Project Part 1: Exploratory Analysis of Collaborative Study in Genomics Revolution

Md. Rafiul Amin and MD Tanim Hasan
University of Houston
Texas, USA

January 22, 2021

Contributions from Each Member

Md. Rafiul Amin

1) Graph files and figure generation.
2) Plot with Gephi.
3) Critical thinking over all the generated plots in the report.
4) about 50% of the report writing.

MD Tanim Hasan

1) Probability Distribution figure generation.
2) Network analysis.
3) Critical thinking over all the generated plots in the report.
4) about 50% of the report writing.

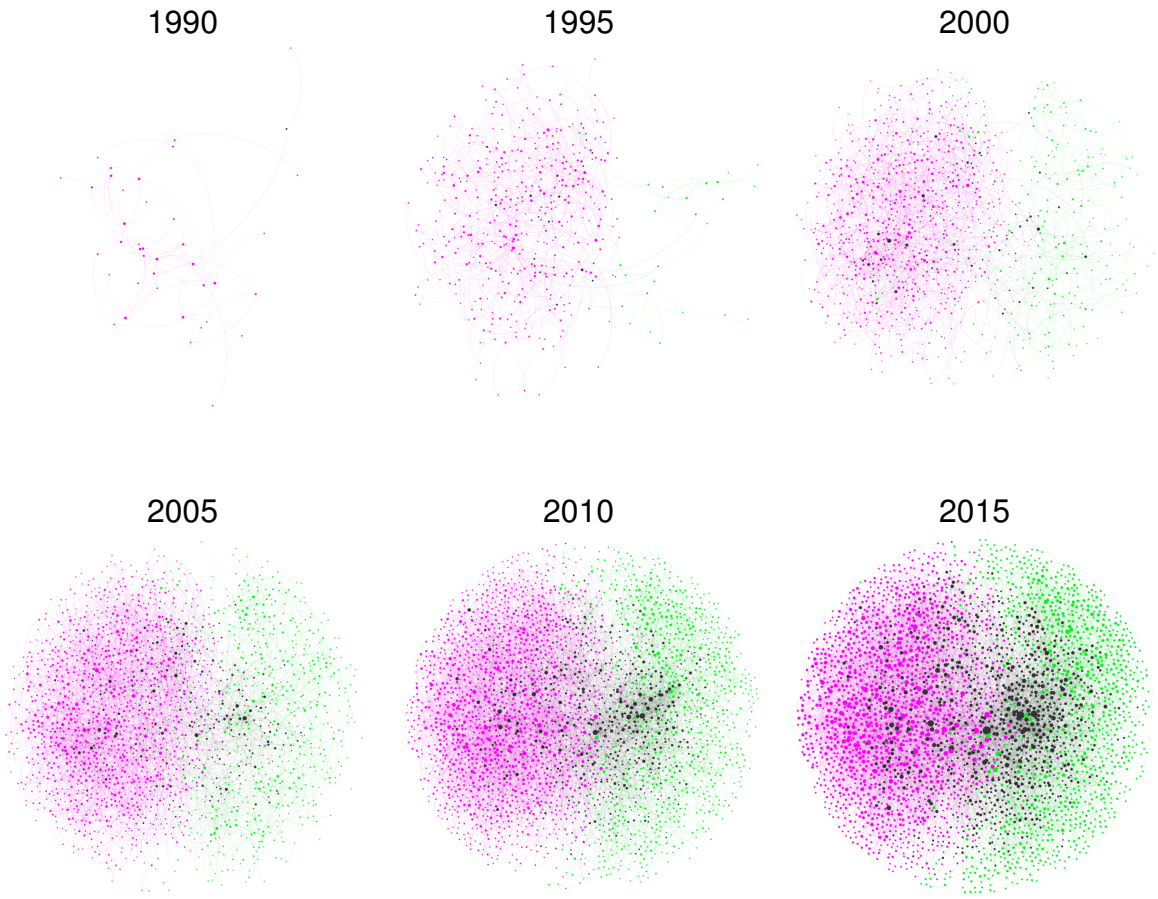

Fig. 2(A). Growth of cross-disciplinary social capital: Evolution of the giant component in the U.S. biology-computing network. Green and magenta nodes represent faculties with biology and computer science affiliation, respectively; black nodes represents that the corresponding faculty has cross disciplinary collaboration by the respective time; node sizes are set proportional to the logarithm of the faculty's degree centrality at respective time.

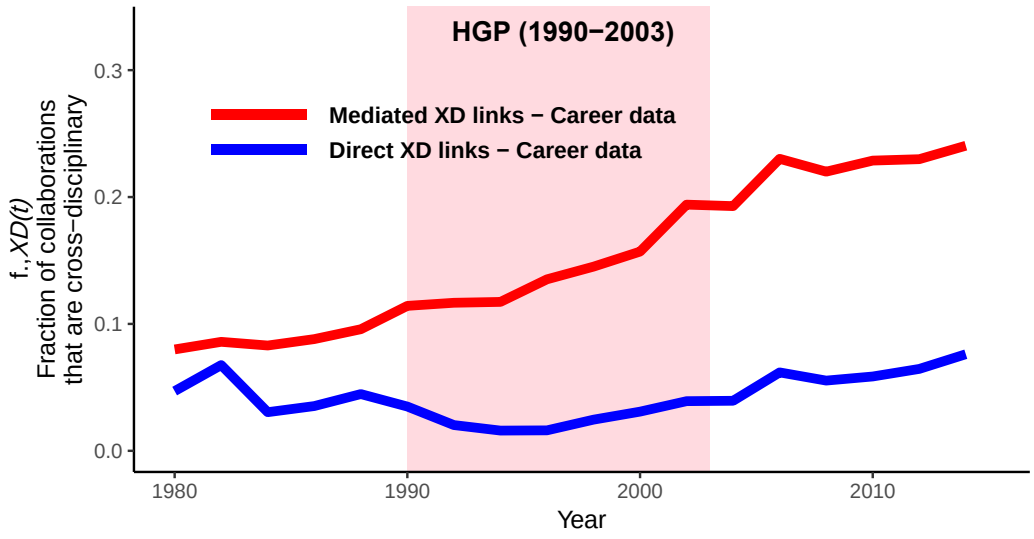

Fig. 2(B). Growth of cross-disciplinary in biology-computing network: Evolution of fraction of cross-disciplinary collaboration according to the published paper. The fraction of cross-disciplinary links $f_{.,XD}(t)$ that are either direct (blue curve) or mediated (red curve) have been calculated for time t (every two years). The link is annotated as a Direct XD link if a direct collaboration took place between two faculties. If the collaboration is through a pollinator (non faculty author) then the link is annotated as mediated XD link.

From Fig. 2(A) and 2(B) we can observe that the number of cross-collaborator increased significantly over time. Furthermore, the larger size of the cross-collaborator in the year 2015 suggests that the average degree centrality increased significantly in cross-collaborators compared to the previous years. This suggests that cross-collaborator faculties started to become more successful over time and became a leader of this knowledge graph.

Although exploratory analysis suggest that the cross-collaborator have more average degree centrality, this result might not suggest the actual scenario of the knowledge graph. If a senior faculty successful faculty have more degree centrality, then the probability of having a cross disciplinary link (mediated or direct) is higher for him. Where it is more intuitive that the faculty should be annotated as a cross-collaborator if the percentage of the cross-collaboration exceeds a certain threshold (for example 10% rather than just one node in his lifetime).

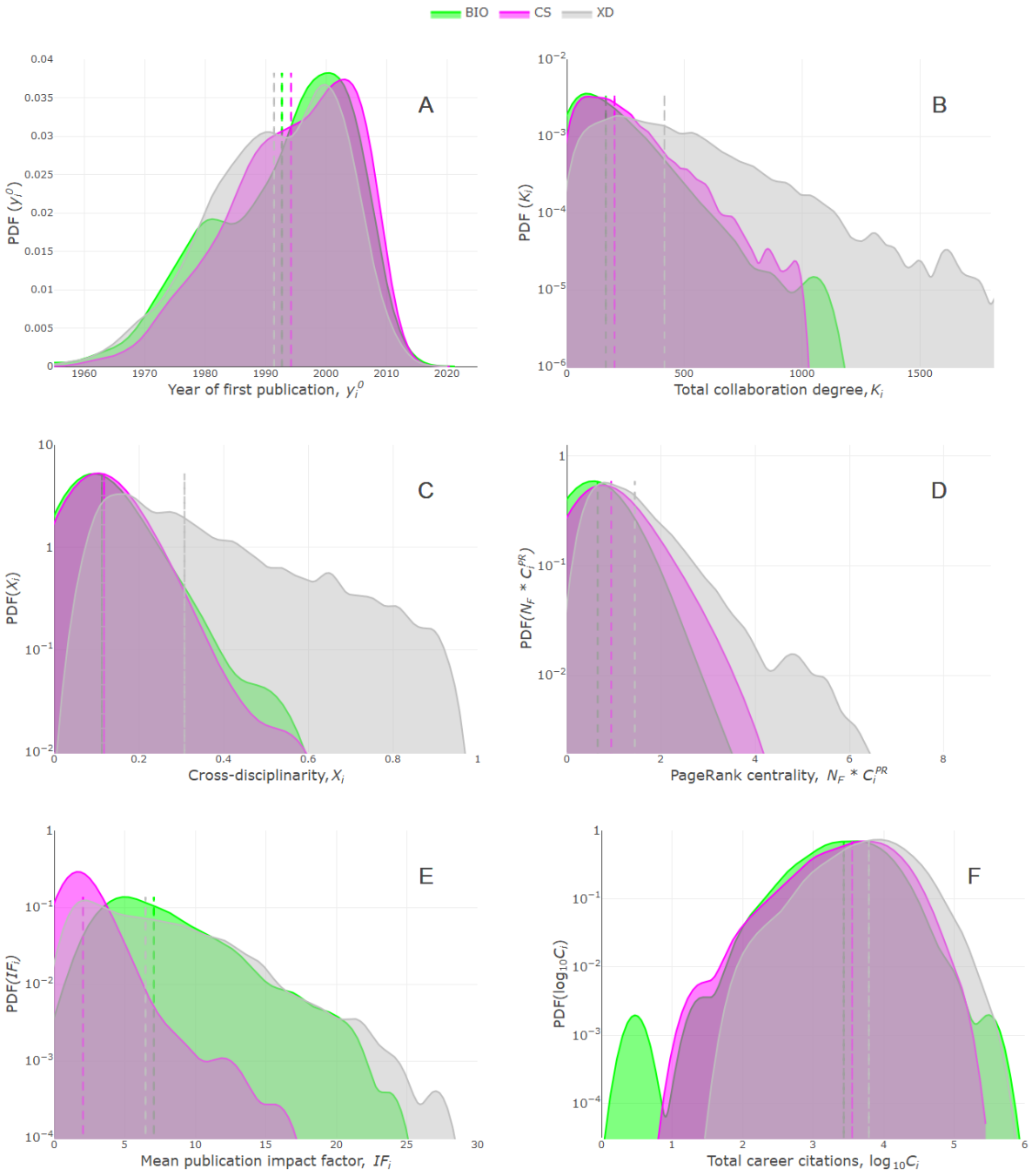

Fig. 3. Probability distribution of career data set. The vertical dashed lines indicate the mean of corresponding subsets. **(A)** Probability distribution of the year of first publication y_i^0 by F_i. **(B)** Probability distribution of K_i, the total number of collaborators for a given F_i. **(C)** Probability distribution of X_i, the fraction of the collaborators of F_i who are cross-disciplinary. **(D)** Probability distribution of C_i^{PR}, the PageRank centrality of ; it is scaled by N_F, the number of faculties, so that the mean value of this scaled quantity across all F_i, independent of the discipline subset, is 1. **(E)** Probability distribution of the mean impact factor (IF_i) of the publication record of F_i. **(F)** Probability distribution of the total citations $\log_{10} C_i$ of F_i.

From Fig. B, C, D, and F we can see that the means corresponding to XD are

greater in BIO than in CS. Which suggests that faculty who are working collaboratively are more successful. Moreover, the distribution in Fig. A, the mean of year of first publication, the XD faculty have a lower mean. Which implies that, faculty who started early eventually joined the XD group. Furthermore, XD collaborators have a relatively high mean of impact factor. All these distributions summarizes that, XD collaborators are more successful or more successful faculties eventually joined the XD group.

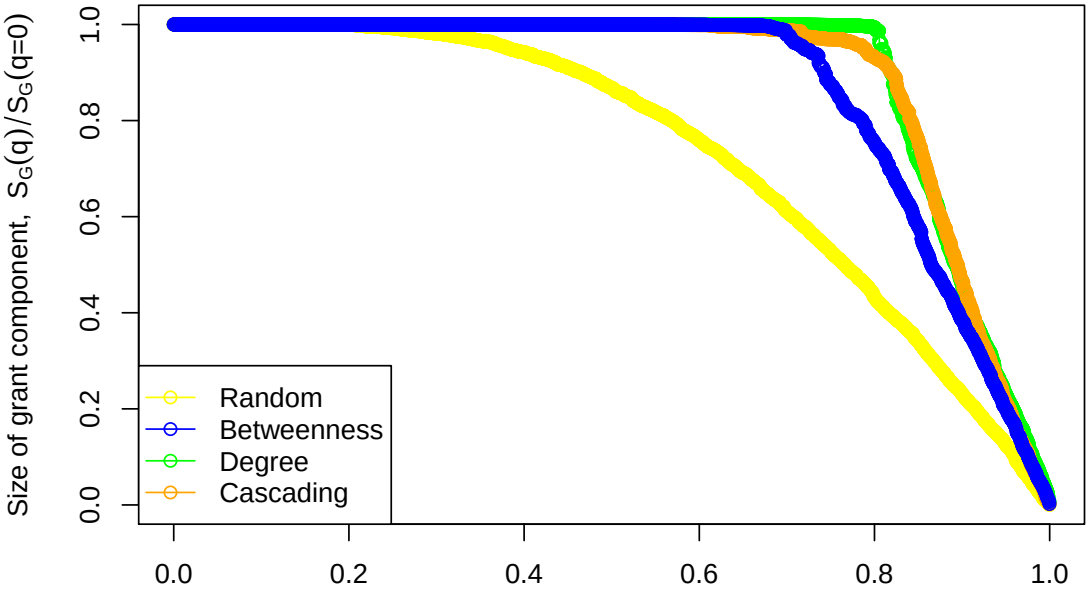

Fig. S1. Robustness of the F network with respect to link removal. The ratio $\frac{S_G(q)}{S_G(q=0)}$ measures the size of the largest remaining fragment $S_G(q)$ relative to the size of the initial giant component $S_G(q=0)$.
We can observe that the ratio $\frac{S_G(q)}{S_G(q=0)}$ as a function of q, demonstrating the robustness of the collaboration network - even after the majority of the links are removed, roughly 50% or more of the F are still connected within the network. The number of iterations for random error assessment is taken as 1 for this case.
We can conclude from this that the network is very well connected. The network has high resistance to random breakdowns or not vulnerable to intentional attacks.

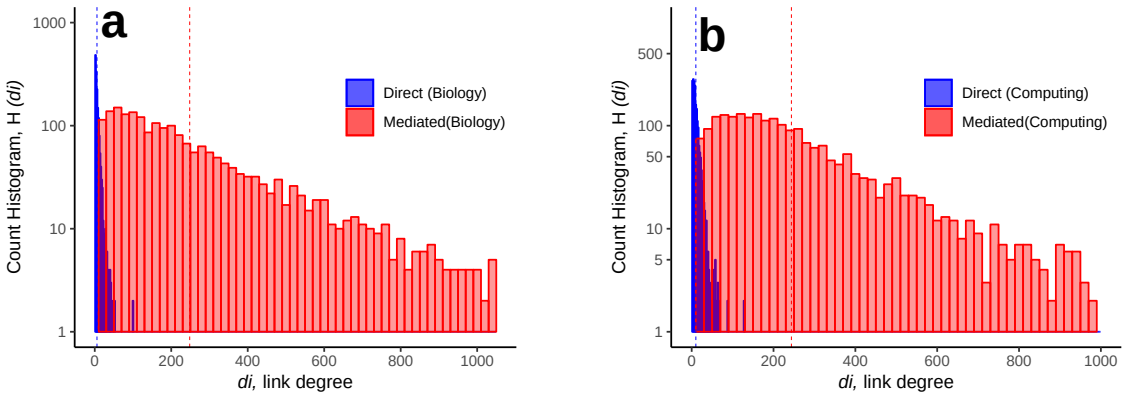

Fig. S2. F network distributions for direct and mediated associations for Biology(A) and Computer Science(B). From this figure, we see each panel shows the frequency distribution (counts) of faculty **F** with a given link degree counting the number of links for a given node, $d_i \equiv C_i^D(t)$, within a particular definition of the F network. The direct sub-networks only include direct links, which are established when two **F** collaborate on at least one publication. The mediated sub-networks only include indirect links between two **F** who have both collaborated with a common pollinator. Vertical dashed lines indicate distribution means.
In our observation, authors are mostly pollinator co-authors in both biology and computing, i.e., researchers not included in the **F** set.

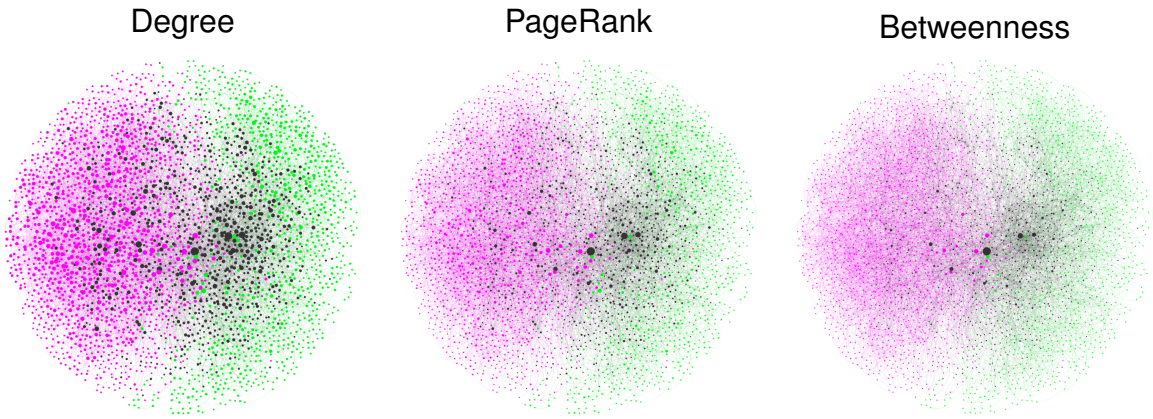

Fig. S3. Faculty Network Graph in 2015 for Different Node Centralities. Shown is the giant connected component of the faculty network **F** using all data from 2011 to 2015. The nodes and links across each network are fixed, only the node sizes vary according to the indicated centrality measure. For better visualization and comparison of the centrality measures same logarithmic transformation has been performed for three of the figures.
We can see a gradual disappearance of the medium size nodes from left two right in these three figures. This suggests the each of the exponential distributed centrality has an increase in the rate parameter in the probability distribution function.
We can hypothesize from this observation that, if a faculty is very important in the network he is more likely to have a collaboration with another important faculty in the knowledge graph.

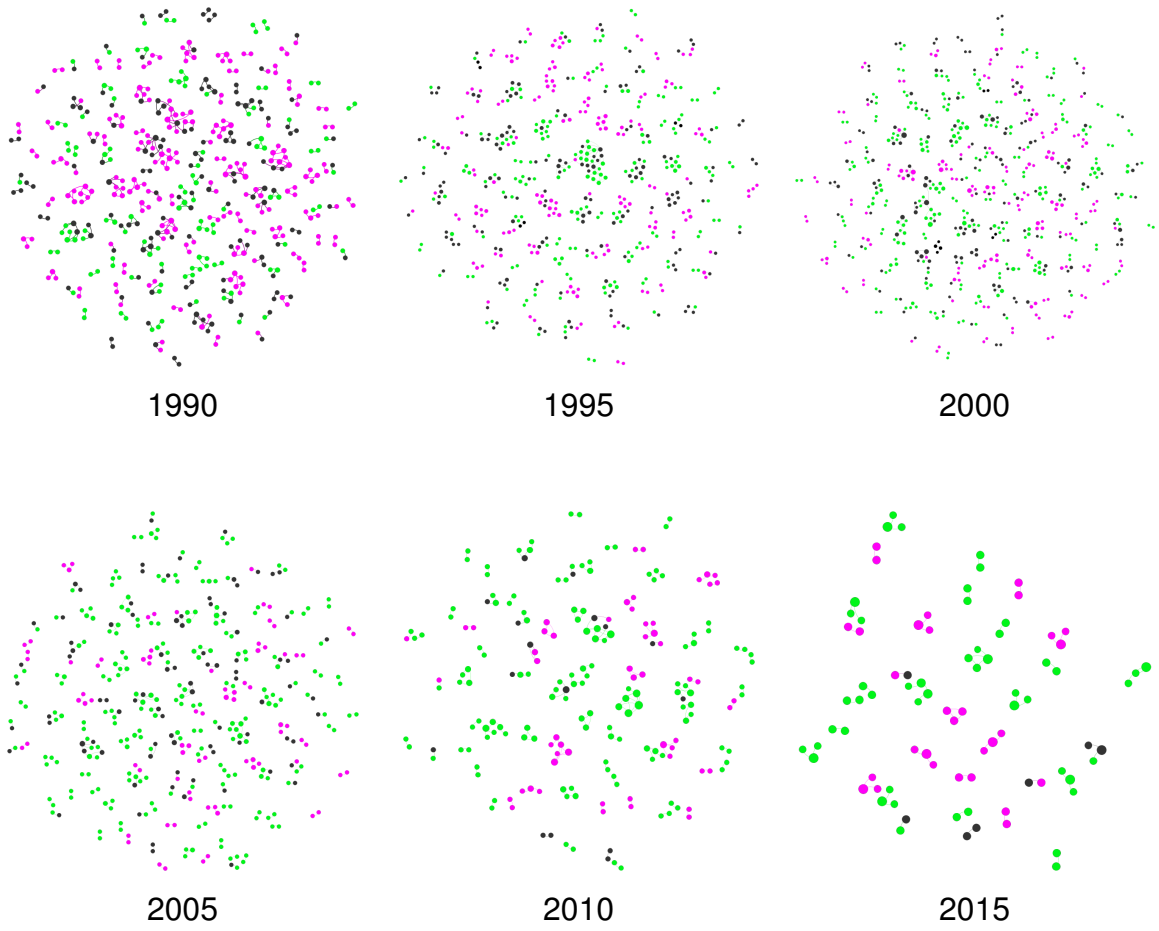

Fig. S4. Evolution of nongiant component in the faculty network F. Green and magenta nodes corresponds to the faculty F_i in **F** with BIO_F and CS_F, respectively. Black nodes represents the faculty that have at least one publication with another faculty with opposite affiliation by the time t (years) and thus joined XD_F group.
We can observe that there is a decrease in the number of non-giant components over the time except for year 1990. There is less number of non-giant component in year 1990 because the knowledge graph has less number of faculties. We can hypothesize that, over the time the collaborative research increases.