Extension Network Analysis

Deanna Schneider

March 20, 2019

Background and Data Collection

University of Wisconsin-Madison Division of Extension has 676 employees located throughout the state. Our "out-state" employees (those not located on a campus), work in one of 22 geographically-defined administrative areas.

To complete this analysis, I surveyed staff from 4 of the areas: 1, 2, 7 and 10. These 4 areas are all located in the northern half of the state but are geographically separated in 2 clusters (Figure 1). Areas 1 and 2 are in the far northwestern corner of the state, while Areas 7 and 10 are in the east-central part of the state. Each area in the state is comprised of 1 to 6 counties or tribal nations. The specific counties included in this analysis were:

Area 1:

- Ashland
- Bayfield
- Douglas
- Iron

Area 2:

- Barron
- Burnett
- Rusk
- Sawyer
- Washburn

Area 7:

- Clark
- Marathon
- Portage
- Wood

Area 10:

- Calumet
- Outagamie
- Waupaca
- Winnebago

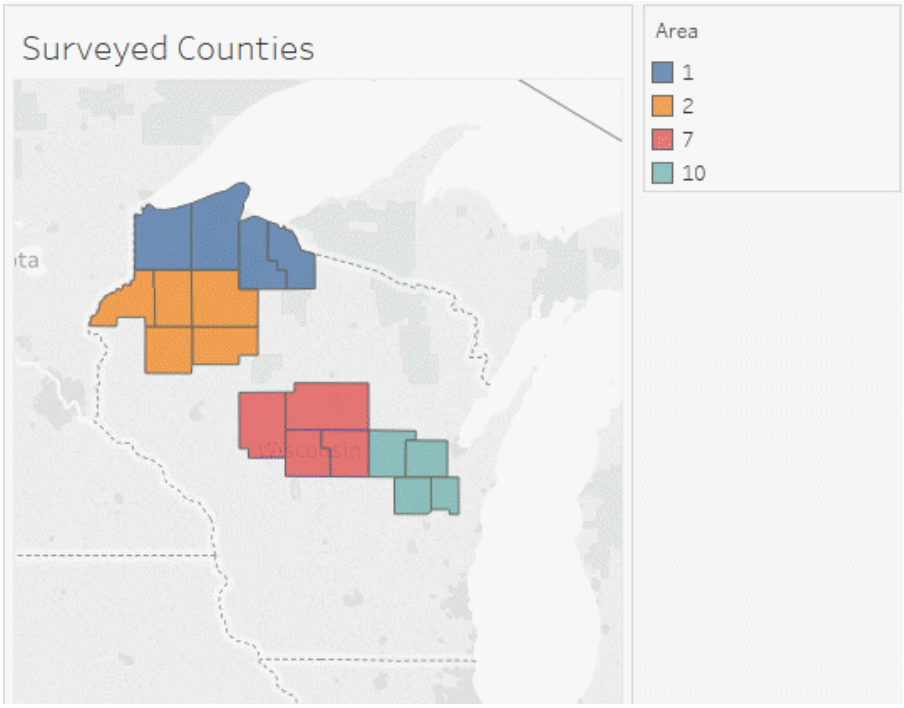

Figure 1- Surveyed Areas

These areas were chosen based on the recommendations of Extension leadership.

I surveyed a total of 80 employees, with 51 responding for a 63.8% response rate. I asked each employee to identify the level of communication and collaboration they engage in with the remaining 79 employees. I used the following definitions:

Communication: any direct written or verbal contact. This could include emails, chat or text messages, phone calls, or in-person conversations.

Collaboration: actively working together on something, with a common goal. This could include coauthoring documents, teaching together, writing curriculum together, working on evaluations together, co-facilitating a meeting, or planning an event.

For each, I asked them to rate their frequency of engagement, using the following scale:

- **Never**: You do not communicate directly or collaborate with this person.
- **Sometimes**: You communicate or collaborate a few times a year to a couple of times per month.
- **Often**: You communicate or collaborate every week or daily.

After the data was collected, I cleaned the data in Excel to have 2 socio-matrices, one for collaboration and one for communication. Additionally, I extracted data from our Human Resources System and Zoom collaboration platform for each employee surveyed. I included a total of 14 node attributes in the network.

- Department: The top-level programmatic affiliation
- Institute: Programmatic affiliation with one of 6 institutes or Administration
- Program: Programmatic affiliation, hierarchically below institute
- PA: Historical programmatic affiliation
- Employee_Class: The type of employee (Faculty, Academic Staff, or Limited)
- FTE: percentage appointment (1 = full time appointment)
- Sex: Identified sex
- Jobcode: Official title code
- Area: Geographic area (1,2,7,10) in which the person works
- Location: County in which the person works
- YearsInJob: Number of years in this appointment
- ZoomMeetings: The number of Zoom online meetings initiated by this employee since Zoom rollout
- ZoomParticipants: The number of participants in meetings initiated by this employee since Zoom rollout
- ZoomMinutes: The total number of meeting minutes for meetings initiated by this employee since Zoom rollout

I made the network non-directed, under the assumption that if colleague A collaborates with colleague B, colleague B is also collaborating with colleague A. Assuming that the network is non-directed lets us use all 80 of the nodes and retain an N x N socio-matrix, filling in the blanks by symmetrizing the matrix. If 2 colleagues scored their level of engagement differently, I retained the higher of the 2 scores. In an ideal situation, I would have had the time to clean up those discrepancies. But, for simplicity, I chose to make the matrix symmetrical and maximized the level of collaboration and communication.

Basic Network Statistics

When working with any network, it always makes sense to start with some basic descriptive statistics: size, density, number of components, size of largest component, total isolates, diameter, and the clustering coefficient. Our network size is 80 nodes and 339 collaboration edges. The rest of the statistics are in the table below:

Basic Description of Extension Network

Density	Number of Components	Size of Largest Component	Number of Isolates	Diameter	Transitivity
0.1072785	2	79	1	6	0.3348214

Density is the proportion of observed ties to possible ties, and it ranges from 0 to 1. Our graph has a density of .107, which suggests that just 10% of all possible ties exist in our network. While this doesn't sound like a very connected network, 2 of our other descriptive statistics give us a different picture.

First, our network is made up of just 2 components. Components are groups of nodes where all members are connected, either directly or indirectly. In our network, one of our components has 79 members, while the other has just 1. That single member is known as an isolate - a node that has no connections. Secondly, our diameter is 6, meaning that from any node, you can reach any other node in 6 or fewer steps. Our network (minus the isolate) is the real-life enactment of the 6 Degrees of Kevin Bacon game.

Finally, our clustering coefficient (also known as transitivity) is the proportion of closed triangles to the total number of open and closed triangles. In other words, this measures how often Employee A collaborates with Employee B and Employee C, and Employee B and C also collaborate with each other. Our network has a transitivity of .334, suggesting a moderate amount of clustering.

Visualizing the Network

One of the goals of our recent reorganization was to foster more collaboration outside of our local offices and programmatic areas. We can use visualization to get a sense of how people collaborate: by our new institutes, by location, and by area.

Institute Affiliation

In Figure 2, node size is proportional to the number of degrees. Black lines represent frequent collaborations and grey lines represent infrequent collaborations. Node color represents institute affiliation. Leadership hopes there will be collaboration not only within institute, but across areas.

Visually, it's challenging to pick out what's going on here.

To make it easier to see what's happening, we can separate out the 2 levels of connection (sometimes and often). The following graph (Figures 4) shows each set of edges separately but maintains the node positions in the same place. Nodes are not sized by degree in this visualization.

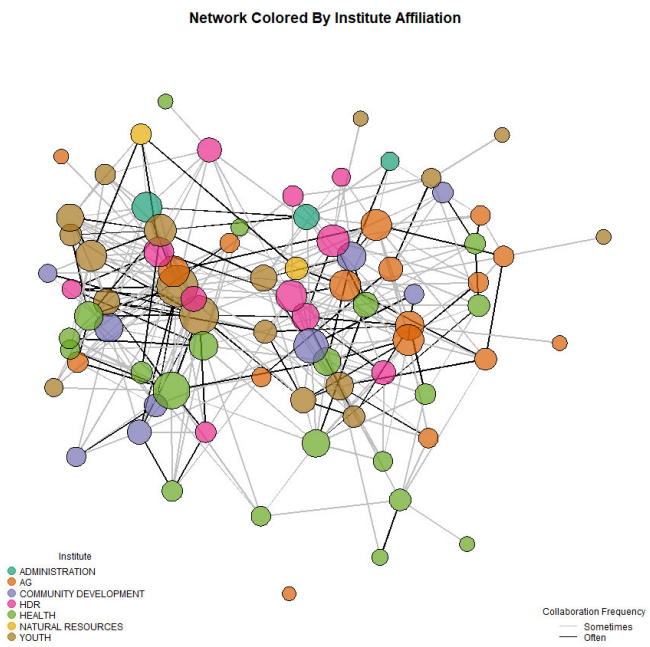

Figure 2- Extension Network by Institute Affiliation

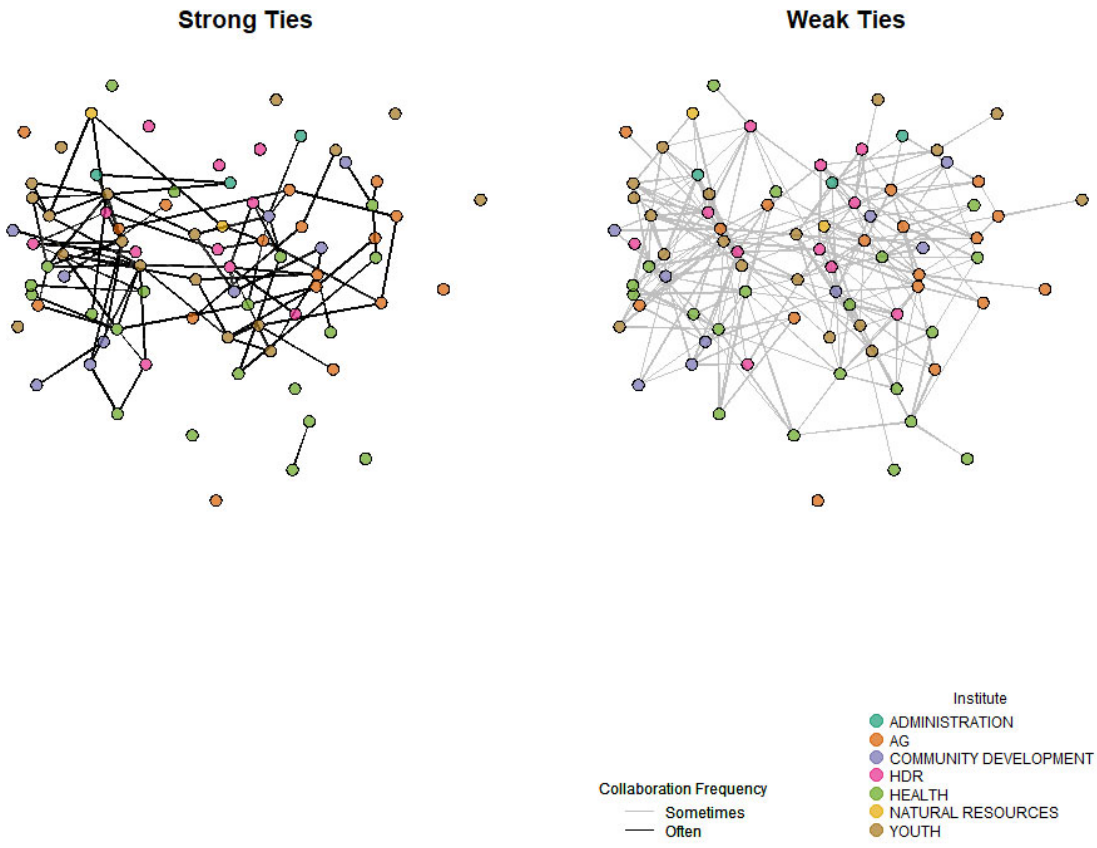

Figure 3- Strong and Weak Ties by Institute Affiliation

When you isolate the frequent collaborators, you can begin to see some tendency for matching institutes to collaborate. For instance, you can see the 3 orange (AG) nodes on the right side of the graph, and a cluster of 11 brown (YOUTH) nodes that are all interconnected. You can also see a handful of scattered green (HEALTH) nodes that are connected with each other, as well.

You can also see intra-institute matches among the weak collaboration ties, particularly among the youth and agriculture institutes. Here, we also pick up at least one triangle among colleagues in the Community Development institute, at the bottom left.

Location

Each person in the dataset works in a local county office. It's reasonable to assume that we might see strong collaboration within the local county office level. Looking at Figure 4, it appears that there is some amount of in-county collaboration happening, but it does not appear that county borders are limiting collaboration either.

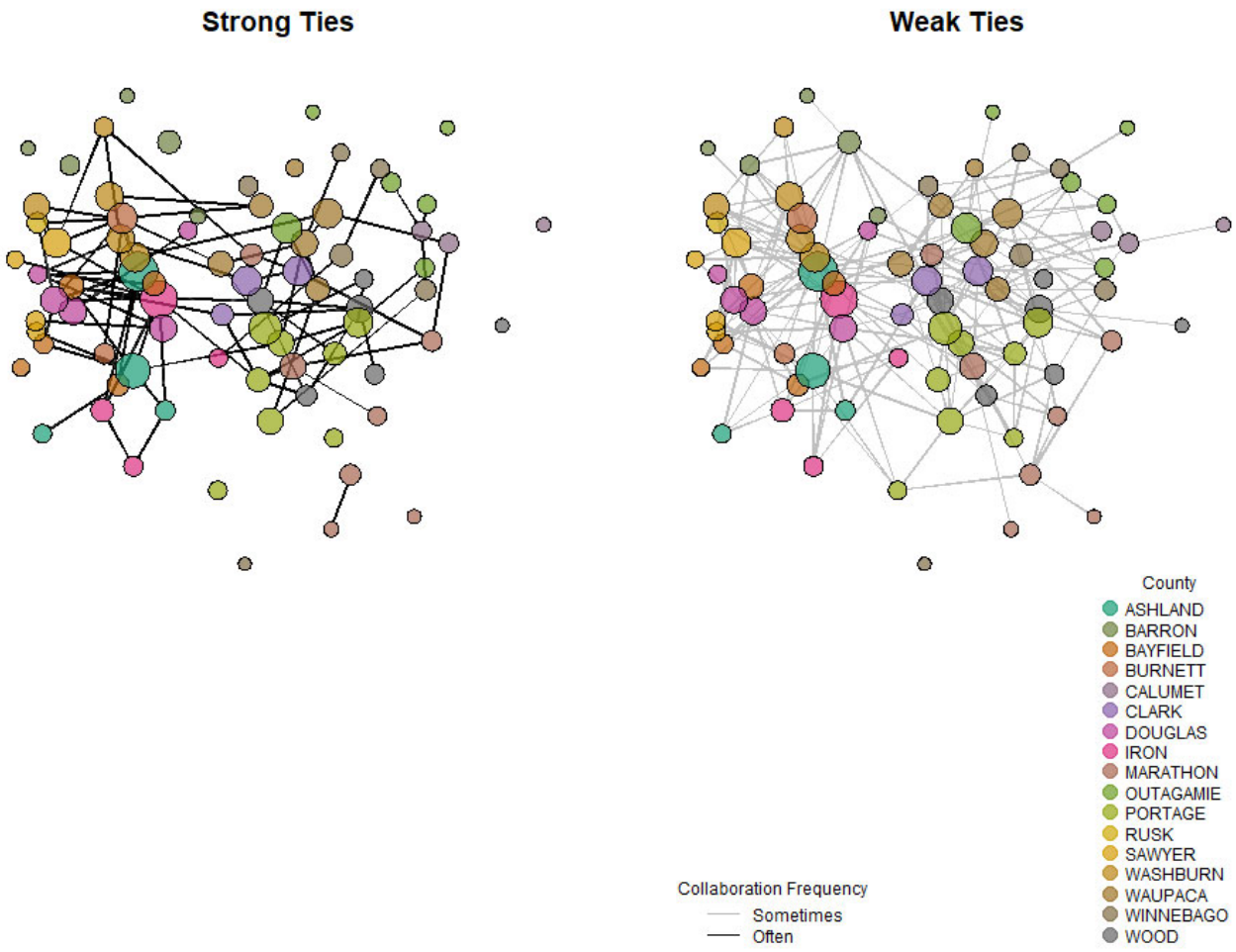

Figure 4 - Strong and Weak Ties by County Affiliation

Area Affiliation

Finally, we have 4 geographical areas. We can visualize the network with nodes colored by these four areas (Figure 5). Area appears to have the most visual cohesion. Additionally, there seems to be fairly strong cohesion within the 2 geographical clusters (areas 1 and 2 vs. areas 7 and 10), though there are multiple connections that bridge that gap.

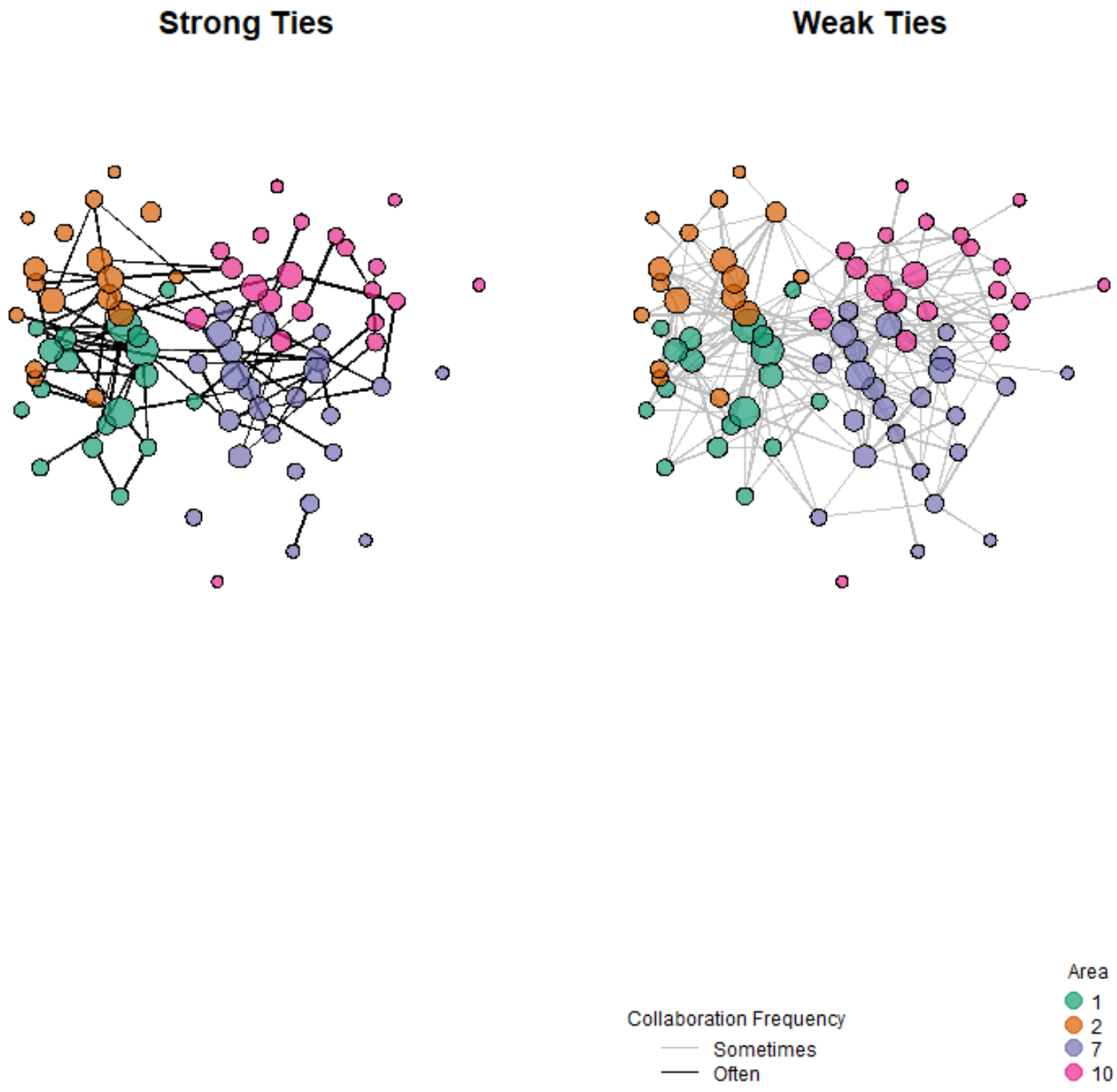

Figure 5 - Strong and Weak Ties by Area Affiliation

Actor Prominence

Another way to visualize the network is to look at who our prominent or important actors are (Figure 6). There are multiple ways to determine importance. We'll look at two: Degree centrality and Betweenness Centrality.

Degree Centrality identifies the actors with the most connections to other people. Having multiple connections means that these people have multiple ways to receive information and accomplish collaborations. We have nodes with degrees ranging from 0 to 24. I rescaled all degrees to be between 1 and 6 and chose prominent actors as those with a rescaled degree above 5.

With *Betweenness Centrality*, we are most interested in the people that sit "between" parts of our network. These are the people that are key for information flow. Without these people, we would have a harder time disseminating information. For Betweenness, we have node scores between 0 and 636. Again, I rescaled the Betweenness scores between 1 and 6, and identified prominent actors as those with a rescaled Betweenness score above 5.

The numbers on the graphs below indicate the ID of the individual at that node. You can see that slightly different people are identified as important using each method.

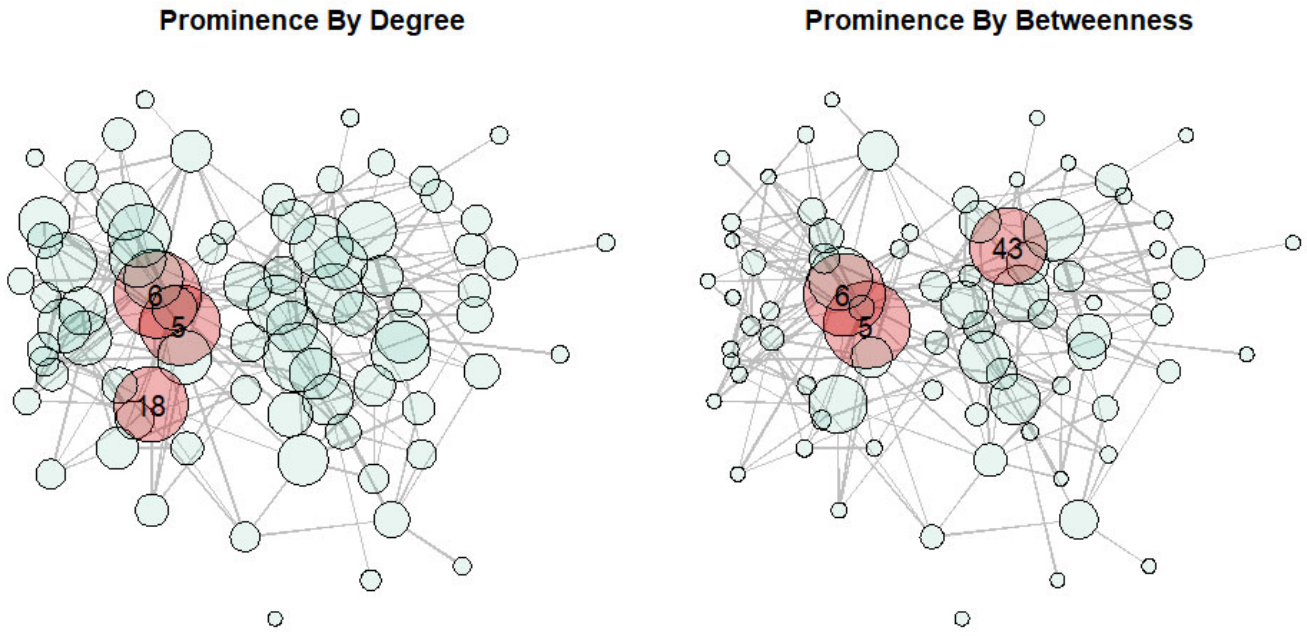

Figure 6 - Actor Prominence by Degree and Betweenness

Characteristics of Important Actors

ID	Area	Location	Department	Institute	Sex	Years in Job
5	1	IRON	YFCD	YOUTH	M	6.69
6	1	ASHLAND	YFCD	YOUTH	M	13.32
18	1	ASHLAND	YFCD	HEALTH	F	0.1
43	10	OUTAGAMIE	YFCD	HDR	F	28.53

Notable is that 3 of our important actors are from the same area, and all of them are from the same department. There is a broad range of years in job (though this is slightly misleading in that the person with a short time in job held a similar position in Extension for the 5 years prior).

If we wanted to do further analysis, we could reach out to these people to do focused interviews about their collaboration approaches to help understand what makes them key players in our network.

Identifying Subgroups

While we could continue looking at various plots to try to visually assess network patterns, our time is better spent using the tools in the igraph package to look for specific types of subgroups.

Clique Identification

Typically, cliques are rare. A clique requires all possible connections in a set of nodes to be there. If even one connection is missing, the clique doesn't exist. For example, for a group of 7 nodes, all 21 possible ties must exist between all seven nodes.

Our network has not 1 but 2 7-node cliques.

Figure 7 compares our network to 1000 randomly generated models using the same number of nodes and edges, 1000 small world models using the approximate degree distribution of our observed network, and 1000 models using the approximate power law of our observed network.

The red horizontal line in each of these graphs indicate our observed network's largest clique and total cliques. You can see that our observed network has far more cliques and larger cliques than expected in either a random network, small world network, or power law network.

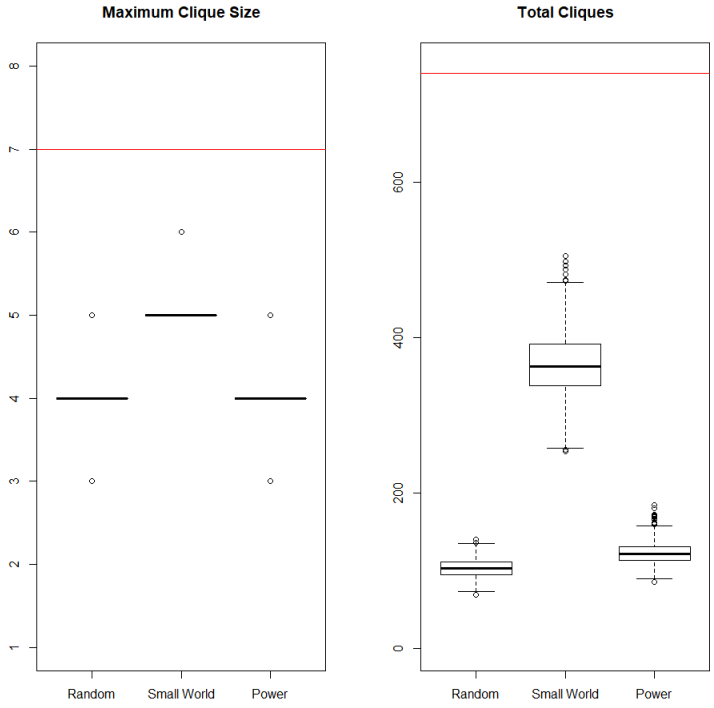

Figure 7 - Clique Comparison Between Observed Network and Network Models

Modularity

Modularity is a measure of the network structure in which there is greater density within a cluster than outside of it. We can look at the various node attributes within our network to see which ones show the highest modularity score. As seen earlier in our visualizations, Area shows the highest modularity. Institute and Location are the next 2 highest modularity scores, but they are quite a bit less modular than Area.

Modularity scores for node attributes of the Extension Network.

	Score
Area	0.4386622
Institute	0.2415137
Location	0.2169186
Program Area	0.1897869
Program	0.1294628
Department	0.1181203
Jobcode	0.0261789
Sex	0.0114165

Automated Community Detection

There are several ways we can do automated community detection using igraph. A weighted, non-directed graph can be tested with Edge-betweenness, Fast-greedy, Louvain, and Infomap algorithms. We can compare these automated approaches to our manual approach of using Area to identify a community.

Community Detection Results for the Extension Network

	Area	Edge Betweenness	Fast Greedy	Louvain	Infomap
Modularity	0.439	0.401	0.412	0.433	0.425
Total Communities	4.000	12.000	6.000	5.000	6.000

Each of these ways uses a slightly different algorithm but results in a similar number of communities - between 4 and 6 except for the Edge Betweenness algorithm which defines 12. The modularity scores are similar (between .401 and .439). The Louvain method achieves the highest level of modularity of the automated algorithms, but simply using Area to define our communities has a higher modularity score.

The 4 visualizations in Figure 8 compare each community (excluding the least modular Edge-betweenness Community) in decreasing order of modularity. You can see that they are quite similar.

Given that the Area communities have the highest modularity score, they are the most logical choice if we were interested in doing community analysis. These communities are understandable and recognizable for the people in Extension, and it is not surprising that they are cohesive communities given the emphasis placed on intra-Area collaboration.

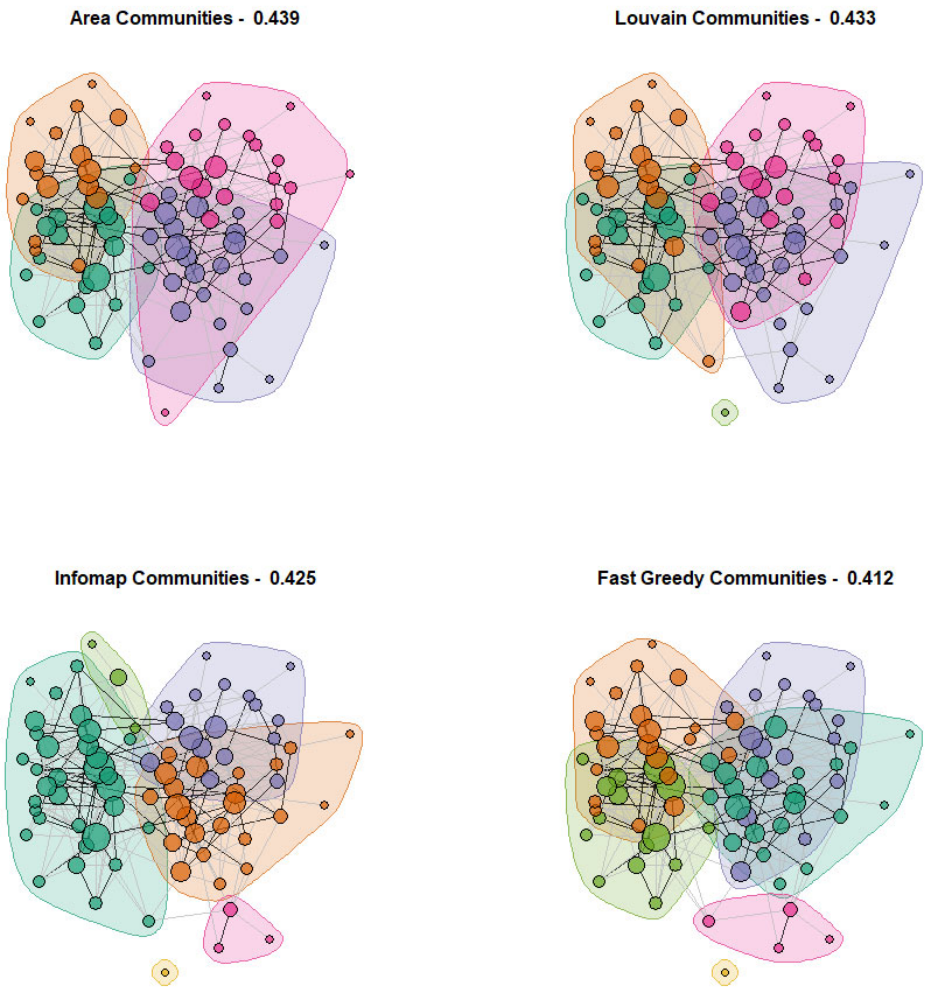

Figure 8 - Comparison of Possible Communities

Modeling

Statistical Network Models

ERGM (exponential random graph models) can be used to build and test hypotheses about networks. ERGMs can help us test hypotheses about both nodes and the overall network, such as diameter or degree distribution.

I created a total of 10 models, testing most of the available parameters to determine their significance as it related to the likelihood that 2 colleagues would form a collaborative relationship. A detailed discussion of each model is included in the accompanying code. Here, I will focus on just the final model, the output of which is below.

```
==========================
Summary of model fit
==========================

Formula:   cnet ~ edges + nodecov("ZoomMinutes") + nodecov("ZoomParticipants") +
    nodefactor("Employee_Class") + nodematch("Employee_Class") +
    nodematch("Area", diff = T) + nodematch("Location") +
    nodematch("Department") + nodematch("Institute") +
    edgecov(contactNet, attr = "weight") + gwesp(1.2, fixed = TRUE)

Iterations:  3 out of 20

Monte Carlo MLE Results:
                              Estimate   Std. Error  MCMC % z value  Pr(>|z|)
edges                      -6.0198056    0.1235571      0 -48.721  < 1e-04 ***
nodecov.ZoomMinutes         0.0015947    0.0006077      0   2.624  0.00868 **
nodecov.ZoomParticipants   -0.0698162    0.0270832      0  -2.578  0.00994 **
nodefactor.Employee_Class.FA  0.6368449  0.0770620      0   8.264  < 1e-04 ***
nodefactor.Employee_Class.LI  0.8557039  0.1891760      0   4.523  < 1e-04 ***
nodematch.Employee_Class   0.3055619    0.1230498      0   2.483  0.01302 *
nodematch.Area.1           1.6459646    0.1513240      0  10.877  < 1e-04 ***
nodematch.Area.10          1.3329971    0.1500353      0   8.885  < 1e-04 ***
nodematch.Area.2           1.6425401    0.1568823      0  10.470  < 1e-04 ***
nodematch.Area.7           1.1366766    0.1340088      0   8.482  < 1e-04 ***
nodematch.Location         1.9273653    0.1841571      0  10.466  < 1e-04 ***
nodematch.Department       0.2871169    0.1312750      0   2.187  0.02873 *
nodematch.Institute        1.7435603    0.1382755      0  12.609  < 1e-04 ***
edgecov.weight             0.3389310    0.1067891      0   3.174  0.00150 **
gwesp.fixed.1.2            0.5946841    0.0507851      0  11.710  < 1e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null Deviance: 4381  on 3160  degrees of freedom
 Residual Deviance: 1267  on 3145  degrees of freedom

AIC: 1297    BIC: 1388    (Smaller is better.)
```

Choosing Predictors

Node Predictors

For the 2 Zoom-related parameters, our educational technology team asked me to test whether the number of Zoom meetings, length of Zoom meetings, and total Zoom participants were significantly associated with likelihood that colleagues would form collaborative ties. The number of Zoom meetings is

not significantly associated, but the duration and number of participants are both significantly associated. The duration is positively associated, and the number of participants is negatively associated. This suggests that longer meetings with fewer participants may be more conducive to collaborative relationships.

Our faculty members have a formalized mentorship program and collaboration is a stated job expectation. I hypothesized that Employee Class would have a significant relationship with collaboration. In fact, it does, meaning that both faculty and limited staff are more apt to collaborate than academic staff.

Dyadic Predictors

Not only are faculty and limited staff more apt to collaborate than academic staff, but two employees of the same class are more likely to collaborate with each other than with someone from a different employee class. This was verified using a node match predictor on employee class.

We had already seen strong modularity by Area and programmatic affiliation. Indeed, each of these node match parameters were positively and significantly associated with the likelihood that 2 individuals would collaborate.

Relation Predictors

We had a secondary network using the same nodes – the frequency of communication between individuals. It is reasonable to assume that more frequent communication would lead to more collaboration. Our model shows that frequency of communication is positively and significantly associated with collaboration.

Structure Predictors

Finally, we had seen earlier that our network had many cliques. The geometrically weighted edgewise shared partners (GWESP) predictor measures the effect of employees having shared edgewise partners. For instance, if I collaborate with Jane, and you collaborate with Jane, does that make us more likely to collaborate with each other, because we share a common link (Jane).

This predictor is a bit complicated in that it requires an alpha parameter. The alpha parameter discounts the effect of shared partners as the number of shared partners increases. The closer the alpha is to zero, the more dramatic the effect is discounted. I tested this model with a variety of alpha parameters and found that 1.2 gave the best results in terms of both AIC and Goodness of Fit measures. This suggests that the effect of multiple shared partners is relatively strong in our network.

An additional structure predictor was also tested Geometrically Weighted Degree Distribution (GWDEG). While this predictor was significant, it did not improve the model (the AIC increased). I removed it for a more parsimonious model.

Goodness of Fit

Before using a model for prediction, it's important to validate the goodness of fit of the model. Our model shows a very good fit. Out of all the statistics tested, only 2 show a p-value under .05 - Degree of 6 and edgewise shared partners of 10.

We can look at the results of these tests with plots in Figure 9. We can see that the average value generated by our model is well within the confidence bounds for most of the statistics tested, suggesting that our model fits our data well.

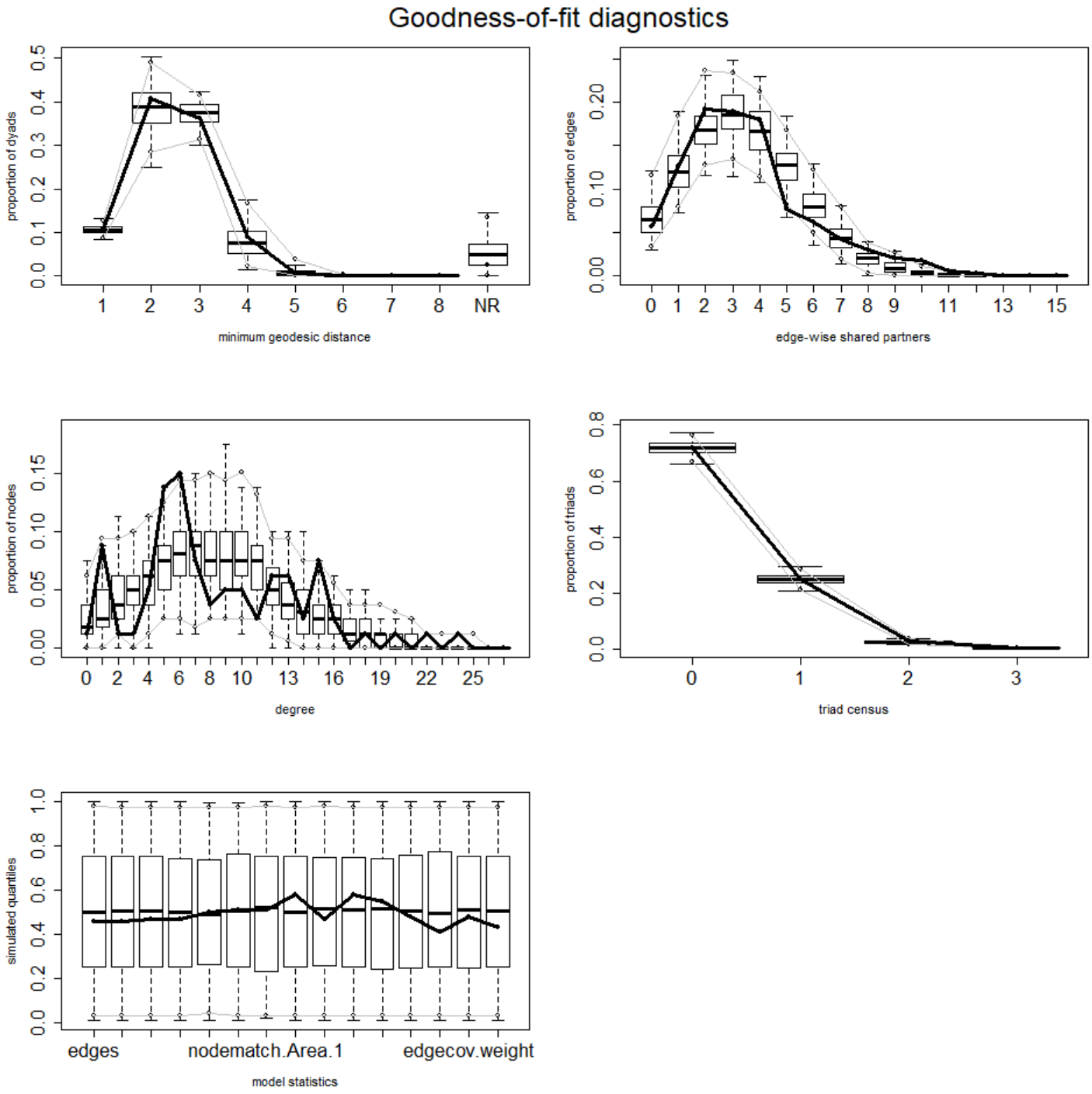

Figure 9 - Diagnostic Plots Indicate Good Fit

Estimating Likelihood of Collaboration

With a well-fitted model, we can estimate the likelihood of collaboration between any two employees. We do this by applying a logistic function to the sum of the coefficients estimates in the model for our example nodes.

For example, we can calculate the likelihood that the following 2 employees will collaborate.

Employee 1:

- Faculty member in the Agriculture and Natural Resources Department
- Agriculture Institute
- Area 10
- 1000 minutes in Zoom meetings with 20 participants

Employee 2:

- Faculty member in the Agriculture and Natural Resources Department
- Natural Resources Institute
- Area 2
- 100 minutes in Zoom meetings with 2 participants

These 2 Faculty members share 3 edgewise partners and have contact often, resulting in the following formula:

```
plogis(-6.0198056 + #edges
         2*0.6368449 + #faculty
         0.3055619 + #both faculty (node match)
         0.2871169 + #matching departments
         1100*0.0015947 + #zoom minutes
         22*-0.0698162  + #zoom participants
         0.5946841*(1-exp(-1.2)^3) + #edgewise shared partners = Cgwesp * (1-exp(-alpha)^ESPij)
         2*0.3389310 #contact often (2)
         )
```

The results of this formula suggest that these 2 colleagues would be 6.4% likely to collaborate. Our overall network density is 10.7%, which suggests that these two colleagues are less likely to collaborate than might be expected in our network.

If these two colleagues were in the same area (say, both in Area 10) and the same county, we would use the following formula:

```
plogis(-6.0198056 + #edges
         2*0.6368449 + #faculty
         0.3055619 + #both faculty (node match)
         0.2871169 + #matching departments
         1100*0.0015947 + #zoom minutes
         22*-0.0698162  + #zoom participants
         0.5946841*(1-exp(-1.2)^3) + #edgewise shared partners = Cgwesp * (1-exp(-alpha)^ESPij)
         2*0.3389310 + #contact often (2)

         1.3329971 + #matching Area 10
         1.9273653 #matching location
         )
```

These two colleagues would be 64.1% likely to collaborate, which is much more than we would expect based on our overall network density.

Estimating Odds

This model gives us the following odds ratios:

Odds Ratios for Model 8

	Lower	OR	Upper
edges	0.0019	0.0024	0.0031
nodecov.ZoomMinutes	1.0004	1.0016	1.0028
nodecov.ZoomParticipants	0.8844	0.9326	0.9834
nodefactor.Employee_Class.FA	1.6258	1.8905	2.1983
nodefactor.Employee_Class.LI	1.6251	2.3530	3.4070
nodematch.Employee_Class	1.0668	1.3574	1.7272
nodematch.Area.1	3.8567	5.1860	6.9736
nodematch.Area.10	2.8273	3.7924	5.0869
nodematch.Area.2	3.8018	5.1683	7.0259
nodematch.Area.7	2.3973	3.1164	4.0511
nodematch.Location	4.7911	6.8714	9.8548
nodematch.Department	1.0305	1.3326	1.7232
nodematch.Institute	4.3613	5.7177	7.4958
edgecov.weight	1.1385	1.4034	1.7300
gwesp.fixed.1.2	1.6408	1.8125	2.0021

These odds ratios represent the likelihood of a tie with respect to the reference group for a categorical variable. For example, there are three employee classes: Academic Staff (the reference category), Faculty, and Limited staff. These odds ratios suggest that a Faculty member is 1.9 times more likely to form a collaboration tie than an academic staff member. If both members of the tie are the same employee class, they are 1.4 times more likely to form a collaborative tie than members of different employee classes.

For numerical variables, the odds ratios represent the likelihood of a tie for a unit change in the predictor. For example, for each additional minute spent in Zoom meetings, employees are 1.00016 times more likely to form a collaborative tie.

Summary and Next Steps

By using descriptive statistics, visualizations, and modeling, we have gained knowledge about the collaboration network of Extension employees in Areas 1, 2, 7 and 10. In general, we have learned that there are strong collaborative ties within each area – strong enough that they form sub-communities that could be further analyzed. Additionally, we see that matching by location and institute are our strongest predictors of 2 employees collaborating, and that some areas are more prone to in-area collaboration than others (notably Area 1). Finally, the model suggests that academic staff are less prone to collaboration than faculty or limited staff.

We can use this information to adjust support opportunities around collaboration, for instance providing academic staff with more collaboration opportunities. Or, we could use this information as the basis for further research - either focused interviews with our important actors or additional in-depth analysis of the factors that allow greater faculty collaboration. Finally, this analysis supports continued data collection in our Zoom collaboration platform to see if the weak but present effect of smaller, longer meetings continues to hold true as we complete the Zoom rollout.

Next steps could include a larger survey, including more areas or an examination of the data as a directed network, to see if there are differences in perceptions about collaborative relationships.

References

Butts, C. T., & Hunter, D. R. (n.d.). Introduction to Exponential-family Random Graph (ERG or (p^*)) modeling with ergm. Retrieved from https://statnet.csde.washington.edu/trac/raw-attachment/wiki/Sunbelt2014/ergm_tutorial_abridged.html

Harris, J. K. (2014). An introduction to exponential random graph modeling.

Harrison, M. M. (n.d.). ColorRampPaletteAlpha. Retrieved March 9, 2019, from https://github.com/mylesmharrison/colorRampPaletteAlpha/blob/master/colorRampPaletteAlpha.R

Hunter, D. R. (2007). Curved exponential family models for social networks. Social Networks, 29(2). Retrieved March 19, 2019, from https://www-sciencedirect-com.ezproxy.library.wisc.edu/science/article/pii/S0378873306000396.

Luke, D. A. (2015). A User's Guide to Network Analysis with R. Cham: Springer International Publishing.