
INF

FAKULTÄT FÜR
INFORMATIK

# VAST Challenge 2020
# Mini-Challenge 1: Graph Analysis
## Summary and Data Understanding

Supervisor: Prof. Dr.-Ing. Bernhard Preim
Dr.-Ing. Monique Meuschke
M.Sc. Uli Niemann

Presenter: Seyed Behnam Beladi
Atrayee Neog
Xiongjun Wang

# Agenda

- Problem Overview
- Mini-Challenge 1
- Exploring the data
- Question 1 overview
- Question 2 overview

# Problem Overview

- Numerous “white hat” hacker organizations **protected** the Internet
- One **anonymous** hacker organization, **accidentally** launched a cyber event that **took down** the global Internet.
- The group has to be **found**
- Center for Global Cyber Strategy(CGCS) is the key
- CGCS maintains offline **databases** (donated for research) of anonymized data including the responsible group
- Goal is to identify candidate groups that authorities could approach for assistance in restoring the internet.

# Mini-Challenge 1 (Overview)

- **One profile** has been identified by CGCS as **most likely** to resemble the structure of the group responsible for internet outage
- **Our task** is to identify the **groups** who **resemble** the identified profiles

# Mini-Challenge 1 (Data overview)

- A subgraph **Template** representing the structure of the group **identified** by CGCS
- 5 **candidate** subgraphs
- A very **large graph**
- A list of 3 **Seeds**, or IDs that can provide starting points for exploring the large graph.

# Exploring the data (Overview 1)

- **All** graph files contain the following columns:
  - **Source**: an integer Id of the source of the communication (could have different meanings based on the eType column)
  - **eType** (edge type): a number between 0 and 6 (inclusive)
  - **Target**: an integer Id of the source of the communication (could have different meanings based on the eType column)
  - **Time**: Time is in seconds from 12:00 AM Jan. 1, 2025, time span related to the cyber event are exactly one year

# Exploring the data (Overview 2)

- Channels are defined based on the eType column.
- **Many** of the channels also include:
  - **Weight**: float values with different meaning based on the channel
  - **SourceLocation**: integer values between 0 and 5 representing countries
  - **TargetLocation**: integer values between 0 and 5 representing countries
  - **SourceLatitude**: latitude locations within the country
  - **SourceLongitude**: longitude locations within the country
  - **TargetLatitude**: latitude locations within the country
  - **TargetLongitude**: longitude locations within the country

# Exploring the data (Overview 3)

- Data types for each column when loading the file:
  - Source int64
  - eType int64
  - Target int64
  - Time int64
  - Weight float64
  - SourceLocation float64
  - TargetLocation float64
  - SourceLatitude float64
  - SourceLongitude float64
  - TargetLatitude float64
  - TargetLongitude float64

# Exploring the data (Channels)

- The data can be classified into 6 different channels
- Each channel represents a different kind of transaction between two nodes.
- These are the channels:
  - Communication
    - Email
    - Phone
  - Procurement
  - Co-Authorship
  - Demographic
  - Travel

# Exploring the data (Channels)

| Channel Name | eType | Representation | Location | Weight | Source | Target | Notable points |
|---|---|---|---|---|---|---|---|
| **Communications (phone and email)** | 0 & 1 | Direct connections between two persons | Some | Always 1 | person | Person | Phone and email channels not clear |
| **Procurement** | 2 & 3 | Buying and selling an item | no | Value of the item | person | Item | For each sell row exists: a buy row |
| **Co-authorship** | 4 | publication of scientific or technical articles | no | Fraction of the authors | Person (author) | Publication | Date must be ignored (not relevant) |
| **Demographics** | 5 | spending characteristics of a person | no | Money spent | person / category | person / category | 29 categories |
| **Travel** | 6 | Connecting people by location | yes | Length of trip(days) | person | location | Some weights are negative |

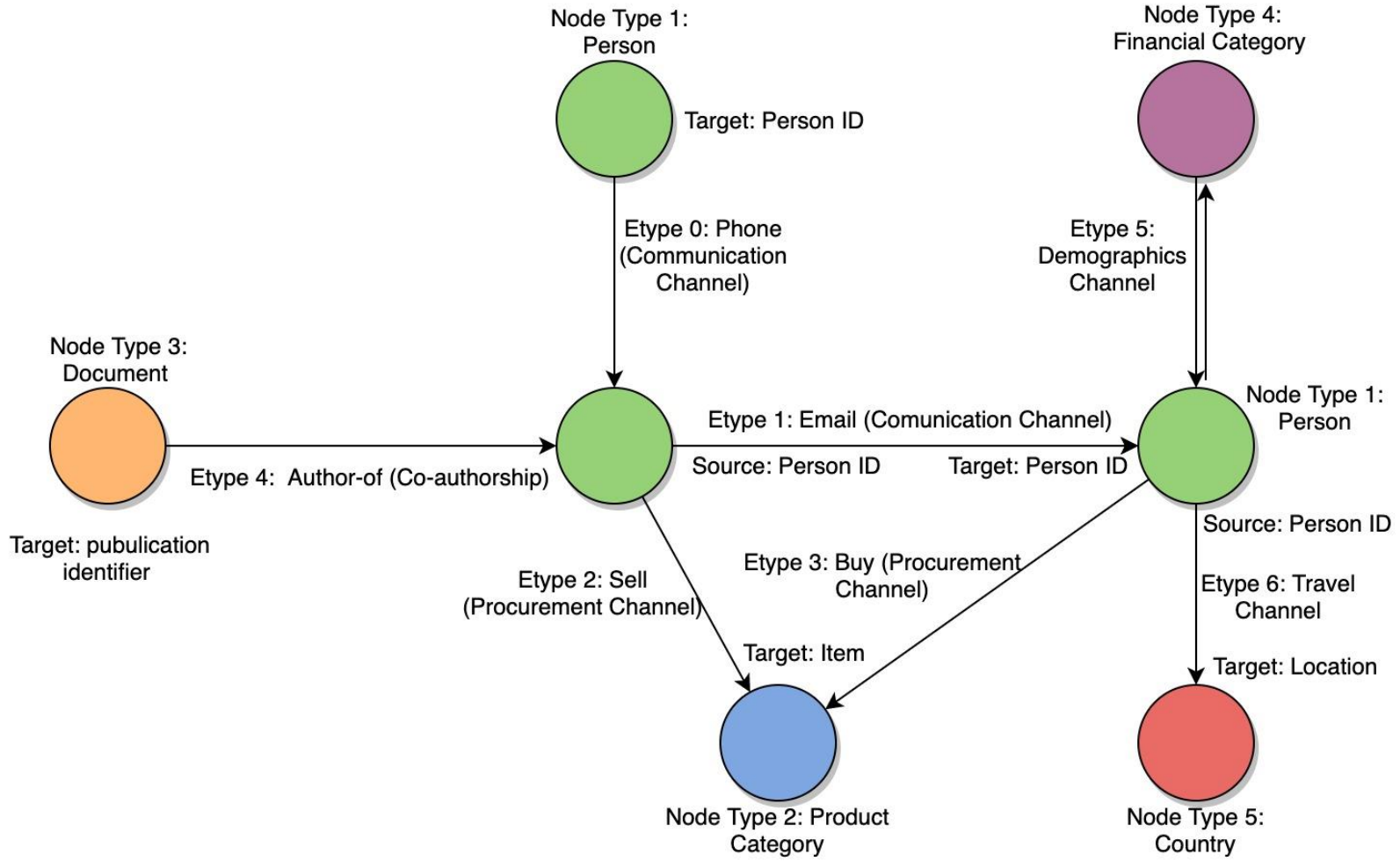
# Exploring the data (Node)

- Each Source and Target Id represents a node
- There are 5 Node type:
  1. Person
     - used in all channels(all eTypes), only nodes with a spatial location
  2. Product category
     - for the procurement channel, eType = 2, 3
  3. Document
     - from the co-authorship channel, eType = 4
  4. Financial category
     - from financial demographics channel, eType = 5
  5. Country
     - from the travel channel, eType = 6

# Exploring the data (Edge)

- Each row is an edge between two nodes
- At least one person is connected to each node
- 7 Edge type (eType):
    1. Email
    2. Phone
    3. Sell (procurement)
    4. Buy (procurement)
    5. Author-of
    6. Financial (income or expenditure, depending on direction)
    7. Travels-to

# Connection between nodes and edges:



# Exploring the data (Template)

- Edge list graph with the same format as the large graph data (.csv)
- Was built by CGCS to represent suspicious activity associated with the hack
- It is a reference pattern for looking for the suspicious activities
- Details:
  - File name: CGCS-Template.csv
  - 1325 rows
  - 301 have location data, none have longitude and latitude
  - The co-authorship channel is replaced by -99

# Question 1

Using visual analytics, compare the template subgraph with the potential matches provided. Show where the two graphs agree and disagree. Use your tool to answer the following questions:

a) Compare the five candidate subgraphs to the provided template. Show where the two graphs agree and disagree. Which subgraph matches the template the best?

b) Which key parts of the best match help discriminate it from the other potential matches?

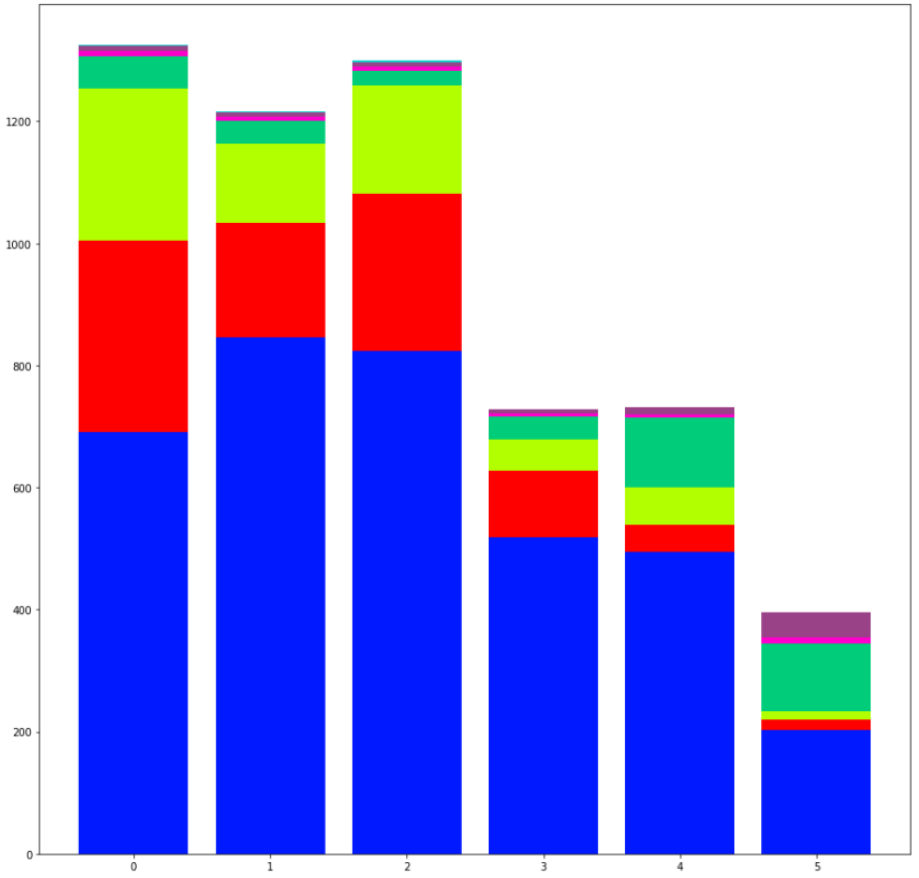
# Question 1 files

- 5 Candidate Subgraphs:
- The subgraphs are the same format as the Template with some differences:

| File name | #rows | #unique Nodes | #location Info | #longitude, Latitude |
|---|---|---|---|---|
| Q1-Graph1.csv | 1216 | 93 | 168 | 168 |
| Q1-Graph2.csv | 1300 | 87 | 201 | 201 |
| Q1-Graph3.csv | 729 | 79 | 88 | 88 |
| Q1-Graph4.csv | 732 | 87 | 176 | 176 |
| Q1-Graph5.csv | 395 | 86 | 124 | 124 |
| CGCS-Template.csv | 1325 | 88 | 301 | 0 |

# Question 1 files comparison based on data distribution on eType

| Name | #rows | #eT0 | #eT1 | #eT2 | #eT3 | #eT4 | #eT5 | #eT6 |
|---|---|---|---|---|---|---|---|---|
| Graph 1 | 1216 | 187 | 131 | 7 | 7 | 1 | 846 | 37 |
| Graph 2 | 1300 | 258 | 177 | 7 | 7 | 4 | 823 | 24 |
| Graph 3 | 729 | 109 | 51 | 6 | 6 | 1 | 519 | 37 |
| Graph 4 | 732 | 45 | 61 | 5 | 12 | 0 | 494 | 115 |
| Graph 5 | 395 | 17 | 14 | 11 | 40 | 0 | 203 | 110 |
| Template | 1325 | 314 | 249 | 9 | 9 | 1 | 691 | 52 |



# Question 2

- CGCS has a set of “seed” IDs that may be members of other potential networks that could have been involved. Take a look at the very large graph. Can you determine if those IDs lead to other networks that matches the template?

# Seeds

- They will act as a starting point for finding hacker groups
  - They only have one line
  - No location information
- Members of our potential groups in the large dataset(availability checked)
- These are the values of the 3 seed files

| File name | Source | eType | Target | Time | Weight |
|---|---|---|---|---|---|
| Q2-Seed1.csv | 600971 | 4 | 579269 | -685755382 | 0.166667 |
| Q2-Seed2.csv | 538771 | 4 | 473043 | -623491200 | 0.0909091 |
| Q2-Seed3.csv | 574136 | 2 | 657187 | 1991785 | 633 |

# Very large Graph (1)

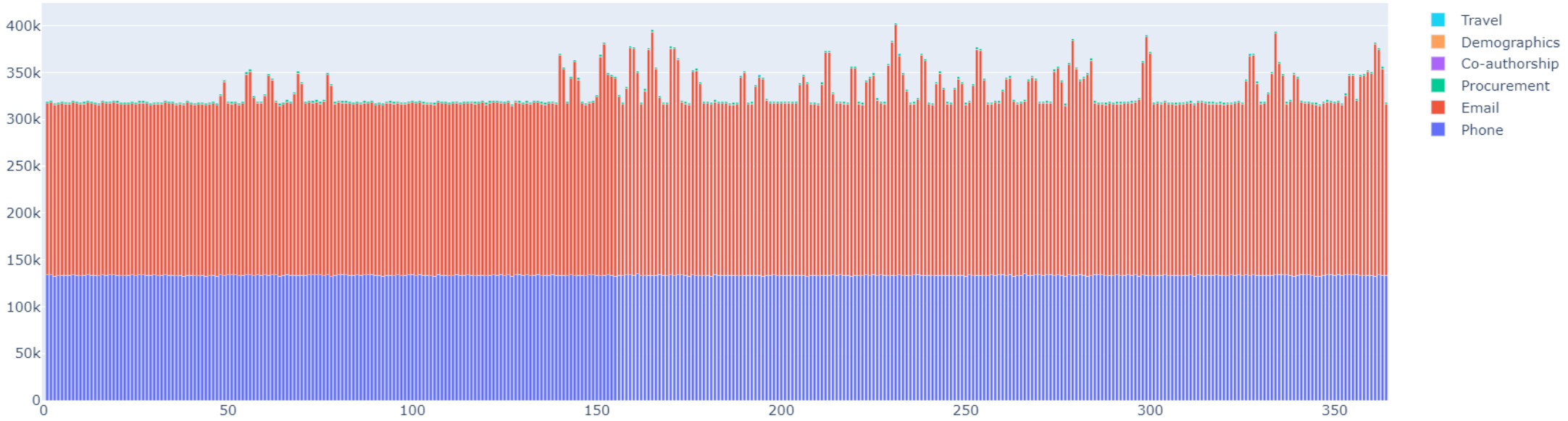
- Filename: CGCS-GraphData.csv
- Containing the data of all the hacker groups
- Connection between different groups are not clear
- Around 124 million rows(edges) and 200860 unique Ids(nodes)
- Around 70 million rows have location information

# Very large Graph (2)

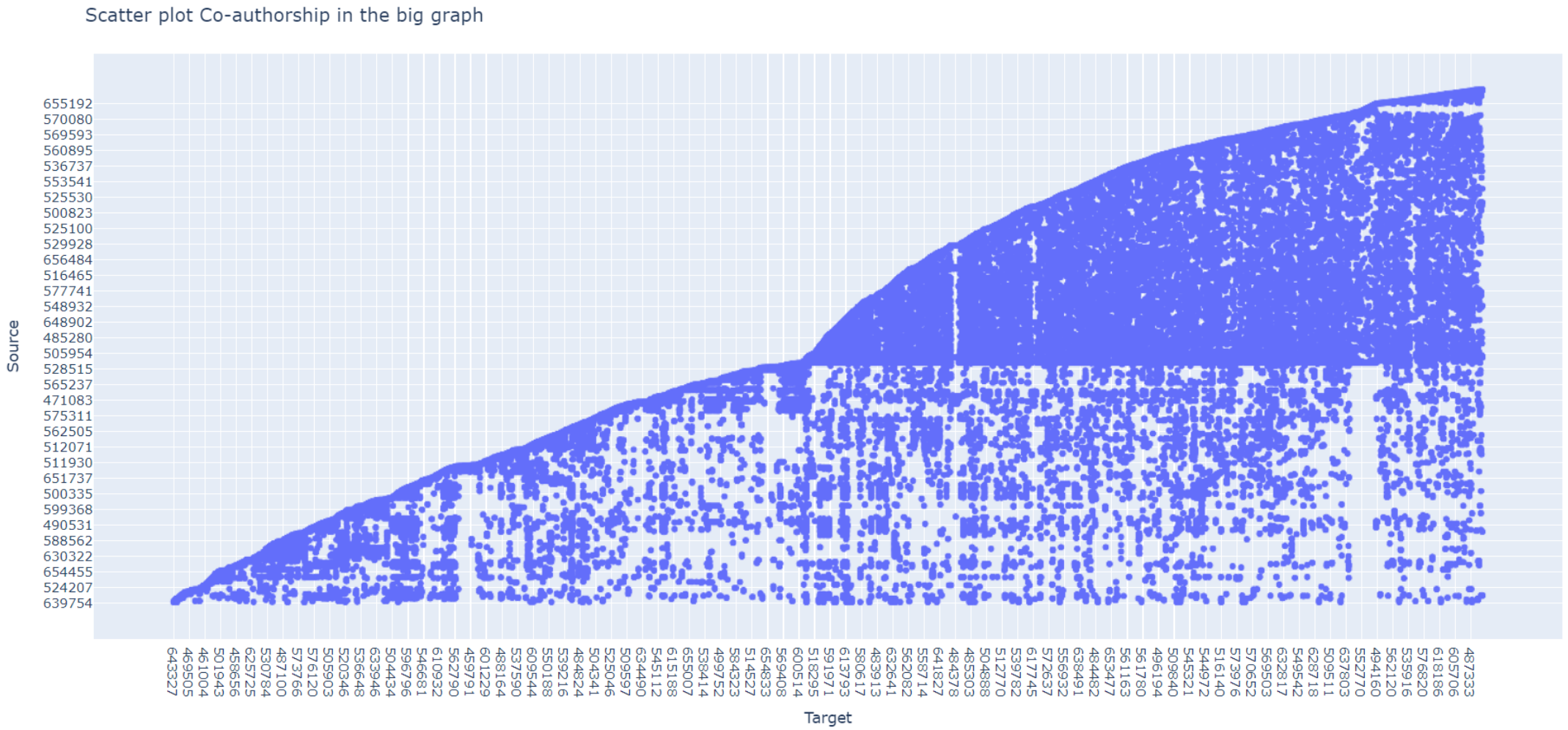
- Number of rows based on eType:

| eType | Number of Rows | Unique Source Rows | Unique Target Rows |
|---|---|---|---|
| 0 | 48662103 | 100000 | 100000 |
| 1 | 70661593 | | |
| 2 | 389211 | 3814 | 2721 |
| 3 | 389211 | | |
| 4 | 259304 | 66173 | 33570 |
| 5 | 2041841 | 100003 | 100027 |
| 6 | 1491998 | 50189 | |
| Total | 123895261 | 164537 | 136324 |

# Very large Graph (3)



# Very large Graph (4) Co-Authorship



# Similarity Measures

- Connection Analysis:
  - Group Analysis : Cluster Coefficient (Transitivity)
  - Network Analysis : Density, Average Path Length, Degree Distribution
- Positional Analysis:
  - Degree : In Degree, Out Degree, All
  - Closeness Centrality
  - Betweenness Centrality
  - Eigen Vector Centrality

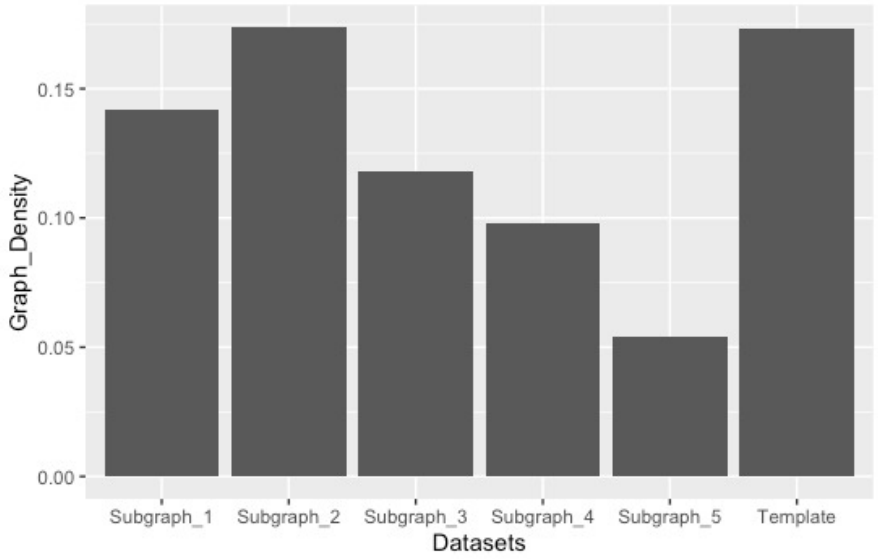
# Connection Analysis

## Density

- Actual Connections/Potential Connections

| Template | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| 0.1730669 | 0.1421225 | 0.1737503 | 0.1183057 | 0.0978348 | 0.05403557 |

- Similarity: Graph2 > Graph1 > Graph3 > Graph4 > Graph5



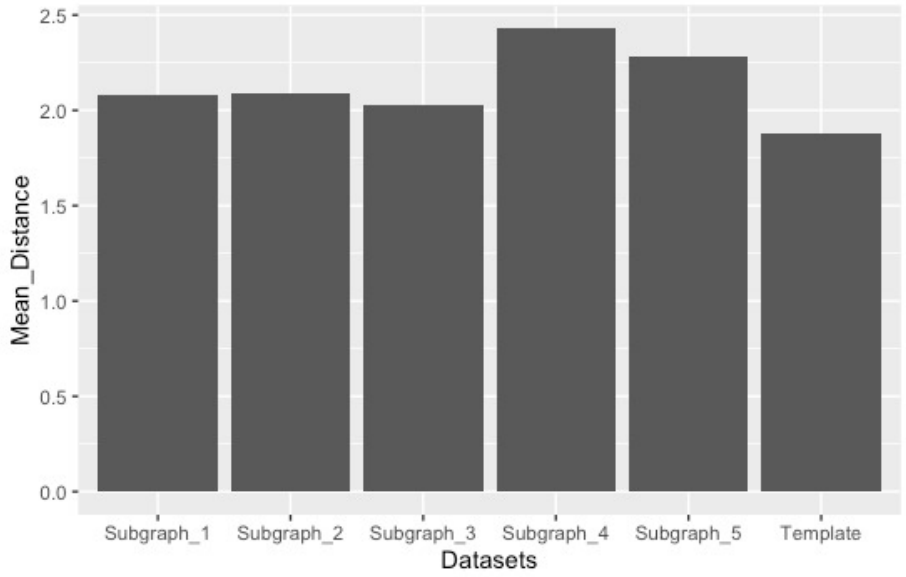
# Connection Analysis

## Average Path Length

- Mean Shortest Path between all nodes.

| Template | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| 1.874689 | 2.083075 | 2.085761 | 2.026447 | 2.429907 | 2.283071 |

- Similarity: Graph3 > Graph1 ~ Graph2 > Graph5 > Graph4



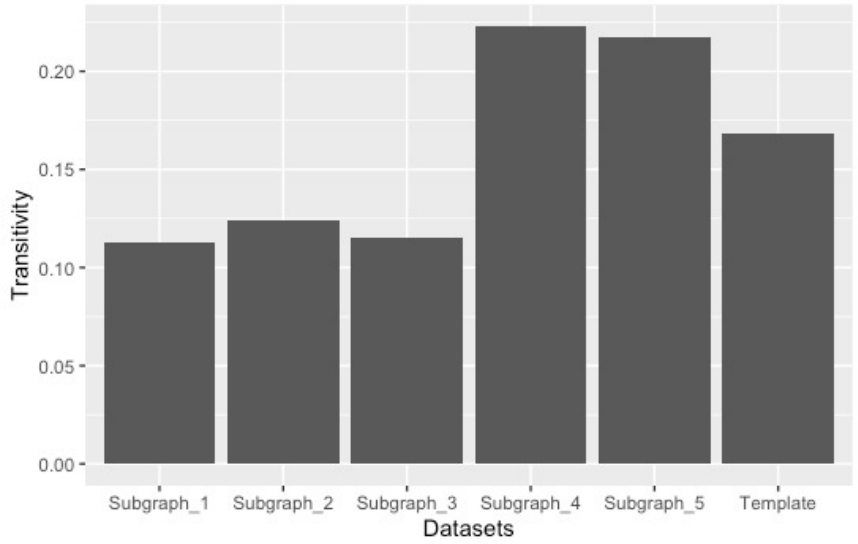
# Connection Analysis

## Cluster Coefficient (Transitivity)

- Measure of the degree to which nodes tend to cluster together.

| Template | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| 0.1685912 | 0.1130306 | 0.1238481 | 0.1151288 | 0.2228648 | 0.217119 |

- Similarity: Graph2 > Graph3 > Graph1 ~ Graph4 > Graph5



# Position Analysis

## Degrees:

- In Degree:

| Template | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| 0-72 | 0-64 | 0-96 | 0-36 | 0-35 | 0-30 |

- Out Degree

| Template | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| 0-136 | 0-77 | 0-96 | 0-50 | 0-59 | 0-67 |

- All

| Template | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| 1-208 | 1-135 | 1-192 | 1-68 | 1-64 | 1-72 |

- Similarity: Graph2 >Graph1 > Graph3 > Graph4 > Graph5

# Position Analysis

## Closeness:

- reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph.
- Thus, the more central a node is, the closer it is to all other nodes.

| Template | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| 0.002949853-0.007042254 | 0.002197802-0.006369427 | 0.003472222-0.007352941 | 0.002624672-0.006802721 | 0.004016064-0.007407407 | 0.003937008-0.007462687 |

- Similarity: Graph3 >Graph1 > Graph2> Graph4 > Graph5

# Position Analysis

## Betweenness:

- The betweenness centrality for each vertex is the number of shortest paths that pass through the vertex.

| Template | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| 0.00-682.1806727 | 0.00-1012.780577 | 0.00-893.3087233 | 0.00-466.4803012 | 0.00-912.652983 | 0.00-912.652983 |

- Similarity: Graph2 >Graph3 > Graph4 = Graph5 > Graph1

# Position Analysis

## Eigen Vector Centrality:

- **Eigenvector centrality** (also called **eigencentrality**) is a measure of the influence of a node in a network.
- It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes

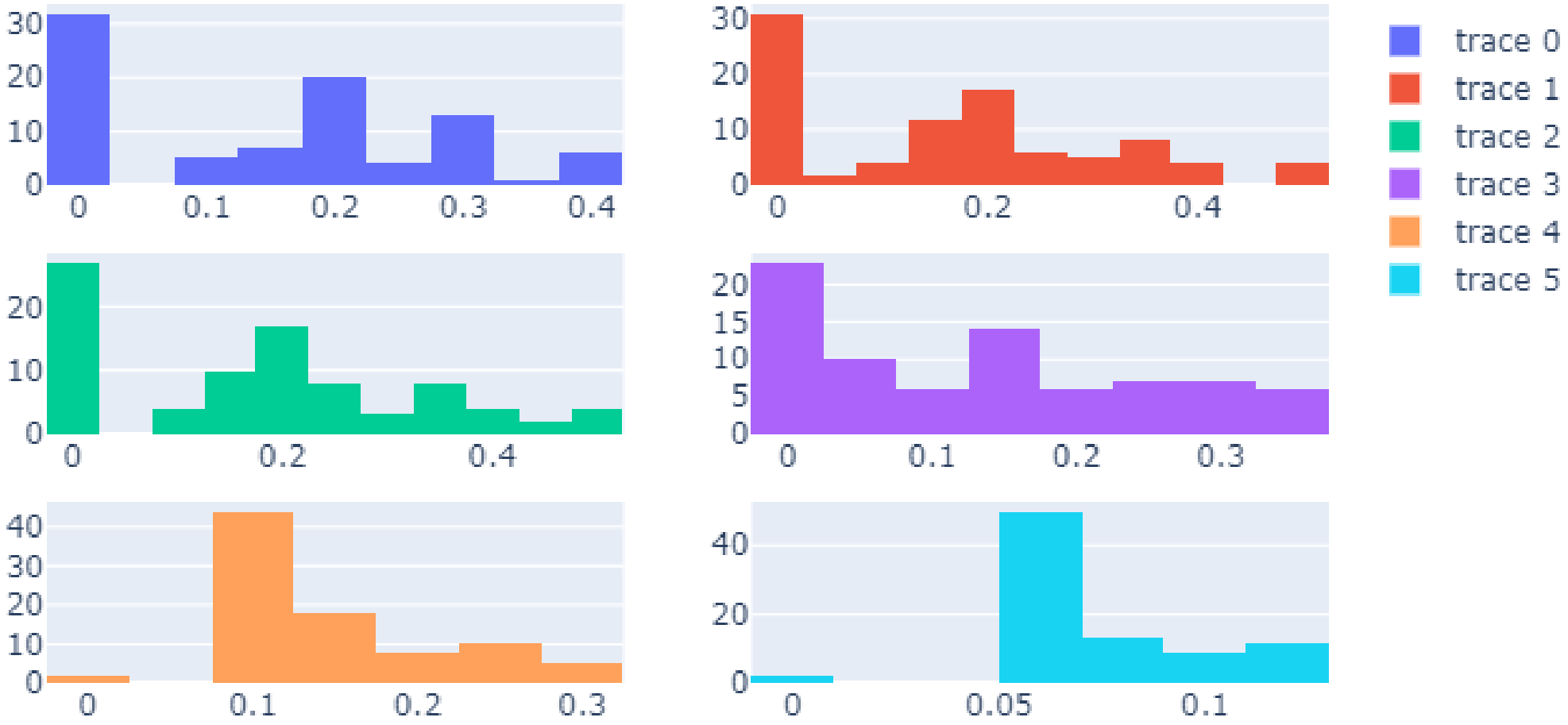
| Template | G1 | G2 | G3 | G4 | G5 |
|---|---|---|---|---|---|
| 2.889807e-05-1.000000 | 8.486762e-07-1.000000 | 0.000202756-1.0000000 | 1.370524e-06-1.0000000 | 0.008849692-1.0000000 | 0.006716513-1.0000000 |
| 6 Nodes > 0.5 | 6 Nodes > 0.5 | 6 Nodes > 0.5 | 3 Nodes > 0.5 | 8 Nodes > 0.5 | 8 Nodes > 0.5 |

- Similarity: Graph2 >Graph1 > Graph3 > Graph4 = Graph5

# Closeness Centrality

To determine the central nodes in networks, the closeness centrality measure considers the nodes that have the smallest average path length (sequence of relationships) for the nodes that are linked to other nodes.

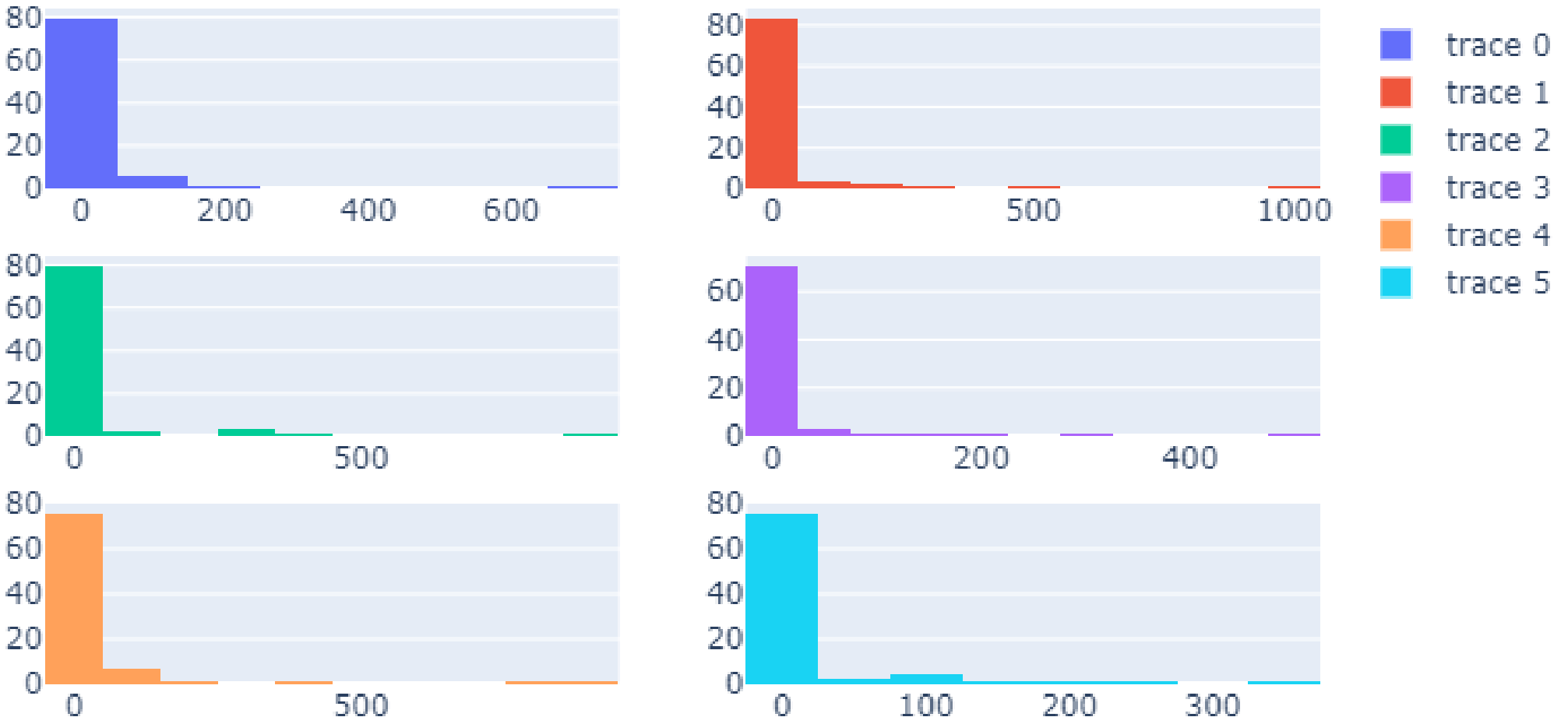
# Closeness Centrality



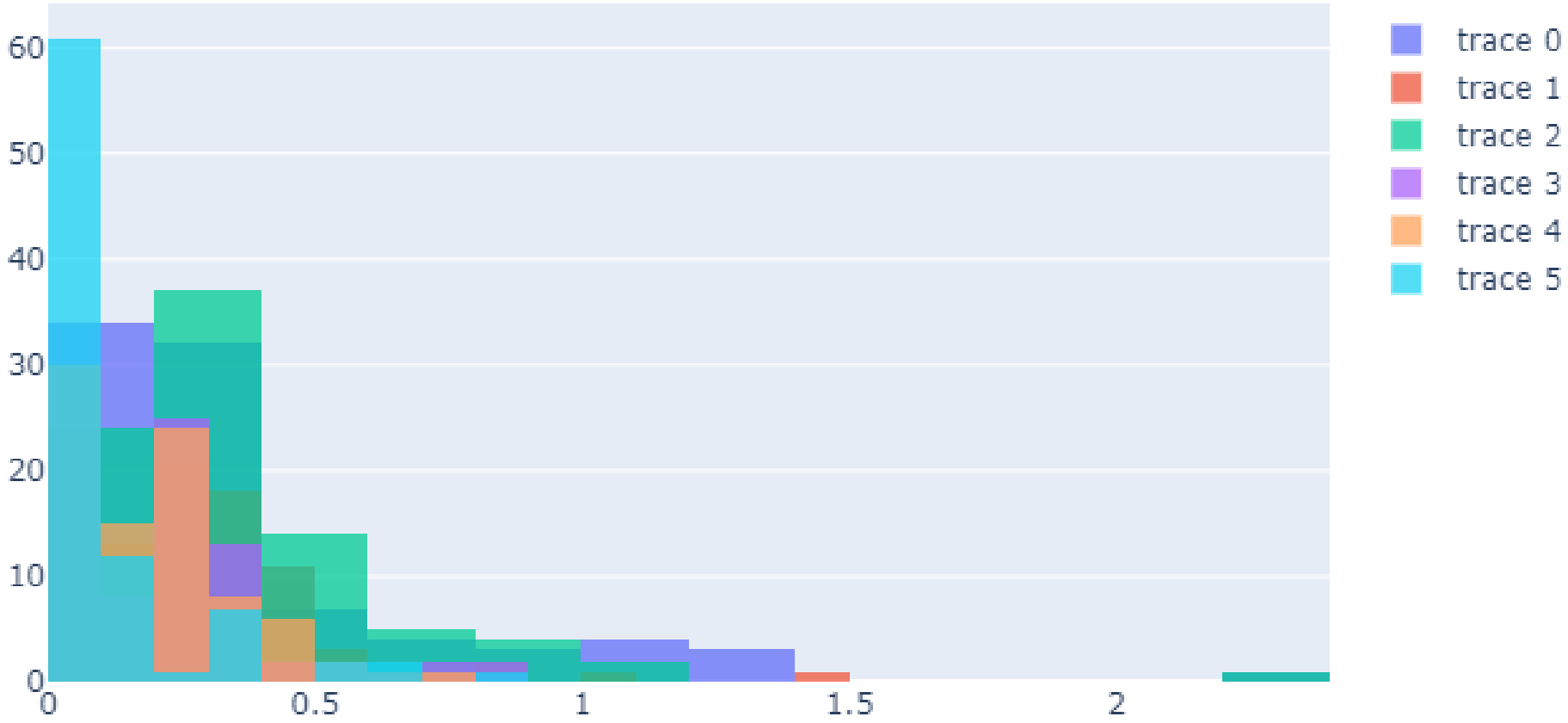
# Betweenness

Nodes that occur on many shortest paths between other nodes in the graph have a high betweenness centrality score.

# Betweenness



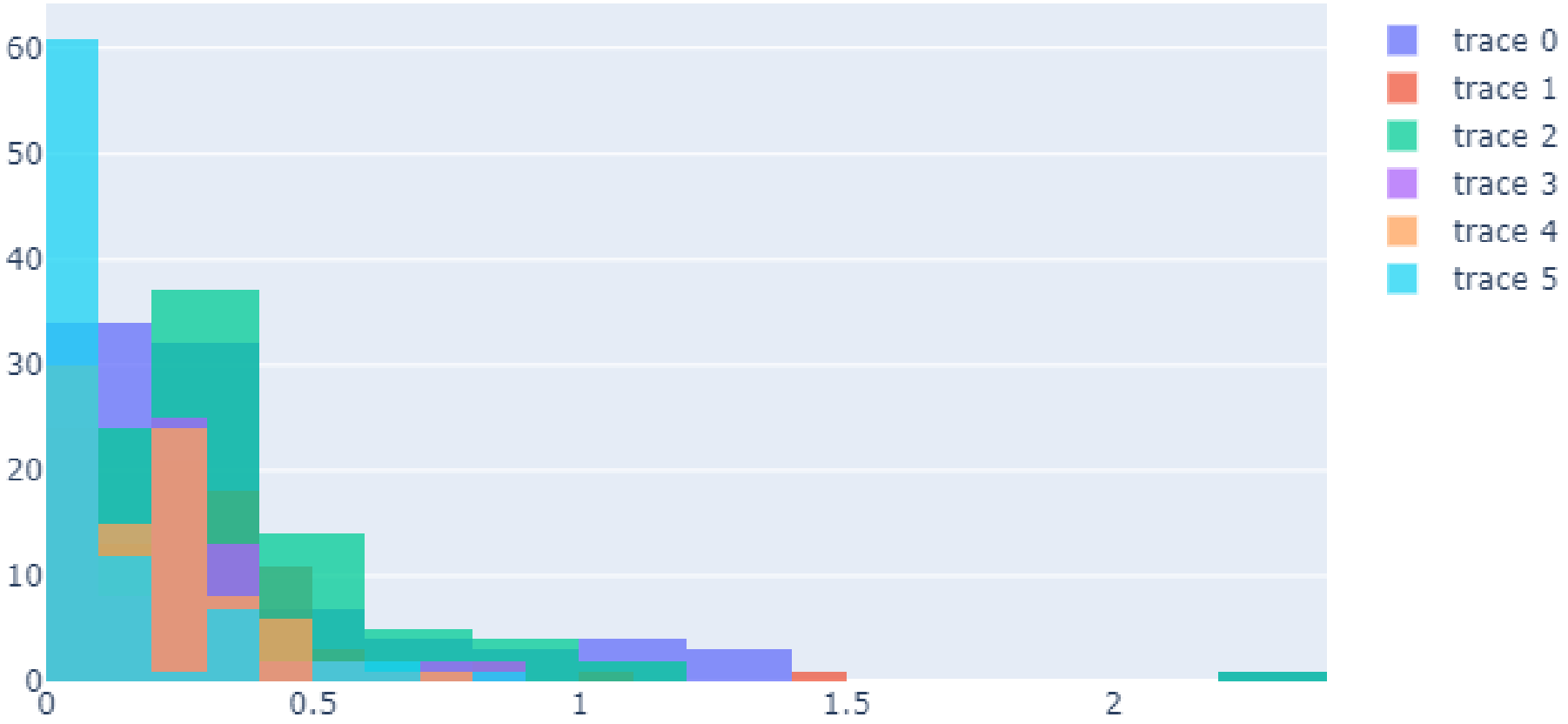
# Degree


60
50
40
30
20
10
0
0
0.5
1
1.5
2
trace 0
trace 1
trace 2
trace 3
trace 4
trace 5

# Degree Centrality

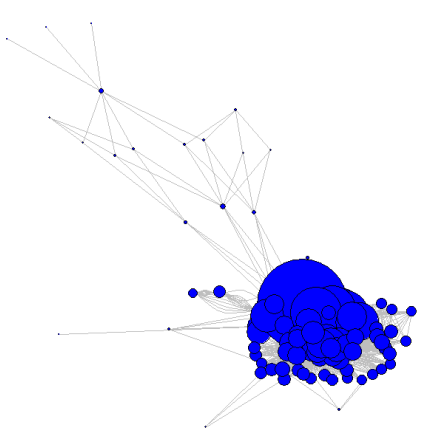
Degree centrality considers the node with the highest degree (largest number of
connections) as the most central node in the network. Degree centrality focuses on
individual nodes—it simply counts the number of edges that a node has.

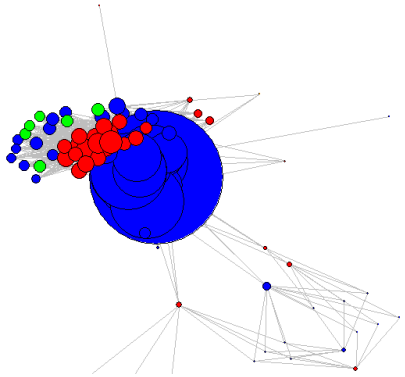
# Degree Centrality

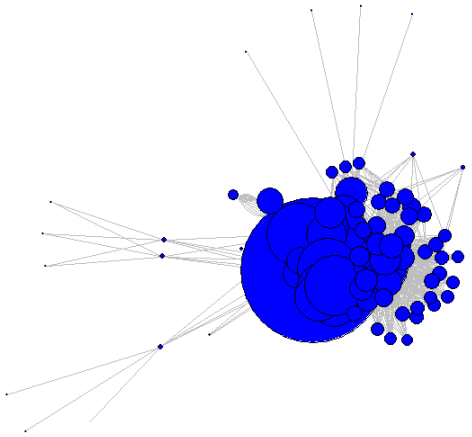


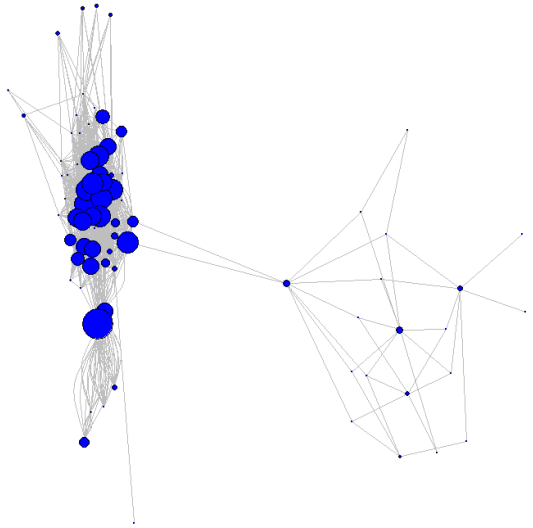
# Visualisations

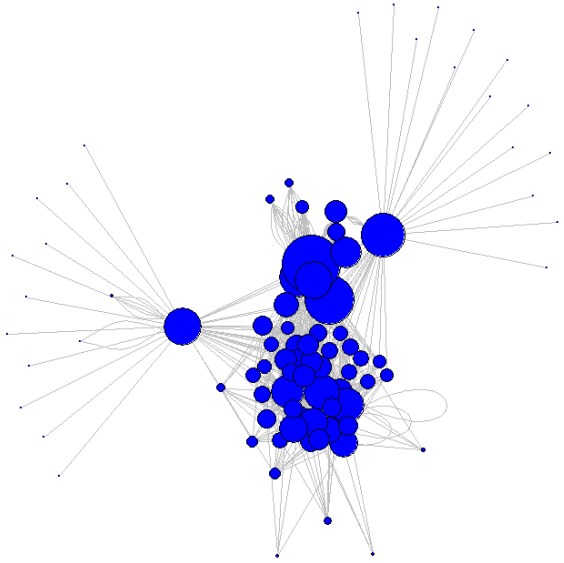
## Degree (All)

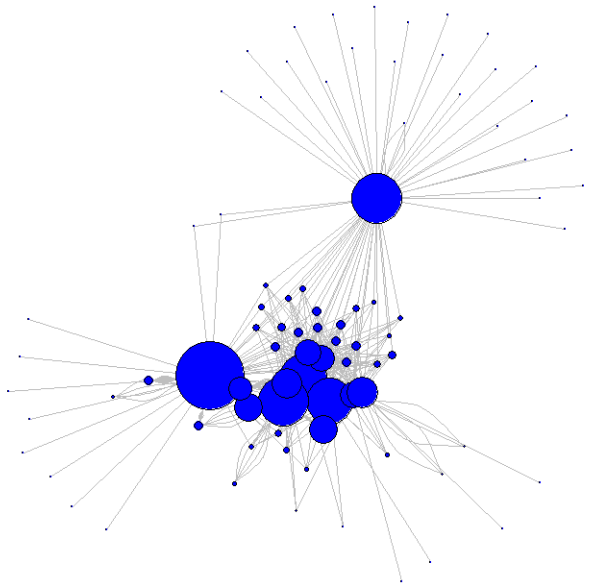






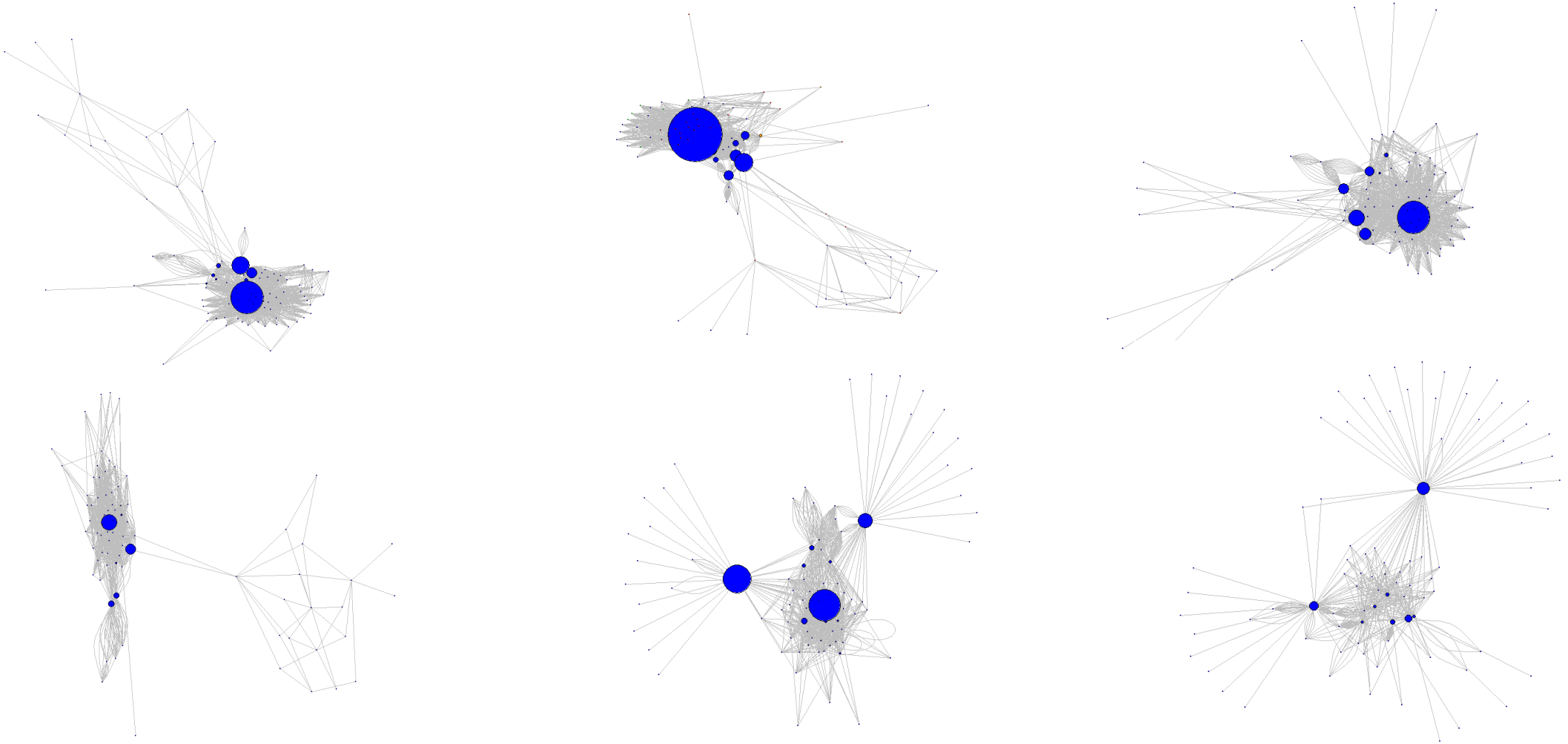






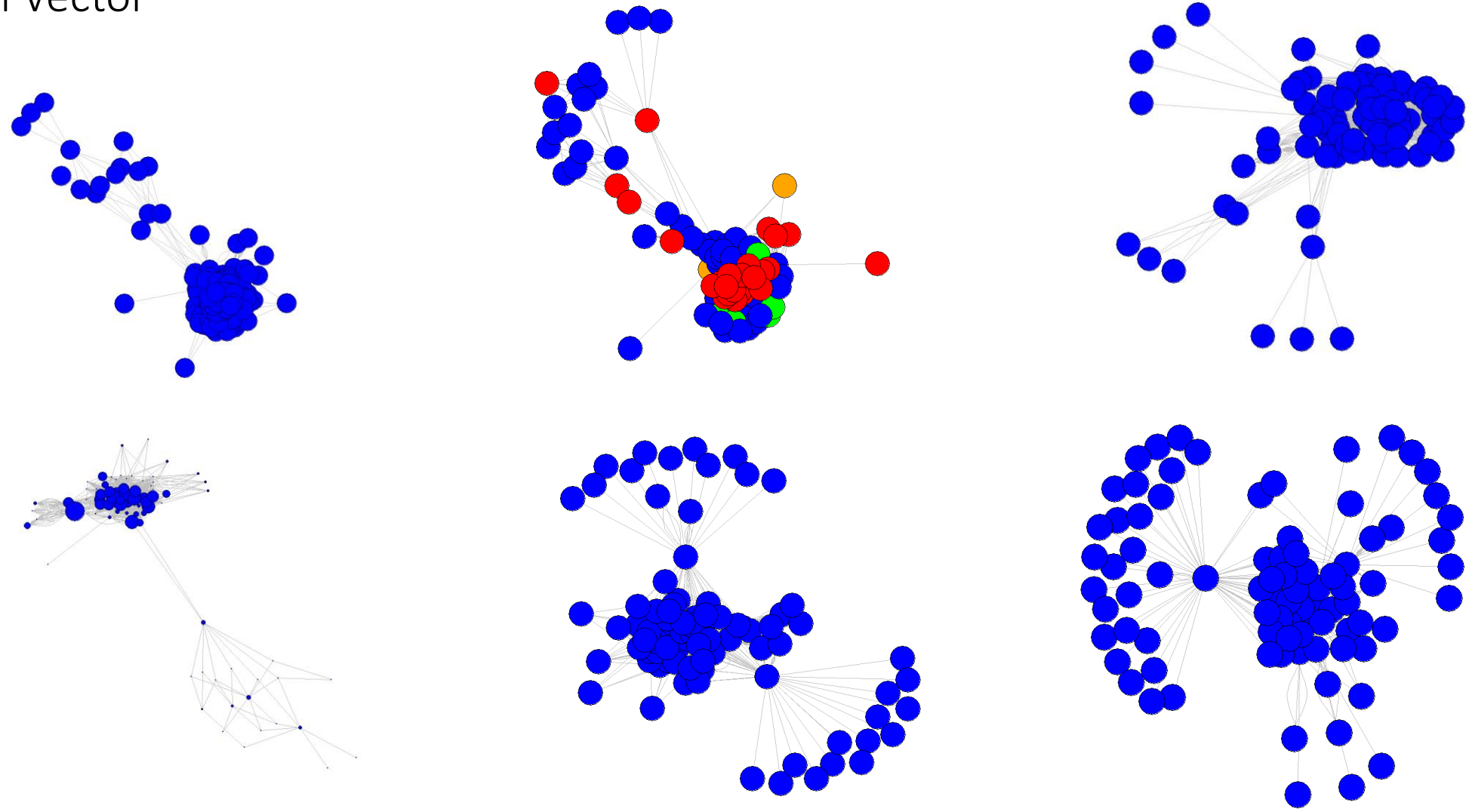
# Visualisation

## Betweenness



# Visualisation

## Eigen Vector




Gephi is an open-source network analysis and visualization software package

# Discussion

- Literature and keywords
- Questions were sent to the committee
- Time based Graph in Very large graph (here)
- Seeds for question 2:
  - They are most probably connected to most of the data
  - Assume they are in a hacker group or
  - Assume they are a group that could be hackers

# What next

- Other similarity measures
- Parallel coordinates
  - Extract interesting measures
  - Comparison
- Question 2: further analysis of seeds
  - Use some of the similarity measure as thresholds
  - Ego graph
- Analysis based on channels

# Thank you for your time