network from a table of patents (one patent per row with columns representing patent ID, title, inventors, year, references, etc.), a computer program might take two column as input: one with the patent ID and a second one with the patent IDs of all cited patents. The program then determines the set of unique patent IDs and their interlinkages via citations.

To extract a co-author (i.e., author co-occurrence) network from a table, a computer program is given a column with all authors on each scientific paper (row). The algorithm determines the set of unique authors and their co-author linkages. The resulting network will show a fully connected clique for each set of papers listed on a paper, e.g., if one paper has n=100 authors then each node of the corresponding 100 node network is connected to each of the other 99 nodes, i.e., there are n^2-n/2 or 4,950 edges.

Multi-node type networks might represent patents linked to companies or scientists linked to scientific papers. Here, a computer program is given two table columns as input, e.g., one column containing all patent IDs and another containing all inventors. The program determines the list of unique inventors and for each inventor the patents they are listed on. The resulting network has two types of nodes: patents and inventors. Patents with many links (high degree) have many inventors; inventors with many links hold many patents. As inventors tend to have expertise in a certain S&T area, there might be topic-based clusters of inventors and patents. Clusters are likely linked by patents from inventors with different expertise and/or by inventors that hold patents in different S&T areas.

Networks take time to grow. A co-author network for a set of publications with the very same publication year will likely have many small unconnected networks. A paper citation network for the very same dataset will have few or no linkages.

Subsequently, we discuss general units of analysis and aggregations. Then we explain different linkage types and network types, and aggregations.

Basic units of analysis form the basis of reporting and may include:

- *Paper:* published article or conference paper
- *Patent:* issued patents
- *Funding awards:* grants, etc.
- *Funding institution:* agency, division, foundation, etc.
- *Scientist*: Scientific paper author, funding award recipient (PI), or patent inventor
- *Institution*: research group, university, corporation, city, state, region, nation, etc.
- *Topics:* terms, either from keywords lists or extracted from titles, abstracts, and other text

Each unit of analysis might be situated in time (e.g., date of publication), geospatial space (ZIP code), and a contextual topic space (e.g., keywords). Dates allow direction, or causal linkages, to be inferred. For instance, if a grant predates a publication by several years, it may have generated the publication, where the converse cannot be true.

Linkages can represent very different connections among basic units of analysis. For example, two patents might be similar and connected by a link because they have the same classification or set of keywords, were written by the same inventors, are owned by the same company or country, or do cite each other. Common linkage types include, but are not limited to, the following:

- Grant-to-paper and derivative linkages such as author-to-PI.
- Paper-to-patent and derivative linkages such as author-to-inventor.
- Grant-to-patent and derivative linkages such as PI-to-inventor.

- Paper-to-non-scientific, public policy issue through words/terms or regulations through standards/numbers.
- Collaborations: institutional, cross-sector, e.g., government, academia, etc.

There exist direct linkages, co-occurrence linkages, and co-citation linkages as discussed subsequently.

Direct linkages occur in Document-Document (Citation) Networks, Author-Author (Citation) Networks, Source-Source (Citation) Networks, and Author-Paper (Consumed/Produced) Networks discussed subsequently.

In Document-Document (Citation) Networks, papers (but also patents) cite other papers via references resulting in an unweighted, directed paper citation graph. The direction of information flow from older to younger papers/patents can be indicated via arrows. This representation allows for the tracking of citation networks chronologically, yielding a better understanding of the influence of previous research on subsequent research. Citations made to a paper support the forward traversal of the graph. Citations made by a paper to earlier publications support backward traversal and search.

In Author-Author (Citation) Networks, authors cite other authors via document references forming a weighted, directed author citation graph. Like document-document networks, author citation networks represent the flow of information over time. Unlike document citations, however, these networks have weighted edges representing the volume of citations from one author to the next.

In Source-Source (Citation) Networks, source nodes (e.g., papers or journals) cite each other via directed citation links. These networks can reveal both the relative importance of certain publications, and the underlying connections between disciplines. These networks are directed and weighted by volume of citations between journals.

In Author-Paper (Consumed/Produced) Networks, active nodes, e.g., authors, produce and consume passive nodes, e.g., papers, patents, datasets, software. The resulting networks have multiple types of nodes, e.g., authors and papers. Directed edges indicate the flow of resources from sources to sinks, e.g., from an author to a written/produced paper to the author who reads/consumes the paper.

A second set of networks is constructed using co-occurrence linkages. Examples include Author Co-Occurrence (Co-Author) Networks, Word Co-Occurrence Networks, Document Cited Reference Co-Occurrence (Bibliographic Coupling) Networks, Author Cited Reference Co-Occurrence (Bibliographic Coupling) Networks, and Journal Cited Reference Co-Occurrence (Bibliographic Coupling) Networks as discussed here.

Author Co-Occurrence (Co-Author) Networks assume that having the names of two authors (or their institutions, countries) listed on one paper, patent, or grant is an empirical manifestation of collaboration. The more often two authors collaborate, the higher the weight of their joint co-author link. Weighted, undirected co-authorship networks appear to have a high correlation with social networks that are themselves impacted by geospatial proximity.

Word Co-Occurrence Networks calculate topic similarity of basic and aggregate units of science via an analysis of the co-occurrence of words in associated texts. Units that share more words in common are assumed to have higher topical overlap and are connected via linkages and/or placed in closer proximity. Word co-occurrence networks are weighted and undirected.

Document Cited Reference Co-Occurrence (Bibliographic Coupling) Networks assume that papers, patents or other scholarly records that share common references are coupled bibliographically. The bibliographic coupling (BC) strength of two scholarly papers can be calculated by counting the number of times that they reference the same third work in their bibliographies. The coupling strength is assumed to reflect topic similarity. Co-occurence networks are undirected and weighted.

Author Cited Reference Co-Occurrence (Bibliographic Coupling) Networks assume that authors who cite the same sources are coupled bibliographically. The bibliographic coupling (BC) strength between two authors can be said to be a measure of similarity between them. The resulting network is weighted and undirected.

Journal Cited Reference Co-Occurrence (Bibliographic Coupling) Networks are analogous to document and author bibliographic coupling networks but assume that journal cited reference co-occurrences provide a measurement of similarity between journals. Edge strength between two journals is determined by the summation of unique references both journals cite.

A third set of networks uses co-citation linkages to extract networks. Two scholarly records are said to be *co-cited* if they jointly appear in the list of references of a third paper. The more often two units are co-cited the higher their presumed similarity. Examples are Document Co-Citation Networks (DCA), Author Co-Citation Networks (ACA), and Journal Co-Citation Networks (JCA) discussed subsequently.

Document Co-Citation Networks (DCA) are a logical opposite of bibliographic coupling. The co-citation frequency equals the number of times two papers are cited together, i.e., they appear together in one reference list.

Author Co-Citation Networks (ACA) assume that if authors of works repeatedly appear together in references-cited lists then they must be related. Clusters in ACA networks often reveal shared schools of thought or methodological approach, common subjects of study, collaborative and student-mentor relationships, ties of nationality, etc. Some regions of scholarship are densely crowded and interactive. Others are isolated or nearly vacant.

Journal Co-Citation Networks (JCA) are analogous to DCA but nodes are journals and linkage co-citation frequency equals the number of times two journals are cited together. Slicing these networks by time can reveal the evolving structure of scientific disciplines. Like DCA and ACA, they are undirected and weighted.

Network Preprocessing

Network processing comprises entity resolution and data aggregation as discussed here.

Entity Resolution. Most raw datasets do not have unique identifiers for terms, papers, journals, authors, institutions, countries, etc. However, it is important that each occurs exactly once in a network. Prior work showed that misspellings of author names cause up to 45 percent of underestimation of citation counts (Wouters 1999). As an example, in the Thomson Reuters database in 1999, the name of Derek John de Solla Price was recorded as DeSollaD, DeSollaPD, DeSollaPDJ, Price D, Price DD, Price DDS, Price DJ, Price DJD, and Price DS. If these misspelling are not corrected, then Price's works will be attributed to nine instances of his name. If left uncorrected, nine nodes will be used to represent him in a co-author network affecting the structure of the network and all subsequent network analyses, e.g., counting the number of Price's co-authors, major clusters, etc.

If citation linkages are used to derive citation, co-citation, or bibliographic coupling networks, then the set of unique references has to be identified and references need to be matched to paper records.

If terms are used to extract topical relationships, then the term source needs to be decided, e.g., title, abstract, or keywords. Paper titles are often flashy instead of descriptive and terms extracted from titles will be problematic. Keywords provided by authors might not match a controlled vocabulary. Classification terms provided by publishers might be optimized for retrieval but not for topical analyses.

Typically, there are many more terms in a paper than authors. That is, if an author, paper, term network is extracted from a set of papers then the different authors and their papers are most likely connected by many more shared terms than any other type of linkage.

Data Aggregation. Basic units of analysis can be aggregated in time, geospatial space, or topic space. Common aggregations include, but are not limited to, the following:

- Publications by topic, journal, research area, discipline.
- Temporal over date, month (volume), year.
- Geospatial by zip code, city, congressional district, state, region, country.
- Topical by paper, journal, scientific field, scientific discipline.
- Institutional by sector (academia, industry, government).

Networks can be compiled for aggregate units of analysis to gain a more global understanding of the structure and dynamics of S&T.

Network Analysis

Networks can be analyzed at the node/edge level and at the network structure level. Basic node properties comprise degree centrality, betweenness centrality, or hub and authority scores. Edge properties include type (directed or not, weighted or not), strength (weak or strong), density (how many potential edges in a network actually exist), reachability (how many steps it takes to go from one "end" of a network to another), centrality (whether a network has a "center" point), and quality (reliability or certainty).

Network properties refer to the number of nodes and edges, network type (directed or not, weighted or not), network density, average path length, clustering coefficient; number of isolate nodes, parallel edges, and self loops; number and size of unconnected network components; but also distributions from which general properties such as small-world, scale-free, or hierarchical can be derived. Identifying major communities via community detection algorithms and calculating the "backbone" of a network via pathfinder network scaling or maximum flow algorithms helps to communicate and make sense of large scale networks. For details, please see textbooks mentioned in *Introduction*.

Network Visualization

Modular visualization design comprises nine distinct decisions, shown below. Each of the nine decisions will impact all other eight.

- **Reference Systems:** How will the space be organized? Time lines, scatter plots, geospatial base maps, or topical science maps are common choices.
- **Projections/Distortions:** Will the reference system be modified to help emphasize certain areas or provide focus and context? For example, global world maps project Earth's surface onto two dimensions.

- **Raw Data**: What raw data will be visualized? Placing it in a reference system will reveal first spatial clusters and outliers.
- **Graphic Design**: How will data attributes be encoded visually using qualities such as size, color, and shape coding of nodes, linkages, or surface areas?
- **Aggregation/Clustering:** What analysis techniques are applied to identify data entities with common attribute values or dense connectivity patterns and how will these clusters be represented, e.g., using color coding or lines that indicate cluster boundaries?
- **Coupled Windows:** Will more than one view of the data be shown, e.g., a timeline, geospatial map, and a topic map? If yes, how will these data views be coupled using global search or brushing and linking (i.e., if a record in one window is selected then the record is automatically highlighted in all other windows).
- **Interactivity:** In many cases, it is desirable to interact with the data, e.g., to zoom, pan, filter, search, and request details on demand.
- **Legend:** Delivers guidance on the purpose, generation, and visual encoding of the data. Mapmakers should proudly ‘sign’ their visualizations, adding credibility as well as contact information.
- **Deployment:** How will the visualization be delivered to the user? Common choices are paper printouts (affordable, high resolution, static); online animations or interactive visualization on handheld, laptop or desktop screens (lower resolution, high bandwidth required, dynamic); or three-dimensional, audiovisual environments (expensive, multi-user, advanced interface).

Note that decisions made regarding the visualization design interact with other workflow elements such as analysis algorithms (e.g., network analysis algorithms that compute additional node/edge attributes for graphic design, clustering techniques that indentify cluster boundaries), or layout algorithms (e.g., network layouts that compute a spatial reference system).

Static Network Visualizations

How to read data charts and geospatial maps is taught in school. Reading and interpreting network visualizations is a skill that many people do not yet have. In general, reading any visualization requires five steps:

1. Data source: understand the quality and coverage of the data.
2. Data mapping: how was the source data was represented as a network, i.e., what do the nodes and edges represent.
3. Visual encoding: how are additional node/edge attributes encoded by sizes, colors, shapes, etc.
4. Data manipulation: what filtering, clustering, other interactivity options exist.
5. Author: Credibility and funding source of the author—visualizations are designed, i.e., they show a subset of the data and omit other parts of the data.

Reading and interpreting a network visualization might start with the identification of the number of nodes and edges (few or many?), network density and distribution analysis (many or few edges; are there any clusters?), identification of nodes with many edges (high degree) or nodes that interconnect network clusters (high betweenness centrality), calculation of network clusters (major components) and major linkages that might make up the ‘backbone’ of the network. Most real world networks will have the majority of their nodes linked into one large subnetwork—the so called ‘giant component.’ All other nodes will be part of much smaller networks or are unconnected ‘isolates’.

To support network reading and interpretation, it is advantageous to use a graphical language to help people decode different network visualizations. People nodes might be represented by circles (individuals), papers and patents by rectangles (letters), funding by triangles (piles of money). Different types of linkages, e.g., direct, co-occurrence, or co-citation, can be indicated by different line types, e.g., solid, dashed, dotted. Continuous numeric properties such as number of citations for a paper or the number of times two people co-invented are best encoded by node area size and edge width. Categorical properties, e.g., male vs. female, or type of company, are often encoded by node/edge color.

Dynamically Evolving Network Visualizations

The visualization of evolving network substrates and activity patterns unfolding over them is a key research challenge in information visualization research. On a local level, the number of different nodes and node types but also their properties and linkages might change over time. Derived node attributes such as the number of links per node or degree, the centrality of a node or edge might alter as well. On a global network level, the number of subnetworks or components, the density of the network, and the location, structure, and strength of the network’s “backbone” might change. Visualization solutions should attempt to increase our understanding of the interdependence of network structure and function.

Common visualization approaches comprise

- Temporal plots of derivative statistics, e.g., changes in network densification over time,
- Networks that use a fixed layout with changing node/edge attributes per time frame.
- Networks that are laid out perfectly for each time frame including changing node/edge attributes.

Each of the three approaches can be presented as

- multiple static snapshots,
- pre-rendered animations, or
- interactive services which can be started, stopped, fast forwarded, or rewind interactively.

Time frames can be disjoint, overlapping, or cumulative. Time frames can have identical or different length. Ideally, time frames are defined in a data driven way based on key events, e.g., company mergers or acquisitions, stock market changes, introduction of new governmental policies.

The visualization of multiple evolving networks and their interdependencies poses even more challenges. Often, a combination of (1) textual value listings, (2) simple charts, and (3) visual depictions of evolving networks works best to communicate key network changes over time.

Interpretation

Each step in a network study impacts the end result: Incomplete or incorrect data can lead to wrong conclusions; selecting the wrong set or sequence of analysis algorithms will impact the value and accuracy of the result; different visualizations provide different views of the data that are more or less appropriate for addressing a specific information need. In general, the complete workflow used to create an analysis result should be documented in a fashion that enables anybody with access to the same data sets and algorithms to arrive at the very same results. This also supports the replication and improvement of results by others and ultimately the credibility of results.

The last but very important step in the general process of network analysis is the interpretation of results by domain experts and other stakeholders. Face to face meetings in which the datasets, algorithms and parameter values used are presented, different visualizations of the data are shown and discussed; paired with informal or structured feedback from experts work well. Frequently, interpretation of results and follow-up questions inspire future studies.

Network Analysis Tools

An overview of existing network analysis and visualization tools is given in (Börner, Huang, Linnemeier et al. 2010), Table 2. It compares Pajek, UCINet, Visone, GeoVISTA, Cytoscape, Tulip, iGraph, Gephi, CiteSpace, HistCite, R, GUESS, GraphViz, NWB Tool, BibExcel, or Publish or Perish and many other tools. The tools were originally developed in the social sciences, scientometrics, biology, geography, and computer science. Many but not all tools are open source, about half of them run on Windows only while others run on all major platforms. Tools like Network Workbench, TexTrend, or the Science of Science Tool that use OSGi (http://www.osgi.org) in combination with the CIShell algorithm architecture (http://cishell.org) make it possible to share algorithm plugins across different projects.

Future Directions

Many areas of science have adopted network science approaches and are advancing the state of the art. Theoretical advances are made in social science, communication studies, informetrics, webometrics, scientometrics, physics, and biology to name just a few. Tool and algorithm advances are mostly made in social science, biology, physics, computer science, and information science. Practical advances can be seen in education, political science, healthcare, and homeland security.

Theoretical advances comprise the extension of algorithms originally developed for unimodal networks (one node/edge type) to heterogeneous networks (multiple node/link types); the study of the interplay of network structure and network function—the structure of the network supports certain collaborations/information flows yet the usage of the network likely leads to structural changes. Another line of research aims to extend theories and models developed for the study of single entities, e.g., inventors, to the study of multiple entities, e.g., inventor teams. Plus, there is increasing interest to study S&T at multiple levels of aggregation—from micro/individual via meso/local to macro/global, see (Börner, Contractor, Falk-Krzesinski et al. 2010; Falk-Krzesinski, Börner, Contractor et al. 2010) and Table 1.

Tool development aims to increase the availability of existing and new network science algorithms developed in different areas of science using different programming languages. Desirable properties of these tools comprise:

- ease of use—most network science researchers, educators, and practitioners do not program or script.
- open source code—anybody can check and improve the code.
- easy integration of new algorithms—they become quickly available outside the original lab.
- customizability—support specific algorithm sets and workflows without programming.

Practical advances comprise new strategies for assembling study teams, treatment smoking habits, or preventing terrorist attacks.

References and Further Readings

Atkins, Daniel E., Chair, Kelvin K. Droegemeier, Stuart I. Feldman, Hector Garcia-Molina, Michael L. Klein, David G. Messerschmitt, Paul Messina, Jeremiah P. Ostriker, and Margaret H. Wright. 2003. "Revolutioning Science and Engineering Through Cyberinfrastructure." *Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. .*

Barabási, A. L. 2002. *Linked: The New Science of Networks*. Cambridge, UK: Perseus.

Barabási, Albert-László. 2010. *Bursts: The Hidden Pattern Behind Everything We Do*. New York, NY: Dutton.

Barrat, Alain, Marc Barthélemy, and Alessandro Vespignani. 2008. *Dynamical Processes on Complex Networks* New York, NY: Cambridge University Press.

Borgman, C.L. and J. Furner. 2002. "Scholarly Communication and Bibliometrics." in *Annual Review of Information Science and Technology*, edited by B. Cronin and R. Shaw. Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.

Börner, Katy. 2010. *Atlas of Science: Visualizing What We Know*. Cambridge, MA: MIT Press.

Börner, Katy, Noshir S. Contractor, Hally J. Falk-Krzesinski, Stephen M. Fiore, Kara L. Hall, Joann Keyton, Bonnie Spring, Daniel Stokols, William Trochim, and Brian Uzzi. 2010. "A Multi-Level Systems Perspective for the Science of Team Science." *Science Translational Medicine* Vol. 2(49).

Börner, Katy, Weixia (Bonnie) Huang, Micah Linnemeier, Russell Jackson Duhon, Patrick Phillips, Nianli Ma, Angela Zoss, Hanning Guo, and Mark Price. 2010. "Rete-Netzwerk-Red: Analyzing and Visualizing Scholarly Networks Using the Network Workbench Tool." *Scientometrics* 83:863-876.

Börner, Katy, Soma Sanyal, and Alessandro Vespignani. 2007. "Network Science." Pp. 537-607 in *Annual Review of Information Science & Technology (ARIST)* vol. 41, edited by B. Cronin. Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.

Bornholdt, Stefan. 2003. *Handbook of Graphs and Networks From the Genome to the Internet*. Weinheim, Germany: Wiley-VCH

Boyack, Kevin W., Katy Börner, and Richard Klavans. 2009. "Mapping the Structure and Evolution of Chemistry Research." *Scientometrics* 79:45-60.

Burt, Ronald S. 2010. *Neighbor Networks: Competitive Advantage Local and Personal*. New York, NY: Oxford University Press.

Callon, M., J. P. Courtial, W. Turner, and S. Bauin. 1983. "From Translations to Problematic Networks: An Introduction to Co-Word Analysis." *Social Science Information* 22:191-235.

Carrington, P., J. Scott, and S. Wasserman. 2005. *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.

Christakis, Nicholas A. 2009. *Connected: The Surprising Power of Out Social Networks and How They Shape Our Lives*. New York, NY: Little, Brown and Co.

Easley, David and Jon Kleinberg. 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. New York, NY: Cambridge University Press.

Falk-Krzesinski, Holly J., Katy Börner, Noshir S. Contractor, Jonathon Cummings, Stephen M. Fiore, Kara L. Hall, Joann Keyton, Bonnie Spring, Daniel Stokols, William Trochim, and

Brian Uzzi. 2010. "Advancing the Science of Team Science." *Clinical and Translational Science* Vol. 3(5).

Freeman, Linton C. 2004. *The Development of Social Network Analysis: A Study in the Sociology of Science*. Vancouver, BC Canada: Empirical Press.

Kleinberg, J.M. 2002. "Bursty and Hierarchical Structure in Streams." Pp. 91-101 in *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: ACM Press.

Mane, Ketan K. and Katy Börner. 2004. "Mapping Topics and Topic Bursts in PNAS." *Proceedings of the National Academy of Science of the United States of America* 101:5183-5185.

Merton, Robert. 1973. *The Sociology of Science*. Chicago: University of Chicago Press.

Monge, Peter R. and Noshir S. Contractor. 2003. *Theories of Communication Networks*. New York: Oxford University Press.

Newman, M.E.J. 2010. *Networks: An Introduction*. New York, NY: Oxford University Press.

Newman, Mark, Albert-László Barabási, and Duncan J. Watts. 2006. *The Structure and Dynamics of Networks*. Princeton, NJ: Princeton University Press.

Nicolaisen, Jeppe. 2007. "Citation Analysis." Pp. 609-641 in *Annual Review of Information Science and Technology*, vol. 41, edited by B. Cronin. Medford, NJ: Information Today, Inc.

Nooy, W. D., Andrej Mrvar, and Vladimir Batagelj. 2005. *Exploratory Social Network Analysis with Pajek*. Cambridge, UK: Cambridge University Press.

Pastor-Satorras, Romualdo and Alessandro Vespignani. 2004. *Evolution and Structure of the Internet: A Statistical Physics Approach*. New York, NY: Cambridge University Press.

Salton, Gerard and C.S. Yang. 1973. "On the Specification of Term Values in Automatic Indexing." *Journal of Documentation* 29:351-372.

Scott, J. P. 2000. *Social Network Analysis: A Handbook*. London: Sage Publications.

Shiffrin, Richard M. and Katy Börner eds. 2004. "Mapping Knowledge Domains." *PNAS* 101.

Shneiderman, Ben. 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations." Pp. 336-343 in *Proceedings of the 1996 IEEE Symposium on Visual Languages (VL '96)*. Boulder, CO, September 3-6: IEEE Computer Society.

Thelwall, M., L. Vaughan, and L. Björneborn. 2005. "Webometrics." Pp. 179-255 in *Annual Review of Information Science and Technology*, vol. 39, edited by B. Cronin. Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.

Wasserman, Stanley and K. Faust. 1994. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences, 8)*: Cambridge University Press.

Watts, D.J. 1999. *Small Worlds*. Princeton, NJ: Princeton University Press.

Watts, Duncan. 2003. *Six Degrees: The Science of a Connected Age*. New York: W. W. Norton & Company.

Weingart, Scott, Hanning Guo, Katy Börner, Kevin W. Boyack, Micah Linnemeier, Russell J. Duhon, Patrick A. Phillips, Chintan Tank, and Joseph Biberstine. 2010, "Science of Science (Sci2) Tool User Manual" *Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, Bloomington,* Retrieved (http://sci.slis.indiana.edu/registration/docs/Sci2_Tutorial.pdf).

White, Howard D. and Katherine W. McCain. 1989. "Bibliometrics." Pp. 119-186 in *Annual Review of Information Science and Technology*, vol. 24, edited by M. E. Williams. Amsterdam: Elsevier.

Wilson, Concepcion S. 1999. "Informetrics." *Annual Review of Information Science & Technology (ARIST)* 107-247.

Wouters, Paul. 1999. "The Citation Culture." Doctoral Thesis, University of Amsterdam, Amsterdam.

Biography

KATY BÖRNER is the Victor H. Yngve Professor of Information Science at the School of Library and Information Science, Adjunct Professor at the School of Informatics and Computing, Adjunct Professor at the Department of Statistics in the College of Arts and Sciences, Core Faculty of Cognitive Science, Research Affiliate of the Biocomplexity Institute, Fellow of the Center for Research on Learning and Technology, Member of the Advanced Visualization Laboratory, and Founding Director of the Cyberinfrastructure for Network Science Center (http://cns.slis.indiana.edu) at Indiana University. She is a curator of the Places & Spaces: Mapping Science exhibit (http://scimaps.org). Her research focuses on the development of data analysis and visualization techniques for information access, understanding, and management. She is particularly interested in the study of the structure and evolution of scientific disciplines; the analysis and visualization of online activity; and the development of cyberinfrastructures for large scale scientific collaboration and computation. She is the co-editor of the Springer book on 'Visual Interfaces to Digital Libraries' and of a special issue of PNAS on 'Mapping Knowledge Domains' (2004). Her new book 'Atlas of Science: Guiding the Navigation and Management of Scholarly Knowledge' published by MIT Press will become available in 2010. She holds a MS in Electrical Engineering from the University of Technology in Leipzig, 1991 and a Ph.D. in Computer Science from the University of Kaiserslautern, 1997.