

Evolving and Emerging Populations and Topics

White Paper for CISE/SBE Advisory Committee on Research Portfolio Analysis

Katy Börner and Angela Zoss

Cyberinfrastructure for Network Science Center

School of Library and Information Science, Indiana University, Bloomington, IN

{amzoss, katy}@indiana.edu

2010.10.06

1. Executive Summary

The CISE/SBE Advisory Committee is charged with “*identifying and demonstrating techniques and tools that characterize a specific set of proposal and award portfolios. It is asked to identify tools and approaches that are most effective in deriving knowledge from the data provided, i.e., most robust in terms of permitting program officers to visualize, interact, and understand the knowledge derived from the data.*”

This work starts with the questions: What computational tools do NSF program officers want, need, and desire? Interviews with 38 science policy makers, conducted in 2008, revealed that most of them want (1) easy access to high quality data from different sources (expert opinions, papers, patents, awards, etc.), (2) analysis approaches and tools that are vetted and peer-reviewed by the scientific community, and (3) transparent (open data, code, tools) and replicable results. The interviews also resulted in many concrete insight needs which cannot be addressed by today’s tools because the tools either do not read the format in which the data exists, are too difficult to use, or simply do not fit the concrete needs and the daily workflow of NSF and other program officers.

SciSIP funding and funding for the CISE/SBE Advisory Committee supported the development and application of the Science of Science (Sci²) Tool that runs on all major software platforms—Windows, Linux, Mac. (Source code and tutorials are online at <http://sci.slis.indiana.edu/sci2>.) Sci² Tool was applied to NSF awards and proposals data with the special focus of identifying evolving and emerging populations and topics. First results comprise:

- A theoretically grounded and a practically useful methodology to analyze and group NSF proposals by the scientific domains they are citing (drawing from). Specifically, the RefMapper as part of the Sci² Tool was applied to identify the journals cited by proposals and to communicate the results via a science map data overlay for fast visual examination and review of hundreds of proposals. The RefMapper can also be applied in order to: explore the interdisciplinarity of proposal sets as an indicator of emerging research areas; understand how various science fields are interlinked; examine the range of topics in a given award portfolio (IIS, SciSIP); assess the amount of interdisciplinary research; identify areas that are funded by multiple directorates or divisions; and identify emerging areas for future funding solicitations.
- Temporal analysis techniques that can be used to identify emerging trends based on topic bursts, i.e., sudden increases in the usage frequency of terms or phrases. NSF program officers might like to be aware of these bursts of activity and might like to ‘seed’ them via workshop money or sugar grants or consult key experts in these emerging areas when compiling new solicitations.
- Evolving collaboration structures at different levels of analysis. Science is driven by scholars that have evolving network associations with other experts at the same or different institutions, in the same or different topic areas, and via co-author, co-PI and other relationships. Existing (successful) collaboration structures can be indicative of the future success of investigators and their proposals.

Details are provided below.

2. Nontechnical Description of Techniques and Tools

Using the open source Science of Science (Sci²) Tool and NSF data, the following four analyses were performed, see [Borner-Zoss-Slides-June7.2010.pdf](#) for concrete results. Please see 12 NIH tutorials at <http://sci.slis.indiana.edu/sci2> for other analyses (e.g., geospatial maps) that might be relevant for NSF program officers.

2.1 Measuring the Interdisciplinarity of NSF Proposals

Some NSF solicitations, e.g., *Cyber Enabled Discovery and Innovation (CDI)*, invite proposals that combine works from many different fields of science. Thousands of proposals might be submitted, and they need to be grouped for review. Proposals that cover many different topics are said to be more interdisciplinary than

those that cover only a few. Interdisciplinarity can be measured using text analysis or linguistic techniques on title, abstract, or full proposal texts. However, writing styles and word usage differ widely across disciplinary boundaries, making it difficult to identify exactly what areas of knowledge a proposal draws from or contributes to.

The RefMapper tool was developed to support the automatic detection, mapping, and clustering of grant awards and proposals based on citation references. It might be used to group proposals for review or to communicate the topic coverage of a proposal/funding portfolio. The tool uses a master list of 18,351 journal names that are indexed by Scopus and Thomson Reuters (ISI SCI, SSCI, and A&H Indexes) and a lookup table of 57,860 different abbreviations for those journal names. Given a set of journal names extracted from proposal references, it science-locates identified journals on the 554 scientific areas of the UCSD Map of Science (Klavans and Boyack 2007). Each of the 13 main scientific disciplines is labeled and color coded in a metaphorical way; e.g., Medicine is blood red and Earth Sciences are brown as soil. Journal names and the number of references published in each are listed together, along with all unidentified references (PhD theses, books, etc.). The RefMapper also identifies clusters based on reference co-occurrence similarity. Analysis results are rendered into a pdf file (see Fig. 1 for an example). The file contains an (a) Overview of the number and topical coverage map of all PDF files on the first page, (b) Visual Index that shows a thumbnail map for each PDF file to support visual comparisons, (c) Detail page with map, journal listing, and unidentified references for each PDF file, and (d) Top-10 Most Similar page for each PDF file to help identify clusters and show disciplinary focus/interdisciplinary spread. The file can be printed, thumbed through, and annotated.

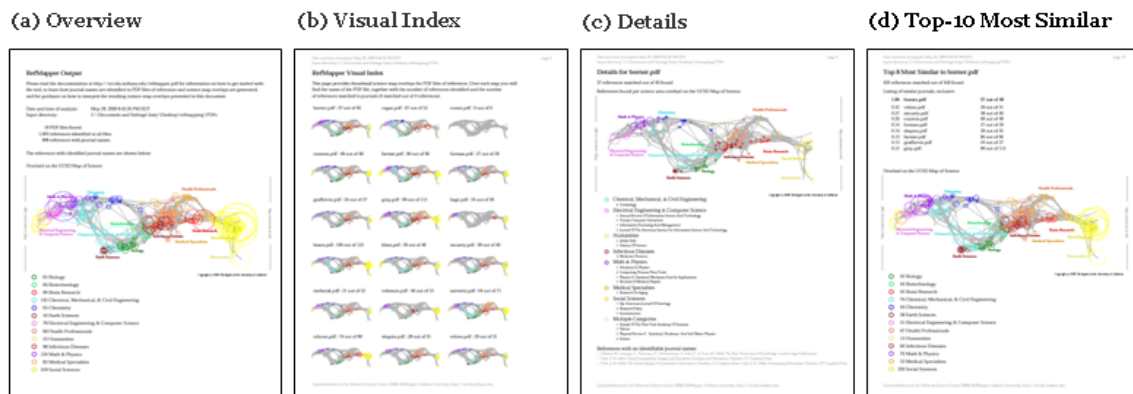


Figure 1: RefMapper presentation of analysis results as static paper printouts providing (a) an overview of all proposals, (b) a visual index of each single proposal, (c) detailed map and listing of matched and unmatched references for each proposal, and (d) the most similar proposals for each proposal.

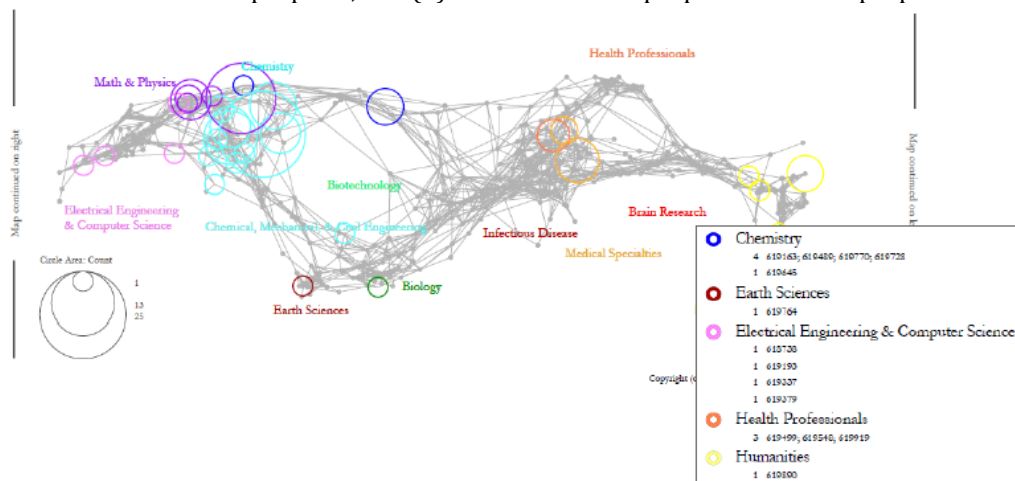


Figure 2: UCSD Map of Science overlay of 118 MRI awards (see section 3.1 for a description of the data). See [Borner-Zoss-MRI-Scimap.pdf](#) and [Borner-Zoss-MRI-Scimap.pdf](#) for high resolution images and details.

Note that the topic coverage and interdisciplinarity of proposals could also be measured using additional data such as

- Keywords provided by investigators, e.g., for CDI proposals.
- Background/departments/publications/prior funding of PI/Co-PIs—requires unique people IDs and resume-like information, e.g., using data from institutional databases of record via VIVO (<http://vivoweb.org>).
- Publications and other results reported in NSF progress reports.
- Topics extracted from titles and abstracts; see Figures 2 and 3.

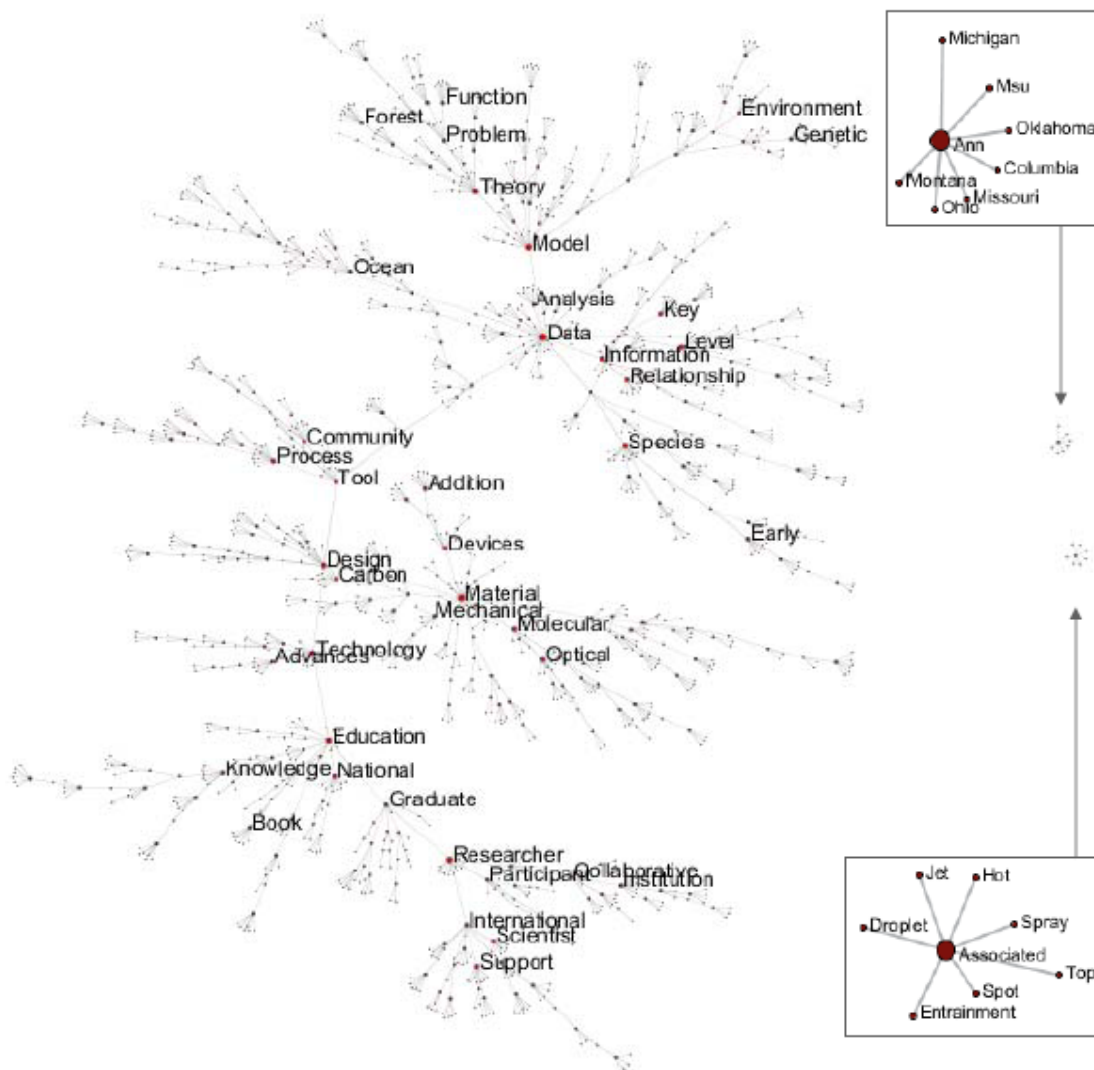


Figure 3: Topics for 50,608 of the 51,322 NSF awards from the Sharepoint site (see section 3.1) were extracted by Dave Newman and are shown here. There are 1,592 topic nodes and 10,389 co-occurrence edges in 3 components. MST-Pathfinder Network Scaling was applied to reduce the number of edges to 8,560. Nodes with degree > 7 are red and labeled. See [Borner-Zoss-Slides-June7.2010.pdf](#) for high resolution images and details.

2.2 Identification of Emerging Trends via Burst Analysis

This part of the project aims to answer questions such as

- What are the main research topics in proposals submitted, and what are the main research topics in awards made by different NSF programs?
- How do these topics evolve over time?

- What bursts of activity exist?

It applies Kleinberg’s burst detection algorithm (Kleinberg 2002) to identify words in titles, keywords, abstracts, summary, or full text (or topics extracted from those texts) that experience a sudden increase in usage frequency.

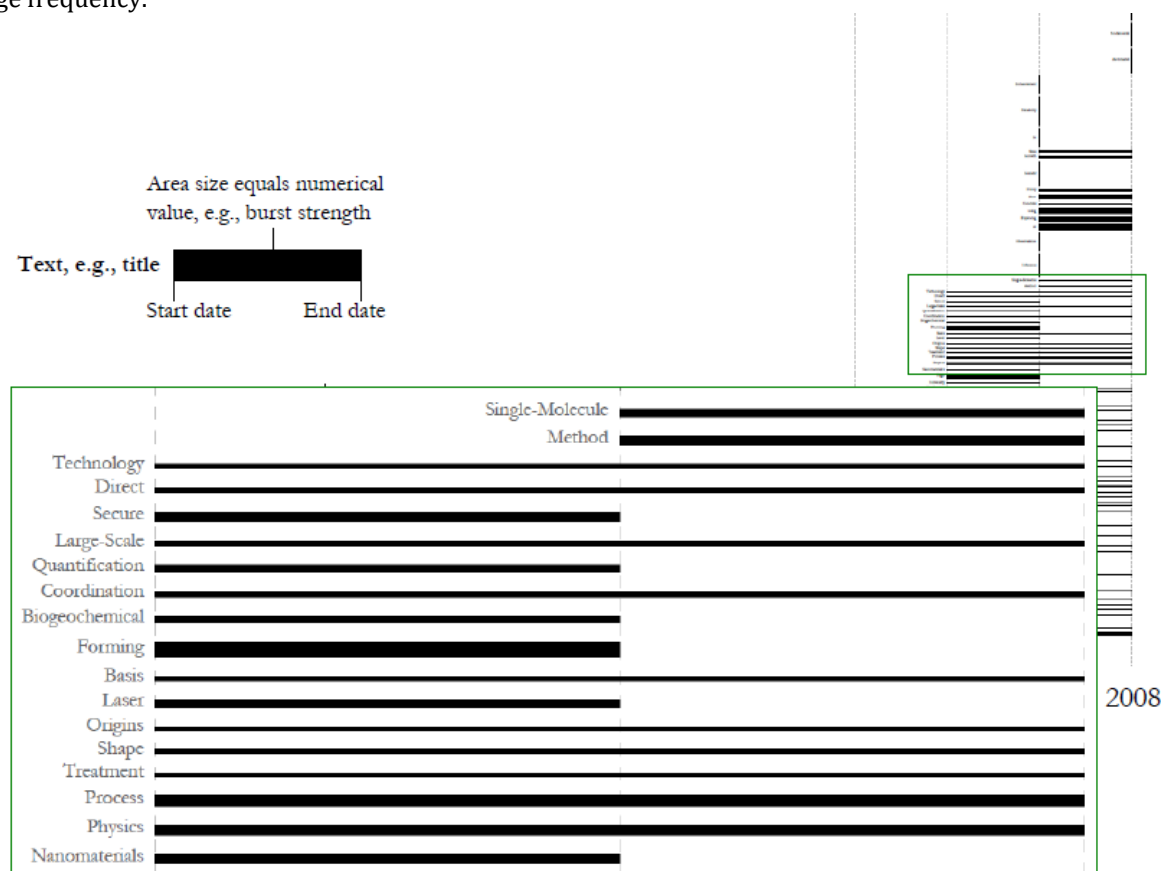


Figure 4: Temporal analysis of bursty terms in titles of Career awards (see section 3.1 for a description of the data). See [Borner-Zoss-Slides-June7.2010.pdf](#) for high resolution images and details. Burst results are rendered as horizontal bar graphs that encode burst strength by area size. We provide a zoom into bursts during years 2006-2008.

2.3 Evolving Collaboration Networks

These analyses aim to answer questions such as

- What funding programs fund what institutions?
- How do PIs from funded institutions collaborate?
- What is the “backbone” of the collaboration network?
- What clusters exist and what topics are they specializing in?

The structure and dynamics (e.g., temporal evolution) of these networks can be analyzed and visualized using more than 60 different network extraction, preprocessing, analysis, and visualization algorithms available in the Sci² Tool.

Using additional data such as publication records for all investigators, the co-authorship network of the co-PIs can be extracted—prior co-authorship might be an indicator for future joint papers. If citation data is available, then highly cited (joint) papers by the investigators might be an indicator future success. Furthermore, one could identify what communities work on similar topics but do not cite each other or collaborate much (see work by Noshir Contractor).

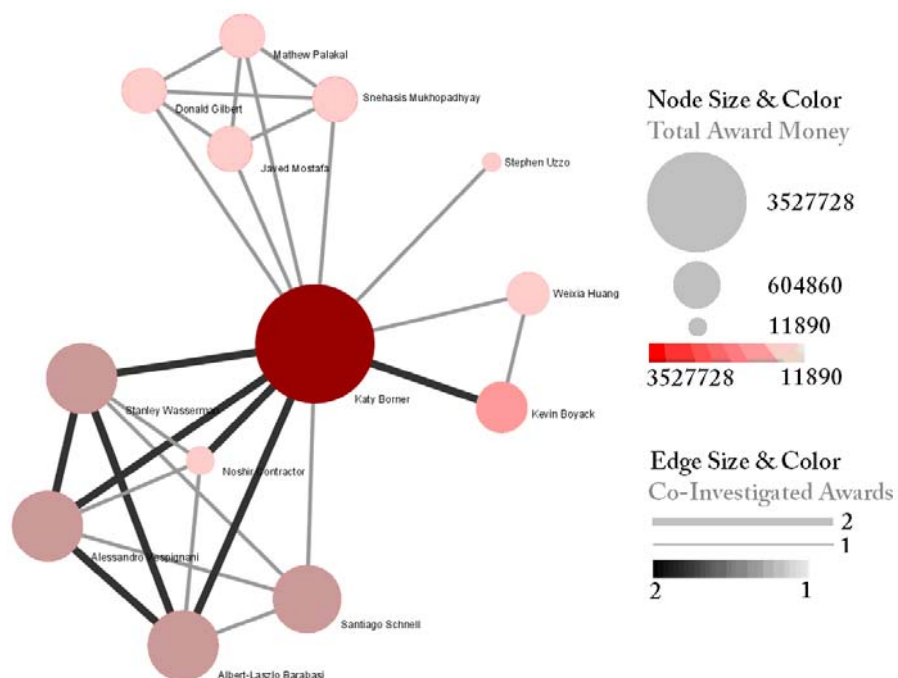


Figure 5: NSF Co-investigator network of one ego, here Katy Börner. Data was retrieved using a query for all funding awarded to the ego. The funding records for all other people in the network are most likely incomplete. See Mapping Collaboration, Publication and Funding Profiles of One Researcher, section 5.1.1 in Sci² Tool Tutorial.

3. Technical Description of Techniques and Tools

The Sci² Tool comes with 110+ pages of detailed documentation and 24 hours of tutorials designed to introduce the tool to the Reporting Branch at NIH, all available online at <http://sci.slis.indiana.edu/sci2>. The utility and robustness of existing tools and approaches seems to be considerably reduced by the quality and coverage of the available data, and the remainder of this white paper will focus on this aspect.

Two data sets were made available to the *CISE/SBE Advisory Committee*:

- 51,322 NSF awards for 6 years (only 4 years are complete) via Sharepoint.
- 151,554 NSF proposals for approximately 3.5 years via a Data Enclave.

Both datasets are discussed below.

Please note that I deeply appreciate all the effort that went into providing access to this data. Additionally, I see the application of different tools and approaches to the same datasets as one of the best ways to compare approaches and to identify those that work or don't work. However, I am asked to provide input in how decision making by NSF program officers can be improved via robust data mining and visualization techniques, and one of the key obstacles appears to be access to high quality and high coverage data.

3.1 NSF Awards Made Available via Sharepoint

All NSF awards that have been active at any time between Oct 2005 and Jan 2010 were provided by Paul Markovitz. The query was not limited by scientific term, program or program officer. The data was retrieved from the Research Spending and Results (RS&R) service on Research.gov:

http://www.research.gov/rgov/anonymouse.portal?nfpb=true&pageLabel=page_research_funding_search&nfls=false and provided in 3 spreadsheets named

- NSF awards Oct 2005_to_June 2007 as of Feb03.xlsx (16,762 records)
- NSF awards July 2007_to_Sept 2008 as of Feb03.xlsx (18,000 records)
- NSF awards Oct 2008_to_Jan 2010 as of Feb03.xlsx (16,561 records)

The annual distribution is shown in Table 1. The total count of all grants is 51,322. Note that complete data is only available for four years.

Table 1. Number of NSF grants per calendar year.

Year	Count	First Awarded Date	Last Awarded Date
2005	831	10/5/2005	12/30/2005
2006	10942	1/3/2006	12/29/2006
2007	12031	1/3/2007	12/31/2007
2008	11622	1/2/2008	12/31/2008
2009	15312	1/2/2009	12/31/2009
2010	584	1/4/2010	2/2/2010
Total Count	51322		

As no NSF program officer has detailed knowledge on all awarded grants and could help validate an analysis of all grants, we extracted program specific subsets:

- IGERT awards (87 records)
- Career awards (2409 records)
- MRI awards (1114 records)

Next, we compared the resulting three datasets with queries run on NSF's award search (<http://www.nsf.gov/awardsearch>) on 2010.05.13 and 14:

- IGERT: NSFawardsearch retrieves **235** for "IGERT" in title with unchecked Historical Awards, Active Awards Only, Expired Awards Only. Excluded one research award on the impact of IGERTS. **114** awards start before October 2005, **121** after.
- Career: NSFawardsearch retrieves **more than 3000** hits for active awards that have "career:" in the title. In *Feb03.xlsx there are **30** awards that started before Oct 2005. In the NSFawardsearch result there are **512** record that start in 2008 while the *Feb03.xlsx files show 530 records.
- MRI: NSFawardsearch retrieves **1746** for "MRI" in title with unchecked Historical Awards, Active Awards Only, Expired Awards Only.

The counts are radically different from counts obtained from the Sharepoint data. These differences need to be resolved before any analysis results are used for real-world decision making.

3.2 NSF Proposals Available via Data Enclave

All 151,554 proposals submitted between November, 2006, and April, 2010, were made available via a Data Enclave. For each proposal, there exist four PDF files: Cover Sheet, Project Summary, Project Description, and References. The cover sheet data were used to define groups of proposals (e.g., those submitted to a particular NSF Organization) and to explore the profiles of individuals and institutions constructing the proposals. The reference data were used as a proxy metric of interdisciplinarity for the groups of proposals.

Sorting of the files was facilitated by simple operations executed via the Windows Command Line utility. To analyze the cover sheets, we applied the parser developed by Lee Giles' team at Penn State to extract and save fielded data as XML. We wrote a Java program that reads these XML files and compiles a comma-separated listing of all fields that can be saved as single CSV file (with one row per proposal). Once the proposals are listed in a tabular data format, it was trivial to iterate through and add a flag that indicates that the proposal ID also appears in the list of awarded proposals that was provided. Additional operations such as cleaning mistakes from the parsing were also applied to records within Excel.

At this point, the cover sheet data can either be mined for lists of proposal IDs that represent interesting groups or they can be imported directly into Sci² for additional analyses and visualizations. Initially, the sample lists of IDs we used enabled us to separate PDFs of References of groups into individual folders for each group. The RefMapper plugin for the Sci² tool then pulls text out of the PDFs from a specified folder, searches for journal names within the references, and matches the proposal references to the UCSD map of science. The plugin outputs not only a visualization of the map of science activations but also a list of the journals found, a list of the references that were not matched, and a list of the number of reference matches for each of the 13 major scientific domains (see section 2.1). The table of reference matches can also be visualized with bar charts for a series of smaller comparisons.

Direct analysis of the parsed cover sheet PDFs took the form of textual and network analyses. Sci² can take any tabular data with a column of text and a date column and analyze how words change in frequency (burst) over time. Likewise, Sci² can generate networks of individuals where links occur between people who have worked together as PIs. For the purposes of anonymization, co-PI relationships were abstracted up to

the institution level, yielding a network of institutions where links occur between the institutions of PIs who have worked together. (Institutional affiliation information is currently only available for the primary PIs. To determine institutional affiliations for as many co-PIs as possible, co-PI names were cross-referenced with primary PI names and were retained if institutional affiliation could be determined. This approach considerably reduces the number of investigators captured in the visualization.)

We are currently working on using the topical descriptions generated by Dave Newman and a recently developed reference parser from the Penn State team. The topical descriptions could be incorporated as an additional property of the proposals and either added to network analyses or compared to the map of science matches. A more accurate reference parser might improve matches between the Reference PDFs and the map of science.

4. Discussion of Scalability and Maturity

The analyses proposed in section 2.1-2.3 use advanced data analysis and visualization algorithms and have been shown to be highly scalable to tens of thousands of records. A regular update of the UCSD map of science is desirable to ensure that new journals, e.g., PLoS, are captured and respective references can be science located. The loading of the 13 different disciplines and 554 scientific fields can also be communicated via horizontal bar charts together with percentages and absolute numbers. If two science maps are compared, it is desirable to present both on one map to make differences easy to spot.

Approaches that combine textual analysis and linkage analyses should be explored further as first results are promising.

The utility of existing tools is reduced, however, by the quality and coverage of data and restrictions imposed by the Data Enclave. The former were discussed above. The latter relate to the requirements that for the

- analysis of interdisciplinarity introduced in section 2.1., only scientific fields that are cited by at least ten PIs can be revealed.
- temporal analysis discussed in section 2.2, only those topic bursts which occur in at least ten proposals can be shown.
- collaboration network analysis in section 2.3, no investigator names can be shown – only institutions that have at least ten lead PIs (institution information is only known for the lead PI).

These restrictions considerably limit the data for the current study and hence the utility of the resulting analyses and visualizations.

5. Recommendations for NSF

Table 2 shows improvements to data quality and coverage that can be made now and in the future to improve decision-making at the award portfolio (less than 100 awards/proposals) and the multiple award portfolios (more than 100 awards/proposals) levels. The distinction by level was made based on feedback by program officers during the virtual presentations on September 22 and 24. There appears to be a major need by program officers to take all proposals in a portfolio; group them; assign keywords; retrieve those that match particular search queries, topic terms, or years; keep a history of prior searches to run the same search on a new proposal/reviewer set; access details on a particular proposal or PI; split sets of proposals among two POs; save a subset into an Excel readable file; and communicate the coverage of a portfolio to others. All this functionality is provided by MS Sharepoint and similar content management tools. A user simply uploads a CSV file with columns such as start/end date, title, keywords, PIs, institutions, etc. and the software supports the generation of faceted search and browsing, custom views of the data, and the continuous update of data and views.

An analysis at a larger scope—multiple award portfolios—might be needed when comparing, merging, splitting, or creating programs. Here, an overview of the total topic coverage, existing general (topic) trends and bursts of activity, as well as existing collaboration networks are beneficial, and tools such as those developed by the *CISE/SBE Advisory Committee* members become relevant.

Table 2. Recommendations to improve data quality and coverage now and in the future for proposal.

	Now	Future
Data Quality	<ul style="list-style-type: none"> • Examine data issues discussed in section 3.1. Why do the counts not match? • Extract cover sheet information (using Giles' parser) and generate topics (using Newman's approach) for all proposals/awards and use them in your daily work. Provide detailed feedback to programmers. 	<ul style="list-style-type: none"> • Store metadata, e.g., from cover sheet, in fielded format for further analysis. • Use unique identifiers for investigators and automatically pull demographic info, e.g., from CVs or VIVO. See ORCID effort. • Ask PIs to upload references in common bibliographic tool formats like BibTeX or Endnote so that they can be parsed. • Link awards to data from progress reports so that the impact of a project can be analyzed algorithmically.
Award Portfolio Level Most relevant for POs	<ul style="list-style-type: none"> • Consider using Confluence, MS Sharepoint, or JIRA to keep track, manage, and view award portfolios. Can be implemented in one day if software is in place. 	<ul style="list-style-type: none"> • Extend Confluence, MS Sharepoint, or JIRA by adding data analysis and visualization functionality, e.g., to identify high degree or high centrality PIs, to identify bursts, draw networks, etc.
Multiple Award Portfolios	<ul style="list-style-type: none"> • Explore using existing tools in your daily work. Provide detailed feedback to programmers. 	<ul style="list-style-type: none"> • Extend system to support both levels • If a base map of science is used, update it each five years.

References

Klavans, R. and K. W. Boyack (2007). [Is There a Convergent Structure to Science?](#) Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics, Madrid, CSIC.

Kleinberg, J. M. (2002). [Bursty and Hierarchical Structure in Streams.](#) 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press.