Network analysis of Titanic passengers

Vinay Sanga

1 Introduction

There is barely someone who doesn't know about the Titanic disaster, more people died than survived. But it begs an interesting question. "How did social networks, or connections between passengers, influence survival rates aboard the Titanic?". Was it just a chance of fate or there was a pattern which led to some people having more chances of survival than the others. In this report, we will take a look at the same things using network analysis and touch upon some interesting facts about the disaster.

2 Methodology

We will first perform data preprocessing, followed by visualization, construction of the network, survival analysis and so on. However, we will not perform any statistical analysis and model building as we are not predicting anything.

2.1 Dataset

We begin by downloading the dataset which is available freely as a part of challenge on Kaggle. The dataset contains the following columns:

- **PassengerId**: An identifier for each passenger.
- **Survived**: Indicates if the passenger survived (1) or not (0).
- **Pclass**: The class of the ticket the passenger purchased (1st, 2nd, or 3rd).
- **Name**: The name of the passenger.
- **Sex**: The gender of the passenger.
- **Age**: The age of the passenger.
- **SibSp**: The number of siblings or spouses the passenger had aboard the Titanic.
- **Parch**: The number of parents or children the passenger had aboard the Titanic.
- **Ticket**: The ticket number.
- **Fare**: The fare the passenger paid.
- **Cabin**: The cabin number where the passenger stayed.
- **Embarked**: The port where the passenger embarked the Titanic.

2.2 Data Preprocessing

We clean and preprocess the Titanic dataset, ensuring it's ready for network analysis. There were an astonishing 77% missing data in Cabin column and 20% people have no info about their about. As the Cabin column in not much useful, we will skip it. But we impute the Age column. After this, we create a column for family names. Using family names will be useful to find the relation between the people.

2.3 Data visualization

Visualizing the data gives us a better overview of the data and reveals some interesting facts. We find from fig. 1 that the no. of survivors was almost more than half the no. of those who did not survive. The survival by passenger class chart (fig. 2) indicates that a greater proportion of first-class passengers survived than those in third class. This might suggest that the first-class people had more access to the lifeboats or they were evacuated on priority. The age distribution from fig. 3 suggest that while there is some overlap in the age distributions of those who survived and those who did not, younger passengers appear to have a slightly higher survival rate, as indicated by the median age being

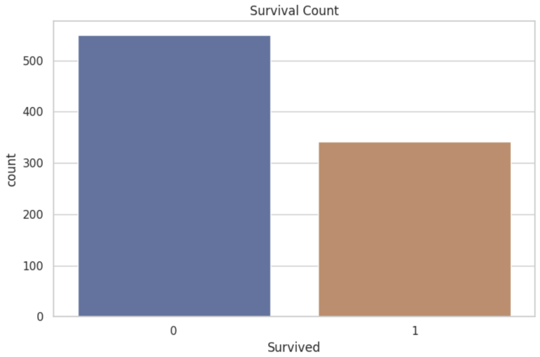

Figure 1: No. of people who survived

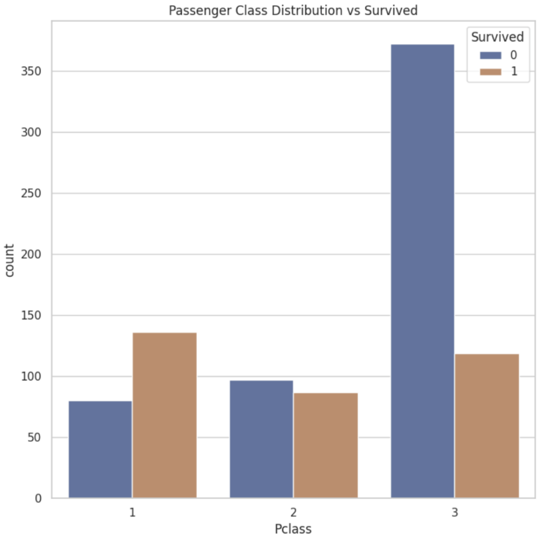

Figure 2: Distribution of passenger class vs. survival

lower for survivors. Maybe the younger people were given more preference while evacuating or maybe they were more capable of surviving as compared to the older people. Following this, we can see from gender-wise survival plot in fig. 4 that more females survived as compared to males. Overall, these visualizations align with historical accounts that women, children, and first-class passengers had higher survival rates on the Titanic.

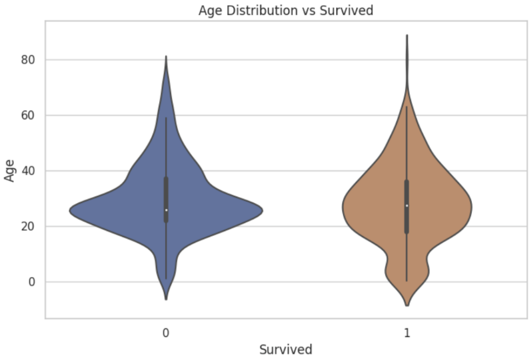

Figure 3: Distribution of age vs. survival

2.4 Network Construction

We created a network (fig. 5) where possible connections include family relationships (siblings, spouses, parents, children), friendship or acquaintance connections, and proximity connections (fig. 6) (passengers with the same ticket id).

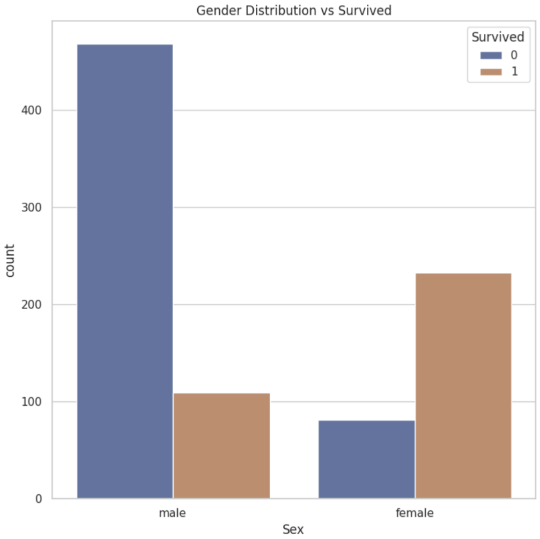

Figure 4: Plot of gender vs. survival

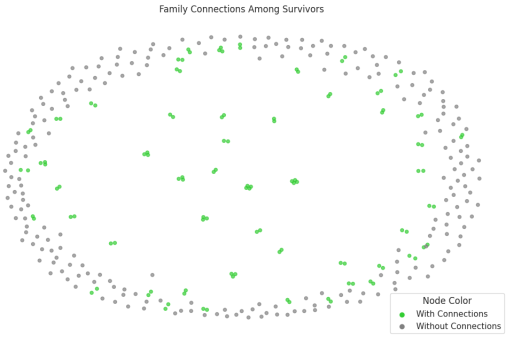

Figure 5: Graph with family connection

2.5 Survival Analysis

We performed survival analysis based on the centrality of the graph. They are discussed as follows:

- **Degree centrality:** The fig. 7 shows the top passengers in the family network with the highest degree centrality. Degree centrality measures the number of direct connections a node has. A higher degree centrality indicates a passenger was connected to more family members, suggesting a larger family or a more extensive family network on board. Fig. 8 illustrates the top passengers in the travel companion network with the highest degree centrality. Like in the family network, this metric highlights those with the most connections within their travel group, indicating passengers who were part of larger groups or had more companions.

- **Betweenness centrality:** Many values are 0 for family companion connections

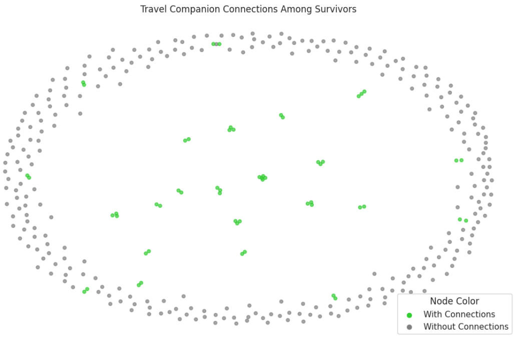

Figure 6: Graph with travel companionship

which indicates that few passengers significantly influenced the information flow in the family network. Similarly, the low values for travel companion connections suggest a limited role for most passengers in this aspect too.

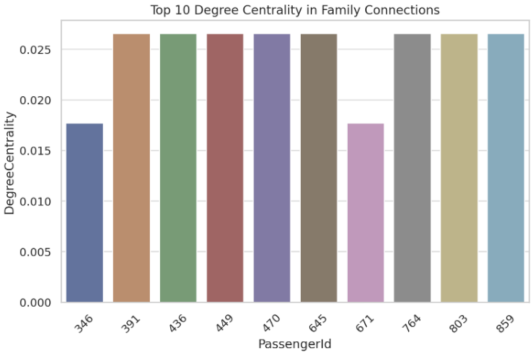

Figure 7: Degree centrality for family

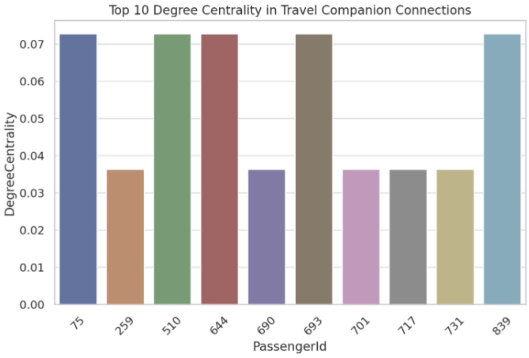

Figure 8: Degree centrality for travel companion

3 Conclusion

Among the passengers who had either family or travel companion connections, about **45%** survived. On the other hand, in the passengers who did not have any connections (neither family nor travel companions), about **32%** survived. In conclusion, this highlights the impact of social networks and connections on survival outcomes during the Titanic disaster. We also saw how age, passenger class and other factors enhance the survival rates.