Research Project Plan

Social Network Analysis for Gun Violence

Team: Alora Alexander, Jimmy Harris, Kamila Jusino-Melendez, Kabilan Selvakumar

Senior Data Fellow: Dr. Jeff Walker

Research Project Plan

Action Item	Week							
	10/19	10/26	11/02	11/09	11/16	11/23	11/30	12/07
Define the project goal and execution plan *Complete project plan, complete prep work, form teams and communication schedule*	X							
Understand the problem and project goal *Review project scope of work, reading materials and datasets*	X	X						
Define specific data and tool(s) needs *What data do you need? How and where will you retrieve it?*	X	X						
Get and understand the data *Gather the data from multiple data sources and define the various data points*	X							
Explore and clean the data *Clean and format the data, preparing it for analysis*	X	X	X					
Find insights *Conduct exploratory analysis as you work to answer questions in the project scope*		X	X					
Validate data *Did you answer the right question? Did you interpret the data correctly?*			X	X	X			
Visualize the data *Tell the story through your project deliverables (i.e., interactive dashboard, report, presentation, etc.)*				X	X			
Package project deliverables					X	X		

Finalize report and presentations and prepare to present								
Present findings *Present project in class*								X

Glossary

The following terms and acronyms are defined to ensure clarity

- **SNA**: Social Network Analysis, a methodology used to analyze the structure of social relationships within a network.
- **GNN**: Graph Neural Network, a type of neural network designed for processing data structured as graphs.
- **IDE**: Integrated Development Environment, a software suite that consolidates the tools required for coding and debugging.
- **PI**: Principal Investigator, Dr. Jeff Walker
- **Nodes**: Individuals or entities (e.g., victims, shooters, friends) represented as points in the social network.
- **Edges**: Relationships or connections (e.g., friendships) between nodes in the social network.

Project goal and execution plan

The main goal is to create an SNA with the data provided by our PI. Then we would need to see if anything exists on the social media pages (may or may not depend on what they have in their posts). It will most likely have to include a direct contact project to get to the people and get more context for their interactions (as stated by our PI). That will be down the road once we have the data where we need it and can run some initial analyses. The latter part is likely beyond the scope of this project, but it's in the general plan for our PI. The main priority is to complete the SNA. Another possibility is creating a proof of concept GNN that accurately "predicts" past crimes. With the mixed skillsets of our group, and the amount of interest there is from others in joining, we believe we can at least get started with this stage of the project.

Understanding the Problem and Project Goal

The SNA treats each person in the dataset as a node, and the connections (interaction and following on social media) are treated as edges. With this structure, you can assign centrality and betweenness values to each person. Centrality is the most straightforward value to find, as it just represents the number of people each person is connected to. This doesn't consider the strength of those connections in any way. The betweenness value represents how often a node will have a message passed through it when a message gets passed through the network. A node with a high betweenness value and low centrality value can be thought of as a bridge between two heavily populated areas.

Specific data and Tool(s) Needed

a. Network X python Library & possibly pytorch GNN if network analysis is completed ahead of schedule
    i. https://networkx.org/

ii. https://github.com/pyg-team/pytorch_geometric

b. Set up IDE environment

i. Determined by preference of each member

c. Python 3.9 or later

d. Project Repository – Code Version Control

i. GitHub

e. Data (From Dr. Walker): Victims + Shooter + Friends list

i. Already obtained

Get and understand the data on gun violence

We've already gotten our data, and our PI explained it to us. We will be meeting with him on October 16 to ask more questions about the data.

Explore and Clean the Data

Address missing data and consider integrating newer social media profiles with PI input.

Emphasis should be on preparing data for Network X compatibility.

Find Insights

Once centrality and betweenness values are assigned to each node, we'll re-examine the social media profiles to get a feel for what these values look like in the real world. Our PI will examine these results and provide some direction in interpreting them.

Validate Data

We'll use pandas for data integrity checks: missing nodes/edges and duplicate ID need to be removed from the dataset before sharing results and creating any visualizations. Our PI will likely give guidance on validation criteria.

Visualize the Data

The goal is to create an interactive network chart (via Matlotlib or Plotly) that shows key nodes through color/size based on centrality metrics. We could include heatmaps or line graphs.

Package Project Deliverables

The main deliverable will be our GitHub Repository with documentation on how to automate the extraction of the data and comments.

Present Findings

** will have more than this (waiting on more data from our PI) ** We'll be presenting a Jupyter notebook that demonstrates the findings of the SNA.