**Presentation**

**Slide – 1**

Hello, my name is Ajay and today I will talk about Drugnet network.

**Slide - 2**

The objectives of the project are - to find dominant people in the network to reduce the spread of drugs and diseases. Many of the bloodborne infection can spread through repeated use of same needle by different people. So, finding and educating dominant people in the network might reduce this.

In this project I also studied the local structures which forms the global pattern and in the end, I simulated the network and compared with the observed network.

**Slide - 3**

The network which I have used for this project is **Drugnet** network. The data is present on **UCINET website**. The network is present in the form of adjacency matrix which represents interaction between **293 people**. The network also has information about **ethnicity** (weather a person is Africana American, Latino and white or other) and **Gender**.

The data is collected for **Hartford, CT** to understand about **people's drug habits**. This data was collected for two years **May 1998 to December 1999**. The data collected through a combination of**, drug-site tracking and social network interviewing** . The survey sample was constructed through **two primary methods**. **The majority (55%) was recruited through street outreach in neighborhoods of high drug-use activity**. **The rest of the cohort was referred** into the study by survey participants. Eligibility criteria for the survey participants recruited through any method included being at least **18 years of age and reported active use of heroine, cocaine/crack or other injected illicit drugs**.

**Slide - 4**

As I mentioned the network has information about two attributes – Gender and Ethnicity. This plot 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..

The other plot 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%.

**Slide – 5**

This stacked bar plot represents the categories of gender by ethnicity. 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.

**Slide - 6**

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 network is directed network consists of 293 nodes and 337 edges. Also, we can see that there is connected component and number of isolates. Total count of isolates in the network is 81.

**Slide – 7**

To explore more about network, I also have studied strongly connected component and weakly 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 usually one large weekly 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 weekly connected component and a strongly connected component which is highlighted by maroon color in figure 4.

There are 193 nodes present in the weakly connected component of the network which is about 65% of the total nodes. From those 193 nodes there are 15 participants belongs from White or other ethnicity, 69 participants from African American ethnicity and 109 participants from Latino ethnicity. Also, from 193 participants in weak component 148 are male participants and 43 are female participants.

In strongly connected component 27 nodes are present and from 27 participants all participants belong from Latino ethnicity and 24 of the participants are male.

**Slide – 8**

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.

We can easily observe that most of the nodes are triangles which represent male participant. 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 in the isolated nodes. The distribution of the female participants 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 ethnicities. Thus, from the figure 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.

**Slide – 9 and 10**

In the 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 having interactions with him about drug use and locations. If we find out people who have high degree centrality and educate them about disease spread, then there is a large probability that they can influence other people. It can be observed that all the nodes which have higher in-degree centrality are male.

To understand in-degree centrality in depth I built an interactive plot using visNetwork library and 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 table. So, I 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.

**Slide – 11**

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. 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.

**Slide – 12**

In this project, we tried three different models – Erdos-Renyi model (p0 model), Holland and Leinhardt's p1 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 are 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 p1 model can be represented by edges, mutuality, sender effect and receiver effect. We tried to simulate the p1 model, but due to computational limitations, it was not possible to complete the execution. The p1 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.

**Slide - 13**

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 highly significant as the p-value is 1e-04 which is less than 0.05 and the value of the coefficient is -5.53290. 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 * 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.

**Slide - 14**

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.

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.

**Slide - 15**

The 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.

**Slide - 16**

I 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.

**Slide - 17**

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. Also, all the attributes are strongly significant.

**Slide – 18**

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 observer that both the lines are very close to each other. So, overall this model is a better fit.

**Slide – 19**

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 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.

**Slide – 20**

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. So, there are total 12 nodes which we 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 underlaying structures in the network and compared simulated models with observed network. From this analysis we found that edges, mutuality, asymmetric dyads, isolates, homophily and ties in simmelian triads are the important local structures which represents some part of the global pattern in the network.