Understanding the NERSC Collaboration Network through Graph Analysis

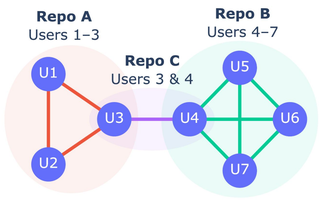
Coby Sandholtz[1], Camelia D. Brumar[2], Sumaiya Azad[3], Kevin Gott[4], Rebecca Hartman-Baker[4]

[1]Brigham Young University, [2]Tufts University, [3]University of Utah, [4]Lawrence Berkeley National Laboratory

Research Question

Which users occupy the **most structurally influential positions** in the NERSC collaboration network?

Background



Within Lawrence Berkeley National Laboratory (LBNL), NERSC provides HPC resources to support research funded by the DOE Office of Science.

When a NERSC user receives an allocation, data such as their name, the git repository their allocation is tied to, the project title, and their organization is recorded as "allocation data."

This analysis investigates using this data to **model the network of NERSC users** (see the above graphic for details).

Method

Individual yearly allocation datasets were combined into a cumulative dataset representing all allocations from 2020 until 2026. Python libraries cugraph and igraph were used to construct a graph (hereafter "NERSC graph") based on the cumulative allocation data. Three analyses were performed:

Preliminary Analysis	What do the NERSC graph's global properties reveal about the NERSC collaboration network?
Degree Distribution	How do the NERSC graph's collaborative properties compare to other scientific networks?
Network Representation	Which graphical representation methods best preserve connection strength?

Acknowledgements:

My gratitude goes to Kevin Gott and Rebecca Hartman-Baker for their generous mentorship, to NERSC staff for providing context and feedback, and to the Computing Sciences Summer Program for support.

References:

[1] M. E. J. Newman, "The structure of scientific collaboration networks," Proceedings of the National Academy of Sciences, vol. 98, no. 2, pp. 404–409, 2001, doi: 10.1073/pnas.021544898.
[2] M. E. J. Newman, "Scientific collaboration networks. I. Network construction and fundamental results," Physical Review E, vol. 64, Art. no. 016131, 2001, doi: 10.1103/PhysRevE.64.016131.
[3] M. E. J. Newman, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," Physical Review E, vol. 64, Art. no. 016132, 2001, doi: 10.1103/PhysRevE.64.016132.

Preliminary analysis

The NERSC graph was compared to graphs representing scientific collaboration networks based on publication data from three databases [1]:

1. **Stanford Public Information Retrieval System (SPIRES):** a repository of pre-prints and papers in high-energy physics. 1974-2001.
2. **Networked Computer Science Technical Reference Library (NCSTRL):** A repository of pre-prints in computer science. 1990-2001.
3. **Los Alamos e-Print Archive (LANL):** A repository of pre-prints in physics. 1992-2001.

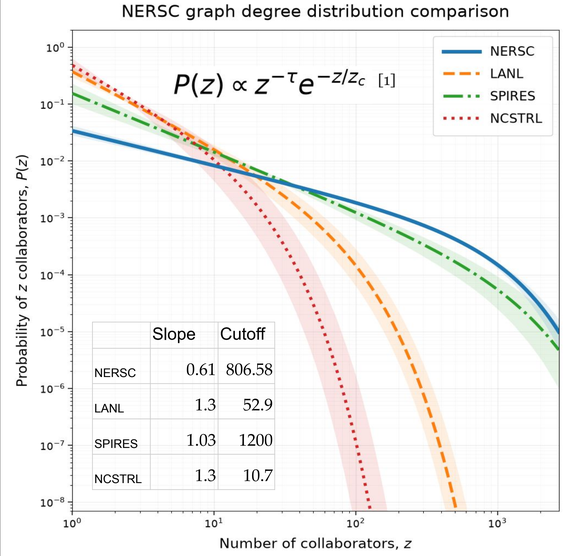
Discussion:

- Aside from the conglomerate all-offices NERSC graph, the two **highest scoring graphs in collaborators per scientist** (HEP and SPIRES) are **both in high-energy physics.** High-energy physics is known for having much larger scientific collaborations than other fields. [2]

Metric	LANL	SPIRES	NCSTRL	Avg non-NERSC	HEP (NERSC)	FES (NERSC)	NP (NERSC)	NERSC (All offices)
Total edges	98,502	66,652	13,169	59441	160	238	133	2,607
Total nodes	52,909	56,627	11,994	40,510	5,419	1,801	1,463	23,286
Mean edges per node	5.10 ± 0.20	11.60 ± 0.50	2.55 ± 0.05	6.42	1.38	1.58	1.26	1.63
Mean nodes per edge	2.53 ± 0.01	8.96 ± 0.18	2.22 ± 0.01	4.57	46.74	11.97	13.89	14.58
Neighbors per node	9.70 ± 0.20	**173.00** ± 6.00	3.59 ± 0.05	62.1	**443.79**	51.47	58.39	179
Mean distance	5.90 ± 0.20	4.00 ± 0.10	9.70 ± 0.40	6.53	2.56 ± 0.04	3	4.97	3.55 ± 0.05
Clustering coefficient	0.430 ± 0.010	0.726 ± 0.008	0.496 ± 0.006	**0.551**	0.787	0.645	0.825	**0.766**

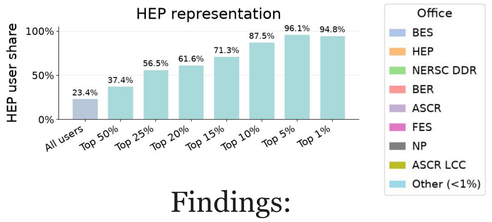
Legend: HEP: *High-energy Physics.* BES: *Basic Energy Sciences.* NERSC DDR: *NERSC Director's Discretionary Reserve.* BER: *Biological and Environmental Research.* ASCR: *Advanced Scientific Computing Research.* FES: *Fusion Energy Sciences.* NP: *Nuclear Physics.* ASCR LCC: *ASCR Leadership Computing Challenge.*

Degree Distribution Analysis



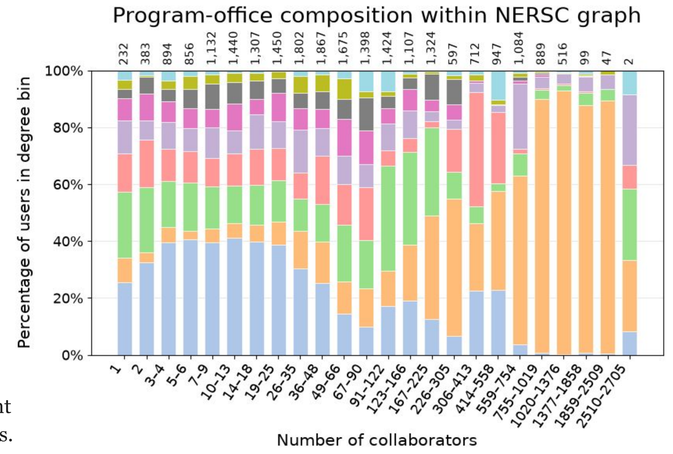
Which users dominate the NERSC network?

"Degree Distribution" is a breakdown of the nodes in a graph by their level of connectedness.

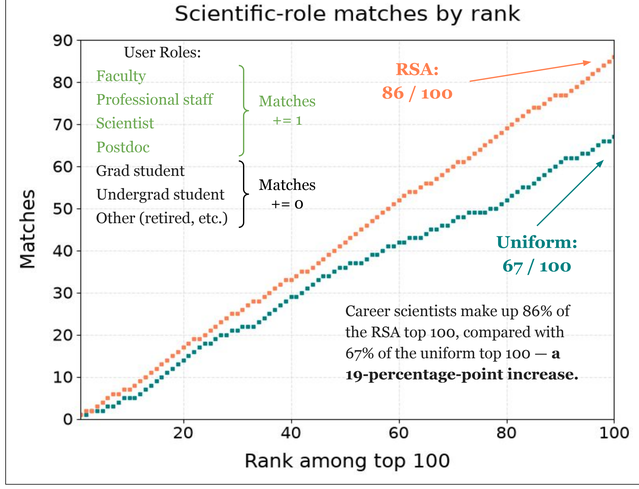


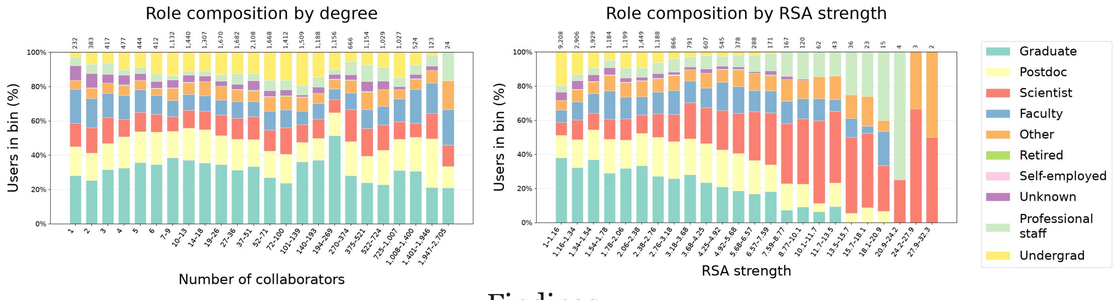
Findings:

- The NERSC network's share of highly-connected users is **greater than any other scientific network considered.**
- Users affiliated with **high-energy physics** account for **96.1%** of NERSC's top 5% most connected users.



Network Representation



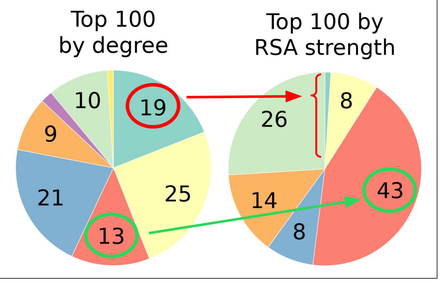


Can NERSC's graph representation be improved?
Consider two edge weightings:

1. *Uniform.* Edges have a weight of 1.
2. *Repository Size Adjusted (RSA).* Edge weight is inversely proportional to the project's total number of collaborators. [3]

Findings:

- The RSA weighting **better represents the strength of scientific relationships** than the uniform weighting.
- NERSC users who are **well connected** (based on RSA degree) are **more likely to be career scientists.**



Conclusions and Future Work

- The **strongest connectors** in the network are **career scientists** and users associated with **High-Energy Physics.**
- HPC network data, **unweighted connectivity is not a reliable measurement** of network influence or cross-disciplinary reach.
- Allocation networks may provide a **useful proxy for coauthorship networks** in the same field.

Next steps: Use repo-size-adjusted (RSA) edge weighting in the program office-focused degree distribution analysis to test the dominance of high-energy physics.

Analyze the composition of code usage in logarithmic degree bins.

Artifact: