Friendship Patterns within Business Analytics Majors Report

Group5 Andrew Renga, Violet Chen, Jose Llamas, Chengzong Zhang, Maria Urdaneta, Alejandra Taboada Donado

2022-12-21

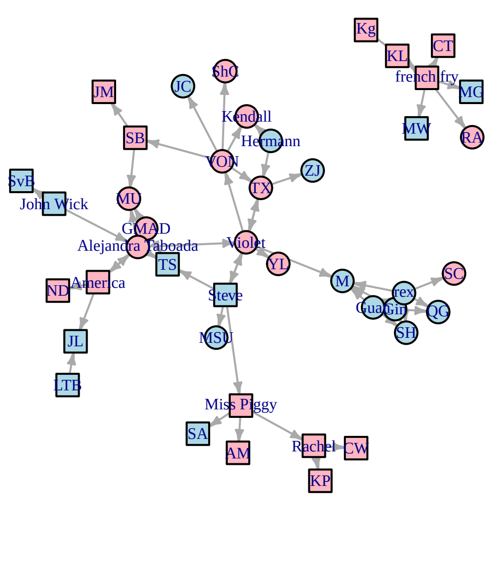
Introduction

The project aims to analyze the friendship patterns between international students and domestic students within Business Analytics majors. Data are collected by interviewing and sending questionnaires, which includes friendship level (1 to 4 scale where 4 = best friends), citizenship, hometown, major, graduation year, and gender. The data are collected on an edge list and paired with the attribute data gathered through a Google form sent to 3 Business Analytics class. The findings are expected to show that international students are more likely to make connections with people from their home country while domestic students do not necessarily follow the same pattern.

This report will provide an in-depth analysis of the degree distribution, centralities, correlation of centralities, subgroups, roles, brokerage, and social capital within the network. Additionally, the report will include the statistical models (ERGM) and interpretations testing for homophily. However, it is important to note that not every student participated in the survey which could have resulted in a less accurate representation of the results. Therefore, it is important to consider the potential limitations of the survey when interpreting the results.

Additionally, the project will explore whether there is a relationship between a domestic student network and an international student network based on citizenship and if it is possible to predict the probability of a connection in a network based on organizations using a network based on majors. The results of this project could provide valuable insights into the dynamics of friendship formation among international and domestic students, and could be used to inform strategies for fostering a more inclusive and diverse learning environment.

A plot of network is shown below with color light pink as female, light blue as male and shape circle as domestic student,square as international student.



Centralities

	degree	closeness	betweenness	eigenvector
Violet	8	0.0105263	129	0.1865989
Alejandra Taboada	8	0.0067114	69	0.0937720
Steve	5	0.0096154	60	0.0782656
Miss Piggy	4	0.0769231	45	0.0050051
VON	6	0.0073529	43	0.0471562
TX	5	0.0077519	38	0.0579650
Rachel	3	0.1666667	20	0.0003318
SB	3	0.2500000	14	0.0081154
America	4	0.0047847	6	0.0376121
french fry	6	0.1000000	5	0.0000000

After plotting the network and analyzing its centralities, we conclude the following:

The first centrality analyzed is *Degree Distribution*. Degree distribution looks at the probability that a randomly chosen node has degree k. *Degree* is the count of edges a node has; in other words, when looking at the network, an edge exists between two nodes if there is a frequent association between the students represented by those nodes. The students with the highest degree are Alejandra Taboada and Violet with a degree of 8. This, essentially, means that Alejandra and Violet have 8 connections with other students across the three classes, making them the most "popular."

We proceeded to compute the *closeness* centrality, which indicates how close a node is to all other nodes in the network. This centrality measures the mean distance from a vertex to other vertices, looking at the average of the shortest path length from the node to every other node in the network. In other words, this centrality measures the communication among networks. In the student network, student named LTB has the highest closeness centrality value at 0.5, meaning he is the best placed to influence the entire network quickly.

Moving on to *betweenness*: this metric computes the importance of nodes in a number of flow based networks. It is a measure of the extent to which a certain vertex lies on the shortest paths between other vertices. Vertices with high betweenness are considered to have considerable influence within a network. Student named Violet has the highest betweenness value at 129. This is expected because the same student scored very high for degree and is, therefore, popular, which means it has great influence within the network.

Then I looked at the *eigenvector* centrality, which measures actor influence or popularity: a high score means that an actor is connected to many nodes that also have high scores. When looking at the student network, student named Rex has the highest eigenvector centrality of 1.

	degree	closeness	betweenness	eigenvector
degree	1.00	NA	0.68	0.47
closeness	NA	1	NA	NA
betweenness	0.68	NA	1.00	-0.03
eigenvector	0.47	NA	-0.03	1.00

These findings led us to correlate these centralities, and we see a high correlation of 0.68 between degree and betweenness for the student network. This makes much sense because the student with the highest degree is Violet, which is also the student with the highest degree centrality. This makes sense because there is overlap between having the most connections and being popular.

Subgroup

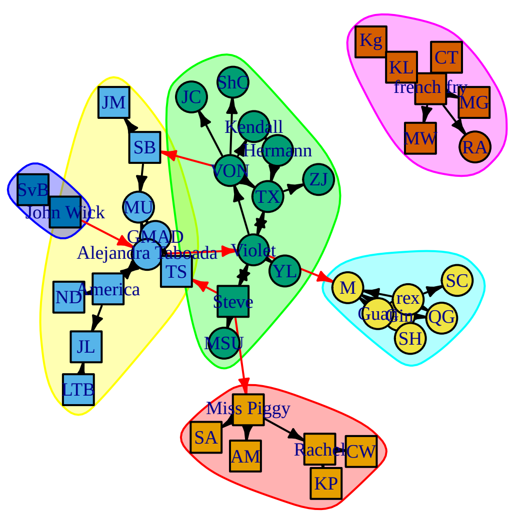
Observing subgroup allows us to study the relationship of social network, the individuals can be formed into groups. Since the community is a closely connected group, everyone will share friends more or less. Besides through an intermediary, two unrelated people can be connected together, Although such a connection may will not be too strong.

Based on our prediction, some people may connect to more than one person since they have differnet friendships. More links mean that they are centralized nodes in this social network. Subgroup allows us to see whether the relationship between the network is tight or sparse. Also if there is isolated communities which means it has no connection with other clusters. We predict that based on the collected data, the subgroup should show a strong correlation.

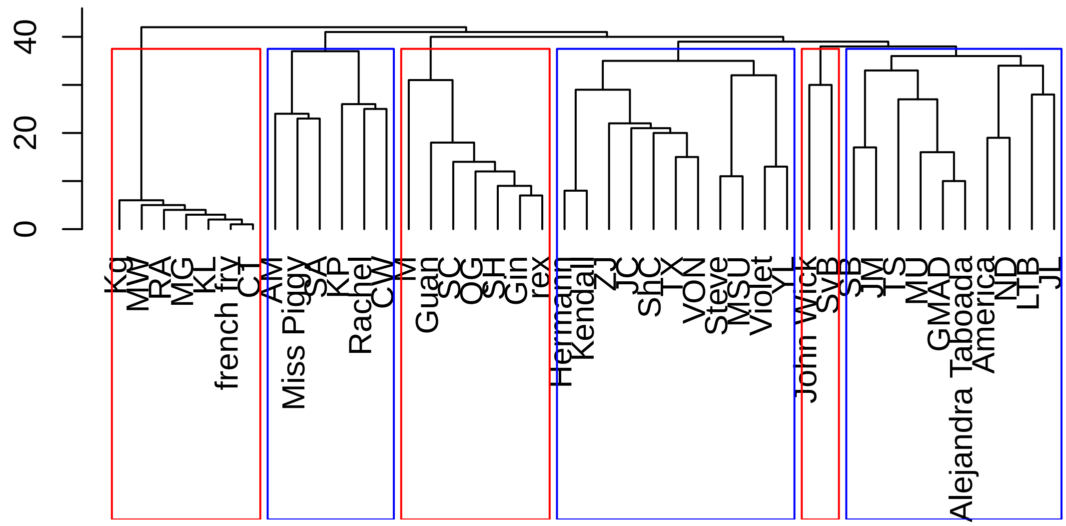
```
## Community sizes
##  1  2  3  4  5  6
##  6 10 11  7  2  7
```

The network is classified into six clusters. The output above shows the number of people and the relationship of each cluster. The fifth cluster only has has two people. Most of the other clusters have similar sizes. The largest is the third cluster which has 11 people. We also calculated that the modularity of this community network is 0.69. According to the concept of modularity, this value should be between -1 and 1. Firstly, it proves that our social network is applicable. Secondly, this result is close to 1, which means that our community network are relatively strong.

Then we use cross table to find the various combinations that already exist in the data and arrange them. It can be seen from the results that most of the combinations are false. Only six combinations are true which means that in this social network, they are the nodes could connect with the other clusters, people in combination have relationship and it helps to connect the different groups. Finally, we plot the graph of the subgroup, although the sizes of the six clusters are different, including the number of nodes and the degree of connection. The degree of connection between the second and third clusters is the largest in the entire community, because there are many two-way relationships, such as Violet and Alejandra both play relatively important roles in their respective sub-communities.



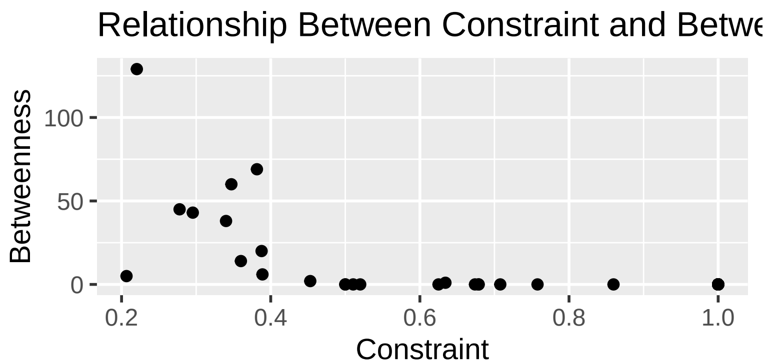
We also use dendrogram to express the structure of the community and the degree of connection. But cluster 6 has little or no relationship with the entire community structure, although they are connected to each other. The reason may be due to the incompleteness of the data and the reliability of the source, because we collected it through the survey. However, the community structure described by subgroup and the value of modularity can prove that our network is relatively strong.



The Constraint Function vs. Betweenness Function

	Constraint	Betweenness	Internationality
Violet	0.2204419	129	International
french fry	0.2066116	5	Domestic
Kg	1.0000000	0	Domestic
John Wick	0.6250000	0	Domestic
GMAD	0.7078000	0	International
Gin	0.6342679	1	International
Guan	0.6738933	0	International
SB	0.3600000	14	Domestic
Steve	0.3472222	60	Domestic
Miss Piggy	0.2777778	45	Domestic
TX	0.3400693	38	International
Hermann	0.5000000	0	International

When referring to *betweenness* centralities and the constraints of the nodes that make up the network, each have different meanings. Betweenness centrality locates the brokers with respect to all the other actors in the network. Burt's constraint however, is a local measure of brokerage based on the level of triadic closure. Referring to the nodes, students with high a betweenness centrality intern have a low constraint. Inversely, students with a high constraint have a low betweenness centrality. This makes sense, as nodes such as *GMAD*, *Guan*, and *Gin* are very constrained to the extent that they have invested a significant portion of their resources (in this case class connections) back into a small number of students. These students are very limited in other possible class connections as they are all only connected with a small number of other students in the entire network. Thus, their high constraint leads to low betweenness as they represent students who are separated from their peers in the network. This concept is supported by the following graph visualizing this relationship:



Burt defines social capital as a function of brokerage opportunities. Supporting his article, "Structural Holes and Good Ideas," families who act as bridges between clusters are the least constrained amongst the network and the most powerful brokers. For example, students such as *Alejandra Taboada* and *Violet* bridge the gap between the main network and students such as *ND*, *YL*, and many more who lack numerous class relationships with other students. Since their class connections are diversified, *Alejandra Taboada* and *Violet* are not constrained to one class connection. Rather, their multiple bridging relations signify their social power as students to connect with, as well as for other characteristics such as internationality and gender.

The Brokerage Function

```
##                     w_I w_O b_IO b_OI b_O  t
## Violet               13   0    0    0   0 13
## french fry            5   0    0    0   0  5
## Gin                   1   0    0    0   0  1
## SB                    2   0    0    0   0  2
## Steve                 3   0    0    0   0  3
## Miss Piggy            3   0    0    0   0  3
## TX                    5   0    0    0   0  5
## Rachel                2   0    0    0   0  2
## VON                   4   0    0    0   0  4
## Alejandra Taboada    11   0    0    0   0 11
## America               2   0    0    0   0  2
## rex                   3   0    0    0   0  3
```

According to "Structures of Mediation: A Formal Approach to Brokerage in Transaction Networks" by Gould and Fernandez, a brokerage is a process by which intermediary actors facilitate transactions between other actors lacking access to or trust in one another. A broker within a network is a node with high betweenness centrality and/or low Burt's constraint score, which both, bridges the gap between disconnected stakeholders as well as receives/sends many relational ties out to the other stakeholders. Based on the Brokerage Function from the class data, all impactful nodes belong within the same subgroups and are categorized as coordinator roles. In this fashion, coordinators oversee the nodes of their clusters similar to how real-world coordinators coordinate the administration of a brokerage. Specifically, *Violet* and *Alejandra Taboada* assume the role more often than any other student. Even based on network location, the students are located at centers of clusters, thus their ability to facilitate actions between nodes is heightened.

In this network, social capital exists where students have an advantage because of their location within the network structure. Referencing Burt's article, nodes who are very connected across different groups are more familiar with alternative ways of thinking than nodes who are less connected. Amongst the students, those who have a higher total number of times as a coordinator resemble entities who are impactful members of their own subgroups as well as neighboring ones, thus more connected overall. For example, *Violet* maintains a large amount of social capital solely based on the fact that she is a broker with several connections with other subgroups of students. In this sense, *Violet*'s dense student and class-related relationships, while also acting as a bridge between subgroups, increase her social power amongst students within GW's Business

Analytics major. Interestingly, the students who act as brokers and have the highest social capital are also of international descent. While the exact reason why this feature is unknown, it may be linked to the number of international students in the Business Analytics major with similar ethnicities which supports homophilic subgroups with brokers, such as *Violet* or *Alejandra Taboada*, connecting them together.

Exponential Random Graph Models (ERGM)

```
##    edges triangle
##       58       18
```

As the summary shown above, we have 58 edges and 18 triad in our network. Before testing for homophily, we run a null model, which is the simplest Exponential Random Graph Models (ERGM) with only the edges term. Shown below, the edge has a Estimate of -3.4058. That is to say the log-odds that a edge is present is -3.4058*1. The corresponding probability is obtained by taking the inverse logit – exp(-3.4058)/(1 + exp(-3.4058)) = 0.032 – the 0.032 is the density of the network.

```
## Call:
## ergm(formula = net ~ edges)
##
## Maximum Likelihood Results:
##
##       Estimate Std. Error MCMC % z value Pr(>|z|)
## edges  -3.4058     0.1335      0  -25.52   <1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 2504  on 1806  degrees of freedom
##  Residual Deviance:  513  on 1805  degrees of freedom
##
## AIC: 515  BIC: 520.5  (Smaller is better. MC Std. Err. = 0)
```

We use mixing matrix to examine the number of connected dyads for each possible combination of levels for categorical node attribute if_international. Out of total 58 edges, large number of edge clustered within the Domestic-Domestic and International-International section of the mixing matrix, 17 edges and 30 edges respectively. Based on the exploratory analysis and before fitting the model, we make the assumption that if_international attribute has a significant impact on the friendship formation.

	Domestic	International
Domestic	17	6
International	5	30

We want to look at the attribute data and see how those attribute impact the formation of friendship. We used edges, graduation_year, if_international, gender, and origin to fit a ERGM model (the details of fit shown below). graduation_year and origin are categorical, while if_international and gender are considered as binary in this data set. Not surprisingly, if_international and origin are statistically significant, both with a p-value smaller than 0.05. The coefficients in ERGM represent the change in the (log-odds) likelihood of a edge for a unit change in a predictor. A interpretation of 1.03975 of nodematch.if_international.International can be – holding all other variable constant, a person being international student will change the odds of the dyad formation by a factor of e^1.03975 = 2.8285 over the reference level. The corresponding probability of tie formation is 0.7388 or 73.88%. This support our hypothesis that international student are more likely to gather with other international.

```
## Call:
## ergm(formula = net ~ edges + nodefactor("grad_yr") + nodematch("if_intl",
##     diff = T, keep = 2) + nodematch("gender", diff = T, keep = 2) +
##     nodematch("origin"))
```

```
##
## Maximum Likelihood Results:
##
##                               Estimate Std. Error MCMC % z value Pr(>|z|)
## edges                         -4.49408    0.36376      0 -12.355  < 1e-04 ***
## nodefactor.grad_yr.2024        0.04922    0.22121      0   0.222 0.823929
## nodefactor.grad_yr.2025       -0.06534    0.50135      0  -0.130 0.896299
## nodefactor.grad_yr.2026       -0.74347    1.04024      0  -0.715 0.474785
## nodematch.if_intl.International  1.03975    0.30551      0   3.403 0.000666 ***
## nodematch.gender.Male          0.35606    0.32214      0   1.105 0.269038
## nodematch.origin               1.32035    0.28756      0   4.592  < 1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 2503.6  on 1806  degrees of freedom
##  Residual Deviance:  465.5  on 1799  degrees of freedom
##
## AIC: 479.5  BIC: 518  (Smaller is better. MC Std. Err. = 0)
```

Based on the formula, we can predict the connecting probability of two students. For instance, if we have 2 following student:

Stdnt	grad_yr	if_intl	gender	origin
1	2024	International	Female	China
2	2024	International	Female	China
2	2024	International	Female	Honduras
4	2024	Domestic	Female	United States

While graduation year and gender remain unchanged. the probability of Student 1 connect to Student 2 is 11.16%, Student 1 to Student 3 is 3.21%, and Student 1 to Student 4 is only 1.16%. This example illustrate how student form friendship with other students within Business Analytics. And it again tells that international students are more likely to make connections with people from their home country.

Conclusion

Based on the analysis of the friendship patterns among Business Analytics majors at GW, it was found that international students were more likely to form connections with people from their home country. The degree distribution and centrality measures showed that Alejandra Taboada and Violet were the most popular and influential students in the network, while LTB had the highest closeness centrality, indicating his ability to quickly influence the entire network. The analysis of subgroups, roles, brokerage, and social capital within the network revealed the presence of strong and weak ties, as well as bridging and bonding social capital. The statistical models and interpretations tested for homophily and found that there was a relationship between a domestic student network and an international student network based on citizenship, and that it was possible to predict the probability of a connection in a network based on organizations using a network based on majors.

Overall, these results provide a comprehensive understanding of the friendship patterns among Business Analytics majors at the university and the factors that influence the formation of these friendships. The findings of this study have important implications for educators and faculty seeking to foster a more inclusive and diverse learning environment. By understanding the dynamics of friendship formation among international and domestic students, educators and faculty members can develop strategies to promote cross-cultural friendships and facilitate the integration of international students into the university community.It is worth noting, however, that the results of this study should be interpreted with caution due to the potential limitations of the survey. Not every student participated in the survey, which could have resulted in a less accurate representation of the results. Additionally, the use of self-reported data may be subject to limitations in the validity of the responses. Therefore, it is important to consider these limitations when interpreting the results of the study.

In conclusion, the analysis of the friendship patterns among Business Analytics majors at a university provides valuable insights into the dynamics of friendship formation among international and domestic students and the factors that influence these friendships. The results of this study can inform strategies for fostering a more inclusive and diverse learning environment and promoting cross-cultural friendships at the university.