CHAPTER 7

Gap Analytics

On September 27, 2013, the Office of the Director of National Intelligence of the Intelligence Advanced Research Projects Activity (IARPA) issued a request for information (RFI).[1] The RFI is seeking information on methods for analyzing and forecasting trends and milestones in science and technology.

The RFI provides a list of illustrative examples of the types of questions they would like to address:

1. "Which companies lead the world in organic light-emitting diode (OLED) manufacturing?"
2. "What is the probability of a 10 cm carbon nanotube being fabricated before December 31, 2014?"
3. "Among abstracts accepted for the 2015 International Conference on Machine Learning (ICML) conference, will the number containing the term 'deep learning' exceed the number containing the term 'support vector machine(s)'?"
4. "How many unique assignees will have at least two USPTO patent applications published using the term 'Type III Secretion System' in their title/abstract/background/claims between October 1, 2013 and September 30, 2014?"
5. "By December 31, 2017, how many FDA-approved products will be based on RNA interference?"
6. "Will there be reported shortages of technetium-99 m in the United States in 2015?"

Let's pick question 2 and consider how we may address this type of question.

The key to addressing this question is to be able to estimate the maximum length of a carbon nanotube at some point in the future. Similarly, questions may limit the scope to a particular country, such as the United States or China, impose an upper limit of the cost, or specify other types of constraints. In its simplest form, a good strategy would be to first find out the state of the art worldwide. Next, we would find out the timeline of how fast advances have been made in the past 5 or 10 years in

[1]http://www.iarpa.gov/RFI/rfi_sti.html

The Fitness of Information: Quantitative Assessments of Critical Evidence, First Edition. Chaomei Chen.

making longer carbon nanotubes. Then, we would need to identify existing challenges that prevent anyone from making a 10 cm carbon nanotube and what techniques are currently available to address individual challenges. Finally, if there is no known solution to an existing problem, how likely is it to find a feasible solution within the specified time frame? To answer the seemingly simple question 2, it appears that we need to be able to answer a series of questions. A capable method, therefore, will likely be an integration of multiple methods.

When I tried to type the query "carbon nanotube length record" on Google, someone else had already searched with the exact query. Perhaps many people are considering how to answer question 2. The first link returned by Google was dated July 10, 2013. It reported that Tsinghua University in China was able to grow a carbon nanotube over 50 cm long (Zhang et al., 2013), which seems that its length has already met the requirement in question 2. We have two options. One is to answer question 2 with a probability of 1 that a 10 cm carbon nanotube can be made by the end of 2014 because there is at least one piece of evidence to support the prediction. The other option is to be more cautious and do more research before answering the question. Is the record-length carbon nanotube reproducible? Is it reproducible by researchers in other countries? What we have learned from OPERA's over-the-speed-of-light finding at the beginning of this book is that we need to ask if we are seeing the tip of an iceberg.

Searching for the world record of the length of a carbon nanotube through unstructured text is not straightforward. Searching for the world record within a specific time period is even harder. Question 2 has revealed several profound challenges. Similar problems may be found in electronic health-care records and how clinical trials specify inclusion and exclusion criteria in freely flowing natural languages. The need to assess the state of the art and track the gap between where we are and where we want to be is a common core in portfolio analyses.

7.1 PORTFOLIO ANALYSIS AND RISK ASSESSMENT

Portfolio analysis has a critical role in strategic planning, risk assessment, policy making, and performance evaluation. It is concerned with a broad spectrum of scientific and technological domains. The primary goal of a portfolio analysis is to assess the performance of a unit of interest, such as an individual, an organization, or a discipline, and identify its strengths and weaknesses regarding a baseline such that strategic adjustments can be made accordingly. Obtaining a holistic picture of the unit of analysis as a complex adaptive system is therefore of profound significance. Methodologies of portfolio analysis can be applied to a wide range of application domains, including gap analysis, situation awareness, competitive intelligence, and research evaluation and assessments.

An important characteristic of a complex adaptive system is that the whole is usually greater than the sum of its parts. In addition to studying individual components of such a system, it is essential to study how individual components are interrelated and how such interrelationships change over time in response to external events and internal perturbations. To be able to cover the structure and dynamics both at the component

level and at the system level, analysts face a tremendous challenge to associate patterns identified at one level with patterns identified at another level. In this chapter, we will focus on how this issue can be addressed in the context of portfolio analysis of publications produced by a unit of interest, including individual scientists, university colleges, research institutes, funding organizations, and scientific fields.

The notion of global science maps and local science maps has been seen in the literature, especially in information science and information visualization. Global science maps, for example, focus on depicting interrelationships of disciplines, whereas local science maps often focus on a specific field of study or a specialty. These existing approaches to the global and local science mapping are limited in terms of the types of organizing frameworks that can be offered to accommodate a portfolio analysis. A typical use of a global science map is to provide a base map over which a layer of additional information, or an overlay, can be superimposed. While existing solutions such as interactive overlays can provide insightful findings of research groups, many potentially significant analytical tasks are not readily supported. For example, each instance of citation in a publication involves a source and a target. The source is the article that initiates the citation, whereas the target is the reference that is being cited. To our knowledge, no global map overlays explicitly depict sources and targets of citations simultaneously.

We will first review an example of a portfolio evaluation in the context of a funding program. Then, we will introduce interactive overlays on journal-based global maps of science. Furthermore, we will introduce interactive overlays on a dual-map design and follow by a few examples to demonstrate how this method can be applied to the analysis of a portfolio.

7.1.1 Portfolios of Grant Proposals

The NSF CISE/SBE Advisory Committee formed a Subcommittee on Discovery in Research Portfolios between 2009 and 2010. The subcommittee was charged with identifying and demonstrating techniques and tools that can facilitate the assessment and evaluation of grant proposals and award portfolios. The subcommittee was asked to identify tools and approaches that are most effective in deriving knowledge from a diverse range of data. The tools should enable program directors visualize, interact, and understand the knowledge derived from the data. Subcommittee members were asked to apply their research tools to structure, analyze, visualize, and interact with datasets provided by the NSF.

Grant proposals submitted to the NSF consist of a number of components, including a cover page, a one-page project summary, a project description up to 15 pages, a list of references, 2-page biographies of investigators, and budget information. The abstracts of awarded projects are publicly available on NSF's website. A set of proposals were selected and made available to the members of the subcommittee for a limited period of time, but reviews of proposals were not accessible. All the results discussed in this book regarding the proposal dataset have been approved by a specific clearance procedure, which was in place to safeguard the privacy and security of the proposal dataset.

We will focus on two types of questions. At the individual proposal level, the main questions are: What is a proposal about in a nutshell? How does one proposal differ from other proposals in terms of their nutshell representations? At the portfolio level, the questions focus on characteristics of a group of proposals. What are the computational indicators that may differentiate awarded and declined proposals? What are the indicators that may identify transformative proposals in a portfolio?

Identifying the Core Information of a Proposal

We make no assumption about the structure or content of text documents. Proposals are processed as unstructured text. We expect that the amount of core information in a proposal is likely to be more than a one-page summary but shorter than a 15-page full-length proposal. If this is true, then it would make sense to reduce the text in a proposal to its core information and still preserve the essence of the full proposal. It would be also reasonable to expect that the shortened representation is probably more coherent than the original full-length document.

One way to extract the core information from a full-length proposal is as follows. First, divide the full-length project description of the proposal into a series of passages of text (known as segments) so that each passage corresponds to an underlying theme or topic. Next, construct a network of these passages based on their similarities. Finally, select passages of high centrality scores in the network to represent the core information of the proposal.

Figure 7.1 illustrates this process. The full-text document was divided into a sequence of segments. The plot in the middle depicts similarities between adjacent segments using a sliding window. A network of segments was generated based on intersegment similarities. The segments in darker shade played a central role in the network. Both of them can be selected to represent the essence of the full proposal.

It is possible to divide text into segments based on the degree of cohesiveness. A passage of text can be partitioned in many ways when the number and sizes of the segments are considered. Given a particular partition, the internal cohesiveness of each segment can be measured by techniques such as latent semantic indexing. The optimal partition would be the one that maximizes cohesiveness of all individual segments. An alternative way to optimize the partition is to ensure that the internal cohesiveness of text within a segment is higher than between segments. The process is known as *text*

FIGURE 7.1 The procedure for identifying core passages of a full-length document.

segmentation. The basic assumption is that a text document usually represents a series of subtopics and it is possible to detect the boundary of a subtopic based on the change of text similarities. Marti Hearst developed a flexible text segmentation algorithm, which can be used for this purpose. Optimizing the parameters for the algorithm is crucial, but currently computationally optimizing the configuration does not appear to be available. Using interactive visualization is a practical solution. Users could explore different configurations and select a good one.

Hearst's text segmentation algorithm detects a shift of topic based on lexical differences of a set of *n*-grams of text. The value of *n* ranges from 10 to 200 units. We implemented this method and provided an interactive visualization of lexical differences to help users to find the optimal parameter combinations. *Window size* and *step size* are two parameters crucial in configuring the text segmentation process. Window size is the number of tokens, mainly terms excluding stop words, in a sequence of tokens, and step size is the number of sequences of token in a block that are used for block–block similarity comparison. The similarity between two blocks is measured by a normalized inner product: given two blocks, b_1 and b_2, each with *window-size* token sequence, where $b_1 = \{\text{token-sequence}_{i-k}, \ldots, \text{token-sequence}_i\}$ and $b_2 = \{\text{token-sequence}_{i+1}, \ldots, \text{token-sequence}_{i+k+1}\}$. Our test with a small number of texts found that window size = 100 and step size = 20 yielded the best results in narrative texts similar to the NSF proposals.

The next step is to select the most representative segments as the core information of a proposal. Once text segments are identified, the similarity between any two segments can be calculated, including vector space models, latent semantic indexing, probabilistic models, or more recent topic models. A network of segments can be constructed based on intersegment similarities. Choosing the most representative block of text is equivalent to choosing the segments with high centrality measures in the network. A central topic is expected to be highly connected with other topics in the same proposal. Metrics such as PageRank are able to rank segments in the order of such centrality. One may choose one or multiple top-ranked segments to represent the proposal for any subsequent text analysis, clustering, or visualization. Our study evaluated candidate ranking metrics against our own proposals and found that segments selected by PageRank are more meaningful than other options.

Information Extraction

Information extraction is process unstructured text and select words that are most relevant to our interest. Natural language processing (NLP) techniques are commonly used in information extraction. It is generally believed that noun phrases are more meaningful and interpretable than single words. In the extraction step, noun phrases are extracted from core segments identified in a proposal.

Noun phrases consist of multiple words with a noun as the final word (known as the head noun). For example, the head noun in the noun phrase *supermassive black hole* is black hole. A *term* may be a single word or multiple words, but does not necessarily contain nouns. Sometimes terms are also referred to as *n*-grams, with *n* as the number of component words. Noun phrases are considered in general better lexical units to represent concepts than words and terms because noun phrases tend to be more meaningful and self-contained.

In order to identify noun phrases in unstructured text, the first step is to tag the type of each word in a text passage, including nouns, verbs, and adjectives. This step is called part-of-speech (POS) tagging. NLP tools are available to perform POS tagging and process tagged text, for example, the GATE system and the Stanford NLP Toolkit. Since NLP tools tend to be built with particular training text, the quality of tagging varies from target datasets. However, we found that using regular expression is the most flexible, customizable, and extensible approach. We experimented with several types of noun phrases in terms of the number of nouns because a general form of a noun phrase is word–word–word–noun and it is possible that the words are also nouns themselves, for example, word–word–noun–noun as in *rapidly increased climate change*. We allow users to filter noun phrases in terms of the number of nouns in a phrase. We tested analytical and statistical results across noun phrases with different word counts.

Detecting Hot Topics

Hot topics are defined in terms of the frequency of noun phrases found in project descriptions, project summaries, or other sources of text. Generally speaking, high-frequency noun phrases are regarded as indicators of a possible hot topic. The most valuable information about a hot topic is since when it becomes hot and how long it will last.

Burst detection determines whether the frequency of an observed event is substantially increased over time with respect to other events of the same type. The types of events are generic, including the appearance of a keyword in newspapers over a period of 12 months and the citations to a particular reference in papers published in the past 10 years. The data mining and knowledge discovery community has developed several burst detection algorithms. There are many techniques for detecting the emergence of a hot topic, notably Kleinberg's burst detection algorithm.

Two temporal properties of the burst of a noun phrase are the waiting time to burst and the duration of burst. The waiting time to burst is how much time has elapsed between the initial appearance of a noun phrase in a set of proposals and when a burst is detected by the algorithm. The duration of a burst is the time elapses between the beginning of the burst until either the burst is over or the end of the time frame of the analysis is reached. These properties are used in a survival analysis to differentiate awarded and declined proposals. Since these properties are domain independent, this method is applicable to a wide range of domains.

Identifying Potentially Transformative Proposals

We envisage that transformative research should be computationally detectable along two dimensions: *synthesis distance* and *structural divergence*. The synthesis distance characterizes a particular scientific contribution in terms of the conceptual distance between component topics it synthesizes and integrates. It is harder to conceive a long-range synthesis than a short-range one, but a synthesis over a long distance is likely to have a higher level of novelty. The structural divergence measures the extent to which a particular scientific contribution departs from the state of the art. As illustrated in Figure 7.2, groundbreaking ideas are likely to have a distant synthesis distance and a large structural divergence.

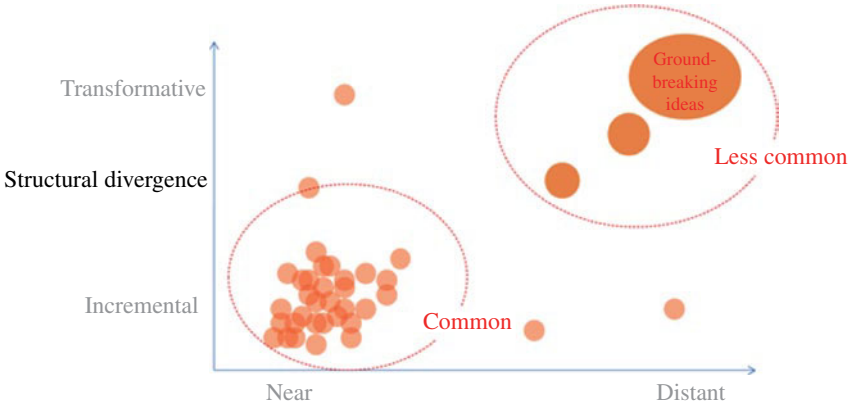

FIGURE 7.2 An illustration of the distribution of transformative research along two dimensions.

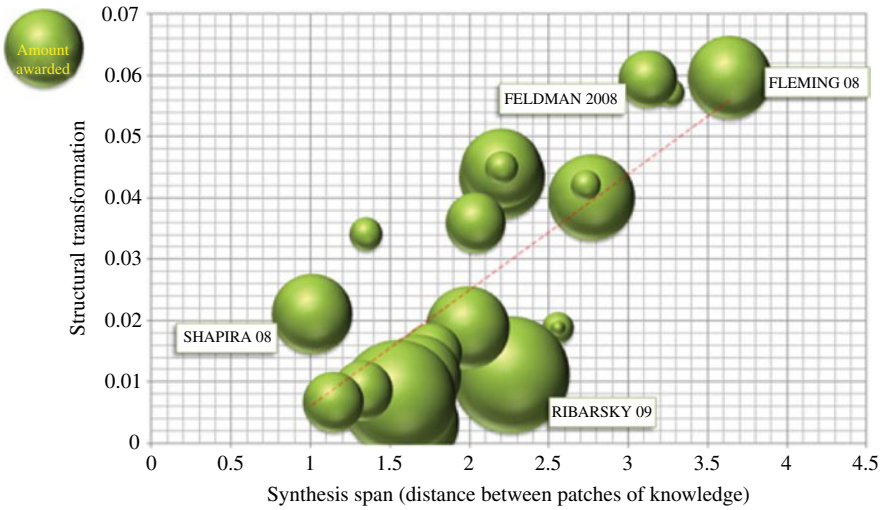

FIGURE 7.3 Transformative potentials of awards. Source: Publicly available NSF award abstracts of the SciSIP program.

Figure 7.3 shows the results of an analysis of proposals awarded in the NSF SciSIP program. Each proposal was represented by its publicly available abstract. The position of an award is determined by its synthesis distance and structural divergence scores. The size of each award represents the amount awarded. Further details of the four awards annotated in Figure 7.3 are listed in Table 7.1, including investigators' names, the year of the award, and the title of the project.

We also analyzed 200 proposals (100 awarded and 100 declined) randomly sampled from 7345 proposals of a different NSF program. The core information of each proposal was represented by three core segments with the highest PageRank scores.

Table 7.1 Awards labeled in Figure 7.3

Investigator	Year	Title
Lee Fleming	2008	DAT: Creating a Patent Collaboration Network Database to Examine the Social Production of Knowledge
Feldman Maryann	2008	State Science Policies: Modeling Their Origins Nature Fit and Effects on Local Universities
Martin Ribarsky	2009	DAT: A Visual Analytics Approach to Science and Innovation Policy
Philip Shapira	2008	MOD Measurement and Analysis of Highly Creative Research in the United States and Europe

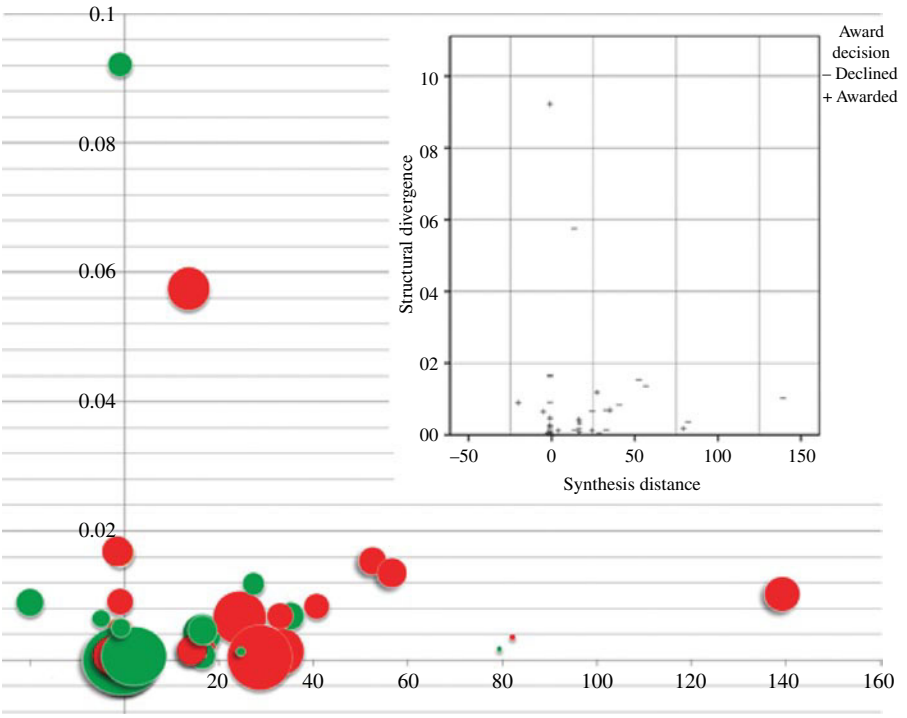

FIGURE 7.4 Transformative potentials of proposals (+: awarded; –: declined). Size = the amount requested. Data Source: 200 randomly sampled proposals.

Noun phrases of one to four nouns were extracted by CiteSpace to estimate the transformative potential properties for each proposal. A total of 141 proposals (70.5%) were found to have positive readings on the chart (see Figure 7.4).

A lesson learned from the analysis of grant proposals was that a broader context is needed to characterize not only the position of an idea at a particular time, but also the trajectory of where it has been. In the next section, we will turn to the design of

interactive overlays on a global map of scientific literature and demonstrate how to analyze a portfolio in terms of the movement of an underlying organization.

7.2 INTERACTIVE OVERLAYS

A commonly used strategy in cartography is to superimpose a thematic layer over a geographic map. For instance, store locators typically display the locations of their stores on a road map or a satellite map. The base map usually refers to the generic geographic map, whereas a thematic overlay refers to a layer of a special type of geographically distributed information. Multiple thematic overlays can be used to display related or nonrelated information. For example, tourists in a city may find it useful to look at an overlay of restaurants and an overlay of ATMs simultaneously. A major advantage of the overlay strategy is that we are familiar with how a map should be used, which tends to reduce the perceived complexity substantially. Sometimes this advantage may become crucial because we tend to be put off by something that appears too complicated. If we can bear with the initial impression of the complexity, then we are more likely to stay and start to learn something new.

7.2.1 Single-Map Overlays

Interactive overlays have been used with a variety of nongeographical base maps. The advantage of having a well-understood base map is still valid. Once we have learned how information is organized in a nongeographical map, we are able to focus on what information in an overlay brings to us and how it fits with our existing understanding of the base map.

Researchers have been developing ways to generate global science maps and use them to facilitate the analysis of issues concerning interrelated disciplines and the interdisciplinarity of a research program. Ismael Rafols, Loet Leydesdorff, and Alan Porter have been studying interdisciplinary research, especially topics that have profound societal challenges such as climate change and the diabetes pandemic. Addressing such societal challenges requires communications and incorporations of different bodies of knowledge, both from disparate parts of academia and from social stakeholders.

Interdisciplinary research involves a great deal of cognitive diversity. Rafols, Porter, and Leydesdorff (2010) developed a method based on overlaying science maps to study interdisciplinary research. The overlay method has two steps: (1) creating a global map of science as the base map and (2) superimposing a layer of a set of publications, for example, from a given institution or on a chosen topic. They have developed a toolkit and made it available for everyone to use.[2] A collection of interactive science overlay maps are also available on the Web.[3] With these interactive maps, one can explore how publications of an organization are distributed in various

[2] http://www.leydesdorff.net/overlaytoolkit
[3] http://idr.gatech.edu/maps.php

disciplines. These maps have been used to study the interdisciplinarity of research (Porter & Rafols, 2009), to compare research outputs of universities and large corporations (Rafols et al., 2010), and to trace the diffusion of research topics over science (Leydesdorff & Rafols, 2011).

The interactive overlay maps created by Rafols, Porter, and Leydesdorff are based on a single base map of scientific journals. Their base maps represent a network of journals based on their citing patterns or based on their cited patterns. In each view, there is only one representation of the global structure of scientific knowledge. This design has some limitations. The major limitation is the separation of citing and cited aspects of a citation. Each overlay can only represent one aspect of a citation, which does not seem to be feasible to address questions such as the following:

- Given a distribution of a set of articles on a base map of citing journals, what is their distribution on the *counterpart* cited journal base map?
- Which journals were cited by articles published in a particular journal?
- How do two organizations differ in terms of where they publish *and* what they cite?

In order to address these questions and other questions central to portfolio analysis, we have introduced a novel design that uses dual-map overlays. Dual-map overlays can address all the questions that single-map overlays can answer. In addition, dual-map overlays provide insights that would not be possible with a single-map overlay.

7.2.2 Dual-Map Overlays

Many citation-based maps are designed to show either the sources or the targets of citations in a single display but not both. A primary concern is that representing a mixture of citing and cited patterns simultaneously may considerably increase the complexity for users. HistCite, for example, represents the citation relationship of articles as a graph. A node in a graph cites nodes in previous years, whereas the node itself may be cited by nodes in the following year. The structure of such a graph is determined by the set of articles being considered. Understanding the structure of one dataset does not help us much to understand the structure of a different dataset. Furthermore, representing all the citation links in such graphs may suffer from a reduced clarity and make it harder for us to accomplish our analytic tasks. Although it is conceivable that a combined structure may be desirable in situations such as a heated debate, researchers are in general more concerned with differentiating various arguments before considering how to combine them.

The Butterfly system designed in the mid-1990s by Jock Mackinlay and his colleagues at Xerox displayed both citing and cited information in the same view, but each view was centered around a single article instead of presenting a holistic view of a collection of articles or journals (Mackinlay, Rao, & Card, 1995). Eugene Garfield's HistCite depicted direct citations in the literature. However, as the number of citations increase, the network tends to become cluttered, which is a common problem to network representations (Garfield, Pudovkin, & Istomin, 2003; Lucio-Arias & Leydesdorff, 2008).

We will introduce a novel design that uses dual-map interactive overlays to reveal additional insights into the structure and dynamics of citation patterns. The dual-map overlay design has several advantages over a single-map overlay. First, it represents the entirety of a citation instance. One can see where a citation originates and where it leads to in a single noninterrupted view. Second, it is easier to compare patterns of citation links than patterns of distributed dots. Third, it provides a framework for a new type of portfolio analysis with reference to a variety of evidence that characterize the movement of multiple populations across an adaptive landscape of scientific knowledge. Fourth, it opens up more research questions that can be addressed in new ways of analysis. For example, it becomes possible to study the interdisciplinarity at both source and target sides. It becomes possible to track the movements of scientific frontiers in terms of their footprints in both base maps.

The new design resembles the metaphor of fitness landscapes (Wright, 1932) in many ways. We can naturally introduce the notion of a trajectory of a collection of scientific publications, a set of patentable ideas, or a series of decisions made by the Supreme Court. A trajectory of a set of publications can be computed at the level of journals or at a disciplinary level.

The entire set of journals can be partitioned into groupings of journals, either by citing and cited patterns. These groupings may represent disciplines or research fields, depending on the granularity chosen for the partition process.

We used Blondel et al.'s algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) to partition the networks of journals in examples described in this chapter. We refer to such clusters as Blondel clusters. Ludo Waltman and Nees Jan van Eck recently announced an algorithm, called the smart local moving (SLM) algorithm, for community detection in large networks. The algorithm maximizes the modularity of a network to find the best partition. According to Waltman and van Eck, the SLM algorithm has been successfully applied to networks with tens of millions of nodes and hundreds of millions of edges (Waltman & van Eck, 2013). It would be a strong candidate for this type of algorithms.

A trajectory of an agent is the path that the agent has traversed in a space of latent variables, which could be a high-dimensional space as well as a 2- or 3-dimensional space. The agent could be an individual, a subpopulation (e.g., an organization), or the entire population. A trajectory *path* in a dual-map visualization is a function defined on a subset of the space × time domain to a subset of the fitness surface: path($x(t)$) = ($x(t)$, fitness($x(t)$)). At a given time t, $x(t)$ is the position of the agent on the base map, thus fitness($x(t)$) represents the fitness value of the agent at time t. The point ($x(t)$, fitness($x(t)$)) is a location on the fitness landscape that is occupied by the agent at time t.

The unit of time t can be year, month, hour, or minute. In our case, the unit is either year or month because of the resolution of the data. Each overlay layer represents a set of publication. As we will see shortly, such a set defines a portfolio of publications, depending on how the dataset is constructed. An overlay dataset D consists of n articles a_i, $i = 1, 2, \ldots, n$. Each article a_i appears in a journal b_j. The positions of the journal on the two base maps are citing(b_j) and cited(b_j), respectively. All the articles in D that are published in year t form $D(t)$, which is a subset of D. The positions of $D(t)$ on the base maps are the weight centers of {citing(b_j)} and {cited(b_j)}. The use of a weight center

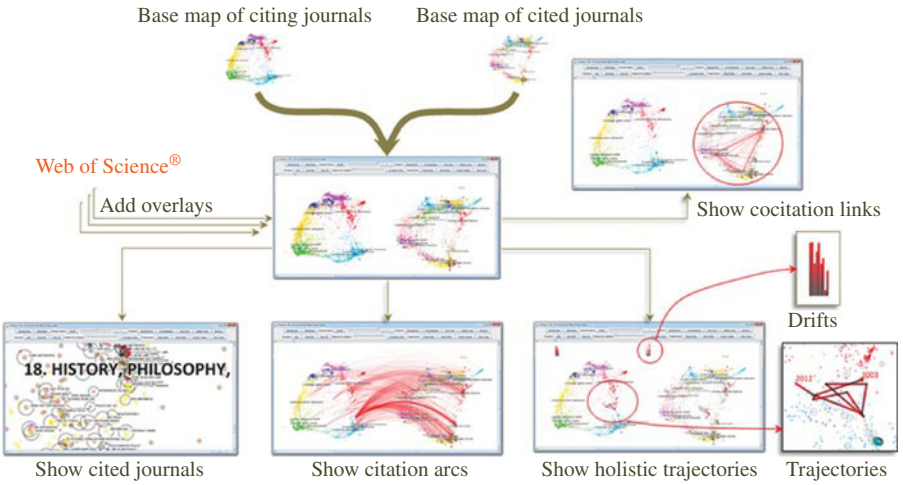

FIGURE 7.5 An overview illustrates the construction and use of dual-map overlays. Citation arcs, cocitation links, and trajectories over time facilitate the study of multiple sets of publications at an interdisciplinary level, an organizational level, and the individual publication level.

is found in the literature in aggregating the information from multiple points, for example, the computation of a Barycenter (Jin & Rousseau, 2001).

The construction of a dual-map base shares the initial steps of interactive overlay maps but differs in later steps. Once the coordinates are available for both citing and cited matrices of journals, a dual-map overlay can be constructed. It is not necessary to have cluster information, but additional functions are possible if cluster information is available. We assume that at least one set of clusters are available for each matrix. In this example, clusters are obtained by applying the Blondel clustering algorithm. Figure 7.5 shows an overview of the new method for a dual-map based portfolio analysis of scientific publications. The method is extensible to other types of global base maps, but here we will limit our descriptions to base maps generated from JCR journals. Details of the base map generation can be found in (Leydesdorff, Rafols, & Chen, 2013).

In the context of scientific publications, each member of a portfolio is a source article, also known as a citer, or a citing article. The journal in which a source article is published is called a citing journal. A reference cited by a source article is called a cited article, or a target article of an instance of citation, which may or may not be a source article in its own right. The journal in which a reference is published is called a cited journal. References cited by the same source article are called cocited references. The publication date of a source article can be identified either as the year only or the year and the month of the publication. The publication date of a cited reference is the year in which it is published. A portfolio represents the output of a research unit, whereas the references it cites as a whole represent the knowledge base on which the research activity is built.

A portfolio is a set of publications of interest. A portfolio can be constructed in a wide variety of ways. Portfolios are commonly defined by a common basis of authorship.

For example, an individual scientist's portfolio consists of all the publications authored or coauthored by the same author. A university's portfolio consists of all the publications by members of the same university. A country's portfolio similarly consists of all the publications by authors in the same country. On the other hand, a set of publications can be collected such that they represent the output of scientists in a particular field of study as a whole. Similarity, one can compile a dataset that represents the activity of a discipline. We can also construct a portfolio with a seed publication and group all the publications related to the seed in one way or another. These arguments can be made for constructing portfolios of patents invented by NCI employees, or on carbon nanotubes or telescopes, and grant proposals funded by the same program.

Given a base map of citing journals and a base map of cited journals, the two base maps are presented in the same user interface. The source of an overlay is a set of bibliographic records retrieved from the Web of Science and stored in a file directory on a computer where the software runs. For each overlay, the user may designate a specific color to distinguish citation arcs that belong to different overlays (see the lower middle part of Figure 7.5). The color chosen by the user will be also used for the trajectories of the overlay. Each trajectory is depicted in a bar chart that shows the pace lengths of the moves made by the trajectory and a trajectory plotted on the two-dimensional base maps (shown in the lower right part of Figure 7.5). The starting time and the ending time of each trajectory are marked. Each segment of a trajectory points from the end with a darker color to the end with a brighter color. The circled area in the upper right part of Figure 7.5 shows cocitation links from an overlay.

Given an overlay, journals involved in the citing and cited base maps are marked with circles (as shown in the lower left part of Figure 7.5). All the journals on a base map are assigned to clusters obtained by the Blondel algorithm (Leydesdorff et al., 2013). Major clusters are labeled by terms chosen from the titles of journals in corresponding clusters. The label terms are selected by a log-likelihood ratio test algorithm implemented in CiteSpace (Chen, 2006; Chen, Ibekwe-SanJuan, & Hou, 2010). For example, the cluster in the lower left part of Figure 7.5 is labeled by terms such as history and philosophy.

Multiple overlays can be superimposed onto the dual-map base one by one. An existing overlay can be removed. The user may use a number of controls such as buttons and sliders to select various types of information to be displayed. The following examples used the same base maps with the Blondel clustering configuration. Each side of the dual-map base depicts over 10,000 journals.

Figure 7.6 shows an annotated user interface. It shows both the citing and cited base maps side by side. The citing base map of 10,330 citing journals is on the left and the cited one of 10,253 cited journals is on the right. Each dot is a journal. Its color denotes its Blondel cluster membership. Various controls are available, for example, switching between Blondel clusters and VOSviewer's clusters, switching the unit of time between yearly and monthly (YR≥MTH and MTH≥YR), and switching between the calculation of trajectories at the cluster or journal level (C≥J and J≥C). The link style at the upper right controls the style of citation links. Our current design provides two types of styles, namely, curves and arcs. The arc style depicts a citation link as a parabolic arc. The curve style depicts a citation link as a spline curve running

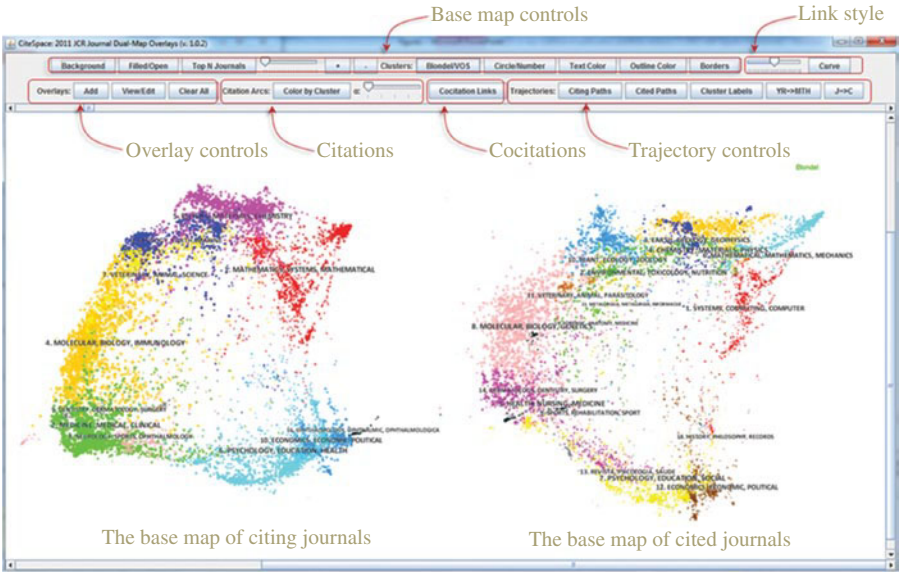

FIGURE 7.6 The initial appearance of the Dual-Map user interface, showing both citing and cited journal base maps simultaneously. The base map of 10,330 citing journals is on the left. The base map of 10,253 cited journals is on the right. The colors depict clusters identified by the Blondel clustering algorithm. (*See insert for color representation of the figure*.)

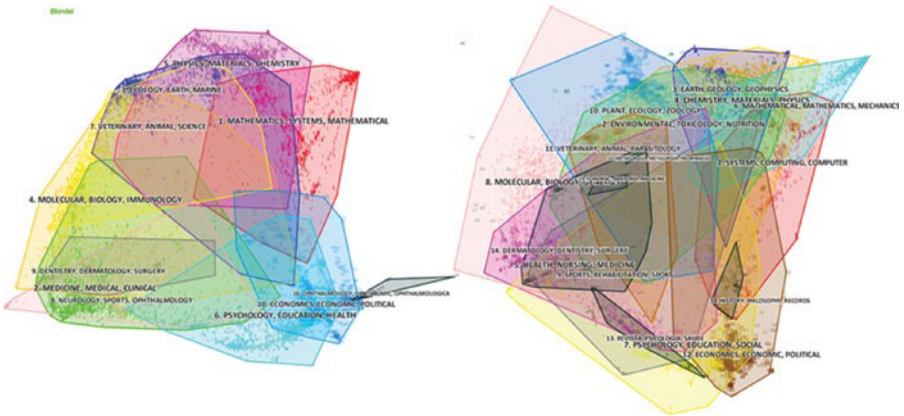

FIGURE 7.7 The boundary of each cluster is shown to depict how its members are distributed. Clusters in both base maps overlap substantially. (*See insert for color representation of the figure*.)

from the source journal to the target journal of the citation. The curve style is designed to improve the clarity of the visualization of a large number of citation links.

Clusters with less than five members are not shown labels. The label of a cluster is represented by terms selected from the titles of journals in the cluster. The label of a cluster is placed at the cluster centroid. As shown in Figure 7.7, the boundaries of

Blondel clusters in both base maps have considerable overlaps between multiple clusters. It is also clear that journals are not evenly distributed. Since cluster memberships are exclusive, reducing the amount of overlapping would be a preferable move if layout algorithms can effectively separate nodes in different clusters (Dwyer et al., 2008). However, a study of this issue is beyond the scope of this study. It should be noted that in this study we make no assumptions concerning the presence or absence of overlapping clusters.

7.3 EXAMPLES OF DUAL-MAP OVERLAYS

We demonstrate the use of dual-map overlays with examples of different types in terms of how a portfolio of publications is constructed. The first type is a *single-source overlay*, which represents portfolios that are generated with a single seed article, that is, a portfolio of this type consists of all the articles that cite the seed article. We will include one example of a single-source overlay on the topic of autism and vaccines. The second type is an *organizational overlay*, which represents a portfolio of an organization, including a department in a university, a corporate research lab, or a national laboratory. Examples of organizational overlays include publication portfolios of three iSchools in the United States and publication portfolios of three well-known corporations. The third type is *subject matter overlays*, which are defined by the relevance to an underlying subject matter. Examples of subject matters include regenerative medicine, mass extinctions, visual analytics, and articles that cited the *Journal of the American Society for Information Science and Technology* (*JASIST*).

These examples are representative of common types of scenarios in the context of portfolio analysis, and we will demonstrate prominent patterns revealed by the new method. For instance, the single-source overlay example is essentially originated from a single publication. The three corporations in the organizational overlay examples are widely known. Examples of subject matter overlays are topics we either have previously studied or are familiar with. The diversity of these examples illustrates the scope and flexibility of the new method.

7.3.1 Portfolios of a Single Source

In 1998, a paper by Wakefield et al. appeared in *Lancet* (Wakefield et al., 1998). It suggested a possible link between a combination of vaccines against measles, mumps, rubella, and autism. The study had drawn a considerable amount of attention from researchers and the general public. As a result, many parents in England decided to skip these vaccines for their children. In 2004, *Lancet* partially retracted the Wakefield paper. In 2010, the journal finally retracted the paper altogether. The Wakefield paper had been controversial for years prior to its retraction. The *Lancet*'s retraction notice in February 2010 noted that several elements of the 1998 paper were incorrect, contrary to the findings of an earlier investigation, and that the paper made false claims of an "approval" of the local ethics committee.

According to the Web of Science, the Wakefield paper is the most cited article that has been retracted. Its impact had evidently reached over 740 articles that cited the controversial paper. Some of the one-citation-away citing articles became highly cited themselves later on. For example, among the articles that cited Wakefield et al.'s study, one was cited by 384 articles and the other by 360 citations. Articles that cited Wakefield et al.'s article further amplified the influence of the later retracted study to an even broader context—the 740 one-citation-away articles were cited by more than 6600 articles in the Web of Science. These two-citation-away articles cited an even larger body of literature of over 12,000 references. The original paper's citation count peaked in 2002. A detailed analysis of citation contexts associated with retracted articles, including Wakefield article, can be found in our study of retracted scientific articles (Chen, Hu, Milbank, & Schultz, 2013).

A single-source overlay represents citation patterns of articles concerning a seed article. All the articles that cite the same seed articles are used to form the overlay. A seed article can be a groundbreaking article that represents a scientific breakthrough or a transformative discovery. A seed article could be a controversial or even retracted article of interest.

We explain one example of single-source overlays in detail here. The seed article in the example, Wakefield et al. (1998), is a highly cited retracted article, which has profound implications on public health, especially on vaccine uptakes from children.

Autism and Vaccines

We use the Wakefield paper as an example to illustrate various patterns that can be discerned from a dual-map overlay (Wakefield et al., 1998). The source of the overlay is a set of 405 articles that cited the Wakefield paper (this is a subset because of the limitation of our Web of Science subscription). Figure 7.8 shows the overlay with annotations to key points of interest. The bar charts near the top of the figure depict stepwise drifts in trajectories aggregated from the citing behavior of the 405 articles. The first bar on the left of the bar chart represents the amount of shift in 1999 with reference to the weight center of the disciplines involved in 1998. The chart shows that the distance of the shifts increased substantially between 2005–2006 and 2011–2012. Given that the Wakefield paper was partially retracted in 2004 and fully retracted in 2010, is it reasonable to hypothesize that a significant change in relevant scientific disciplines may attract new publications from new perspectives? New perspectives would result in publications in different journals.

Citations made by these source articles are shown as the spline waves, which are primarily in yellow, green, and cyan. Each spline curve starts from a citing journal in the base map on the left and points to a cited journal in the base map on the right. Labels in the vicinity of the launching areas indicate corresponding disciplines in which citing articles were published. Each label is centered at the cluster centroid of the corresponding journals. In this example, relevant disciplines include medicine, clinical, biology, immunology, psychology, education, and health on the citing side of the dual map. The majority of the citations were directed to disciplinary areas such as health, nursing, and medicine in the cited base map. Cocitation links that connect different disciplines can be displayed as dashed lines.

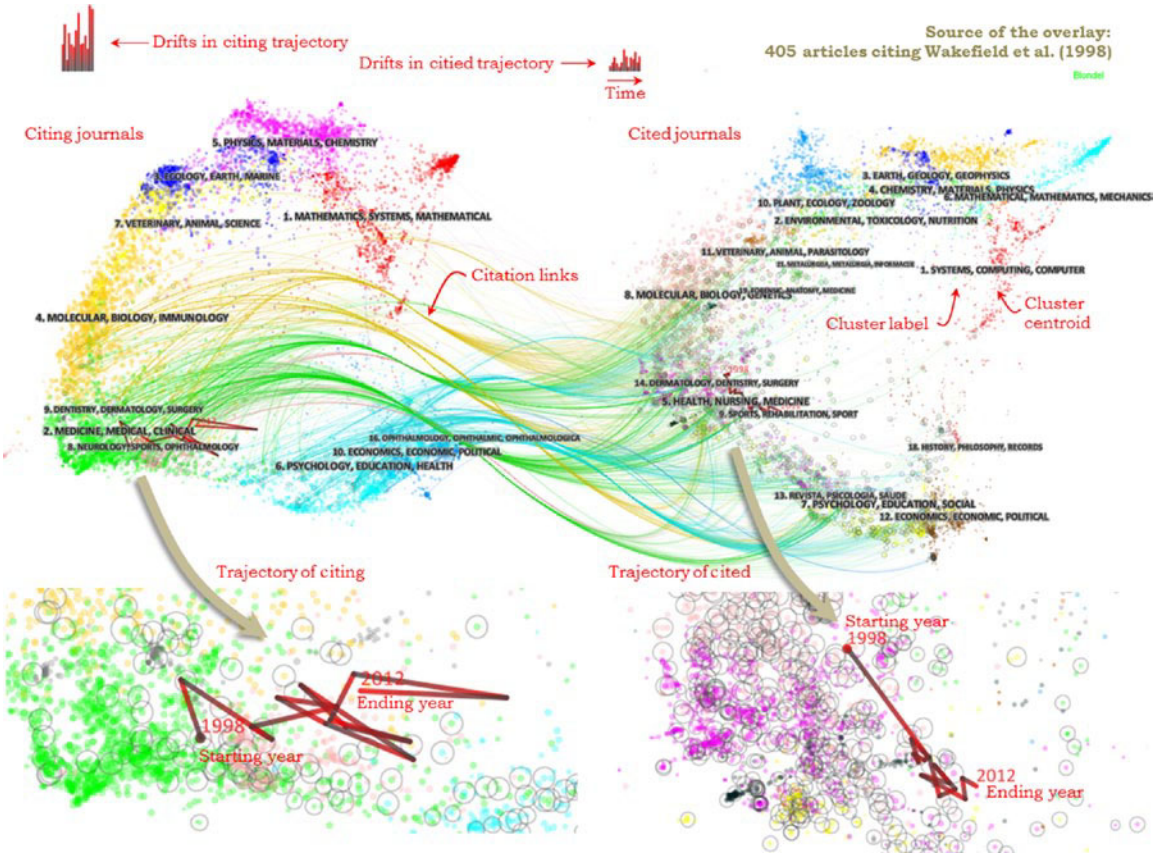

FIGURE 7.8 Citation patterns in an overlay of 405 articles that cited the Wakefield paper. (*See insert for color representation of the figure.*)

The lower half of the figure shows the trajectory of citing patterns on the left and the trajectory of cited patterns on the right. Properties of a citing trajectory can tell us about the dynamics of publications concerning the Wakefield paper at a disciplinary level. For example, if the citing trajectory shows a shift from one region to another on the base map of citing journals, we would know that there was a change of the primary disciplines in terms of relevant articles were published in a different set of journals. In this case, the citing trajectory is drifting toward the right-hand side of the citing base map. Based on the citation links shown in the upper half of the overlay map, the starting position of the citing trajectory is predominated by publications in the discipline of medicine and clinical medicine, whereas the ending position of the trajectory appears to be influenced by activities in areas near to the disciplines of psychology, education, and health.

The most active citing journal after 2004 is *Social Science Medicine*. For example, a paper published in 2005 in the journal reported an ethnographic study of the choice of vaccine (Poltorak, Leach, Fairhead, & Cassell, 2005): "In the context of the high-profile controversy that has unfolded in the UK around the measles, mumps and rubella (MMR) vaccine and its possible adverse effects, this paper explores how parents in Brighton, southern England, are thinking about MMR for their own children."

7.3.2 Portfolios of Organizations

The construction of an organizational overlay is based on a search in the address field in the Web of Science. For example, the portfolio of the College of Information Science and Technology[4] at Drexel University can be constructed by searching for bibliographic records that have the name of the college in the Address field. A portfolio of an individual researcher can be obtained by adding the author's name to the search.

Three iSchools

Publications with author affiliations involving one of the three iSchools in the United States are used as the source of three overlays, one for each iSchool. The window of analysis starts from 2003 and finishes at the end of 2012.

Three threads of citations stand out in Figure 7.9. The blue thread connects the cluster of mathematics and systems in the citing base map to the cluster of systems and computing in the cited base map. Representative journals of this thread include *Data and Knowledge Engineering*, *IEEE Computer Graphics and Applications*, and *IEEE Computer*. The two threads following the red lines are also prominent. The upper thread of the two essentially connects the library and information science in the citing journal base map to computing and information systems in the cited journal base map. Representative citing journals include *JASIST*, *Information Processing and Management*, and *Journal of Informetrics*. The lower thread of the two represents citations from journals such as *Journal of Computer-Mediated Communication*,

[4]The new name of the college is the College of Computing and Informatics after September 2013.

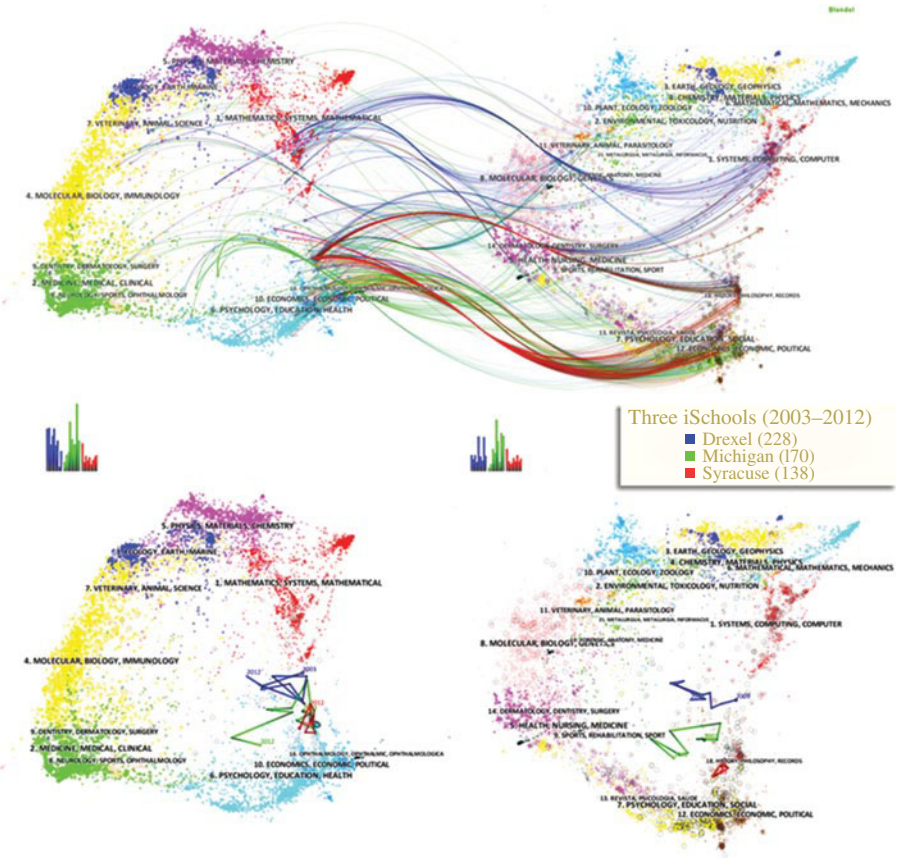

FIGURE 7.9 Overlays of three iSchools show major threads of citations that may characterize the publication portfolios of these institutions. The lower half of the figure shows the citing and cited trajectories in each of the base maps. (*See insert for color representation of the figure.*)

Computer Human Behavior, and *Government Information Quarterly* to the corresponding cluster in the cited base map. These disciplinary patterns provide useful insights into the nature of iSchools in terms of the core values perceived by the iSchools, namely, information, users, and technology.

The citing trajectories of the three iSchools have different patterns. The red trajectory had been concentrating on a small set of journals, whereas the blue and green trajectories had made long jumps across a larger group of journals. Their cited trajectories demonstrated similar patterns. In a portfolio analysis, one can investigate further to uncover the reasons that led to the differences at the global level and take actions accordingly. For example, if a trajectory with long jumps turns out to reflect the diverse expertise of new faculty members, then one may take this into account when considering hiring new faculty members. Multiple concentrations on the citing map may trace to the same source on the cited map. If that is the case, then authors

of the seemingly diverse concentrations may share more common grounds than they probably realized.

The examples we present here are intended to illustrate the new method of portfolio analysis rather than present the results of a portfolio analysis because the data was collected from the Web of Science only. For an actual portfolio analysis, it is advisable to construct a comprehensive portfolio from multiple sources.

Three Corporations

In the following example, portfolios of publications by authors from three corporations, Google, Microsoft, and IBM, were constructed from bibliographic records in the Web of Science. Publications for each corporation were identified based on the address field of these bibliographic records. In order to concentrate on original research outputs from these corporations, we retained publications of type Article only for the study. Other types such as Review or Note were excluded.

Google's portfolio consisted of 620 source articles, which appeared in 550 unique journals and cited 8724 journals. Microsoft's portfolio had 1,968 papers from 1,050 journals and cited 21,193 journals. IBM's portfolio is the largest of the three, containing 3,965 articles from 1,593 journals and cited 27,617 journals (see Table 7.2).

Each corporation was assigned a distinct color: Google (blue), Microsoft (red), and IBM (green). These colors were used in their trajectories, bar charts, citation arcs, and cocitation links. As shown in Figure 7.10, the trajectories of all three organizations were located in the upper right region. In particular, IBM's trajectory (green) appeared slightly higher up in both base maps, relatively closer to disciplines such as physics and mathematics. Microsoft's and Google's trajectories appeared slightly lower in the map, relatively closer to psychology and other humanity-related disciplines.

The middle row in Figure 7.10 shows the trajectories of the three corporations. Citing trajectories are shown in the image on the left. Cited trajectories are on the right. Overall, the citing trajectories of Google and Microsoft appeared in areas near to each other, whereas the trajectory of IBM was farther away from them. At the global level, these patterns suggest that Google and Microsoft are more similar than IBM in terms of where they publish. Their trajectories in the cited map were separated apart evenly, indicating that they built on distinct areas of prior knowledge for their work.

The bar charts shown at the bottom of Figure 7.10 represent the length of each move in their trajectories. A short bar indicates a near-range move. A tall bar indicates a long-range move. All the three corporations made near-range moves in their citing trajectories, suggesting that they published in a relatively stable set of journals over

Table 7.2 Portfolios of three organizations' publications during 2008 and 2012

Organization	Color	Articles	Citing journals	Cited journals
Google	Blue	620	550	8,724
Microsoft	Red	1968	1050	21,193
IBM	Green	3965	1593	27,617

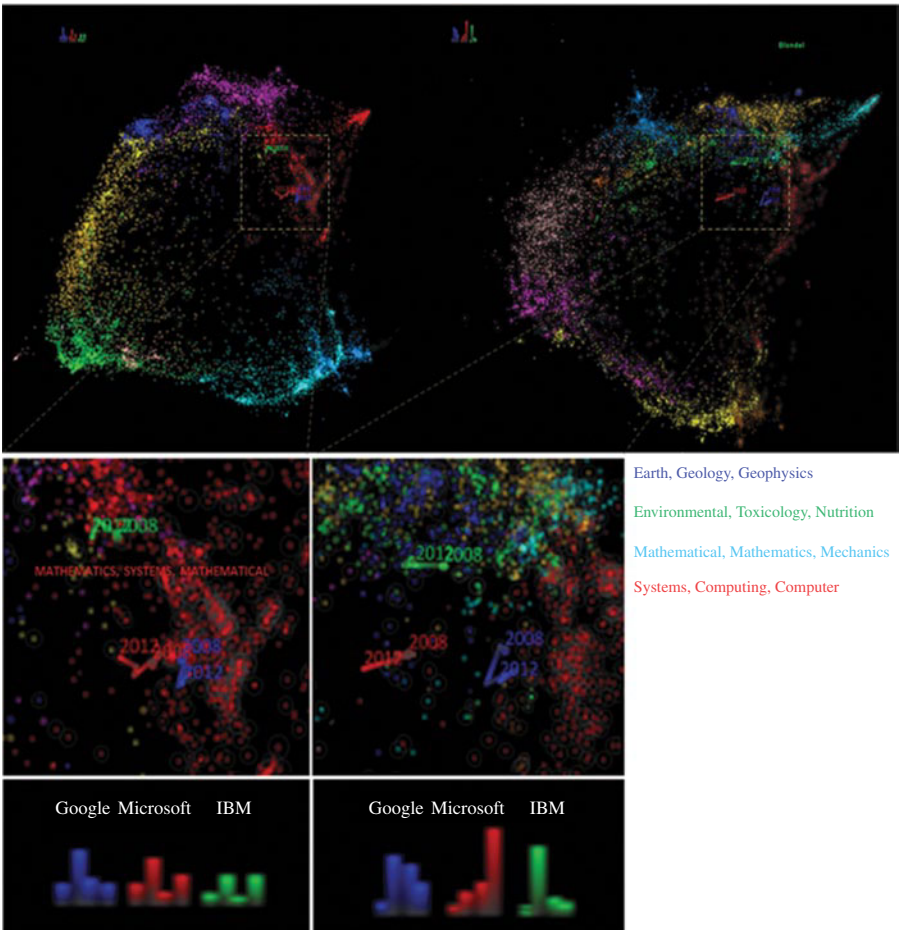

FIGURE 7.10 Trajectories of Google (blue), Microsoft (red), and IBM (green). (*See insert for color representation of the figure*.)

the years. In contrast, their cited trajectories revealed longer jumps. For example, Microsoft made a substantial shift in the most recent year and IBM made a large shift in the second year of the window, which means they had changed substantially what they chose to cite.

Figure 7.11 shows the citation overlays of the three organizations. On top of the figure, in the base map of citing journals on the left, citations made by Google mostly originated from the area labeled as mathematics and systems (not shown in the figure, but accessible interactively). The majority of the citation arcs led to the corresponding area of the same discipline in the base map of cited journals on the right. The overlay in red below Google's overlay is from Microsoft. In addition to the same citation passageway as Google's citations, Microsoft's citation arcs followed a wider range of citation passageways. For example, Microsoft's citations reached several areas

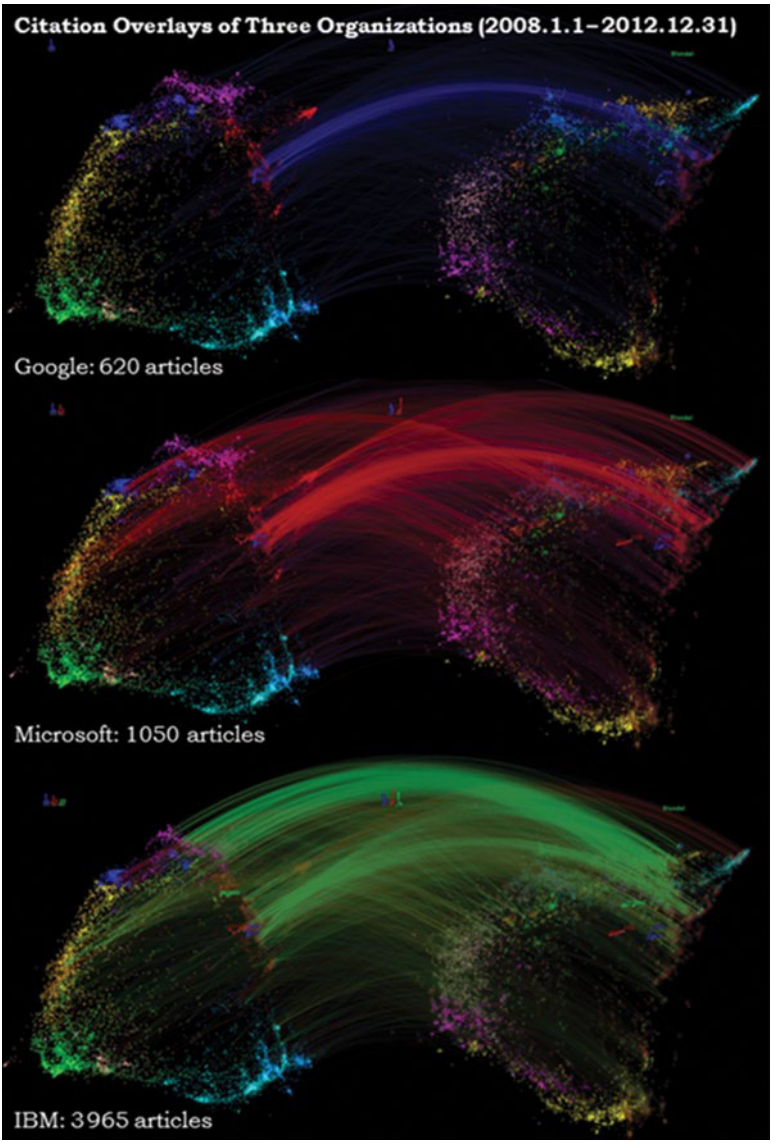

FIGURE 7.11 Citation overlays of three corporations. (*See insert for color representation of the figure*.)

located in the upper left part of the citing map, whereas these areas were not active in Google's portfolio. IBM's portfolio shows an even broader scope. The prominent trail of green arcs on the top of IBM's overlay chart highlighted some of the IBM's major competencies with more hardware-related areas. In contrast, this passageway was not prominent in the portfolio overlays of Google and Microsoft, which were more active in software-oriented areas.

With portfolio overlays and aggregated trajectories of organizations, one can quickly glean insightful patterns of these organizations. Furthermore, these patterns draw our attention to a subset of publications in a portfolio. We can then pursue more detailed information about publications associated with a particular pattern. Overlay and trajectory patterns at a macroscopic level provide a useful gateway to the study of the dynamics at both macroscopic and microscopic levels.

7.3.3 Portfolios of Subject Matters

A subject matter portfolio consists of publications relevant to the subject matter. Such portfolios can be constructed by a topic search in the Web of Science. A portfolio of research in regenerative medicine, for example, can be obtained by searching for bibliographic records relevant to "regenerative medicine" in the Web of Science.

Regenerative Medicine

Regenerative medicine is a rapidly growing area of research. It has many clinical implications and potentials. In a study published in 2012, we found that the topic of induced pluripotent stem cells (iPSCs) plays a leading role in regenerative medicine research (Chen, Hu, Liu, & Tseng, 2012). iPSCs research was awarded the 2012 Nobel Prize in Medicine. Figure 7.12 shows the trajectories of regenerative medicine. We updated the dataset with a new topic search for "regenerative medicine" between 2005 and 2012 in the Web of Science. A total of 3559 records found in this time frame were used to generate the overlay.

The bar charts of the trajectories, shown on the top of the figure, indicate that the trajectories are stable. The citing trajectory on the left closely tracked the disciplines along the disciplinary region labeled by terms such as molecular, biology,

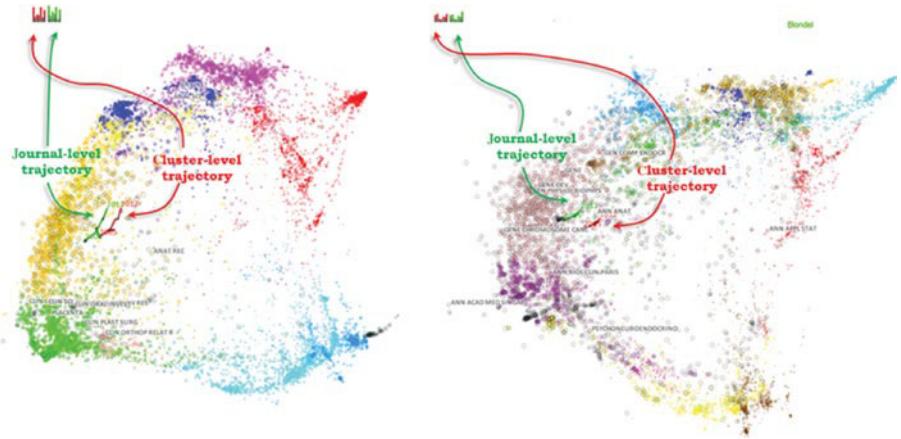

FIGURE 7.12 Trajectories of regenerative medicine research (2005–2012). The citing trajectory remains to be in the disciplinary area labeled as molecular, biology, and immunology throughout the entire course. (*See insert for color representation of the figure*.)

Table 7.3 Journals involved in the regenerative medicine dataset (2005–2012)

Year	2005	2006	2007	2008	2009	2010	2011	2012
Citing	70	81	144	192	222	265	325	319
Cited	1712	1912	2029	2321	2304	2379	2571	2618

and immunology. Throughout the entire course, the citing trajectory remains in the same discipline.

Table 7.3 lists the number of citing and cited journals per year in the regenerative medicine dataset. The few journals at both ends are likely to contribute to the instability of the trajectory.

Mass Extinctions (1975–2010)

Positions of trajectories in previous examples are calculated at the journal level. In this example, we calculate positions of trajectories at the disciplinary level. At a particular year, the positions of journals are matched to the cluster centroids of their corresponding clusters. Trajectories at the discipline level are expected to be more stable than trajectories at the journal level because many journal-to-journal movements within the same disciplinary cluster would be absorbed to a stable centroid of the same cluster.

Figure 7.13 shows the citing trajectory of mass extinctions research at the discipline level. The trajectory has a core discipline almost right at the center of the area labeled as ecology, earth, and marine. The trajectory spent most of the time in this area, except a long-range movement along the shape of a long triangle between 1977 and 1980. The longest distance it has moved was from the discipline labeled as physics, materials, and chemistry to the center of the region labeled as molecular, biology, and immunology. (See the schematic sketch on the top of Figure 7.13.) The long-distant move returned to the core of the trajectory next year. What specific papers caused the long-distant move? What kept the trajectory to such a compact core discipline for so many years? Studies of the structure and dynamics of specialties at a lower level of granularity are more appropriate to address this type of questions. For example, in a previous study of mass extinctions, we identified turning points in mass extinctions research (Chen, Cribbin, Macredie, & Morar, 2002). The overlay example here demonstrates how it may be integrated with the study of the dynamics of a specialty.

Visual Analytics (2006–2012)

The third example of a subject matter overlay is based on publications on visual analytics between 2006 and 2012. Figure 7.14 shows that the majority of visual analytics publications originated in the discipline of mathematics and computer science (threads in red originated from the red cluster in the citing base map on the left). The way in which visual analytics connects various disciplines was highlighted by cocitation links between disciplines—dashed lines connecting the centroids of clusters.

Cocitation links between clusters of cited journals show that visual analytics is primarily drawn upon the work in disciplines such as (1) computing and systems, (2) psychology and sociology, (3) economics and politics, and (4) plant, ecology, and zoology.

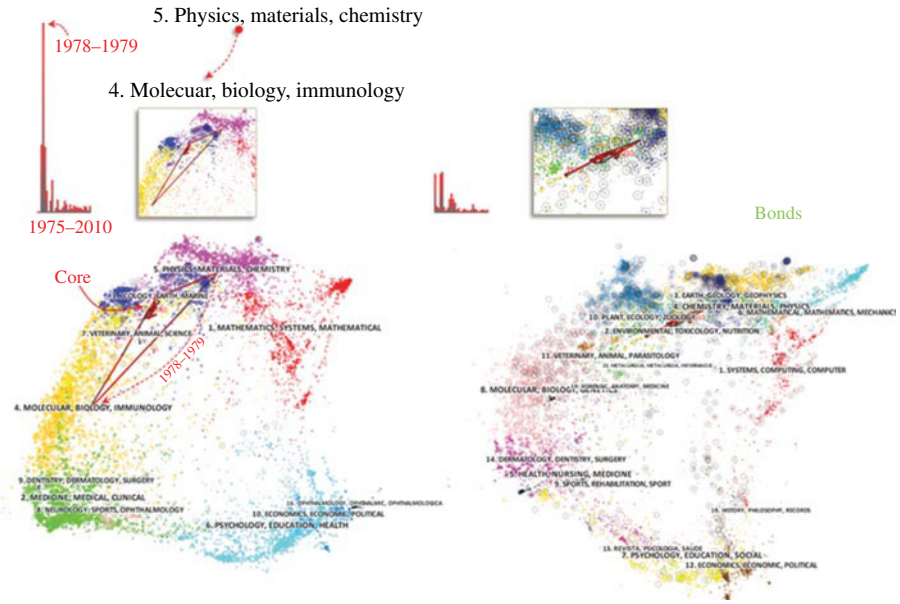

FIGURE 7.13 Trajectories of research in mass extinctions (1975–2010) at the discipline level. The core discipline of the research is identified as the Blondel cluster 3 on ecology, earth, and marine. The longest single-year shift occurred between 1978 and 1979 as the disciplinary center of the journals shifted from the Blondel cluster 5 on physics, materials, and chemistry to the Blondel cluster 4 on molecular, biology, and immunology. (*See insert for color representation of the figure*.)

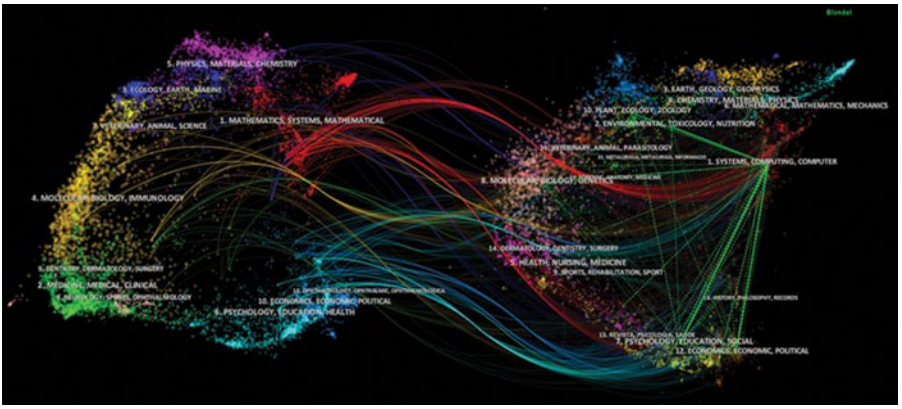

FIGURE 7.14 An overlay of publications in visual analytics (2006–2012). Wavelike curves depict citation links. They are colored by their source clusters. Dashed lines depict cocitation links across disciplinary boundaries. (*See insert for color representation of the figure*.)

Articles Citing JASIST Publications (2002–2011)

Articles that cite the *JASIST* between 2002 and 2011 were retrieved from the Web of Science (Table 7.4). An overlay was generated to reveal the impact of the journal. Figure 7.15 shows the same overlay in two different styles. The style used in the

Table 7.4 The number of journals involved in articles citing *JASIST*

	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Citing journals	8	40	61	100	131	179	250	336	374	439
Cited journals	58	289	655	999	1480	1763	2211	3176	3230	3711

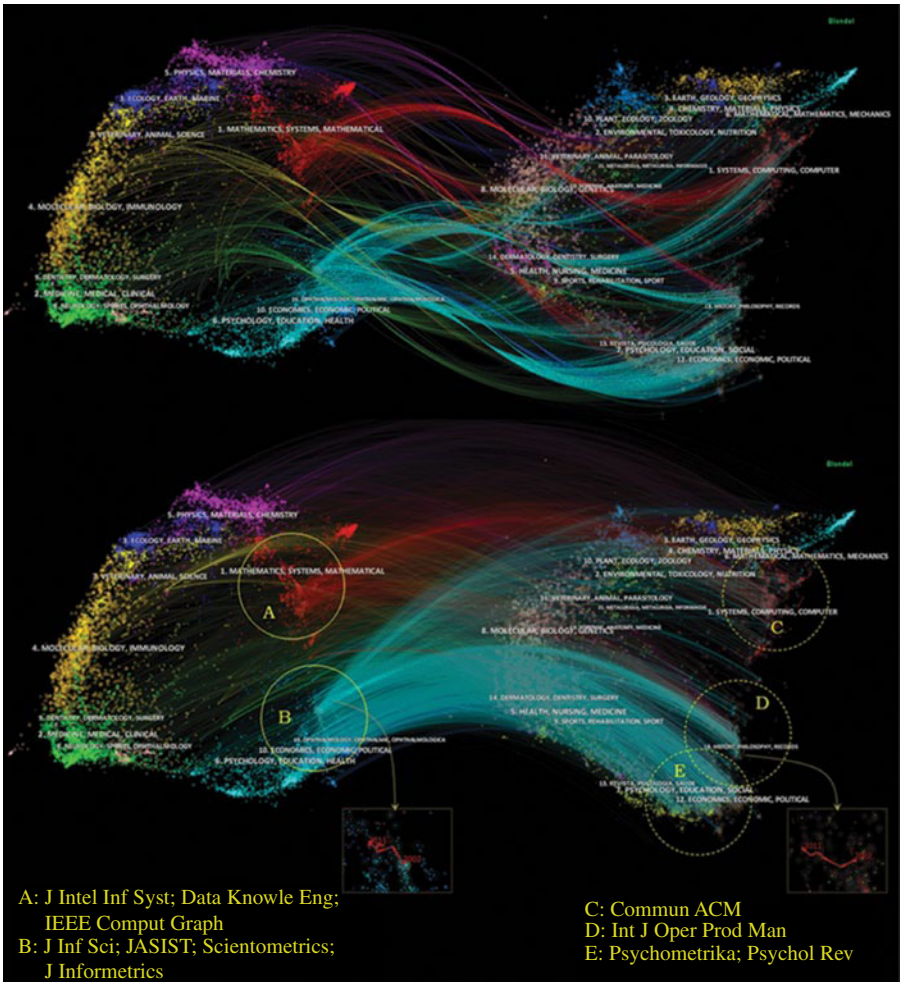

FIGURE 7.15 An overlay of articles citing *JASIST* (2002–2011). (*See insert for color representation of the figure.*)

upper overlay depicts citation links in spline curves, whereas the style used in the lower overlay shows citation links in arcs. Both styles of citation links are colored by the corresponding source clusters of journals.

The overlay map shows that two areas on the citing map are particularly active, namely, area A (computer science) and area B (information science). The most frequently

seen journals in area A include *Journal of Intelligent Information Systems*, *Data and Knowledge Engineering*, and *IEEE Computer Graphics and Applications*. The top journals in area B include *Journal of Informetrics*, *JASIST*, and *Scientometrics*. The citation arcs reveal three areas C, D, and E. The patterns revealed by citation arcs connecting disciplinary areas in the two base maps are straightforward to interpret once the user becomes familiar with the "geography" of the base maps.

7.3.4 Patterns in Trajectories

Trajectories visualized with the dual-map overlays enable the study of the dynamics of a complex system characterized by a portfolio as a whole. Some of the patterns may provide useful signs to guide more in-depth pursuits at lower levels of granularity. A trajectory is a sequence of disciplines or journals that are most representative of the collective publishing behavior of scientists. Trajectories can be formed at the level of individual journals and at the level of clusters of journals to represent disciplines. We will focus on the disciplinary trajectories in the rest of the section, although the metrics and interpretations can be translated to the journal-level trajectories.

The set of unique disciplines in a trajectory provides useful information to differentiate different trajectories. For example, two trajectories may involve the same set of disciplines. Some trajectories may essentially deal with one discipline, whereas other trajectories may deal with a large number of disciplines. Given the same set of disciplines, trajectories may differ in terms of the order these disciplines appear and how often each individual discipline appears.

We define the distance between two disciplines in terms of the distance between the centroids of the two disciplines. The centroid of a discipline is defined as the weight center of the journals that belong to the same Blondel cluster. The proximity of two disciplines on the base maps reflects the strength of the interdisciplinary connection between them because the base maps are built on citing and cited patterns of journals. Using a combination of the disciplines involved and the proximity of these disciplines, we can derive a number of patterns.

Core Disciplines—If a trajectory either repeatedly enters or persistently remains in the same discipline, it may represent the collective behavior of scientists who mostly publish in the same disciplines.

Multiple Disciplines with Variations of Proximity—A trajectory may involve multiple disciplines. These multiple disciplines may be traversed in a variety of order. The proximity of adjacent disciplines may change slowly or rapidly. For example, a rapid change of proximity indicates a long-range move or a shift. A small step in a trajectory represents a drift or a short-range move.

Closed and Open—A closed trajectory may drift and/or shift back and forth multiple times between disciplines in a small group, whereas in an open trajectory, the number of disciplines involved is constantly growing. Since the total number of Blondel clusters in our study is fixed, an open trajectory is relatively speaking. If a scientific breakthrough is made and it has a broad impact on other disciplines, we would expect that the number of disciplines involved will increase rapidly in the following years. On the other hand, the trajectory of a research program addressing a highly unique problem may be limited to a small number of disciplines and show a closed trajectory.

We illustrate some of these patterns with disciplinary trajectories of two productive and high-impact scholars, Edward Witten, a physicist, and Ben Shneiderman, a computer scientist. Both of them have published hundreds of articles in their fields. According to the Web of Science, Witten has 228 publications indexed by the Science Citation Index (SCI) with an *h*-index of 115. The 228 publications have been cited by 37,523 records without self-citations. His average citation per item is 279.50. Shneiderman has 156 publications in the Web of Science with an h-index of 30. The 156 publications have been cited by 2489 records with an average of 18.67 citations per item. Note that these statistics are by no means comprehensive and they are likely to be biased by the source of data. For example, Google Citation Profile lists of an h-index of 89 and 46,324 citations for Shneiderman, whereas Witten does not have set up his Google Citation Profile.

We retrieved bibliographic records of the type article from the Web of Science for each of them. The article type corresponds to original research publications. 206 articles and 132 articles were retrieved for Witten and Shneiderman, respectively. Trajectories were generated at both discipline and journal levels. Figure 7.16 illustrates the trajectories of the overlays generated from the two sets of publications. Witten's publications projected a compact and stable citing trajectory in areas

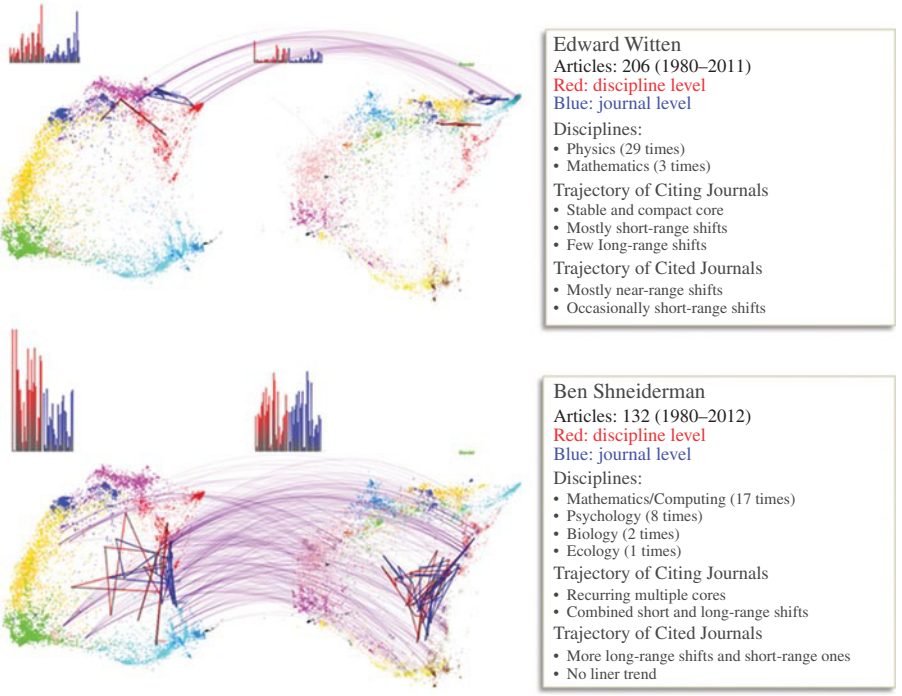

FIGURE 7.16 Characteristics of trajectories of Witten's publications and Shneiderman's publications. Citing trajectories of overlays are shown on the left. Cited trajectories are shown on the right. (*See insert for color representation of the figure.*)

corresponding to physics. The moves of Witten's trajectory are mostly short-range moves. The journal-level trajectory shows some brief episodes of publications in mathematics. No other disciplines were involved. In contrast, the citing trajectory of Shneiderman shows that he routinely published in two disciplines (mathematics/computing and psychology/education/health) and occasionally published in another two disciplines (biology and ecology).

Table 7.5 lists the nearest disciplines of the citing trajectories of both scholars over time. The names of disciplines are represented by terms chosen from titles of journals in the corresponding Blondel cluster of citing journals. Witten's publications

Table 7.5 Citing trajectories at the discipline level

	Disciplines of publications (Blondel cluster labels)	
Year	Edward Witten's publications	Ben Shneiderman's publications
1980	Physics, Materials, Chemistry	Psychology, Education, Health
1981	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
1982	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
1983	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
1984	Physics, Materials, Chemistry	
1985	Physics, Materials, Chemistry	Psychology, Education, Health
1986	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
1987	Physics, Materials, Chemistry	Psychology, Education, Health
1988	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
1989	Physics, Materials, Chemistry	
1990	Physics, Materials, Chemistry	
1991	Mathematics, Systems, Mathematical	Mathematics, Systems, Mathematical
1992	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
1993	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
1994	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
1995	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
1996	Physics, Materials, Chemistry	
1997	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
1998	Physics, Materials, Chemistry	Psychology, Education, Health
1999	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
2000	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
2001	Physics, Materials, Chemistry	Psychology, Education, Health
2002	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
2003	Physics, Materials, Chemistry	Molecular, Biology, Immunology
2004	Physics, Materials, Chemistry	Psychology, Education, Health
2005	Physics, Materials, Chemistry	Ecology, Earth, Marine
2006	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
2007	Mathematics, Systems, Mathematical	Mathematics, Systems, Mathematical
2008	Mathematics, Systems, Mathematical	Psychology, Education, Health
2009	Physics, Materials, Chemistry	Mathematics, Systems, Mathematical
2010	Physics, Materials, Chemistry	
2011	Physics, Materials, Chemistry	Molecular, Biology, Immunology
2012		Psychology, Education, Health

are remarkably persistent, containing an almost noninterrupted stream of the same discipline, physics. In contrast, Shneiderman's publication trajectory has shown an oscillation between mathematics and computing, and psychology, education, and health. Blank rows in the table indicate that the Web of Science has no records of Shneiderman's publication indexed during that year, which doesn't indicate whether or not he published outside the coverage of the Web of Science. Shneiderman is widely known for his work on human–computer interaction; thus, we expect the recurring psychology/education/health reflects his work in that aspect. He is also well known for his work in information visualization and other computing areas, which would correspond to the recurring mathematics, systems, and mathematical in his trajectory. More interestingly, the trajectory appears to merge two streams together such that one episode of one discipline is followed by an episode of another discipline. The switch between the two disciplines almost appears to be periodic. For example, the trajectory has repeated the shifts from psychology, education, and health to mathematics and systems for six times since 1980.

The contrast between the two illustrative examples shows that one can discern useful information from trajectories of portfolios of individual scientists at disciplinary and journal levels. Examining further details at lower levels of granularity is feasible, especially in connection with more traditional scientometric studies. The same type of analysis of trajectories can be used to study portfolios of a wide range of units of analysis at organizational and international levels.

7.4 SUMMARY

We have demonstrated the potential of simultaneously displaying two global maps of science at the discipline level. The dual-map design enables an explicit, intuitive, and easy-to-interpret representation of citations made by a wide variety of portfolios of publications. The dual-map space provides a flexible and extensible framework to support a new set of visual analytic tasks that are essential for portfolio analysis, gap analysis, and competitive intelligence. The notion of an aggregated trajectory of a portfolio provides an additional new gateway from the study of macroscopic patterns to the dynamics at microscopic levels.

Several issues need to be addressed and have room for improvement in the future. One issue that we have not addressed in the development of the dual-map overlay design is the stability of global science maps, and how their stability would influence the validity of patterns revealed. Pragmatically, how often do we need to update the underlying base maps in order to maintain the reliability of patterns of an overlay? Although the issue is concerned with science mapping in general, the role increasingly played by thematic overlays in portfolio analysis highlights the need to investigate the issue in particular. Another issue is related to the layout of the base maps. Our visualization has revealed a substantial amount of overlaps among Blondel clusters in both citing and cited base maps. Future research may investigate feasible trade-offs between the layout of base maps and their role as a gateway to integrate analytical tasks at various levels of granularity.

Our examples have demonstrated the flexibility of global maps of science at the level of journals and clusters of journals. A related issue is to what extent the new method introduced here can be applied to other types of global maps of science, such as a global map of science constructed at higher or lower levels of granularity than journals. In particular, a topic map of science derived from promising techniques such as topic modeling may apply. Leydesdorff and his colleagues have extended the base map construction process from scholarly publications to patents (Leydesdorff, Kushnir, & Rafols, 2014). The method described here is applicable to a dual-map overlay of patent portfolio analysis and even to a hybrid publication and patent portfolio analysis. We are actively pursuing an extension of the dual-map method to patent portfolio analysis. Our experience with the dual-map overlays also suggests that it may be worth considering multimap overlays for a comprehensive portfolio analysis that may involve multiple types of entities and relations, such as publications, patents, and grants.

The dual-map design enables analysts to perform several new and intuitive types of visual analytic tasks for portfolio analysis, including comparing dynamic patterns and trends of multiple portfolios at multiple levels of granularity from individual citation arcs, to dynamic patterns of trajectories aggregated over an entire portfolio. The dual-map design provides a new conceptual framework in which one can derive a variety of new metrics and algorithms. For example, we have introduced the concept of structural variation and its implications on detecting and predicting potentially transformative contributions of scientific publications in the framework of a complex adaptive system (Chen, 2012). The dual-map design provides an opportunity to study the predictive effects of structural variation from an alternative perspective. We will pursue this opportunity in our subsequent studies.

We have introduced a new method for portfolio analysis based on a dual-map design. The potential of the method is demonstrated through a series of examples of a variety of portfolios of publications, ranging from individual scientists, organizations, and subject matter-focused fields of research. We have shown that multiple overlays on the dual-map visualization can facilitate the analysis of portfolios in terms of identifying the areas of competencies and patterns of movements with reference to multiple disciplines. The dual-map overlays provide an intuitive gateway to integrate the study of scientific disciplines at a macroscopic level and the study of more specific specialties at a lower level of granularity. We can expect that the new method may lead to fruitful routes of further research and enrich the available methodologies for portfolio analysis, gap analysis, and competitive intelligence.

7.5 CONCLUSION

We begin the book with questions on what attracts our attention and what makes something interesting. A gap between what we know and what we don't know has emerged as one of the major driving forces behind answers to these questions. The magnitude of these gaps matters. If the gap is too small, then we are likely to use the perspective that we are currently using to answer the question. If the gap is too large,

then we may not even consider making any effort to bridge the gap. With a gap of an "interesting" length, we are attracted to learn and potentially find a new perspective to see the world.

In Chapter 2, we devoted to the role of mental models in guiding our thinking and decision making and discussed the risks of having a wrong mental model. Ironically, it often takes little effort to build up a mental model, but it may take much more effort and much stronger evidence to make us realize that we probably ought to update or discard our current mental model. We take many things for granted. Sometimes, it may be one thing too many.

In Chapter 3, we consolidated the mental model theme further through a series of examples that underlined the subjectivity of evidence. The same information may serve as evidence for different purposes and may support different arguments. We characterized the ambiguity with a metaphor of the tips of quite different icebergs. The examples of making arguments to prove or disprove that a proposition is beyond the reasonable doubt highlighted the challenges we face when we deal with the complexity of various plausible scenarios and our beliefs.

In Chapter 4, we summarized the CiteSpace system that we have developed to reduce the level of complexity in understanding the structure and dynamics of scientific knowledge. CiteSpace supports a generic process of systematically analyzing the structure of a complex adaptive system.

In Chapter 5, we reviewed the origin of the fitness landscape paradigm in evolutionary genetics and its subsequent variations and extensions beyond biology. The evolutionary thinking behind fitness landscapes is one of the few recurring themes throughout the book. Evolution offers an inspirational framework to rethink the nature of explorations, optimizations, recombinant search, and many profound phenomena.

In Chapter 6, we turned our attention to the exploratory, optimal, and recombinant search mechanisms manifested in inventing new ideas in high-impact radical patterns. We included several detailed examples to highlight the wisdoms of how to be creative. The implications of evolution theories are evident in examples that have gone beyond the realm of the genotype and phenotype. The notion of gaps between where we are and where we want to be is illustrated in generalized fitness landscapes.

In Chapter 7, we introduced a dual-map overlay technique for a new type of portfolio analysis. The concepts of landscapes, trajectories, and movements are integrated to reduce the complexity of detecting early signs of critical transitions in a complex adaptive system to trajectories that one can monitor and track as closely as possible.

There are a few grand challenges ahead. If we can resolve these challenges, then we will be able to make tremendous progress in improving our ability to handle the complexity of reality and a variety of issues concerning a complex adaptive system.

The first grand challenge is reducing the complexity of the information we need to deal with. How can we cut to the chase in assimilating a large amount of complex, uncertain, incomplete, and ambiguous information? Is it possible to reduce the complexity of how we represent and communicate scientific knowledge today? Fitness landscapes suggest a potentially fruitful direction to pursue because fitness landscapes encourage us to hide various details but direct our focus to where we are and what we can do to achieve our goals.

The second grand challenge is the increased preparedness for cognitive misperception. We are often unprepared for many things we take for granted. The worst-case scenario is that we are unprepared but we thought we were well prepared because we had the wrong mental model. What can we do to increase our consciousness that there may be alternative interpretations of what we see? How can we reduce the risk of missing subtle but critical warning signs?

The third grand challenge is to be able to detect early signs of critical transitions in a complex adaptive system. This is probably the most challenging area and also the one that we will benefit the most from if the challenge is resolved.

What these grand challenges have in common is the fitness of information. We have visited a number of principles, procedures, and concrete examples in this book. The theme of the fitness of information represents itself in many variations, notably including the genetic fitness of a population based on combinations of genes at the genotypic level, the fitness of a compound as a drug candidate as a function of its molecular structure, the fitness of a country in terms of its life expectancy as a function of its multidimensional array of economic indicators, and the fitness of an idea in terms of its predictive potential as an early warning sign for a critical transition in scientific knowledge. This diverse range of ways to characterize the fitness of information underlines the fundamental value of an evolutionary paradigm in dealing with the complexity of data science. The fitness of information leads naturally to the notion of a generalized fitness landscape and a tightly coupled relationship between the information we are analyzing and actions we may pursue to accomplish our long-term goals.

Figure 7.17 illustrates an ideal fitness landscape of scientific information. Many scientific inquiries can be conceptualized as a creative search on such a landscape. In some areas, scientists have consistent findings. In other areas, there may be contradictions, such as competing paradigms and conflicting mental models. Some areas may be well defined and populated, whereas other areas may be ambiguous, uncertain, or totally uncharted. There are peaks resulted from the Matthew Effect as well as valleys and no man's land where groundbreaking ideas may rise. The fitness

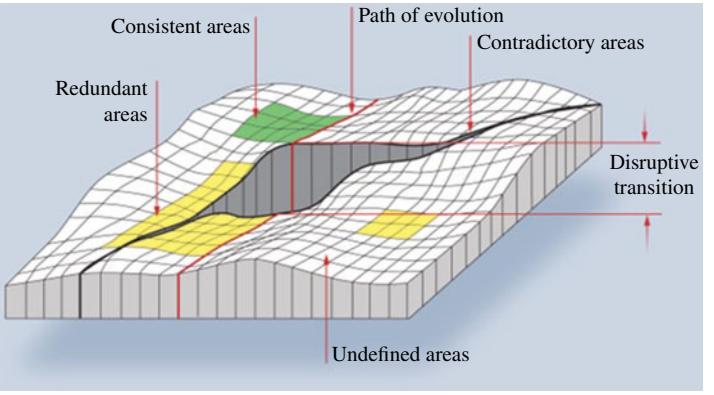

FIGURE 7.17 A fitness landscape of scientific inquires. (*See insert for color representation of the figure.*)

landscape provides an analytic framework for studying the behavior of a complex adaptive system at both microscopic and macroscopic levels. Ideally, the fitness landscape should provide an accurate representation of the history, the state of the art, and foresight of science. To paraphrase IARPA's RFI, what is the probability of having a fitness landscape of science of this kind on our desktop within five years?

In this book, we have explored and examined several important aspects of measuring the fitness of information in a diverse range of settings. The fitness value of information derived from complex, incomplete, ambiguous, and heterogeneous streams of data plays an essential role in maintaining our awareness of an evolving situation. The fitness of information also lets us focus on the basis of an evolutionary framework for gap analytics and portfolio analysis in a wide variety of domains.

The value of an evolutionary perspective is that it allows us to assess the fitness of information in the context of an optimization process, which can be a recombinant search, a boundary-spanning synthesis, a gap-bridging theory, a successive hill-climbing trajectory, or an exploration across no man's land.

As the tension increases between the size of the information haystack and the types of needles that may hold the key to critical transitions in a complex adaptive system, a focus on the fitness of information provides a good place to start.

BIBLIOGRAPHY

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *8*(10), 10008.

Chen, C. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology*, *63*(3), 431–449.

Chen, C. M. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377.

Chen, C., Cribbin, T., Macredie, R., & Morar, S. (2002). Visualizing and tracking the growth of competing paradigms: Two case studies. *Journal of the American Society for Information Science and Technology*, *53*(8), 678–689.

Chen, C., Hu, Z., Liu, S., & Tseng, H. (2012). Emerging trends in regenerative medicine: A scientometric analysis in CiteSpace. *Expert Opinions on Biological Therapy*, *12*(5), 593–608.

Chen, C., Hu, Z., Milbank, J., & Schultz, T. (2013). A visual analytic study of retracted articles in scientific literature. *Journal of the American Society for Information Science and Technology*, *64*(2), 234–253.

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. *Journal of the American Society for Information Science and Technology*, *61*(7), 1386–1409.

Dwyer, T., Marriott, K., Schreiber, F., Stuckey, P. J., Woodward, M., & Wybrow, M. (2008). Exploration of networks using overview + detail with constraint-based cooperative layout. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1293–1300.

Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology*, *54*(5), 400–412.

Jin, B., & Rousseau, R. (2001). An introduction to the Barycenter method with an application to China's mean centre of publication. *Libri*, *51*(4), 225–233.

Leydesdorff, L., & Rafols, I. (2011). Local Emergence and Global Diffusion of Research Technologies: An exploration of patterns of network formation. *Journal of the American Society for Information Science and Technology*, *62*(5), 846–860.

Leydesdorff, L., Kushnir, D., & Rafols, I. (2014). Interactive Overlay Maps for US Patent (USPTO) data based on International Patent Classifications (IPC). *Scientometrics*, *98*(3), 1583–1599.

Leydesdorff, L., Rafols, I., & Chen, C. (2013). Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal-journal citations. *Journal of the American Society for Information Science and Technology*, *64*(12), 2573–2586.

Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path-dependent transitions in HistCite (TM)-based historiograms. *Journal of the American Society for Information Science and Technology*, *59*(12), 1948–1962.

Mackinlay, J. D., Rao, R., & Card, S. K. (1995). *An organic user interface for searching citation links*. Paper presented at the SIGCHI'95, May 7–11, 1995, Denver, CO.

Poltorak, M., Leach, M., Fairhead, J., & Cassell, J. (2005). 'MMR talk' and vaccination choices: An ethnographic study in Brighton. *Social Science Medicine*, *61*(3), 709–719.

Porter, A. L., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719–745.

Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, *61*(9), 1871–1887.

Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., et al. (1998). Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children (Retracted article. See vol 375, pg 445, 2010). *The Lancet*, *351*(9103), 637–641.

Waltman, L., & Van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B*, *86*(11), 471.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In: *Proceedings of the sixth international congress on genetics* (pp. 356–366), August 1932, Ithaca, NY.

Zhang, R., Zhang, Y., Zhang, Q., Xie, H., Qian, W., & Wei, F. (2013). Growth of half-meter long carbon nanotubes based on Schulz–Flory distribution. *ACS Nano*, *7*(7), 6156–6161.