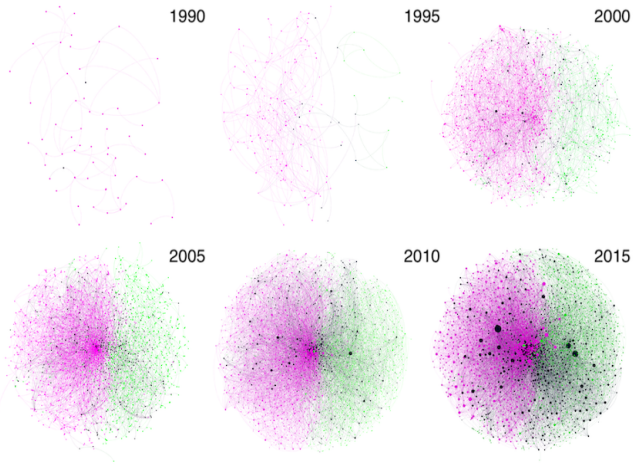
COSC6323: Group Assignment

Reproducing figures 'Cross-disciplinary evolution of the genomics revolution'

Vitalii Zhukov, Eljose E Sajan, Jonathan Plata

3/8/2019

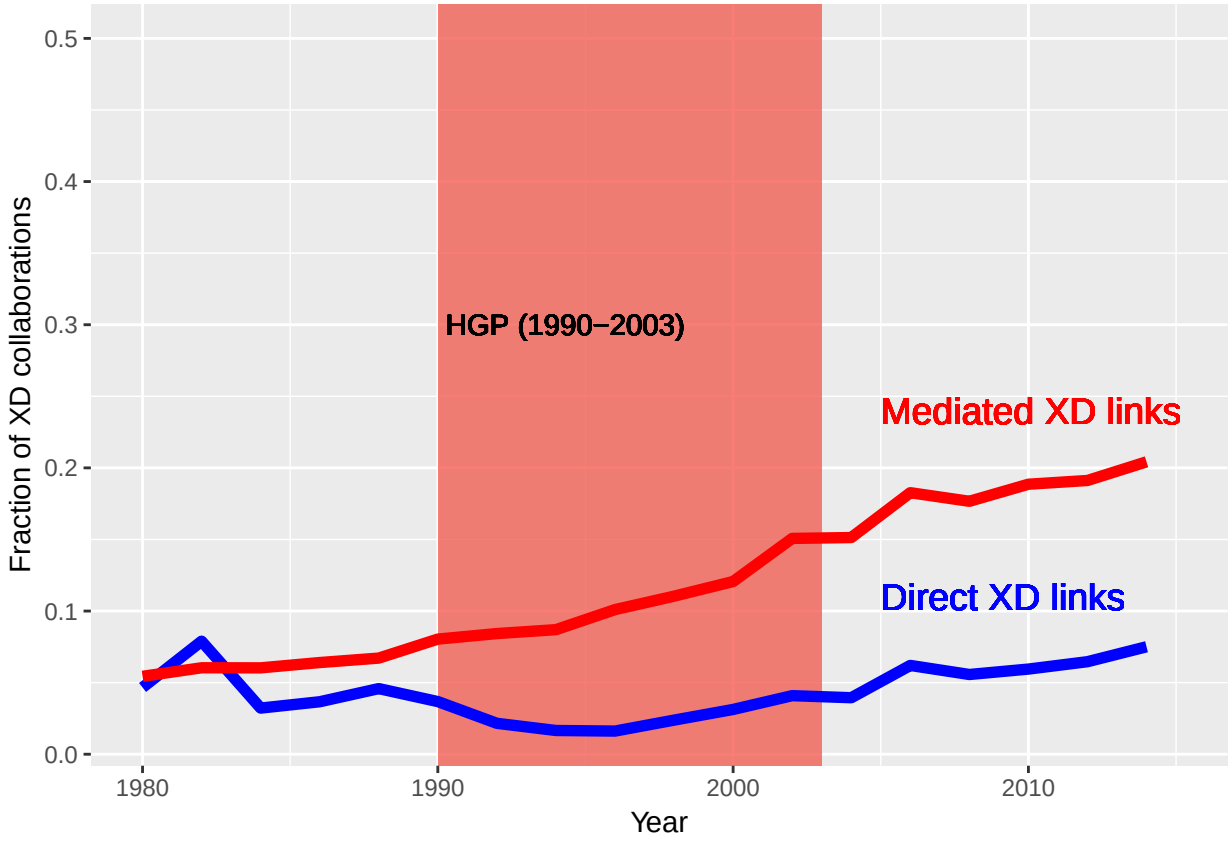
Fig 2a Evolution of the giant component in the U.S. biology-computing network.



Comments: These figures demonstrate the growth of the collaboration network from before the start of the HGP until 2015. Green nodes represents the faculties that belongs to the BIO group, while the magenta nodes represents the CS group.
The black nodes corresponds to the XD group. Faculty nodes are classified as XD after the first year in which cross-disciplinary collaboration was recorded. Nodes are sized according to their degree.

Conclusions: The XD group clearly grows in size beginning with the start of the HGP, becoming better connected as time passes. Since the graphs show only giant components, the growth of the XD network implies that researchers became more cross-disciplinary and better connected. Non-giant components decrease as time goes by and the XD group becomes more homogenous, implying that the set of giant components swallowed up the non-giant components as well.

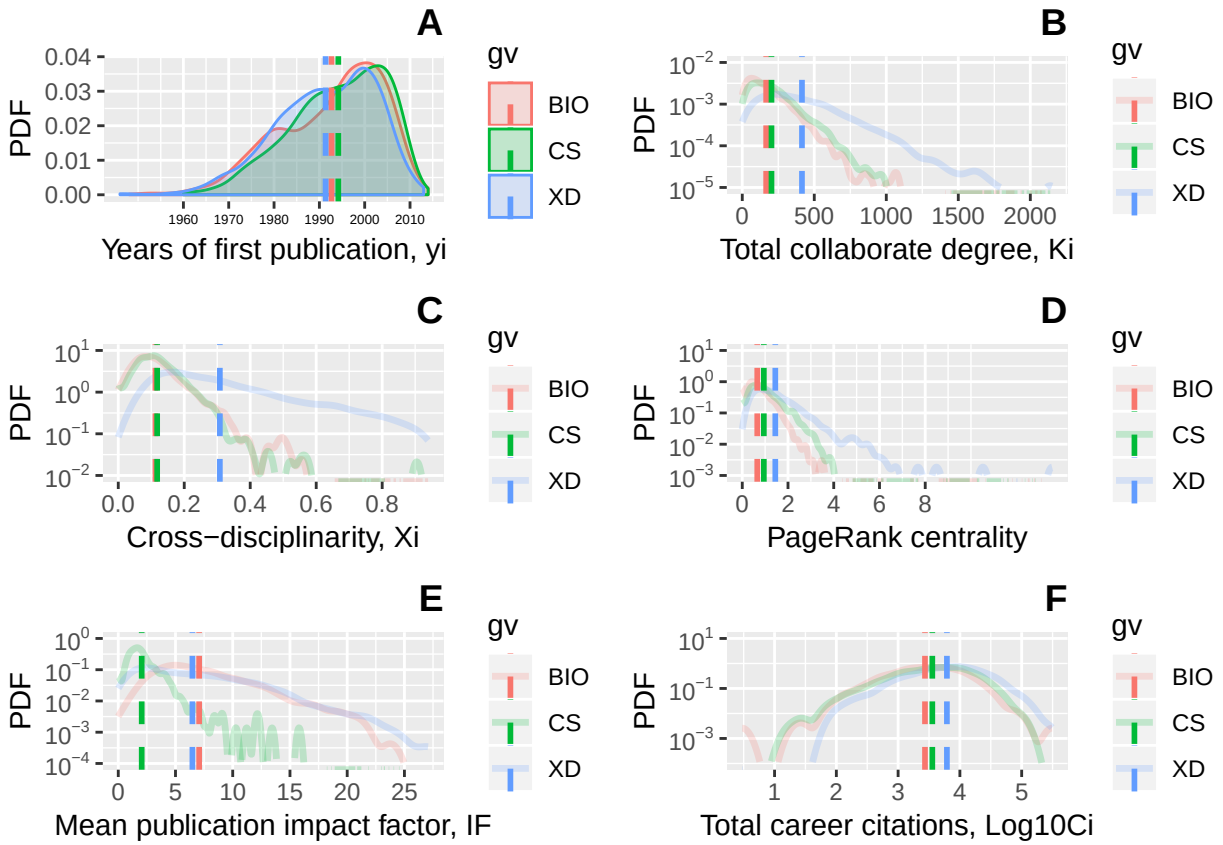
Fig 2b Fractions of collaborations that are cross-disciplinary.



Comments: Evolution of the fraction of collaboration links in the F network that are cross-disciplinary. We calculated Direct XD links between faculties (blue line) and association links mediated by polinators (red line). Orange area marks the HGP project period.

Conclusions: We tried to quantify emergence and centrality of cross-disciplinary scholars in the network during and after HGP. We can notice marked growth during and in the wake of HGP project period, which greatly illustrates the common trend of cross-disciplinarity growth in the field of genomics as time passes.

Fig 3 Descriptive statistics for the career data set

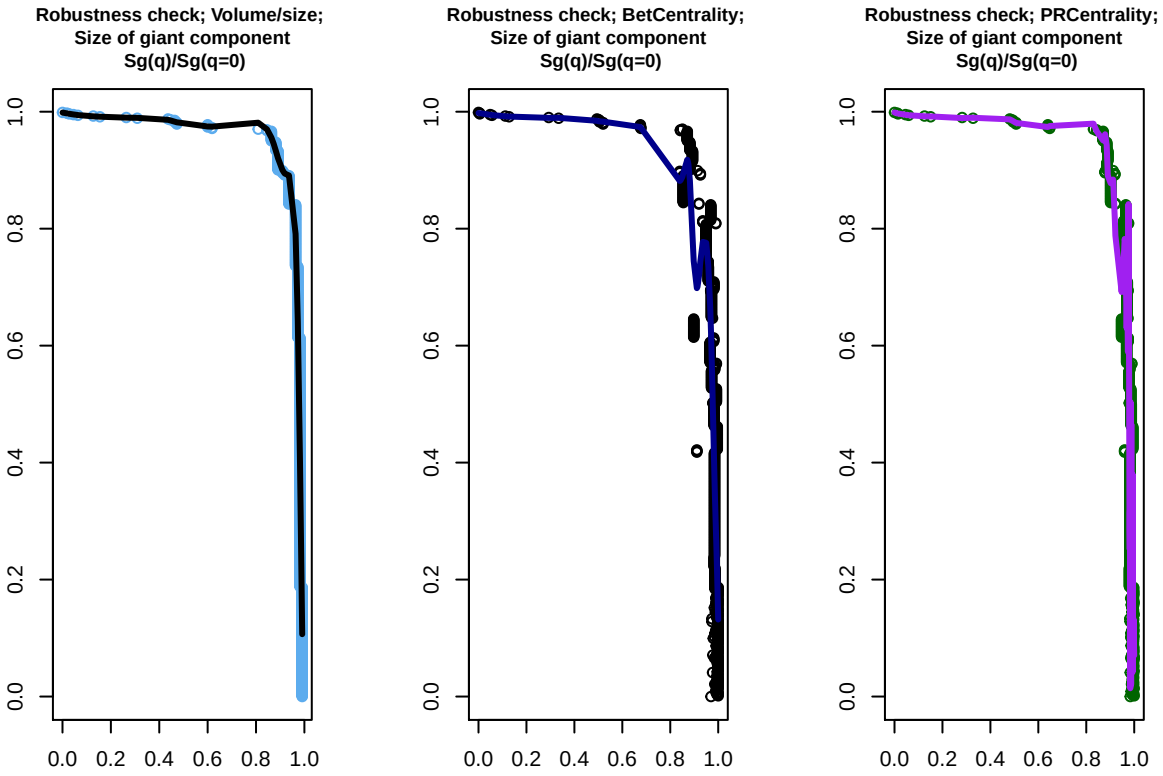


Comments: (A) Probability distribution of the year of first publication by Fi. (B) Probability distribution of Ki, the total number of collaborators for a given Fi. (C) Probability distribution of Xi, the fraction of the collaborators of Fi who are cross-disciplinary. (D) Probability distribution of Fi, the PageRank centrality of Fi; (E) Probability distribution of the mean impact factor of the publication record of Fi. (F) Probability distribution of the total citations log10 Ci of Fi.

Conclusions:

- Fig3A shows that typical Fi, either in BIO or CS department,started his/her career in the 1990s. The HGP initiative started in early 1990, so this set of Faculties are an ideal group for this study on the effect of the genomic revolution on cross-disciplinary evolution.
- Fig3B demonstrates that we have significantly higher degree of cross-collaboration with in the XD group, compared to BIO, CS.
- Fig3C shows that the XD group has a significantly higher degree of cross-disciplinarity than CS and BIO groups. The measure of Chi represents the fraction of her/his collaborators who are cross-disciplinary.
- Fig3D shows that the mean PageRank centrality of XD group is significanly higher than the mean centrality of BIO and CS. The PageRank centrality provides information about the weightage of each link thereby providing an indication on the quality of that link. This figure clearly shows that the XD Faculties have links with more importance.
- Fig3E shows that XD faculties have similiar publishing behavior as BIO faculties (high-impact factor journals). We calculated the mean Journal Citations Report (JCR) impact factor among the publication set of each faculties.
- Fig3F shows that XD group has higher mean citation impact (log10) compared to XD and BIO faculties.

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

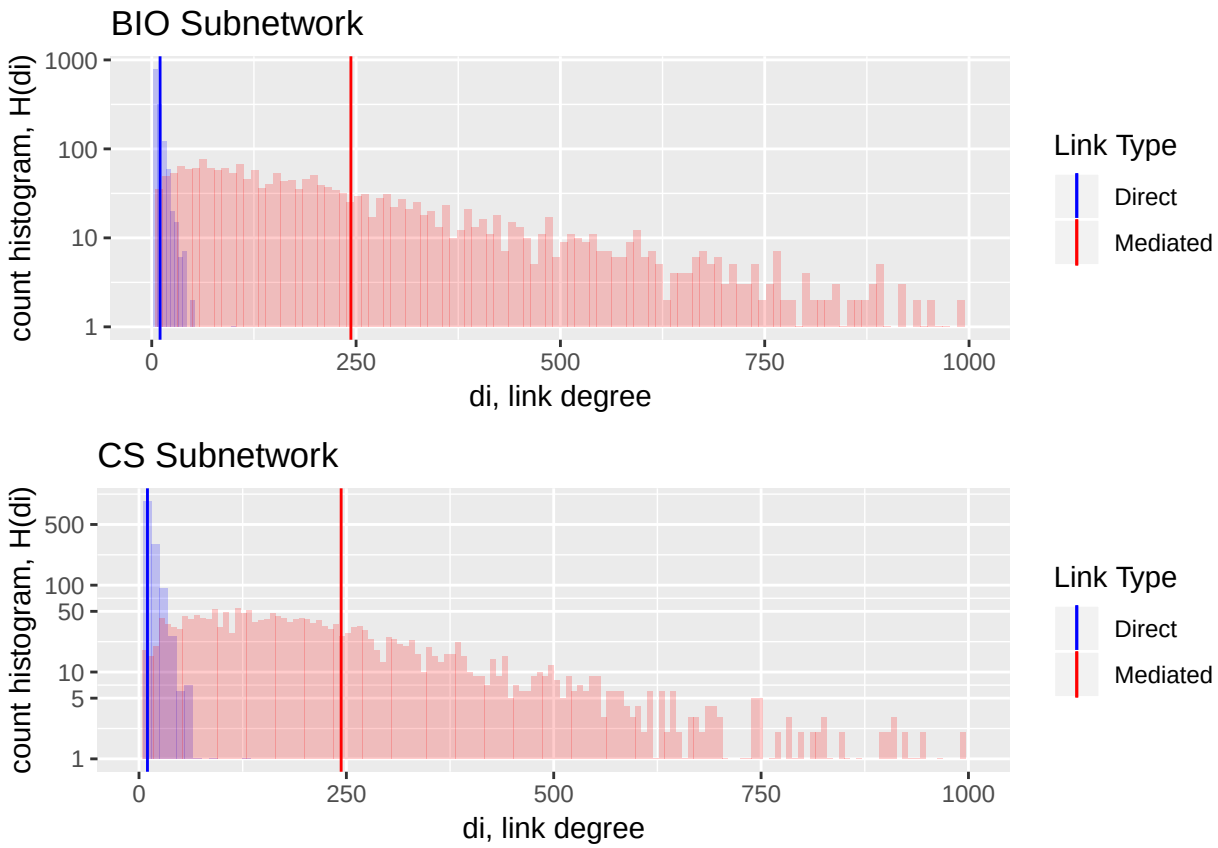


Comments: Robustness check of network. The ratio Sg(q)/Sg(q=o) measures the size of the largest remaining fragment Sg(q) after link removal, relative to the size of the initial giant component Sg(q=0). Compared subnet volumes, BetCentrality and PRcentrality values. The vertical axis corresponds to the afore mentioned ratio and the horizontal axis corresponds to the amount of links that are removed.

Conclusions: A higher value in the vertical axis implies that the largest remaining fragment of the network and the initial giant component are comparable to each other. The increasing value on the horizontal axis corresponds to the fraction of the links that are removed so as to test the network robustness.

The slow decay of the network on link removal,until q=0.6-0.8 indicates that this network is robust to variation in the connectivity of scholars to a large extent. The network starts deteriorating significantly only after more than 60% of the components in our network is removed. This implies that our network mainly consists of relativly big components, which is quite significant for our study, since the data under study is of the type of Knowledge network, which by nature should be higly connected and robust. This test gives conclusive proof that the dataset choosen isby nature a solid knowledge network.

Fig S2



Comments: This figure is used to represent the distribution of the associations between the faculty members of the dataset. A direct association is established between 2 Faculty members when, both have co-authored or collaborated on at least one publication. A mediated link on the other hand is established when bot the Faculty had collaborated with a common co-author; thereby forming a triadic closure. In the figure, the horizontal axis corresponds to the number of links for a given node (i.e. the link degree). The vertical axis corresponds to the count of these link degrees.

Please note, since the vertical axis is in the logarithmic base 10 scale, the y-axis starts from 1. Hence an empty histogram count represents a minimum value of 1.

Conclusions: The count of direct link degree for both the BIO and CS subnetworks is the highest at the lower end of the horizontal axis. From the data that is available, the number of publications where 2 Faculty collaborate on directly are limited, as represented by the x-axis. This represents that individual Faculty members collaborated directly on a publication only a limited number of times. The mediated links degree on the other hand shows the strong association between the Faculty members in our dataset, if we consider it as a knowledge network. The association between Faculties becomes much larger if we include just one common pollinator between them. The count(y-axis) and values (x-axis) of the mediated degree links provide solid evidence that the dataset that is chosen; even when considering BIO and CS colleges separately, form a concrete Knowledge network.

In addition: The data transformations that were done to reproduce this graph was taken from Faculty_GoogleScholar_Funding_Data_N4190 data set, specifically the KMediated and KDirect entries. But by the data description, the KMediated represents the number of pollinator co-authors and not the actual degree of mediated links of that Faculty. So we designed a new method that could be used to count the mediated degree of the faculty. But the limitation of this method was that the GoogleScholar_paper_stats dataset abstracts the pollinators of the invisible college into 3 groups 0,1,2 respectively depending on their

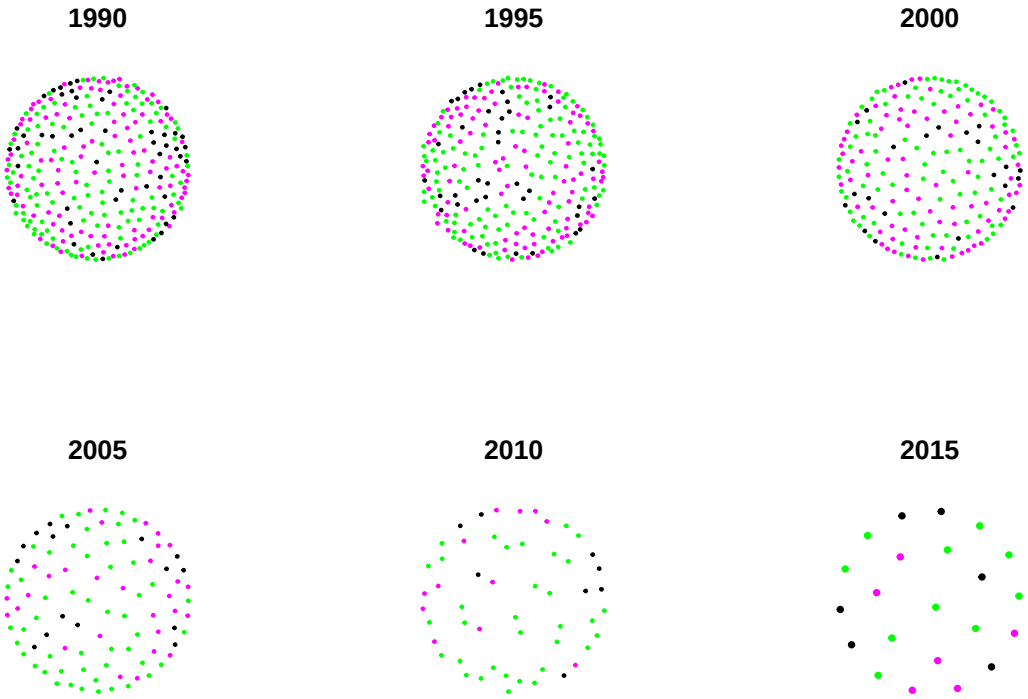
college affiliation. This makes it difficult to distinguish between the pollinators of the invisible college and hence our method provides a largely different result from the paper. The code for this is nonetheless provided for reference.

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

Comments: This figure uses the dataset generated by the code shown in the Figure 2a snippet, specifically the final 2015 dataset. Nodes remain fixed in position while the sizing varies on centrality measure: degree, PageRank and betweenness.

Conclusions: Visually, the three centrality measures do not appear to much different with respect to node size. This implies that while they measure different node properties, they appear to be similar or correlate to each other. Prominent and well-connected members of the XD group remain around the same size in all three centrality measures.

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



Comments: Green and magenta nodes represent faculty Fi with BIO and CS affilation respectivly, black nodes represent faculty Fi that by time t collaborated with at least one faculty from the opposite department

and thus joined XD group.

Conclusions: FigS4 shows that with time giant component of the F network significanly growing in size, cross-disciplinary nodes also grows in numbers. The amount of small subnetworks reduces significantly during and after HGP period.