COSC 6323 - Statistical Methods in Research
Project Phase - 1

Members: Team-8

1. Md Rafiqul Islam Rabin, ID:1797648, mrabin@central.uh.edu
2. S M Salah Uddin Kadir, ID:1800503, ssalahuddinkadir@uh.edu
3. Farah Naz Chowdhury, ID:1798957, fchowdhury4@uh.edu@

March 08, 2019.

Contributions: In the first week of the project, we worked on Fig 3. At first, we drew Fig 3 with ggplot, but as we faced a problem with y-log transformation, we moved to plot_ly. We all equally contributed to Fig 3. After that, we sit together several times to discuss the required data for Gephi and Supplementary plot. Then we individually collect data (i.e. XDIndicator, Node, Edge, Etc.) from given CSV files. We cross-checked our data for confirmation. So, we all also equally contributed to the data processing. Now, at that moment, we divided our remaining tasks among us: Fig 3(density), 2B, and S2 for Rabin; Fig 3(layout), 2A and S4 for Farah; Fig 3(mean), S1 and S3 for Salah. Although we divided our tasks at that stage, we always shared our progress/problem and helped each other during implementation. We created a git repository and distributively contributed for the project. As we all completed our tasks and actively involved with each other all the times, we almost equally contributed to this phase of project.

Fig. 2(A) :

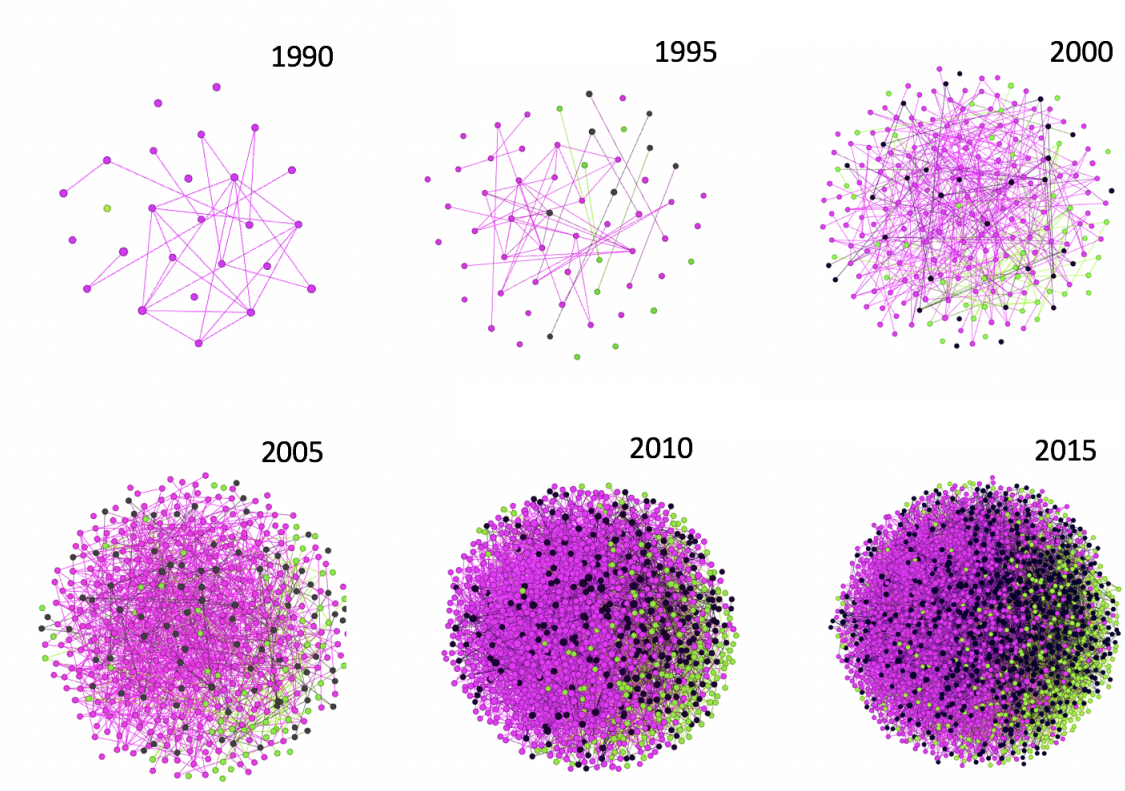

Fig. 2. Growth of cross-disciplinary social capital. (A) Evolution of the giant component in the U.S. biology-computing network.

Description of figure content:

Here, in Fig 2A, it shows how the faculty in the computing and biology department have collaborated over time and formed a giant community which can be defined as U.S. biology-computing network. Green node is for faculty in the biology department and magenta node is for computing department for the same. Black nodes stand for cross disciplined faculty which means by the time of our observation period that faculty has published at least one cross disciplinary publication and thus has joined in the cross disciplinary faculty group. The size of the node is proportional to the degree of their collaboration by that particular time of our observation.

Observations, conclusions, and hypotheses:

- It is evident from the graph that XD nodes have become more dominant over the years. 2011- 2015, it is the most prominent one.
- In the previous year range both the XD and intra discipline publication is very few compared to the later years.
- The degree of collaboration which is shown by the size of the node has increased over time.

Fig. 2(B) :

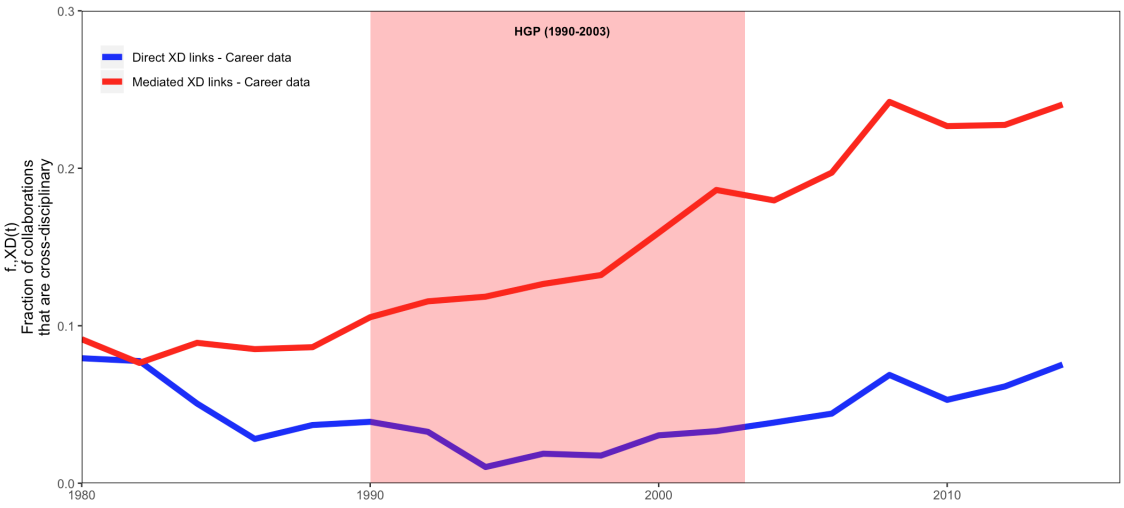

Fig. 2. Growth of cross-disciplinary social capital. (B) Evolution of the fraction of collaboration links in the F network that are cross-disciplinary.

Description of figure content:

Here, in Fig 2B, we plotted the fraction of cross-disciplinary collaboration for each nonoverlapping 2-years period through (1979-1980) to (2013-2014). The blue and red line represents the trend of Direct-XD links and Mediated-XD links, respectively. For each nonoverlapping 2-years period, we collected all the publication data from **GoogleScholar_paper_stats.csv** file. Then, for the Direct-XD links, we count total direct (F-F) links and total cross-discipline direct links for each period. Finally, we calculated the fraction of direct cross-disciplinary collaboration by **total F-F direct links / total cross-discipline direct links**. Similarly, for the Mediated-XD links, we count total mediated (F-P-F) links and cross-discipline mediated links for each period. Finally, we calculated the fraction of mediated cross-disciplinary collaboration by **total F-P-F mediated links / total cross-discipline mediated links.**

Observations, conclusions, and hypotheses:

- The first thing we can see that, the growth is comparatively higher for mediated cross-disciplinary collaboration but slower for direct cross-disciplinary collaboration. This highlights that pollinators are more likely to be cross link than direct faculty persons.
- Both the direct and mediated cross-disciplinary collaboration has increased during HGP(1990-2003) period and the increasing rate also continues upward after the HGP as well. So, the impact of HGP on cross-disciplinary collaboration is clearly visible.
- The increasing rate of cross-disciplinary collaboration continues. In recent years (i.e. by 2015) the direct cross-disciplinary collaboration becomes almost 10% of total direct collaboration, and the mediated cross-disciplinary collaboration becomes almost 30% of total mediated collaboration (whether within-discipline or cross-discipline).

Fig. 3 :

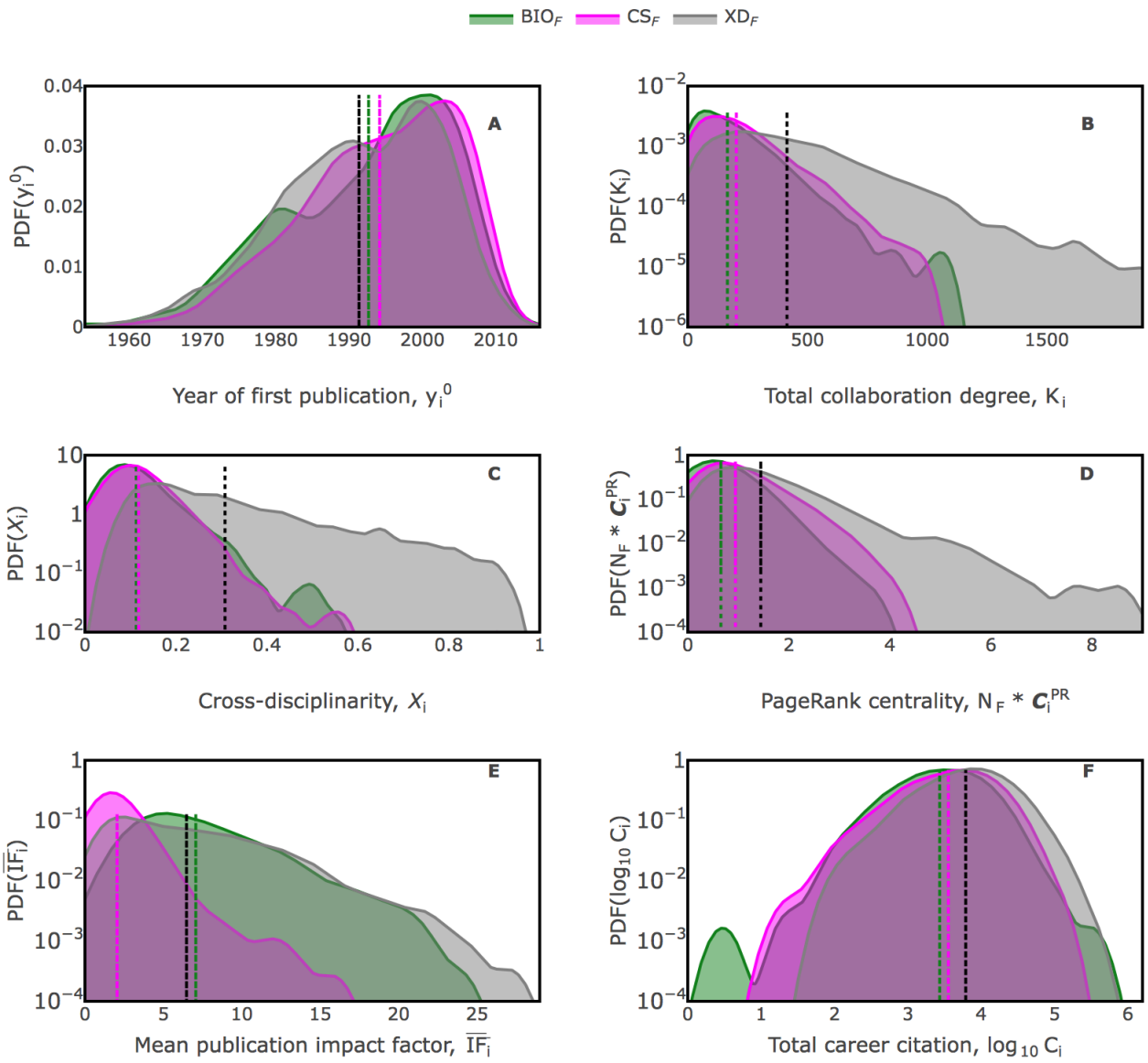

Fig. 3. Descriptive statistics for the career data set.

Description of figure content:

In Fig-3, Green color represents the Biology, Majenta color represents Computer Science, and Gray color represents the Cross-platform distributions. The vertical lines denote the mean values of the corresponding distribution. The Fig-A describes the probability distribution of the year of first publication, y_i^0, by faculty members. Fig-B, the probability distribution of a total number of collaborators for each faculty member. Fig-C describes the ratio of cross-disciplinary collaborator faculties. Fig-D, the probability distribution for page rank centrality, where N_F describes the number of faculty members in the corresponding department. Fig-E, the probability distribution of mean impact factor, $\overline{IF_i}$, of publications of faculty members. Fig-F describes the probability distribution of total citations of each faculty member in **log10** scale.

Observations, conclusions, and hypotheses:

In Fig-3, we have several statistical comparisons on research activities among Biology (BIO), Computer Science (CS), and Cross-Platform (XD) faculty members.

- In Fig-A, we can see that the average starting publication time in the early 1990s. The rate of adding new professors in research in all platforms are similar.
- Fig-B shows that Cross-platform (XD) faculty members have a significantly higher number of collaborations than Computer Science (CS) or Biology (BIO) groups.
- Fig-C indicates that the Cross-platform (XD) group has a significantly higher degree of cross-disciplinarity than the other two groups.
- Fig-D shows that the mean centrality value for google page rank is significantly higher in cross-disciplinary faculty members than other groups.
- Fig-E shows the mean publication impact factor. The figure shows that the Biology and Cross-disciplinary groups have similar publishing pattern.
- In Fig-F, we can see that the mean value of total career citations is similar in all disciplinary faculty members.

Fig. S1 :

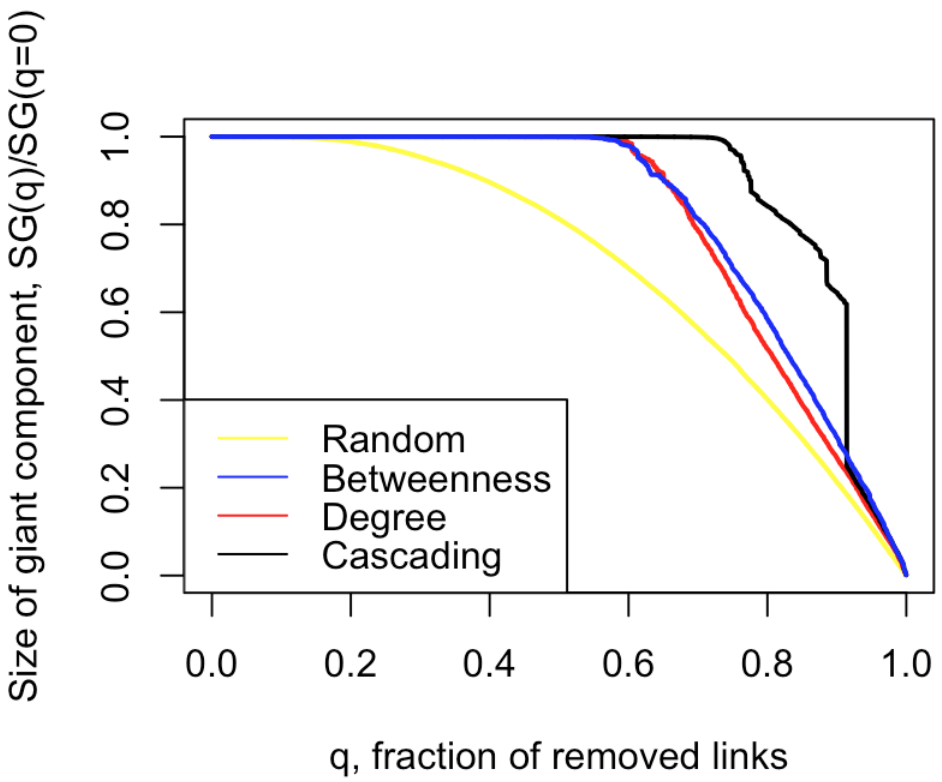

Fig. S1. Robustness of the F network with respect to link removal.

Description of figure content:

This figure shows the Giant component size decreasing rate with the removal of links. We removed links randomly for Uniform distribution. We also removed the links according to increasing weights and also according to the inverted weights. We have got four plots, those are for Uniform distribution, betweenness centralities that the loss of connectivity induced between nodes i and j, Overlap fraction in the first degree neighbors of node i and j, and for cascading attack.

Observations, conclusions, and hypotheses:

From the graph, we can see that the size of the Giant component did not change very much upto 60% of the connections. Again, there were about 60% nodes were connected even after removing 80% of the links. So maximum faculty members are connected in large groups. For the given q, we repeated the fragmentation process for 10 iterations. We tried for 40 iterations, waited about 18 hours to complete the process, but it did not complete and we had to stop the process.

Fig. S2 :

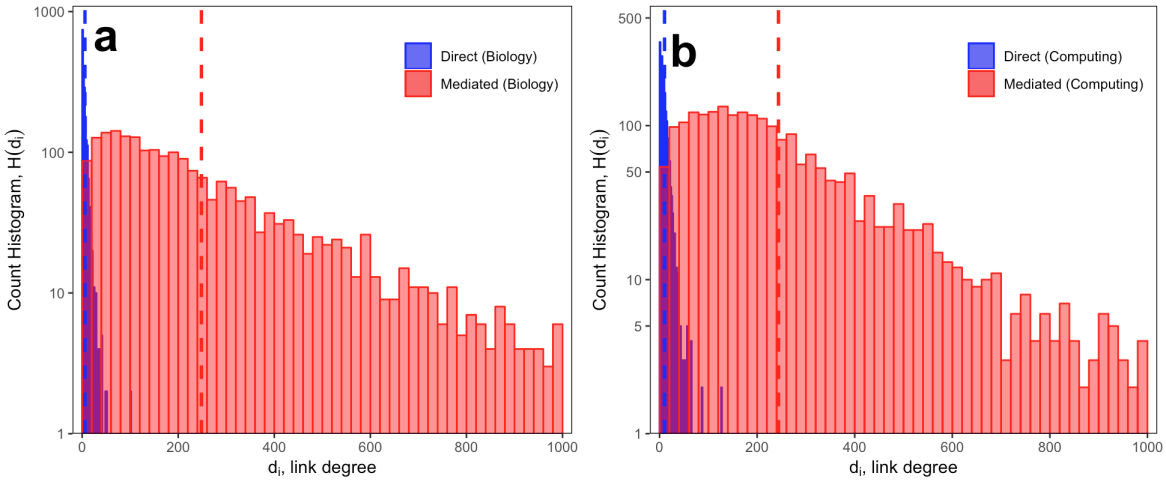

Fig. S2. ***F*** **network distributions for direct and mediated associations.**

Description of figure content:

Here, in Fig S2, we showed the frequency distribution (counts) of faculty within a given link degree. Histogram **(a)** for biology department and **(b)** for computing department. At first, we read the **KDirect** and **KMediated** data from **Faculty_GoogleScholar_Funding_Data_N4190.csv** file for **BIO** and **CS** department, separately. Then we used binwidth=20 for mediated counts and binwidth=2 for direct counts, for both biology and computing department. The blue and red bar represent the counts for direct and mediated links, respectively. The blue and red vertical lines indicate the distribution means for direct and mediated links, respectively.

Observations, conclusions, and hypotheses:

- In higher link degree (i.e after 200 link degree), the mediated link counts are still significant but there is almost no direct link counts, for both biology and computing department.
- More than 90% co-authors are pollinators, for both biology and computing department.
- The mean line, for both biology and computing department, clearly shows that the mediated link degree are higher than the direct link degree.
- Finally, the significant impact of pollinators in network distributions is clearly visible.

Fig. S3 :

Fig. S3. Three perspectives on the centrality of F_i in the direct collaboration network.

Description of figure content:

Here, in Fig S3, three criteria for centrality is considered that indicates which nodes play an important role in the direct collaboration network. These three figures consider all the data collected till 2015.Here the node and edge position is in fixed position for the three figures, only the size of the node is correspondent to the three-different centrality measures. They are respectively degree, PageRank and Betweenness centrality.

Observations, conclusions, and hypotheses:

It is evident from the graph that in all the centrality measures the central corresponding nodes are almost similar. That is the person who is most important in one centrality measure, holds the same importance in other centrality measures in the network. That is those faculties who has high degree measure also has high PageRank measure and high betweenness measure.

Fig. S4 :

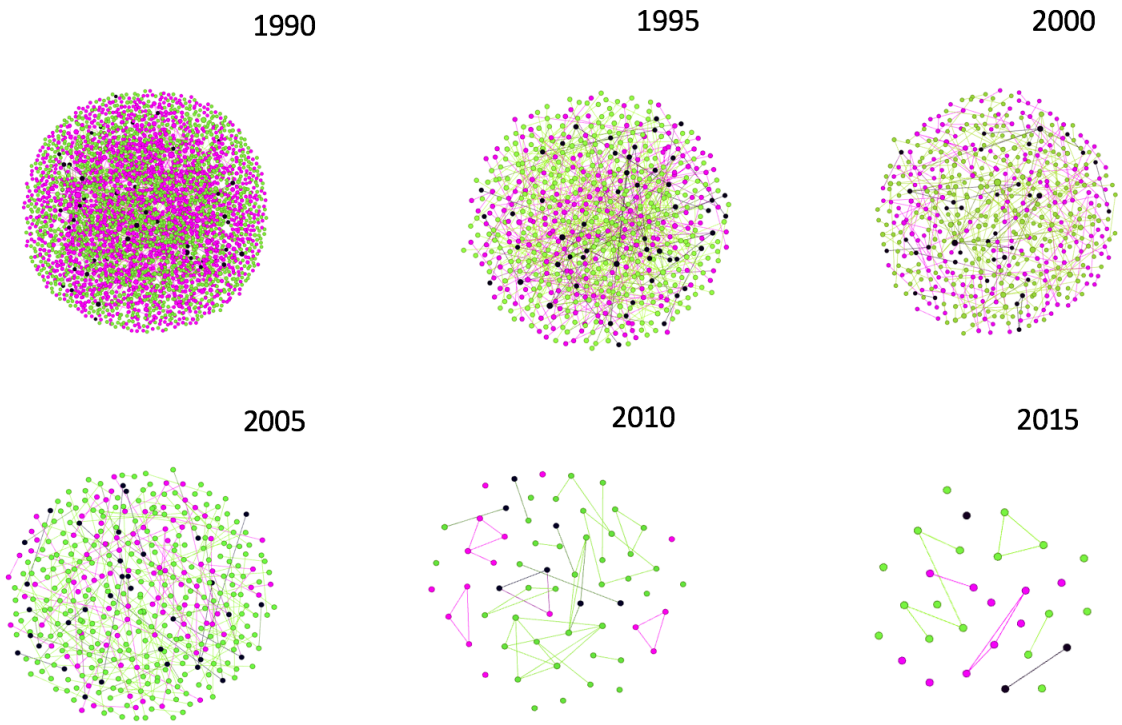

Fig. S4. Evolution of the nongiant components in the F network.

Description of figure content:

Here, in Fig S4, non-giant component is shown for faculty in the biology and computing department. Green node represents faculty of Biology and the magenta node represents faculty of computing department. Black node stands for faculty who has joined the cross disciplinary group by the time of our observation.

Observations, conclusions, and hypotheses:

Over the years the network has become densely connected and if we remove the giant component from the year range the graph becomes less dense as the components which are not part of the giant component are very few. For example, In the year range 1986-1990 the graph seems denser even after giant component removal because at that period of time there were few faculty who were strongly connected with many other faculty.