NETWORK ANALYSIS

ST. LOUIS CRIME

GROUP 10

August - Chrisanna - Danielle - Maria - Moneeca - Sabrina

IT-UNIVERSITETET I KBH

AGENDA

- Introduction – Data and Network
- Network Statistics & EDA
- Research Question
- Analysis & Results
- Conclusion
- Limitations
- Questions

INTRODUCTION - THE DATA

- Derived from police records using Snowball sampling from five initial homicides in St. Louis in the 1990s

- Metadata includes:
  - Name
  - Gender
  - Person's role in the crime

- Roles include:
  - Victim
  - Suspect
  - Victim/suspect
  - Witness

INTRODUCTION - THE NETWORK

- Undirected, unweighted, bipartite network
- 1380 nodes
  - 829 people
  - 551 crimes
- 1476 edges
- Nodes = people and crimes
- Edges = connect people to particular crime events

NETWORK STATISTICS

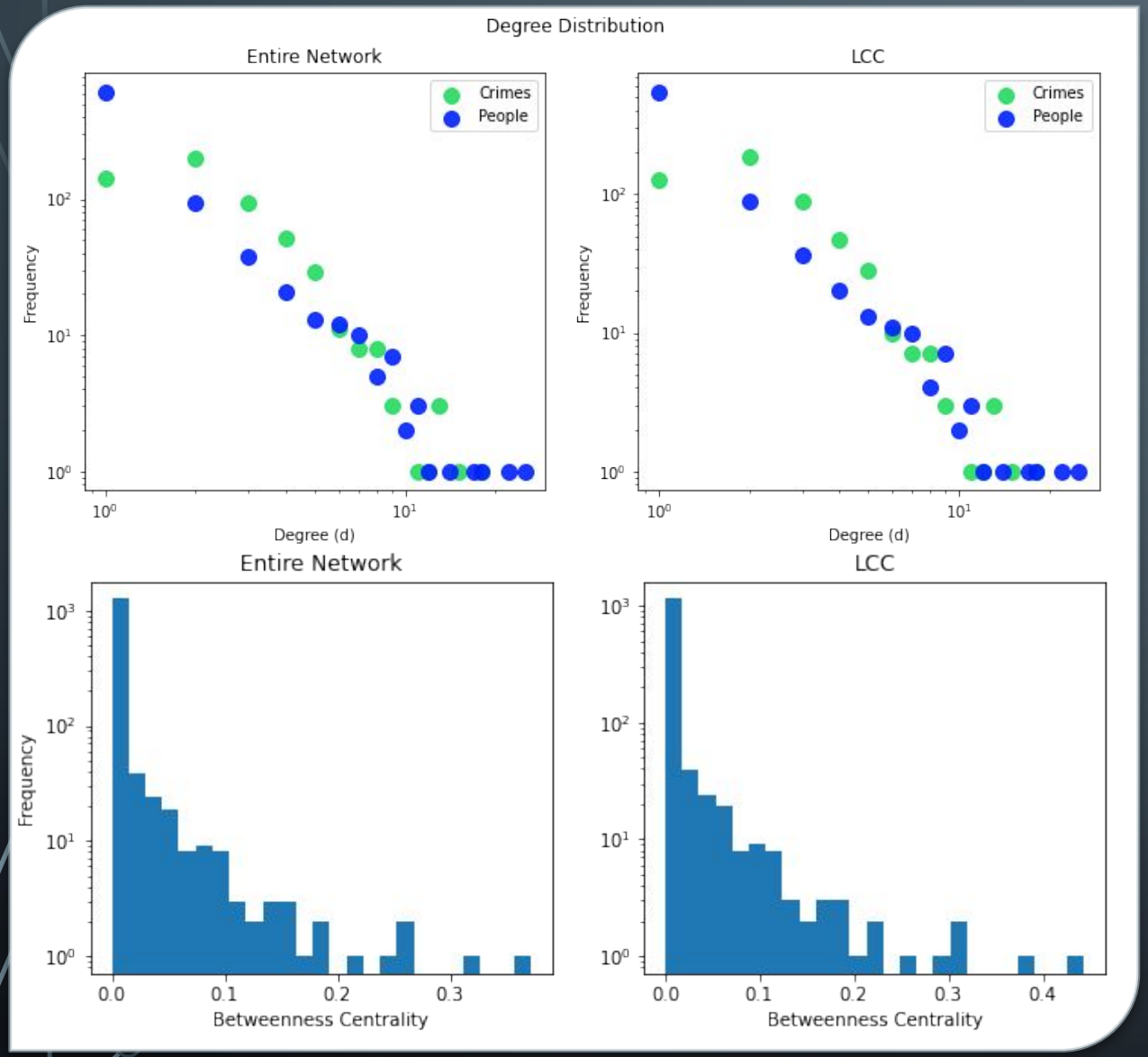
Created a configuration model based on the same degree distribution to compare basic statistics of our network with a random model.

	Crime Network	Configuration Model
• Density:	0.003	0.003
• Average Clustering:	0.427	0.335
• Global Clustering:	0.0505	0.005

Higher than random global clustering → makes sense with social network

RESEARCH QUESTION

Is it possible to find people of interest in a crime network, using link prediction and community detection?



ANALYSIS - LCC

- 20 connected components
- The largest connected component (LCC):
  - 1263 nodes (out of 1380)
  - represents 91.5% of all nodes

LCC representative of entire network

ANALYSIS - PROJECTION

Create a unipartite graph of people

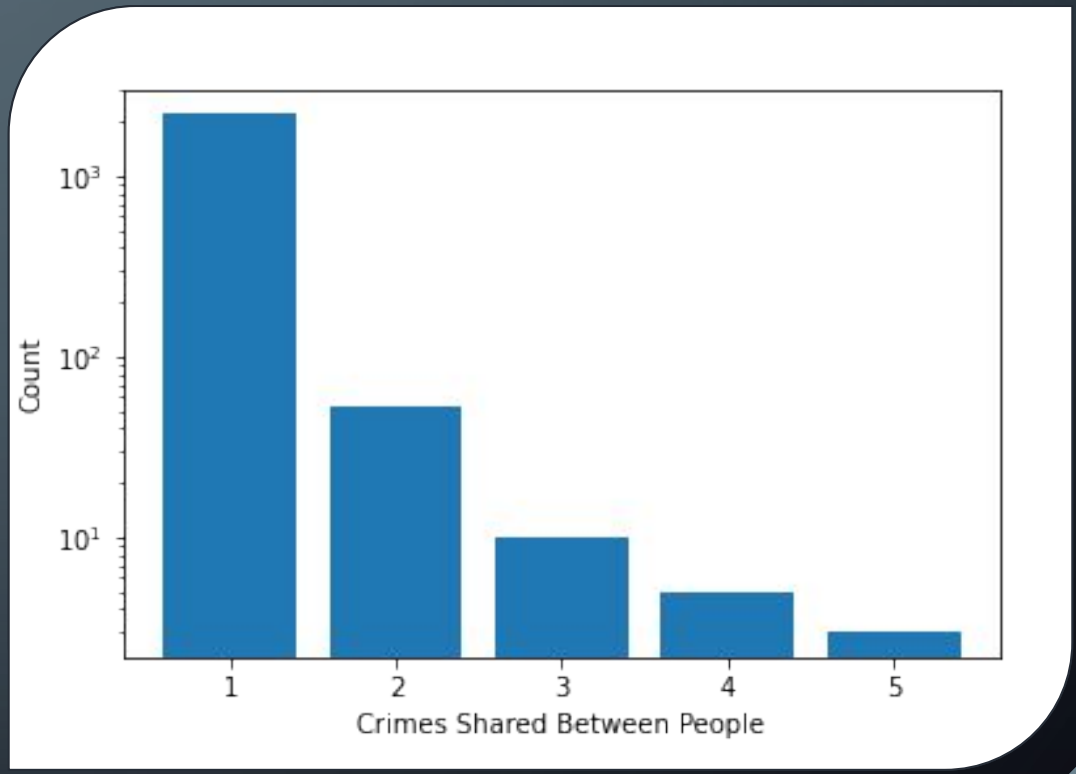
Reasons:

- To detect communities of people
- To predict new edges between people
- Metadata only available for people

Simple weighting is sufficient:

- Edge weight = number of crimes shared between two people

Number of crimes that connect two people



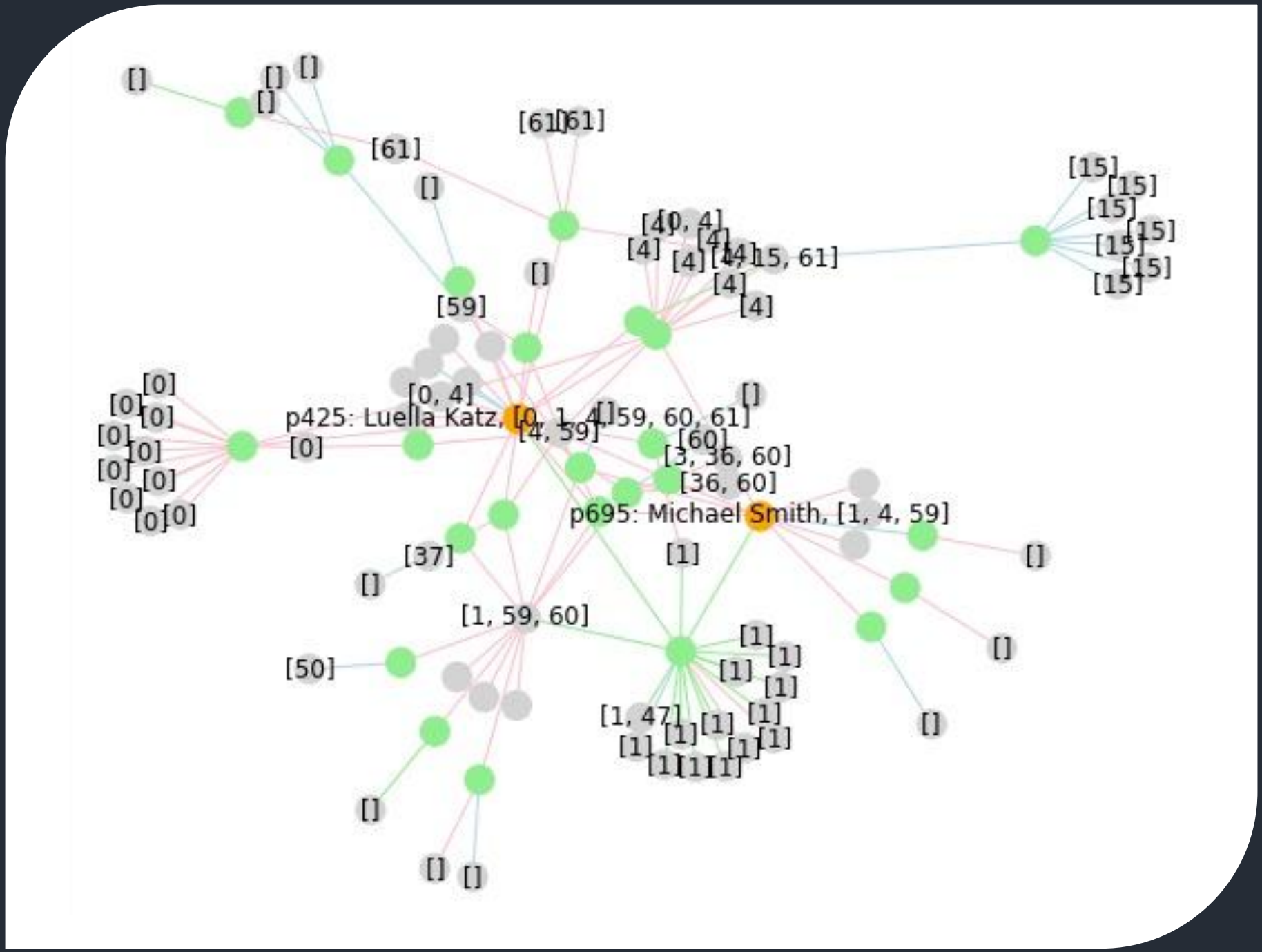
- Most of the people are linked by one crime and max value is five crimes

ANALYSIS – COMMUNITY DETECTION

Assumption: People in same crime and same community → they know each other!

- Used two algorithms for community detection :
  - greedy_modularity_communities → highest modularity score (separate)
  - K-cliques → overlapping communities

Community detection alone is not insightful

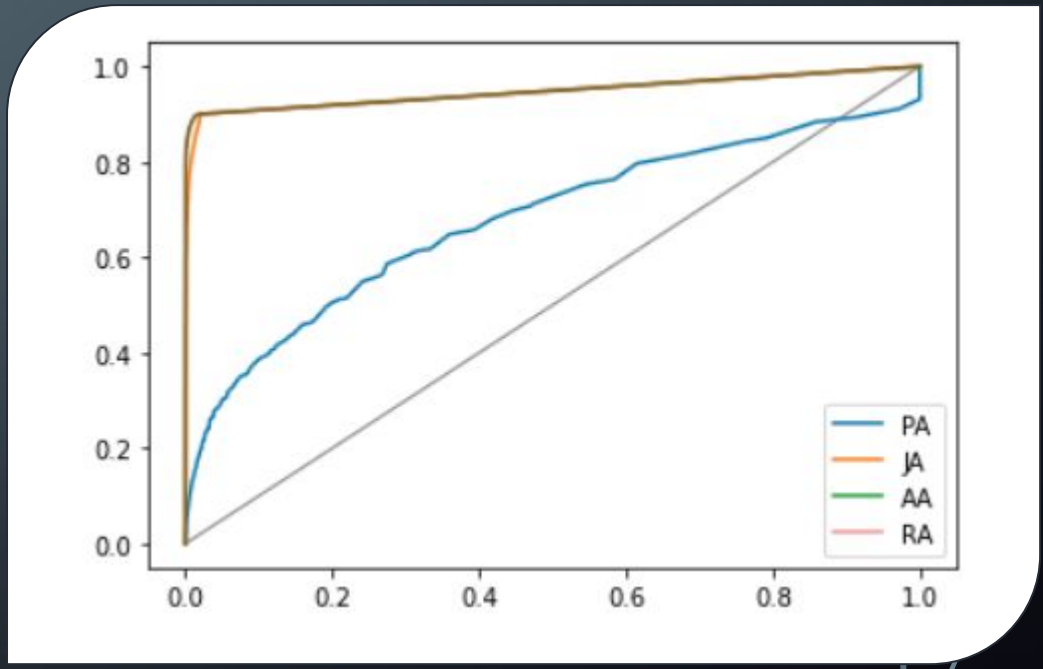

p425: Luella Katz, [0, 1, 4, 59, 60, 61]
p695: Michael Smith, [1, 4, 59]
[0, 4]
[4, 59]
[3, 36, 60]
[36, 60]
[1, 59, 60]
[1, 47]
[50]
[37]
[59]
[61]
[4]
[15]
[60]
[1]
[0]
[]

ANALYSIS – LINK PREDICTION

Method: **RESOURCE ALLOCATION INDEX** - allocates connections to nodes whose common neighbors have low degree.

Reasons:

1. Limit number of crimes a person can participate in
2. Consideration of distance between nodes
3. Size of the network - log scale



ROC curves

→ Performance evaluation with k-Fold Cross Validation

- AUC = 0.95

RESULTS - LINK PREDICTION

Top 9 of predicted links

	node1	node2	RA_score
0	p336	p815	0.708333
1	p695	p691	0.500000
2	p74	p237	0.500000
3	p132	p690	0.500000
4	p301	p815	0.500000
5	p215	p214	0.500000
6	p293	p797	0.500000
7	p155	p269	0.500000
8	p228	p514	0.500000

Possible people of interest

ID	Name	Gender	Class	Degree
p815	Jenny Willis	F	Criminal	25
p695	Thomas Michael Smith	M	Criminal	11
p336	Liz Hall	F	Criminal	9
p691	Percy Small	F	Criminal	1
p301	Brian Godfrey	M	Innocent	1

p815 - two times in the list of top 9 predicted links and highest value of RA score
p695 - list of people in overlap communities

CONCLUSION

- Community detection alone does not add additional insight.
  - Crime re-detection, not community detection
  - Communities tend to form around 1-2 crimes
  - Those placed in overlapping communities were in multiple crimes
- Through link prediction, possible people of interest could be identified.
  - A predicted link predicts a crime involving the two people
  - Although their role in the possible crime is unknown

LIMITATIONS

- Small sample size
- Lacking information (type of crime, time and date, etc.)
- Missing nature of connection when projecting onto people


Thank You For Your Attention!
Presentation is over!