Impact of a Cyberinfrastructure Enterprise on the Nation's Workforce: Visualizations of a Decade of NCSA's Diaspora

James Howison
University of Texas at Austin

Nicholas Berente
University of Georgia

James Cheng
University of Texas at Austin

Amanda Sutton
University of Texas at Austin

Heath Naquin
University of Texas at Austin

Overview

Finding and holding onto good people is a challenge in cyberinfrastructure ("CI") enterprises. Leading-edge technological skills that are necessary for CI are also in great demand in virtually every industry. Research-oriented CI enterprises that lose personnel to industry often see this as a failure of the system, or of leadership, to hold onto valuable employees. So much time and resources were spent on these personnel to build high levels of expertise in very important technological domains. As those people are hired away, the resulting departures are often considered "lost to science," and must be replaced, often with difficulty (and much expense to taxpayers who have paid for the expert's training).

However, there are benefits to CI enterprises - and to their regions and nation - from the departure of key personnel. This departure can be characterized as a form of "diaspora"[1] – one that has a beneficial side that is rarely articulated – much like emigrant diaspora from developing countries. In particular this diaspora enhances the cutting-edge, highly technical knowledge of the workforce, strengthens capabilities of the leading organizations that hire these experts, contributes to the intellectual property of those organizations, and helps to disseminate technological innovations developed and improved at the CI enterprise.

Conceived in terms of diaspora, the impact of CI enterprises can be thought of in terms of workforce development, strengthening leading-edge industrial knowledge, and building national capabilities. To investigate how we might think through the diaspora of CI enterprises, we conducted a pilot study of one major CI enterprise: the National Center for Supercomputing Applications (NCSA),

[1] Howison, J., Berente, N., King, J.L., 2013. From Loss to Gain: Exploiting Diaspora in Cyberinfrastructure Enterprises. Presented at the Atlanta Conference on Science and Innovation Policy, Atlanta, GA

housed at the University of Illinois at Urbana-Champaign. We compiled and analyzed data from 425 key employees who departed NCSA in the decade from 2003-2013. Often the best way to make sense of complex data involves visualization, and the purpose of this effort was to experiment with visualizations of this diaspora data in an effort to better understand the impact of NCSA's diaspora.

Some key observations from this study include:

- NCSA has an impact on the national workforce, strengthening the workforce of nearly every state.[2]
- NCSA has the greatest impact on the Illinois workforce - both in terms of employees that leave NCSA (the biggest destination is Illinois) and also in terms of attracting key people from other states (attracting personnel from California, for example).
- NCSA contributes to building the capabilities of some of the nation's most important technological organizations: from digital technology firms such as Google and Microsoft to industrial technology firms such as Caterpillar and Northrup Grumman.[3]
- NCSA contributes to the workforce and capabilities of a very large number of small and medium sized organizations (SMEs).[4]
- NCSA impacts a wide array of technologically-intensive industries, and feeds many people back to academia (thus ostensibly contributing to the development of a future workforce).[5]
- NCSA's diaspora is responsible for intellectual property for a number of key U.S. firms, including Apple, Google, Facebook, etc.

Next we present the resulting visualizations in three sections: (1) geographic diaspora; (2) industrial diaspora; and (3) knowledge contributions. The method we followed and challenges faced during the project are detailed in the Appendix.

[2] Interactive visualization of geographic diaspora: https://www.ischool.utexas.edu/~jcheng/mapLinesFromIllinois.html and chord diagrams: https://www.ischool.utexas.edu/~jcheng/blargh.html

[3] Interactive visualization of firm diaspora: https://www.ischool.utexas.edu/~jcheng/firmsRevised.html

[4] Interactive visualization of firm diaspora with SMEs: https://www.ischool.utexas.edu/~jcheng/firms.html

[5] Interactive visualization of industry diaspora https://www.ischool.utexas.edu/~jcheng/industryCondensed.html

Visualizations of Geographic Diaspora

The geographic visualizations were animated to show workers leaving and entering Illinois. The number of workers going to each state is represented by the width of the arrow. Fig. 1 shows the finished geographic visualization.

For an interactive visualization of geographic diaspora online, see:
https://www.ischool.utexas.edu/~jcheng/mapLinesFromIllinois.html

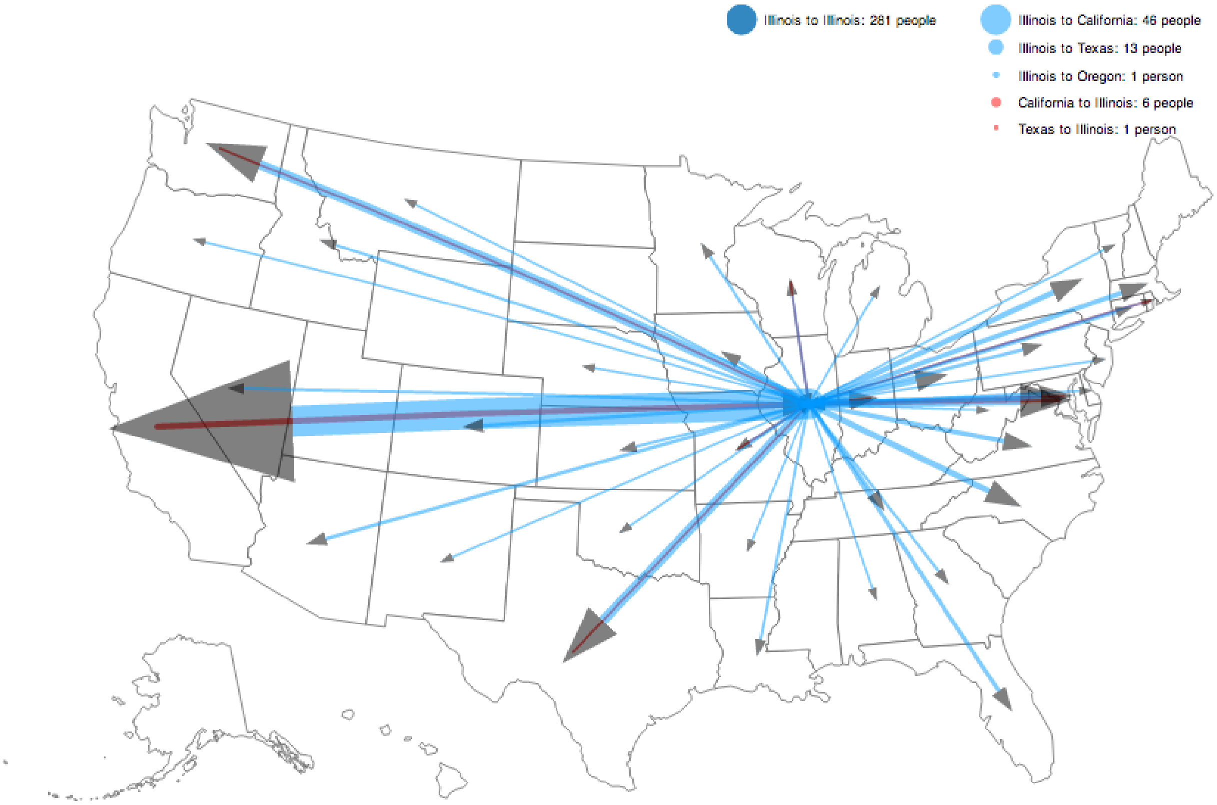

Illinois to Illinois: 281 people
Illinois to California: 46 people
Illinois to Texas: 13 people
Illinois to Oregon: 1 person
California to Illinois: 6 people
Texas to Illinois: 1 person

Fig 1. Visualization of former NCSA employees entering and leaving Illinois.

To further represent the geographic movements of the subjects, we also created Chord Diagrams (Fig. 2). This visualization was inspired by the work of Nikola Sander and associates[6] and is intended to show movement of subjects over time. In the diagram, the darker outside segments are arcs, which represent origins and destinations. Chords are the bands that connect one arc to another, which represent the flow of people from one area to another. The user can hover their mouse over either the arcs or chords to see more information.

For an interactive visualization of chord diagrams of geographic diaspora online, see:
https://www.ischool.utexas.edu/~jcheng/blargh.html

[6] Sander, N., Abel, G.J., Bauer, R. and Schmidt, J., (2014) "Visualising Migration Flow Data With Circular Plots," Vienna Institute of Demography Working Papers, 2014.

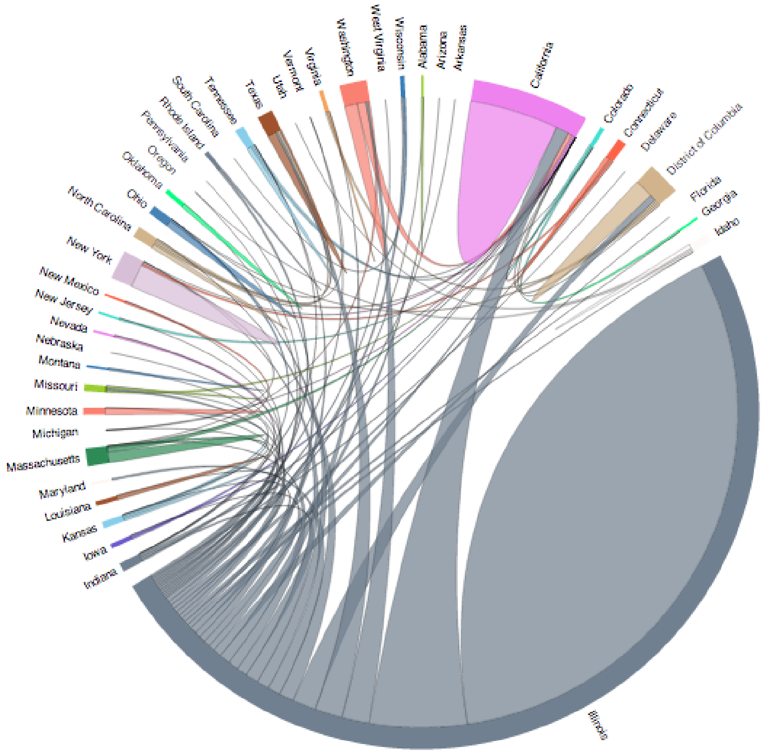
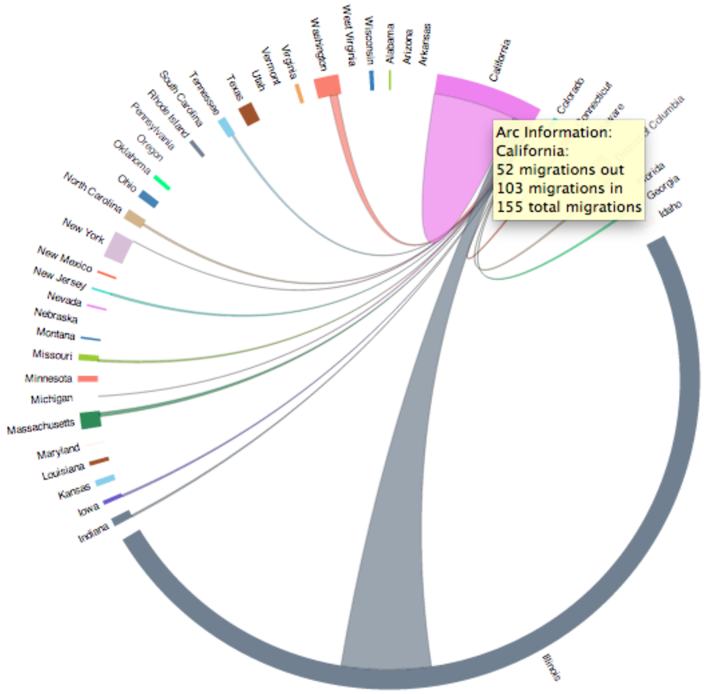


Fig. 2: This chord diagram represents subject geographic migration. The diagram can show all of the migrations, as shown in this figure.



Arc Information:
California:
52 migrations out
103 migrations in
155 total migrations

Fig. 3: When a user hovers his or her mouse over one of the relations within the chord diagram, more information about that particular migration pops up.

Visualizations of Firm Diaspora

Bubble plots (or bubble charts) are the ideal visualization for quickly indicating relative scale through the area of round “bubbles” (the human mind naturally interprets relative scale of circles in terms of their area). We created a bubble chart with only the top employers - firms that hired at least three NCSA alumni - in Fig. 4. (Raytheon, Qualcomm, Motorola, and Intel each hired 3; Caterpillar hired 22.)

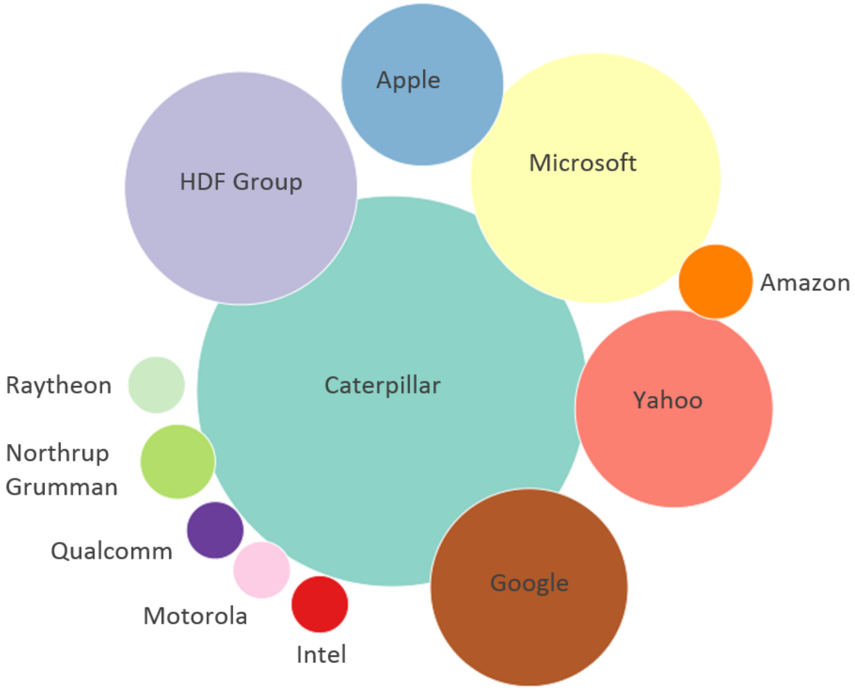


Fig 4: Bubble plot representing the main employer destinations of NCSA diaspora. Range is 3-22; the user can find the company and the number of NCSA employees they hired by hovering over the bubble.

For interactive visualization of top employers, see:
https://www.ischool.utexas.edu/~jcheng/firmsRevised.html

Fig. 5 below is a bubble plot of *all* firms across a variety of categories and names of the very biggest employers. Many went to small and medium sized firms (n=314) and Fortune 100 firms (n=40 in addition to the firms listed).

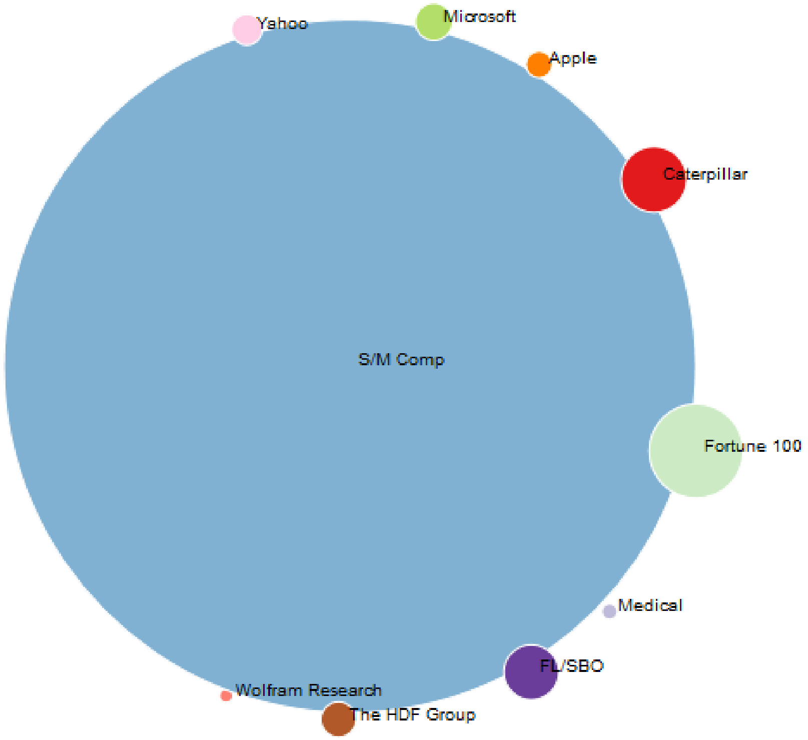


Fig 5: Bubble plot representing the main employer destinations of NCSA diaspora. Range is 3-22; the user can find the company and the number of NCSA employees they hired by hovering over the bubble.

For an interactive visualization of this view of firm diaspora online, see: https://www.ischool.utexas.edu/~jcheng/firms.html

Fig. 6 indicates diaspora by type of firm (through our inductive categorization of industries. This visualization clearly shows that NCSA impacts the workforce across a wide variety of industries. The biggest category is academia, however, indicating (from a workforce development perspective) that the knowledge will further impact the workforce through continued skill development in academia.

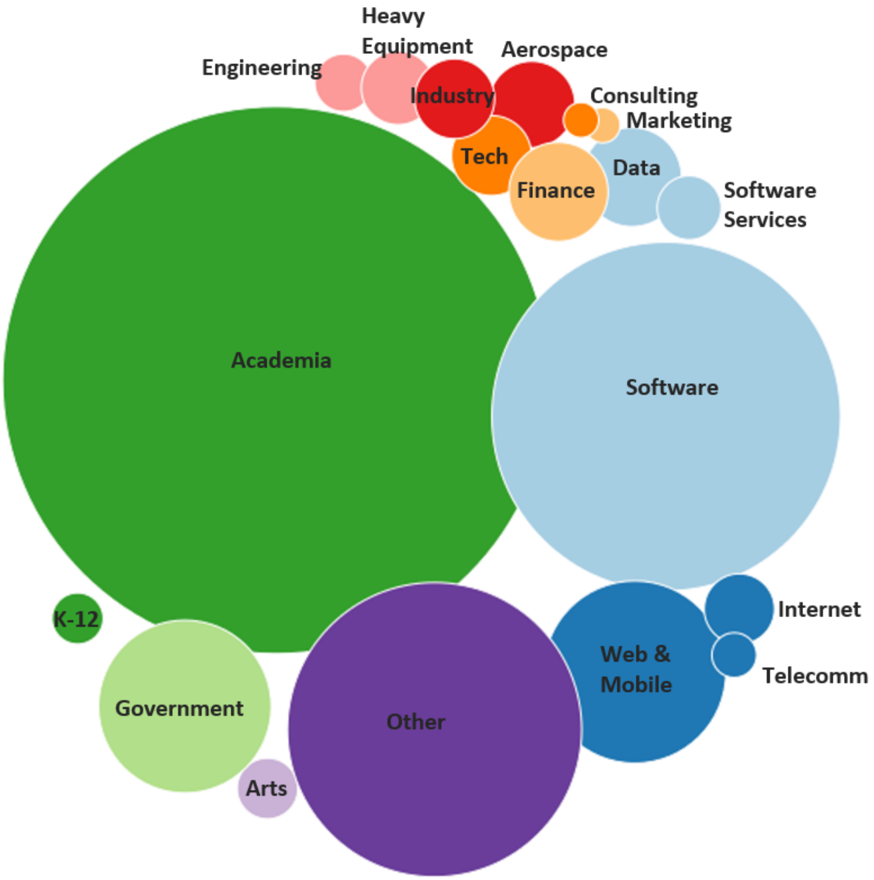


Fig. 6: Industries of the NCSA diaspora. Range 9-169 (Marketing n=9; Consulting n=9; Academia n=169)

For an interactive visualization of industry diaspora see:
https://www.ischool.utexas.edu/~jcheng/industryCondensed.html

For a more fine-grained visualization of industry diaspora, see:
https://www.ischool.utexas.edu/~jcheng/industry.html

Visualizations of Diaspora Intellectual Output

Lastly, using Microsoft Excel, we visualized the subjects' intellectual output. Fig. 7 describes the number of patents and their assignment (Source: Google patents).

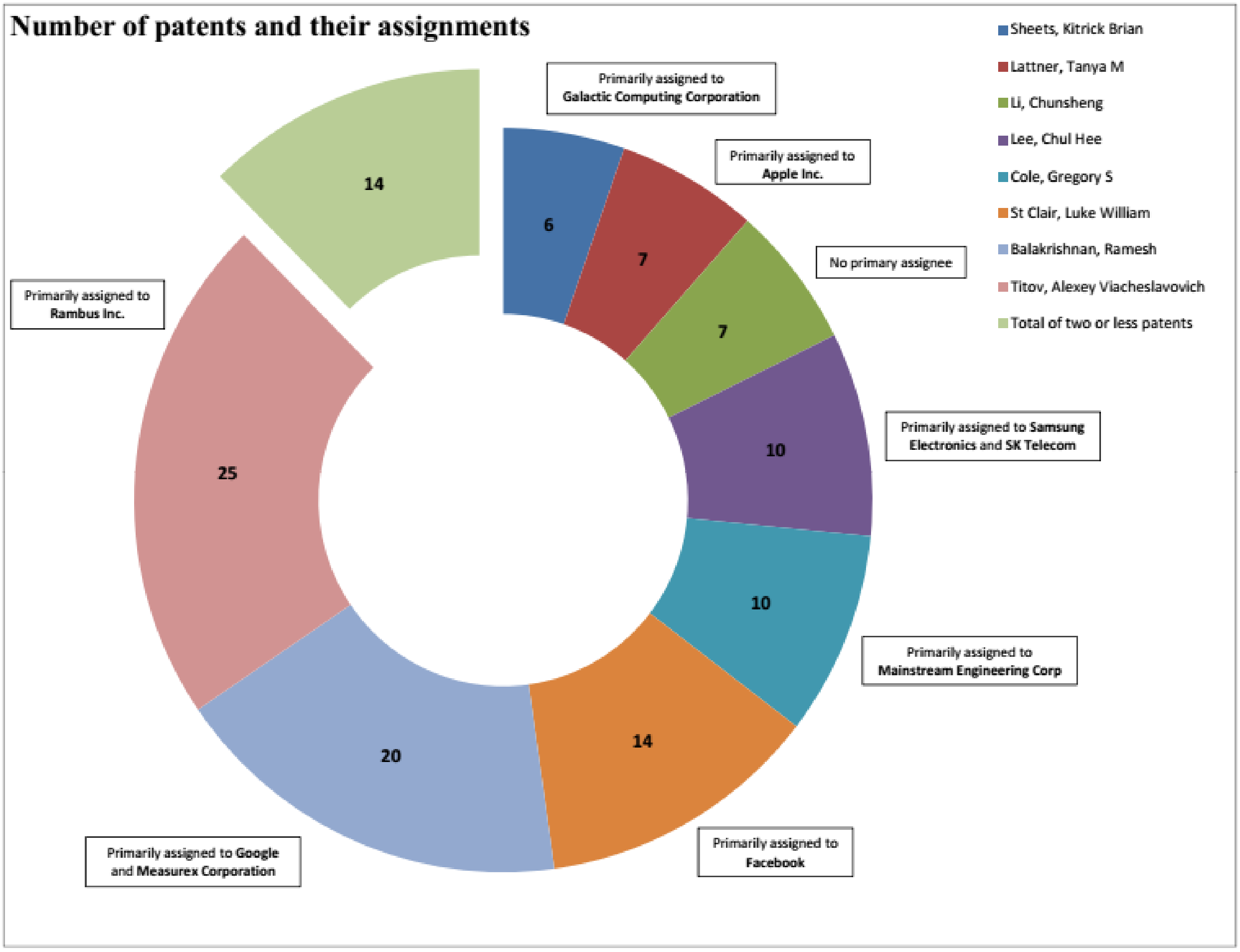


Fig 7: Number of patents made by former NCSA employees and those patents' assignments.

The number of publications by former NCSA employees was graphed in a bar chart (Fig. 8 - source: Google scholar).

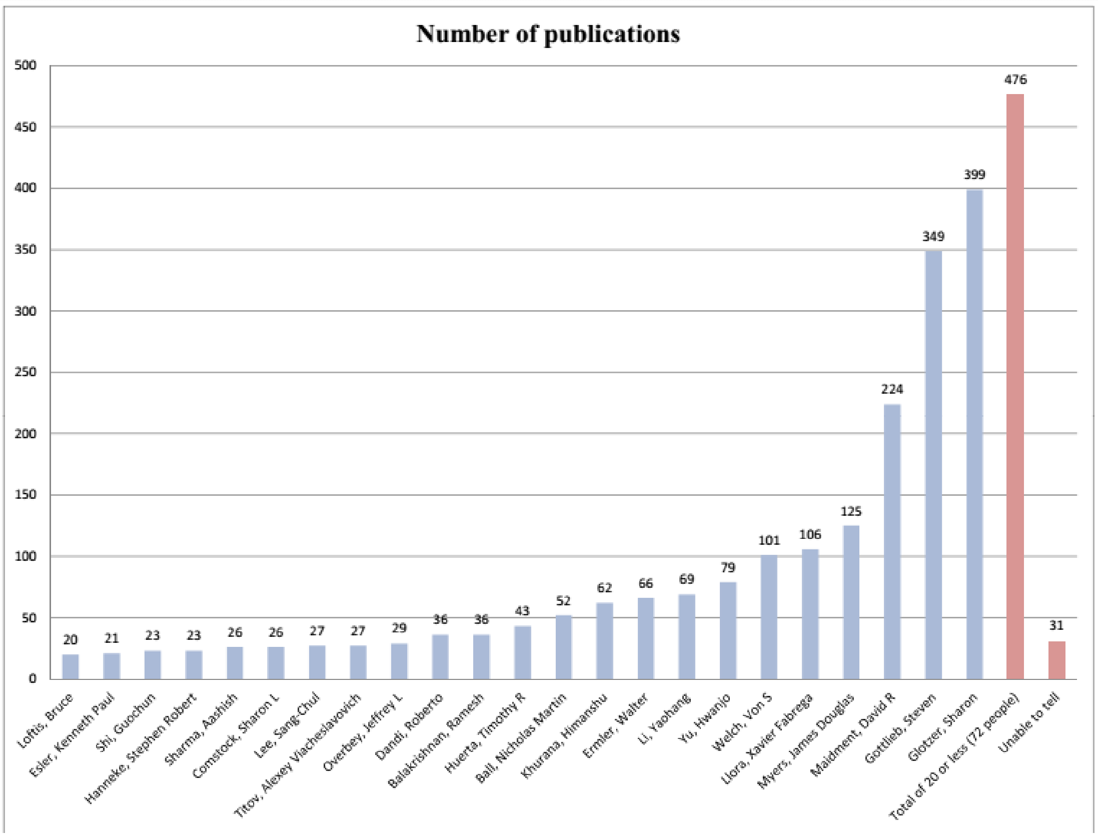


Fig. 8: Number of publications written by former NCSA employees.

Lastly, the researcher created a bar graph to show the amount of grant money awarded to former NCSA employees (Fig. 9 - source: grants.gov).

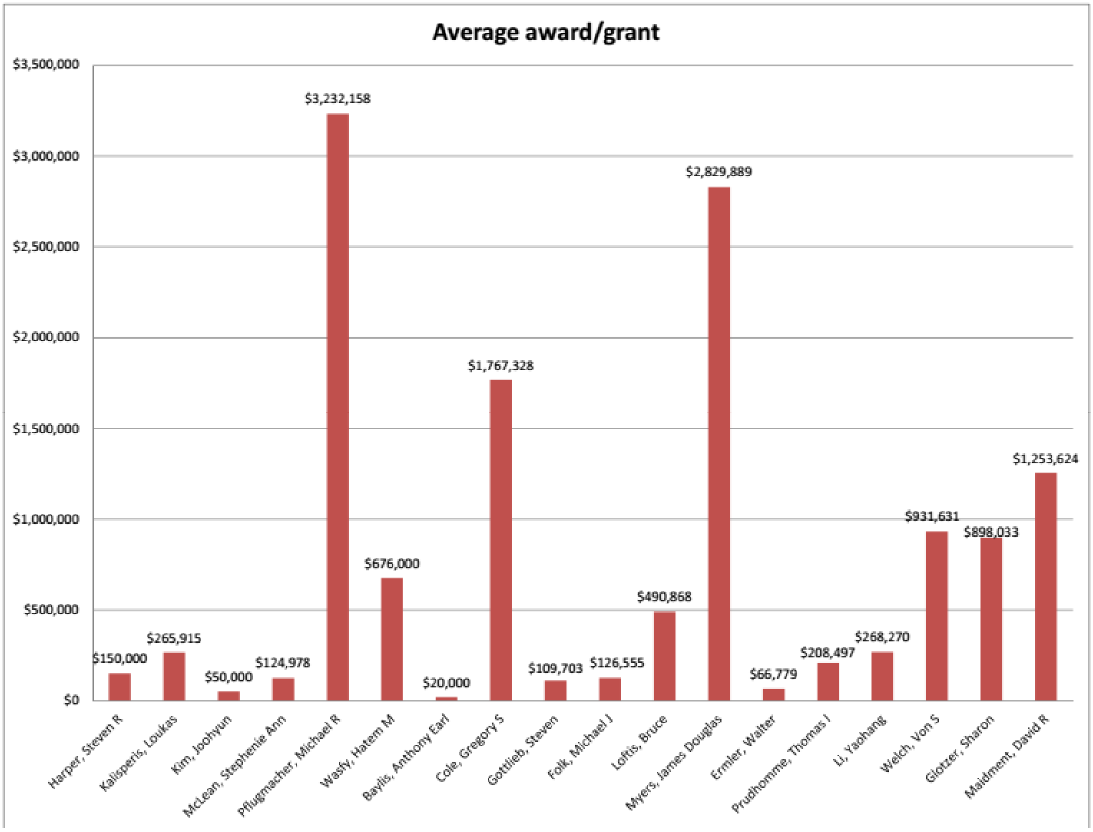


Fig. 9: Grant money received by NCSA diaspora.

Conclusions and Future Directions

This exercise represented an attempt to make sense of NCSA's impact through its diaspora. All of these visualizations when looked at together give a good understanding of the career paths and intellectual outputs of subjects after they left the NCSA.

Certainly, more and different visualizations are possible, and following are some ideas for future work:

- Address a larger sample from NCSA - including a greater time period and/or including more former employees.
- Differentiate among different types of employees and their patterns of diaspora.
- Normalize population figures by state population size or net migration status, as well as providing normalizations or comparison points for other visualizations.
- Some of career paths did not remain within science and computing (for example, one alum of NCSA was with a law firm). This could be an interesting issue to explore separately. Researchers on this project did include all undergraduate student workers in their analysis (at least those earning above $20k), but this is an issue that should be looked at more closely in future studies.
- More complicated visualizations, including maps that have more accurate depictions of time and bubble plots that include more data, could be created by researchers.
- Include other CI enterprises - perhaps go back to the PACI program and understand the diaspora of the NSF's HPC program.

Appendix: Method & Challenges

Pilot Project Research Method

Two graduate students (Cheng and Sutton) conducted the data analysis and developed the visualizations with 80 hours of effort each. The analysis proceeded as follows.

After obtaining information on former staffers from the NCSA, the research team used publicly available data source to put together a picture of former employee careers. The researchers visualized the data to clearly show the migration of workers and their output. By tracking and visualizing four areas of geography, industry, type of company, and intellectual output, researchers attempted to discern how NCSA employees not only add value to the NCSA's activities, but also to the global scientific community and national workforce.

NCSA provided very basic information about employees who left the center between 2003 and 2013. These records gave basic information about former employees, including the employee's name, employment type, annual pay and the date he or she left the NCSA. These records included employees who left between the years of 2003 and 2013. The NCSA provided records for all employees, but the research team decided to limit their tracking to NCSA employees who had an annual salary of $20,000 or higher. This number was chosen as it excluded employees whose time at the NCSA would have been very short. Also, this limited the list of records to 425 subjects, making it a manageable number to research and analyze for a pilot project. In the end, the annual pay amounts ranged from $20,800 to $164,377.90. The subjects' mean pay was $44,267.31, and the median pay was $39,000.

The employee type categories were "Academic Professional," "Academic Hourly," "Post Doc," "Grad Assistant," "Grad Hourly," "Student," and "Extra Help." While some of those titles had consistent annual pay numbers attached to them (in the "Student" and "Grad Assistant" categories, for example), most of the categories had large pay variance within them.

The graduate student assistants used this data provided by the NCSA as a starting point for gathering the rest of the information about the employees' employment histories. The students created a master Google spreadsheet as a place to keep all the information, as it would be easy to collaborate on the spreadsheet online. After entering the information provided by the NCSA, the graduate students would enter all the information about the subjects' jobs after the NCSA in the same spreadsheet.

To find information about the employees employment track, the two graduate student assistants used the LinkedIn Premium service to find subjects' virtual résumés. Some subjects were very easy to find through this process, as they had a full résumé, including their time at the NCSA, on their LinkedIn profile. If a profile was this complete, the searching and data entry process would usually take no more than five minutes. Other subjects were harder to find. This could be because they had a very common name, or they did not include much information about themselves on LinkedIn. If this was the case, the graduate assistants would try to find the subject from various identifying traits, such as any connection to the NCSA or their college institution if the NCSA listed them as a student worker. The graduate assistants ranked the quality of the LinkedIn information on a 3 point scale - 1 was an incomplete or ambiguous profile, 2 indicated a profile that the graduate student was sure of the identity of the subject, but didn't give full information about his or her career, and 3 indicated a full profile. These rankings, while not explicitly used in this study, could be used in future studies to determine the quality of the subject data. A link to the subject's résumé or LinkedIn profile was included in the spreadsheet for future researchers.

Every so often, a subject could not be found on LinkedIn. The graduate students would then search on Google for the person, using their name and, if necessary, an identifier such as "NCSA" or "University of Illinois." Sometimes, subjects would have their résumés on a personal or university website, which then the students would use to fill out the spreadsheet about the subject. If, after about twenty minutes of searching, the graduate assistants were not able to find the subject, the record would be marked as empty in the spreadsheet.

Once the graduate assistants had the employment information data about the subjects, they would enter it in the spreadsheet. For each job, they would include the job's location (city, state and country were later separated to make it easier for visualization), the position title, the position firm and the length of time in years and months that the subject stayed in the position. Also, if the subject included any information about the position in his or her profile, we would copy that into a "job paragraph" section. This paragraph information could be used to help determine the industry of the position, especially when the company or job title was vague.

The graduate assistants would also search for any intellectual contributions the subjects made to the scientific world. The team decided on three intellectual outputs to track - papers published, patents, and the amount of grant money they have received. To search for these items, the graduate students used the Google Patent and the Google Scholar search engine, as well as the grants.gov search engine. The students also included any patents, grants or papers listed on the subjects' websites or LinkedIn profiles. For papers, the students included the link to their Google Scholar search (and Google Scholar profile, if it existed) and noted the number of papers. For patents, the students noted the patent numbers and the number of patents. Lastly, for grants, the students included the grant number, the number of grants and the total amount of money they won.

After all the data was inputted, the students created visualizations to make the data easier to understand. Visualizations showed the geographic migrations of former NCSA employees, the industries that employees worked in and the type of companies they worked for. Also visualized was the intellectual output of the subjects, which included the amount of grants they received, and the number of patents and publications they produced.

Visualizations were created in javascript using the D3.js Library (http://d3js.org) by Mike Bostock.

Challenges

Although the project ran smoothly, researchers did run into a few challenges, especially when researching the subjects. In the future, these issues could be avoided with a few changes to the process of the study.

When asked for former employment data, the start dates of the subjects' employment were not provided. This made it impossible to tell how long the subject worked for the NCSA, unless they explicitly stated this information on their LinkedIn profiles or their websites. As researchers studied the subjects in closer detail, it became clear that some had only been with the NCSA for a simple project or for a very short time. Although this time could have still had a profound effect on the subject's career trajectory, it seems harder to prove. In the future, researchers should have the length of employment at NCSA from the beginning.

Related to this issue is the difference between types of employment. While "academic professional" employees seem more likely to have career trajectories that match the industry of the NCSA, student employees were not as clear. Without knowing what an undergraduate student employee did at the NCSA, it's hard to make assumptions about their career path post-NCSA.

As mentioned earlier, the amount of information available about the subjects varied widely. While some subjects had a full LinkedIn profile, others barely had any information about themselves at all. Researchers addressed this issue by using the ranking system described in the "Methods" section of this report. Although researchers did not use the LinkedIn completeness scores in the final visualizations, future studies should find a way to utilize this quality information.

While coding the subjects' jobs based on industry, researchers had issues determining how specific to be. By being too broad, researchers risked missing interesting trends in the industries subjects moved to after the NCSA, but by being too narrow, the data would get too specific to show large trends. Also, certain large software companies, such as Microsoft or Apple, could be filed under many different industries when they are defined in narrow terms. Researchers should choose a list of industries from the beginning that will suit the data, and then try to stay with those as closely as possible to prevent confusion. Also, this would help to prevent researchers from using different terms for the same category, making for cleaner, more useful data. In this study, researchers went back through the data after everything was inputted to clean up the industry category, but it would be better if these categories could be defined from the beginning.

Lastly, in this study, researchers chose to use the spreadsheet function within Google Drive to hold the information. Google Drive offers flexible tools for free, and, helpful in this case, allow various users to access and edit the files at the same time. Researchers chose Google Drive for that reason, but they quickly ran into the limitations of the online software. As the spreadsheet got larger, it started to be unwieldy and very slow in loading. Also, the searching function would have been easier to use if researchers had created a true database for the data. Although Google Drive was a useful tool for this pilot program, researchers would recommend exploring different, more flexible options for future studies.