

**CSS 692 – Social Network Analysis**  
**Final Project Report**  
**Ajay Kulkarni (G01024139)**

---

### **Abstract**

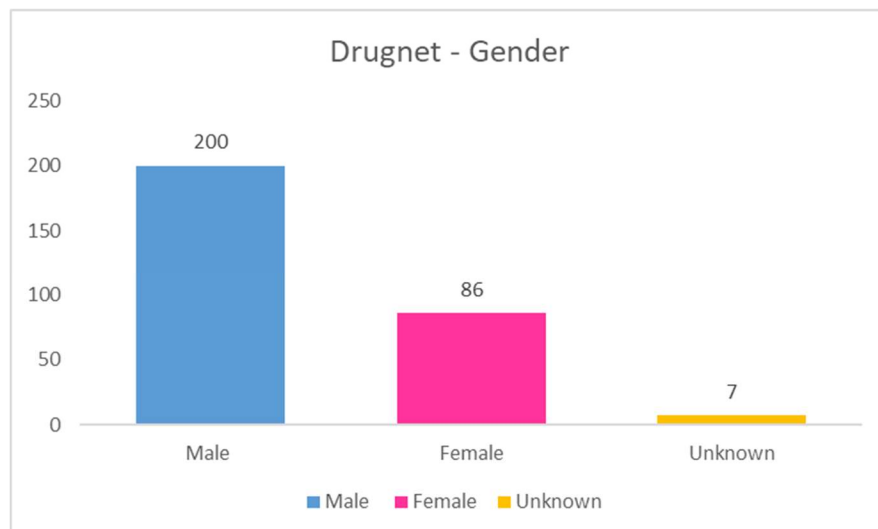
In this project, we have analyzed “Drugnet” network to identify the important nodes and the underlying network structures which forms the global pattern in the network. We have used centrality to find out the important nodes in the network and analyzed the network considering their gender as well as ethnicity. In addition to that, we have used Exponential Random Graph Models to understand the underlying structures in the network. Further, we created a model by selecting attributes, simulated the model and compared the simulated network with original network.

### **Introduction**

The network which we have used for this project is used from UCINET website, and the network is known as “Drugnet”. This network represents the interaction of 293 people in the form of adjacency matrix. The network is directed network and represents the data of 293 people from Hartford, CT. The data collected through a combination ethnographic interviewing, observations as well as drug-use site tracking and social network interviewing and tracking. The survey sample was constructed through two primary methods. The majority (55%) was recruited through street outreach in a neighborhood of high drug-usage. The rest of the cohort was referred to the study by survey participants through a “respondent-driven” sampling process. The eligibility criterion for survey participants recruited through any method included being at least 18 years of age and reported active use of heroin, cocaine/crack or other injected illicit drug.

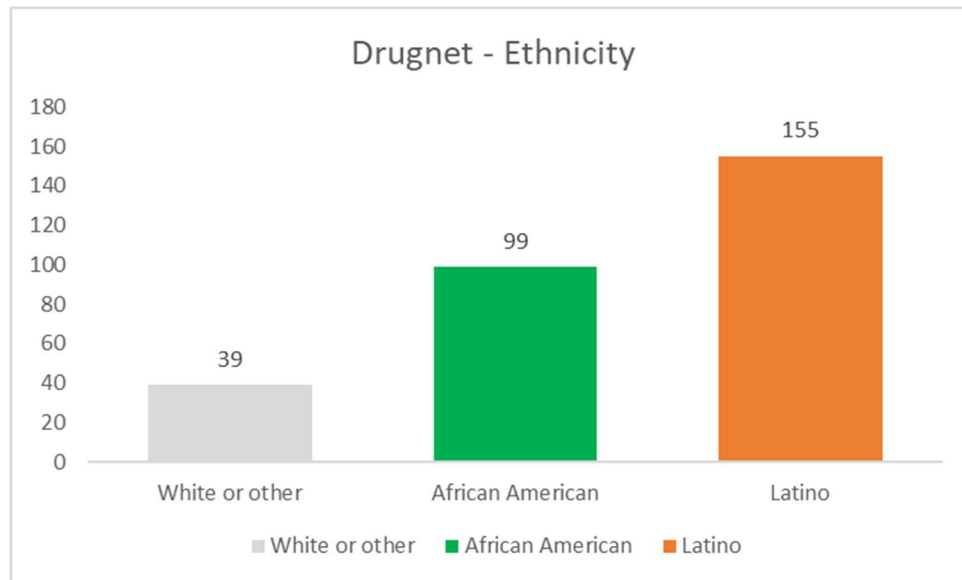
The report is divided into three parts, the first part will explain about network details. The second part will explain about important nodes and people we need to convince to educate about disease spread through the repeated use of same needles. In the last part, we will explain the details about simulated network and comparison with default network.

#### **Part 1: Network attributes**



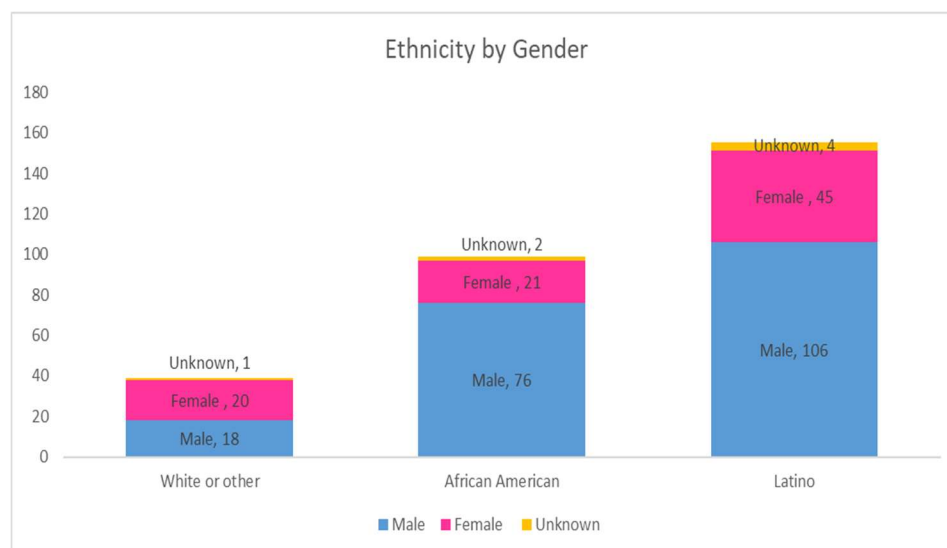
**Figure 1: Drugnet - Gender**

The drugnet network has information about two attributes – Gender and Ethnicity. We have presented statistics about Gender and Ethnicity in the form of bar chart. Plot 1 shows the distribution of gender of participants present in the survey. From 293 participants there are 200 male, 86 female and seven are unknown. This plot also indicates that the network contains about 68% male and they are dominant in the network..



**Figure 2: Drugnet - Ethnicity**

Figure 2 represents the distribution of ethnicity in the network. Three categories are present in the network – White or other, African American and Latino. It can be easily observed that the participants who have Latino (155) ethnicity are dominant as compared to white or other (39) and African American (99) participants in the network. The percentage of Latino participants in the network is about 53% and the percentage of African American participants is about 33%.

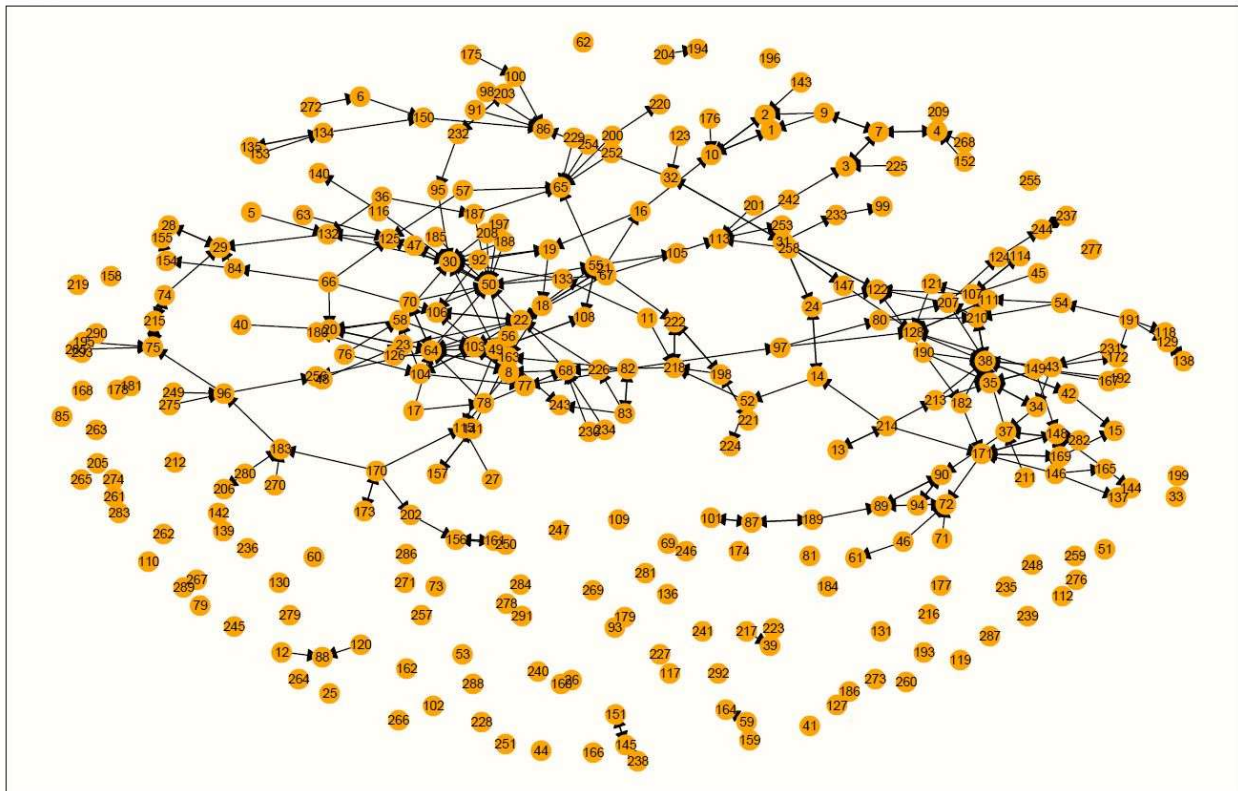


**Figure 3: Drugnet – Ethnicity by Gender**

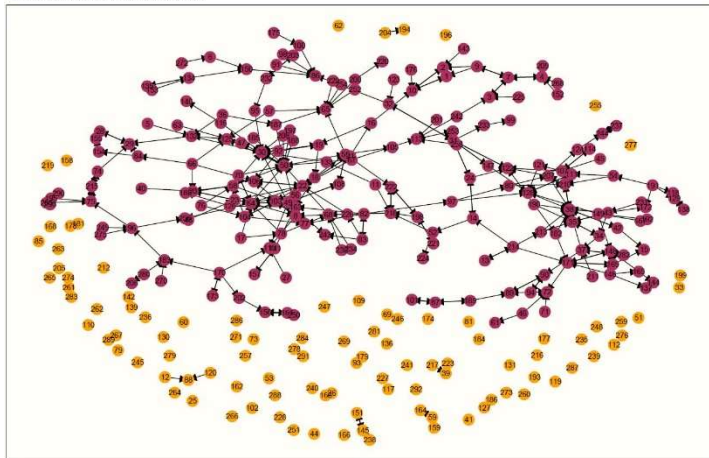
Figure 3 shows the stacked bar chart representing the categories of gender by ethnicity. We have seen that male participants are dominant in the network and observed that count of Latino participants is more as compared to count of other ethnicities. Figure 3 will help us to understand the distribution of ethnicity by gender in the network. It can be easily observed that in Latino and African American ethnicities there are more male participants as compared to female participants. On the other hand, in case of White or other ethnicities, there are more female participants as compared to male participants.

## Part 2: Network structure and centrality

Drugnet



Drugnet (Weekly Connected Component)



Drugnet (Strongly Connected Component)

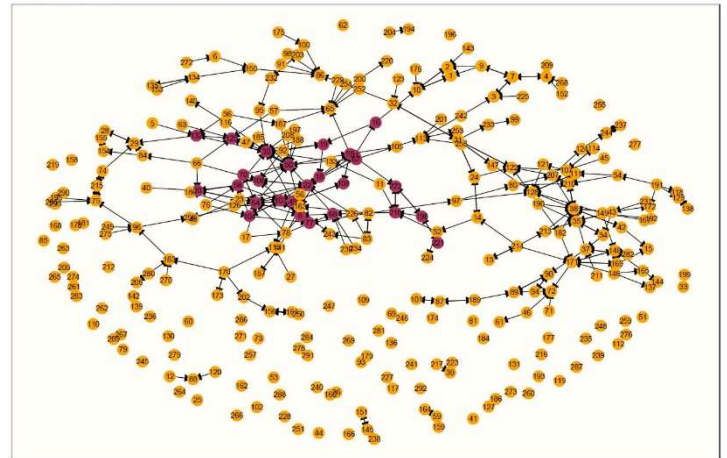
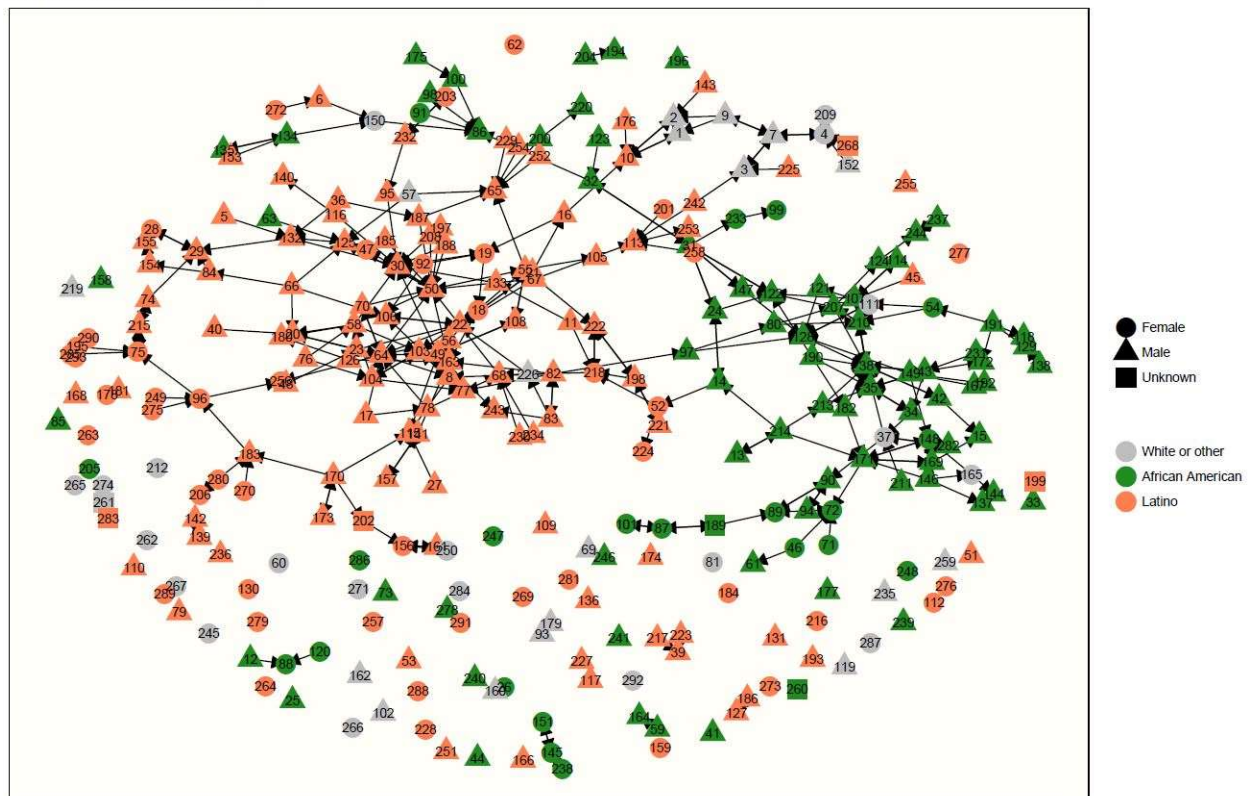


Figure 4: Drugnet – Network structure

The network structure (figure 4) represents the Drugnet network. The network is imported as adjacency matrix in R and then converted into network using “network” library. To make clear and elegant visualization “ggnet” library is used for all the network plots. The above network shows directed network consisting of a large component, small components as well as isolates. The directed network consists of a weakly connected component and strongly connected component. The weakly connected component includes two vertices if they are connected by a path where paths can go either way along the edge. Typically, there is one large weakly connected component present in the network. On the other hand, two vertices are strongly connected if each can reach and is reachable from the other along a directed path. We can easily observe weakly connected component and a strongly connected component which is highlighted by maroon color in figure 4. There are 193 nodes which are present in the weak component of the network which is about 65% of the total nodes. Also, there are 27 nodes which are present in the strongly connected component of the network. It can also be observed that there are 81 isolates present in the network which have degree zero.

Drugnet (Gender and Ethnicity)

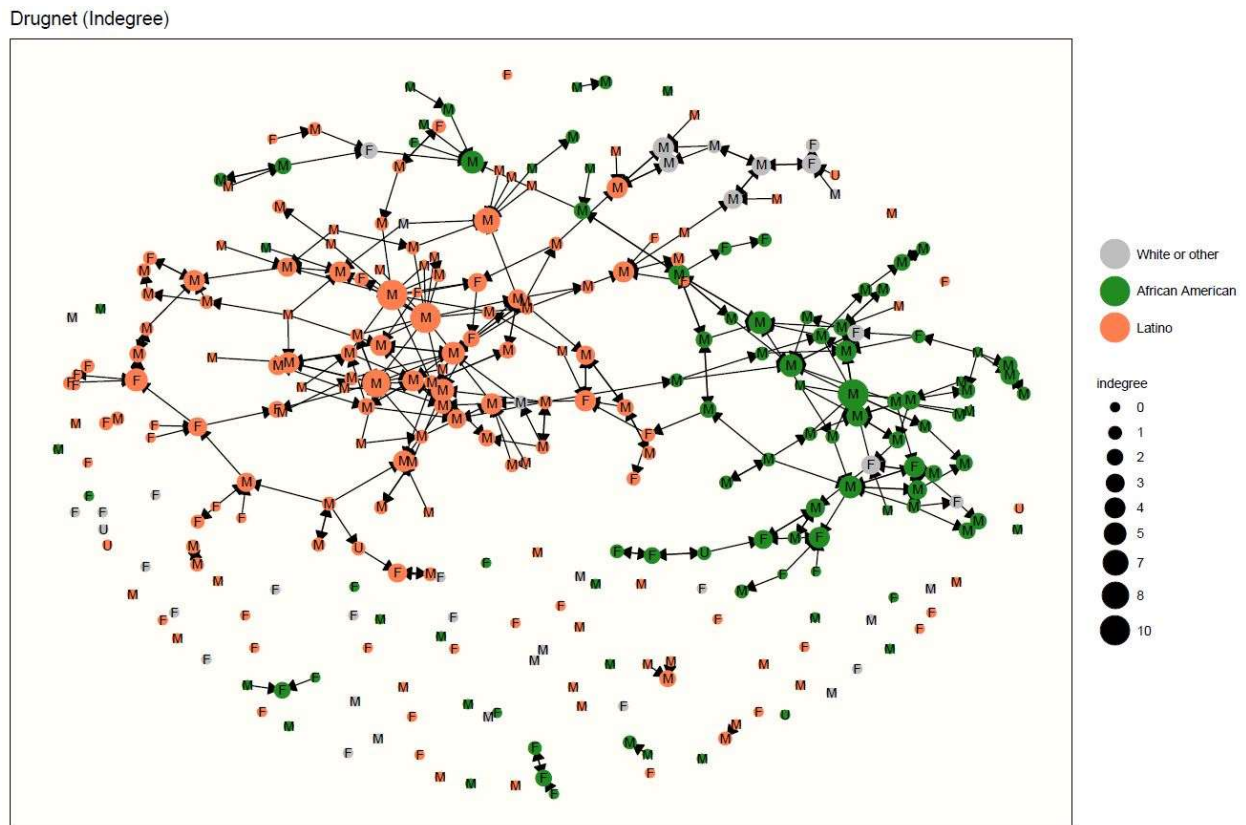


**Figure 5: Drugnet – Gender and Ethnicity**

The above plot represents gender (Male, Female and Unknown) by shape and ethnicity (White or other, African American and Latino) with color. It can be observed that there are two groups present in the network and they are divided by ethnicity. We already have observed that Latino and African American participants are more as compared to other ethnicities. The same can be observed from the above plot. Also, the two clusters are strongly visible with tight grouping within Latino and African American participants. We can also say that many of the white or other ethnicities are not part of the largely connected component and are isolated nodes.



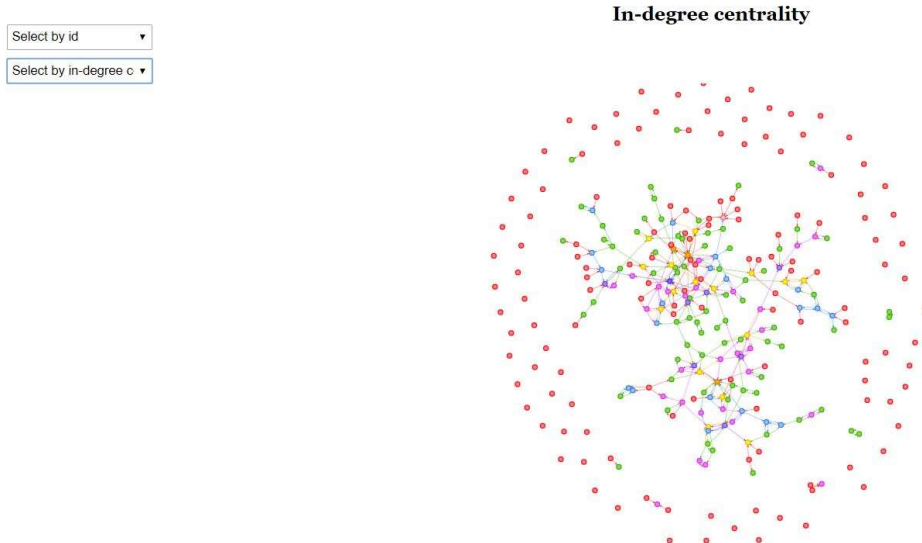
If we observe gender, then we can easily observe that most of the nodes are triangles which represent the male. It can also observe that largely connected component is strongly dominated by male participants with few female participants. We can also observe that majority of the female participants are present as the isolated nodes. The distribution of the females was more as compared to males in white or other ethnicities. So, this can be a reason that there are more female participants which are isolated and also belongs to white or other ethnicity. Thus, from the figure 6 we can say that there are two sub-groups present in the network one for Latino and other for African American. Also, many female participants are present as isolated nodes in the network.



**Figure 6: Drugnet – In-degree centrality**

Figure 6 shows the in-degree for every node in the network. In the above figure, color represents ethnicity and size of the node indicates in-degree of a node. The labels are also provided for every node which indicates the gender of a person. The reason behind the use of in-degree centrality is it indicates the popularity of a person. If a person has larger in-degree, it means that other people are making interactions with him about drug use and locations. If we find out people who have high in-degree centrality and educate them about disease spread, then there is a large probability that they can influence other people.

We can observe in the figure 6 that all the important nodes are present in the two groups – Latino and African American. If we consider a group of Latino people, then we can easily observe that the nodes which are present in the center of a cluster have higher in-degree centrality. It can also be observed that all the nodes which have higher in-degree centrality are male. Similar observations can also be concluded



**Figure 7: Drugnet – In-degree centrality using visNetwork**

for the African American group. To understand in-degree in more depth, we built an interactive plot using visNetwork library and provided two options (id and indegree) to understand centrality. The screenshot of the interactive plot is included and actual plot is saved in html format. We analyzed the network for the nodes which have in-degree centrality between 10 to 5 and included results in the following table.

Degree	Node id	Gender	Ethnicity
10	30	Male	Latino
	38	Male	African American
	50	Male	Latino
8	64	Male	Latino
7	65	Male	Latino
5	22	Male	Latino
	75	Female	Latino
	87	Male	African American
	124	Male	African American
	130	Male	African American
	165	Male	Latino
	173	Male	African American

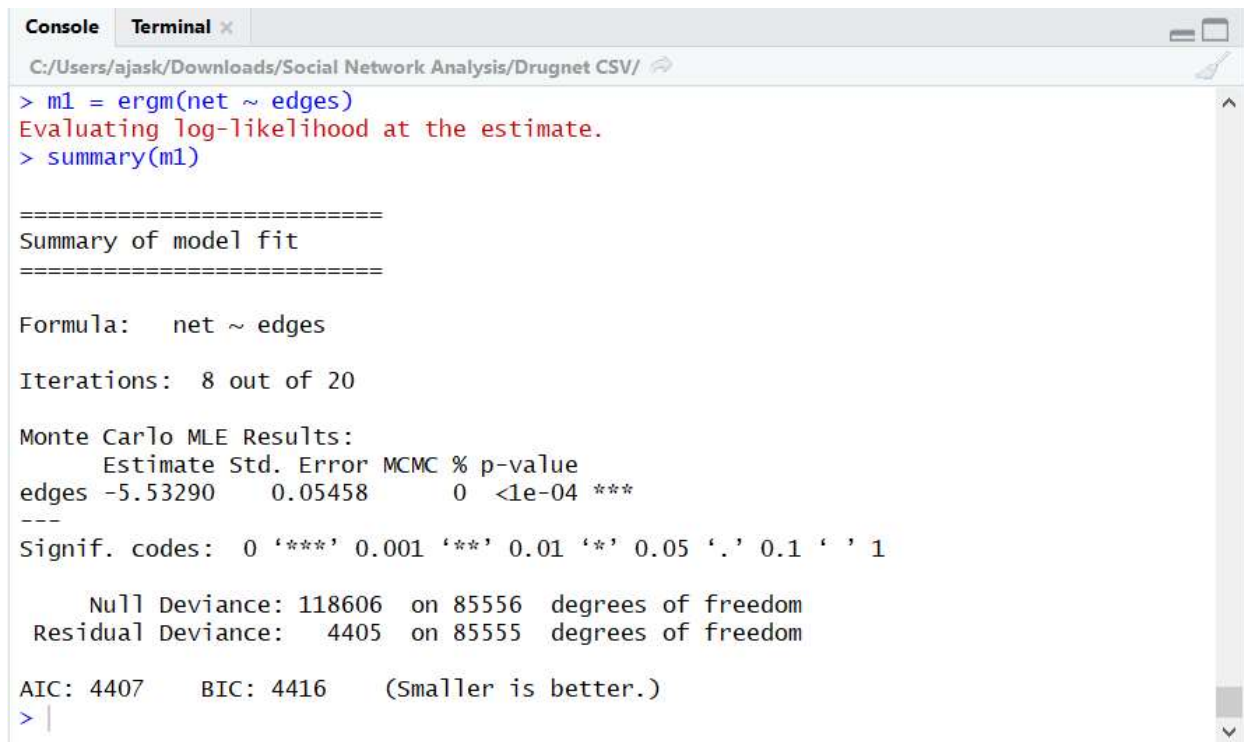
The above table indicates the degree, node id, gender and ethnicity of a person. In the network there are many nodes which are isolates and only three nodes which have degree centrality 10. We also observed that there is only one female participant who have in-degree centrality 5 and majority of the important participant in the network are male and belongs from Latino ethnicity. So, we think that if we want to control the drug usage and educate people, then we need to provide knowledge to all the people who are present in the table.

### Part 3: Exponential Random Graph Models

Exponential Random Graph Models present a way to understand the processes of a network structure emergence and tie formation. It helps us to understand better about local social processes and their interaction to form global network patterns. In ERGM we include a parameter of each configuration, and the parameter values will help us to identify a probability distribution for all graphs of size  $n$ . We estimate the parameters values that best matches the observed network. For estimation, we use MCMC (Markov Chain Monte Carlo) technique. Once we have our probability distributions, we can draw random graphs from it and compare their characteristic with our observed network. These coefficients of the model are very important and indicate the importance of subcomponents in the model. If the coefficient is positive, the probability increases with the feature value, and if the coefficient is negative, the probability decreases with feature value. If the coefficient is zero, then it means that feature does not affect.

In this project, we tried three different models – Erdos-Renyi model ( $p_0$  model), Holland and Leinhardt's  $p_1$  model and then we created our own model by selecting attributes based on Morris et al. (2008). As we know that the network is directed network and there are different terms present for every type of network. We used a table as a reference from Morris et al. (2008) for selecting the attributes for the directed network. The results of all three models are given below,

- **Erdos-Renyi model ( $p_0$  model)**



```
Console Terminal x
C:/Users/ajask/Downloads/Social Network Analysis/Drugnet CSV/
> m1 = ergm(net ~ edges)
Evaluating log-likelihood at the estimate.
> summary(m1)

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

Formula:   net ~ edges

Iterations: 8 out of 20

Monte Carlo MLE Results:
      Estimate Std. Error MCMC % p-value
edges -5.53290    0.05458      0 <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null Deviance: 118606 on 85556 degrees of freedom
Residual Deviance:  4405 on 85555 degrees of freedom

AIC: 4407    BIC: 4416    (Smaller is better.)
> |
```

Figure 8: Erdos-Renyi model

This is the simplest model where we only need to include edges term in the function. The syntax for the model and the calculated coefficient is given in figure x. We can see that the coefficient is strongly

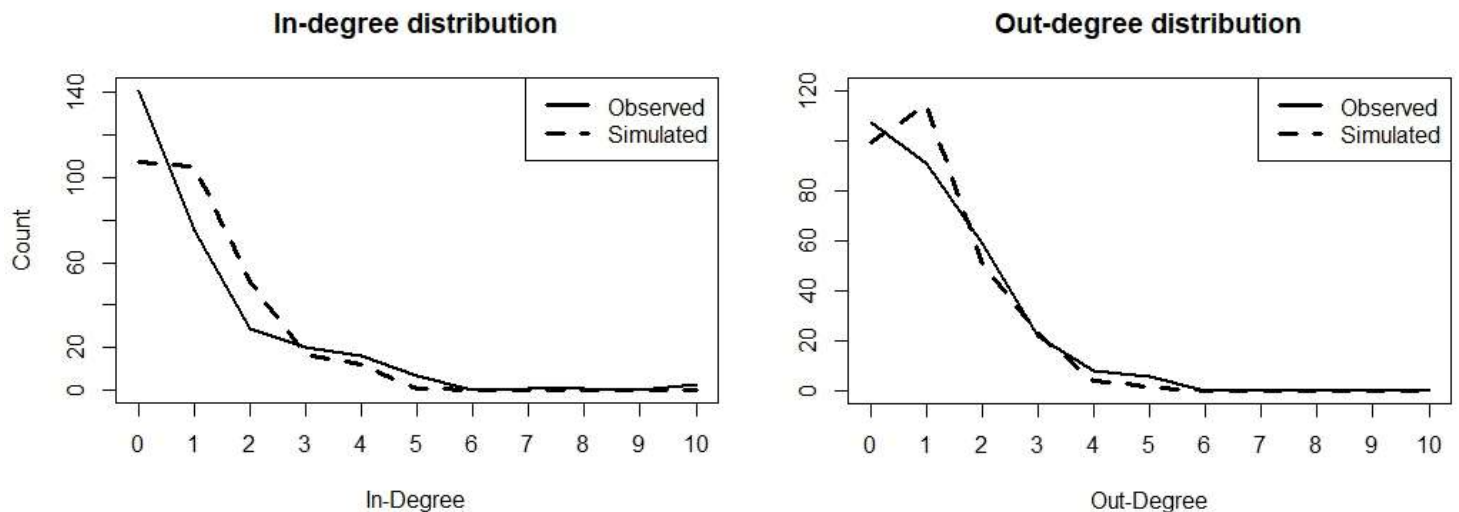
significant as the p-value is 1e-04 which is less than 0.05. Also, the value of the coefficient is -5.53290 and the probability that corresponds to this log-odds is

$$\frac{e^{-5.53290}}{1+e^{-5.53290}} = 0.000394.$$

This model can be interpreted as the log-odds of any tie occurring is

$$-5.53290 * \text{change in the number if tie} = -5.53290 * 1 = -5.53290$$

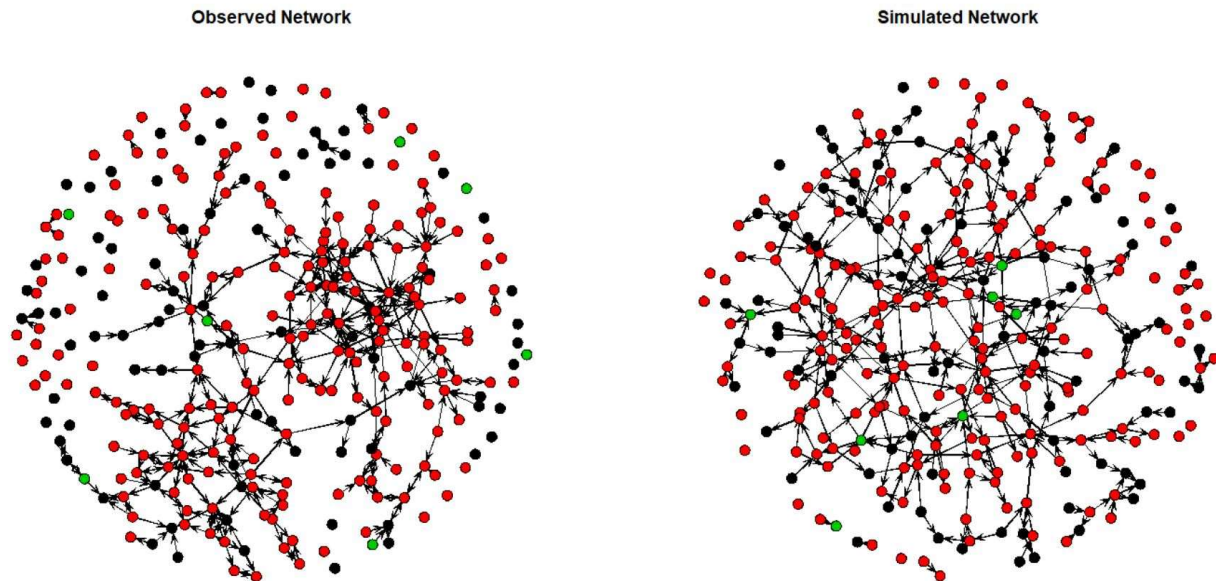
for all ties, since the addition of any tie to the network changes the number of ties by 1. We can also observe the BIC value is 4416 which is higher. The best model should have lower BIC value. ERGM represents the process that governs tie formation at a local level and then these local processes aggregate to produce characteristic of global network. So, we perform one test to see how well it reproduces these global properties. This can be done by choosing a network statistic that is not in the model and comparing the value of this statistic observed in the original network.



**Figure 9: In-degree and out-degree distribution of Erdos-Renyi model**

We can observe from the in-degree distribution that the number of isolates is under-represented by the model and number of nodes with degree 2 are over-represented by the simulated model. We can also observe that for in-degree there is under-representation of the nodes for degree 3, 4, and 5. In the case of out-degree, we can observe that number of nodes with degree 1 are over-represented by the model, and after that, there is a very small under-representation of the nodes for degree 2, 4 and 5. We also have simulated the network and then plotted simulated as well as observed networks on next page.





**Figure 10: Observed and Simulated network**

The above plot shows the simulated network and observed network. The colors of node denote the gender – Male (Red), Female (Black) and Unknown (Green). Two groups can be easily observed in the observed network but in case of a simulated network is difficult to identify the two distinct groups. Also, there are more isolates in the observed network, but a simulated network is unable to replicate that pattern.

- **Holland and Leinhardt's  $p_1$  model**

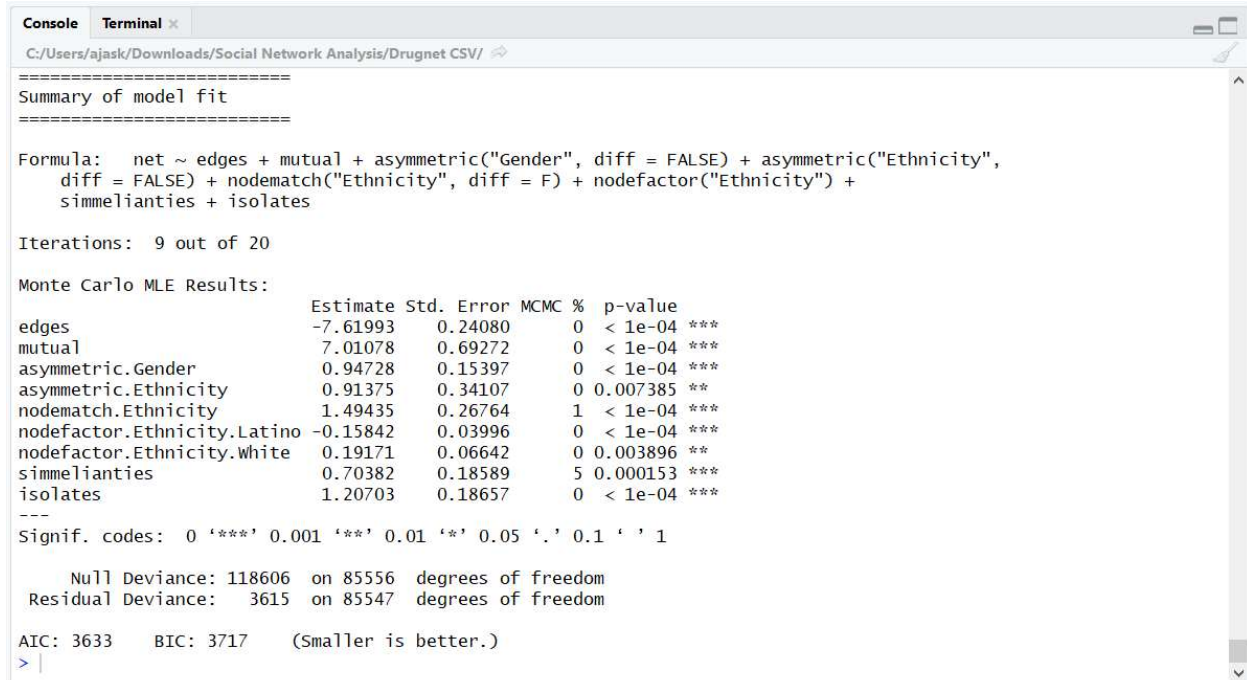
The  $p_1$  model can be represented by edges, mutuality, sender effect and receiver effect. We tried to simulate the  $p_1$  model, but due to computational limitations, it was not possible to complete the execution. The  $p_1$  model is a nice strategy if the directed network has less number of nodes but in a larger network to consider all the sender and receiver effects can be computationally expensive.

- **Model based on selected attributes**

In this case, we prepared our model by selecting statistically relevant attributes from the list provided by Morris et al. (2008). The details and their specifications are given below in the table.

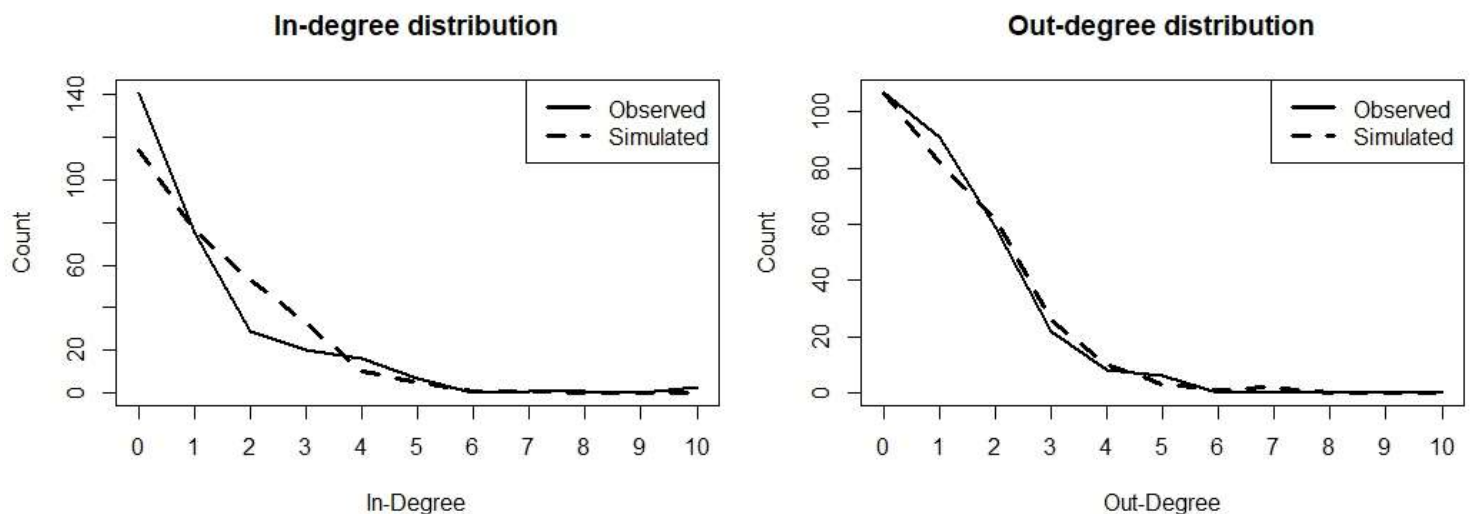
Attribute	Explanation
Edges	Edges in the network
Mutual	Mutuality
Asymmetric	Asymmetric dyads
Nodfactor	Main effect of a factor attribute
Nodematch	Interaction term - homophily
Simmelianties	Ties in Simmelian triads
Isolates	Isolates in the network

We also triad other attributes such as istar, ostar, m2star as well as degree attributes but we found that some attributes are insignificant for the model which resulted into increase in BIC. For other attributes, the model was degenerating. So, we decided to move forward with selected attributes. The model and the coefficients are given below,



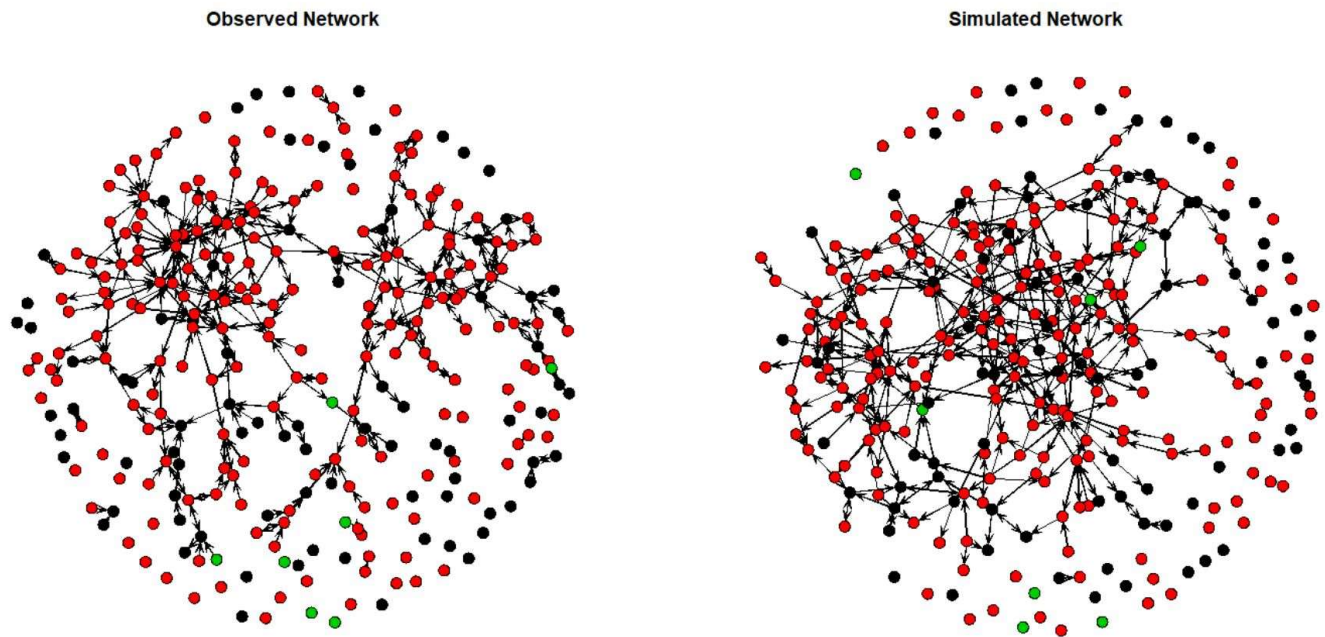
**Figure 11: Summary of the model**

It can be observed that the all the attributes are significant for the network. It can also be observed the BIC score 3717 which lesser than the  $p_0$  model (4416). So, based on AIC as well as BIC scores we can say that this model is a better fit for the network. The in-degree and out-degree distributions are also given below to compare the simulated network and original network. The figure 12 shows the in-degree and



**Figure 12: In-degree and Out-degree distribution**

out-degree distributions of the simulated and observed network. From in-degree network, we can observe that simulated model is underrepresenting the count of nodes of degree 0 and overrepresenting for degree 2 as well as 3. For out-degree, the simulated model can produce a similar pattern and we can observe that both the lines are very close to each other. So, overall this model is a better fit for the network compared to the previous model.



**Figure 13: Observed and Simulated network**

The above plot shows the network structure of the observed and simulated network. We can observe two groups in the simulated network. Those two groups are not much differentiable as compared to the observed network, but we can identify them. Also, the simulated network is looking dense as compared to the observed network. In case of isolates, the model is performing better as compared to the previous model and is a close match to the observed network.

### **Conclusion**

In this project “Drugnet” network has been analyzed to find important nodes as well as underlying structures in the network. We found that the network is male dominated and high percentage of participants belongs from Latino ethnicity. To find popularity of participants in the network we used in-degree centrality. From in-degree centrality we observed that there are many nodes which are isolates and we decided to focus on nodes which have in-degree centrality between 10 to 5. After performing this analysis 12 participants (nodes) were selected and if we educate them about drug usage then there is large probability that others in the network can get influenced. Further, we also used ERGMs for understanding the underlying structures in the network and compared simulated models with observed network. From this analysis we found that edges, mutuality, asymmetric dyads, isolates, homophily and simmelian triads are the important local structures which represents some part of the global pattern in the network.

## References

- 1) WEEKS, M. R., CLAIR, S., BORGATTI, S. P., RADDA, K. & SCHENSUL, J. J. 2002. Social networks of drug users in high-risk sites: Finding the connections. *AIDS and Behaviour*, 6, 193-206.
- 2) Introduction to Exponential-family Random Graph (ERG or  $p^*$ ) modeling with ergm, The Statnet Development Team, August 2017
- 3) ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks, *Journal of Statistical Software*, David R. Hunter et. al. (2008)
- 4) Speciation of Exponential-Family Random Graph Models: Terms and Computational Aspects, *Journal of Statistical Software*, Martina Morris et. al. (2008)
- 5) Social Network Analysis with sna, *Journal of Statistical Software*, Carter T. Butts (2008)
- 6) Exponential Random Graph (ERG or  $p^*$ ) Models, COMM 645: Communication Networks Annenberg School of Communication University of Southern California
- 7) A statnet Tutorial, *Journal of Statistical Software*, Steven M. Goodreau et. al. (2008)
- 8) [https://rstudio-pubs-static.s3.amazonaws.com/157501\\_93a72a58ec614946901e10edf78c1384.html](https://rstudio-pubs-static.s3.amazonaws.com/157501_93a72a58ec614946901e10edf78c1384.html)
- 9) <https://www.sci.unich.it/~francesc/teaching/network/>
- 10) <http://badhessian.org/2012/09/lessons-on-exponential-random-graph-modeling-from-greys-anatomy-hook-ups/>
- 11) [http://www.mjdenny.com/Preparing\\_Network\\_Data\\_In\\_R.html](http://www.mjdenny.com/Preparing_Network_Data_In_R.html)