Final Project Part 2 for DNSC 6215

Group: Daniel Chen, Sara Kalwar, Hanjing Wang, Yunning Zhu

Introduction

This network research analysis is conducted to evaluate the effect of hosting meetings among Cancer researchers and Communicable Diseases (CD) researchers in terms of the degree of their collaboration in both 2015 and 2016. School of Health Policy and Management at George Washington University made intervention among those researchers by arranging meetings for these individuals. We want to see the change in the researchers' interaction by analyzing their network relationship and attributes, especially across-field collaboration.

Objective

Our objective is to evaluate whether there is a significant increase, decrease, or no change between the levels of research collaboration in 2015 and in 2016.

About Datasets

The datasets provided are from four states, including Florida, Vermont, South Dakota and Kentucky. For every state we have six datasets, including information exchange as a sum of send and receive, information exchange in a Binary Adjacency Matrix, and the researchers' attributes for both 2015 and 2016. We combined the send and receive datasets into one dataset by adding them for each state.

While cleaning the attributes datasets, we used different colors to represent different types of Cancer IDs. Blue represents Cancer researchers, Green represents Communicable Disease researchers and Pink represents others.

Density

We conducted the density of the graph for the information exchange in each states in both years and the data is shown in the table below.

For the combined dataset:

Density	2015	2016
Florida	0.05347594	0.1176471
Kentucky	0.04098361	0.1277778
Vermont	0.1011494	0.1117647
South Dakota	0.2904762	0.1580087

As we can see from the data, the density of informational exchange for Florida and

Kentucky increased by two times. For Vermont, there's slightly 10% increase. However, the value for South Dakota decreased by 50%. KY has the lowest density in four states in 2015 so the meeting might has the most impact on KY. SD has the highest density in both years which means researchers in SD collaborate best.

For the "How do you Know" dataset:

Density	2015	2016
Florida	0.09803922	0.1327986
Kentucky	0.06054997	0.1444444
Vermont	0.02643678	0.1747899
South Dakota	0.3238095	0.2662338

Similar with the density for information exchange, the density of the degree researchers know each other for Florida and Kentucky also increased for about 50%. But for Vermont and South Dakota the values decreased. Clearly researchers in SD know each other very well compared to the other 3 states; researchers in VT don't know each other well in 2015 but has the most increase in 2016.

Based on the change in the density for these four states, we can conclude that the level of how close researchers work together increased for Florida and Kentucky, decreased for Vermont and South Dakota.

Reciprocity

We also conducted reciprocity, which is the proportion of reciprocated ties for the networks. For the combined dataset:

Reciprocity	2015	2016
Florida	0.8	0.6212121
Kentucky	0.4387097	0.5465839
Vermont	0.4772727	0.2556391
South Dakota	0.6065574	0.5479452

The reciprocity calculated for four states for the extent of how much researchers exchange information mutually. From 2015 to 2016, the reciprocity value decreased for Florida for about 23%, 45% for Vermont, and 10% for South Dakota. It increased for Kentucky for 25%.

For the "How do you Know" dataset:

Reciprocity	2015	2016
Florida	0.5272727	0.7248322
Kentucky	0.419214	0.5824176
Vermont	0.173913	0.3557692
South Dakota	0.8235294	0.601626

In terms of the reciprocity value of "How do you Know dataset", it is surprising to see

that the value increased for Florida and Kentucky, especially for Vermont. The value decreased for South Dakota.

Centrality Values

Degree

The degree calculation is conducted for all the datasets. The nodes with the highest degree, which are also the most important and active researchers in both information exchange and how much other researchers know each other for four states are shown in the following table. We find that the important nodes, nodes with high degree, are quite different from 2015 to 2016, which means the important researchers don't remain the same in the network.

The combined dataset:

Degree	Highest in 2015	Highest in 2016
Florida	FL 6, FL 25, FL 4	FL51,FL 1,FL 6
Kentucky	KY12, KY2, KY4, KY6	KY56, KY12,KY 31
Vermont	VT7, VT3, VT9	VT48, VT16, VT9
South Dakota	SD9, SD2, SD10, SD16	SD15, SD10, SD28

"How do you know" dataset:

Degree	Highest in 2015	Highest in 2016
Florida	FL 6, FL 25, FL 4	FL51,FL 1,FL 33
Kentucky	KY2, KY12, KY4, KY31	KY12, KY31,KY 24
Vermont	VT30, VT29, VT24...	VT16, VT5, VT38
South Dakota	SD3, SD9, SD16	SD10, SD15, SD14

We also adjusted the size of the vertex by the degree, and it is easier to see what type of cancer researchers stood out from the crowd and have more frequent contact with other researchers. All the graphs are attached in the Appendix. For Florida and Kentucky, blue dots stood out, which means cancer researchers have tighter and more importance in the state. For Vermont and South Dakota, green dots, which are communicable disease researchers, comparably turn out to be more important and contribute more to the collaboration to the network.

Centrality Values for each dataset

The combined dataset:

Centrality	2015			2016		
Type	Closeness	Betweenness	Eigenvector	Closeness	Betweenness	Eigenvector
FL	0.03377716	0.1149	0.8023575	0.05921311	0.1457559	0.6529143
KY	0.02099296	0.06659276	0.8138593	0.08201232	0.1077288	0.6770147
SD	0.1085468	0.08902961	0.5421876	0.2280338	0.1356576	0.5959614
VT	0.3725413	0.1547053	0.738008	0.4988421	0.1069293	0.7041487

The closeness centrality of a node is a measure of centrality in a network, calculated as the sum of the length of the shortest paths between the node and all other nodes in the graph. The graph above shows that KY, SD and VT have an increase closeness from 2015 to 2016; FL's closeness decreases from 2015 to 2016. So we think the centrality increases overall based on the information exchange datasets.

"How do you Know" dataset:

Centrality	2015			2016		
Type	Closeness	Betweenness	Eigenvector	Closeness	Betweenness	Eigenvector
FL	0.06038159	0.159529	0.7824213	0.05817914	0.1405003	0.6302305
KY	0.02435238	0.04924679	0.7570716	0.08125356	0.1026685	0.6158726
SD	0.4456734	0.0714375	0.4915098	0.2262942	0.1175378	0.4780866
VT	0.0275859	0.01375913	0.9015827	0.4897462	0.116527	0.6312527

Betweenness centrality is a measure of centrality in a graph based on shortest paths. Betweenness of KY, SD and VT increased from 2015 to 2016. FL has a decreased betweenness from 2015 to 2016. This result is consistent with the former analysis. So we can conclude that centrality increased in KY, SD and VT indicating researchers collaborate more in 2016 and the intervention may have a positive effect.

ERGM

We conducted ERGM for 4 states both years to predict the probability of 2 nodes forming a tie. For the "How do you Know" datasets we only analyzed based on density and for the combined datasets we took the attributes into consideration and analyzed among different CancerCD groups. When we analyze the results we are focusing on every single parameter and controlling for other variables in the model.

For the combined dataset, first, we analyze parameter for edge. The probability to form a tie for FL in 2015 is 0.017 and 0.097 in 2016; for VT it's 0.07 in 2015 and 0.11 in 2016; for KY it's 0.038 in 2015 to 0.14 to 2016. These 3 states have an increasing probability of exchanging information between people from 2015 to 2016. It shows the opposite in SD. SD's probability decreased from 0.18 in 2015 to 0.14 in 2016. So we can say that the overall collaboration between researchers in 4 states from 2015 to 2016 has increased. However, the values of the probabilities are all quite low, which indicates people still don't communicate much, despite the increased probability.

Then we look at the parameters of 3 CancerCD groups. In state FL, Group 1 has a probability of 0.96 in 2015 and 0.91 in 2016; Groups 2 and 3 have infinite parameters in 2015, meaning that there are very few nodes in these two categories, and which form no within-group ties. The empty cells are what produce the -Inf estimates. We do have many missing values in these two groups, so it's hard to compare the values between 2015 and 2016 for FL. In state KY,

Group 1 has a probability of 0.88 in 2015 and 0.71 in 2016; Group 2 has a probability of 0.57 in 2015 and 0.72 in 2016; Group 3 has a probability of 0.33 in 2015 and 0.32 in 2016. Clearly researchers doing cancer research communicate more than those doing CD research, yet among the CD group, the collaboration probability increased and the cancer group decreased. In SD, Group 1 has a probability of 0.66 to 0.81, Group 2 has a probability of 0.92 to 0.83. Therefore, the CD group communicates better than the cancer group. In VT, the probability remains steady between Groups 1 and 2 and between years 2015 and 2016; Group 1 has a probability of 0.74 in 2015 and 0.61 in 2016, while Group 2 has a probability of 0.74 in 2015 and 0.60 in 2016. We have come to a conclusion that the probabilities of collaboration among the three CancerCD groups in 4 states are quite high, indicating that researchers communicate very well in a certain group. But it's difficult to say if the intervention (meetings) in 2015 had made a positive impact since there are slight increases and decreases from 2015 to 2016 within different groups and different states.

For the "How do you Know" datasets, we only analyzed density. In FL, the probability of forming a tie between two nodes is 0.098 in 2015 and 0.13 in 2016; in KY, it's 0.06 in 2015 and 0.14 in 2016; in SD, it's 0.32 in 2015 and 0.27 in 2016; in VT, it's 0.03 in 2015 and 0.17 in 2016. So 3 out of the 4 states have an increased probability from 2015 to 2016 and, considering the density of 4 states, we find that SD has the highest density in both years and also has the highest probability. Also, the probability trend is consistent with the density trend. Clearly there's a positive relationship between density and probability, researchers who know each other better tend to have a higher probability of developing collaboration.

In summary, the overall communication probability in each state has increased but is at a low level. The collaboration probability is quite high among certain CancerCD groups but it increases or decreases half of the time among all the groups in four states. So we believe there's a positive impact of the meeting that benefits the collaboration but the impact is not quite significant.

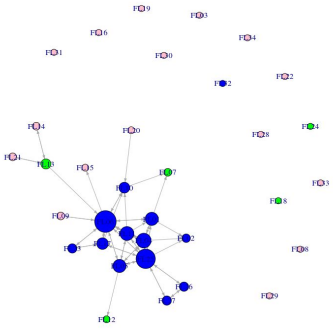
Conclusion

After conducting all the tests, we can't make a definitive conclusion but in general there was a slight improvement in collaboration from 2015 to 2016. There's an overall increase in centrality and density in four states from 2015 to 2016 and the ERGM results indicate that the overall collaboration probability has increased as well. However the increased amounts are not significant and we can't be sure if it's the result of the intervention.

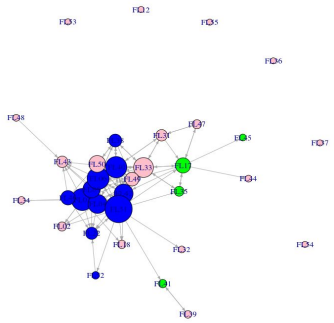
So our advice is that the meeting in 2016 should arrange more interactive activities to help the collaboration between researchers and to try to strengthen the connection between different CancerCD groups like workshops specifically for group cancer and group CD to communicate and discuss. We also think activities targeting the enhancement of the connection

between researchers in FL are needed since the collaboration in FL is the lowest in 4 states, knowing from the probability based only on density in the “send and receive information” combined datasets.

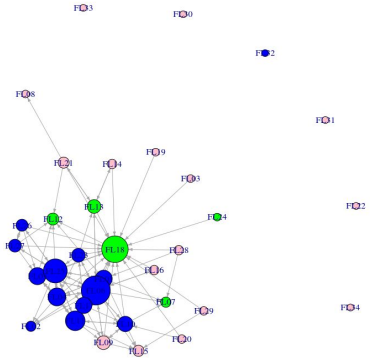
Appendix



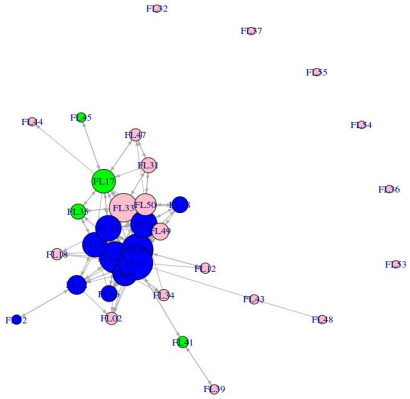
FL combined 2015 with CancerCD and degree



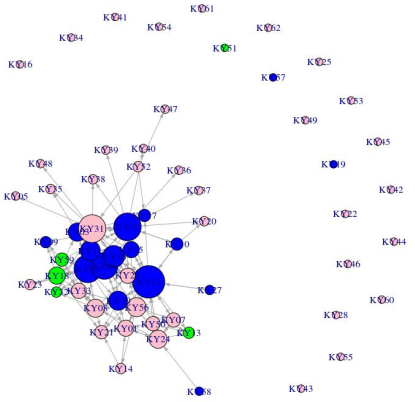
FL combined 2016 with CancerCD and degree



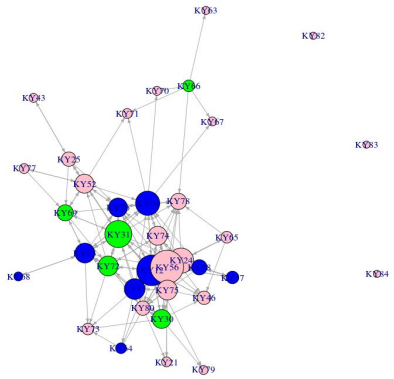
FL how 2015 with CancerCD and degree



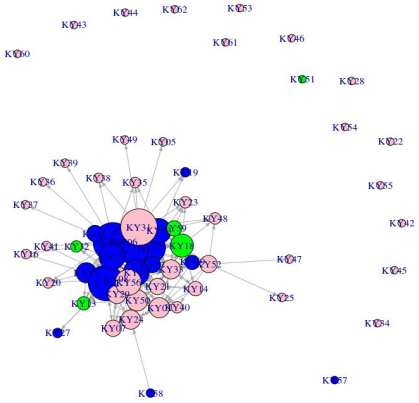
FL how 2016 with CancerCD and degree



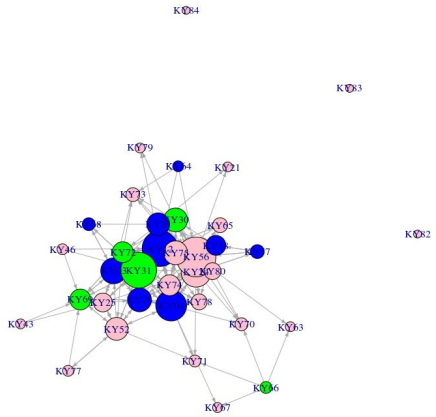
KY combined 2015 with CancerCD and degree



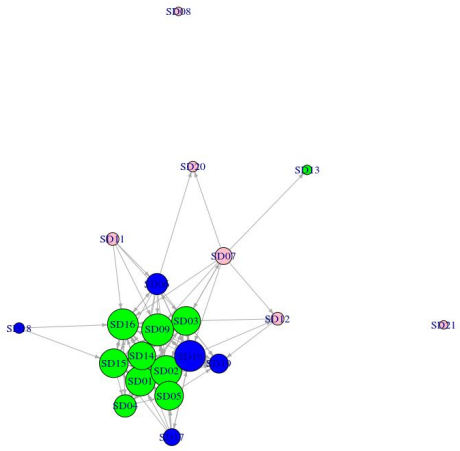
KY combined 2016 with CancerCD and degree



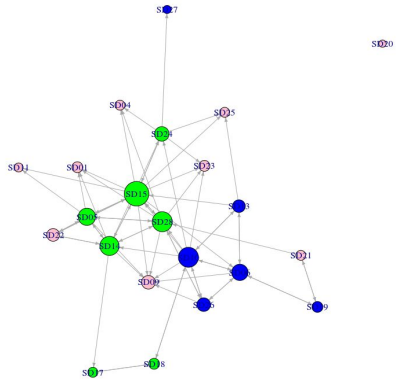
KY how 2015 with CancerCD and degree



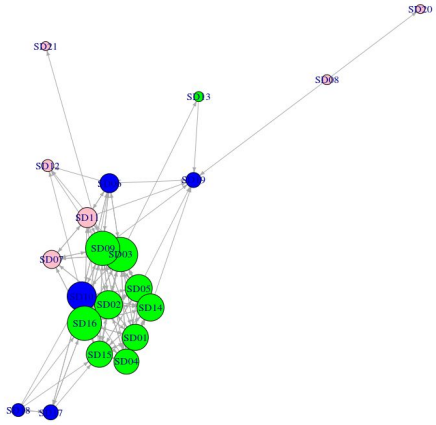
KY how 2016 with CancerCD and degree



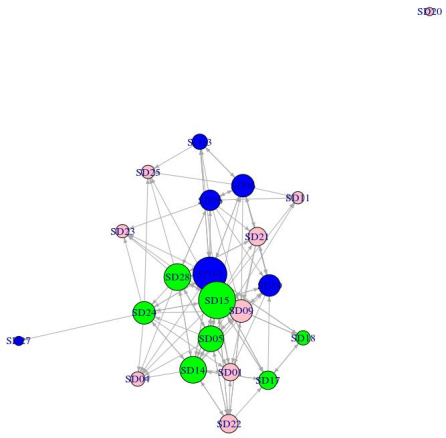
SD combined 2015 with CancerCD and degree



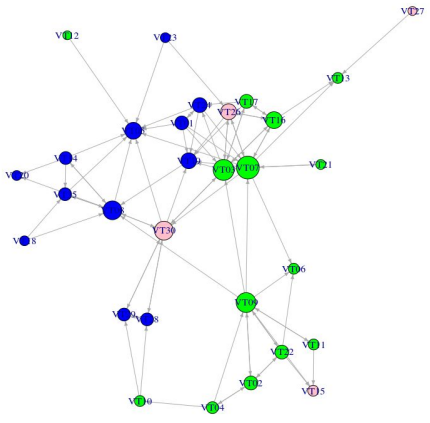
SD combined 2016 with CancerCD and degree



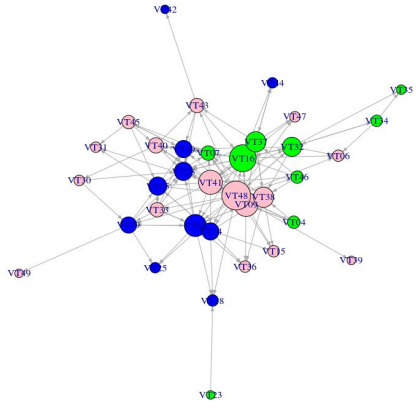
SD how 2015 with CancerCD and degree



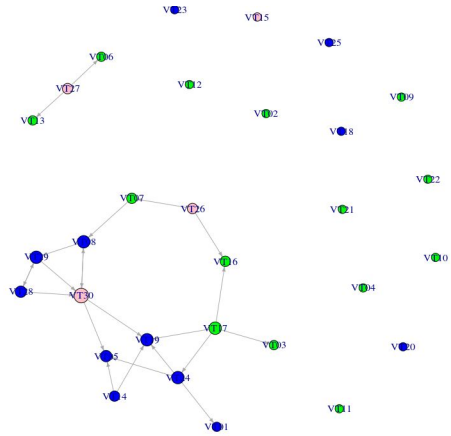
SD how 2016 with CancerCD and degree



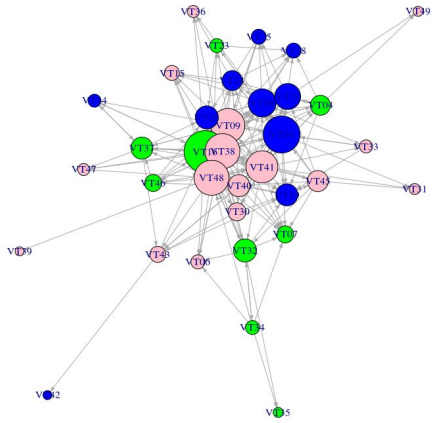
VT combined 2015 with CancerCD and degree



VT combined 2016 with CancerCD and degree



VT how 2015 with CancerCD and degree



VT how 2016 with CancerCD and degree