Collaborative Research Network Analysis: A Case Study of Harvard's Biomedical Research Community

Florian Hillen, Roxanne Rahnama, Saeyoung Rho, Olamide Oladeji, and Chao Zhang

I. MOTIVATION

The high quality of medical care in our society is built upon the creation of scientific knowledge generated from medical research. Over the past decade, the research questions in this field have become exponentially more complex, labor intensive, and experimental, thereby leading to a greater need for expensive research equipment and interdisciplinary collaborations. While there has been a growth of network literature and research examining both citation and co-author networks across various academic fields, there continue to be important questions that remain to be further investigated, including both community detection and time series-based analysis of networks (Newman [2004]). Consequently, we are motivated by these unexplored questions and an urge to better understand the role that academic networks and greater collaboration within the biomedical research field play in enabling successful research. For the purpose of this project, we are focusing specifically on the research community at Harvard Medical School in our analyses around the following set of core research questions:

- Does stronger collaboration in the biomedical research co-author network translate into greater impact?
- What effects, if any, did the NIH's Clinical and Translational Science Award Program (fully implemented in 2012) have on advancing collaborative research?
- In what ways do the examined collaboration networks' centrality measurements and community structures change pre-2012 and post-2012?
- What is the nature of community structures in the biomedical research field? Is there inter- or intra-field research?
- Do researchers in the biomedical network have a tendency to collaborate in an intra or inter-departmental manner?

II. LITERATURE REVIEW

A. General Collaboration Networks

While collaboration has been a longstanding phenomenon within and between academic institutions, the value of understanding the formative properties and functionalities of networks within the university infrastructure has increasingly risen over the last decade among researchers. In an early study of academics from a networked perspective, (Newman [2004]) discovers properties such as a small world effect in scientific collaboration networks. More recently, (Anderson and Krawczyk [2011]) analyze how isolated or connected various fields of studies are from one another, and measure the strength of cross-departmental interdisciplinary collaborations over time at Stanford, finding improvements for Medicine and Engineering. In a much more recent study, (Claudel et al. [2017]) examine collaborative patterns of faculty at MIT by examining their academic output and organizational structures between 2004 and 2014, finding a power law relationship between proximity and collaboration.

B. Co-Author Networks

A more specific sub-set of research on academic collaboration networks include analyses based on co-author networks. While this area of work continues to be rife with questions to examine, there are a number of papers over the last decade that highlight a set of notable findings. In an early paper on this topic, (Newman [2004]) utilized bibliographic databases in the fields of biology, physics, and mathematics to construct co-author networks and explore their collaboration patterns, as measured by the number of papers written, number of co-authors on papers, the distance between authors in a network, and temporal variations in patterns of collaboration. In particular, the paper emphasizes the presence of a higher number of co-authors in biological sciences, as well as an interesting approximate consistency between the collaboration networks and a linear preferential attachment model. In a similar line of analysis that focused on betweenness centrality, (Goh et al. [2002]) and (Holme et al. [2002]) found that the distribution of betweenness scores in the collaboration network appears to follow a power law and that collaboration networks are highly affected by the removal of nodes with the highest betweenness scores, respectively. More recent publications and research have taken these co-author network analyses even further, applying various methods such as Weighted Mutual Influence Rank (WMIRank) to find so-called rising stars in the co-author networks - or authors with the potential to be come experts in the future (Daud et al. [2017]).

C. Social Dimensions of Academic Networks

Other recent studies have continued to examine additional interesting properties of academic networks, including academic social networking sites, as well as variations in gender effects on these sites, and of developmental (ego-centered) networks of academic staff. In relation to the latter topic, (Jordan [2017]) examines the network structures facilitated by Academia.edu and ResearchGate, as well on Twitter, finding that academic SNS networks were smaller and more highly clustered than their Twitter networks. Thelwall and Kousha [2014] additionally study attributes of scholars on

Academia.edu, finding advantages for female faculty compared with males in terms of profile views across law, history, computer science and philosophy. Moreover, (Barthauer et al. [2016]) use ego-networks of PhD students and postdoctorates at German universities with cohesion (density and degree) and brokerage (effectiveness and constraint) as indicators for access to social capital to study gender effects of developmental networks of male and female academic staff.

While the existing set of aforementioned studies paint a telling story of academic networks and subnetworks, the literature is still relatively sparse, particularly as it relates to medical networks, thereby highlighting the important need for additional research on this dynamic area of study.

III. DATA DESCRIPTION AND PREPROCESSING

The data used for this project is primarily derived from two sources - namely Harvard Catalyst and the Web of Science.

A. *Harvard Catalyst*

Harvard Catalyst is funded by the National Institute of Health (NIH) Clinical and Translational Sciences Awards (CTSA) program and includes a platform for downloading the informational profiles of Harvard Catalyst members, including their first and last name, institution, department, and faculty status. For the purpose of this project, we obtained data on 5024 Principal Investigators of medical research labs at Harvard, using this platform.

B. *Web of Science*

The Web of Science is a web-based scientific citation indexing service that provides comprehensive citation information, and further enables access to in-depth citation reports for specialized sub-fields. To obtain the data needed from the Web of Science database, we wrote Python scripts to access the Web of Science Application Programming Interface (API), which provided XML outputs of publication and citation data for author queries. These XML outputs were then parsed through an XML parser and formatted to obtain a more structured dataset. We faced a couple of challenges related to constraints in the number of queries and results that were allowed by the Web of Science API and eventually resolved these by using the Advanced Search and Citation Report features of the Web of Science front end.

C. *Merged Data*

The final merged set of data includes information about the title, year, journal of publication, author names, affiliation (research area), institution, and citation information from 2003 to 2017. Having organized and combined these files, we built a co-author network, working with a few different adjacency matrix versions of the data: (1) the full (weighted and undirected) network, (2) the full (weighted and undirected) network without isolated nodes - i.e. degree 0, (3) the previous two networks with the inclusion of author labels, and (4) a subsetted dataset that includes all papers that have no fewer than two authors within the Harvard Catalyst Principal Investigator list.

IV. METHODOLOGY

A. *Co-Author Network Summary Statistics*

As briefly discussed in the previous section, we used the information on authors to build a co-author network of the medical research community in Boston and Cambridge, Massachusetts. As a first step, we evaluate a core set of summary statistics on both the full (weighted and undirected) network, as well as the full (weighted and undirected) network without isolated nodes, measuring characteristics such the number of nodes, number of edges, average degree, maximum connected component, transitivity (clustering coefficient), average path length, edge density, diameter, degree distribution, betweenness centrality, and eigenvector centrality using the i-graph package in R statistical programming language.

B. *Collaboration and Scientific Impact: Regression Analysis*

As part of our core analysis, the first research question that we aim explore relates to evaluating the existence and strength of the relationship between the level of collaboration and level of impact for any given PI in this biomedical community. In particular, we consider two measures of collaboration: (1) the average number of collaborators per paper for each author and (2) the degree of the researcher in the network. In terms of measuring scientific impact, we consider three key metrics, namely (1) the Hirsch-Index (H-Index) of each researcher, which is explained below; (2) citations per paper; and (3) the number of papers published. The various regressions additionally include a proxy for age of the researcher, which we roughly measure as the number of years since the first publication since there was no age variable to work with in our dataset. We call this variable active years. We moreover include a control for research field or department, as there may be variations in the level of funding or other various forms of biases between research areas that need to be accounted for.

H-Index Computation: The H-Index (Hirsch-index) is a widely accepted author metric that attempts to quantify productivity and citation impact. Specifically, an author with an H-level of k has published k papers, each of which have been cited at least k times. Following this definition, we computed the H-Index for every principal investigator in the network using only publications that are in the dataset.

Besides the h-index, we also computed the other impact and collaboration measures from our dataset. The regression model utilized in our analysis is specified as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Where Y is the relevant impact measure that is computed based on the publications from 2003 to 2017, X_1 represents either degree centrality or collaborators per paper, and the X_2 and X_3 represent the active years and research area controls, respectively. We are particularly interested in the β_1 coefficient values.

C. Temporal Analysis of Network Collaboration Patterns: Case Study of the CTSA

In addition to assessing the relationship between collaboration and scientific impact through regression analysis with the calculated H-Index, we further examined the potential effect of the NIH Clinical Translational Science Awards Program on advancing collaboration in this biomedical co-author network. This program and our approach is described in greater details in the following subsection.

Case Study: Impact of the Clinical and Translational Science Awards Program: In 2012, the National Institute of Health fully implemented its Clinical and Translational Science Award (CTSA), a type of U.S. federal grant dedicated to supporting the creation of integrated centers and resources for clinical and patient care research. The program is comprised of 60 grantee institutions, including Harvard University, and funding is structured in such a way as to encourage collaboration between researchers from diverse scientific fields and backgrounds. This program provides, on average, about $500,000,000 of funding annually, with the latest funding cycle for 2017 reaching up to $516,124,810 in funding. Given this program's particular relevance to the research networks that we are examining and the fact that it funds Harvard Catalyst, we are interested in conducting a form of pre-CTSA, post-CTSA analysis to analyze the ways in which collaboration patterns and centrality measures change, if at all, in our co-author network before and after the full implementation of the program in 2012. We examine these changes through a combination of methods, including (1) a Welch Two Sample t-test on the average edge density for the pre-period (2009-2012) and post-period (2013-2016); (2) a two-sample Kolmogorov-Smirnov test for comparing the degree distributions for the 5024 PIs in the pre-period vs post-period; and (3) an ultimate comparison of community patterns for the pre- vs. post- periods. Findings from this analysis may point at potential efficacies or inefficacies of this NIH program targeting collaboration in the biomedical research community.

D. Community Detection

Lastly, to answer questions about the nature of community structures in the biomedical research field and assess evidence of stronger versus weaker intra- and/or inter field and department collaboration, we apply a community detection algorithm. The detection of community structure in a network can be achieved in a number of ways. For example, graph partitioning is a simple and effective technique for applying community detection. In addition, a more classical and precise way of community partitioning could be based on advanced community detection algorithms such as modularity maximization, maximum likelihood (Newman [2016]), and stability optimization (Le Martelot and Hankin [2013]). For the purpose of our community detection analysis on this biomedical research network, we utilize Louvain modularity, which is defined below and by optimizing the modularity (Q), we detect the communities in the network.

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

In exploring additional questions related to inter-community and inter-department collaboration tendencies, we use the metric 1-Q or one minus the modularity. The data mapping and communication structures are then visualized using Gelphi.

V. Results

A. Co-Author Network Summary Statistics

As previously discussed, we first constructed the network and evaluated several network measures to gain more insights about network structures in biomedical research settings:

TABLE I
SUMMARY STATISTICS

Data Characteristic	Full Network	No Isolated Nodes
Number of Nodes	5024	4768
Number of Edges	61534	61534
Average Degree	24.496	25.8112
Max Connected Component	4764	4764
Clustering Coefficient	0.1849247	0.1849247
Average Path Length	3.247498	3.247498
Edge Density	0.004876771	0.005414567
Diameter	18	18

There are a number of important values in the summary statistics to comment on in the context of our co-author network analysis. In particular, the average degree of approximately 25 implies that each author is, on average, connected with an edge to 25 other authors in the biomedical research network. Moreover, the average path length implies that there are about three steps along the shortest paths for all possible pairs of authors and the diameter of 18 implies 18 steps in the shortest path between the two most distant author nodes. Lastly, the clustering coefficient of 18.5 percent is interestingly higher than the clustering coefficient of seven percent for biology-related co-author networks analyzed in (Newman [2004]). Our clustering coefficient is considerably similar to the clustering coefficient of 17.44 percent found in a co-author network analysis of Stanford University's Academic Network (Anderson and Krawczyk [2011]).

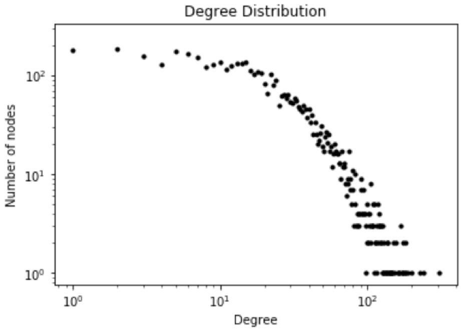


Fig. 1. Degree distribution of the full weighted and undirected network, including isolated nodes, on a log-log scale. Does not appear to follow power law distribution, in line with the finding in (Newman [2001]) of small world properties in scientific collaboration networks. The degree distribution of the full network without isolated nodes follows a nearly identical pattern, so for the sake of space limits, we will only show it for the full network.

In addition to the set of summary statistics, the degree distribution plot above is shown on a log-log scale to demonstrate the finding that the co-author network does not appear to follow a power law distribution, but rather more of a small world distribution, in line with the findings in (Newman [2001]) regarding the distribution properties of scientific collaboration network.

B. Collaboration and Scientific Impact: Regression Analysis

As discussed in the Methodology section, we conduct a number of regressions to measure the effect of collaboration (independent variable) on scientific impact (dependent variable), controlling for the research area of the PI and the age through a proxy measurement of active years. First, we use the total number of papers published in the last 15 years as a dependent variable for assessing scientific impact (see Table II). In addition, when controlling for age (active years) and research area (Column 2) there is a very significant relationship between the degree centrality and number of publications. In particular, an increase of one degree is associated with 1.895 additional publications. Ex-ante, we hypothesized that a higher number of average collaborators might increase productivity, as the work can be better divided among authors. However, the results in the third column of our table does not support this hypothesis, showing no significant relationship between the output and input variables.

Second, we choose citations per paper as another dependent variable measuring research quality (see Table III). Here we can see that one degree in centrality is related to 0.516 more citations per paper, at a high statistical significance. Even more intuitive is the result that an additional average collaborator on a publication is associated with 1.145 more citations per paper. While we do not know the causal mechanism behind this effect, we hypothesize that having additional collaborators may result in better and more highly cited research. Alternatively, a publication may have more exposure due to more collaborators, which could also potentially cause a higher number of citations.

Third, we were interested in analyzing the relationship between the collaboration of the PI and his/her personal impact measurement, the H-index (see Table IV). While the degree centrality has a very significant relation to the H-index (Column 1), the effect size decreases with the inclusion of the controls (Column 2). Intuitively, this makes sense since the longer the PI has been active, the more time he/she had to publish papers and others had to cite his/her work, resulting in a higher H-index. The relationship with avg. collaborators is also significant yet has a small effect size of 0.105 (Column 3). Comparing this to the results for the avg. collaborator variable in Tables II and III, we can hypothesize that this relationship is probably not driven by an increase of paper publications, but rather an increase of citations.

Overall, it seems that scientific impact is significantly affected by the various collaboration properties of the PI and that higher collaboration may translate into greater scientific impact.

TABLE II
THE RELATIONSHIP BETWEEN PAPERS PUBLISHED AND COLLABORATION

	Dependent variable:		
	Number of Paper published		
	(1)	(2)	(3)
Degree Centrality	2.490***	1.895***	
	(0.107)	(0.102)	
Avg. Collaborators/paper			0.122
			(0.095)
Active Years		9.458***	12.860***
		(0.623)	(0.618)
Constant	−2.731	−56.537	−73.768
	(3.869)	(182.336)	(188.600)
Research Area Control	No	Yes	Yes
Observations	5,024	5,024	5,024
Adjusted R^2	0.098	0.266	0.214
Note:			*p<0.1; **p<0.05; ***p<0.01

TABLE III
THE RELATIONSHIP BETWEEN CITATIONS PER PAPER AND COLLABORATION

	Dependent variable:		
	Citations per Paper		
	(1)	(2)	(3)
Degree Centrality	0.525***	0.516***	
	(0.035)	(0.036)	
Avg. Collaborators/paper			1.145***
			(0.028)
Active Years		0.359	1.058***
		(0.253)	(0.215)
Constant	19.961***	22.065	13.177
	(1.322)	(63.784)	(55.690)
Research Area Control	No	Yes	Yes
Observations	4,447	4,447	4,447
Adjusted R^2	0.049	0.045	0.272
Note:			*p<0.1; **p<0.05; ***p<0.01

TABLE IV

THE RELATIONSHIP BETWEEN H-INDEX AND COLLABORATION

	Dependent variable:		
		H-Index	
	(1)	(2)	(3)
Degree Centrality	0.370***	0.282***	
	(0.009)	(0.008)	
Avg. Collaborators/paper			0.105***
			(0.008)
Active Years		1.783***	2.239***
		(0.050)	(0.053)
Constant	4.334***	−8.261	−10.957
	(0.315)	(14.680)	(16.085)
Research Area Control	No	Yes	Yes
Observations	5,024	5,024	5,024
Adjusted R^2	0.265	0.414	0.297
Note:			*p<0.1; **p<0.05; ***p<0.01

C. Temporal Analysis of Network Collaboration Patterns: A Case Study of the CTSA

We approach our temporal analysis of network collaboration patterns from three angles, as discussed under the Methodology section: (1) a Welch Two Sample t-test on the average edge density for the pre-period (2009-12) and post-period (2013-16); (2) a two-sample Kolmogorov-Smirnov test for comparing the degree distribution for the 5024 PIs in the pre vs post-period; (3) an exploratory analysis of the centrality measures of the top five author nodes in the two periods before and after the CTSA program implementation, and (4) an ultimate comparison of community patterns in the two periods. The results are presented below as follows in their respective sub-sections.

Welch T-Test on Average Edge Density: The following image below presents the edge density for each publication year from 2009-2012 and then 2013-2016, with the time separation marking the time in which the NIH Clinical Translational Science Award Program went into full implementation. We observe a significant jump in the average edge density, implying that there are more edges and potentially a greater number of collaborations after 2012.

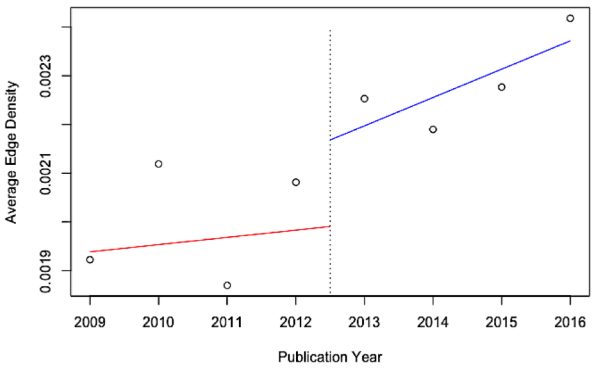


Fig. 2. Edge Density

In our Welch Two-sample t-test on the difference between the average edge density from 2009-12 and 2013-16, we find **t-statistic of -3.7074** and a **p-value of 0.01091**, which implies a strongly statistically significant difference in the average edge density in the pre vs -post CTSA launch periods.

Kolmogorov-Smirnov Two-Sample Test on Degree Distributions: In the next part of our temporal analysis, we compare the degree centrality of each of the 5024 PI's in a panel-analysis format, testing for differences in the degree distribution in the pre-CTSA full launch period (2009-12) and post-CTSA (2013-16) full launch period. In our comparison of the pre- and post- period degree centralities, we find a highly statistically significant **p-value of 2.2e-16**, implying a strong increase in the number connections that the PIs in our dataset hold post 2012.

Exploratory Analysis of Top 5 Author Nodes, Pre- and Post- CTSA: Next, we were interested in investigating the behavior of the top-five nodes in the network in the pre- and post-2012 periods to observe if there were any remarkable differences in their respective propensity to collaborate and connect with more researchers, upon the inauguration of a funding program catered to fostering collaboration in scientific research. The betweenness, closeness, and eigenvector centralities of these nodes are presented in the following pre- and post-2012 period tables:

TABLE V

TOP 5 PIS: PRE-PERIOD (2009-2012)

Name	Btwn	Close	Eigen	Total
Fuch, Charles	1	3	1	5
Zurakowski, David	4	0	0	4
Hunter, David	2	2	0	4
Rodig, Scott	3	1	0	4
Giovannucci, Edward	0	2	1	3

TABLE VI

TOP 5 PIS: POST-PERIOD (2013-2016)

Name	Btwn	Close	Eigen	Total
Getz, Gad	2	4	2	8
Garraway, Levi	0	3	2	5
Zurakowski, David	4	0	0	4
Fuchs, Charles	1	2	1	4
Hu, Frank	1	2	1	4

A key and noticeable takeaway from this analysis shows us that David Zurakowski and Charles Fuch play consistently important roles in the network throughout the entire time period of analysis, with the former researcher holding a persistently high betweenness centrality and the latter a high closeness. In particular, this finding might suggest that Dr. Zurakowski's removal from the biomedical research network would cause a collapse in collaboration patterns. Through some additional research, we found that David Zurakowski is an Associate Professor of Anesthesia and is based at Boston Children's Hospital, holds 895 publications, and is working currently on Biomarkers Predictive outcome algorithms, while Charles Fuch is a Professor of Medical Oncology and Director of the Yale Cancer Center and Physician-in-Chief at the Smilow Cancer Hospital. Moreover, Gad Getz and Levi Garraway appear to be important rising actors in the network after 2012.

We took additional steps to examine the NIH funding database and found that Dr. Getz was found to have received $941,937 in 2017 from the NIH for a project titled Global Infrastructure for Collaborative High - Throughput Cancer Genomics Analysis. Given the focus of this project it seems highly likely that it was funded under the CTSA, which could indicate that the policy had an effect on the research projects of specific researchers. However, it should be noted that it was not explicitly stated in the database whether this NIH funding was under CTSA or not, as we could not access specific information on CTSA recipients.

Community Patterns, Pre and Post 2012: In this ultimate section, which precedes a more in-depth community detection analysis, we specifically analyze the structure and patterns of communities from 2009-2012 and 2013-2016. In a sense, this provides a qualitative indication of higher collaboration between authors in the network as the number of members in one community has increased considerably in the period after the full implementation of the Clinical and Translational Science Award Program of the NIH.

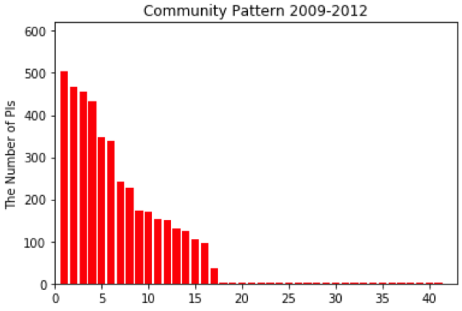


Fig. 3. Community Detection

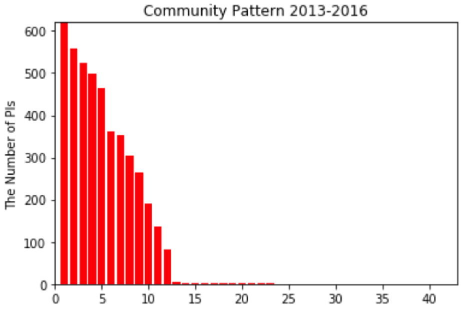


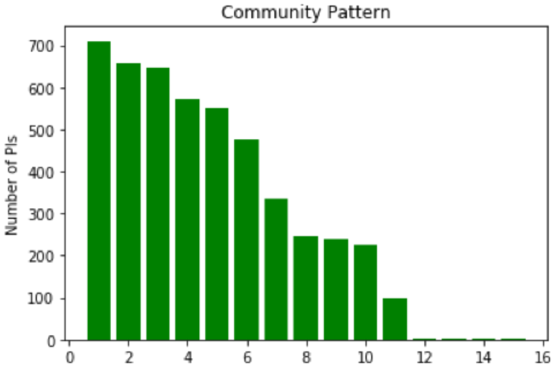
Fig. 4. Community Detection

An interesting finding from observing these two figures is that while the number of communities (x-axis) shrinks after 2012, each community, however, increases in size, as measured by the y-axis (number of PIs).

In sum, we can conclude from our multiple analyses that the CTSA program of the NIH appears to have been significantly successful at promoting collaboration among researchers in the biomedical research community.

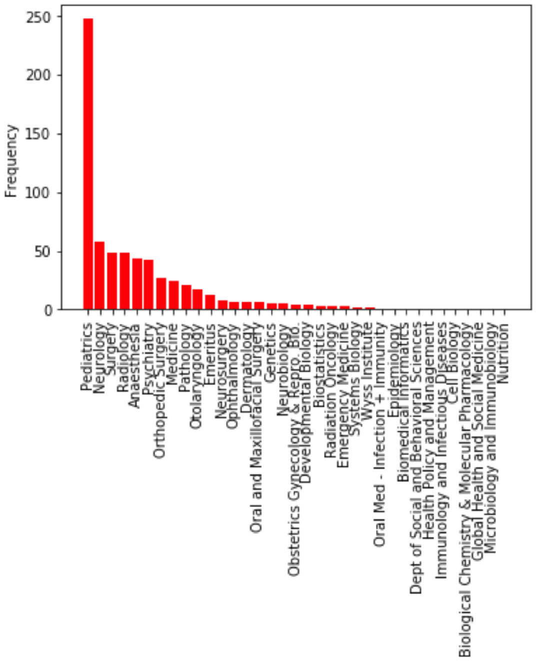
D. Community Detection

In the ultimate section of our analysis, we apply a Louvain modularity community detection algorithm to better understand the nature of community structures in our biomedical research field as well as the patterns of intra and/or inter-community and departmental collaboration. The following table provides an initial high-level overview of the structure of communities in the full biomedical research network with all isolated nodes removed (i.e. on 4768 nodes, as compared with the full network of 5024 nodes):

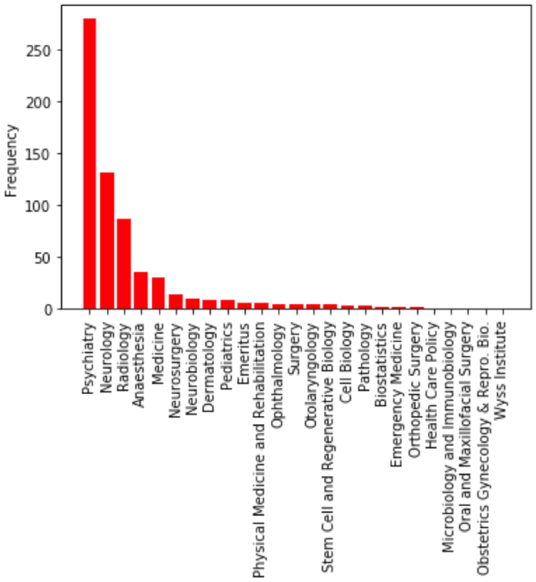


We find a total of 15 communities. On the x-axis, the number represents the community index ranging from 0 to 15. On the y-axis, the size ranges from 2 to 710. The largest community is comprised of approximately 700 researchers. In summary, there are a few communities, among which most of them tend to have a medium to large number of researchers associated and several are very small.

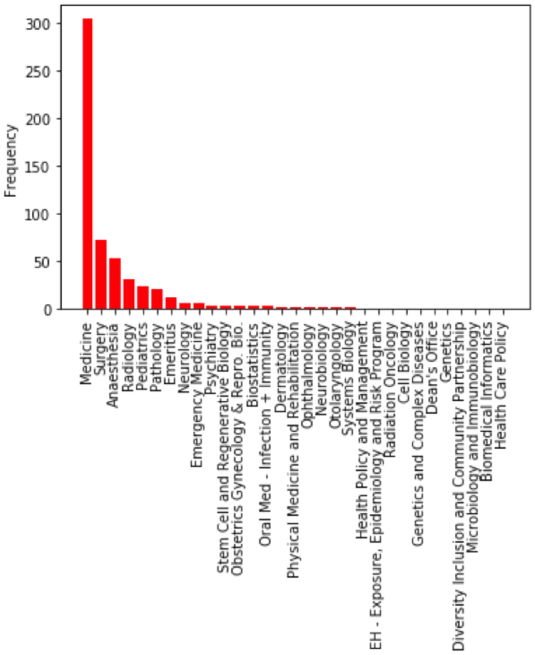
In the next step of the analysis, we dig deeper into community detection, with a specific focus on examining communities through the lens of research fields. In particular, we look at histograms of core research topics examined for the second largest community (which is made up of 656 members), third largest community (which is made up of 649 members), and fourth largest community (which is made up 571 members). The first largest community is dominated by the field of general medicine, thus, considering space limits, we present the results for communities 2 to 4, given that they provide more insights into specialized fields. The figures are displayed and analyzed as follows:



We can see in the figure above that pediatrics is the leading research topic in the second largest community.

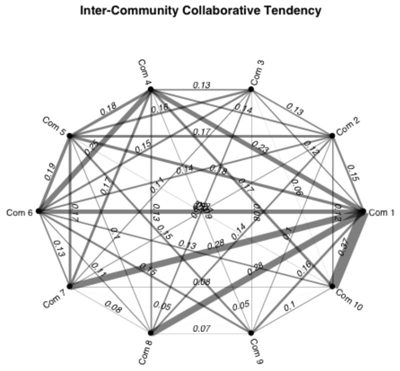


We can further see in the figure above that psychiatry is the leading research topic in the third largest community.

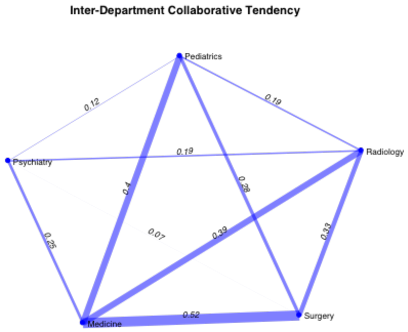


We can see in the last histogram that surgery is the leading research topic in the fourth largest community, after medicine - which is a bit too general of a research field to comment upon on its own. These initial community detection analyses provide us insights about the departments and fields that each of the largest communities in the biomedical network are a part of, helping us to better understand community structures at a high level and make some interpretations about the departments and sub-fields that are more likely to form communities and collaborate with each other.

In our final set of analyses, we examine collaborative tendencies both on an inter-community and inter-department level. In particular, the following figures and analysis examine our research question: do researchers in the biomedical network have a tendency to collaborate in an intra or inter-departmental manner? The following two figures provide a visualization of this topic. In particular, the first inter-community collaborative tendency figure illustrates the top ten communities, their collaboration tendencies with one another, and the strength of their collaborations, as indicated by the thickness of the edges. The numbers along each edge are representative of 1-Q (the modularity) or homophily.



The second figure below, namely the inter-departmental collaborative tendency figure, further extends this analysis to an examination of the top five departments in the PI list based on the number of publications, with thicker edges indicating a higher value of homophily.



Ultimately, community detection is a valuable analytical exercise to carry out, particularly on large-scale networks, due to its strengths in helping to better formulate and describe network structures at a higher level and identify relationships between communities. Specifically in the context of our analysis of the biomedical research network for Harvard University, there are multiple implications that are worth noting with regards to benefits of analyzing communities. For example, by learning community structures, one can potentially forecast and predict how the network will grow over time, which can further provide insights to policymakers and funders on which specific areas of collaborate to target and encourage. Moreover, the analyses on the inter-community and inter-department collaboration tendencies provide insights to academic actors about current interdisciplinary research trends and could inform the ways in which PIs might position themselves in the future in the network to enable greater productivity and collaboration.

VI. CONCLUSION

Through our co-author analysis of the biomedical research network of Harvard University, we arrive at a number of illustrative results about the structure, growth, and impact of collaboration between researchers over the last decade. In particular, our three-pronged methodological approach paves the way to three complementary sets of finding. First, the regression analysis assessing the effects of collaboration (as measured by degree centrality and average collaborations/paper) on impact (as measured by number of papers published, citations per paper, and the H-index in our study time frame) reveal a highly statistically significant effect of collaboration on the success and impact of a researcher across the measures. In the second temporal case study analysis of the effects of the Clinical and Translational Science Award Program, we find that the CTSA seems to have had an impact on collaboration patterns by consolidating communities, and moreover resulted in a strong increase in edge density and degree distribution measures after its full implementation in 2012. We ultimately concluded our analysis of collaboration patterns in the biomedical research community at Harvard by delving deeper into understanding the structures and patterns of collaboration through the implementation of a community detection algorithm using Louvain modularity and measures of homophily. We find that the biomedical research network possesses a strong community structure, in which closely related departments collaborate mostly within communities.

To further increase the robustness and depth of our analysis, in future works we would be interested in developing new metrics that combine collaboration measures with additional impact measures, such as promotions or prizes for researchers, to test the ways in which collaboration may affect success beyond citation-based metrics. Moreover, we would utilize the fraction of collaborative papers out of the total publications as weights on the edges to improve the robustness of our analysis through some level of measurement standardization. Ultimately, we would be interested to extend our analyses to research institutions beyond Harvard, in particular by obtaining hub and authority values for other leading institutions (including Harvard) to analyze what research centres may hold the most prestige in terms of co-authorship and other collaboration-based metrics.

References

Mark EJ Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, 101(suppl 1):5200–5205, 2004.

Ashton Anderson and Stefan Krawczyk. Analyzing stanfords academic network.

Matthew Claudel, Emanuele Massaro, Paolo Santi, Fiona Murray, and Carlo Ratti. An exploration of collaborative scientific production at mit through spatial organization and institutional affiliation. *PloS one*, 12(6):e0179334, 2017.

Kwang-Il Goh, Eulsik Oh, Hawoong Jeong, Byungnam Kahng, and Doochul Kim. Classification of scale-free networks. *Proceedings of the National Academy of Sciences*, 99(20):12583–12588, 2002.

Petter Holme, Beom Jun Kim, Chang No Yoon, and Seung Kee Han. Attack vulnerability of complex networks. *Physical review E*, 65(5):056109, 2002.

Ali Daud, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, Zahid Rafique, Tehmina Amjad, Hussain Dawood, and Khaled H Alyoubi. Finding rising stars in co-author networks via weighted mutual influence. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 33–41. International World Wide Web Conferences Steering Committee, 2017.

Katy Jordan. *Understanding the structure and role of academics' ego-networks on social networking sites*. PhD thesis, The Open University, 2017.

Mike Thelwall and Kayvan Kousha. Academia. edu: social network or academic network? *Journal of the Association for Information Science and Technology*, 65(4):721–731, 2014.

Luisa Barthauer, Daniel Spurk, and Simone Kauffeld. Womens social capital in academia: A personal network analysis. *International Review of Social Research*, 6(4):195–205, 2016.

MEJ Newman. Community detection in networks: Modularity optimization and maximum likelihood are equivalent. *arXiv preprint arXiv:1606.02319*, 2016.

Erwan Le Martelot and Chris Hankin. Multi-scale community detection using stability optimisation. *International Journal of Web Based Communities*, 9(3):323–348, 2013.

Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.