Exploring collaboration patterns of male and female scholars in Physics

Sousan Homaeipour
A Research Paper Submitted to the
Institute for Web Science and Technologies
University of Koblenz-Landau, Germany
shomaeipour@uni-koblenz.de
Supervisor: Dr. Fariba Karimi

January 4, 2019

Abstract

We explored collaboration patterns of scholars who published in American Physical Society (APS) journals between 1893 and 2016. Our analysis showed male and female scientists differ in their collaborative behaviors (degree centrality, K-core, eigenvector centrality, efficiency and clustering coefficient). Women have lower degree centrality and K-core than men and higher clustering coefficient, efficiency and eigenvector centrality than men. In addition, female scientists have higher tendency to associate with same-gender scholars (Gender homophily). Women have higher Homophily than men in this network. Furthermore, we analyzed correlation between collaborative behaviors of scholars and their research performance and found that successful male and female scientists have the same collaboration patterns: they tend to collaborated with more distinct scholars and seek innovations as brokers.

1 Introduction

In this research, we investigate gender differences in collaborative behavior of scientists who published in American Physical Society (APS) journals between 1893 and 2016. In the APS co-authorship or collaboration network, each node represents an author and each edge indicates that two authors have published at least one paper together. Previous research on scientific collaboration showed that men and women have different collaborative behaviors across their career, such as differences in the number of collaborations, position in the network and degree of homophilic behavior [1, 2]. In this paper, we focus on differences in five social network analysis (SNA) measures and homophilic behavior of men and women scientists. In addition, we investigate if successful men and women have the same collaborative behavior by analyzing the correlation between measures of SNA and scholars research performance.

Abbasi et al. [2] explored the collaboration (co-authorship) networks of scholars in information systems. They used measures from social network analysis (SNA) (i.e., normalized degree centrality, normalized closeness centrality, normalized betweenness centrality, normalized eigenvector centrality, average ties strength, and efficiency) for examining the effect of social networks on the performance of scholars. They suggested that research performance of scholars (g-index) is positively correlated with four SNA measures except for the normalized betweenness centrality and the normalized closeness centrality measures. Furthermore, they showed that only normalized degree centrality and efficiency have positive significant influence on the g-index and normalized eigenvector centrality has a negative significant influence on the g-index. Jadidi et al. [1] investigated gender-specific differences in collaboration patterns of computer scientists over 47 years. They explored how these patterns impact scientific success. Their results showed that successful male and female scientists have the same collaboration patterns: they tend to collaborate with more colleagues than other scientists and seek innovations as brokers. They suggested women are on average less likely to adapt the collaboration patterns that are related with success. Furthermore, they found that women have stronger gender homophily.

The structure of this paper is as follows. In chapter 2, we describe datasets and data collection methods. In chapter 3, we present how we inferred gender from scientists' name. In chapter 4, we discuss descriptive statistics of the APS collaboration network. In chapter 5, we investigate the collaboration patterns of male and female scholars. In chapter 6, we discuss correlation between SNA measures and research performance of scholars. In chapter 7, we conclude the research and discuss the possible future work.

2 Data set

The American Physical Society (APS) data set contains bibliographic information on all the articles published by the American Physical Society between 1893 and 2016 [3]. This data set includes 591168 articles consists of metadata such as DOI, journal, volume, issue, title, date and authors [3]. The American Physical Society is the second largest organization of physicists and a member of the society of the American Institute of Physics. The

Physical Review Journal is a collection of leading peer-reviewed research journals such as Physical Review Letters, Physical Review X, Physical Review A, Physical Review B, Physical Review C, Physical Review D and Physical Review E.

3 Inferring Gender From Names

To infer the gender of authors, we applied the method proposed by Karimi et al. [4] that combines the result of name-based (Genderize [1]) and image-based (Faceplusplus [2]) gender detection services. This method has a high accuracy for most countries compared to other name-based methods. However, this approach performs poorly for Chinese and Korean names (table 1). Therefore, Chinese and Korean names are excluded from our gender-specific analyses. To detect Chinese and Korean names, we use a dataset compiled for this purpose by Jadidi et al. [1]. They compiled unique Chinese names from the China Biographical Database Project (CBDB) [3] and compiled Korean names from Wikipedia. They extracted the page titles of all the back links to Korean names [4]. This Asian name detector classifies 88 out of 100 randomly selected scientists correctly. In addition to Chinese and Korean names, we don't include authors with only first initials since we cannot infer their genders. Those authors for which we cannot detect their genders are not considered for our gender-specific analyses, however, included in all other analyses.

	# instances	SSA	IPUMS	Sexmachine	Genderize	Face++	Mixed1	Mixed2
United States	419	0.82	0.76	0.84	0.83	0.91	0.91	0.90
China	113	0.20	0.11	0.67	0.28	0.65	0.50	0.56
United Kingdom	96	0.94	0.92	0.92	0.94	0.81	0.98	0.94
Germany	82	0.87	0.88	0.96	0.94	0.87	0.96	0.93
Italy	75	0.93	0.92	0.94	0.98	0.79	0.99	1
Canada	60	0.87	0.77	0.86	0.91	0.90	0.96	0.93
France	58	0.93	0.92	0.80	0.96	0.81	0.97	1
Japan	56	0.79	0.70	1	0.90	0.62	0.91	0.94
Brazil	44	0.29	0.29	0.15	0.44	0.81	0.90	0.93
Spain	39	0.96	0.92	0.92	1	0.92	1	1
Australia	31	0.89	0.89	0.90	0.86	0.86	0.94	0.93
India	29	0.67	0.17	0.71	0.78	0.83	0.83	0.93
South Korea	27	0.04	0.00	0.58	0.11	0.74	0.37	0.66
Switzerland	25	0.78	0.70	0.56	0.83	0.88	0.90	0.92
Turkey	21	0.43	0.14	0.79	0.81	0.86	1	1

Table 1: Accuracy of gender detection methods. For most countries mixed approaches perform best. [4].

[1]https://genderize.io/.
[2]https://www.faceplusplus.com/.
[3]http://projects.iq.harvard.edu/cbdb/home/.
[4]https://en.wikipedia.org/wiki/Korean _name/.

In the mixed approach proposed by Karimi et al. [4], Genderize is used first. Then for the remaining unidentified names, the image-based method Face++, is used. Genderize utilizes big datasets of information, from user profiles across major social networks and exposes this data through its API. Face++ is an image-based application with high performance. In order to derive the gender for a specific scientist, we collect the first five Google thumbnails using the full name as search query term and then apply image-recognition on the search results. So, we collect a sample of pictures that depict people who are named like the person we searched for. The advantage of this method is that for first names that are ambiguous or unisex, the combination of first and last name is a better indicator of the gender.

4 Descriptive Statistics of the APS Collaboration Network

APS dataset consists of 413455 scientists. After removing initials, Chinese and Korean names, we used the mixed method to infer gender of the 55% of remained names. We inferred gender of 66% of remained names using the mixed method. 85% of authors were identified as men and 15% as women. The largest connected component (LCC) of the collaboration network, consists of 109962 scientists (85% are men and 15% are women). Figure 1 shows the cumulative number of men and women and how the community of scholars who published in APS journals is growing. Corresponding ratio of men to women over 123 years is depicted in Figure 2 and indicates the community is becoming more gender-balanced. In other words, the gender gap is closing over time because men to women ratio decreases from 10 in 1980 to 5.3 up until 2017.

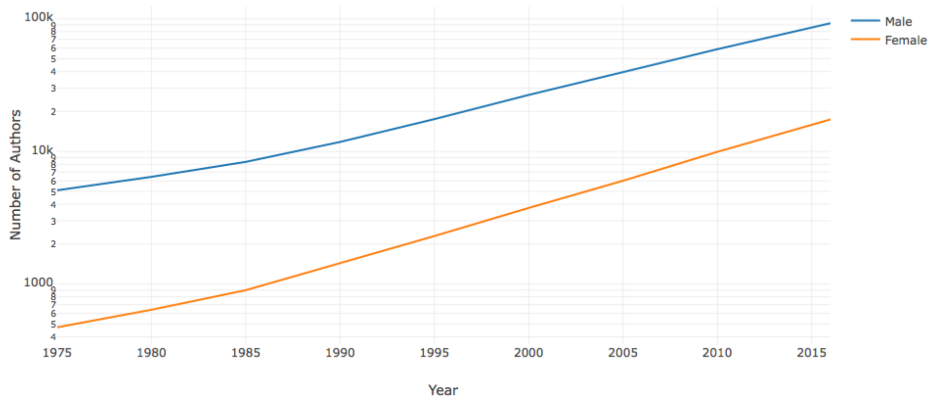

Figure 1: Number of scholars over time.

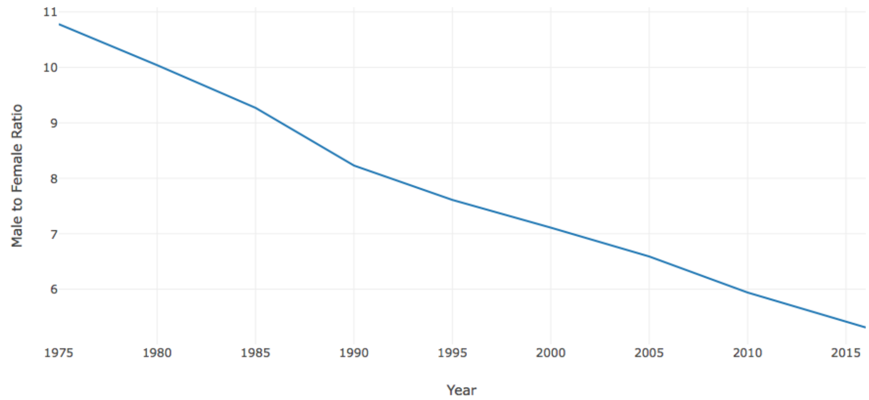

Figure 2: Male to female ratio over time.

5 Collaboration Patterns

5.1 Degree Centrality

Degree centrality for an scholar is the fraction of scholars who co-authored at least an article with him or her. The degree centrality values are normalized by dividing by the maximum degree in the collaboration network (n-1) where n is the number of scholars. In the largest connected component of the APS collaboration network, women have lower degree centrality. Mean of degree centrality of women is 1.65e-05 and mean of men is 1.99e-05. In order to assess the significance of the gap in degree centrality, we used the Mann-Whitney U test. Using this test, we can investigate if the distributions of degree centrality of female and male scholars are equal. We use this test because the distribution of degree centralities is skewed, a small number of researchers have very high degree centralities and a large number of researchers have very low degree centralities.

Mann-Whitney U test (unpaired two-sample Wilcoxon test) is a non-parametric alternative to the unpaired two-sample t-test. It can be used to compare two independent groups of samples when data are not normally distributed. Mann-Whitney U test can be used to determine whether two independent samples were selected from populations having the same distributions. In this test, the alternative hypothesis is that one distribution is statistically greater than the other. However, There are many other ways to formulate the null and alternative hypotheses. A very general formulation is to assume that:

1. All the observations from both samples are independent of each other,
2. The observations are ordinal,

3. The null hypothesis is that the distributions of both populations are equal.
4. The alternative hypothesis is that the distributions are not equal.

Using this test we found that distribution of degree centrality of male scholars is significantly greater than that of female scholars (p-value of Mann Whitney U test $< 2.2e{-16}$). Box plots in figure 3 show distribution of degree centrality of male and female scholars.

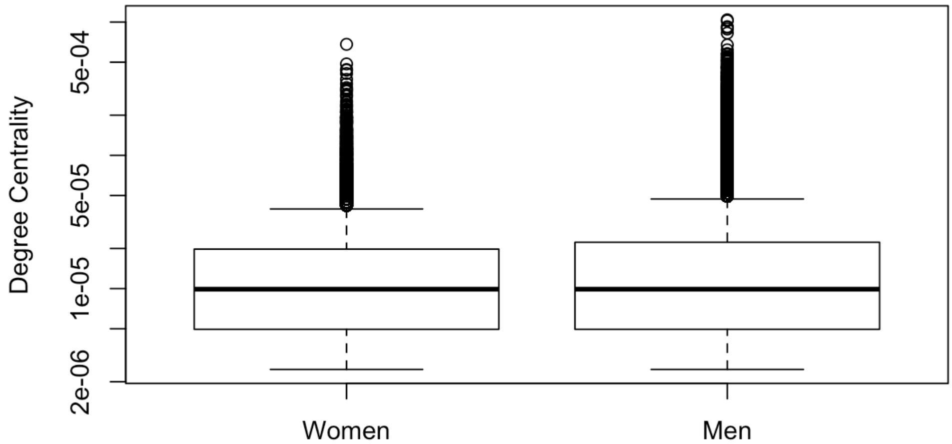

Figure 3: Degree centrality

5.2 K-core

K-core is a maximal subgraph in the APS collaboration network that contains scholars of degree k or more. The core number of a scholar is the largest value k of a k-core containing that scholar [5]. K-core measures cohesion, the extent to which a network has evolved into a hierarchical structure of increasingly dense cores embedding into each other [5]. In the largest connected component of the APS collaboration network, women have lower k-core. Mean of k-core of women is 4.99 and mean of men is 5.13. Using Mann-Whitney U test we found that distribution of k-core of male scholars is significantly greater than that of female scholars (p-value of Mann Whitney U test is 0.0002777). Box plots in figure 4 show distribution of k-core of male and female scholars.

5.3 Clustering Coefficient

Clustering coefficient of a scholar in the APS collaboration network is the fraction of possible triangles through that scholar [6]. Clustering coefficient measures the density of an ego network excluding that ego [6]. In other words, it measures closure, the absence of

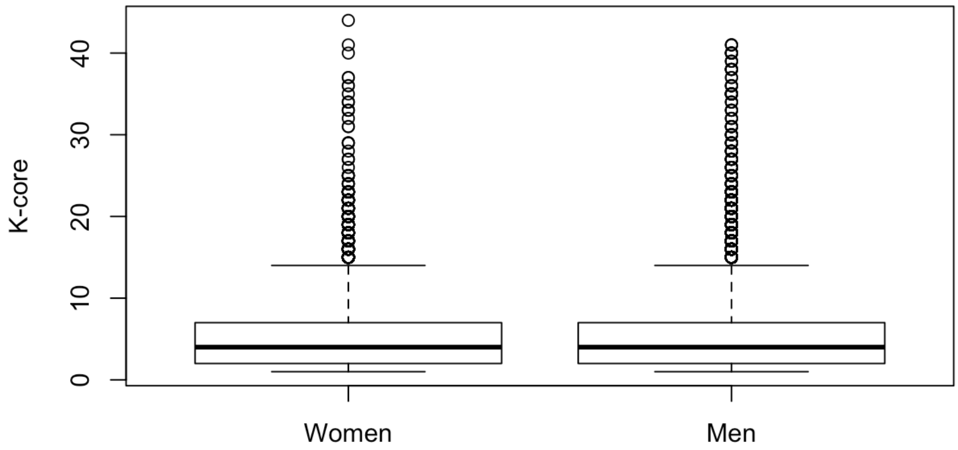

Figure 4: K-core

structural holes. In the largest connected component of the APS collaboration network, women have higher clustering coefficient. Mean of clustering coefficient of women is 0.72 and mean of men is 0.63. Using Mann-Whitney U test we found that distribution of clustering coefficient of female scholars is significantly greater than that of male scholars (p-value of Mann Whitney U test is $< 2.2e-16$). Box plots in figure 5 show distribution of clustering coefficient of male and female scholars.

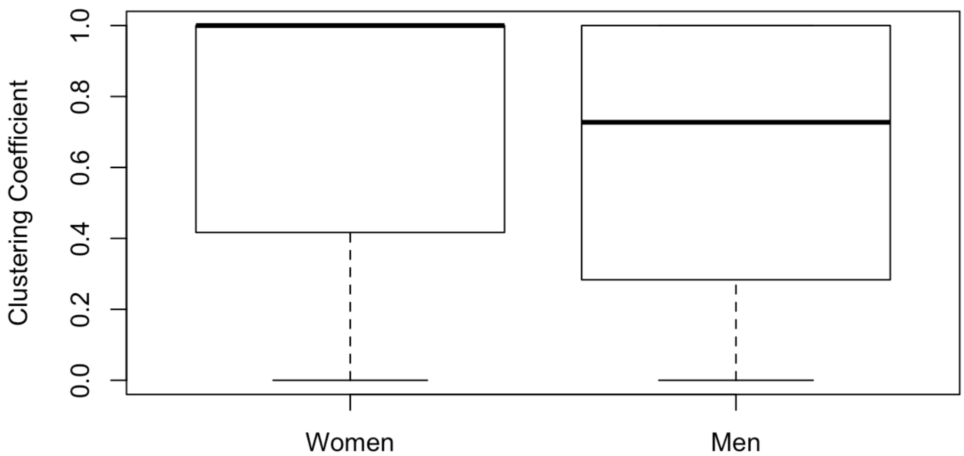

Figure 5: Clustering coefficient

5.4 Eigenvector Centrality

Eigenvector centrality computes the centrality for a scholar based on the centrality of its neighbors in the collaboration network [7]. In the largest connected component of the APS collaboration network, women have higher eigenvector centrality. Mean of eigenvector centrality of women is 9.86477112276e-06 and mean of men is 6.38158053426e-06.

Using Mann-Whitney U test we found that distribution of eigenvector centrality of female scholars is significantly greater than that of male scholars (p-value of Mann Whitney U test = 0.004847). Box plots in figure 6 show distribution of eigenvector centrality of male and female scholars.

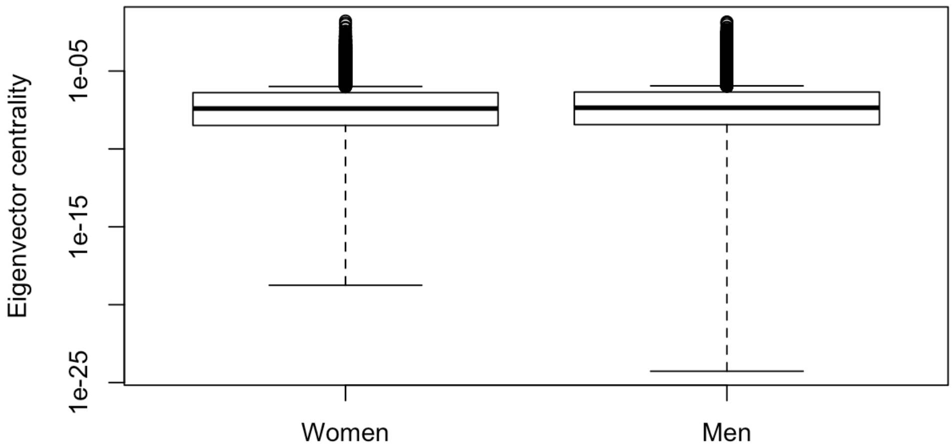

Figure 6: Eigenvector centrality

5.5 Efficiency

Efficiency of a pair of nodes in a graph is the multiplicative inverse of the shortest path distance between the nodes. The local efficiency of a node in the graph is the average global efficiency of the subgraph induced by the neighbors of the node [8]. The average global efficiency of a graph is the average efficiency of all pairs of nodes. Efficiency measures how well the nodes of the graph exchange information as a result of presence of structural holes or possibilities of brokerage [8]. In the largest connected component of the APS collaboration network, women have higher efficiency. Mean of efficiency of women is 0.92 and mean of men is 0.89. Using Mann-Whitney U test we found that distribution of efficiency of female scholars is significantly greater than that of male scholars (p-value of Mann Whitney U test $< 2.2e - 16$). Box plots in figure 7 show distribution of efficiency of male and female scholars.

5.6 Homophily

Homophily refers to a tendency of different types of individuals to associate with others who are similar to themselves [9]. Homophily has been studied across characteristics such as race and gender in co-authorship networks [1]. Here, we investigate gender homophily in the APS collaboration network using two measures. The first one is assortative mixing,

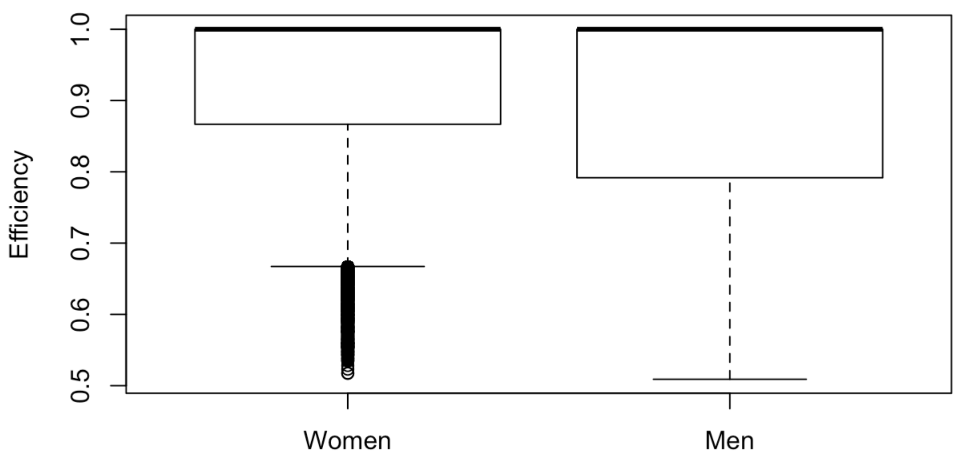

Figure 7: Efficiency

the preferential association of individuals with others who are like them in some way [9]. This measure equals 1 when we have complete assortative mixing (all individuals associate solely with others of the same type as themselves) and 0 when mixing is random (disassortativity). In partial mixing it varies between 0 and 1. We found that gender assortativity is 0.56 in the largest connected component of the APS collaboration network that indicates a positive tendency among scholars to collaborate with similar gender.

In order to measure homophily between women and between men, we used another measure proposed by Coleman [10]. Let s_f denote the number of co-authorships that female scholars have with female scholars, and let d_f be the number of co-authorships that female scholars have with male scholars. Then p_{ww} measures the fraction of the co-authorships of women that are with women.

$$p_{ww} = \frac{s_f}{s_f + d_f} \tag{1}$$

If the fraction of women is denoted by f_w, the Coleman index for women is :

$$C_w = \frac{p_{ww} - f_w}{1 - f_w} \tag{2}$$

Let s_m denote the number of co-authorships that male scholars have with male scholars, and let d_m be the number of co-authorships that male scholars have with female scholars. Then p_{mm} measures the fraction of the co-authorships of men that are with men.

$$p_{mm} = \frac{s_m}{s_m + d_m} \tag{3}$$

If the fraction of men is denoted by f_m, the Coleman index for men is :

$$C_m = \frac{p_{mm} - f_m}{1 - f_m} \tag{4}$$

In the APS collaboration network, $C_w = -0.08$ and $C_m = -0.38$. These results suggest that the homophily among women is higher than the homophily among men in this network.

6 Correlation Between Research Performance and Social Network Analysis Measures

Here, we examine the effect of social network measures (i.e., degree centrality, eigenvector centrality, clustering coefficient, K-core and efficiency) on the research performance (citation count measure) of male and female scholars. For calculating citation count of scholars we constructed the citation network of articles published in APS journals using the APS data set. We used Pearson correlation test to investigate the influence of the five SNA measures on the performance of scholars (citation count). Then, we created regression models that describe the relationship between the collaborative behaviour of male and female scientists and their success.

6.1 Correlation Analysis

The research performance of male and female scholars (citation count) is positively correlated with three SNA measures of degree centrality, eigenvector centrality and K-core and negatively correlated with clustering coefficient and efficiency. The table 2 shows the correlation coefficients of the 5 measures in order of their impact on research performance of scholars. Based on these results, successful male and female scientists have the same collaboration patterns: they tend to collaborated with more distinct scholars than other scientists and seek innovations as brokers. The effect of degree centrality on success of both male and female scholars is higher than the effect of efficiency. These results can explain, to some extent, why men have higher research performance than women. The mean of citation counts of female scholars is 26.91 and the mean of citation counts of male scholars is 50.17. Because research performance is positively correlated with degree centrality and we showed women have lower degree centrality than men. In addition, research performance is negatively correlated with efficiency and as we showed women have higher efficiency than men.

Measure	Men	Women
degree centrality	0.53^{****}	0.55^{****}
efficiency	-0.36^{****}	-0.43^{****}
K-core	0.24^{****}	0.25^{****}
clustering coefficient	-0.23^{****}	-0.25^{****}
eigenvector centrality	0.06^{****}	0.08^{****}

Table 2: Pearson correlation between SNA measures and citation count of scholars.

6.2 Regression Analysis

Here, we model the relationship between the dependent variable of research performance of scholars (citation count) and SNA measures (explanatory variables) using linear regression. Correlation analysis showed all 5 SNA measures are correlated with research performance. However, we found high correlation between clustering coefficient and efficiency and between K-core and degree centrality. Therefore, we only consider degree centrality and efficiency as predictor variables to prevent multicollinearity that happens If the predictors are highly correlated [11]. Correlation coefficients between SNA measures are shown in tables 8 and 9. Before fitting the model, we normalized variables by standardization that makes the values of each variable have zero-mean and unit-variance. In addition, we didn't consider eigenvector centrality as an explanatory variable because removing it didn't change R-squared.

	Citation_Count	Degree_Centrality	Clustering_Coefficient	Eigenvector_Centrality	Efficiency	K_core
Citation_Count	1.000000	0.537856	-0.230303	0.064040	-0.367945	0.248981
Degree_Centrality	0.537856	1.000000	-0.304999	0.175803	-0.643578	0.723497
Clustering_Coefficient	-0.230303	-0.304999	1.000000	-0.022503	0.647343	-0.037906
Eigenvector_Centrality	0.064040	0.175803	-0.022503	1.000000	-0.056659	0.211074
Efficiency	-0.367945	-0.643578	0.647343	-0.056659	1.000000	-0.466746
K_core	0.248981	0.723497	-0.037906	0.211074	-0.466746	1.000000

Figure 8: Multicollinearity of SNA measures of men

The results of our regression analysis are shown in figures 10 and 11. R-squared of the model of male scholars indicates that the model explains 29% and R-squared of the model

	Citation_Count	Degree_Centrality	Clustering_Coefficient	Eigenvector_Centrality	Efficiency	K_core
Citation_Count	1.000000	0.556234	-0.252206	0.083170	-0.439171	0.251031
Degree_Centrality	0.556234	1.000000	-0.249496	0.231224	-0.640689	0.759088
Clustering_Coefficient	-0.252206	-0.249496	1.000000	-0.021183	0.600099	0.023090
Eigenvector_Centrality	0.083170	0.231224	-0.021183	1.000000	-0.056184	0.243833
Efficiency	-0.439171	-0.640689	0.600099	-0.056184	1.000000	-0.418022
K_core	0.251031	0.759088	0.023090	0.243833	-0.418022	1.000000

Figure 9: Multicollinearity of SNA measures of women

of female scholars explains 32% of the variability of the research performance around its mean. The regression models for male and female scholars can be written as:

$$\text{Citation Count}_{men} = 0.56 * \text{Degree centrality}_{men} - 0.04 * \text{Efficiency}_{men} \tag{5}$$

$$\text{Citation Count}_{women} = 0.23 * \text{Degree centrality}_{women} - 0.07 * \text{Efficiency}_{women} \tag{6}$$

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0173	0.003	5.524	0.000	0.011	0.023
Degree_Centrality	0.5614	0.004	142.015	0.000	0.554	0.569
Efficiency	-0.0419	0.004	-10.281	0.000	-0.050	-0.034

Figure 10: Regression model of men

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0946	0.003	-29.860	0.000	-0.101	-0.088
Degree_Centrality	0.2300	0.004	57.334	0.000	0.222	0.238
Efficiency	-0.0712	0.004	-17.271	0.000	-0.079	-0.063

Figure 11: Regression model of women

7 Conclusion

We investigated differences in collaboration patterns between male and female scholars who published in American Physical Society (APS) journals. Our analysis showed male and female scientists differ in their collaborative behaviors (degree centrality, K-core, eigenvector centrality, efficiency and clustering coefficient). Women have lower degree centrality and K-core than men and higher clustering coefficient, efficiency and eigenvector centrality than men. In addition, female scientists have higher tendency to associate with same-gender scholars (Gender homophily). Women have higher Homophily than men in this network. Furthermore, we analyzed correlation between collaborative behaviors of scholars and their research performance and found that successful male and female scientists have the same collaboration patterns: they tend to collaborated with more distinct scholars and seek innovations as brokers.

References

[1] Jadidi, Mohsen, Fariba Karimi, and Claudia Wagner. "Gender Disparities in Science? Dropout, Productivity, Collaborations and Success of Male and Female Computer Scientists." arXiv preprint arXiv:1704.05801 (2017).

[2] Abbasi, Alireza, Jrn Altmann, and Liaquat Hossain. "Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures." Journal of Informetrics 5.4 (2011): 594-607.

[3] https://publish.aps.org/datasets

[4] Karimi, Fariba, et al. "Inferring gender from names on the web: A comparative evaluation of gender detection methods." Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, 2016.

[5] Batagelj, Vladimir, and Matjaz Zaversnik. "An O (m) algorithm for cores decomposition of networks." arXiv preprint cs/0310049 (2003).

[6] Saramki, Jari, et al. "Generalizations of the clustering coefficient to weighted complex networks." Physical Review E 75.2 (2007): 027105.

[7] Newman, Mark. Networks: an introduction. Oxford university press, 2010.

[8] Latora, Vito, and Massimo Marchiori. "Efficient behavior of small-world networks." Physical review letters 87.19 (2001): 198701.

[9] Newman, Mark EJ. "Assortative mixing in networks." Physical review letters 89.20 (2002): 208701.

[10] Coleman, James S. "Relational analysis: the study of social organizations with survey methods." Human organization 17.4 (1958): 28-36.

[11] Hosmer, DWr, and S. Lemeshow. "Applied logistic regression 2nd ed. Chap. 4, Model-Building Strategies." (2000).