Exploring Interdisciplinary Connections in Duke Ph.D. Committees
2018 Scholars@Duke Visualization Challenge

Matthew Epland
Duke University, Durham, NC 27707

January 22, 2018

I. INTRODUCTION

This submission to the 2018 Scholars@Duke Visualization Challenge[1] explored the nature of interdisciplinary research at Duke by studying the connections discovered in Ph.D. committees for the 2013–2017 academic years. By combining the committee membership data with the faculty appointments directory, connections between different academic organizations were found and used to construct an undirected, weighted graph. From this graph communities of closely connected academic organizations were created via the Louvain method, and the level of interdisciplinary activity in each organization was measured by comparing the relative weights of their external and self connections.

II. METHODS

A. Constructing the Academic Organizations Graph

Individual Ph.D. committees were identified in the `dissertation_committees_2012-2017.xlsx` dataset provided by the Graduate School by computing a unique student/committee ID[2]. Incomplete and potentially corrupted committees[3] were removed. Using the `ScholarsAtDuke_Faculty_October2017.xlsx` dataset provided by Scholars@Duke, committee members were matched to faculty appointments via their Duke unique ID numbers (DUID). At Duke, can hold one primary appointment and multiple secondary or joint appointments in other academic organizations[4]. Each time a faculty member appeared on a committee they were replaced by all of the academic organizations where they held appointments. From this committee level list of organizations[5], ie nodes, all possible combinations of two organizations were found. Each combination was saved, along with the degree conferred date, to a list of edges. The final academic organizations graph could then be constructed edge-by-edge, increasing the weight w of a particular edge by 1 each time it appeared in the list. A schematic representation of this process is provided in Figure 1. The graph building code was written in `python` using `pandas` [1] for data management and `networkx` [2] for graph operations. The complete codebase for this analysis can be found on GitHub[6]. The primary Jupyter notebook that generates the final plots is also viewable through Bl.ocks[7].

---

[1] https://rc.duke.edu/scholars-vis-challenge-2018/
[2] {Student random ID}_{Degree Nbr}_{Compl Term}_{Acad Org}
[3] Incomplete committees having less than 4 members, and 1838_2_1420_ELEC&CMP
[4] Administrative appointments and organizations were removed as they did not add to the study of interdisciplinary connections between academic organizations. Many organizational unit numbers were merged to clean the data. Additionally similar, but formally distinct, organizations were merged by hand in order to simplify the number of organizations — in particular the numerous Medical School subdisciplines.
[5] Including duplicates if members shared any common appointments
[6] https://github.com/mepland/vis_challenge_2018
[7] http://bl.ocks.org/mepland/raw/4cf24fbc77944c185d1d27fad64a5dce

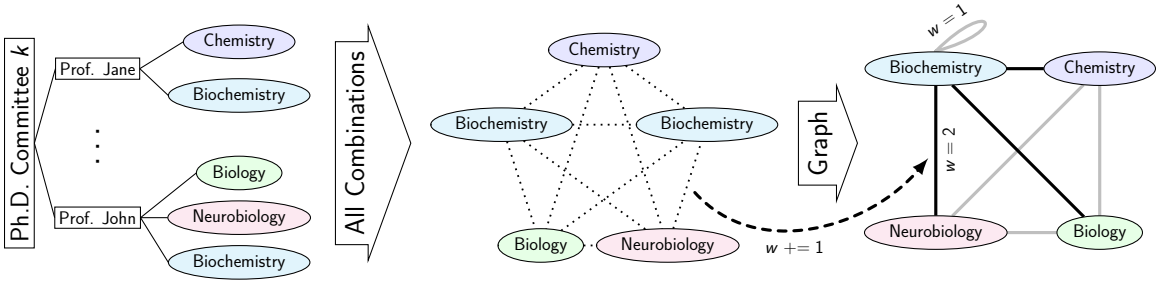

FIG. 1. Schematic representation of the method used to build the weighted academic organizations graph. Two members from a single committee are illustrated for example. In practice the method is applied to all committees and all members.

B. Finding Communities

The academic organizations graph naturally contains sub-groups, or communities, of related disciplines, such as the Physical Sciences or Liberal Arts. These communities can be constructed algorithmically via the Louvain method [3] which optimizes the graph's modularity, a measure of the density of interior to exterior edges of the constituent communities. The modularity Q of graph G can be defined as (1) where w_{ij} is the edge weight between nodes i and j, W_i is the sum of edge weights of node i, W_{G} is the total edge weight of the graph, and c_i is the community of node i.

$$Q\left(G\right) = \frac{1}{2W_{\mathrm{G}}} \sum_{ij \in G} \left(w_{ij} - \frac{W_i W_j}{2W_{\mathrm{G}}} \right) \delta\left(c_i,\, c_j\right) \tag{1}$$

In this analysis the Louvain method was implemented via the `python-louvain` package [4] with the resolution parameter[8] set to the default value of 1.

C. Measuring Interdisciplinary Activity

To measure the interdisciplinary activity of each academic organization a straightforward interdisciplinary fraction f of external and self connections was utilized (2). Here w_{external} is the sum of external edge weights of an organization's node, while w_{self} is the weight of the edge from the node to itself. Binning the academic organizations graph by academic year[9] it is possible to see how f changes for an organization over time.

$$f = w_{\mathrm{external}} / \left(w_{\mathrm{external}} + w_{\mathrm{self}} \right) \tag{2}$$

f works well for Ph.D. granting organizations with good statistics, but frequently breaks down with a value of $f = 1.0$ for non-Ph.D. granting organizations as they do not have multiple faculty members sitting together on their own Ph.D. committees. To help remove such cases from consideration it is required that $w_{\mathrm{total}} = w_{\mathrm{external}} + w_{\mathrm{self}} > 100$ per year, and that an organization have ≥ 3 such years before being displayed.

---

[8] A resolution of 1 corresponds to the standard Louvain method, while diverging from 1 favors communities of different sizes. Other values were tested, but the best results were obtained with a resolution of 1.

[9] With bin edges: 2012–5–1, 2013–8–26, 2014–8–25, 2015–8–24, 2016–8–29, 2017–10–1

III. RESULTS

A. Academic Organizations Graph

In addition to being the base object for later analysis, the academic organizations graph for all years, Figure 2, provides a useful high-level view of the interdisciplinary networks at Duke. At a glance one can see how tightly linked organizations form the core of communities[10] with smaller organizations on the periphery, and the relative separation between the scientific / medical communities and the liberal arts. The binned graphs for each academic year may be found in Appendix A.

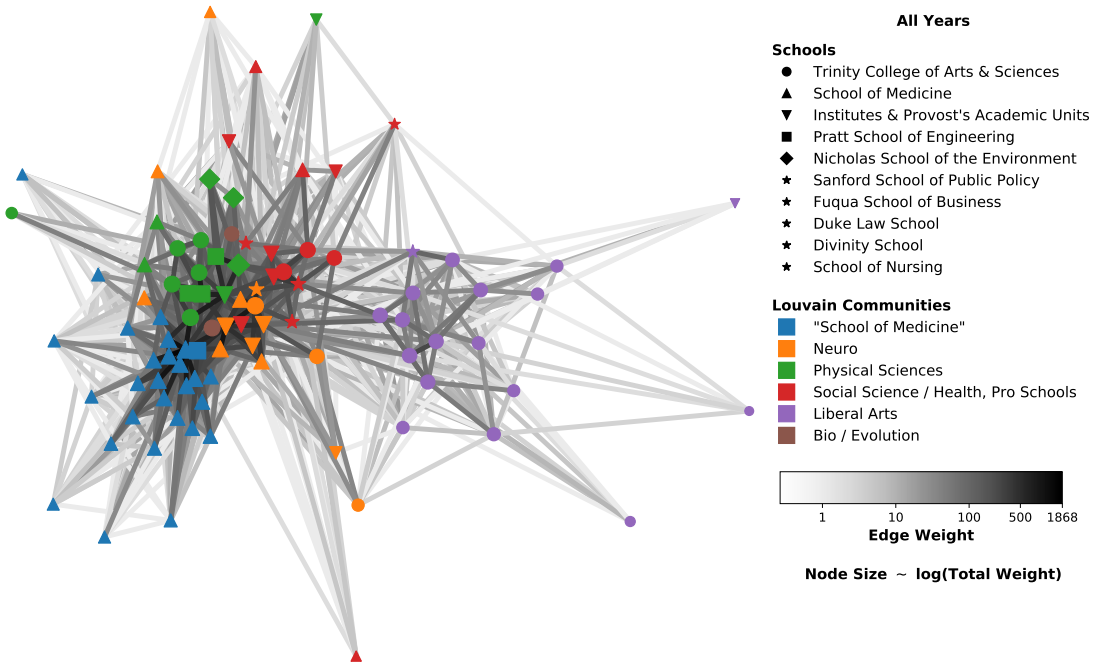

FIG. 2. Academic organizations graph for all years.

The graph for all years may also be viewed interactively online[11], displayed with the `visJS2jupyter` package [5]. There additional details for each node and edge may be viewed by hovering over them, and the nodes may be dragged into new positions to better examine certain areas.

B. Communities

When run on the academic organizations graph for all years, the Louvain method found 6 communities of varying sizes. Each community was then named in order to summarize its constituent organizations; “School of Medicine”, “Neuro”, “Physical Sciences”, “Social Science / Health, Pro Schools”, “Liberal Arts”, and “Bio / Evolution”. Most communities contained the organizations one would expect, with a few random additions. The large Neuro community incorporating organizations from multiple schools across campus was an interesting find, as was the insular Biology / Evolutionary Anthropology paring. Surprisingly the Biology and Evolutionary Anthropology departments did not

[10] The node positions are set via the Fruchterman-Reingold force-directed spring algorithm which shortens high weight edges, and lengthens low weight edges.

[11] http://bl.ocks.org/mepland/raw/598590f30f49b17dc76ea4ed74695252

join any larger community, despite several appearing compatible from a traditional disciplinary point of view, but instead paired with themselves. See Appendix B for a complete listing of organizations in each community.

C. Interdisciplinary Activity

The interdisciplinary fraction f vs year was plotted for the top 10 organizations by total weight in each community, see Figures 3–5 for three interesting examples.

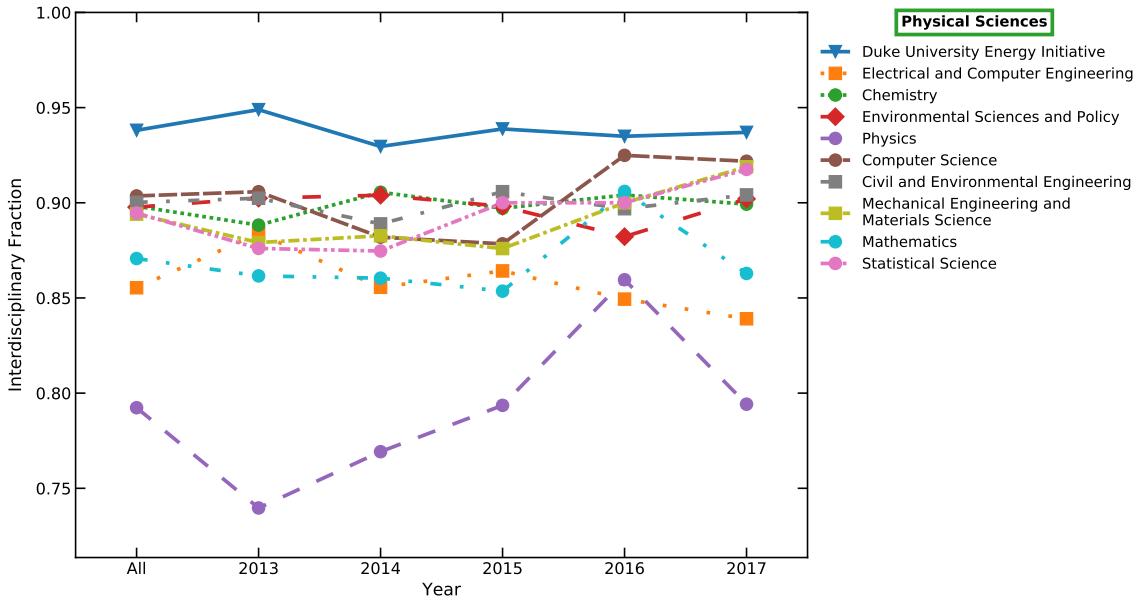

FIG. 3. Interdisciplinary fraction vs year for the Physical Sciences community.

In the Physical Sciences community the majority of the top 10 organizations had fairly steady $f \approx 90-95\%$, with the exception of Physics which had wide variations between $f \approx 75-90\%$. In the Neuro community the majority of organizations fell a bit higher at $f \approx 94-98\%$, with Psychology and Neuroscience, and Philosophy varying between $f \approx 84-94\%$.

The lower f values and increased year-to-year variations in the Physics, Psychology and Neuroscience, and Philosophy departments is intriguing and warrants further investigation. Two hypotheses for why they behave differently from their peers is that these departments have stricter policies regarding faculty holding joint and secondary appointments in other departments, or including multiple Ph.D. committee members from outside the field. Further analysis efforts described in Section IV could help test these hypotheses, as would qualitatively reviewing the department cultures and policies.

In contrast to the Physical Sciences and Nero communities, organizations in the Liberal Arts were lower at $f \approx 75-90\%$, but suffered from low statistics which increased the variance and limited the number of organizations passing the w_{total} selection to only 5. Additional data is needed from these organizations before quality comparisons between the sciences and liberal arts can be made.

The remaining plots of f for each community can be found in Appendix C. Additionally similar plots were produced for the top 10 organizations by total weight in each school, see Appendix D.

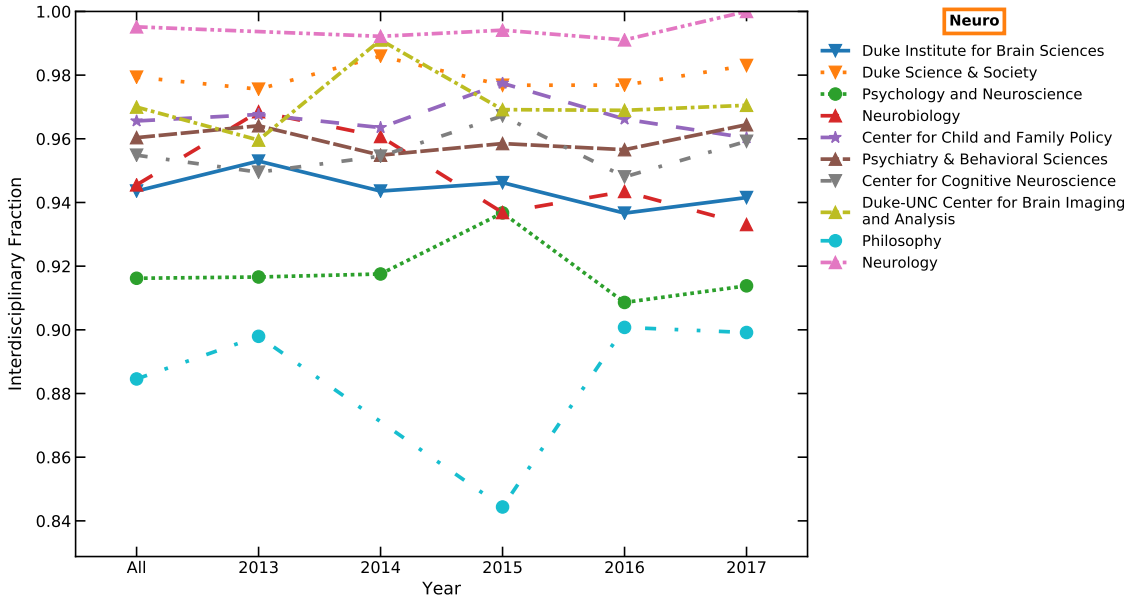

FIG. 4. Interdisciplinary fraction vs year for the Neuro community.

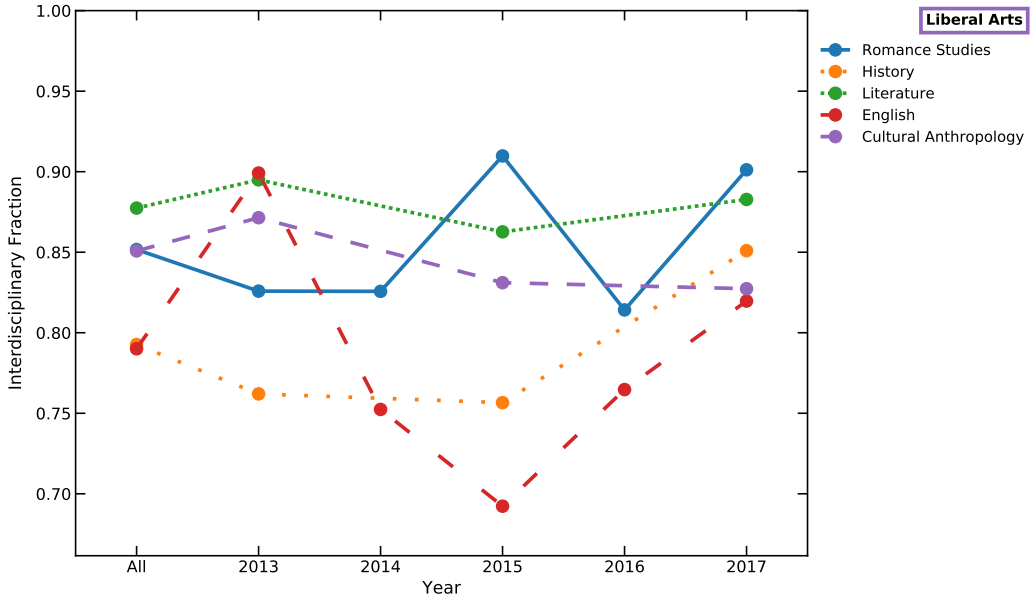

FIG. 5. Interdisciplinary fraction vs year for the Liberal Arts community.

IV. POTENTIAL ISSUES AND FUTURE IMPROVEMENTS

Due to time constraints imposed by the challenge a number of potential issues in and improvements to the analysis were identified but could not be investigated. They are listed here in the spirit of transparency and for possible implementation in the future. Note that some solutions presented here should improve multiple aspects of the analysis simultaneously.

1. *Non-Ph.D. Granting Academic Organizations Underrepresented*

As the `dissertation_committees_2012-2017.xlsx` dataset only contains information on Ph.D. committees, academic organizations such as professional schools who typically grant other kinds of graduate degrees, and interdisciplinary institutes and centers who do not directly grant graduate degrees of any kind, are underrepresented. This leads to poor statistics and frequent unrealistic $f = 1.0$ break downs for these organizations.

An easy solution, provided the data is available, is to request and integrate the non-Ph.D. committee records from the graduate and professional schools. However this does not address the issues with organizations that do not grant any graduate degrees. A potentially wider solution is to switch datasets entirely and utilize the `ScholarsAtDuke_Publications_2012-2017.xlsx` publication data instead. There joint authorship on a paper could be used in the exact same way as joint membership on a committee to construct a new graph using much of the existing procedure and code, but would constitute essentially re-running the entire analysis.

2. *Effects of Joint and Secondary Appointments vs Committee Membership*

The academic organizations graph is currently constructed such that the weight added to an edge of two organizations connected from one faculty member holding appointments in each ($w = 1$) is the same as the weight added to an edge from two faculty members with different appointments serving on the same Ph.D. committee. While there is nothing incorrect with this method a priori, there is also no independent reason for it. Alternative weighting schemes should be devised and tested to determine what works best for this dataset and analysis. Another round of elicitation from the relevant stakeholders would be helpful when forming metrics on which to test the weighting schemes[12], as holding multiple appointments and sitting on an interdisciplinary committee are both interdisciplinary activities, but of potentially different importance.

Two weighting schemes were in fact tested during development, one which only considered primary appointments and the second as presented here in Section II A which weighted primary, joint and secondary appointments equally. The second method was ultimately chosen as it produced a more interconnected graph with reasonable Louvain communities. Other possible weighting schemes to test include weighting joint and secondary appointments at a constant non-zero value less than primary appointments, and normalizing the weights per faculty member such that their primary appointment receives a weight of 0.5[13] while any n joint and secondary appointments receive $0.5/n$ such that each faculty member only contributes a maximum combined weight of 1.

3. *Improved Data Cleaning*

As implemented the process to clean and merge the committee and faculty datasets is fairly strict. Everything is done by the DUID number and if there is a missing or mismatched record the faculty member will be dropped. Some of these cases may be caused by recently retired faculty appearing on past committees, but not in `ScholarsAtDuke_Faculty_October2017.xlsx`; the solution here is to acquire a larger dataset of all faculty from 2012–2017. Others may be due to non-Duke faculty serving as committee members, which is probably intractable with the Duke only sources of data available[14]. Lastly, some faculty mismatches may be the simple result of clerical errors when entering the DUID[15]. In this case a semi-autonomous fallback function could be developed to try to match faculty by name.

---

[12] For example, if a department has restrictive policies regarding joint and secondary appointments, should that be taken as a sign of non-interdisciplinary activity, or be guarded against as a possible source of bias?

[13] Or 1 if $n = 0$ and they only hold a primary appointment.

[14] Barring some extensive publication and web scraping effort.

[15] A handful of committee members have DELETE in their names, so this is a real possibility.

The additional effort needed to improve the data cleaning may not be worth the gain in statistics — particularly if large amounts of new data is being acquired yearly. However, it should at least be studied as a potential source of bias as some academic organizations may be systematically affected by one or more of the above DUID data quality issues.

V. CONCLUSIONS

The nature of interdisciplinary research at Duke was explored at the organizational level by studying connections found in Ph.D. committees from the 2013–2017 academic years. Communities of related academic organizations were created via the Louvain method, most following the typical disciplinary divisions with a few interesting exceptions in Neuro community and Biology / Evolutionary Anthropology paring. The interdisciplinary activity of individual organizations was investigated via the development of interdisciplinary fraction f, which revealed lower values of f with high variances for the Physics, Psychology and Neuroscience, and Philosophy departments. Lastly, future directions and areas of improvement for the analysis were identified, along with possible solutions.

[1] W. McKinney, *Data Structures for Statistical Computing in Python*, in *Proceedings of the 9th Python in Science Conference*. 2010. https://pandas.pydata.org/.

[2] A. A. Hagberg, D. A. Schult, and P. J. Swart, *Exploring network structure, dynamics, and function using NetworkX*, in *Proceedings of the 7th Python in Science Conference (SciPy2008)*.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics: Theory and Experiment **2008** (2008) P10008, http://stacks.iop.org/1742-5468/2008/i=10/a=P10008.

[4] T. Aynaud, *python-louvain, Louvain Community Detection*, https://github.com/taynaud/python-louvain.

[5] S. B. Rosenthal, J. Len, M. Webster, A. Gary, A. Birmingham, and K. M. Fisch, *Interactive network visualization in Jupyter notebooks: visJS2jupyter*, Bioinformatics **34** (2018) 126–128, http://dx.doi.org/10.1093/bioinformatics/btx581.

A. ACADEMIC ORGANIZATIONS GRAPHS BY YEAR

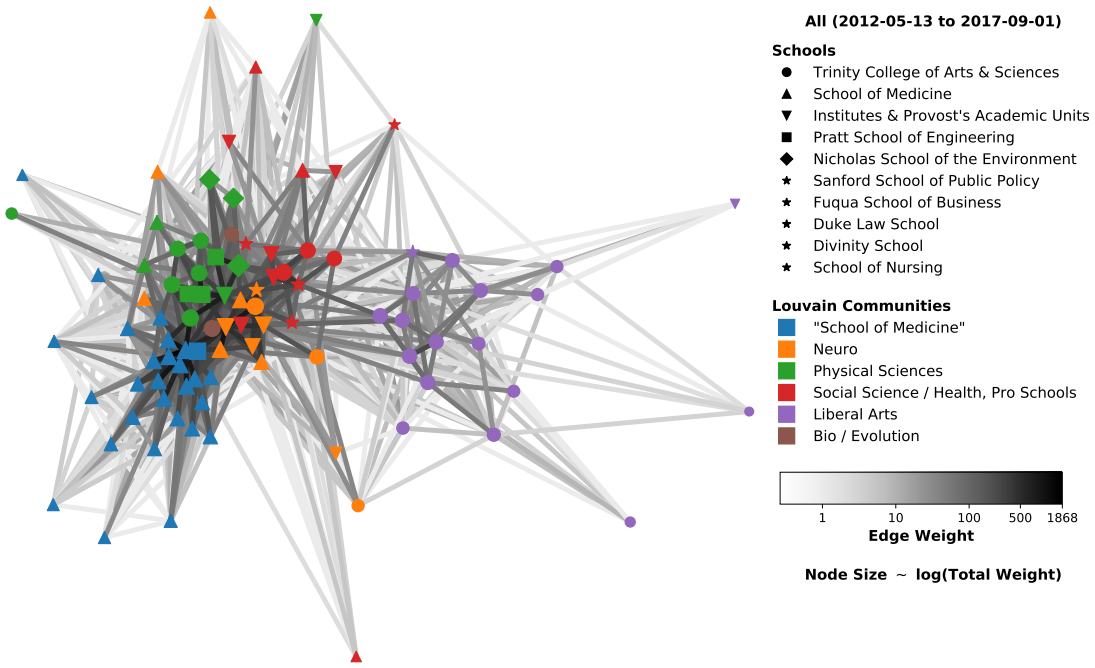

FIG. 6. Academic organizations graph for all years.

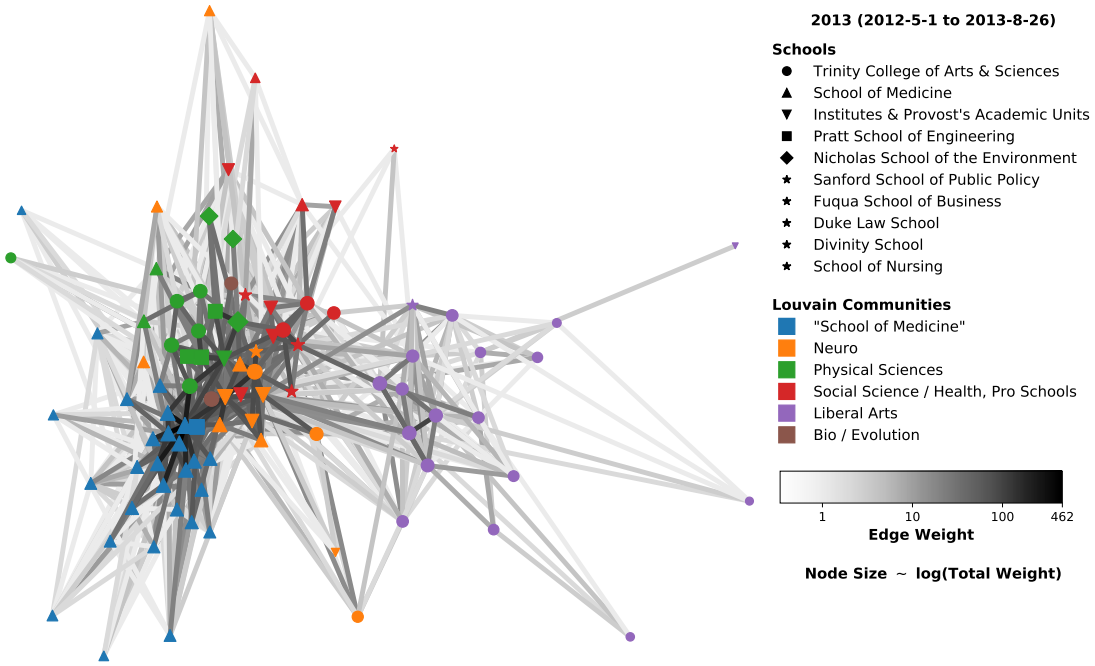

FIG. 7. Academic organizations graph for 2013.

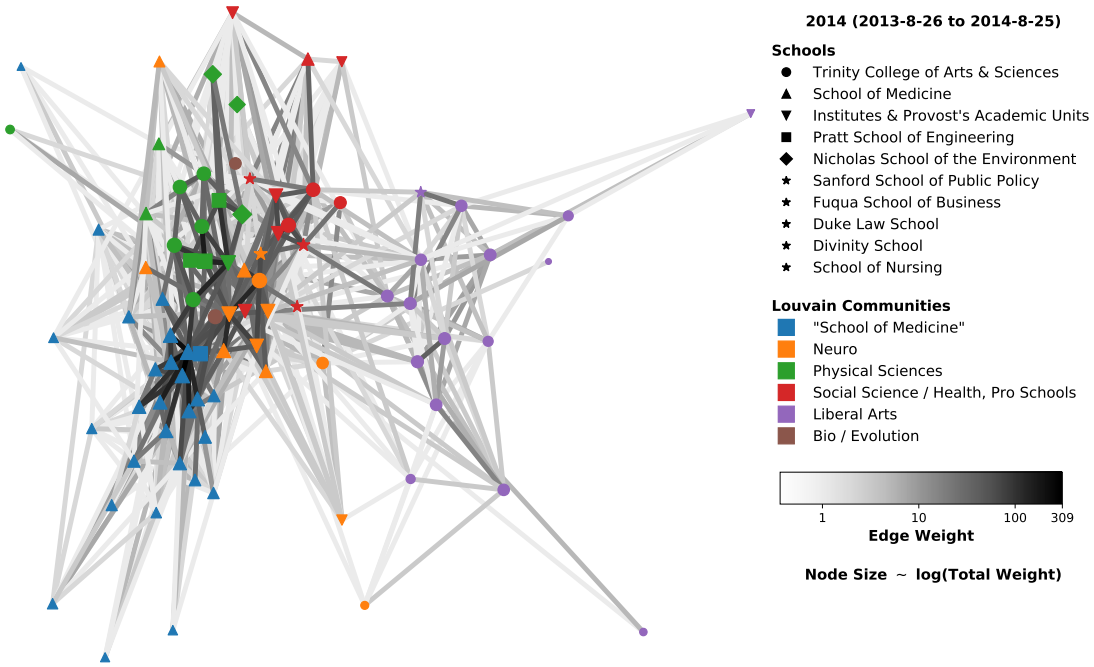

FIG. 8. Academic organizations graph for 2014.

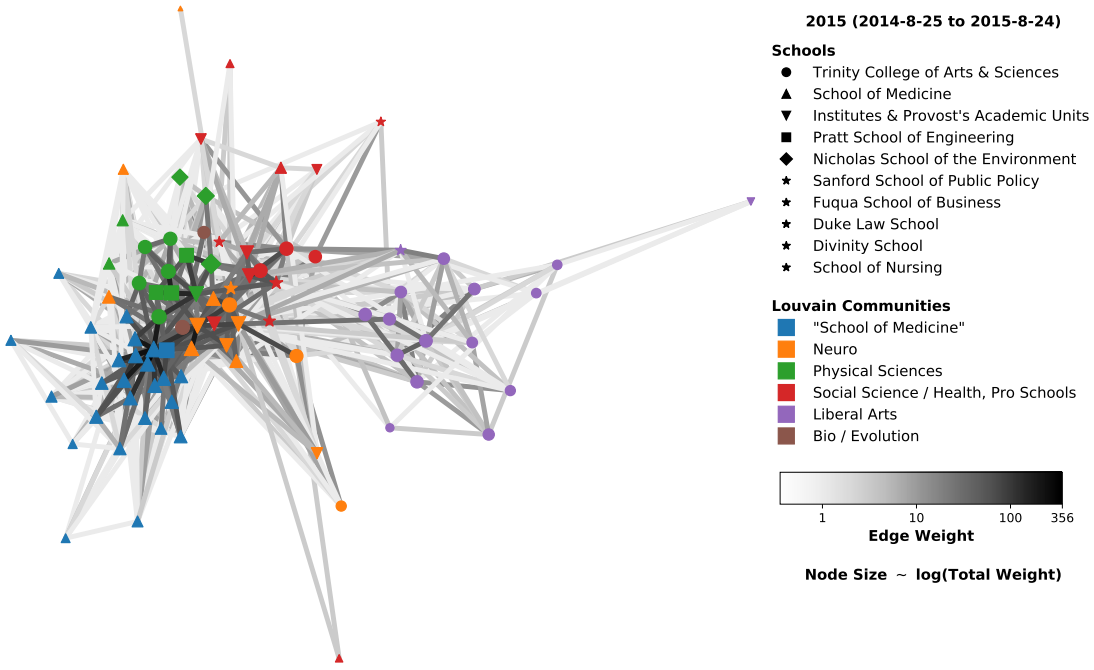

FIG. 9. Academic organizations graph for 2015.

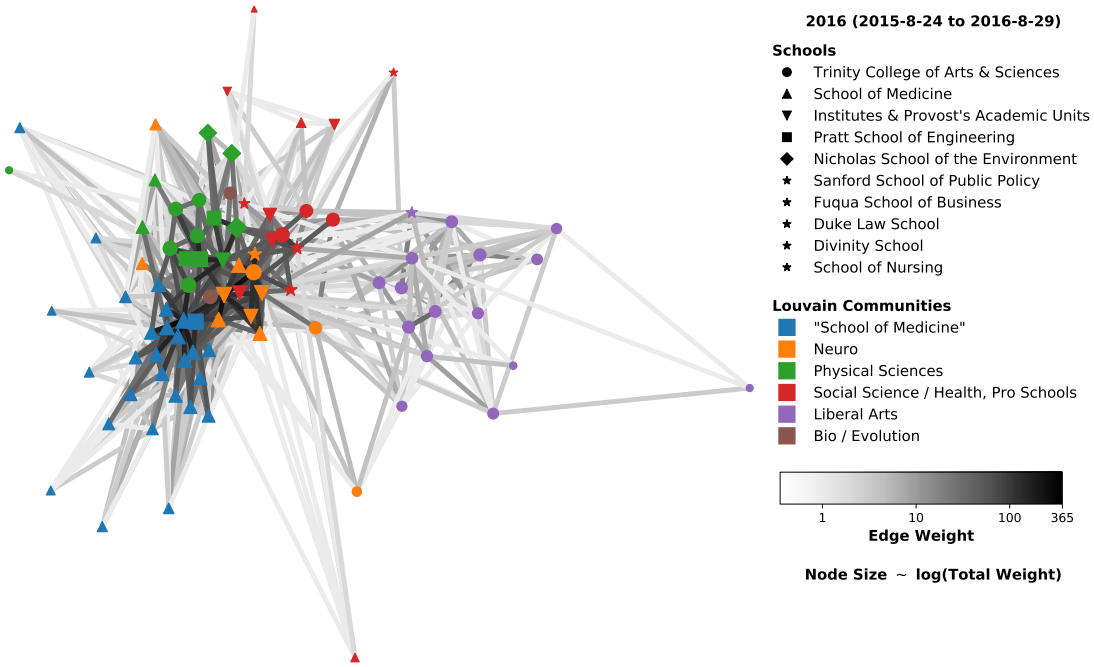

FIG. 10. Academic organizations graph for 2016.

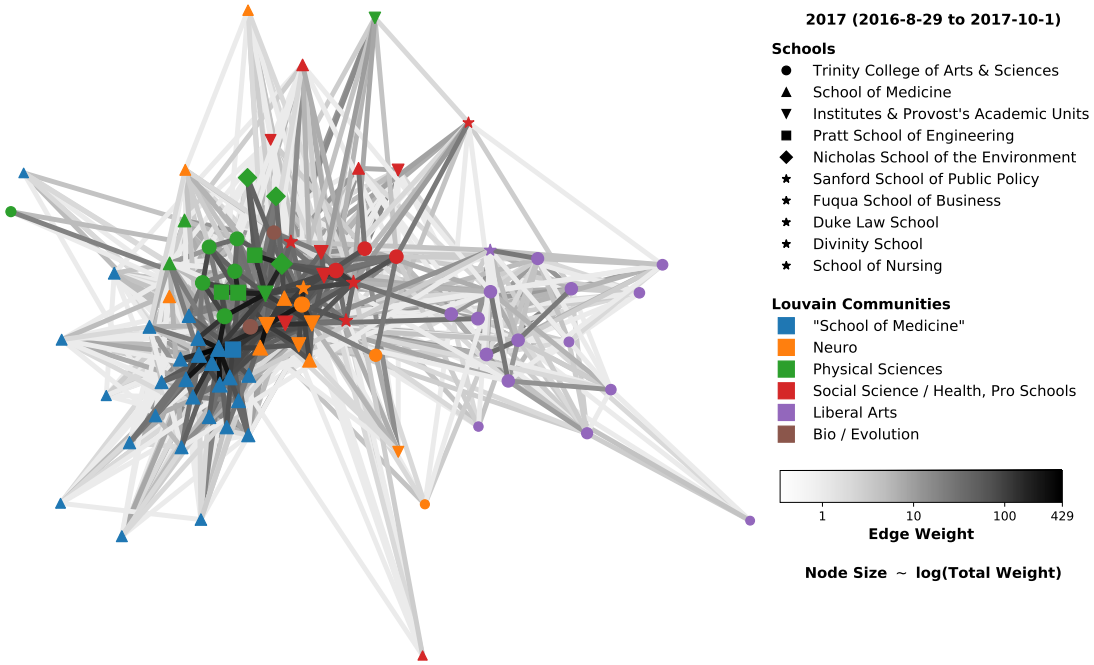

FIG. 11. Academic organizations graph for 2017.

B. LOUVAIN COMMUNITY MEMBERS

Community 0: "School of Medicine"	School
Biomedical Engineering	Pratt School of Engineering
Anesthesiology	School of Medicine
Biochemistry	School of Medicine
Cardiology	School of Medicine
Cell Biology	School of Medicine
Dermatology	School of Medicine
Duke Cancer Institute	School of Medicine
Duke Center for Human Genome Variation	School of Medicine
Duke Molecular Physiology Institute	School of Medicine
Endocrinology, Metabolism, and Nutrition	School of Medicine
Gastroenterology	School of Medicine
Geriatrics	School of Medicine
Hematology	School of Medicine
Human Vaccine Institute	School of Medicine
Immunology	School of Medicine
Infectious Diseases	School of Medicine
Molecular Genetics and Microbiology	School of Medicine
Obstetrics and Gynecology	School of Medicine
Oncology	School of Medicine
Ophthalmology	School of Medicine
Orthopaedics	School of Medicine
Pathology	School of Medicine
Pediatrics	School of Medicine
Pharmacology & Cancer Biology	School of Medicine
Pulmonary, Allergy, and Critical Care Medicine	School of Medicine
Radiology	School of Medicine
Regeneration Next Initiative	School of Medicine
Surgery	School of Medicine

FIG. 12. Members of the “School of Medicine” Louvain community.

Community 1: Neuro	School
Center for Cognitive Neuroscience	Institutes & Provost's Academic Units
Duke Institute for Brain Sciences	Institutes & Provost's Academic Units
Duke Science & Society	Institutes & Provost's Academic Units
Kenan Institute for Ethics	Institutes & Provost's Academic Units
Center for Child and Family Policy	Sanford School of Public Policy
Duke-UNC Center for Brain Imaging and Analysis	School of Medicine
Nephrology	School of Medicine
Neurobiology	School of Medicine
Neurology	School of Medicine
Population Health Sciences	School of Medicine
Psychiatry & Behavioral Sciences	School of Medicine
Linguistics	Trinity College of Arts & Sciences
Philosophy	Trinity College of Arts & Sciences
Psychology and Neuroscience	Trinity College of Arts & Sciences

FIG. 13. Members of the Neuro Louvain community.

Community 2: Physical Sciences	School
Center for Latin American and Caribbean Studies	Institutes & Provost's Academic Units
Duke University Energy Initiative	Institutes & Provost's Academic Units
Earth and Ocean Sciences	Nicholas School of the Environment
Environmental Sciences and Policy	Nicholas School of the Environment
Marine Science and Conservation	Nicholas School of the Environment
Civil and Environmental Engineering	Pratt School of Engineering
Electrical and Computer Engineering	Pratt School of Engineering
Mechanical Engineering and Materials Science	Pratt School of Engineering
Biostatistics & Bioinformatics	School of Medicine
Duke Clinical Research Institute	School of Medicine
Chemistry	Trinity College of Arts & Sciences
Computer Science	Trinity College of Arts & Sciences
Education	Trinity College of Arts & Sciences
Mathematics	Trinity College of Arts & Sciences
Physics	Trinity College of Arts & Sciences
Statistical Science	Trinity College of Arts & Sciences

FIG. 14. Members of the Physical Sciences Louvain community.

Community 3: Social Science / Health, Pro Schools	School
Duke Law School	Duke Law School
Fuqua School of Business	Fuqua School of Business
Center for Population Health & Aging	Institutes & Provost's Academic Units
Center on Biobehaviorial Health Disparities Research	Institutes & Provost's Academic Units
Duke Population Research Center	Institutes & Provost's Academic Units
Global Health Institute	Institutes & Provost's Academic Units
Social Science Research Institute	Institutes & Provost's Academic Units
Sanford School of Public Policy	Sanford School of Public Policy
Center for the Study of Aging and Human Development	School of Medicine
Community and Family Medicine	School of Medicine
General Internal Medicine	School of Medicine
School of Nursing	School of Nursing
Economics	Trinity College of Arts & Sciences
Political Science	Trinity College of Arts & Sciences
Sociology	Trinity College of Arts & Sciences

FIG. 15. Members of the Social Science / Health, Pro Schools Louvain community.

Community 4: Liberal Arts	School
Divinity School	Divinity School
Asian Pacific Studies Institute	Institutes & Provost's Academic Units
African and African American Studies	Trinity College of Arts & Sciences
Art, Art History & Visual Studies	Trinity College of Arts & Sciences
Asian and Middle Eastern Studies	Trinity College of Arts & Sciences
Classical Studies	Trinity College of Arts & Sciences
Cultural Anthropology	Trinity College of Arts & Sciences
Dance	Trinity College of Arts & Sciences
English	Trinity College of Arts & Sciences
Gender, Sexuality & Feminist Studies	Trinity College of Arts & Sciences
Germanic Languages	Trinity College of Arts & Sciences
History	Trinity College of Arts & Sciences
Literature	Trinity College of Arts & Sciences
Music	Trinity College of Arts & Sciences
Religious Studies	Trinity College of Arts & Sciences
Romance Studies	Trinity College of Arts & Sciences
Slavic and Eurasian Studies	Trinity College of Arts & Sciences
Theater Studies	Trinity College of Arts & Sciences

FIG. 16. Members of the Liberal Arts Louvain community.

Community 5: Bio / Evolution	School
Biology	Trinity College of Arts & Sciences
Evolutionary Anthropology	Trinity College of Arts & Sciences

FIG. 17. Members of the Bio / Evolution Louvain community.

C. INTERDISCIPLINARY FRACTION BY COMMUNITY

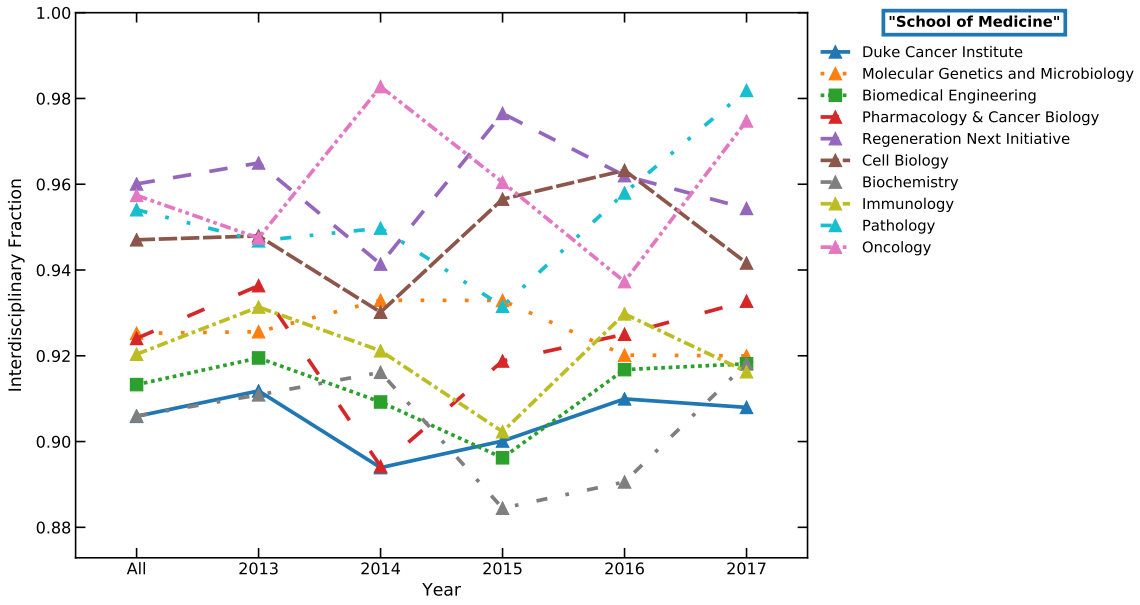

FIG. 18. Interdisciplinary fraction vs year for the "School of Medicine" community.

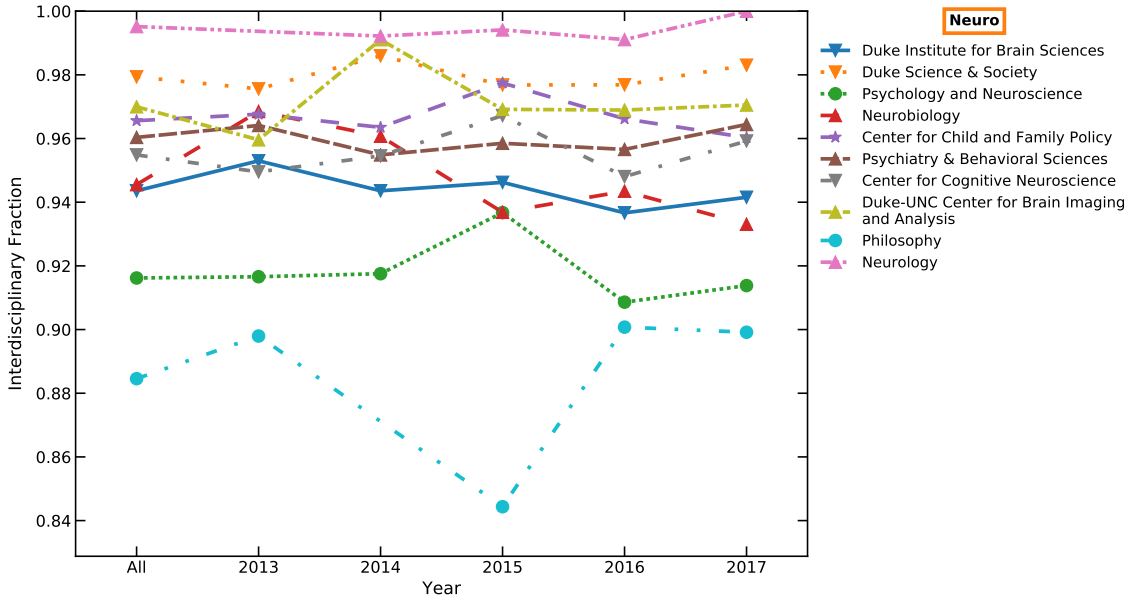

FIG. 19. Interdisciplinary fraction vs year for the Neuro community. Figure 4 reproduced for convenience.

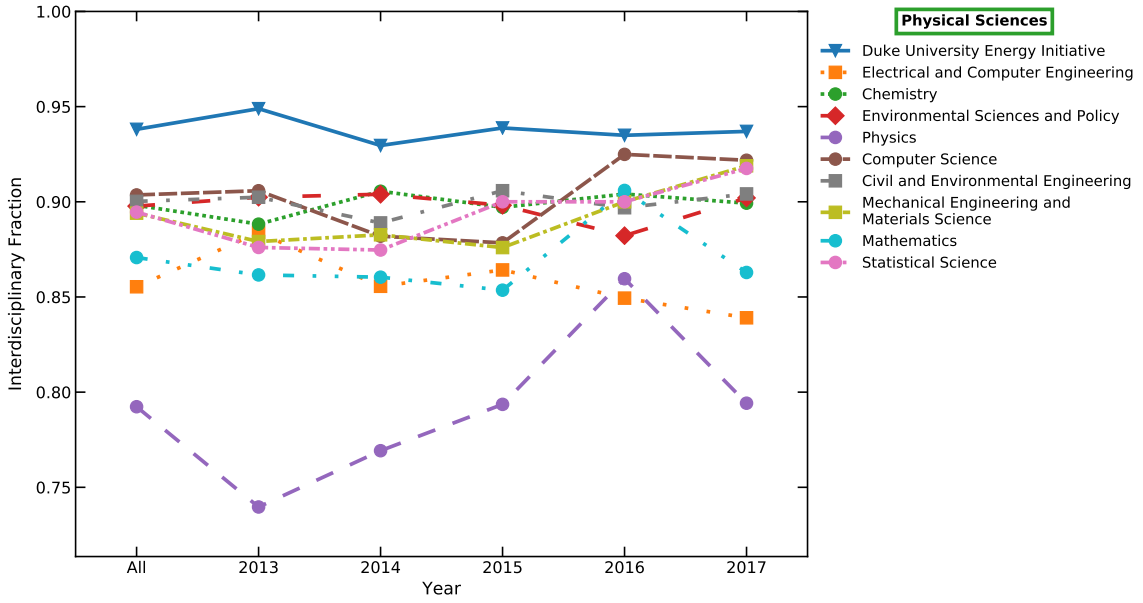

FIG. 20. Interdisciplinary fraction vs year for the Physical Sciences community. Figure 3 reproduced for convenience.

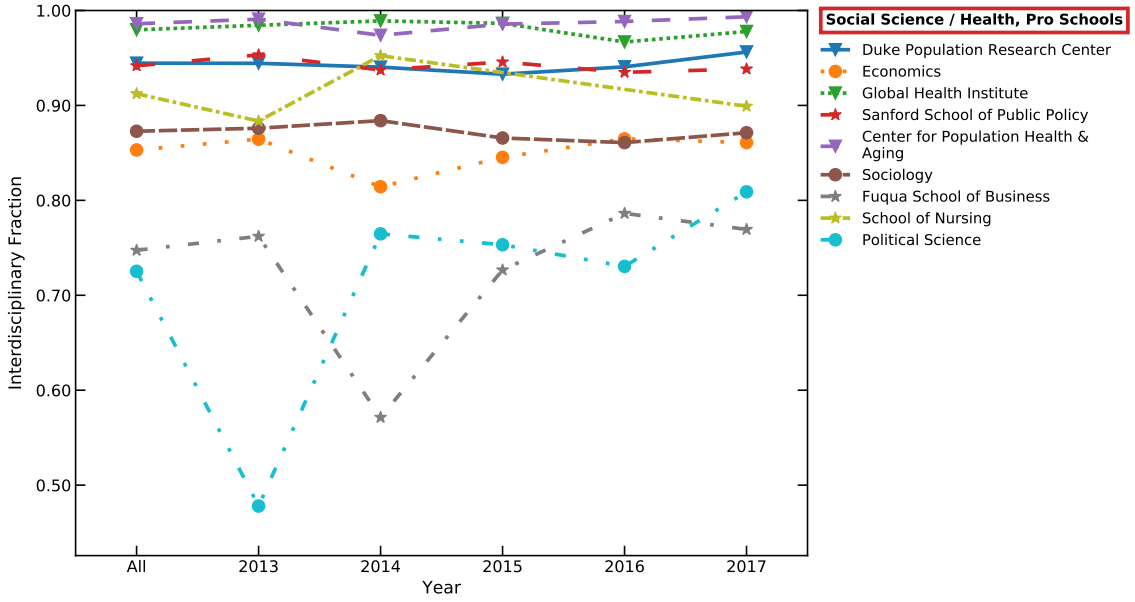

FIG. 21. Interdisciplinary fraction vs year for the Social Science / Health, Pro Schools community.

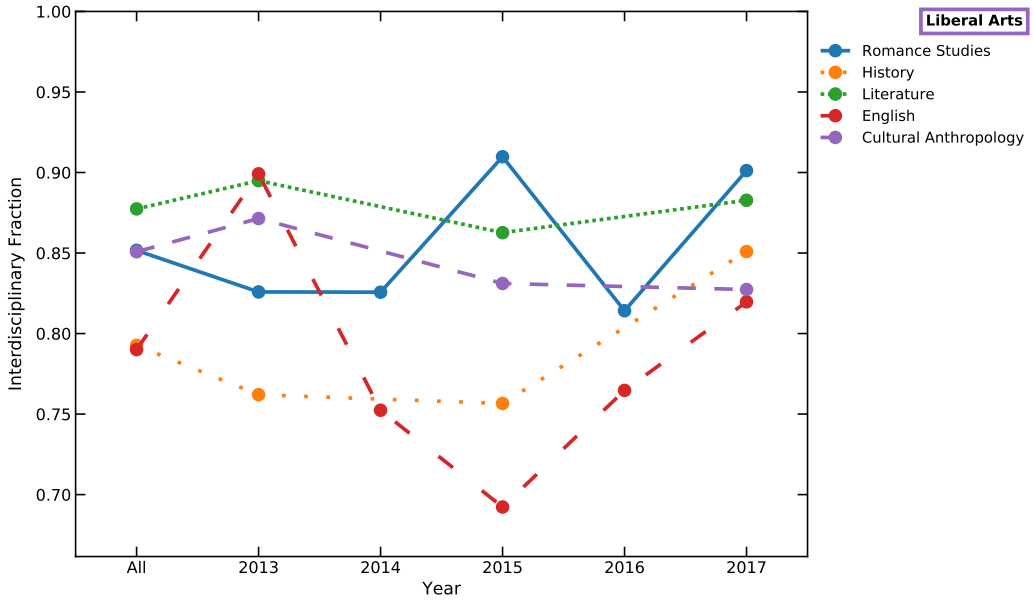

FIG. 22. Interdisciplinary fraction vs year for the Liberal Arts community. Figure 5 reproduced for convenience.

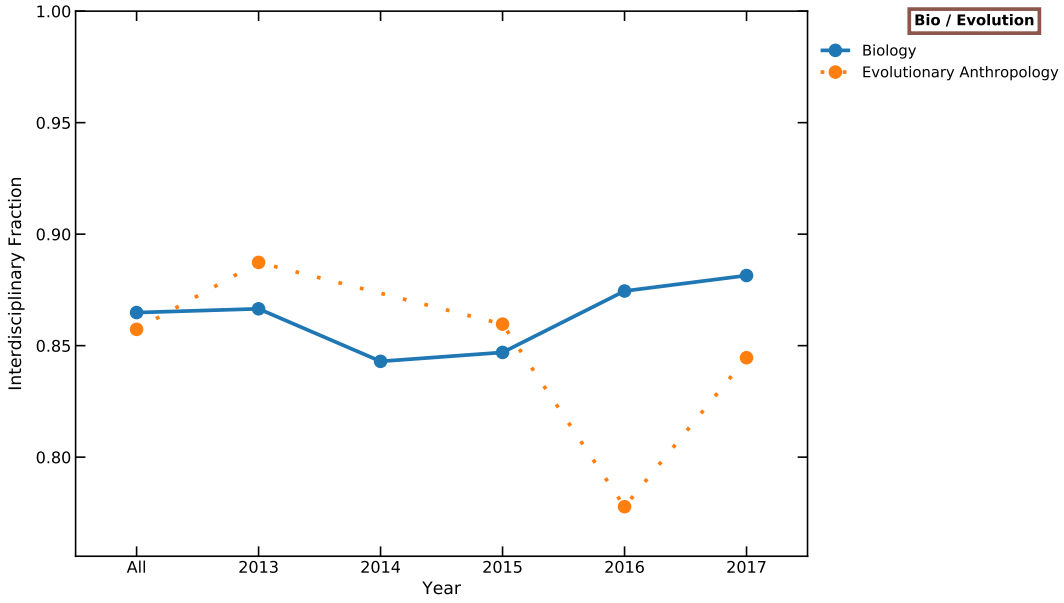

FIG. 23. Interdisciplinary fraction vs year for the Bio / Evolution community.

D. INTERDISCIPLINARY FRACTION BY SCHOOL

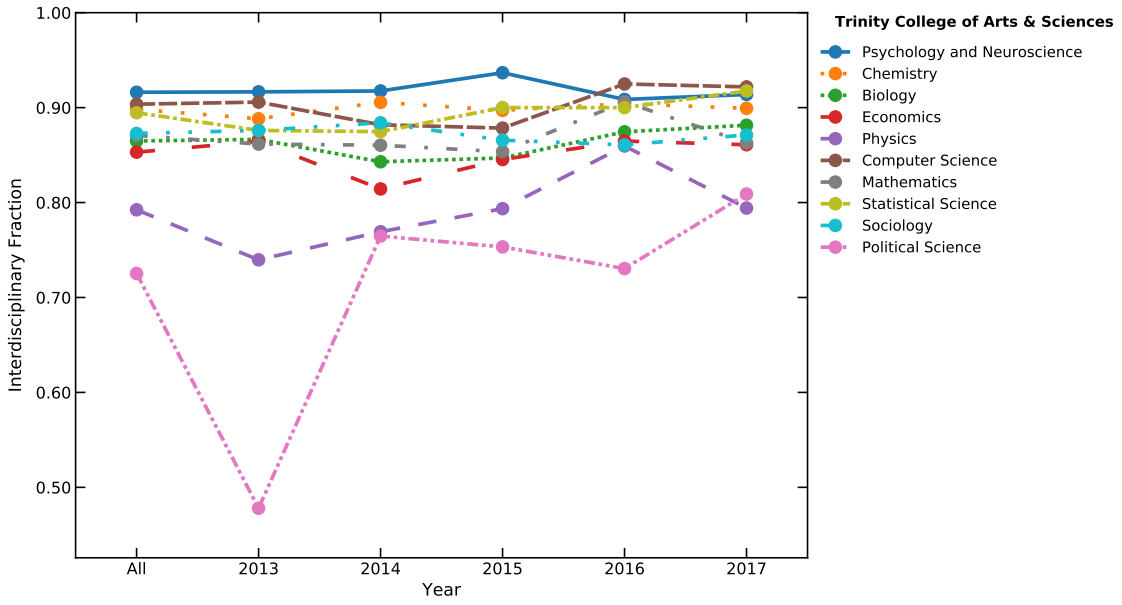

FIG. 24. Interdisciplinary fraction vs year for the Trinity College of Arts & Sciences.

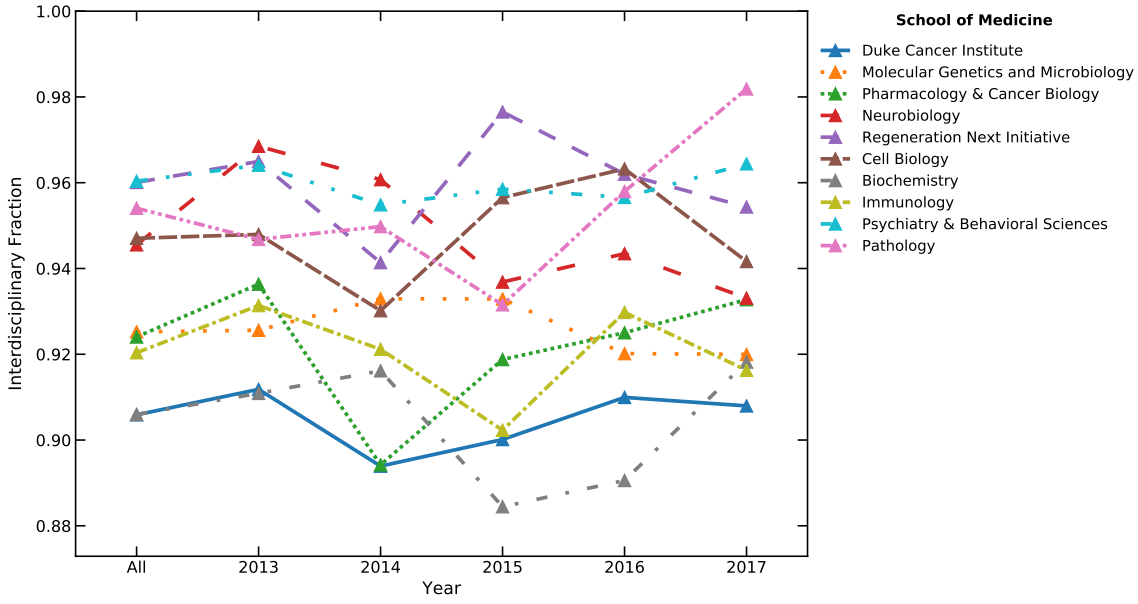

FIG. 25. Interdisciplinary fraction vs year for the School of Medicine.

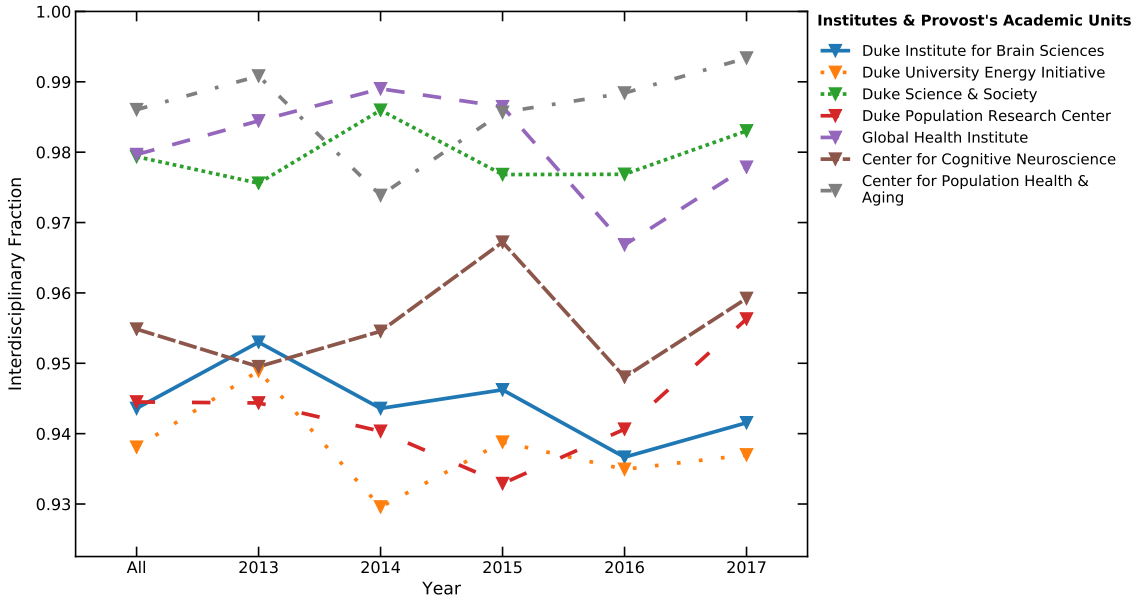

FIG. 26. Interdisciplinary fraction vs year for the Institutes & Provost's Academic Units.

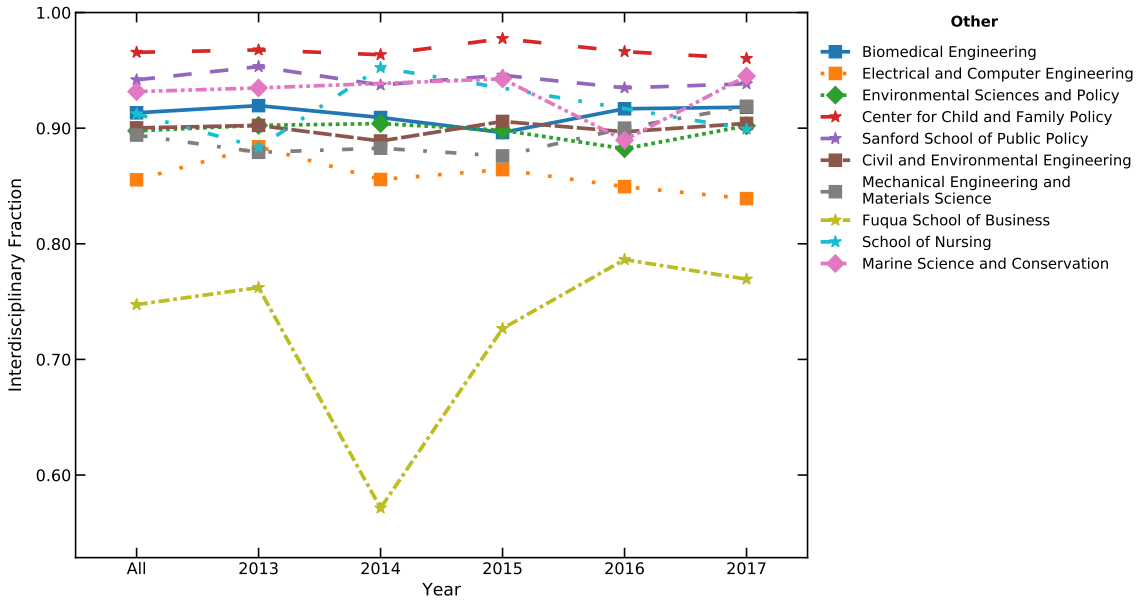

FIG. 27. Interdisciplinary fraction vs year for the remaining schools and units.