

Timeline Visualization of Research Fronts¹

Steven A. Morris², G. Yen, Zheng Wu, Benyam Asnake

School of Electrical and Computer Engineering
202 Engineering So.
Oklahoma State University, Stillwater, Oklahoma, 74078

ABSTRACT

Research fronts, defined as clusters of documents that tend to cite a fixed, time invariant set of base documents, are plotted as timelines for visualization and exploration. Using a set of documents related to the subject of anthrax research, this paper illustrates the construction, exploration, and interpretation of timelines for the purpose of identifying and visualizing temporal changes in research activity through journal articles. Such information is useful for presentation to members of expert panels used for technology forecasting.

1. INTRODUCTION

The objective of this work is to develop a visualization tool that can be used to provide a visual summary of research activity in a technical field to members of expert panels that conduct technology forecasts. Such a tool should help experts to understand changes over time in the technology of interest: identifying specific research areas and showing which areas are growing and which are declining. It is also important for experts to understand the flow of information among research areas: which areas contribute knowledge and which areas are “borrowers” of knowledge.

For this purpose, it is useful to track scientific paradigms as a function of time. Kuhn’s second definition of a paradigm, explained in the postscript to *The Structure of Scientific Revolutions* (Kuhn, 1970), as a set of “concrete puzzle-solutions” will be used

¹ This is a preprint of an article accepted for publication in a special topic issue (Visualization of Scientific Paradigms) of the Journal of the American Society for Information Science and Technology (JASIST) © 2002 John Wiley and Sons, Inc.

² Corresponding author

here. Taking a microscopic view, or more appropriately, a fractal view of research activity, research is conducted by small groups of people attacking equally small and intensely focused sets of shared problems. Kuhn indicates that these small groups of people are comprised of “perhaps one hundred members, occasionally significantly fewer” (Kuhn, 1970, p178). The goal of this work is not to visualize the temporal changes in the paradigms, i.e., the exemplar puzzle solutions, but more importantly, to identify and track the small groups of puzzle solvers and indeed the puzzles themselves, or *research fronts*, as evidenced in the scientific literature.

Timelines, plots where the x -axis variable is time, are a very effective method of displaying trends, building on the human perception of time as linear and taking advantage of the human ability to infer closeness of relation of two items from their spatial proximity. An example of the effective use of timelines to display temporal relations among research topics is seen in Small and Greenlee (1989), where the growth and diversification of research topics in AIDS research from 1981 to 1987 were derived from the co-citation patterns of a collection of research papers and plotted according to time. This timeline, referred to as a *cluster string*, clearly shows how research topics multiplied in AIDS research over time, and showed the interrelation and evolution of the topics. Braam, Moed, and van Raan (1991) used timelines to display time variations of several characteristics, e.g., citation rates to key documents over time, of a collection of documents from an unspecified field, where research topics were derived from the document set by clustering using a combination of document co-citation and analysis of word co-occurrence within individual documents.

This paper introduces a timeline technique wherein documents are initially clustered using *bibliographic coupling* (Kessler, 1962) into research fronts, groups of documents that consistently cite a fixed, time invariant group of base documents. Documents from a research front are plotted by time along a horizontal track in the timeline, with related research fronts being plotted in nearby tracks according to the hierarchical structure produced during clustering. Such a timeline is able to effectively communicate the temporal relations among research fronts and their base documents, and upon exploration using a suitable graphical user interface to the document database, can identify topics of research, experts, centers of excellence, fields of origin of base

documents, and most importantly for technology forecasting, the possible emergence of new research fronts.

The definition used here of a research front differs slightly from that of previous authors. Price (1965), without formal definition, mentions a research front as a “growing tip, or epidermal layer” of current papers and those papers in the immediate past cited by them. Garfield (1994) defines research fronts as “co-citation clusters and the documents that cite them.” Persson (1994) defines a research front as clusters of articles grouped by the co-citation clusters they cite. While the specific details of Garfield’s and Persson’s co-citation based methods of finding research fronts are unavailable, it appears that they involve the following steps: 1) the collection of documents is divided into two groups, highly-cited and not-highly-cited, 2) highly-cited documents are clustered using co-citation, 3) each not-highly-cited document is assigned to the co-citation cluster whose members are best represented in the document’s reference list, 4) Garfield’s method names each co-citation cluster and its assigned not-highly-cited documents as a research front, while Persson’s method names each co-citation cluster as a set of base documents and the assigned not-highly-cited documents as its associated research front.

Bibliographic coupling and co-citation are dual concepts; bibliographic coupling links documents that cite the same references while co-citation links documents that are cited together as references. Clustering using bibliographic coupling groups documents that cite the same groups of references while clustering using co-citation groups documents that are often referenced together.

Temporally, when moving from past to present, bibliographic coupling between two documents is static, since bibliographic coupling is based on the fixed reference lists of the two documents. Bibliographic coupling is immediately available upon publication of the second document of a pair. Co-citation between two documents, however, varies with time as new papers are published and always starts at zero and grows to a stable value over a period of years.

It is the static and immediate nature of bibliographic coupling that makes it convenient to use for constructing timelines. Using bibliographic coupling, the process of finding research fronts and base documents is the dual of the co-citation based process outlined above: 1) documents are clustered into static research fronts using bibliographic

coupling, 2) base documents are assigned to each research front by identifying the references common to a large percentage of the documents in the research front. Note that it is possible for a base document to belong to more than one research front if it is cited heavily from the documents of more than one research front. Also note that the base documents themselves can be assigned membership in a research front.

It is necessary to apply qualifying criteria to bibliographic coupling clusters before labeling them as research fronts. The qualifying criteria assure that a bibliographic coupling cluster is above a given size and possesses a consistent set of base documents before it is considered a research front. The qualifying definition used in this paper is that a research front contains at least 10 documents and possesses at least one base document, where a base document is a document cited by 40% or more of the documents in the research front.

When the bibliographic coupling based process outlined above is applied, base documents tend to precede all the documents in their associated research front. As will be seen in the anthrax example given here, the location of base documents in relation to their associated research fronts can reveal information about the flow of information within a collection of research fronts.

The remainder of this paper is organized as follows: Section 2 of this paper will explain the fundamental assumptions used to justify the use of research front timelines. Next, using a set of anthrax documents as an example, Section 3 of this paper explains the construction of timelines, a method based upon clustering documents by inter-document similarities based on bibliographic coupling. Section 4 follows with a detailed exploration and analysis of the research fronts of the anthrax data set, illustrating the use of the timeline visualization, and the extraction of information using database interface functions to the timeline. Section 5 summarizes results.

2. EXTRACTING AND DISPLAYING RESEARCH FRONTS AS TIMELINES

As noted above, a paradigm is defined as a set of exemplars, i.e., example “puzzle solutions,” while a research front corresponds to a currently unsolved puzzle of interest to researchers. It seems safe to assume that paradigms occur at all scales, and occur as

fundamental and specialized knowledge in even the narrowest technical fields. Paradigms are by definition stable, but when viewed as sets of exemplars it is evident that there will be considerable overlap in the membership of individual exemplars in different paradigms. Important exemplars will be included in the paradigms of many research fronts, while very specialized exemplars will be in the paradigms of only a small number of research fronts. This suggests a hierarchical structure of exemplars and also the possibility of comparing and relating groups of research fronts by the overlap of exemplars included in their individual paradigms.

If the activity within a research front leads to a scientific discovery, that discovery could create new scientific “puzzles” to be solved, in other words, one or more new research fronts could abruptly come into existence. In this case the discovery will most likely be an exemplar in the paradigms of those new research fronts. It is also evident that when the “puzzle” associated with a research front is solved or otherwise becomes moot, the research front will disappear. These ideas imply that research fronts are discontinuous, starting and ending abruptly as scientists move from one puzzle to the next.

Given the above characteristics of paradigms, exemplars, and research fronts, the following assumptions are made about the evidence of research fronts in collections of scientific documents:

- Documents within a research front will tend to cite a set of fixed, time invariant base documents representing the exemplars of the research front’s paradigm.
- Given a set of documents covering a research area, documents will form into several research fronts, hierarchically organized. Documents in nearby branches of this hierarchy will tend to overlap in the base documents they cite.
- Given a set of documents covering a research area, many base documents, and other documents as well, will not be a part of any specific research front. These documents outside the collection of research fronts will be referred to as “external” documents. This

assumption recognizes that some exemplars are drawn from outside the scientific specialty of interest, and that due to overlap in subject areas, some documents in the collection will not be related directly to any identified research front.

The assumptions above suggest that research fronts can be tracked in the literature by hierarchically clustering documents into groups that cite similar sets of documents. A scatter plot of such clustered documents, using publication date on the x -axis, placing the documents into horizontal tracks on the y -axis according to their cluster membership, will generate a timeline of the research fronts. Such a timeline will show the time of appearance and disappearance of each research front, and can display structural information about the collection of research fronts as well. The following paragraphs describe the construction of such a timeline.

As a working definition, define a research front as any collection of 10 or more documents that possess one or more base documents. Define a base document as a document cited by 40% or more of the documents in a research front. Small clusters of documents and large clusters of external documents fall outside this working definition of a research front. Using this definition, a document cluster produced during clustering is a research front if it meets the above requirements. While the research front size and base document definitions are somewhat arbitrary, experience shows that ten documents is a convenient minimum size for a research front, and that smaller size clusters tend to produce anomalous base documents. Additionally, the cutoff figure of 40% for base documents tends to produce between 5 and 10 base documents for each cluster; a smaller cutoff figure produces an unmanageable number of base documents while a larger cutoff figure produces a paucity of base documents.

Bibliographic coupling, normalized to account for differing lengths of the lists of references in individual documents, is a good measure of similarity when clustering documents according to the groups of documents they cite. Bibliographic coupling between two documents is defined as the number of references common to the reference lists of both documents. Bibliographic couplings are converted through normalization to

similarity values between zero and unity, which are then converted to distances for hierarchical clustering by subtracting them from unity.

Based on experience, an agglomerative hierarchical clustering routine, using Ward's method, works well for this application. Ward's method (Gordon, 1999) is not subject to "chaining" artifacts, as is single link clustering, and tends to produce well balanced clusters with generally equal numbers of documents in each cluster. For a set of about 400 documents, it is generally useful to start by generating 10 clusters, possibly modifying this number based on exploration of the data after the timeline is generated.

The agglomerative clustering method produces a dendrogram of the cluster tree structure that maps the clusters in a linear sequence as leaves on the dendrogram. This sequence is used to order the horizontal tracks within which documents from the clusters are plotted on the timeline. The dendrogram is positioned on the left side of the timeline to help visualize the structure of the research fronts, as shown in Figure 1.

Layout of timelines as shown in Figure 1 produces a simple, easily comprehensible visualization. The structure of the research fronts can be inferred from the dendrogram displayed to the left of the plot, while the area to the right of the plot provides a convenient and spacious area for research front labels. Horizontal gridlines are placed on the plot to mark the tracks along which documents are plotted. It is necessary to dither the placement of the documents slightly off the cluster tracks, a technique that provides vertical height to the horizontal clusters, thus allowing visualization of document density in each track as a function of time. Additionally, the size of document markers are made proportional to the number of times a document has been cited, causing important papers, as judged from large citation counts, to stand out on the timeline as shown in Figure 1.

The timeline layout also facilitates the application of interactive exploration functions. The timeline of Figure 1 was produced using DIVA, "Database Information Visualization and Analysis" system, a home-brew research tool for interactively exploring document sets (Morris, 2002). DIVA allows selection of groups of documents on the timeline using mouse pointer functions. Pop-up windows display lists of titles, authors, abstracts, and frequency tables of citations and other relevant data about the selected documents. A proposed interactive feature, not yet implemented, is to add

the ability to expand and collapse branches of the cluster tree to explore the structure of the research front collection.

Using DIVA, citations can be displayed on the map. A citation from one document to another is displayed as a line between those two documents. Functions are provided to display citations from a selected group of documents to its base documents. This greatly facilitates the exploration and interpretation of the timeline and yields insight into the information flow among research fronts.

3. GENERATION OF A TIMELINE OF ANTHRAX LITERATURE

Figure 1 shows an example of a timeline of 347 documents from the ISI Science Citation Index on the subject of anthrax. This document timeline was constructed using the following sequence of steps:

1. Using the term 'anthrax' in the general search function of ISI's Web of Science product, a set of 821 documents was collected corresponding to journal articles published from 1981 to the end of 2001. Document authors, citations, titles, and source journal information were stored.
2. A frequency table of citations was constructed for the 821 anthrax documents. Of the top 50 cited documents, those that were available but that were not captured in the original query were manually retrieved, bringing the total number of documents to 833.
3. A SQL query was used to count the number of total bibliographic couplings for each pair of documents in the set. Any document that did not have at least 5 bibliographic couplings with another document was discarded, reducing the total number of documents to 347. This step removes documents from the dataset that do not have strong similarity to any other document, and helps to produce meaningful clusters. While this step produces a drastic reduction in the size of the document collection, this is to be expected, as the original query would have retrieved many documents not well related to anthrax research.

4. Inter-document similarities were calculated using the cosine coefficient (Salton, 1989):

$$s_{ij} = \frac{bc_{ij}}{\sqrt{N_i N_j}} \quad (1)$$

- Where bc_{ij} is the number of documents cited by both document i and document j , while N_i and N_j are the total number of document citations for document i and document j respectively. Similarities were converted to distances by subtracting them from unity.
5. An agglomerative hierarchical clustering routine was used to group the documents. Linkage was calculated using Ward's method (Mathworks, 2000). The cluster tree was truncated to produce ten clusters.
6. A dendrogram was constructed to produce y -axis placements of the clusters on the timeline plot. Documents were plotted in horizontal tracks using publication date as the x coordinate and cluster leaf position on the dendrogram as the y coordinate. Values of individual document y coordinates were dithered slightly, adding vertical depth to the horizontal clusters.
7. Document titles from each cluster were examined manually for common themes in order to derive labels for each cluster. Upon exploration, it was determined that documents in cluster 9 should be further subdivided. Expansion of cluster 9 into its two sub-branches produced clusters 9a and 9b and was the final step in the construction of the timeline.

Figure 1 shows the resulting timeline as constructed using the steps above. From here the timeline was explored with the assistance of a subject matter expert to label research fronts, identify seminal papers, derive information about the evolution of the anthrax research field and most importantly, to identify possible emerging research fronts in the field.

4. EXPLORATION AND DISCUSSION OF THE ANTHRAX TIMELINE

Anthrax research makes an excellent benchmark for testing the ability to visualize temporal changes in research fronts as they appear in the scientific literature. A great deal of anthrax research has been performed in the past 20 years; it is well documented, and is well covered by the Science Citation Index. A review paper exists (Bhatnagar and Smriti, 2001) that names and discusses many key papers in anthrax research in the past 20 years. From early preliminary anthrax research, several research fronts have emerged with varied growth characteristics. Vaccine and gene sequencing research fronts have proceeded steadily for 15 to 20 years, for example, while research on anthrax toxins shows a pattern of rapid growth and specialization. The topic of anthrax bioterrorism emerged since 1999 in response to perceived threats, and the Fall, 2001 bioterror attacks through the U. S. postal services have already produced measurable changes in the citation patterns of the anthrax bioterrorism research front.

The next four paragraphs give a short summary of anthrax and anthrax research which will aid the reader in evaluating the interpretation of the anthrax timeline that follows. Most of this summary is derived from Bhatnagar and Smriti. Anthrax research has a very long and significant history, and was the disease used by Koch, a contemporary of Pasteur in the late 19th century, to prove the original “germ theory.” Anthrax was originally thought to cause death by blocking of capillaries, but experiments by Smith and Keppie in the 1950’s showed that it kills through the actions of a toxin. The anthrax toxin consists of three parts, *protective antigen* (PA), *lethal factor* (LF), and *edema factor* (EF). Gladstone reported on anthrax protective antigen for the first time in 1946, while Leppla reported in detail on lethal factor and edema factor in a seminal paper in 1982.

Anthrax spores enter the host and are taken up by macrophages, amoeboid cells that attack foreign matter in the host, and are transported to nearby lymph nodes. Anthrax bacilli are protected from the macrophages by a *capsule*, an external covering. Within the macrophages the bacteria germinate and after release from the macrophages the bacilli multiply in the lymph system and eventually enter the blood stream.

Friedlander first reported the importance of macrophages in the spread of the infection in a seminal paper in 1986.

In the bloodstream the bacilli secrete the three-part toxin that attacks cells and eventually kills the host. When attacking cells, protective antigen bonds to a receptor protein on the host cell surface where it cleaves to become the protein PA63 and then forms a portal into the cell through which lethal factor and edema factor pass to do their damage inside the cell. Anthrax treatment usually fails if delayed because, while it is possible to kill off the bacilli with antibiotics, the toxin that was produced before treatment remains to kill off the host.

The following paragraphs describe exploration, labeling and interpretation of the anthrax document timeline of Figure 1. After the generation of the timeline, each cluster of documents was explored to extract cluster labels, identify base documents, and further identify seminal papers that may have resulted in the creation of new research fronts. Visual cues from temporal relations of the plotted research fronts and from displays of document citations were used to investigate sets of documents that could be part of an emerging research front. A subject matter expert was consulted to validate the interpretation of the timeline.

It is not difficult to label the anthrax document clusters by exploring titles of the documents in individual clusters. Exploration of frequency tables of cited documents for each cluster allows identification of base documents. There is some overlap of base documents among clusters that are on nearby branches of the dendrogram.

Through exploration and visualization of document citations, it is possible to trace information flow from the early research fronts to the latest, more specialized research fronts. As shown by the large arrows in Figure 1, the earliest research front, “preliminary research,” dealing mostly with immunology and some preliminary topics on toxins, was the origin of basic information for research that followed. Largely building on the seminal paper by Leppla in 1982, the research front became inactive in the late 1980’s shortly after the publication of the seminal paper by Friedlander on macrophages in 1986. The research front split off into specialties dealing with vaccines, gene sequencing, toxin production, and toxin research.

Toxin research consisted of basic questions of how protective antigen, lethal factor and edema factor worked together to destroy cells. In the mid-1990's the operation of the three part toxin was well enough understood that specialty research fronts emerged dealing with protective antigen and lethal factor. Surprisingly, a research front emerged as scientists began attempting to use anthrax protective antigen's ability to open a portal in cell membranes as a means to mediate delivery of substances into macrophages that could induce immunity to HIV and other diseases (Harvard Medical School, 1996).

A research front dealing with anthrax bioterrorism was likely induced into existence by growing concern over threats from terrorists and rogue nations in the 1990's. Exploration of citations from documents in this group indicates that there is some flow of information from documents in the vaccine research front.

In Figure 1, track 1, marked "external," corresponds to a large group of documents with no identifiable citation pattern. As pointed out in the assumptions, this cluster is expected to exist in the document set and corresponds to documents that are external to any of the research fronts on the timeline. The document most frequently cited from this cluster is only cited by 14% of the documents in the cluster; therefore this cluster falls outside the working definition of a research front given above. Track 7, containing only 7 documents, also falls outside the working definition of a research front.

Track 5, "Preliminary research", is the oldest of the research fronts. Visually, it appears to have become inactive around 1990. Table 1 gives the base documents for this cluster, which date back to the 1950's and have a median publication date of 1974. Many of the documents in this research front deal with immunology and vaccines. There are many base documents for other research fronts in this track.

Track 3, "vaccine research," begins around 1987 and extends to the present. Of the 39 documents in this research front, 19 of them use the term "vaccine" or "immune" in the title, while 8 of the 13 base documents come from the journal "*Infectious Immunity*". Table 2 shows the base documents of this research front. The median base document publication year is 1986. Figure 2 shows the location of the base documents relative to the research front itself. Lines between documents on this figure represent citations from documents in the "vaccine research" research front to base documents. Base documents are located at the vertices of groups of lines on the figure. Most base

documents are within “preliminary research” or within the early papers of “vaccine research” itself, indicating that “vaccine research” is a continuation and redirection of research from “preliminary research.” Friedlander, 1986, the seminal paper on macrophages, is cited heavily from “vaccine research”, and appears shortly before its emergence, implying Friedlander, 1986, was the key knowledge that caused the formation of the “vaccine” research front.

Among the four toxin research fronts, track 9b, “PA mediated delivery,” will be used as an example. Table 3 lists the base documents for this cluster, while Figure 3 shows the position of the base documents relative to the track. The median publication date for these base documents is 1993. Note that many of the base documents are in track 2, “secondary research”. This implies a significant flow of information from “secondary research” into “PA mediated delivery” and suggests that “PA mediated delivery” is a continuation and redirection of research in “secondary research.” Note two base documents, Petosa, 1997, and Duesbery, 1998, in the “external” cluster. Such an anomalous uptake of information from such recent documents possibly signals the impending emergence of a new research front, as this indicates that new exemplars have been added to the research front’s paradigm, signaling a paradigm “shift.” Expansion of track 9b into two sub-clusters would probably reveal a sequence of two research fronts, with the second research front starting in 1997 and including documents that cite Petosa, 1997, and Duesbery, 1998. Anomalies such as this, where documents from a research front start citing new documents heavily, can be brought to the attention of subject matter experts who can evaluate their significance. See the later discussion on the “bioterrorism” research front for a further example of this type of anomaly and an interpretation of its significance.

The paper “Duesbery, 1998,” is worthy of note. This document, published in *Science* in May, 1998, has been cited 72 times as of January, 2002, a large number of citations for that short, 3 year, 8 month period. Excluding external documents, 40% of all documents on the timeline published after “Duesbery, 1998” cite that paper, indicating that it is a seminal paper that will become a base document for the anthrax research front as a whole. The paper’s status as an external document is confirmed by noting that Duesbery was conducting cancer research and serendipitously discovered that anthrax

lethal factor attacks MAPKK, a signaling pathway within cells (Travis, 1998). This finding, which greatly advanced knowledge about how anthrax kills host cells, has been of great utility to anthrax researchers.

Figure 4 shows the position of Duesbery, 1998, and three other important documents that followed it. Note the great number of citations to Duesbery, 1998, from documents in all tracks on the timeline. Two highly cited papers succeeding Duesbery, 1998, also discuss MAPKK. These papers, are Vitale, 1998, and Pellizzari, 1999, and are presumably based on Duesbery, 1998. Also noted in Figure 4, Wesche, 1998, provides information on the mechanism that allows protective antigen to transmit lethal factor and edema factor through the cell membrane. The information available from Wesche, 1998 and Duesbery, 1998 appears to have facilitated the creation of a new research front, shown as track 9a, “specific PA research.”

Track 10, “bioterrorism” is worthy of note. This research front begins around 1997 and consists of 25 documents. Table 4 shows a list of the base documents for this research front, while Figure 5 shows the position of the base documents relative to “bioterrorism” documents. One of the base documents is from track 3, “vaccine research,” while the remaining documents are external. There are no repeat authors in the papers comprising this research front, indicating a lack of authors consistently publishing about anthrax bioterrorism. Among the base documents, “Inglesby, 1999” is worthy of note. This document, published in November, 1999 in the *Journal of the American Medical Association*, has been cited 39 times as of January 2002, a large number of citations for that 2 year, 2 month period. In “bioterrorism,” 60% of the documents cite “Inglesby, 1999,” which is a collection of consensus-based recommendations formulated by the Working Group on Civilian Biodefense.

In “bioterrorism,” six of the documents were published from November 2001 to January 2002 and deal directly with the anthrax bioterrorism attacks through the U. S. postal system that occurred late in the year 2001. Upon exploration, “Dixon, 1999,” from the *New England Journal of Medicine*, is cited by 5 of these 6 documents, but is only cited by 2 of the previous 19 documents in the track. Figure 5 shows these six documents and their citations to Dixon, 1999. Dixon, 1999, is a review article describing in detail the medical aspects of anthrax, e.g., manifestations, diagnosis, testing, and treatment. Its

appearance as a reference in documents after the anthrax attacks reflects the need to reference base knowledge of the treatment of anthrax, a subject that, until the time of the attacks, had been neglected in anthrax research because of the rarity of anthrax cases. The references to Dixon, 1999, possibly indicate the impending emergence of a research front on anthrax treatment.

As a final point of discussion, readers are encouraged to compare the visual information exhibited by the anthrax timeline to Garfield, Malin, and Small's observations about the "life cycle" of some specialties as evidenced in the scientific literature (1978). These authors note that often the following sequence of phases occurs in the literature following a significant scientific discovery:

1. The appearance of a highly cited discovery paper or small series of discovery papers.
2. The appearance, a year or two after the discovery papers, of a series of exploitation papers, highly co-cited with the discovery papers.
3. The death or decline, concurrently with the appearance of exploitation papers, of the old methodological co-citation cluster.
4. The emergence of a period of stability where papers published in the specialty are cited at a normal rate.
5. The decline of the specialty, or, given a new discovery in the specialty, the repeat of the "life cycle" outlined above.

Note in Figure 2 that Leppla, 1982, is highly cited (judging from its relative size), and precedes all the papers in its research front. From this we conclude that Leppla, 1982 is a discovery paper. In the research front, "preliminary research," Leppla, 1982, is followed by a series of highly cited papers that presumably represent a set of exploitation papers. This research front does not experience a period of stability, but ends abruptly after Freidlander, 1986, is published. The Friedlander paper is the discovery paper for the "secondary research" research front, initiating a new "life cycle." Note that the "vaccine research" and "gene sequencing" research fronts exhibit, after initial highly cited discovery or exploitation papers, periods of long stability.

As evidence of the effectiveness of timeline visualization, the discussion above indicates that the anthrax timeline visualizes several of the phases of specialty “life cycles” that are listed in Garfield, Malin and Small’s verbal description. It should be noted however that not all specialties exhibit change as dramatically as this particular example.

5. CONCLUSION

As shown in the previous section, the use of document timelines allows the extraction of useful information from a database of scientific journal papers for presentation to subject matter experts for the purpose of technology forecasting. This information includes:

- Dates of emergence and dates of disappearance of research fronts.
- Possible emerging research fronts and their potential base documents.
- Experts and centers of excellence as indicated by author and institution frequency from documents within the research fronts.
- Hierarchical structure of research fronts within the overall field being investigated.
- Information flow among research fronts and from external fields.

The timeline method presented here is not difficult to implement. The anthrax data set used here, for example, was stored in an MS Access database on a PC workstation. Similarity values were calculated using simple SQL queries, clustering and plotting were done using MATLAB. A visual interface for extracting summary data from the database is somewhat complicated to implement, but if such an interface is unavailable then much of the information needed for timeline interpretation can be generated by SQL queries spilling into printed reports (Morris, 2002).

These timelines are useful for presentation to expert panel members for technology forecasting. The added insight into the structure of research in the field of interest, and

the information on the impending emergence of new research fronts contributes greatly to the production of a high quality technology forecast by well-informed experts.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support from Halliburton Energy Services. Many thanks to molecular biologist Dr. James Harmon of Oklahoma State University for his help in validating and interpreting the timeline of anthrax research fronts.

REFERENCES

- Bhatnagar, R., Smriti, B. (2001). Anthrax toxin. *Critical Reviews in Microbiology*, 27(3), 167-200
- Braam, R. R., Moed, H. F., van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis. II: dynamical aspects. *Journal of the American Society of Information Science*. 42(4), 252-266
- Price, D. de S. (1965). Networks of scientific papers. *Science*, 149, 510-515
- Garfield, E., Malin, M. V. & Small, H. (1978). Citation data as science indicators. In U. Elkana & J. Lederberg & R. K. Merten & A. Thackray & H. Zuckerman (Eds.), *Toward a metric of science: the advent of science indicators*: John Wiley and Sons.
- Garfield, E. (1994). Research fronts. *Current Contents*, 41, 3-7, Oct. 10, 1994.
- Gordon, A. D. (1999). *Classification. Monographs on statistics and applied probability ; 82 (2nd ed.)*. Boca Raton: Chapman & Hall/CRC
- Harvard Medical School Office of Public Affairs. (1996). Turning weapons into vaccines: Harvard researchers are putting anthrax toxin to good use. *Focus*, December 6, 1996
- Kessler, M. M. (1962). Bibliographic coupling between scientific papers. *American Documentation*, 14, 10-25
- Kuhn, T. S. (1970). *The structure of scientific revolutions, second edition*. University of Chicago Press, Chicago, U.S.A.
- Mathworks. (2000). *Statistics toolbox user's guide, version 3*. The Mathworks Inc., 2 Apple Hill Drive, Natick, MA 01760
- Morris, S. , Deyong, C., Wu, Z., Salman, S., Yemenu, D. (to be published in 2002). DIVA: a visualization system for exploring document databases for technology forecasting. *Computers and Industrial Engineering Journal*
- Persson, O. (1994). The intellectual base and research fronts of JASIS 1986-1990. *Journal of the American Society for Information Science*, 45(1), 31-38
- Salton, G., (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley: Reading, Mass., U.S.A
- Small, H., Greenlee, E. (1989). A co-citation study of AIDS research. *Communication Research*, 16(5), 642-666
- Travis, J. (1998). New clue hints at how anthrax kills, *Science News*. 154(10), 299, May 9, 1998

Table 1. Base documents for track 5, “preliminary research”

citation	percent of group citing
RISTROPH JD, 1983, INFECT IMMUN, V39, P483	52%
MIKESELL P, 1983, INFECT IMMUN, V39, P371	47%
PUZISS M, 1963, APPL MICROBIOL, V11, P330	47%
HAINES BW, 1965, J BACTERIOL, V89, P74	42%
THORNE CB, 1957, J GEN MICROBIOL, V17, P505	42%
STANLEY JL, 1961, J GEN MICROBIOL, V26, P49	42%
EZZELL JW, 1984, INFECT IMMUN, V45, P761	42%
LEPPLA SH, 1984, ADV CYCLIC NUCL PROT, V17, P189	42%

Table 2. Base documents for cluster 3. “vaccine research”

citation	percent of group citing
TURNBULL PCB, 1986, INFECT IMMUN, V52, P356	64%
LITTLE SF, 1986, INFECT IMMUN, V52, P509	64%
MIKESELL P, 1983, INFECT IMMUN, V39, P371	56%
IVINS BE, 1986, INFECT IMMUN, V54, P537	48%
IVINS BE, 1992, INFECT IMMUN, V60, P662	48%
BRACHMAN PS, 1962, AM J PUBLIC HEALTH, V52, P632	46%
PUZISS M, 1963, APPL MICROBIOL, V11, P330	46%
TURNBULL PCB, 1988, MED MICROBIOL IMMUN, V177, P293	43%
GREEN BD, 1985, INFECT IMMUN, V49, P291	43%
FRIEDLANDER AM, 1986, J BIOL CHEM, V261, P7123	43%
IVINS BE, 1986, INFECT IMMUN, V52, P454	41%
LEPPLA SH, 1982, P NATL ACAD SCI USA, V79, P3162	41%
IVINS BE, 1990, INFECT IMMUN, V58, P303	41%

Table 3. Base documents for cluster 9b, “PA mediated delivery of other substances”

citation	percent of group citing
LEPPLA SH, 1982, P NATL ACAD SCI USA, V79, P3162	61%
ARORA N, 1993, J BIOL CHEM, V268, P3334	57%
MILNE JC, 1994, J BIOL CHEM, V269, P20607	57%
PETOSA C, 1997, NATURE, V385, P833	57%
LEPPLA SH, 1995, HANDB NAT T, V8, P543	52%
ARORA N, 1992, J BIOL CHEM, V267, P15542	52%
MILNE JC, 1995, MOL MICROBIOL, V15, P661	52%
FRIEDLANDER AM, 1986, J BIOL CHEM, V261, P7123	47%
SINGH Y, 1989, J BIOL CHEM, V264, P19103	47%
DUESBERY NS, 1998, SCIENCE, V280, P734	42%

Table 4. Base documents for cluster 10, “bioterrorism”

Citation	percent of group citing
MESELSON M, 1994, SCIENCE, V266, P1202	76%
FRANZ DR, 1997, JAMA-J AM MED ASSOC, V278, P399	60%
INGLESBY TV, 1999, JAMA-J AM MED ASSOC, V281, P1735	60%
TOROK TJ, 1997, JAMA-J AM MED ASSOC, V278, P389	40%
PILE JC, 1998, ARCH INTERN MED, V158, P429	40%

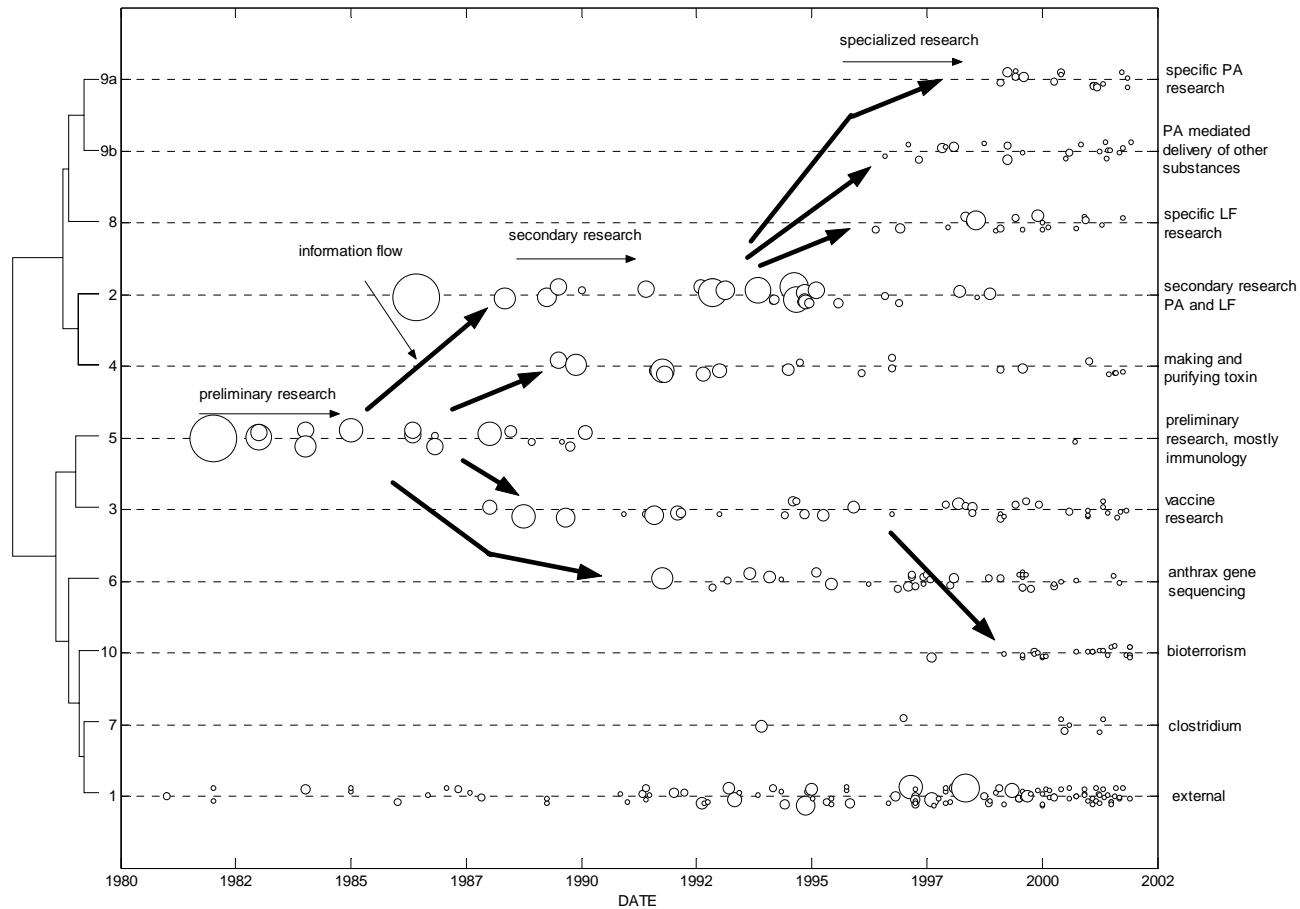


Figure 1. Document timeline for the anthrax dataset. Information flow, derived from exploration of citation patterns, is shown as heavy arrows.

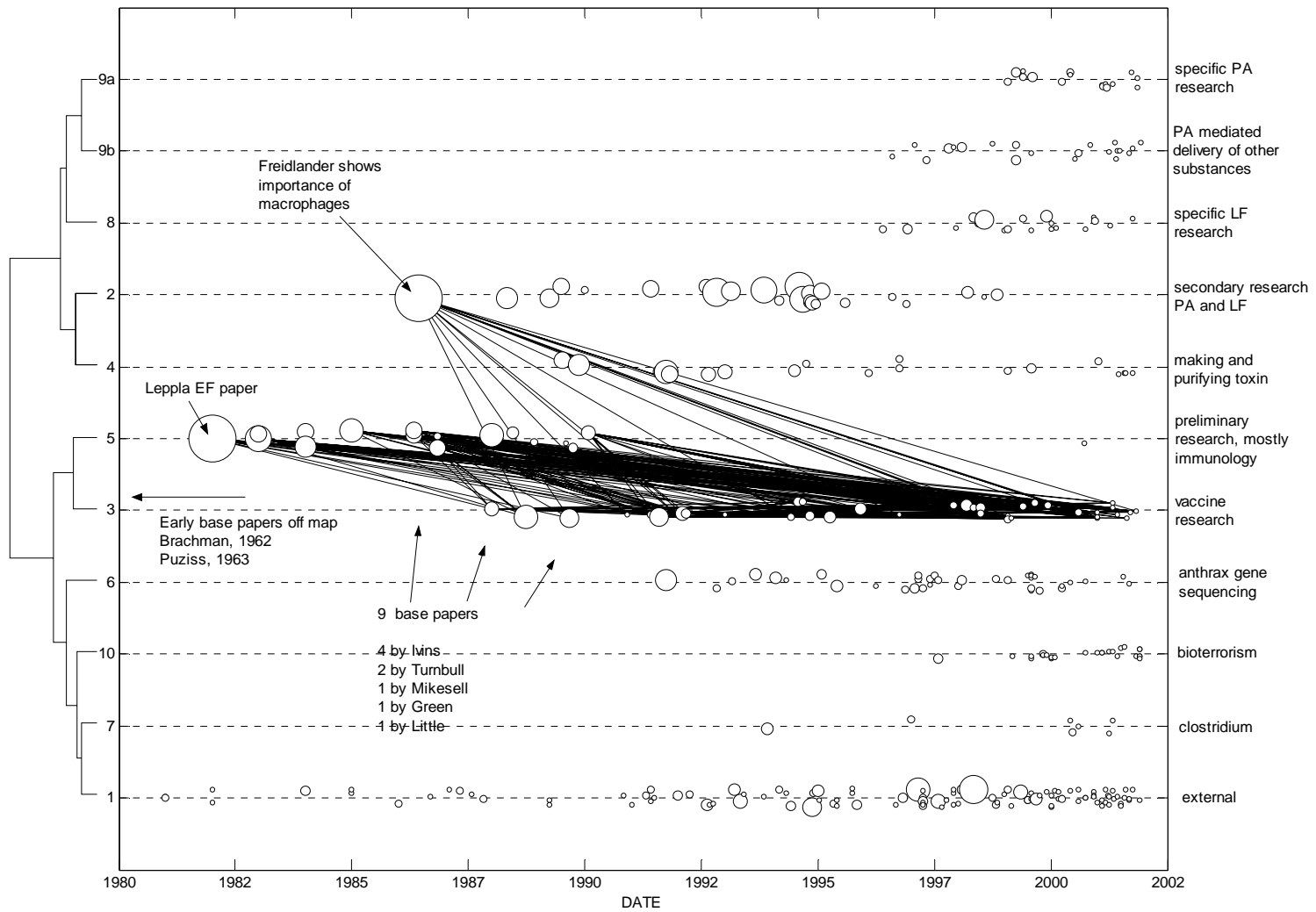


Figure 2. Base documents for track 3, "Vaccine research"

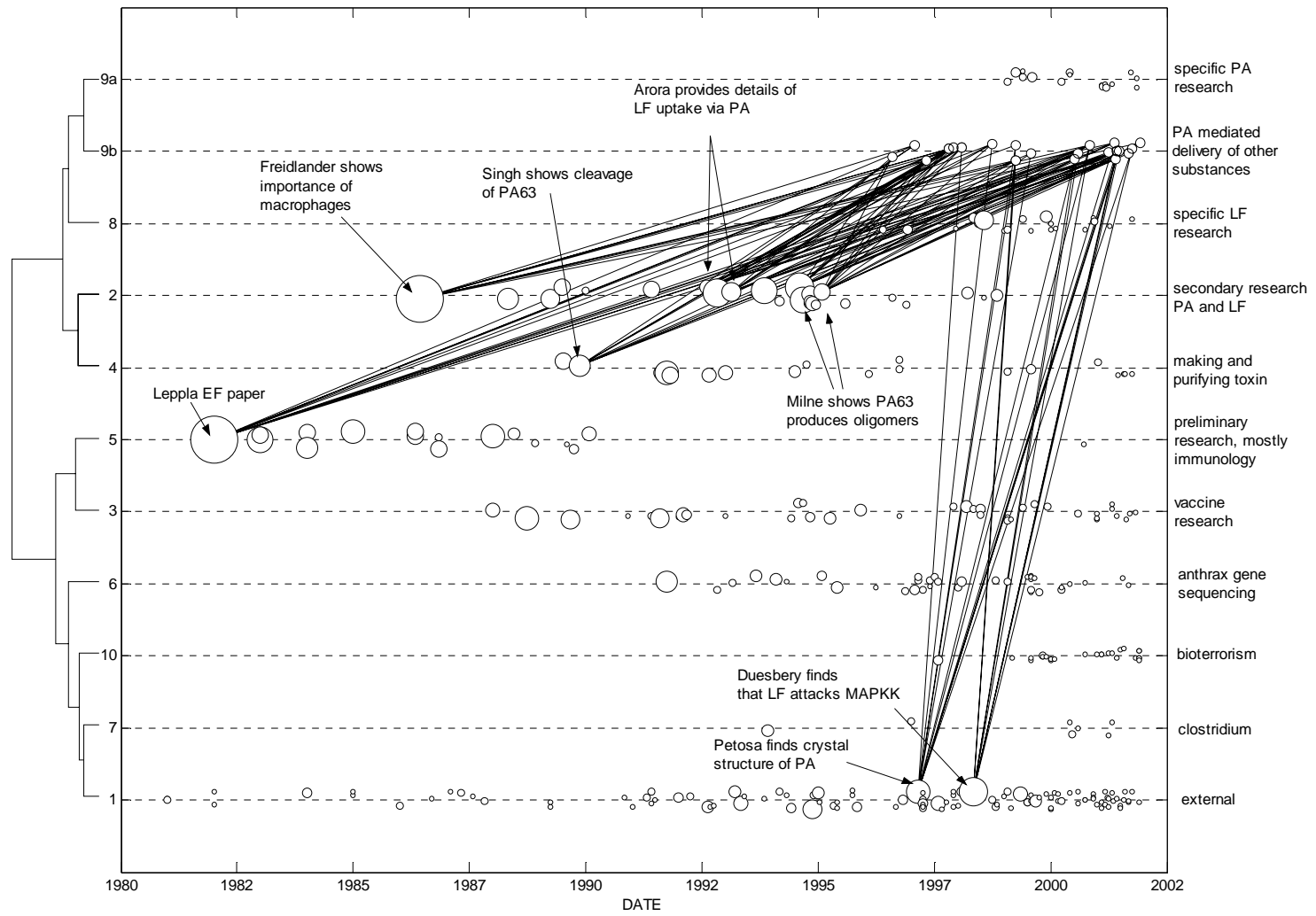


Figure 3. Base documents for track 9b, “PA mediated delivery of other substances”

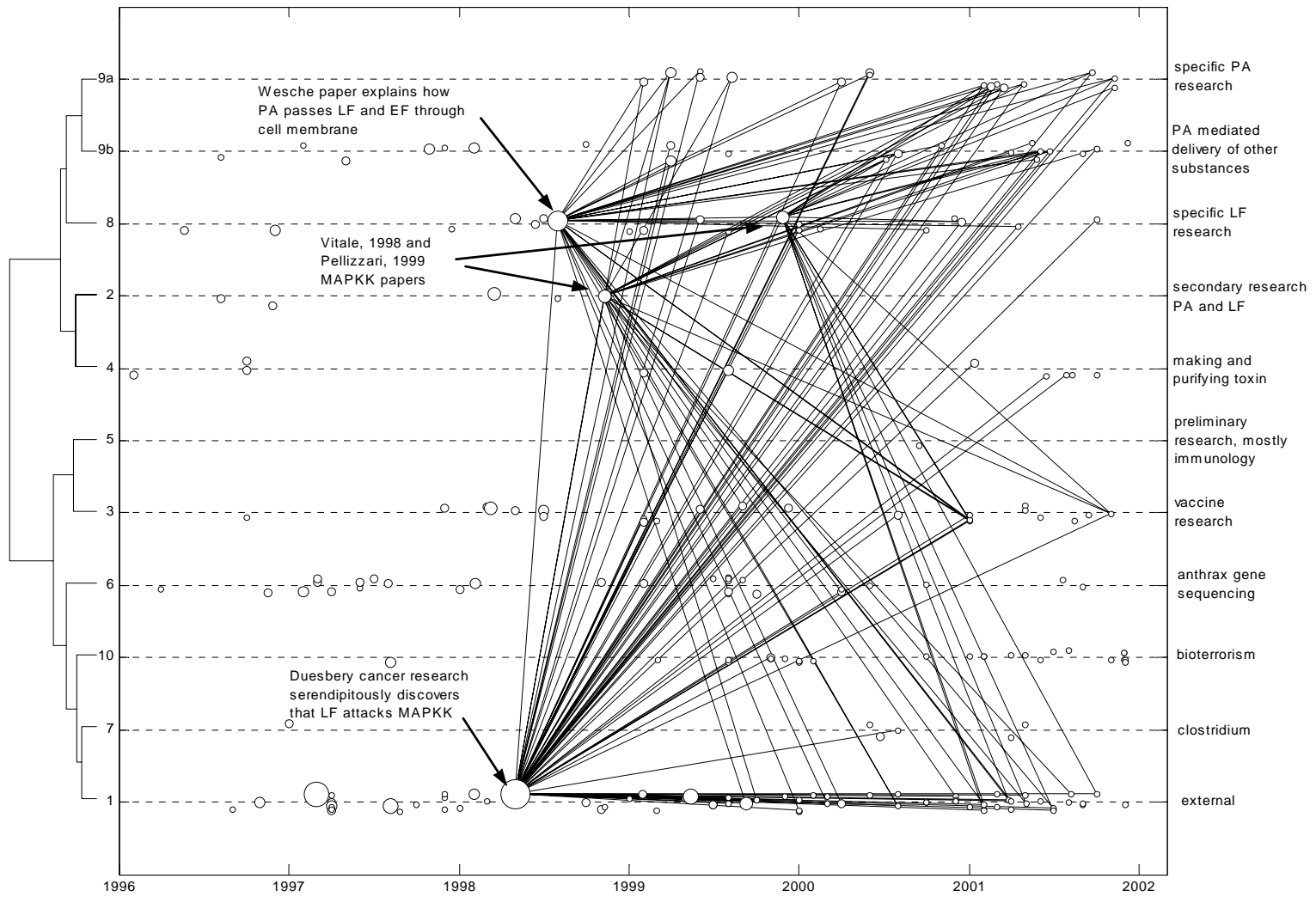


Figure 4. Timeline showing important recent documents.

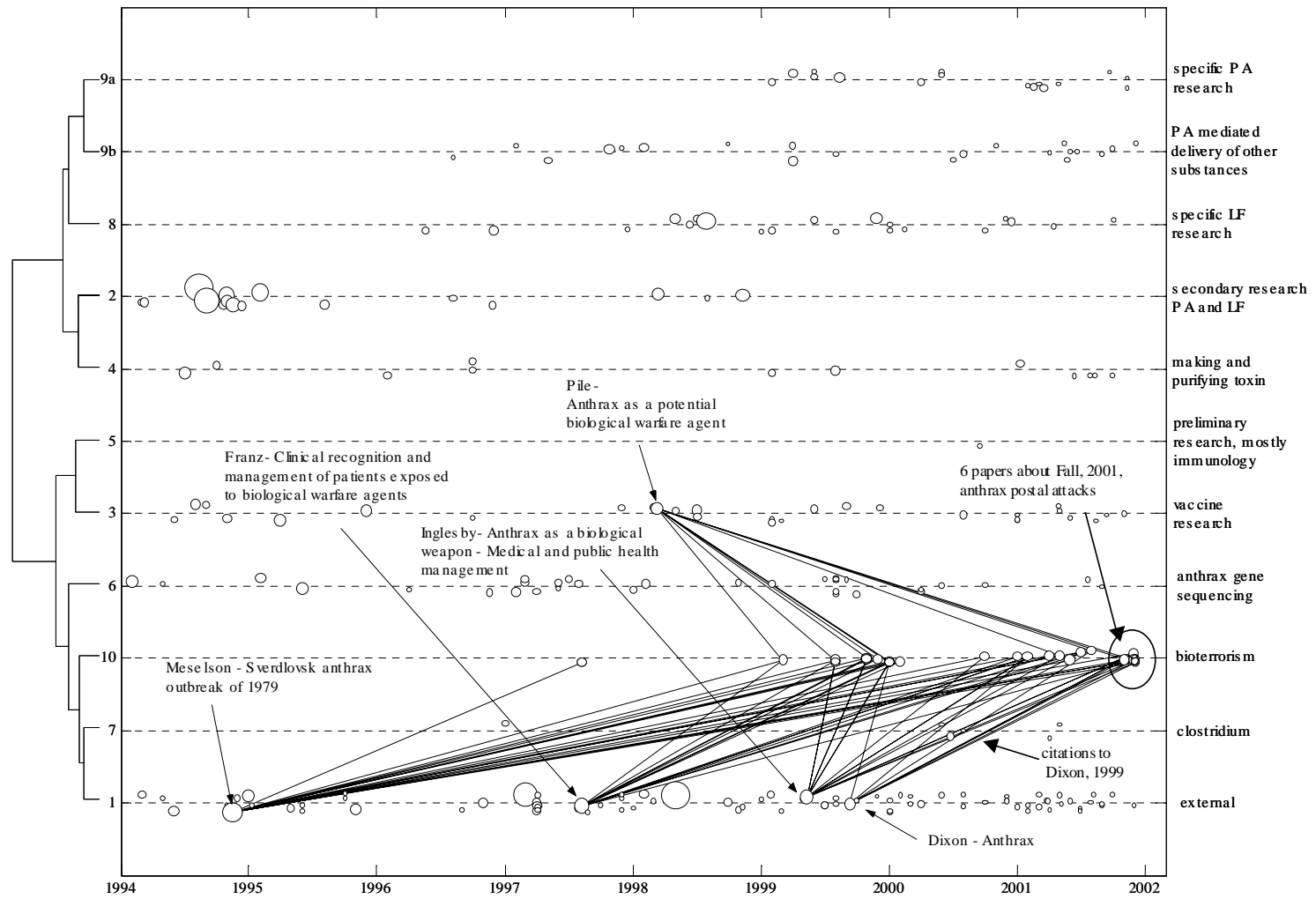


Figure 5. Base documents for track 10, “Bioterrorism.” Recent papers covering 2001 bioterrorism attacks tend to cite Dixon, 1999.