DS 210 Final Project Report

Crime Data from 2020 to Present (Los Angeles Open Data)

Dataset: I used the “Crime Data from 2020 to Present” dataset published by the City of Los Angeles. It contains every reported crime incident in L.A. since 2020, with timestamp, crime type, and neighborhood (AREA_NAME). This real-world data is interesting because it lets us model how criminal activity “spreads” through the city over time and space.

Project Idea and Implementation:

- Nodes: Each unique neighborhood (AREA_NAME) in the dataset.
- Edges: Two neighborhoods are connected if both report at least one crime on the same calendar day.
- Pipeline:
    1. Preprocess (Python): Deduplicate the million-row CSV down to unique (DAY, AREA_NAME) pairs (data/day_area.csv).
    2. Graph Construction (Rust): Read day_area.csv, build an undirected petgraph::UnGraph<String, ()> with one node per neighborhood, fully connect co-occurring areas.
    3. Analysis (Rust):
        - Degree Distribution: Count how many neighbors each node has and print a sorted histogram.
        - Average Shortest-Path (BFS): For each node, run a breadth-first search to sum distances to all others, yielding an overall average path length.
        - Closeness Centrality: Compute for each node the reciprocal of its total distance sum, identify the top 5 “most central” neighborhoods.
        - Connected Components: Count how many disconnected groups appear.
    4. Export:
        - report/metrics.json with all key metrics
        - report/degree_counts.csv for plotting
- Languages & Tools: Rust (Cargo + petgraph) for core pipeline; Python (Pandas/Matplotlib) for preprocessing and plotting.

Results:

- Graph size: 103 nodes, 1 234 edges
- Average shortest-path length: 2.431
- Degree distribution: heavy-tailed (many neighborhoods with few same-day links, a handful with dozens).