Tracking and Predicting Growth Areas in Science

Henry Small

henry.small@thomson.com
Thomson Scientific
3501 Market, Philadelphia, Pa. 19104 (USA)

Abstract
We explore the possibility of using co-citation clusters over three time periods to track the emergence and growth of research areas, and predict their near term change. Data sets are from three overlapping six-year periods: 1996-2001, 1997-2002 and 1998-2003. The methodologies of co-citation clustering, mapping, and string formation are reviewed, and a measure of cluster currency is defined as the average age of highly cited papers relative to the year span of the data set. An association is found between the currency variable in a prior period and the percentage change in cluster size and citation frequency in the following period. Various approaches to measuring growth and change in research fields are discussed.

Measuring the growth of research areas

The identification and tracking of research areas has been a perennial theme in scientometrics. By a research area we mean a group of documents or other bibliometric units that define a research topic and an associated group of researchers who share a research interest in the topic. This definition of research area thus involves both content and social aspects. Earlier work in sociology of research areas emphasized the formal and informal communication ties among the practitioners (Griffith & Mullins), although these ties were not in and of themselves sufficient to define such areas. By the same token, a group of papers on the same topic may not constitute a research area if there is no communication among its practitioners. Various types of data and techniques have been advocated for the purpose of delineating research areas including Garfield's historiographs (2004), document co-citation (Small 1973) or author co-citation (White & Griffith), co-word analysis (Callon, et al), and journal mapping (Leydesdorff). Techniques focus on a wide range of clustering, ordination and multivariate methods.

Having deployed such tools the question often arises whether there is any predictive value in them, that is, can they help predict the emergence or growth of scientific areas? Related questions are whether scientific discoveries can be predicted from the existing state of knowledge (Swanson & Smalheiser) or the more limited question of whether we can predict if an existing research area will advance or grow. Or is the best we can expect to achieve a retrospective monitoring and tracking of developments?

Types of leading bibliometric indicators one might postulate are temporal or structural. Examples of temporal or time based indicators are Price's index (1970), that can be framed in a general way as the percentage of papers or references in a specific time period, the mean or median age of papers such as so-called half-life measures, or immediacy indicators such as the immediacy index (Garfield 1972). Structural indicators, which we will not deal with here, might also be considered, such as the linkage density, centrality, and interdisciplinarity.

The rationale for temporal indicators as predictors of growth is that a new discovery or development is likely to quickly attract the attention of researchers in the field, and the field will expand rapidly as scientists publish new papers building on and citing the original discovery papers. This would suggest that the more current highly cited papers in a research area, the more likely the area will grow rapidly in the near future. We will propose a specific definition of currency based on the mean age of highly cited papers relative to a specific time frame.

Numerous earlier studies have focused in one way or another on this question, without however providing a definitive answer. The logistic curve proposed by Price (1961) and used by Crane (1972) to describe the growth of research areas is in effect a model of how much an area will grow in the

Proceedings of ISSI 2005, pages 13–23.
Edited by P. Ingwersen and B. Larsen.

future. An early attempt to find a leading indicator of growth points in science is found in the work of Meadows (1971). Goffman (1971) proposed an epidemic model of the growth and prediction of research areas in pulses which was later tested by Wagner-Dobler (1999). Tabah (1992) attempted to use chaos theory to model the growth of literature in different areas.

Co-citation clusters of high currency have also been suggested as leading indicators of specialty growth. Griffith and Small (1974) observed that co-discoveries in science are sometimes marked by the emergence of pairs of highly co-cited papers, such as the Temin and Baltimore co-discovery of reverse transcriptase. Merton's observation (1963) that many discoveries are in fact multiples to varying degrees can be interpreted in a bibliometric context to mean that a group of related papers marks the emergence of a significant discovery, rather than an isolated paper.

Methodology for Delineating Research Areas

Of course to study the emergence and growth of research areas, we need a method for delineating them, which in this case is co-citation clustering. In a nutshell highly cited papers, defined as the top one percent (1%) of papers in each of 22 broad disciplines, are the basic units of analysis (for field definitions used see: http://www.in-cites.com/field-def.html). We assemble these units into co-citation networks through a series of clustering operations, yielding clusters at various levels of aggregation. We then track these objects over time by looking at successive time slices of data to determine the pattern of continuing highly cited papers from one set to the next. The threads of continuity are referred to as cluster strings (Small 1977).

Three data sets of co-citation clusters were used representing three overlapping six year time frames: 1996-2001, 1997-2002 and 1998-2003. Both cited and citing items are restricted to these time spans. All co-citations among the selected highly cited papers are computed. To perform clustering, a threshold is set on the normalized co-citation coefficient (cosine similarity) which determines the selection of linked papers. The single-link cluster analysis gathers together the links that share common papers. To prevent chaining a maximum cluster size is set with a provision for re-clustering at a higher threshold.

Table 1 gives statistics on the three data sets used: the number of clusters, highly cited papers, average citations per paper, and average publication year of papers.

Table 1: Statistics on clusters in three time periods

	1996-2001	**1997-2002**	**1998-2003**
# Clusters	5,005	5,221	5,269
# Papers	20,395	21,183	21,315
Cites per paper	70.6	74.9	76.9
Average year	1998.5	1999.5	2000.6

Once a cluster of co-cited documents is formed, a map or visualization can be created. On these maps the area of the circle is proportional to the citation frequency of the paper and the arrangement of circles reflects an attempt to minimize the net force on each node. Each co-citation link is modeled as an attractive force proportional to normalized co-citation and varying linearly with distance between nodes (Fruchterman & Reingold 1991). In addition there is a repulsive force among all nodes varying as the inverse square of the distance. Nodes are moved iteratively to minimize the forces acting on them, giving an average residual force per node. For a front of 30 or so papers there might be a few hundred co-citation links that go into determining node positions. However, for clarity only the strongest links for each node are displayed, and dotted lines are used to represent a minimal spanning tree between any remaining separate components. Figure 1 shows a map for a cluster on webometrics from the period 1997-2002 containing 13 papers. Papers from the most recent year (2001) are darker in color.

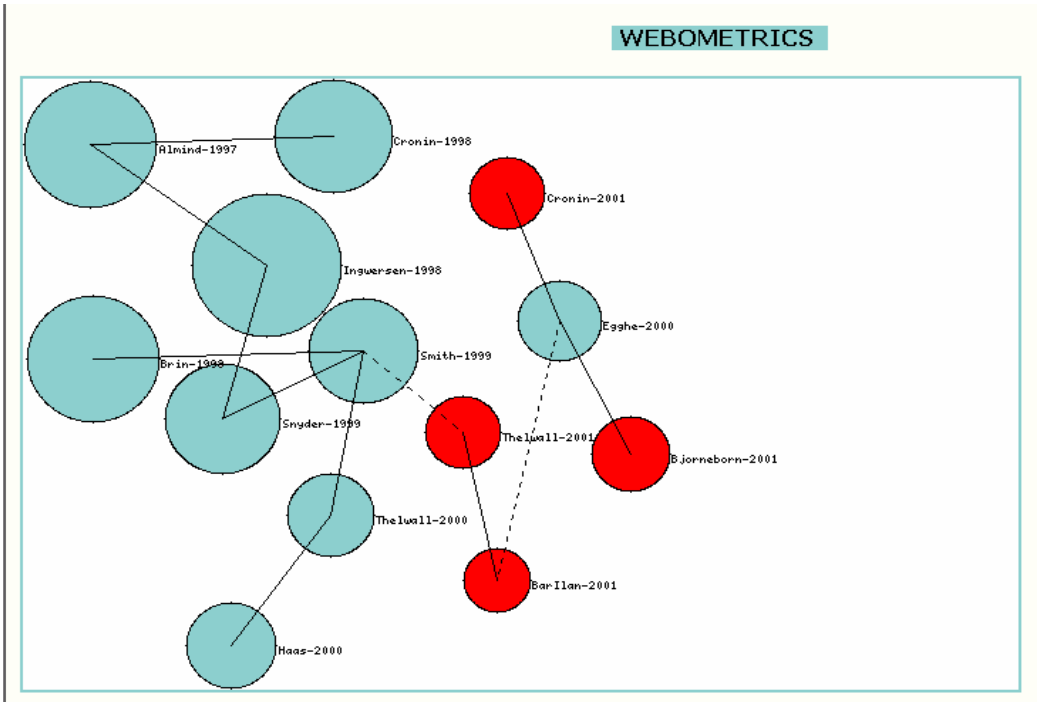

Figure 1 – Map of webometrics 1997-2002 (force/node = 0.28)

After a mapping is carried out in one time period, change over time is studied by moving the time frame period by period, and tracking the highly cited papers from one period to the next to create cluster strings.

The majority of strings are simple, linear continuations that extend over the three time periods. Two examples of simple strings (Figure 2) relevant to the field of scientometrics are small-world networks and webometrics. Both show substantial growth and have a high initial currency as defined below. The string diagram is chronological from left to right, and the size of the circle is proportional to the number of highly cited papers in the cluster (shown in the center of the circle).

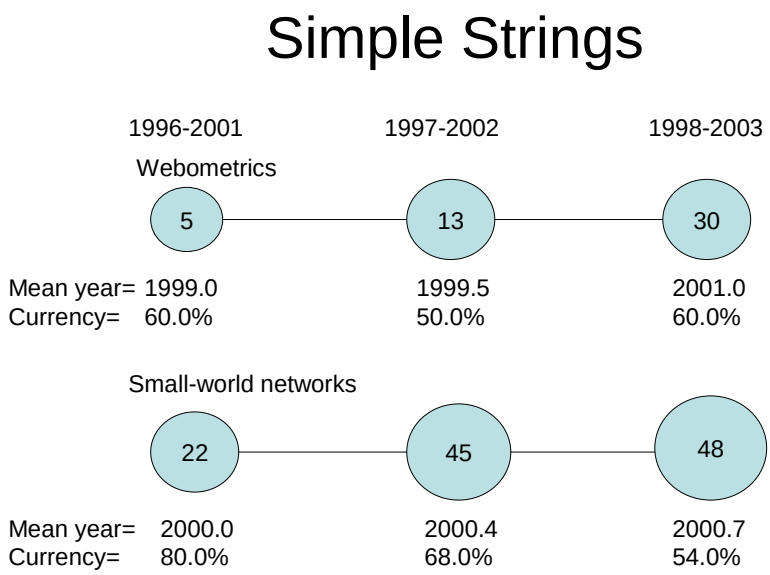

Figure 2 – Two simple strings in scientometrics [maps available from author]

Higher Level Clusters

The first level document clustering and mapping described above shows the structure of specialized scientific areas. To see the structure of larger fields or disciplines, we carry out another clustering and mapping process. Higher level structure is obtained by treating each initial level cluster as an object, re-computing the links between pairs of these objects, and re-clustering. This provides, in effect, a telescoping of levels progressing from documents to specialties to disciplines, in a hierarchical schema. Not all first level objects have links to other objects at that level, reflecting differing degrees of isolation of research areas. It is possible to reintroduce isolates at each higher level to increase recall. To illustrate a hierarchy we start with a cluster on carbon nanotubes show in Figure 3.

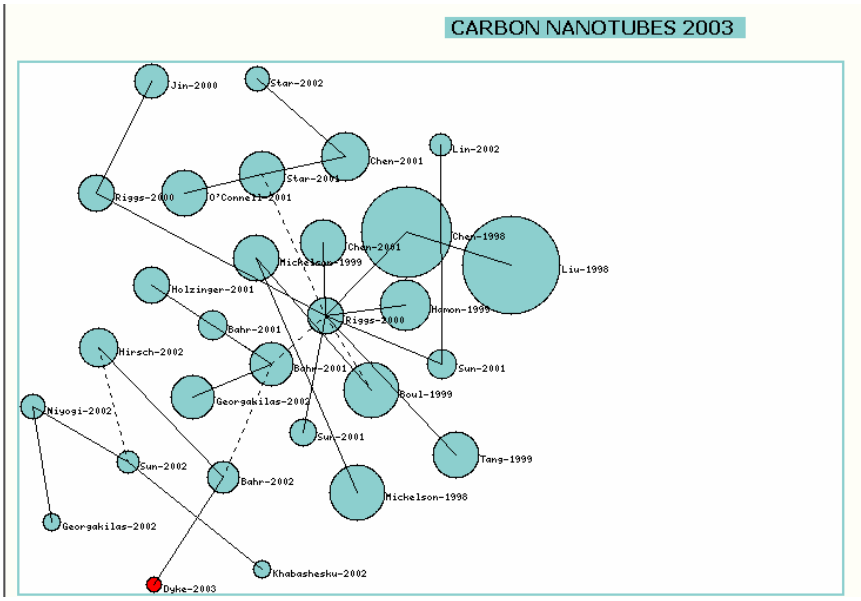

Figure 3 – Carbon nanotube map 1998-2003 (force/node = 0.28)

Treating this group as a single object and collapsing its nodes, we compute its links to other objects at the same level. With another iteration of clustering the subject matter broadens, identifying a group of specialized areas in nanoscience generally (Figure 4), containing the above nanotube cluster as a single circle.

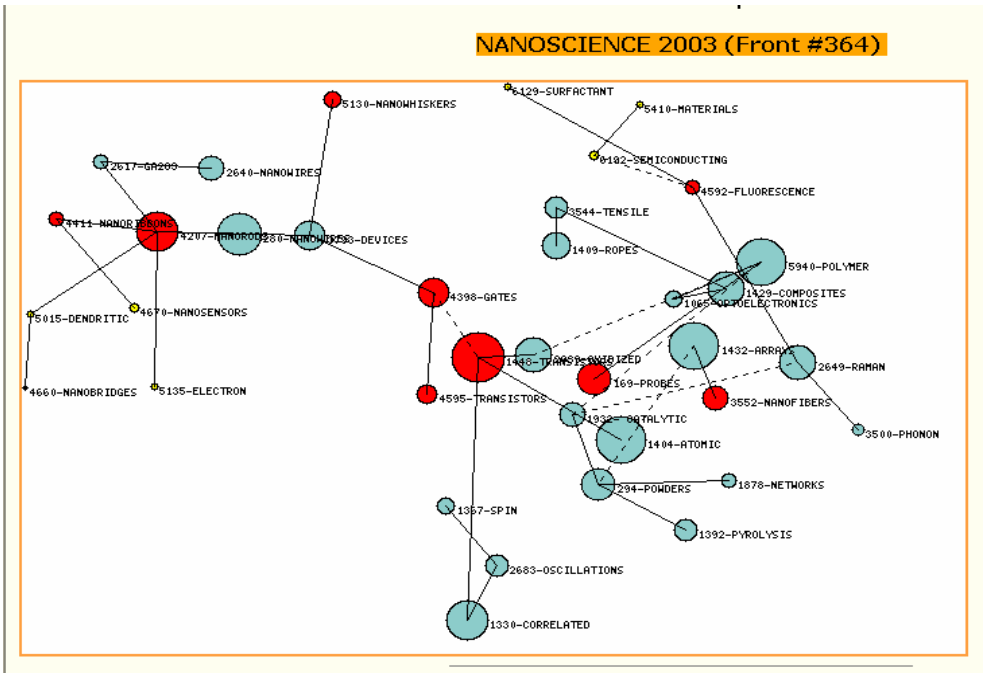

Figure 4 – Level 2 map for nanoscience 1998-2003

This second level map shows areas dealing with a variety of nanostructures, including wires, rods, ribbons, and other nanostructures with electronic or chemical properties. The first level carbon nanotube cluster is an object on this higher level map (#5940). Another iteration of clustering leads to a disciplinary view, in this case of materials science (Figure 5) containing the level 2 cluster as a single circle (#364).

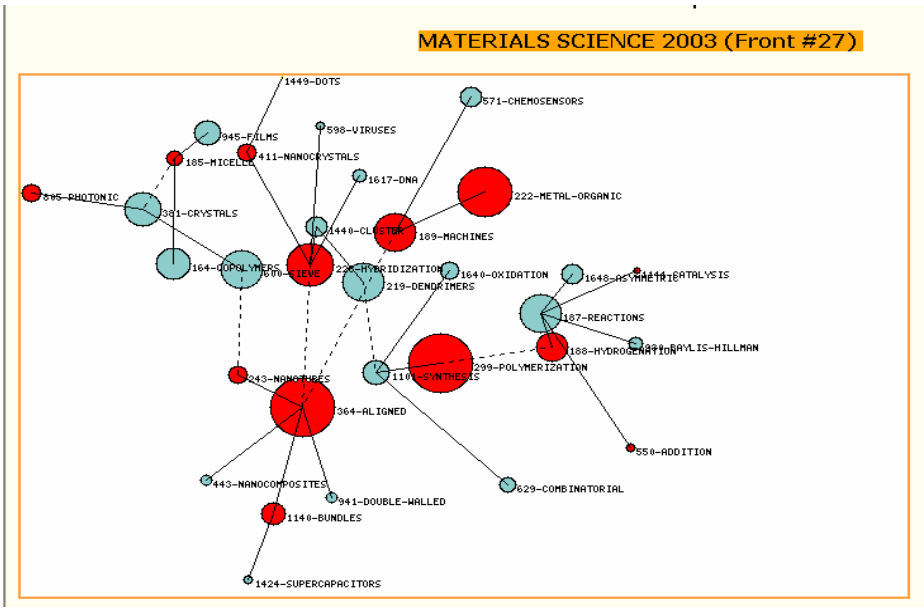

Figure 5 – Level 3 map on materials science 1998-2003

On this third level map we find regions dealing with molecular machines, multilayer films, block copolymers, catalysis, crystals, synthetic chemistry, and DNA as an electronic device. At yet higher levels we would reach a map of science, consisting of other cohesive segments of science that have residual co-citation links.

Cluster strings can also be formed for higher level clusters in the same manner as first level strings. These higher level strings for nanoscience display a splitting or twigging pattern across the three time periods coincident with the emergence of the second level nanoscience maps. In an earlier study on AIDS (Small & Greenlee) this twigging phenomenon was related to the progression of the research area from a single specialty within immunology to a separate biomedical discipline with its own higher level map.

New Clusters

At the opposite extreme of continuity we have new clusters which have no continuing papers from the prior period. Comparing the 1997-2002 dataset against 1998-2003 reveals 37% "new" clusters in the later period. Table 2 lists the five largest new clusters in 1998-2003 by number of highly cited papers, excluding "single issue" cases as explained below.

Table 2: New clusters 1998-2003

Cluster description	**Papers**
SEVERE ACUTE RESPIRATORY SYNDROME (SARS)	25
SONOGASHIRA COUPLING REACTION	23
ORAL DIRECT THROMBIN INHIBITORS	14
HEAT-SENSITIVE TRP CHANNELS (TEMPERATURE SENSING)	11
TWO-DIMENSIONAL GAS CHROMATOGRAPHY	11

The largest new area to emerge in 2003 was SARS consisting of 25 highly cited papers. All but two of its papers are published in 2003. Figure 6 is a map of this cluster with 2003 papers a darker color.

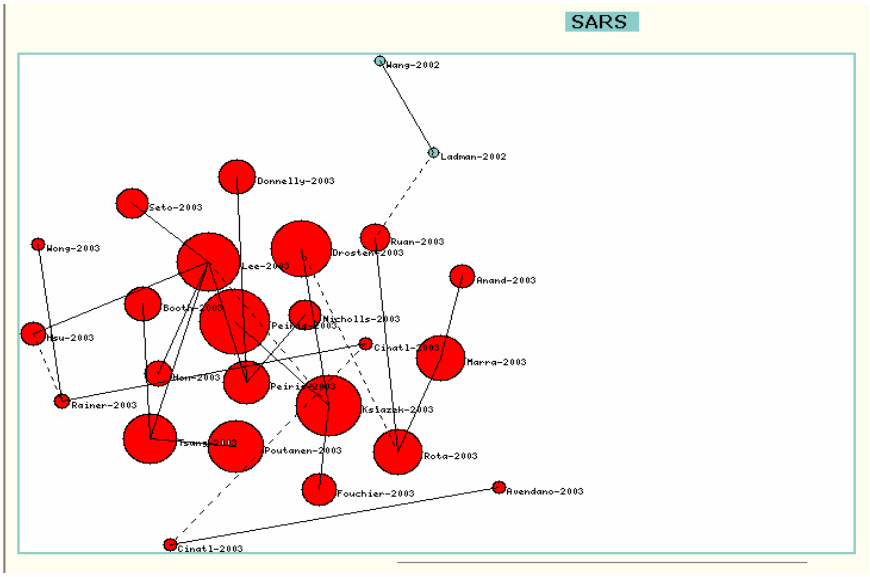

Figure 6 – Map of emerging cluster on SARS 1998-2003

The map has a central region with papers describing the initial clinical outbreak, and an outer region with papers on the genome structure of the coronavirus. Moving up one level in the hierarchy we see SARS (near the top) in the context of other infectious diseases such as pneumonia and influenza (Figure 7).

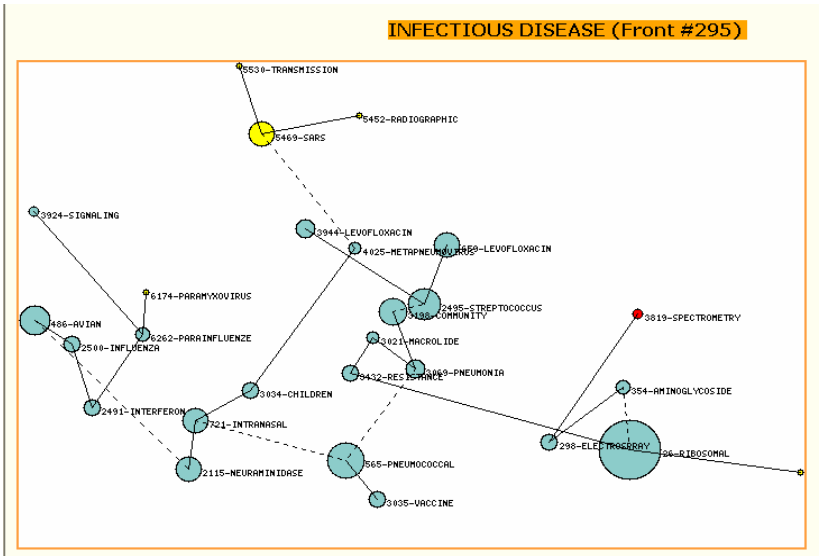

Figure 7 – Level 2 map on infectious diseases containing SARS

Before concluding that a new cluster represents a significant new area of research, it is necessary to examine whether it is the result of a publication artifact such as a special issue of a journal. This comes about when authors of papers in a special issue cite one another's papers within the issue. About 20 of the 90 largest new fronts in the 2003 file having 6 or more papers are to some degree "special issue" clusters, although this fraction declines as the cluster size diminishes. Such clusters usually do not change over time.

Strings and the Analysis of Cluster Change

As noted above, tracking fronts from one time period to the next is based on common or continuing highly cited papers. However, the evolution of fronts is often not a simple process of one area leading to another in a linear fashion. Patterns of development can be complex and involve a combination of branching and merging of clusters, and the death and birth of others.

A strategy to deal with this complexity is to create strings of clusters across time that share common cited papers including all the branching and merging (Small 1977). Strings are created by a separate single-link cluster analysis based on inter-year cluster overlaps. Performing this analysis on the three time periods, and taking into consideration all year to year exchanges of highly cited papers at an inter-year coefficient of 0.1 (cosine coefficient), we obtain 3,651 strings involving first level clusters. The right side of Figure 8 shows the number of strings of different size, that is, the number of first level clusters in the string.

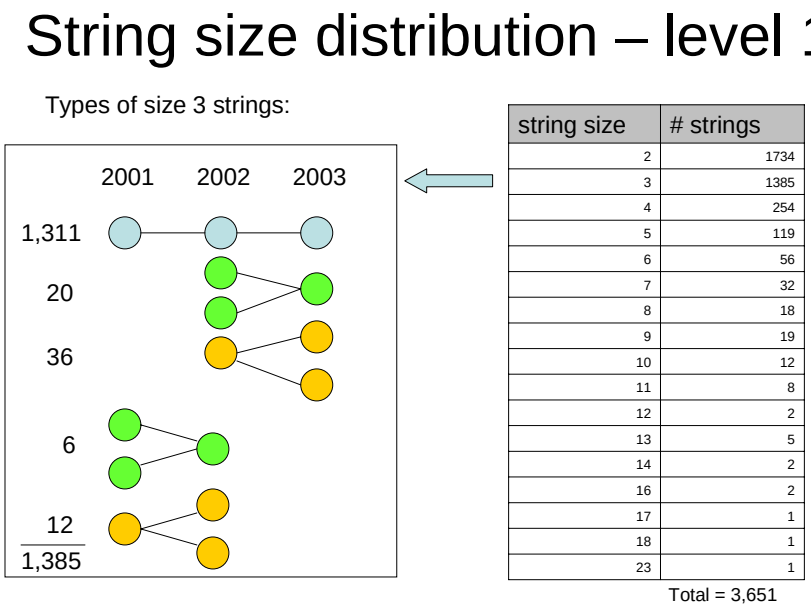

Figure 8 – Statistics on first level cluster strings (2001 – 2003)

Focusing on strings of size three, the left-hand side of Figure 8 shows the different developmental forms possible, with 95% continuing in a simple linear fashion across time. Linear sequences comprise 83% of all 3,651 strings in the three time periods, including strings of two clusters which are by definition linear.

Table 3 gives the topics of the seven largest strings and the number of fronts comprising them. We see a mix of pure and applied physical sciences as well as biomedical sciences. Five of the seven largest strings are from the physical sciences, and two are from biology. Three of the physical science areas are oriented to applied physics and engineering: molecular machines, vertical-cavity lasers, and plasma mass spectrometry.

Table 3 - Largest level 1 strings: 2001-2003

String Description	**# clusters**
STRING THEORY OF D-BRANES	23
LINEAR MOLECULAR MACHINES AND MOLECULAR ELECTRONIC DEVICES	18
ALZHEIMER'S DISEASE-ASSOCIATED PRESENILIN-1 PROTEIN	17
QUANTUM WELL VERTICAL-CAVITY LASERS	16
PEROXISOMAL PROLIFERATOR-ACTIVATED RECEPTOR GAMMA (PPAR GAMMA)	16
LASER ABLATION INDUCTIVELY COUPLED PLASMA MASS SPECTROMETRY	14
POSITIVE MUON ANOMALOUS MAGNETIC MOMENT	14

As an example of a complex string, Alzheimer's disease shows the sudden emergence of a new area in 2002, consisting of 21 papers on the role of the presenilin-gamma-secretase complex (Figure 9).

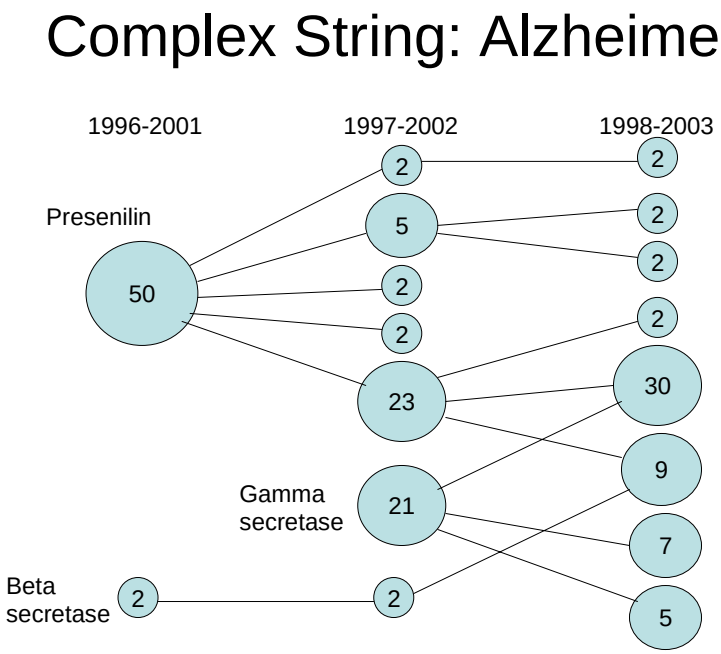

Figure 9 – Complex string on Alzheimer's disease

The map for the gamma secretase cluster (Figure 10) shows a high influx of 2002 papers shown darker in color.

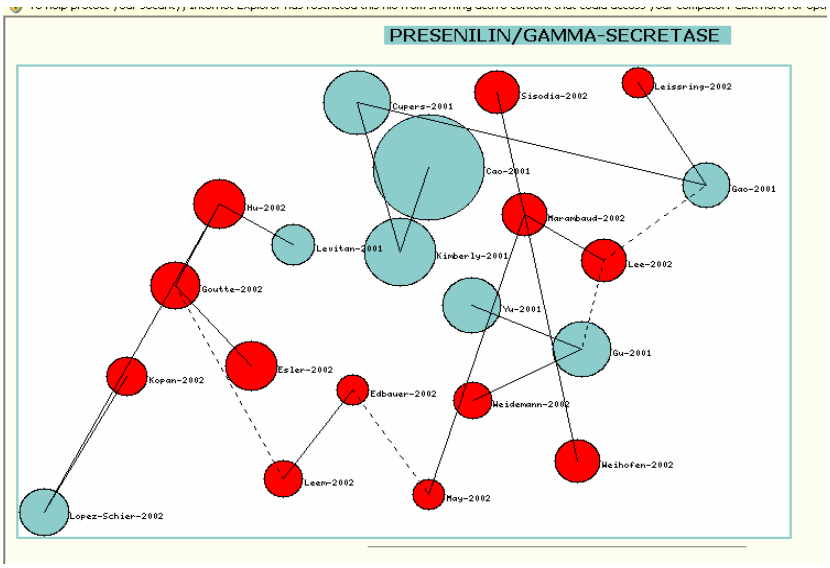

Figure 10 – Map of gamma secretase in Alzheimer's 1997-2002

The Alzheimer's string also displays twigging, the splitting of clusters into various branches. The separation of lines of research corresponds to a proliferation of research topics.

It is often not easy to see whether strings are growing or declining when they contain multiple branching and merging patterns. We can overcome this by summing the sizes of each cluster connected in the string in each time period, and computing the aggregate rate of change of the connected areas. The most rapidly growing areas are shown in Table 4 among all strings in the three

time periods. The fastest growing of these is a physics topic on "time-dependent string backgrounds" which emerged in the 2002 period and grew dramatically in 2003. The table shows the year the first front in the string appeared. The slope is defined as the change in size per year.

Table 4: Fast growing strings: 2001 - 2003

String description	**Slope**	**1st Year**
TIME-DEPENDENT STRING BACKGROUNDS	29.0	2002
BULK METALLIC GLASS	24.0	2002
SMALL INTERFERING RNAS	22.0	2002
COLLOIDAL SEMICONDUCTOR NANOCRYSTALLINE QUANTUM DOTS	21.0	2002
SUPERMASSIVE BLACK HOLES	20.5	2001
FAT-DERIVED HORMONE ADIPONECTIN AND INSULIN RESISTANCE	20.5	2001
SUPERSYMMETRIC EXTRA DIMENSIONS	20.0	2001

Predicting Growth Areas using Strings

To test whether the number of recent papers in an area signals subsequent growth, we first define a measure of this "currency" factor. Since a cluster consists of a set of papers from various earlier years within a defined period, a simple measure of currency is the average year of publication. For comparisons over time or data sets, this mean year is normalized to the year span of the data:

Currency of cluster = 100*(([year span] - 1) – ([most recent year] – [mean year]))/([year span] - 1)

Currency is 100% if the all the papers in the cluster are from the most recent year of the time window, and 0 if all the papers are from the first year of the window.

Table 5 relates the percentage change in string size to the currency of the prior year string, using all year-to-year transitions within strings of all degrees of complexity. To compile these data, the clusters were aggregated by year within each string, summing their sizes, total citations, and the average publication year for the pooled set of papers in the year. The first column is the range of currency grouped in one year intervals. The second column is the number of consecutive year-to-year transitions. The third column is the average percentage increase in string size (number of highly cited papers) from one period to the next. The fourth column is the percentage change in the sum of sizes across all strings in the range. The fifth column is the mean currency of the later year. The last column is the percentage change in total citations. We see for example that strings with a currency between 80 and 100 percent, of which there are 616 cases have on average a growth rate in highly cited papers of 35.4% and 306.8% in citations, while strings having a currency 20% or less decline in size by an average of 11% with only a 5.3% increase in citations. The results suggest that strings with higher prior year currency have a larger percentage increase in size the next year than strings with lower prior year currency. We also note that currencies in general diminish in the second year if the currency is higher the prior year.

Table 5: All strings 2001-2002-2003

Currency Range in Year-1	**#string-Year Pairs**	**Mean % Change In Size**	**% Change Total Size**	**Mean Currency In Year-2**	**% Change In Citations**
<=100% & >80%	616	+35.4%	+19.9%	73.8%	306.8%
<=80% & >60%	1,119	+18.8%	+12.6%	55.6%	89.1%
<=60% & >40%	1,489	+9.4%	+0.21%	37.9%	45.9%
<=40% & >20%	1,511	+0.33%	-12.4%	20.4%	23.3%
<=20% & >0	742	-11.0%	-25.9%	6.7%	5.3%

Since cluster size varies with currency, the fourth column in Table 5 shows the expected changes in size for each currency range which in general are less than the actual average percent changes in size.

Further analysis shows that strings at least doubling in size from one period to the next have an average prior year currency of 66%, while strings that decline in size by more than one-half have a prior year currency of 46%. However, of those that double in size, 5% have a currency of 20% or less, and 8% of strings that decline by more than one-half have a currency greater than 80%. Overall we find a correlation coefficient of only +0.20 between the prior year currency and percentage increase in size, and a chi square test rejects independence of these variables. There is an even stronger tendency for cluster of high prior year currency to receive more citations the following year with a correlation of +0.43, reflected also by the last column in Table 5. This latter finding is however expected due to the commonly observed peak in citations two to three years after publication.

Conclusions

Co-citation clustering, mapping and cluster string formation allow us to study the emergence and growth of research areas from several perspectives. First level clusters and strings can show significant linear growth as in the case of webometrics and small-world networks. More complex and articulated growth is shown by nanoscience which, due to the splitting off of subspecialties, has formed a second level cluster, marking the transition from specialty to discipline. At the other extreme are cases of entirely new areas which appear quite suddenly. An example is SARS in the 1998-2003 period which resulted from a concerted global effort to halt the spread of the virus and elucidate its structure. This cluster showed a remarkable currency in its set of highly cited papers.

The study of change in large and complex strings is facilitated by a separate clustering process on the inter-period linkages. The large Alzheimer's string unexpectedly contained a significant new cluster which emerged suddenly and had a significant effect on the field the next year. A map of this area showed a high proportion of papers from the most recent year. Like nanoscience and AIDS, the Alzheimer's string displayed a high degree of twigging, suggesting the emergence of higher level structure. The growth rate of such complex strings is measured by aggregating the highly cited papers in each branch of the string by year, and looking at the time series of aggregate cluster sizes.

Since many cases of emerging and growing clusters have a high concentration of recent highly cited papers, this feature may be a leading indicator of growth. A measure of currency was defined as the average publication year normalized with respect to the year span of the data set. The percentage increase in year-to-year cluster size was then compared with various ranges of the prior-year currency measure. The result was that there is a slight tendency for clusters of high currency to grow more rapidly the following year than clusters of low currency, but the overall correlation was low. The

correlation is higher for the increase in citations. Thus, it appears then that currency has some limited value as a short term predictor of year to year growth in research areas.

The relationship of currency to growth can be understood if we consider that an area with new discoveries or findings will expand more rapidly from a smaller base than an older area whose rate of development may have slowed. The latter would be expected to grow only if it experienced rejuvenation through new findings or grew though a merger or combination with other older areas. Further work will be needed to identify other factors that might contribute to short term growth, and to understand the longer term dynamics of research areas.

References

Callon, M., Courtial, J.P., Turner, W.A., Bauin, S. (1983). From translations to problematic networks – an introduction to co-word analysis. *Social Science Information*, 22(2), 191-235.

Crane, D. (1972). *Invisible Colleges: Diffusion of Knowledge in Scientific Communities.* Chicago: The University of Chicago Press.

Fruchterman, T.M.J., Reingold, E.M. (1991). Graph drawing by force-directed placement. *Software – Practice and Experience*, 21(11), 1129-1164.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178(4060), 471 - 479.

Garfield, E. (2004). Historiographic mapping of knowledge domains literature. *Journal of Information Science*, 30(2), 119-145.

Goffman, W., Harmon, G. (1971). Mathematical approach to prediction of scientific discovery. *Nature*, 229 (5280), 103-104.

Griffith, B.C., Mullins, N.C. (1972). Coherent social groups in scientific change. *Science*, 177, 959-964.

Griffith, B.C., Small, H., Stonehill, J., Dey, S. (1974). The structure of scientific literatures II: toward a macro- and microstructure for science. *Science Studies*, 4, 339 – 365.

Leydesdorff, L. (2004). Top-down decomposition of the journal citation report of the social science citation index: graph- and factor-analytical approaches. *Scientometrics*, 60(2), 159-180.

Meadows, A.J., O'Connor, J.G. (1971). Bibliographical statistics as a guide to growth points in science. *Science Studies*, 1, 95-99.

Merton, R.K. (1963). Resistance to the systematic study of multiple discoveries in science. *European Journal of Sociology,* 4: 237-249. Reprinted in: *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: University of Chicago Press, 1973, pp. 371-382.

Price, D.J.D. (1961). *Science Since Babylon.* New Haven, Conn: Yale University Press.

Price, D.J.D. (1970). Citation measures of hard science, soft science, technology, and nonscience. In: *Communication Among Scientists and Engineers*. Nelson, C.E., Pollack, D., eds. Lexington, Mass, D.C. Heath & Co.

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 265-269.

Small, H. (1977). A co-citation model of a scientific specialty: a longitudinal study of collagen research. Social Studies of Science, 7, 139-166.

Small, H., Greenlee, E. (1990). A co-citation study of AIDS research. In: *Scholarly Communication and Bibliometrics*. C. Borgman, ed. London: Sage Publications, pp. 166-193.

Swanson, D.R., Smalheiser, N.R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial intelligence*, 91(2), 183 – 203.

Tabah, A.N. (1992). Nonlinear dynamics and the growth of literature. *Information Processing and Management*, 28(1), 61-73.

Wagner-Dobler, R. (1999). William Goffman's "Mathematical approach to the prediction of scientific discovery" and its application to logic, revisited. *Scientometrics*, 46(3), 635-645.

White, H.D., Griffith, B.C. (1981). Author cocitation: a literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32, 163-171