Creation of a highly detailed, dynamic, global model and map of science

Kevin W. Boyack* and Richard Klavans**

* kboyack@mapofscience.com
SciTech Strategies, Inc., Albuquerque, NM 87122 (USA)

** rklavans@mapofscience.com
SciTech Strategies, Inc., Berwyn, PA 19312 (USA)

Abstract
The majority of the effort in metrics research has addressed research evaluation. Far less research has been done to address the unique problems of research planning. Models and maps of science that can address the detailed problems associated with research planning are needed. This article reports on the creation of an article-level model and map of science covering 16 years and nearly 20 million articles using co-citation-based techniques. The map is then used to define discipline-like structures consisting of natural groupings of articles and clusters of articles. This combination of detail and high-level structure can be used to address planning-related problems such as identification of emerging topics, and the identification of which areas of science and technology are innovative and which are simply persisting. In addition to presenting the model and map, several process improvements that result in higher accuracy structures are detailed, including a bibliographic coupling approach for assigning current papers to co-citation clusters, and a sequentially hybrid approach to producing visual maps from models.

Introduction
The majority of the effort in metrics (sciento-, biblio-, infor-, alt-) studies has been aimed at research evaluation (B. R. Martin, Nightingale, & Yegros-Yegros, 2012). Examples of evaluation-related topics include impact factors, the h-index and other related indices, university rankings and national level science indicators. The 40+ year history of the use of document-based indicators for research evaluation includes publication of handbooks (cf., Moed, Glänzel, & Schmoch, 2004), the introduction of new journals (e.g., Scientometrics, Journal of Informetrics) and several annual or biannual conferences (e.g., ISSI, Collnet, STI/ENID) specifically aimed at reporting on these activities. The evaluation of research using scientific and technical documents is a well-established area of research.

Far less effort in metrics research has been aimed at the unique problems of research planning. Planning-related questions (Börner, Boyack, Milojević, & Morris, 2012) that are asked by funders, administrators and researchers are different from evaluation-related questions. As such, they require different models than those that are used for evaluation. For example, planning requires a model that can predict emerging topics in science and technology. Funders need models that can help them identify the most innovative and promising proposals and researchers. Administrators, particularly those in industry, need models to help them best allocate their internal research funds, including knowing which existing areas to cut. To enable detailed planning, document-based models of science and technology need to be highly granular, and while based on retrospective data, must be robust enough to enable forecasting. Overall, the
technical requirements of a model that can be used for planning are unique. Research and development of such models is an under-developed area in metrics research.

To that end, this article reports on a co-citation-based model and map of science comprised of nearly 20 million articles over 16 years that has the potential to be used to answer planning-related questions. Although the model is similar in concept to one that has been previously reported (Klavans & Boyack, 2011), it differs in several significant aspects: it uses an improved current paper assignment process, it uses an improved map layout process, and it has been used to create article-level discipline-like structures to provide context for its detailed structure. In the balance of the article, we first review related work to provide context for the work reported here. We then detail the experiments that led to the improvements in our map creation process. This is followed by introduction and characterization of the model and map, along with a description of how the map was divided into discipline-like groupings. The paper concludes with a brief discussion of how the map will be used in the future for a variety of analyses.

**Background**

Science mapping, when reduced to its most basic components, is a combination of classification and visualization. We assume there is a structure to science, and then we seek to create a representation of that structure by partitioning sets of documents (or journals, authors, grants, etc.) into different groups. This act of partitioning is the classification part of science mapping, and typically takes the majority of the effort. The resulting classification system, along with some representation of the relationships between the partitions, can be thought of as a model of science inasmuch as it specifies structure. The visualization part of science mapping uses the classification and relationships as input, and creates a visual representation (or map) of that model as output.

Mapping of scientific structure using data on scientific publications began not long after the introduction of ISI’s citation indexes in the 1950s. Since then, science mapping has been done at a variety of scales and with a variety of data types. Science mapping studies have been roughly evenly split between document, journal, and author-based maps. Both text-based and citation-based methods have been used. Many of these different types of studies have been reviewed at intervals in the past (Börner, Chen, & Boyack, 2003; Morris & Martens, 2008; White & McCain, 1997). Each type of study is aimed at answering certain types of questions. For example, author-based maps are most often used to characterize the major topics within an area of science, and to show the key researchers in those topic areas. Journal-based models and maps are often used to characterize discipline-level structures in science. Overlay maps based on journals can be used to answer high level policy questions (Rafols, Porter, & Leydesdorff, 2010). However, more detailed questions, such as questions related to planning, require the use of document-level models and maps of science. The balance of this section thus focuses on document-level models and maps.

When it comes to mapping of document sets, most studies have been done using local datasets. The term ‘local’ is used here to denote a small set of topics or a small subset of the whole of science. While these local studies have successfully been used to improve mapping techniques, and to provide detailed information about the areas they study, we prefer global mapping
because of the increased context and accuracy that are enabled by mapping of all of science (Klavans & Boyack, 2011).

The context for work presented here lies in the efforts undertaken since the 1970s to map all of science at the document level using citation-based techniques. The first map of worldwide science based on documents was created by Griffith, Small, Stonehill & Dey (1974). Their map, based on co-citation analysis, contained 1,310 highly cited references in 115 clusters, showing the most highly cited areas in biomedicine, physics, and chemistry. Henry Small continued generating document-level maps using co-citation analysis (Small, Sweeney, & Greenlee, 1985), typically using thresholds based on fractional citation counting that ended up keeping roughly (but not strictly) the top 1% of highly cited references by discipline. The mapping process and software created by Small at the Institute for Scientific Information (ISI) evolved to generate hierarchically nested maps with four levels. Small (1999) presents a four level map based on nearly 130,000 highly cited references from papers published in 1995, which contained nearly 19,000 clusters at its lowest level. At roughly the same time, the Center for Research Planning (CRP) was creating similar maps for the private sector using similar thresholds and methods (Franklin & Johnston, 1988). One major difference is that CRP’s maps only used one level of clustering rather than multiple levels.

The next major step in mapping all of science at the document level took place in the mid-2000’s when Klavans & Boyack (2006) created co-citation models of over 700,000 reference papers and bibliographic coupling models of over 700,000 current papers from the 2002 fileyear of the combined Science and Social Science Citation Indexes. Later, Boyack (2009) used bibliographic coupling to create a model and map of nearly 1,000,000 documents in 117,000 clusters from the 2003 citation indexes. Through 2004, the citation indexes from ISI were the only comprehensive data source that could be used for such maps. The introduction of the Scopus database in late 2004 provided another data source that could be used for comprehensive models and maps of science. Klavans & Boyack (2010) used Scopus data from 2007 to create a co-citation model of science comprised of over 2,000,000 reference papers assigned to 84,000 clusters. Over 5,600,000 citing papers from 2003-2007 were assigned to these co-citation clusters based on reference patterns.

All of the models and maps mentioned to this point have been static maps – that is they were all created using data from a single year, and were snapshot pictures of science at a single point in time. It is only recently that researchers have created maps of all of science that are longitudinal in nature. In the first of these, Klavans & Boyack (2011) extended their co-citation mapping approach by linking together nine annual models of science to generate a nine-year global map of science comprised of 10,360,000 papers from 2000-2008. The clusters in this model have been found to be highly recognizable to subject matter experts (Klavans, Boyack, & Small, 2012). The fact that these small partitions are recognizable units of science suggests that these structures are reasonable representations of the actual topics in science.

More recently, Waltman & van Eck (2012) at CWTS clustered nearly 10,000,000 documents from the Web of Science (2001-2010) using direct citation and a modularity-based approach that is similar to their familiar VOS method. This new CWTS approach has advantages over other approaches: it can be used to generate a multi-level hierarchical clustering, and it can handle very
large document sets with relatively modest computational requirements. Although it has been used with direct citation similarities, there is no reason it could not be used with similarities generated from other methods, such as co-citation, bibliographic coupling, or even text-based or hybrid similarity measures.

**Improvements in Co-citation Model and Map Creation**

We have used co-citation analysis for our models of science for many years because of a combination of several features:

1. Co-citation analysis generates clusters of sufficient accuracy (Boyack & Klavans, 2010).
2. Co-citation analysis enables fractional assignment of current papers, which allows clusters to be linked (within an annual model) through article overlaps (Klavans & Boyack, 2010).
3. On a conceptual level, co-citation analysis generates models where the cognitive structures tend to be highly unstable (Garfield, Malin, & Small, 1978). Instability is a very useful characteristic of a model if it is to be used specifically for the early identification of emergent phenomena.
4. Reference papers (in co-citation clusters) are a natural and accurate basis for linking annual clusters into longitudinal structures (Klavans & Boyack, 2011) that enable exploration of the dynamics of science.

Although our modeling choices have been based on experimentation, it is also true that more feature space remains to be explored, and likely that more accurate and useful models could be generated through additional exploration of that space. Given that we expect our maps to be used to help answer planning-related questions, we continuously seek to improve our modeling and mapping processes to make them more accurate. In this section we report on three efforts to improve the accuracy and usefulness of our co-citation models and maps.

Co-citation analysis consists of two high-level steps – 1) clustering of reference papers into co-citation clusters, and 2) assigning of the current papers to those co-citation clusters. We have recently explored changes in each of these two steps in an attempt to increase the accuracy of our models. Our third effort to improve our processes addresses the step in which a visual map is created from a model.

**Co-citation thresholds**

The first step in any co-citation clustering or co-citation analysis activity is to select the references that are to be clustered. As mentioned previously, early maps of science (cf., Small et al., 1985) used high thresholds, resulting in maps containing approximately the top 1% highly cited references by discipline. The reasoning behind using high thresholds is that the co-citation clusters would be focused on the hottest topics in science. We have used much less stringent thresholds because of our desire to model the average and cold science in addition to the hottest science.

The effect of cited reference thresholds on the accuracy of the resulting co-citation clusters has not been systematically investigated. We thus designed an experiment whose primary intent was to allow us to determine the approximate fraction of references that should be retained to maximize accuracy. In addition, since reference age is known to have an effect on structure (van
Raan, 2005), we also ran several cases designed to show the effect of retaining only the most recent references. Six test cases using different thresholding criteria were chosen, and each was tested using the 2008 fileyear of Scopus data. These six test cases were:

1) CCA-01: include the top 1% highly cited references, regardless of age.
2) CCA-01R: include the top 1% highly cited recent references, from 2002-2008 only.
3) CCA-05: include the top 5% highly cited references, regardless of age.
4) CCA-STS: our published thresholding criteria (Klavans & Boyack, 2011) – include references cited five or more times in a single year; include references published within 3 years with at least \(age + 1\) cites; include references of age 0 cited at least twice. These criteria result in retaining approximately the top 12% highly cited references.
5) CCA-R: include all recent references cited at least twice, from 2002-2008 only.
6) CCA-H: include the most recent half of the references for each citing paper.

All six test cases excluded references that were cited more than 100 times in the 2008 fileyear (about 4,000 references). These references are excluded because they lead to over-aggregation within a cluster solution. We did not apply field normalization for these tests because our current criterion (CCA-STS) already has a relatively low citing threshold. This set of six cases tests the effect of threshold percentages (CCA-01, CCA-05, and CCA-STS), and also tests the effect of recency of references (cases CCA-01R, CCA-R, and CCA-H). We ran these six sets of references through our co-citation analysis process (Klavans & Boyack, 2011), which includes clustering of references, followed by fractional assignment of current papers to the clusters of reference papers using reference lists.

The accuracy of each cluster solution was measured by calculating textual coherence using our established method (Boyack & Klavans, 2010). The premise behind this measure is that a cluster solution whose clusters have more coherent (or less divergent) word distributions is more accurate than a cluster solution with lower coherence. Said differently, this measure assumes that clusters with less textual variation are “better” clusters than those with more textual variation. Coherence is based on the Jensen-Shannon divergence (JSD) (Lin, 1991), which quantifies the divergence between two probability distributions – in this case between the word distribution for a document and the word distribution for the cluster in which the document resides. Calculation of coherence for a full cluster solution requires several steps. First, JSD is calculated for each document and then averaged over all documents in a cluster. Second, since JSD is cluster size dependent, a coherence gain for the cluster is calculated by subtracting the JSD value for a random cluster of the same size. Finally, the coherence for the entire solution is calculated as the cluster size-weighted average of the cluster coherence gains.

A variety of quantities are shown in Table 1, which gives the numerical results of this thresholding study. From the standpoint of usefulness to creating co-citation models of science, we are interested in the solution that produces the best mix of high accuracy (Coh-Pap) and high coverage (%Pap). Coh-Pap is the coherence of the cluster solution based on the cluster assignments of papers from 2008, while %Pap is the percentage of papers with references that are included in the solution.
Table 1. Results of the co-citation analysis thresholding study on six test cases.

<table>
<thead>
<tr>
<th></th>
<th>CCA-01</th>
<th>CCA-01R</th>
<th>CCA-05</th>
<th>CCA-STS</th>
<th>CCA-R</th>
<th>CCA-H</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Ref</td>
<td>208,353</td>
<td>71,217</td>
<td>878,499</td>
<td>2,434,899</td>
<td>2,909,539</td>
<td>9,669,673</td>
</tr>
<tr>
<td>%Ref</td>
<td>1.02%</td>
<td>0.35%</td>
<td>4.32%</td>
<td>11.96%</td>
<td>14.29%</td>
<td>47.50%</td>
</tr>
<tr>
<td>#Clust</td>
<td>20,421</td>
<td>8,230</td>
<td>57,279</td>
<td>97,276</td>
<td>103,090</td>
<td>216,511</td>
</tr>
<tr>
<td>Ref/Clust</td>
<td>10.20</td>
<td>8.65</td>
<td>15.34</td>
<td>25.03</td>
<td>28.22</td>
<td>44.66</td>
</tr>
<tr>
<td>#Pap</td>
<td>1,190,189</td>
<td>683,457</td>
<td>1,383,053</td>
<td>1,479,574</td>
<td>1,418,618</td>
<td>1,575,590</td>
</tr>
<tr>
<td>%Pap</td>
<td>73.65%</td>
<td>42.29%</td>
<td>85.59%</td>
<td>91.56%</td>
<td>87.79%</td>
<td>97.50%</td>
</tr>
<tr>
<td>Pap/Clust</td>
<td>58.28</td>
<td>83.04</td>
<td>24.15</td>
<td>15.21</td>
<td>13.76</td>
<td>7.28</td>
</tr>
<tr>
<td>Coh-Pap</td>
<td>0.07114</td>
<td>0.07321</td>
<td>0.08097</td>
<td>0.08190</td>
<td>0.07846</td>
<td>0.06976</td>
</tr>
</tbody>
</table>

Table 1 shows that coherence and coverage both increase with the fraction of references that are included in the solution (compare CCA-01, CCA-05, and CCA-STS). Both quantities are at their highest values when ~12% of references are included in the solution. The data also show that limiting the set to recent references has a deleterious effect in terms of coverage – some current papers are not included in the solution because they don’t reference recent work. The effect of limiting to current references when using large numbers of references also seems to negatively affect the coherence – CCA-R and CCA-H both have lower coherence values than the CCA-STS set. We note that the coherence values for the ~5% threshold and the ~12% threshold are within one percent of each other – a very small difference. Thus, it is very possible that the peak coherence lies somewhere between the two. However, it is also clear that increasing the number of references included in the calculation also increases coverage of current papers, and coverage is important when using the model to answer detailed questions related to planning. The data show that the CCA-STS thresholding criterion (our current baseline) provides the best solution among the criteria tested in that it has both the highest coverage and the highest coherence solution.

Incorporating bibliographic coupling in the assignment of current papers

The next area of investigation stemmed from a lingering question that was raised in a previous study that compared the accuracy of the cluster solutions calculated using different citation-based techniques (Boyack & Klavans, 2010). One of the unexpected outcomes of that study was that the document clusters created by co-citation analysis had slightly lower coherence than those created by bibliographic coupling. As such, we needed to determine if this justified a complete shift to creating yearly models using bibliographic coupling. It is important to realize that clustering of current papers using bibliographic coupling is done in one step, while clustering of current papers using co-citation analysis is done in two steps – 1) clustering of reference papers into co-citation clusters, and 2) assigning of the current papers to those co-citation clusters. The lower coherence of co-citation solution could have been a consequence of either step.

We focused on whether the reason for lower coherence was due to the second step – the assignment of current papers to co-citation clusters. In the past we have done this using simple fractional assignment based on the distribution of references to clusters for each paper. For example, for a paper with 10 references, if seven of those references appeared in one cluster, and
three in a second cluster, the paper would be assigned to those two clusters with fractions of 0.7 and 0.3, respectively.¹

To the best of our knowledge, all researchers who have assigned current papers to co-citation clusters have used either dominant assignment (assigning the current paper to the single cluster it references most) or some form of simple fractional assignment. We assume this because no mention is made of other assignment techniques. Since different assignment techniques have not been published, we decided to design and test an alternate technique in hopes that it would increase the accuracy (as measured by textual coherence) of the cluster solution. Simple fraction assignment is essentially fractional assignment based on direct citation. Given our previous results showing that bibliographic coupling produced a more accurate clustering than direct citation (Boyack & Klavans, 2010), we decided to create a current paper assignment approach that uses bibliographic coupling. A brief version of the approach (details are in Appendix B) is as follows:

- First, current papers are assigned using the simple fractional assignment (FA) process.
- Second, a bibliographic coupling (BC) solution is created using the methodology from Boyack & Klavans (2010). This gives two distinct solutions of current papers to clusters.
- Third, the two cluster solutions are compared at the paper level. The resulting FA:BC cluster pair sums over papers are used to create new fractional assignments.
- Fourth, for papers that only appeared in one of the two solutions, the assignment from that single solution is used.

We tested this new bibliographic coupling-based assignment approach against the simpler fractional assignment approach using a set of 2.15 million documents (2004-2008) that intersect the Scopus and Medline databases. We used this set, rather than the Scopus 2008 set that was used for the co-citation threshold study, because we had previously calculated both co-citation and bibliographic coupling solutions for these data (Boyack & Klavans, 2010), and they were thus available for use without needing additional work.

To maintain the same basis of comparison that was used in our previous studies (Boyack & Klavans, 2010; Boyack et al., 2011), the accuracies of the two solutions were compared using two different metrics – textual coherence (as detailed above), and a concentration index based on NIH grant-to-article linkages from Medline. The premise for using grant-to-article linkages as a metric for measuring the accuracy of a cluster solution is the assumption that the articles acknowledging a single grant (especially small grants, e.g., R01 grants) should be highly related, and should be concentrated in a cluster solution of the document space. Using this approach, a cluster solution that does a better job of concentrating articles associated with individual grants is more accurate than one that does not concentrating these articles as well. One positive benefit of using grant-to-article linkages is that they are an extrinsic measure of quality – they are not used in the clustering in any way and thus do not bias the results.

¹ Additional detail is needed to replicate the simple fractional assignment process. To avoid assigning papers to very large numbers of clusters with very small fractions each, if a paper has at least two references in any one cluster, the tail of the distribution is truncated by ignoring clusters where the paper only has one reference. We do this because we assume single counts to be potentially random events, while signals based on two or more counts are considered to be non-random events.
The concentration measurement is limited to those grants that can show a concentrated solution. We limited the grant-to-article linkage set to those grants that have produced a minimum of four articles. The resulting test set consisted of 571,405 separate links between 262,959 unique articles (over 12% of the corpus) and 43,442 NIH grants. We use a standard Herfindahl index as a measure of concentration. This is calculated for each grant \( i \) as

\[
H_i = \frac{\sum (n_{i,j} / n_i)^2}{n_i}
\]

where \( n_{i,j} \) is the number of articles acknowledging grant \( i \) in cluster \( j \), and \( n_i \) is the total number of articles acknowledging grant \( i \). An overall value for each cluster solution is then calculated as the weighted average over all grants,

\[
H = n_i H_i / \left( \sum n_i \right)
\]

Table 2 shows that the coherence and the concentration metrics are both significantly higher for our new bibliographic coupling-based assignment approach than for the simple fractional assignment (CC-FA) approach. We label this new approach as CC-BC because it combines standard co-citation clustering of references with a bibliographic coupling-based assignment of current papers. The coherence for the CC-BC approach is also higher than that for a standard bibliographic coupling (BC) approach. Bibliographic coupling alone has a higher concentration index than the CC-BC approach, but not by much. Given the size (over 2 million articles) and scope of the dataset, we take these results as an indicator that a combined co-citation/bibliographic coupling approach is preferable to using either approach separately, and have revised our approach to modeling accordingly.

<table>
<thead>
<tr>
<th></th>
<th>BC</th>
<th>CC-FA</th>
<th>CC-BC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coh-Pap</td>
<td>0.08599 (+5.3%)</td>
<td>0.08167</td>
<td>0.08865 (+8.5%)</td>
</tr>
<tr>
<td>Concentration (H)</td>
<td>0.28486 (+19.8%)</td>
<td>0.23778</td>
<td>0.27516 (+15.7%)</td>
</tr>
</tbody>
</table>

**Sequentially hybrid map layout**

Once a model (i.e., in our case, a cluster solution) has been created from a set of documents, it is often useful to create a visual map of the model. One straightforward way of doing this is to create a graph layout of the clusters in the model. A variety of methods have been devised for this type of visualization. However, the most common layout algorithms in use today (e.g., Fruchterman-Reingold, Kamada-Kawai) are typically only used to generate layouts for small datasets (100s of objects). We use the OpenOrd (formerly DrL) algorithm to generate layouts for sets of hundreds of thousands of objects (S. Martin, Brown, Klavans, & Boyack, 2011).

For many years we have generated visual maps of our large-scale co-citation models using co-citation between pairs of clusters. This choice was based on tradition – for example, Small (1999) used sequential co-citation analysis steps to create a hierarchical set of document clusters. To do this, one takes the original list of citing-cited article pairs, replaces the cited articles with their cluster numbers, and then runs the same algorithm used in the original co-citation similarity calculation, but with cluster numbers as the cited items. This results in a set of cluster-cluster
similarity values based on co-citation at the cluster level. This set of similarity values can then be used as input to a graph layout algorithm to calculate the cluster coordinates which, when plotted, form a visual map of the model. As an example, Figure 1 (left) shows the map of 116,163 article clusters (co-citation clusters) from the 2010 fileyear of Scopus data created using this approach. Although this map shows the relative positions of major areas of science, we have never found this type of map to be as appealing and informative as we would like because everything is so bunched together. There is very little white space in this visual map; there are few grouping of clusters that differentiate themselves from the overall mass that could represent discipline-level structures.

Prior research has shown us that different similarity types give rise to different types of visual cluster patterns. Typical patterns range from the space-filled mass exemplified in Figure 1 (left) to concentrated areas tied together with stringy pathways to very tight clusters with no pathways between them (Boyack, Klavans, & Börner, 2005). White space increases along this continuum. Aesthetically, we favor the middle ground – i.e., cluster solutions that show sets of islands connected by pathways. These types of visual structures imply that there are multiple high level areas in science that are differentiated from each other, and that these are connected.

A test was run to see if a cluster-cluster similarity based on textual analysis might create a different visual cluster pattern, one that was both more appealing and more accurate. Text-based similarity measures require much more computation than citation-based measures because the
cluster-cluster matrix is typically much less sparse for text than for citations. Calculating a cluster-cluster similarity matrix based on textual characteristics for a hundred thousand clusters or more is thus a significant undertaking. Many different text-based measures are available for such a calculation. We chose to use the BM25 measure (Sparck Jones, Walker, & Robertson, 2000a, 2000b) because it is simple to calculate, is among the least computationally expensive text approaches, and is among the most accurate text measures (Boyack et al., 2011; Lin & Wilbur, 2007). Each cluster was represented textually as being comprised of the titles and abstracts of its papers. BM25 was then used to calculate cluster-cluster scores. The BM25 similarity between one object \( q \) and another object \( d \) is calculated as:

\[
s(q,d) = \sum_{i=1}^{n} \frac{IDF_i n_i (k_1 + 1)}{n_i + k_1 \left(1 - b + b \frac{|D|}{D} \right)}
\]

where \( n_i \) is the frequency of term \( i \) in object \( d \). Note that \( n_i = 0 \) for terms that are in \( q \) but not in \( d \). Typical values were chosen for the constants \( k_1 \) and \( b \) (2.0 and 0.75, respectively). In our formulation each cluster was treated as if it were a single document. Document length \( |D| \) was estimated by adding the term frequencies \( n_i \) per document. Average document length \( \overline{|D|} \) is computed over the entire set of documents. The IDF value for a particular term \( i \) is computed as:

\[
IDF_i = \log \frac{N - n_i + 0.5}{n_i + 0.5}
\]

where \( N \) is the total number of documents in the dataset and \( d_i \) is the number of documents containing term \( i \). Each individual term in the summation in the first formula is independent of document \( q \). To remove the influence of high frequency terms all terms with IDF scores below 2.0 were discarded.

Figure 1 (right) shows the visual map of co-citation clusters that resulted from using BM25 to calculate cluster-cluster similarities. All similarity values were not used in the calculation. Rather, the similarity file was filtered to the top-N similarities per cluster, and layout was done with OpenOrd using the default edge cutting parameter. The number of similarities \( N \) per cluster varied between 5 and 15, and was scaled to \( \text{sum(BM25)} \) using the top50 similarities per cluster. These same steps were also used for the co-citation map at the left of Figure 1, enabling a comparison of the two maps that can be definitively tied to the differences in similarity measures.

The similarities and differences between these two maps are quite striking. Once the layout is given the same orientation (Physics at the top and the Social Sciences on the left), the relative location of each field is almost exactly the same. Biology is in the middle, Chemistry is on the right, and Medicine is at the bottom. Earth Sciences is above Biology and borders Computer Science and Engineering. Brain Research borders the Social Sciences, Health Sciences and Medicine in both maps. The ordering and relative locations of large fields are not affected by the use of a co-citation vs. text-based cluster-cluster similarity measure. This is to be expected given the consensus in high level map structures that has been recently noted by multiple researchers (Klavans & Boyack, 2009; Leydesdorff & Rafols, 2009).
The major difference between these two maps is the existence of interior white space in the text-based map on the right. Interior white space is an extremely attractive characteristic in that it may represent the between-cluster gaps that are at the heart of a network-based theory of innovation (Chen et al., 2009). The BM25 similarity values are an order of magnitude higher than the co-citation similarity values, as shown in Figure 2, which suggests that the lower similarities lead to much more even spacing between nodes in the map, while higher similarity values create a map with much more well defined groupings of clusters. This should also not be surprising. The first level of clustering in both maps was based on co-citation, using up a high fraction of the variance in the system that could be accounted for using co-citation. A second level of similarity using co-citation should thus have far less signal available with which to link clusters than would textual analysis, the majority of whose signal would still be available.

![Figure 2. Similarity value distributions for the two map layout methods.](image)

We find the text-based map to be far more visually compelling than the co-citation map because of the localized density of clusters, the greater amount of white space, and the visible pathways between localized areas that indicate connections between discipline-like structures. This new mapping technique can be considered as a hybrid technique. Although the first level similarity metric is not a citation+text hybrid, this technique uses a citation-based method to generate clusters, followed by a text-based method to generate a cluster layout, and is thus a *sequentially hybrid* map layout technique.

The arguments above speak to the aesthetic nature of the sequentially hybrid map, and give anecdotal arguments as to why this approach may capture more of the variance in the system than using two sequential co-citation steps. In addition to anecdotal evidence, we present some numerical evidence that the sequentially hybrid layout is more accurate than the map based purely on co-citation analysis. For each map in Figure 1, all possible pairs of papers with a common author were located. For instance if a researcher authored 4 papers (A,B,C,D), the set of
distinct pairs of those papers are AB, AC, AD, BC, BD, CD. A total of 12,812,133 distinct (and
de-duplicated) pairs of papers linked through common authors were identified, and the distances
between each pair in map units were calculated. Map units are shown in Figure 1. Each map was
normalized to 1000 units in both the x and y directions to allow direct comparison between the
two maps. Figure 3 shows that a significant fraction of the distances in the sequentially hybrid
(BM25) layout are shorter than those in the citation-based layout. The distribution for the text-
based layout is bimodal, with the first peak at 1.5 log units and the second peak at 5.5 log units,
while the citation-based layout has only a single peak at 5.3 log units with a tail at the leading
end. To get a sense of what the log units mean in practice, Figure 1 (right) shows log unit scales;
1.5 log units has a length on the map of about 1/5 the length of the 3.2 log unit bar. The
sequentially hybrid layout, on average, clearly places papers by the same author closer to each
other within the map than does the citation-based layout. Assuming that multiple papers by an
author should be closer together on a map rather than further apart, this suggests that the
sequentially hybrid map is more accurate, as well as more appealing, than the citation-based
map.

![Figure 3. Distribution of distances between papers authored by the same researcher for the
two map layout methods.](image)

**A Dynamic Global Model and Map of Science**

*Model creation and characterization*

The improvements to science mapping described in the previous sections can be used to generate
more accurate models and maps of science. These improvements have been incorporated into our
processes, and have been used to create a new model and map of science using a 16-year (1996-
2011) set of Scopus data comprised of nearly 20 million documents. Although the entire Scopus
data from those years contains 25.6 million records, only 20.6 million of those have references,
thus comprising our basis set.
The model and map were created in several steps, which are briefly mentioned here. Detailed protocols can be found in Appendixes A and B.

- Annual models are created for each year. Each annual model is calculated by first generating clusters of cited references (co-citation clusters), and then by assigning the current year papers to the co-citation clusters using the bibliographic coupling approach introduced earlier in this paper.
- Annual models are then linked into longitudinal structures (called *threads*) by linking clusters of documents from adjacent years together using overlaps in the cited references belonging to each cluster.

Sixteen annual models were created using this process for the 16-year period from 1996-2011. Table 3, which contains numbers of papers and clusters by year, shows that the process is relatively stable in terms of cluster sizes and the fraction of articles covered by each annual model. Roughly 4% of the articles containing references are missing from the models. The majority of these articles were missing because they did not contain any references that were included in the co-citation calculations.

<table>
<thead>
<tr>
<th>Year</th>
<th>#Clust</th>
<th>#Pap</th>
<th>Pap/Clust</th>
<th>%Pap</th>
<th>#Ref</th>
<th>Ref/Clust</th>
<th>FwdCos</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>54,221</td>
<td>752,442</td>
<td>13.88</td>
<td>95.2%</td>
<td>1,072,014</td>
<td>19.77</td>
<td>0.2593</td>
</tr>
<tr>
<td>1997</td>
<td>56,225</td>
<td>774,390</td>
<td>13.77</td>
<td>95.3%</td>
<td>1,108,296</td>
<td>19.71</td>
<td>0.2578</td>
</tr>
<tr>
<td>1998</td>
<td>57,434</td>
<td>788,643</td>
<td>13.73</td>
<td>95.3%</td>
<td>1,149,310</td>
<td>20.01</td>
<td>0.2572</td>
</tr>
<tr>
<td>1999</td>
<td>59,048</td>
<td>808,027</td>
<td>13.68</td>
<td>95.3%</td>
<td>1,215,370</td>
<td>20.58</td>
<td>0.2605</td>
</tr>
<tr>
<td>2000</td>
<td>64,072</td>
<td>876,335</td>
<td>13.68</td>
<td>95.4%</td>
<td>1,333,079</td>
<td>20.81</td>
<td>0.2575</td>
</tr>
<tr>
<td>2001</td>
<td>70,680</td>
<td>965,106</td>
<td>13.65</td>
<td>95.7%</td>
<td>1,447,172</td>
<td>20.47</td>
<td>0.2540</td>
</tr>
<tr>
<td>2002</td>
<td>74,207</td>
<td>1,004,837</td>
<td>13.54</td>
<td>96.0%</td>
<td>1,541,707</td>
<td>20.78</td>
<td>0.2547</td>
</tr>
<tr>
<td>2003</td>
<td>79,657</td>
<td>1,080,103</td>
<td>13.56</td>
<td>96.2%</td>
<td>1,665,590</td>
<td>20.91</td>
<td>0.2535</td>
</tr>
<tr>
<td>2004</td>
<td>90,074</td>
<td>1,212,349</td>
<td>13.46</td>
<td>96.1%</td>
<td>1,854,537</td>
<td>20.59</td>
<td>0.2500</td>
</tr>
<tr>
<td>2005</td>
<td>98,848</td>
<td>1,332,524</td>
<td>13.48</td>
<td>96.2%</td>
<td>2,058,536</td>
<td>20.83</td>
<td>0.2451</td>
</tr>
<tr>
<td>2006</td>
<td>107,197</td>
<td>1,451,006</td>
<td>13.54</td>
<td>96.1%</td>
<td>2,243,455</td>
<td>20.93</td>
<td>0.2385</td>
</tr>
<tr>
<td>2007</td>
<td>113,426</td>
<td>1,546,811</td>
<td>13.64</td>
<td>95.6%</td>
<td>2,360,593</td>
<td>20.81</td>
<td>0.2357</td>
</tr>
<tr>
<td>2008</td>
<td>121,595</td>
<td>1,645,524</td>
<td>13.53</td>
<td>95.6%</td>
<td>2,539,626</td>
<td>20.89</td>
<td>0.2294</td>
</tr>
<tr>
<td>2009</td>
<td>130,701</td>
<td>1,754,603</td>
<td>13.42</td>
<td>96.0%</td>
<td>2,759,731</td>
<td>21.11</td>
<td>0.2230</td>
</tr>
<tr>
<td>2010</td>
<td>135,836</td>
<td>1,807,757</td>
<td>13.31</td>
<td>96.4%</td>
<td>2,930,351</td>
<td>21.57</td>
<td>0.2227</td>
</tr>
<tr>
<td>2011</td>
<td>151,305</td>
<td>2,004,176</td>
<td>13.25</td>
<td>96.6%</td>
<td>3,277,735</td>
<td>21.66</td>
<td></td>
</tr>
<tr>
<td>TOT</td>
<td>1,464,526</td>
<td>19,804,633</td>
<td>13.52</td>
<td>95.9%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The 16 annual models of Table 3 were then linked together into a longitudinal model of science. Pairs of clusters from adjacent years whose cosine (based on overlap of their cited references) was greater than the FwdCos value listed in Table 3 were linked. Linked sets of clusters are called *threads*, examples of which are shown in Figure 4.
Figure 4. Examples of threads – a) three year, b) four year, c) five year with a split, d) four year with a merge.

Table 4 shows the age characteristics of the resulting threads. Roughly 40% of the annual clusters are what we call *isolates*. These are clusters that link neither forward nor backward within the overall model (above a specific threshold – the FwdCos column in Table 3), and can thus be thought of as one-year threads. These are research problems that do not have enough momentum to continue into a second year. We note that the exact fraction of isolates is a function of the linking thresholds used in the model (see Appendix A). However, this fraction does not decrease dramatically with a lowering of the threshold values. *Isolates* are typically among the smallest clusters, while the longest threads are comprised of larger clusters on average. 46% of annual clusters are in threads that last 3 years or longer.

Table 4. Characteristics of threads in the linked co-citation/bibliographic coupling model.  
Thr = threads, $P_{cont}$ = probability of continuing to the next year.

<table>
<thead>
<tr>
<th>Age</th>
<th>#Thr</th>
<th>#Clust</th>
<th>Clust/Thr/Yr</th>
<th>Pap/Clust</th>
<th>$P_{cont}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>596,807</td>
<td>596,807</td>
<td>1.000</td>
<td>10.31</td>
<td>0.265</td>
</tr>
<tr>
<td>2</td>
<td>92,191</td>
<td>189,268</td>
<td>1.026</td>
<td>12.08</td>
<td>0.606</td>
</tr>
<tr>
<td>3</td>
<td>35,309</td>
<td>113,688</td>
<td>1.073</td>
<td>13.44</td>
<td>0.748</td>
</tr>
<tr>
<td>4</td>
<td>17,569</td>
<td>79,481</td>
<td>1.131</td>
<td>14.70</td>
<td>0.829</td>
</tr>
<tr>
<td>5</td>
<td>11,687</td>
<td>69,842</td>
<td>1.195</td>
<td>15.66</td>
<td>0.859</td>
</tr>
<tr>
<td>6</td>
<td>7,914</td>
<td>59,895</td>
<td>1.261</td>
<td>16.50</td>
<td>0.885</td>
</tr>
<tr>
<td>7</td>
<td>5,969</td>
<td>56,049</td>
<td>1.341</td>
<td>17.00</td>
<td>0.896</td>
</tr>
<tr>
<td>8</td>
<td>4,406</td>
<td>50,106</td>
<td>1.422</td>
<td>17.54</td>
<td>0.911</td>
</tr>
<tr>
<td>9</td>
<td>3,812</td>
<td>49,078</td>
<td>1.431</td>
<td>18.26</td>
<td>0.894</td>
</tr>
<tr>
<td>10</td>
<td>3,024</td>
<td>43,337</td>
<td>1.433</td>
<td>18.55</td>
<td>0.889</td>
</tr>
<tr>
<td>11</td>
<td>2,562</td>
<td>41,209</td>
<td>1.462</td>
<td>18.86</td>
<td>0.878</td>
</tr>
<tr>
<td>12</td>
<td>1,791</td>
<td>31,008</td>
<td>1.443</td>
<td>18.86</td>
<td>0.883</td>
</tr>
<tr>
<td>13</td>
<td>1,201</td>
<td>22,785</td>
<td>1.459</td>
<td>19.16</td>
<td>0.894</td>
</tr>
<tr>
<td>14</td>
<td>738</td>
<td>14,995</td>
<td>1.451</td>
<td>19.27</td>
<td>0.911</td>
</tr>
<tr>
<td>15</td>
<td>538</td>
<td>11,734</td>
<td>1.454</td>
<td>19.24</td>
<td>0.916</td>
</tr>
<tr>
<td>16</td>
<td>1,440</td>
<td>35,244</td>
<td>1.530</td>
<td>21.01</td>
<td></td>
</tr>
<tr>
<td>TOT</td>
<td>1,464,526</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This thread model reflects a combination of stable and unstable science. Over 28% of its papers are in threads that are 8 years or older. This forms a very stable core of science at the level of research problems. Only 31% of papers are in isolates; thus, less than one-third of the papers published each year are experiments which are not followed up cognitively in the next year. We realize that many researchers may feel that science is more stable than suggested by this model. It is true that science is quite stable when measured at the level of disciplines. However, it is also true that many scientists work on a variety of research problems and move from problem to problem in search of problems that have staying power. It is at this level that science is unstable.

Table 4 also shows the importance of topic momentum. The “P_cont” column shows the probability that a thread will continue to the next year as a function of thread age. Only 26.5% of threads that are newly born (first year) will continue to a second year. Once a thread is two years old, it has a 60.6% chance of continuing to a third year. Survival rates increase with age. We note that the survival rate decreases slightly for several years after the 8th year. This is an artefact of our threading criterion (see Appendix A) of not allowing two long threads to be retrospectively linked together.

Visual map
A visual map (Figure 5) has been created from the model using the text-based layout method described above. BM25 coefficients were calculated between pairs of threads using their titles and abstracts. Isolates were not included in this calculation; only the 190,151 threads lasting two years or longer were included in the layout. Each article was assigned a color based on the color-to-journal scheme used by the UCSD map of science (Börner, Klavans, et al., 2012), and each thread was colored by dominant article color. Isolates were added into the map later by positioning each isolate next to the thread with which it shared the largest set of fractional paper assignments.

This map is very similar in terms of high level structure to the consensus view of science (Klavans & Boyack, 2009). The so-called hard sciences, engineering, and computer science all appear in the top half of the map, while the medical and related sciences appear in the lower half of the map. Chemistry (blue) is to the right, and the Social Sciences (light orange) are to the left. The map can easily be time-sliced to show growth in the various areas of science over time. Although this is not dynamic in the sense that it does not allow clusters to move in the map from year to year, it does show births and deaths in clusters. Although this map is shown here at high level, smaller sections can be enlarged to show greater detail. It can also be used as a template on which additional information, such as cluster ages, or output from a particular author, journal or institution, can be overlaid (Rafols et al., 2010).
Figure 5. Visual map of the CC-BC 16-year model of science.

**Definition of discipline-like structures**

To make the map more useful for answering high-level questions, and to enable it to be compared with existing discipline level maps, we have partitioned the map into discipline-like groups using a semi-automated process. To do this, we overlaid the map with an 85x85 grid structure, and “grew” disciplines from grid cells that were chosen as discipline seeds. The principle behind this process is to take a grid cell (or a group of adjacent grid cells) and link to it the grid cell with which it has the greatest overlap, and then to repeat that process iteratively until all grid cells have been linked into a group. Since square grid cells have no inherent overlap, for the growth calculation we define the grid cells as being circular such that the square grid cell would fit entirely with its corresponding circular grid cell (see Figure 6). Using circular grid cells, most threads and isolates belong to more than one cell, thus creating natural overlaps between grid cells that can be used as a linkage mechanism. For example, the center cell in Figure 6 has potentially substantial overlaps with the cells above, below, and to either side of it. In our process, if the center cell were chosen as the seed, the cell with the largest number of articles in the overlap space with the center cell would be linked to it. In the second iteration, the
cell with the largest overlap with either the center cell or the first one that was connected to it would then be linked to the group.

Figure 6. Sample of a grid structure showing square and round representations of the grid cells, and overlaps between round grid cells.

We started by identifying a set of grid cells as discipline seeds, as shown at the top left of Figure 7. The first set of seeds was identified automatically by choosing the grid cell in each 5x5 subsection of the map with the largest number of articles. This starting set was modified by hand, deleting seeds where two seeds were too close to each other, and adding seeds on “islands” in the map that did not have seeds. Disciplines were then grown from the seeds over the course of 40 iterations using the linkage process described above. A constraint was placed on the first 20 iterations that required linkages to be between grid cells of the same color. This ensures that the initial, large growth of disciplines was within established high level grouping of science. This restriction was lifted for the final 20 iterations, thus allowing the disciplines to grow in an interdisciplinary way where appropriate.

Figure 7 shows instances the results of the growth process at three different points – after seed identification, after 10 iterations, and after all 40 iterations. After 40 iterations, no more grid cells were being linked to existing groupings, and the calculation was stopped.

With the calculation complete, threads and isolates were assigned to disciplinary groups. To avoid problems with multiple assignments, each thread and isolate was assigned to the single square grid in which it occurs. Circular grid cells were not used for this assignment process. The bottom right panel shows the resulting 211 discipline-like groups that were created using this process. Approximately 250 grid cells with very low paper counts, comprising a mere 0.4% of
the articles in the model, did not end up being linked into any disciplinary group due to lack of overlap. These will be assigned to existing disciplines as needed.

Figure 7. Creating disciplines by dividing the map of science into a series of fine grids, specifying grid points as discipline seeds, and iteratively linking adjacent grid points to those already chosen. Grid dots reflect the number of papers in each grid. Final disciplines and relative sizes are shown at the bottom right.

We realize that the disciplinary structure defined here is somewhat arbitrary, depending largely on the number of seeds that were used. However, it is less arbitrary than one might think. Close examination of the bottom left panel of Figure 7 shows that the density of articles in different grid cells varies widely. It is natural to assume that disciplinary structures will be based on dense
areas within the map, and separated by areas with much lower density. Comparison of the two bottom panels of Figure 7 shows that the majority of the disciplines are associated with areas of high article density, and are separated by areas of lower density. This is also consistent with the principle behind creation of journal-based disciplines, which is that journals are grouped such that the within-group citation density is high and the between-group citation density is low, thus suggesting that the disciplinary groupings shown in Figure 7 are reasonable.

A detailed description of the resulting disciplinary structure is beyond the scope of this article. Nevertheless, we give examples of some disciplinary groups that exemplify different combinations of properties (Table 5). These disciplines were chosen because they provide examples of different types of extremes with respect to properties. For example, Superconductivity and Elementary Particle Physics are examples of disciplines with low growth rates, relatively high stability (a high percentage of long duration threads), and mono-disciplinarity (high correlation with journal-based disciplines). Sleep; Environmental, Energy and Economic Policy; Business Ethics and Strategy; and Literature and Language are examples of the exact opposite combination of properties – high growth rates, relatively low stability, and multi-disciplinarity. In general, growth rates tend to be inversely correlated with stability. Some interesting exceptions to these trends are Psychoanalysis, which has low growth and low stability, and Power Transmission and Distribution, which combines high growth with relatively high stability while being very mono-disciplinary. Nanotechnology exhibits high growth with mid-range stability, while Endocrinology exhibits low growth with mid-range stability.

Table 5. Examples of disciplines identified from the CC-BC map of science. Mono = level of monodisciplinarity, Growth = \#Pap_{2011} / \#Pap_{2007}, \%Unstable = percentage of threads that are isolates, \%Stable = percentage of threads lasting at least 8 years.

<table>
<thead>
<tr>
<th>Discipline Name</th>
<th>#Pap</th>
<th>Growth</th>
<th>Mono</th>
<th>%Unstable</th>
<th>%Stable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superconductivity</td>
<td>111,980</td>
<td>0.79</td>
<td>0.183</td>
<td>61.1%</td>
<td>5.6%</td>
</tr>
<tr>
<td>Psychoanalysis</td>
<td>24,073</td>
<td>0.98</td>
<td>0.159</td>
<td>82.1%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Endocrinology</td>
<td>88,542</td>
<td>1.00</td>
<td>0.021</td>
<td>75.7%</td>
<td>3.0%</td>
</tr>
<tr>
<td>Elementary Particle Physics</td>
<td>108,123</td>
<td>1.07</td>
<td>0.299</td>
<td>57.2%</td>
<td>9.2%</td>
</tr>
<tr>
<td>Sleep</td>
<td>79,948</td>
<td>1.57</td>
<td>0.020</td>
<td>79.4%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Environment, Energy and Economic Policy</td>
<td>97,436</td>
<td>1.59</td>
<td>0.029</td>
<td>82.6%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Nanotechnology</td>
<td>83,449</td>
<td>1.60</td>
<td>0.038</td>
<td>73.7%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Business Ethics and Strategy</td>
<td>182,689</td>
<td>1.64</td>
<td>0.034</td>
<td>82.6%</td>
<td>1.2%</td>
</tr>
<tr>
<td>Literature and Language</td>
<td>108,493</td>
<td>1.71</td>
<td>0.017</td>
<td>84.9%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Power Transmission and Distribution</td>
<td>174,313</td>
<td>1.85</td>
<td>0.163</td>
<td>65.1%</td>
<td>3.3%</td>
</tr>
</tbody>
</table>

The mono-disciplinarity index of Table 5 was calculated for each of the disciplines in the map of Figure 6 by locating their articles in 554 journal-based disciplines from the UCSD map of science (Börner, Klavans, et al., 2012), and then calculating a Herfindahl (concentration) index on the resulting fractions. A high value indicates a high degree of overlap with a journal-based disciplinary structure, while a low value indicates the opposite. 153 of the 211 disciplines have a concentration value of less than 0.1 (the highest value is 0.56 for Astronomy), which shows that this article-based disciplinary structure is significantly different than a journal-based disciplinary
structure across a wide range of science. Given that this model was built from the ground up based on interactions at the article level, and thus reflects the structure of science at a much more topical level than can be found in a journal-based model, these results once again call into question the validity of journal-based structures of science (Boyack & Klavans, 2011).

Implications and Future Directions
This article reports on a 16-year model of science containing nearly 20 million documents that was created using a combined co-citation and bibliographic coupling process. The bibliographic coupling based process to assign current papers to co-citation clusters is new, and is an advance in process that significantly increases the accuracy of the model. A sequentially hybrid approach to producing useful visual maps from models was introduced, and was shown to likewise increase accuracy. The resulting cluster structure shows areas of high density with pathways between those areas of high density. This structure is amenable to segmentation into discipline-like groupings, and a process was introduced to split the map into disciplines.

This highly detailed, dynamic model and map of science were created primarily to use in answering planning-related questions. Despite that, it can be used for evaluation functions. For example, segmentation of the map into discipline-like groupings enables calculation of field-normalized metrics using an article-level, rather than journal-level, set of standards. Given the recent success of citation metrics based on citing-side normalization (Zitt, 2010) or fractional citation counting (Waltman et al., 2012), it is certainly worth investigating if field normalization based on an article-level classification system would be a suitable alternative to existing normalization methods.

We have a variety of efforts underway that use the map to answer detailed questions raised by research planners. For example, the threads from this map have been combined with a direct citation approach to successfully identify recent emerging topics in science (Small, Boyack, & Klavans, 2013). We are investigating new types of metrics to identify innovative (rather than impactful) articles and researchers, and are working with funders to correlate data associated with grant applications with metrics associated with applicants, referees and review panels. Initial indications of the results of these new investigations are all positive. These investigations all rely upon the ability to model science at the highly detailed level of the model and map reported here, and as such, suggest that this model will prove useful to answering detailed policy-related questions.

We note that these efforts are retrospective rather than predictive. However, they are moving us toward prediction in the sense that we are developing a detailed understanding of how science operates and how those operations are embodied in a model. This understanding is one key to prediction – a better the understanding of fundamental operations in science and technology and how those impact structure will, in time, allow us to project from current models into the future.

Improving science models is a continuous process. The model presented here can be improved in several ways. As it relates to planning applications, this model is not global enough in that it does not include enough technology (patents) or other technical literature that is not indexed by the large citation databases. It includes very little non-English literature. For example, the Chinese scientific and technical databases have millions of additional documents that, if added,
could substantially change our perception of the landscape of science and technology. In addition, these maps do not include funding sources, which often can be thought of as early indicators of where science is moving. Finally, as models increase in size and complexity, these increases must be accompanied by increasing numbers of validation studies.

Acknowledgements
We thank Michael Patek for extracting the necessary data and creating and running the scripts that create our models of science. The sequentially hybrid layout work described here was supported by the Center for Scientific Review (CSR) at the U.S. National Institutes of Health.

References


 Appendix A – Co-citation model protocol

This protocol first appeared in Boyack & Klavans (2010). The data from each source publication year is considered separately so that annual models can be built for each year. To create an annual model from a single year’s data, we use the following method:

- An age-dependent threshold is applied to identify a set of highly cited references. A relatively low threshold is used – references are included if cited five or more times in a single year. Recent references have a lower threshold – those published within 3 years must have at least \( \text{age} + 1 \) cites, while those of age 0 must have been cited twice. This threshold maximizes the number of cited references that are retained in the calculation, subject to the current limits of our clustering approach.

- Co-citation frequencies, \( C_{ij} \), between pairs of reference documents \( i \) and \( j \) were calculated from the citing:cited pairs list.

- Each co-citation frequency was modified using

\[
F_{ij} = \frac{1}{\log(p(C_{ij}+1))} \quad \text{where} \quad p(C_{ij}+1) = \frac{C_{ij}(C_{ij}+1)}{2}.
\]

(1)

- K50 (modified cosine) values were calculated from each \( F_{ij} \) value as:

\[
K50_{i,j} = K50_{j,i} = \max \left( \frac{F_{i,j} - E_{i,j}}{\sqrt{S_i S_j}}, \frac{F_{j,i} - E_{j,i}}{\sqrt{S_i S_j}} \right)
\]

(2)

where

\[
E_{i,j} = \frac{S_i S_j}{SS - S_i}, \quad S_i = \sum_{j=1}^{n} F_{i,j}, \quad j \neq i,
\]

\[
SS = \sum_{i=1}^{n} S_i.
\]

\( E \) is an expected value of \( F \), and varies with \( S_j \); K50 differs from most other measures in that it is a relative measure that subtracts out the expected value. Thus K50 will only be positive for those reference paper interactions that are larger than expected given the matrix row and column sums. Note also that although \( E_{i,j} \neq E_{j,i} \), the differences in these
values are typically too small to be of consequence for paper-level similarities, even though they can be quite large for journal-level similarities.

- Given the current practical limits of the DrL layout algorithm, we filter the relatedness matrix, keeping only the top-n most related references (and their relatedness values) for each reference. We vary the n in top-n from 5 to 15, and scale its value on log(degree), where degree is the number of other references to which the reference is linked through co-citation. The result of this step is a file with pairs of reference papers and their similarity values.

The reference clustering process is comprised of the following steps:

1) The OpenOrd (formerly DrL) graph layout routine (S. Martin et al., 2011) was run using a similarity file as input, and using a cutting parameter of 0.975 (maximum cutting). DrL uses a random walk routine and prunes edges based on degree and edge distance; long edges between nodes of high degree are preferentially cut. A typical DrL run using an input file of 2M articles and 15M edges will cut approximately 60% of the input edges, where an edge represents a single document-document similarity pair from the original similarity file. At the end of the layout calculation, each article has an x,y position, and roughly 40% of the original edges remain.

2) Papers were assigned to clusters using an average-linkage clustering algorithm (Klavans & Boyack, 2006). The average-linkage clustering algorithm uses the article positions (x,y) and remaining edges to assign papers to clusters. Once the clusters are generated, the full list of pairs of papers that co-occur in a cluster are generated for each solution. For example, if papers A, B, C, and D are in a cluster together, the set of pairs will be AB, AC, AD, BC, BC, and CD.

3) Steps (1-2) were run 10 separate times using 10 different random starting seeds for DrL, and thus giving rise to 10 unique cluster solutions for the same similarity file. Different starting seeds (i.e., different starting points for the random walk) will give rise to different graph layouts and different (but typically highly overlapping) sets of remaining edges. We use these differences to our advantage in this clustering process.

4) Those paper pairs that appear in 6 or more out of the 10 DrL solutions are considered to be the robust pairs, and are listed in a separate file. This list of pairs is then used as the input edges to the same average-linkage clustering algorithm used in the previous step. Using this input, the algorithm essentially finds and outputs all distinct graph components. Each separate component is a cluster, and these clusters are referred to as level 0 clusters.

5) Logic dictates that a cluster should have a minimum size; otherwise there is not enough content to differentiate it from other clusters. In our experience, a cluster should contain a minimum of approximately five papers per year (or 25 papers over the five year length of the corpus) to be considered topical. Thus we take all clusters with fewer than 25 papers, and aggregate them. This is done by calculating K50 similarities between all pairs of level 0 clusters, and then aggregating each small cluster (< 25 papers) with the cluster to which it has the largest K50 similarity until no clusters with < 25 papers remain. K50 values are calculated from aggregated modified frequency values (the \(1/\log(p(C+1))\) values) where available, and from the aggregated top-n similarity values in all other cases. The resulting aggregated clusters are known as level 1 clusters.
Annual models are then linked into a longitudinal model of science by linking clusters of documents from adjacent years together using overlaps in the cited references belonging to each cluster. For each pair of years, linking is done using the superset of references from the two years’ models. Typically, $\frac{1}{3}$ of the references are present in both models, $\frac{1}{3}$ are only present in the first year’s model, and the other $\frac{1}{3}$ are only present in the second year’s model. The majority of the references that are only in one model are missing from the other only because they did not meet the citation threshold, and can be easily added to the other model using their reference lists. This process generates augmented reference lists for each model. For a given model, the augmented reference list used for linking to the prior year’s model will be somewhat different than the augmented reference list used for linking to subsequent year’s model.

Using these augmented reference lists, clusters from adjacent models are linked if a simple cosine index based on the number of overlapping references is above a threshold. Although it would be nice to set this cosine threshold based on theory, in practice we find it requires a heuristic approach. If the cosine threshold is set too low, a giant component quickly emerges and the longitudinally-linked sets of clusters become so large as to no longer represent research problems. If the cosine threshold is set too high, there is very little linking between clusters. We also found that, given that science is growing and that the number of clusters increases each year, using a single cosine threshold for all pairs of years created less linking for later years (late 2000s) than for earlier years (late 1990s). This was an undesirable effect. To create consistency in the linking patterns over time, we ran linking calculations at a variety of linkage fractions, where linkage fraction is defined as the fraction of clusters in a given year that link to a cluster in the subsequent year. This required calculation of the cosine threshold for each year that would return the desired linkage fraction. For each set of calculations we examined the average and maximum numbers of forward links per cluster (of those that have forward links), along with the maximum thread size. We chose a linkage fraction of 0.48, meaning that only 48% of the clusters link to a cluster in the subsequent year. This gave us average and maximum numbers of forward links per cluster of 1.1 and 5, respectively. The corresponding forward linking cosine threshold values for each year are listed in Table 3, and range from 0.260 to 0.223, typically decreasing over time.

When linking clusters this way over many years, one additional problem can arise – two long threads can be linked together in a later year if the cosine value is high enough, creating artificially large threads. For example, using the method and thresholds detailed above, the largest thread had 872 clusters, or 54 clusters per year. Although this type of linkage can reflect the history of how topics link together from a retrospective point of view, it does not necessarily reflect how each thread grows. We thus implemented an additional step in our threading calculation. For cases where a single cluster merges two threads, where one thread is at least 8 years old, and the other is at least 5 years old, the cluster is assigned to the shorter thread and not to the longer one, despite the fact that the cosine threshold is met in both cases. This criterion still allows very long threads to form, but it does not allow retrospectively joining of long existing threads. The effect of this criterion on size was to reduce the largest thread to 66 clusters (or 4 per year). This is a significant improvement in our view; the majority of the threads remain thin – they are not dominated by branching – and thus represent coherent research problems as they move through time.
Appendix B – Bibliographic coupling assignment process

The detailed process for assigning current papers to co-citation clusters is as follows:

1) A bibliographic coupling solution for the current papers was calculated using the methodology from Boyack & Klavans (2010). At this point there are two solutions – the original co-citation solution which fractionally assigns papers to clusters (PID CC wt), and the bibliographic coupling solution which singly assigns papers to cluster (PID BC).

2) The current papers (PID) are divided into 3 groups:
   a. Group A – those that are in both solutions
   b. Group B – those that are only in the BC solution
   c. Group C – those that are only in the CC solution

3) For all papers in group A, a figure of merit (FOM) based on a combination of the co-citation (CC) and bibliographic coupling (BC) clusters was calculated:
   a. The BC cluster was assigned to each PID in the CC solution to create a table with entries (PID CC BC wt)
   b. Weights were summed up by CC:BC pair (CC BC sumccbc)
   c. Weights were summed by BC cluster (BC sumbc) and added to the table in 3b (CC BC sumccbc sumbc)
   d. Divide sumccbc by sumbc to get FOM for each CC:BC pair (CC BC FOM)
   e. This figure of merit replaces the original weights for each PID within a CC:BC pair (PID CC BC FOM)
   f. FOM are summed and normalized so that they sum to 1.0 for each PID (PID CC FOMnorm). These FOMnorm become the new weights for PID in the co-citation clusters, replacing the old weights.

4) Bibliographic coupling clusters some papers that are not clustered by the co-citation solution. These papers (group B) were assigned to co-citation clusters by:
   a. Creating a CC:BC cosine relatedness matrix from the FOM in step 3d where the matrix values are cos = FOM/sqrt(rowsum*columns)
   b. Linking PID to CC using their BC clusters (PID CC BC cos)
   c. Singly assigning the PID to the CC cluster with the highest cosine value (PID CC 1.0)

5) Co-citation clusters some papers that are not clustered by the bibliographic coupling solution (group C). The simple fractional assignments from co-citation analysis are used for these papers.