Entry Name: **"OVGU-Beladi-MC1"**
VAST Challenge 2020
Mini-Challenge 1

Team Members:
Seyedbehnam Beladi, Otto von Guericke University Magdeburg, seyedbehnam.beladi@ovgu.de
PRIMARY
Atrayee Neog, Otto von Guericke University Magdeburg, atrayee.neog@ovgu.de
Xiongjun Wang, Otto von Guericke University Magdeburg, xiongjun.wang@st.ovgu.de
Uli Niemann, Institute of Technical and Business Information Systems, University of Magdeburg, Germany, uli.niemann@ovgu.de
Kai Lawonn, Institute of Computer Science, University of Jena, Germany, kai.lawonn@uni-jena.de
Bernhard Preim, Department of Simulation and Graphics, University of Magdeburg, Germany, bernhard@isg.cs.uni-magdeburg.de
Monique Meuschke, Department of Simulation and Graphics, University of Magdeburg and Institute of Computer Science, University of Jena, Germany, meuschke@isg.cs.uni-magdeburg.de

Student Team: NO

Tools Used:
R/RStudio
Python/Colab/Jupyter/dash/Networkx
Tableau
Gephi

Approximately how many hours were spent working on this submission in total?
900 hours in total

May we post your submission in the Visual Analytics Benchmark Repository after VAST Challenge 2020 is complete? *YES*

Video
https://youtu.be/unTamN8LeDU

Link to entry form on GitHub:
https://atrayeeneog.github.io/VAST-Challenge-2020-MC1/

Center for Global Cyber Strategy (CGCS) researchers have used the data donated by the white hat groups to create anonymized profiles of the groups. One such profile has been identified by CGCS sociopsychologists as most likely to resemble the structure of the group who accidentally caused this internet outage. You have been asked to examine CGCS records and identify those groups who most closely resemble the identified profile

Questions

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 subgraph. Show where the two graphs agree and disagree. Which subgraph matches the template subgraph the best? Please limit your answer to seven images and 500 words.

Answer:

The starting point to compare the subgraphs is using their graph attributes as similarity measures. Figure 1 visualizes eight measures in a parallel coordinate plot (PCs) for each subgraph. Each subgraph was then compared to template

subgraph. Subgraph 1 and subgraph 2 appear to have more patterns matching the template compared to other subgraphs.

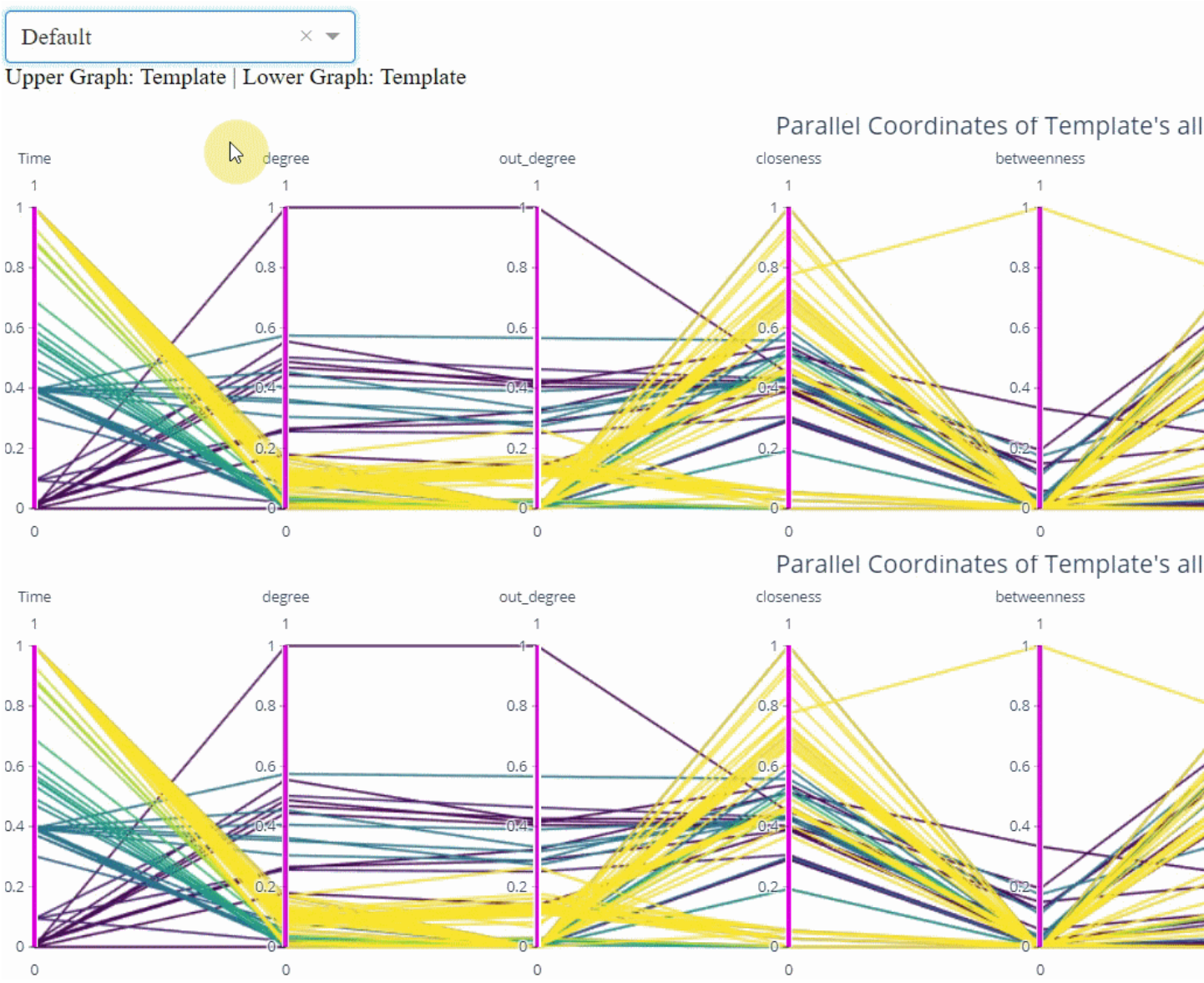

Figure 1. Parallel coordinates allow us to observe the distribution of all the similarity measures at the same time. Color-scale is based on time (0-1): (1.1.2025-1.1.2026)

To quantify these comparisons, a Wasserstein-based test (Ramdas et al. 2017) was done between each subgraph and template for each measure. The Wasserstein metric measures the similarity between two probability distributions by estimating the cost of turning one distribution into another. The test outputs a Wasserstein metric, p-value (depicting confidence in the similarity) and three

contributing factors: network size, network shape and nodes location in the network.

Most measures in Figure 2 show smaller Wasserstein metrics with higher confidence factor denoted by the p-value and more similar distributions between subgraph 1 and subgraph 2 and template.

Figure 2. Comparison of subgraphs for each distribution measure. Distribution curves are compared in each cell. Wasserstein metric (W) and p-value (p) are also written. Contribution percentage of the factors (location, size and shape) are presented by pie-charts.

Based on high p-values, three measures were selected to further investigate similarities and dissimilarities between the subgraphs. Figure 3, Figure 4 and

Figure 5 visualize the subgraphs network. Node sizes are proportional to the magnitude of the measures.

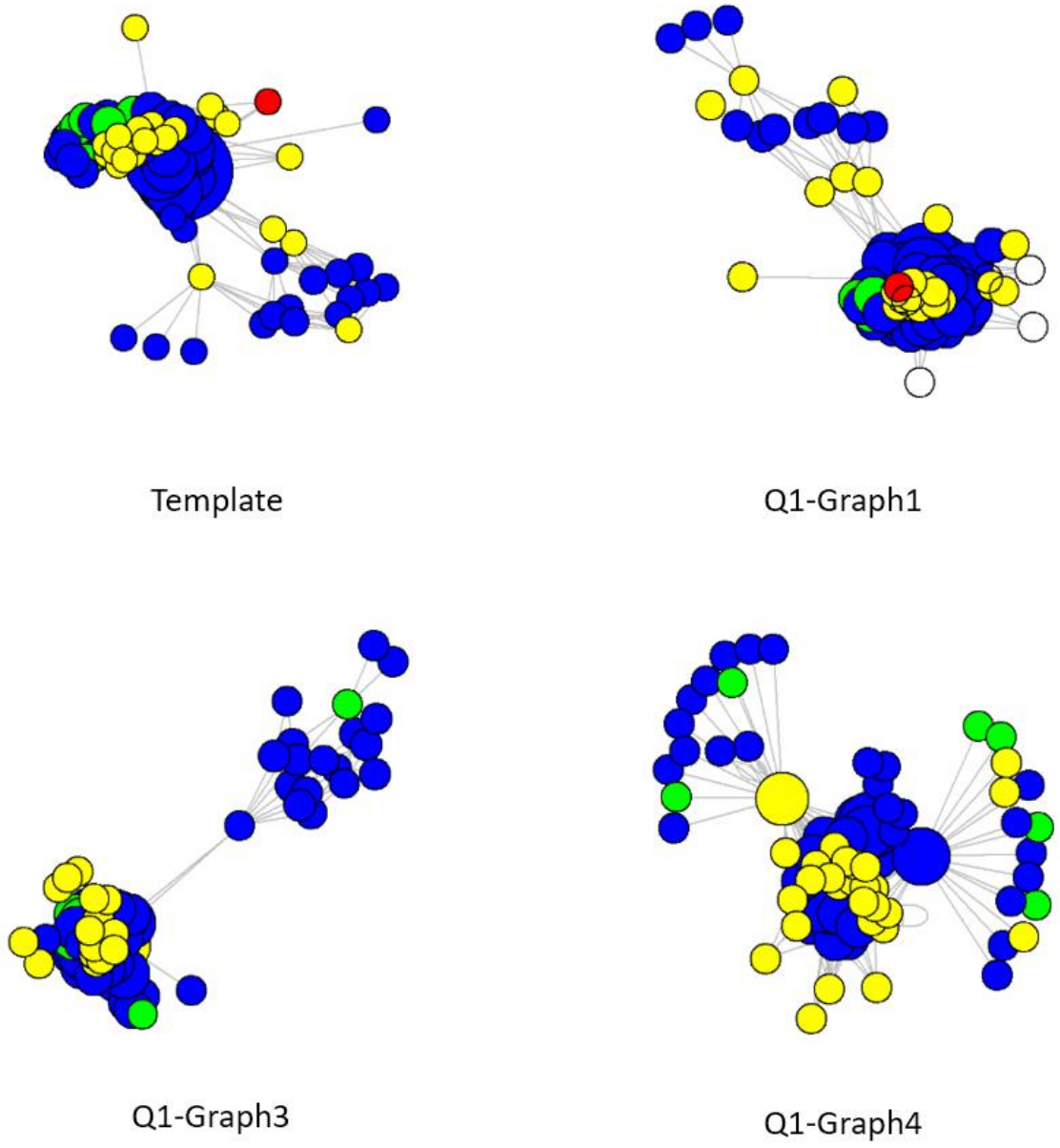

Figure 3. ***Out-Degree*** *measures the number of edges going out from each node in a directed graph. In this particular case, out-going nodes usually belong to person IDs. Since the core of a hacker group is its people, Out-Degrees help emphasize their activities in the network.*

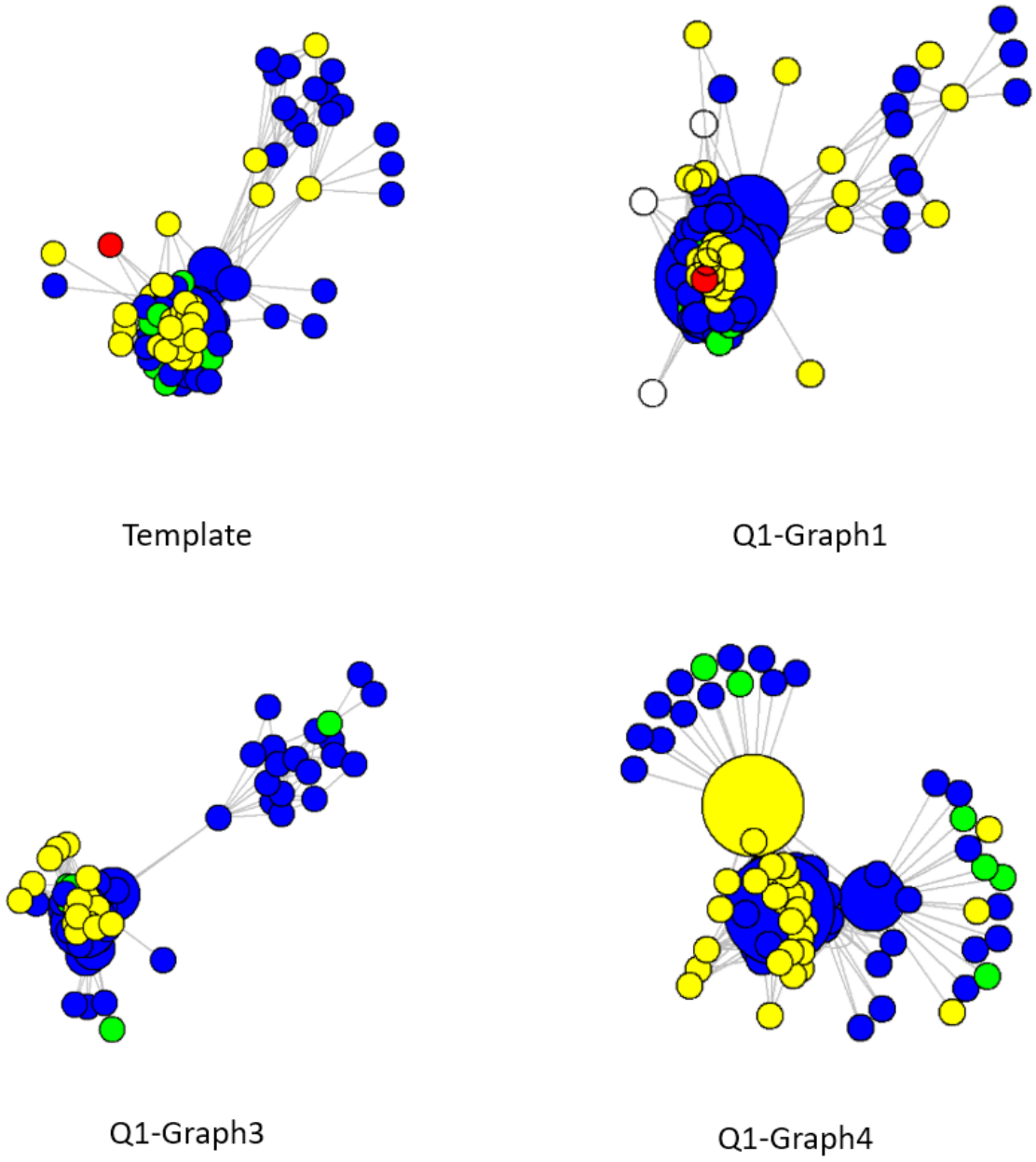

Figure 4. ***Betweenness centrality*** *measures the number of shortest paths passing through a node and the importance of it based on how central it is in the network. Therefore, this measure helps us compare central and important nodes.*

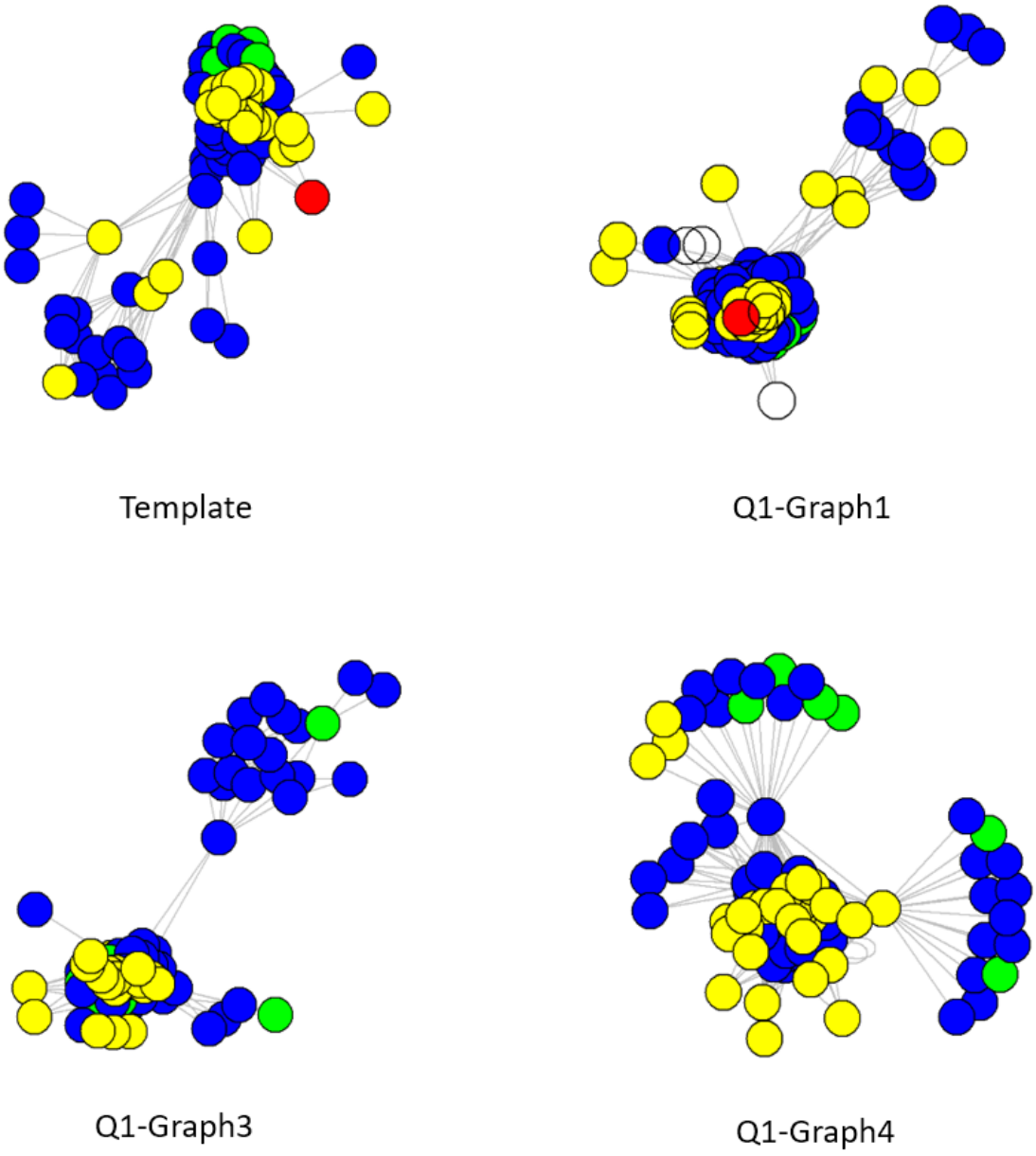

Figure 5. ***Eigenvector centrality****. In order to extract a network of hackers, who caused the outage, it is important to understand how the central nodes are connected to each other. Nodes with high eigenvector score are connected to many nodes which themselves have high scores.*

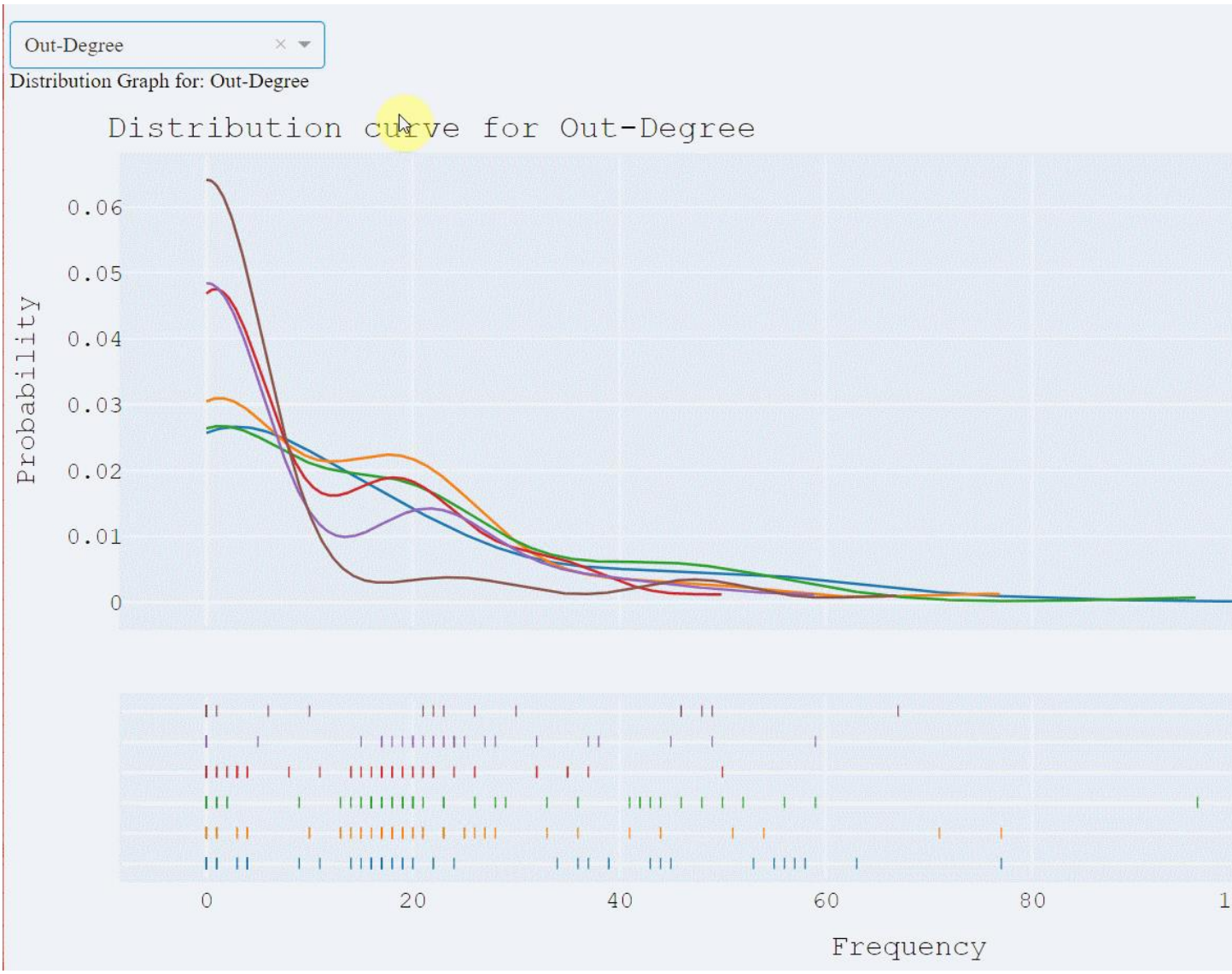

Figure 6. Visualization of the estimated probability density and individual values as "rugs" for out-degree, betweenness centrality and eigenvector centrality.

Eigenvector Centrality measure shows dissimilarity between template and subgraph 2. As seen from the pie charts in Figure 2, the shape of the networks is the largest contributing factor for this difference, meaning the central nodes are distributed differently in template and subgraph 2. The Page Rank (normalized

form of Eigenvector Centrality) nullifies this difference so the similarity between the two is again high for this measure.

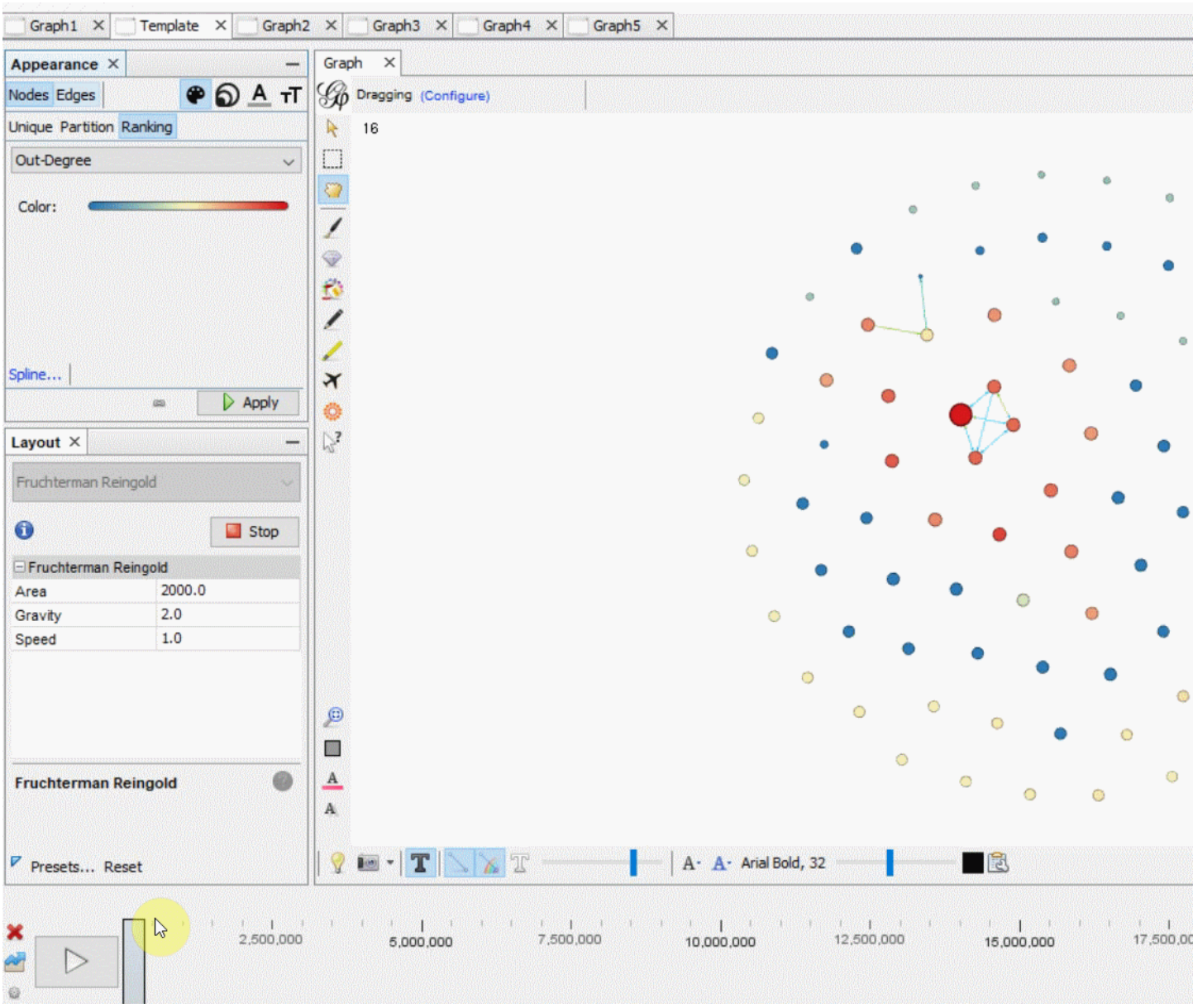

Figure 7. Visualizing and comparing the network graphs, network analysis and changes based on time in Gephi 0.9.2 201709241107 software.

With all the above rationale in mind, we can conclude that ***subgraph 2 is the most similar to the template subgraph.***

b. Which key parts of the best match help discriminate it from the other potential matches? Please limit your answer to five images and 300 words.

Answer:

To further analyze the key similarities and differences between subgraphs, an in-depth analysis was done based on each channel. Certain patterns emerged when visualizing these channels which further proved that subgraph 2 is the most similar to the template.

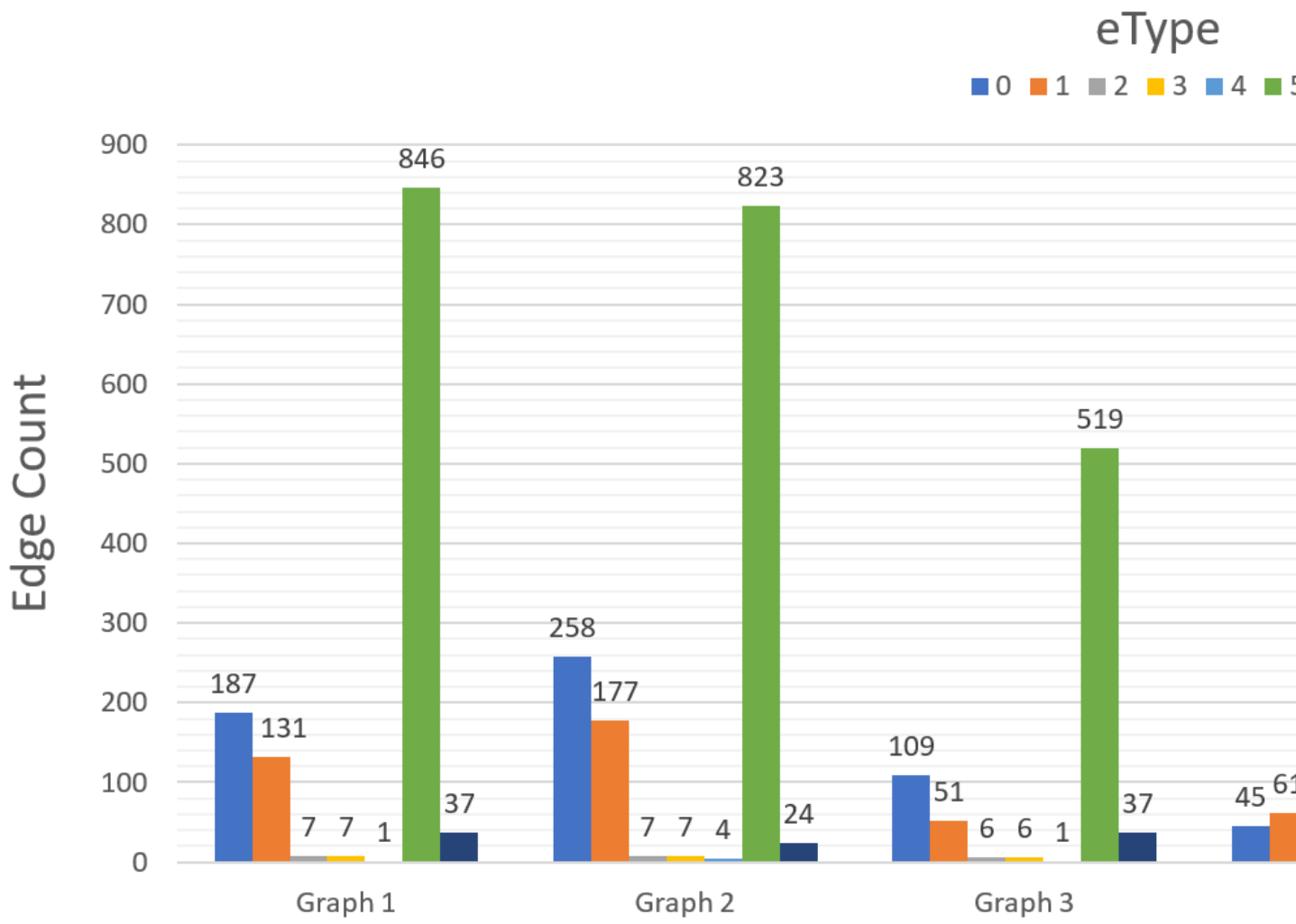

Figure 8. Overview of count of edges for each "eType". The plot helps visualize the relative importance of each channel in the various graphs and is a starting point for our analysis. For instance, subgraph 4 and subgraph 5 don't have a co-authorship channel unlike the others and hence this channel could be ignored.

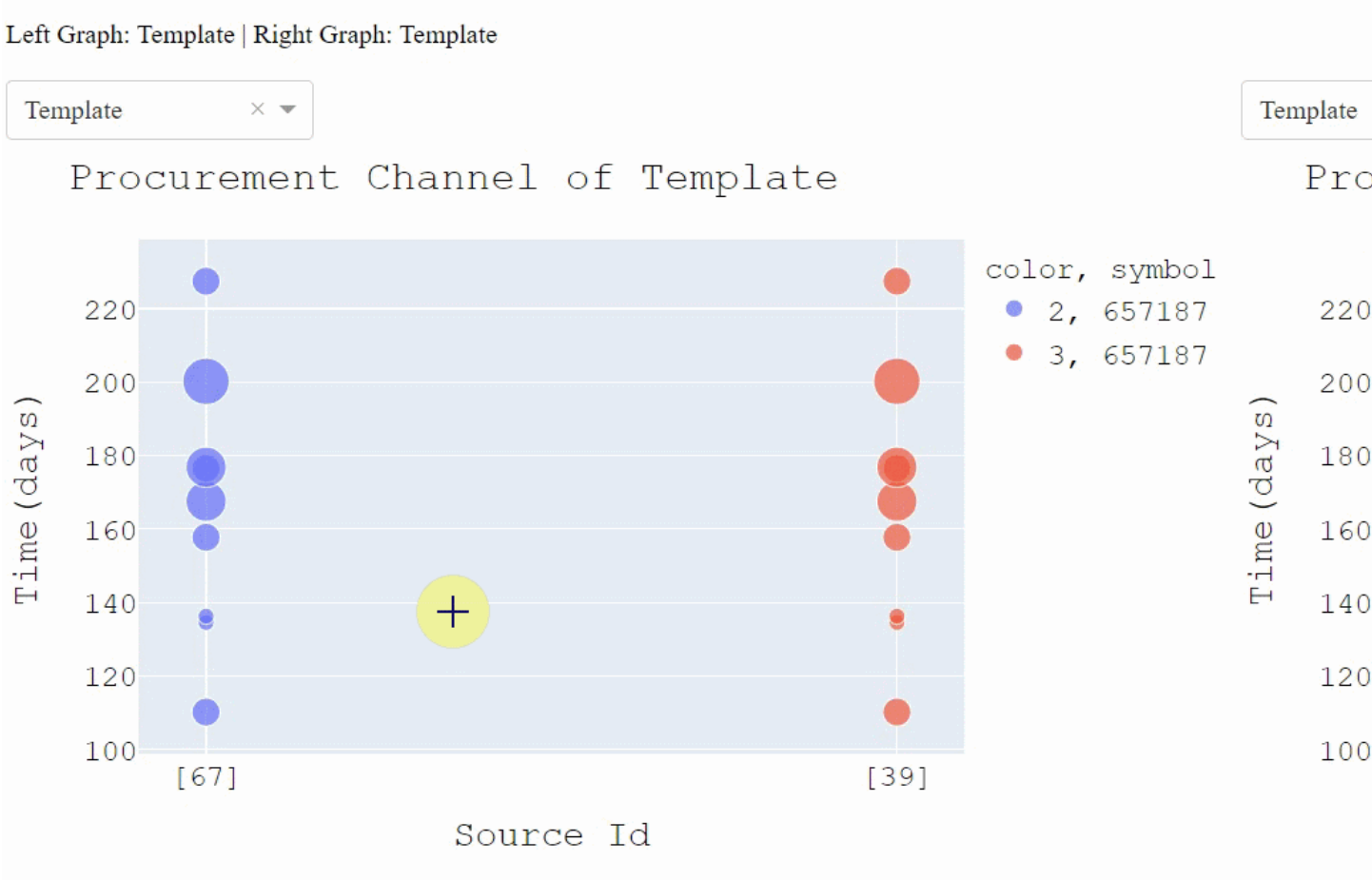

Figure 9. ***Procurement channel****. In subgraph 1, subgraph 2 and subgraph 3 there is only one item sold from one seller (Source) to one buyer (Source) which is same as the template. There were no such distinct transactions for subgraph 4 and subgraph 5. Color: "eType", shape: "item_ID", size: "weight" (cost)*

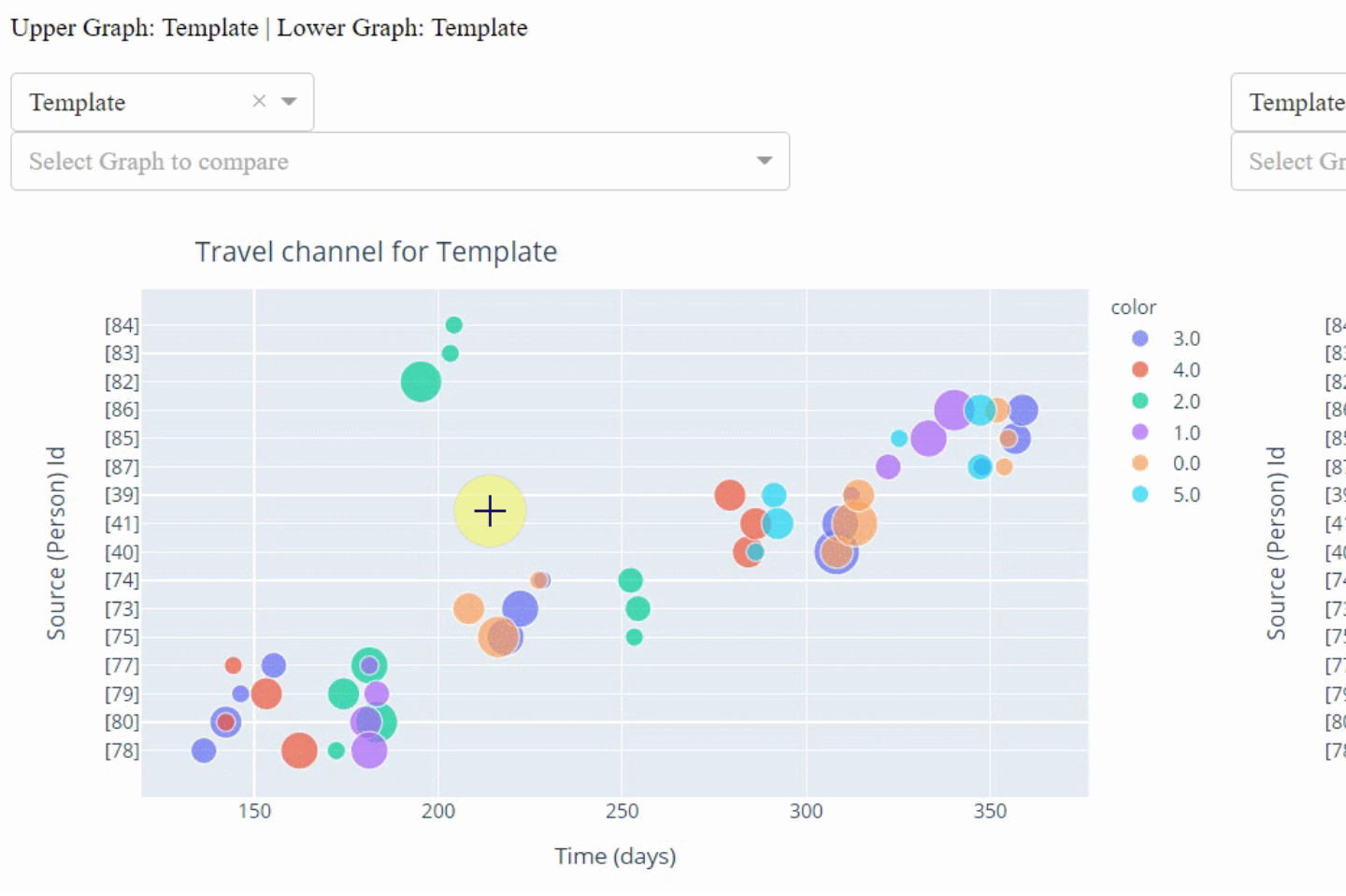

Figure_10. ***Travel Channel****. Colors distinguish countries, size the length of travel and travel records are being filtered based on targets. Similar clusters seem to appear in the template, subgraph_1, subgraph_2 and subgraph_3 with people travelling to the same "TargetLocation" from the same "SourceLocation" at close points in time.*

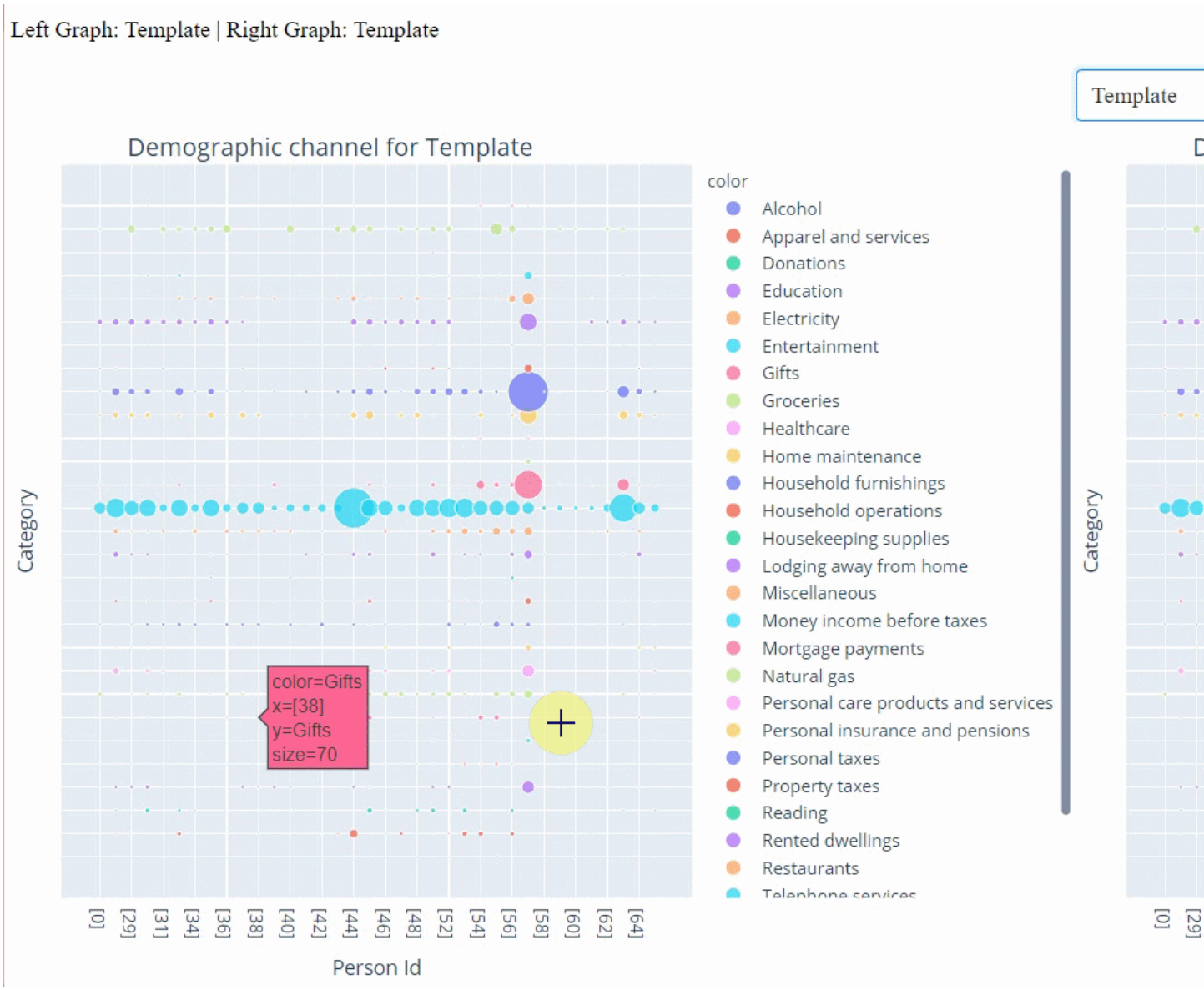

Figure_11. ***Demographic Channel****. Data was reported for the end of the year. Subgraph_2 and subgraph 3, show similar patterns to template in the various categories unlike other subgraphs.*

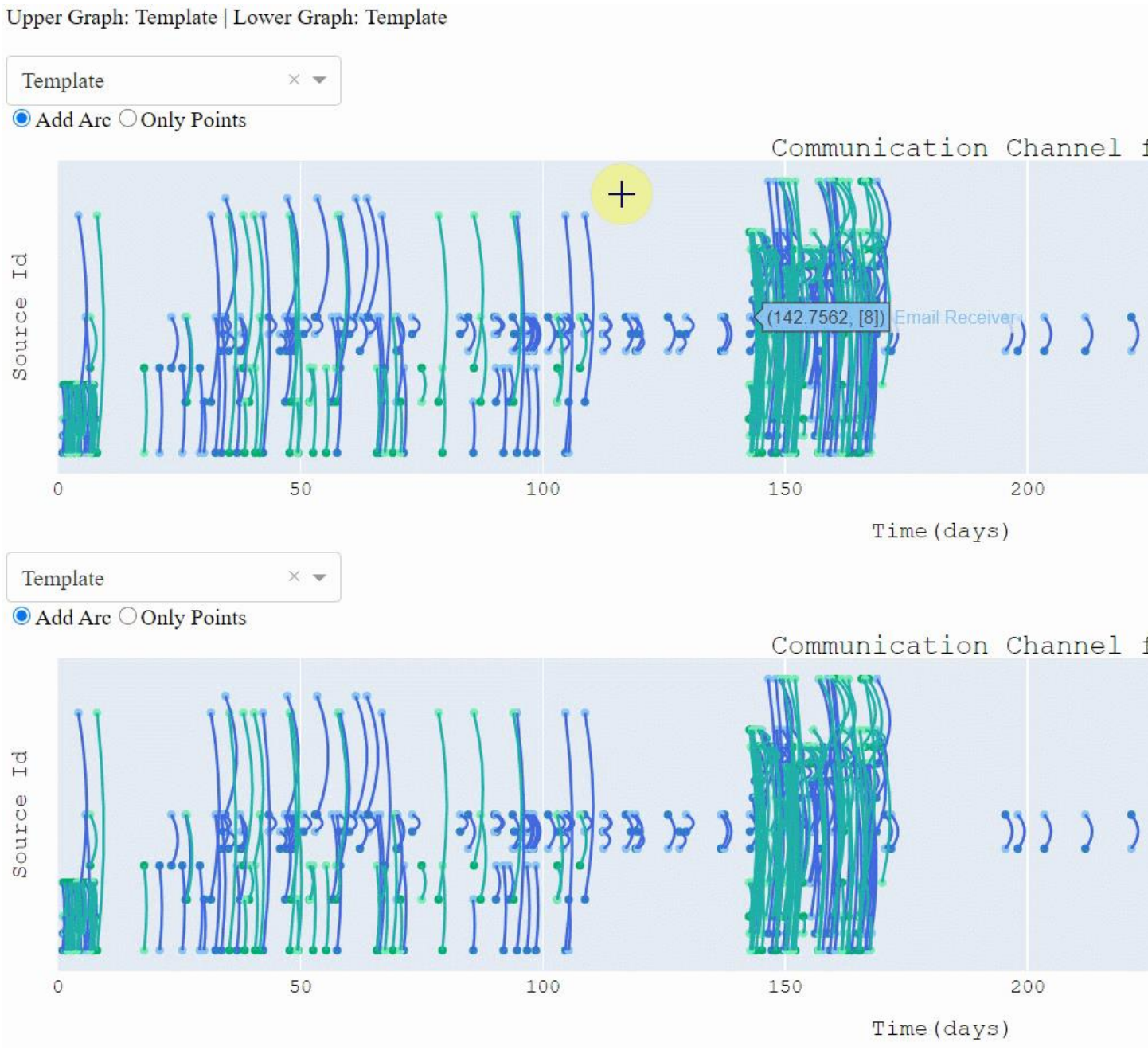

Figure_12. ***Communication Channel****. Each arc represents a call or an email exchange between Sources. Communication's frequency based on time best fits the subgraph_2 to the template while subgraph 1 and subgraph 3 show the same patterns unlike subgraph 4 and subgraph 5.*

Subgraph 1, subgraph 2 and subgraph 3 each could match the template to some extent. However, due to the importance of the dominating channels like communication alongside the aforementioned network analysis, we believe that **subgraph 2 is the best match and most similar to the template subgraph.**

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 subgraph? Describe your process and findings in no more than ten images and 500 words.

Answer:

Our process of building the graphs from the seeds could be generalized into four main stages:

- **Stage 1:** The seed was located in the large graph.
- **Stage 2**: All nodes directly connected to person's IDs in the seed, were found.
- **Stage 3**: Nodes corresponding to people were separated based on channels. Then nodes were analyzed based on channels and only the most prominent nodes were kept. For instance, for the Communication channel, nodes with less than or equal to the minimum Closeness centrality were considered less prominent and removed. Finally, every channel data for the remaining person nodes were added and an extended graph was created.
- **Stage 4**: Based on existing patterns in the template's channels, the extended graph was then reduced in each channel to best resemble the template.

Seed 2 did not go beyond the third stage because almost all the retrieved nodes had only co-authorship data.

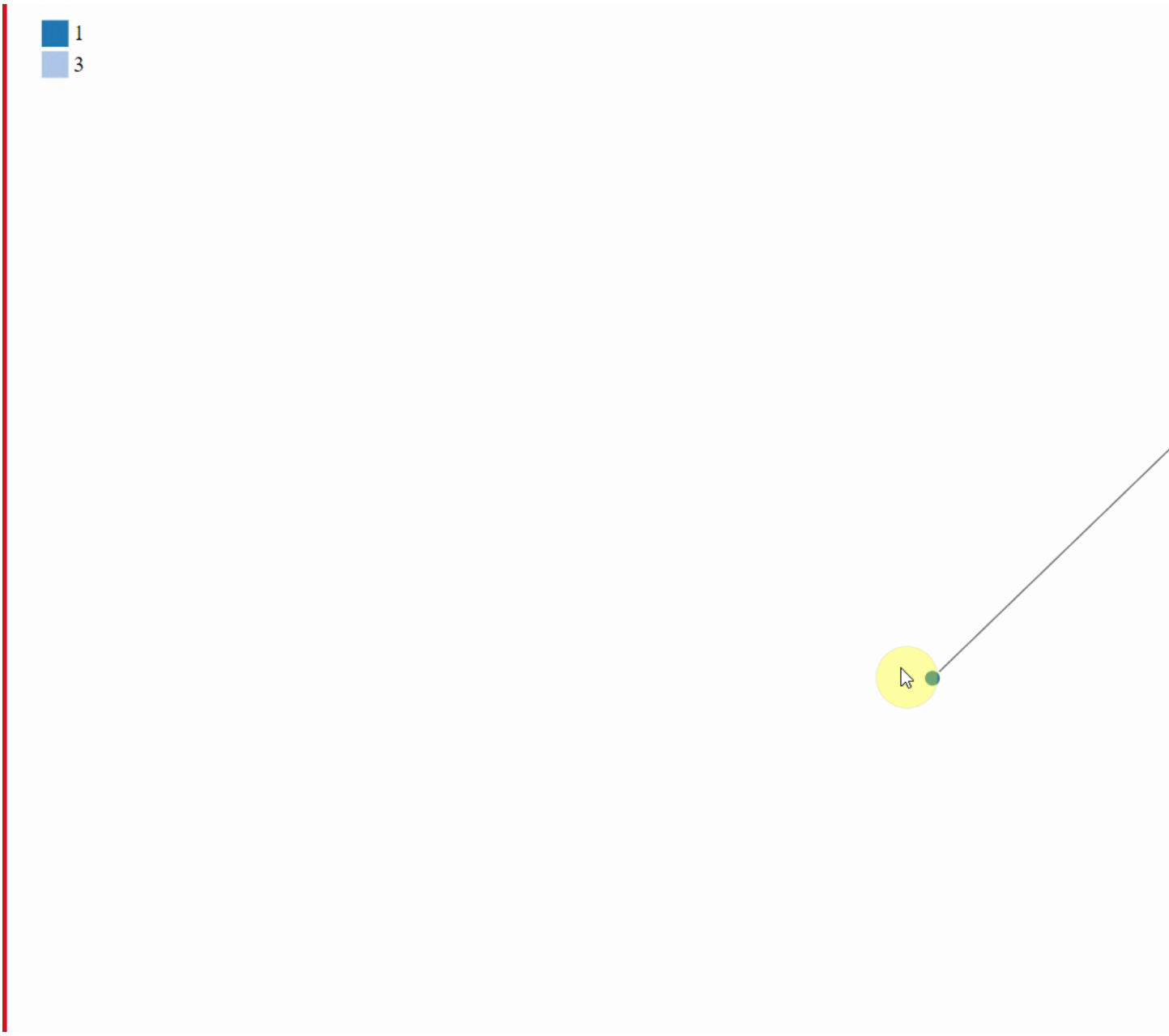

*Figure 13. Animated creation of a network graph based on **seed 1**.*

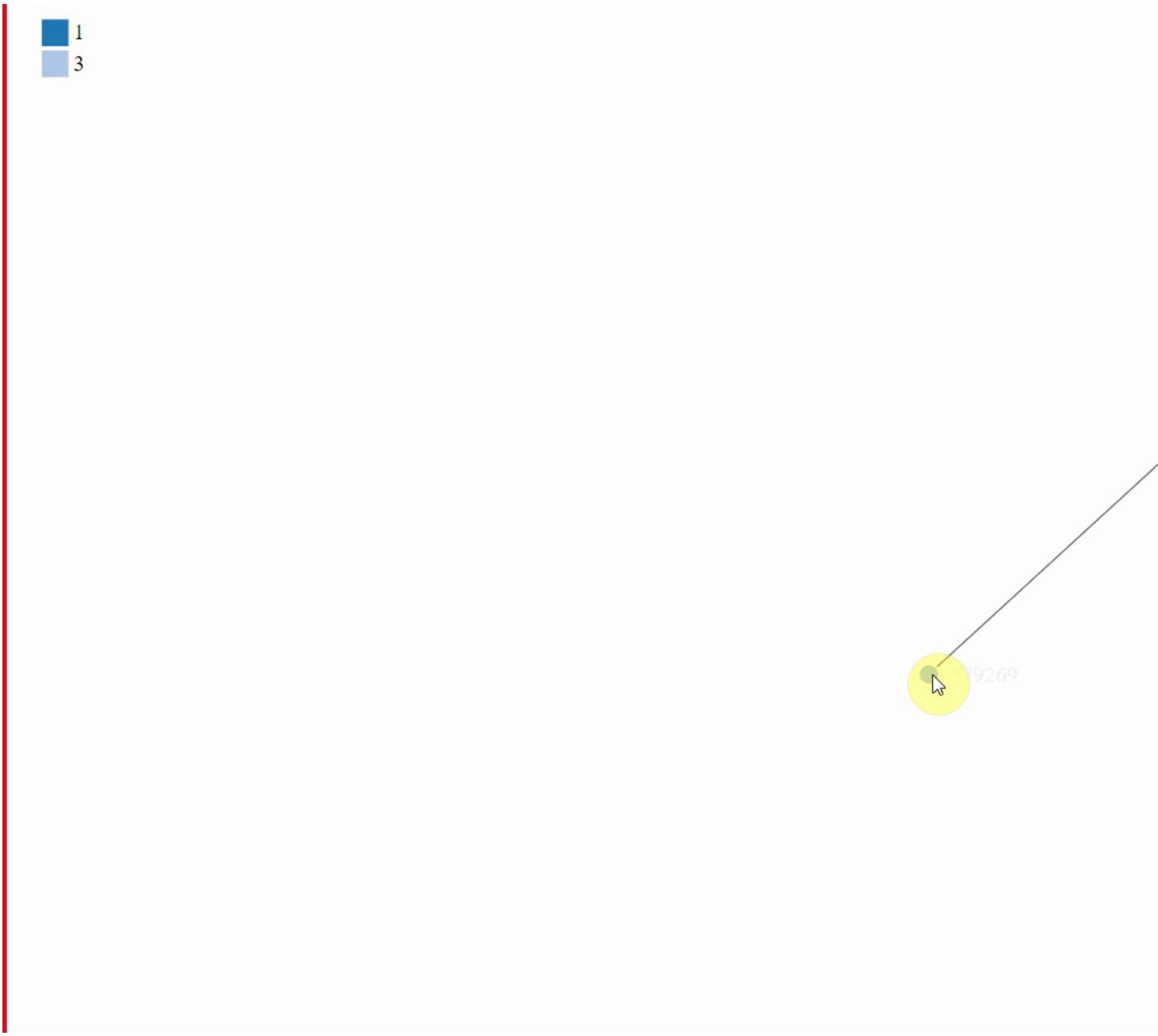

*Figure 14. Animated creation of a network graph based on **seed 2**. Only two nodes are connected to channels other than co-authorship.*

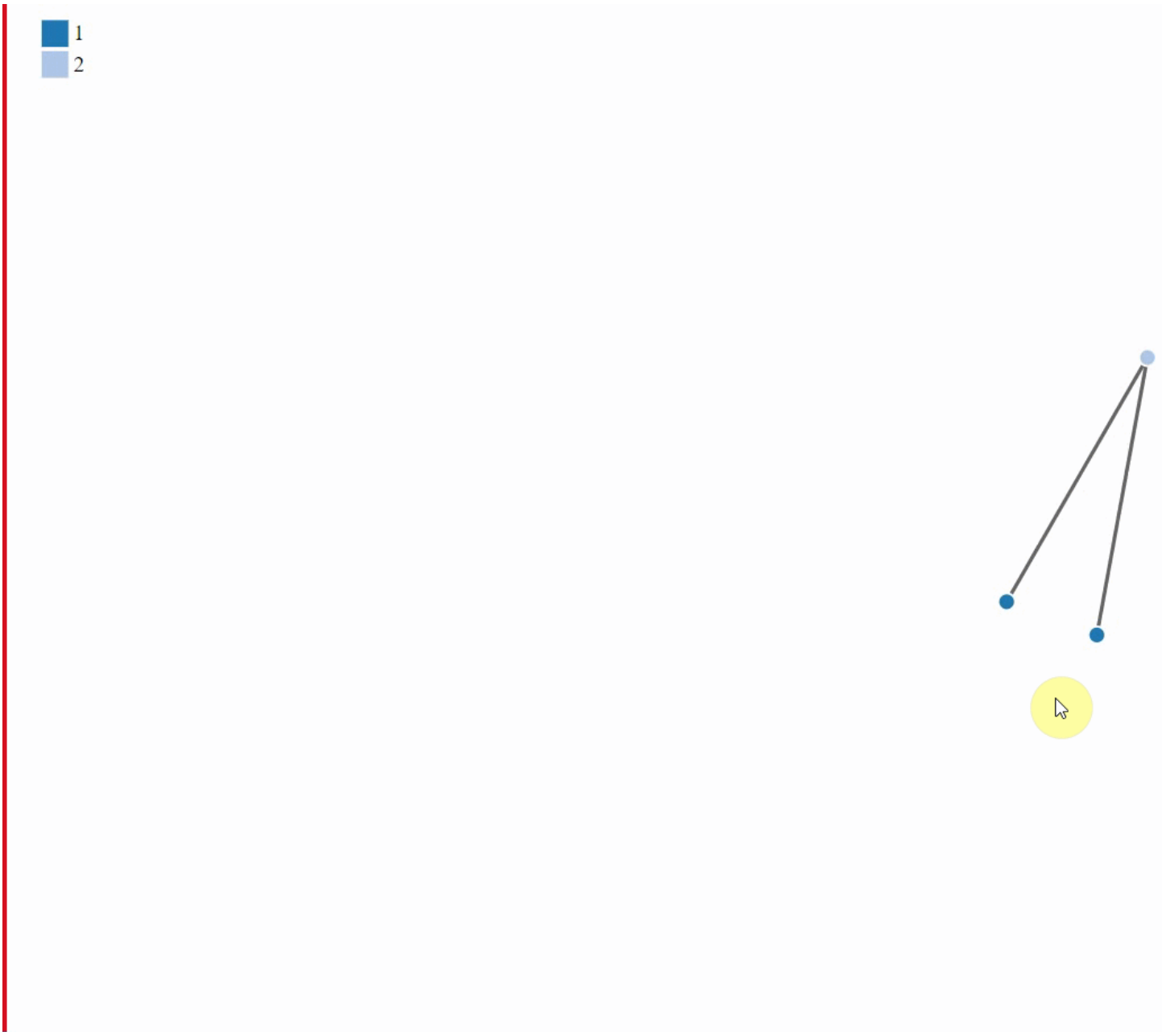

*Figure 15. Animated creation of a network graph based on **seed-3**.*

The retrieved subgraphs for seed 1 (Q2-Graph1) and seed 3 (Q2-Graph3) have been compared to the template subgraph and the best result from the first question (subgraph 2) based on each channel.

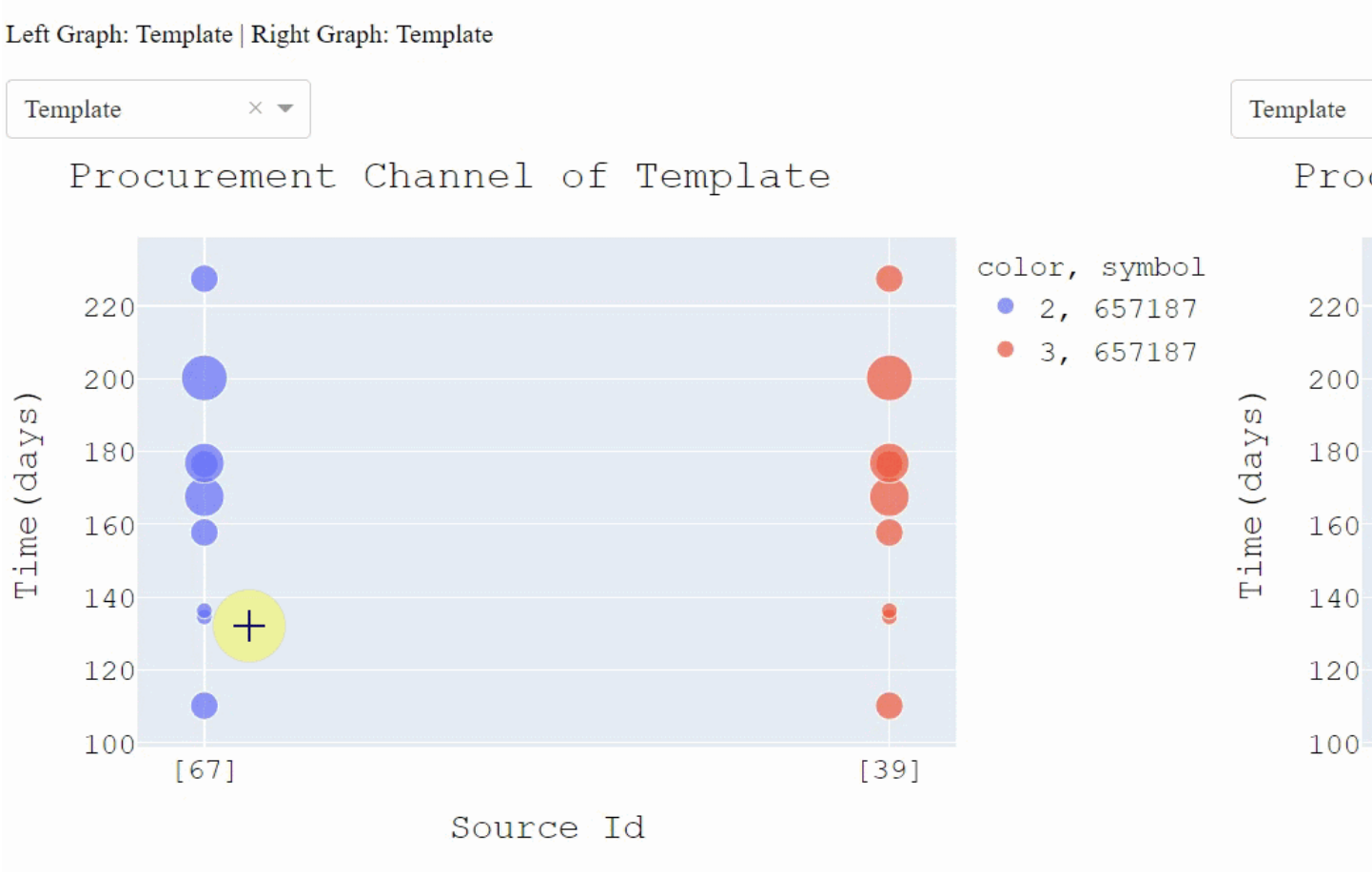

Figure 16. ***Procurement Channel****: seed 1 and seed 3 subgraphs have a maximum of two transactions for the same item which is lower than the template, but there are transactions nonetheless.*

*Figure 17. **Travel Channel**: Similar to template and subgraph 2, seed 1 and seed 3 subgraphs include nodes which have travelled from one "SourceLocation" (depicted by color) to the same "TargetLocation" at close points in time.*

Figure 18. ***Demographic Channel****: seed 1 and seed 3 subgraphs show similar income and expense patterns as the template and subgraph 2.*

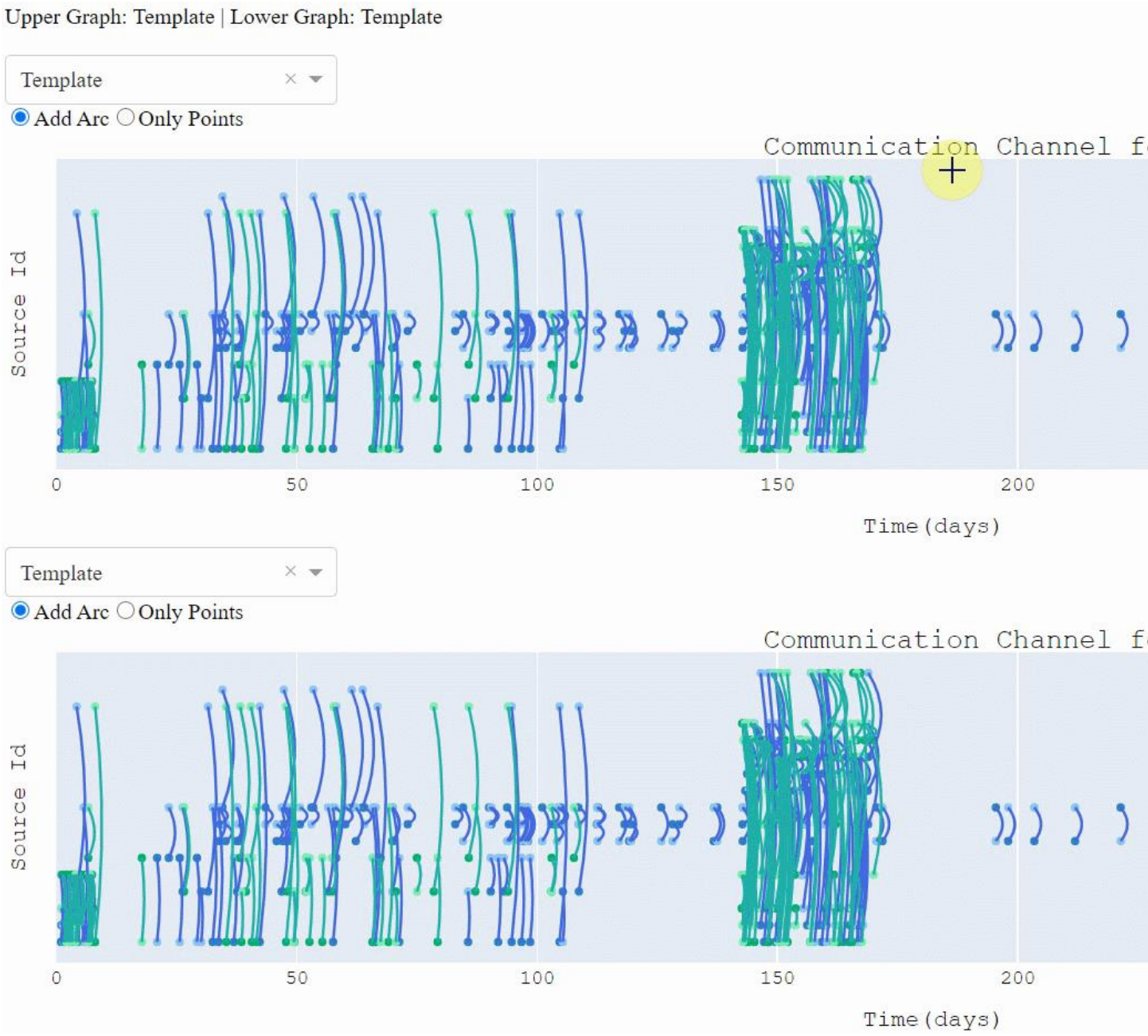

*Figure 19. **Communication Channel**: seed 1 and seed 3 subgraphs frequencies of communication between sources are compared to template.*

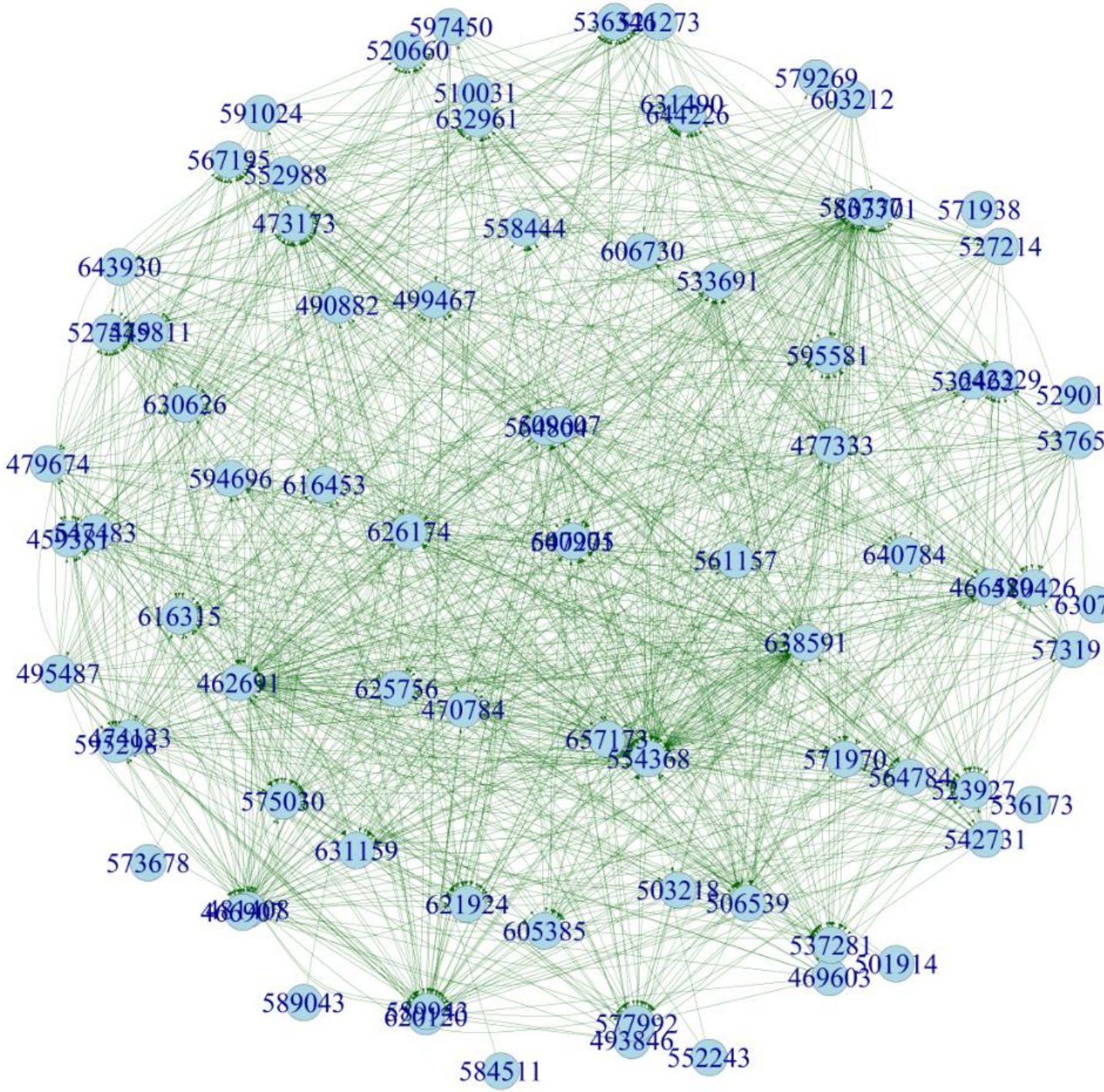

Figure 20. Network graph displaying person's ID of Q2-Graph1.

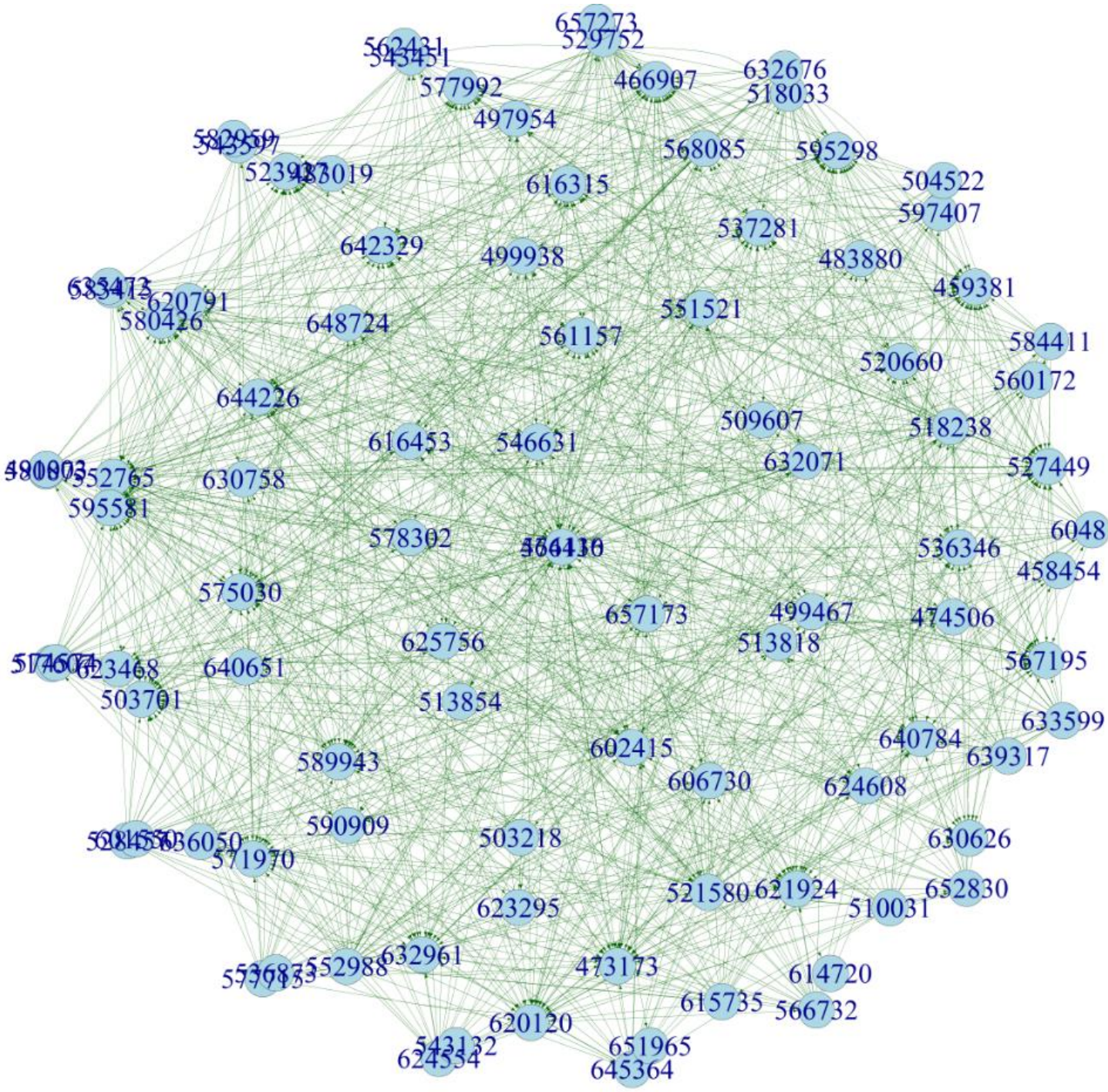

Figure 21. Network graph displaying person's ID of Q2-Graph3.

All in all, seed 1 (Q2-Graph1 in Figure 20) and seed 3 (Q2-Graph3 in Figure 21) seem to have a lot of similarities to the template based on the channels and therefore they should be considered potential hacker groups.

3 – Optional: Take a look at the very large graph. Can you find other subgraphs that match the template subgraph provided? Describe your process and your findings in no more than ten images and 500 words.

Answer:

In order to find other subgraphs from the large graph that resemble the template subgraph, we use the patterns from each channel and try to extract the sources which show similar patterns to the template subgraph for each channel. The process can be described by four stages:

Stage 1: A list of sellers and buyers who made transactions equal to or more than seven times (same as the template subgraph) was retrieved from the large graph. A set of frequent items list was created for items which appeared in both the sellers and buyers' list. Each of the transactions for these items were then analyzed and there emerged three sets of source and targets, that showed similar trends as the template subgraph, which is, there were frequent transactions between these three sets of buyers and sellers for the same target item.

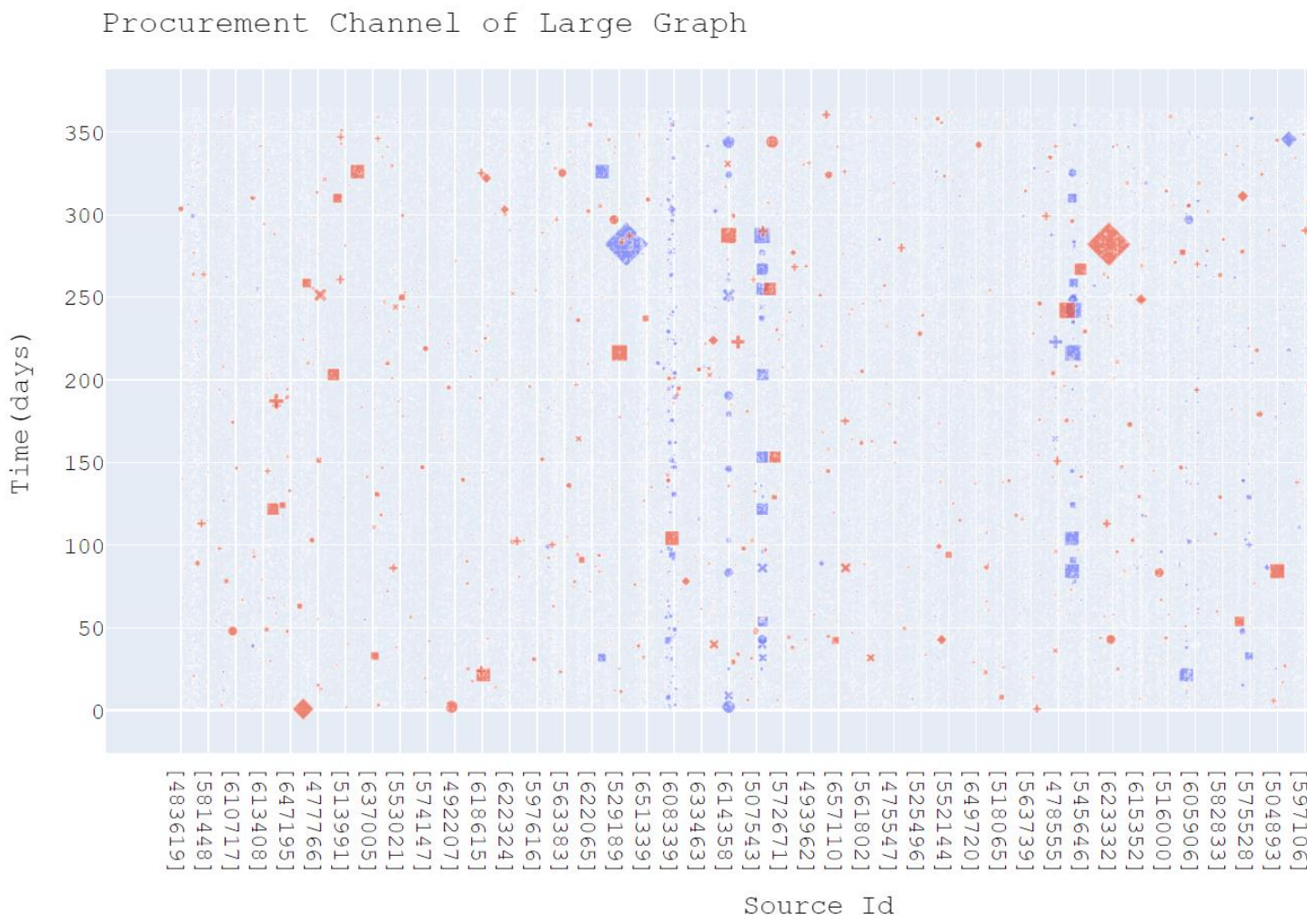

Figure 22. Procurement channel of large graph was visualized and analyzed to find patterns. Color: "eType", symbol: items

Stage 2: Tableau Desktop (version 2019.4.7, Tableau Software Inc., Mountain View, CA, USA) was used to extract the nodes which have travelled together from the same source location to the same target location at the same time from the large graph. For each target location all the sources belonging to these clusters were kept in a list, List 1.

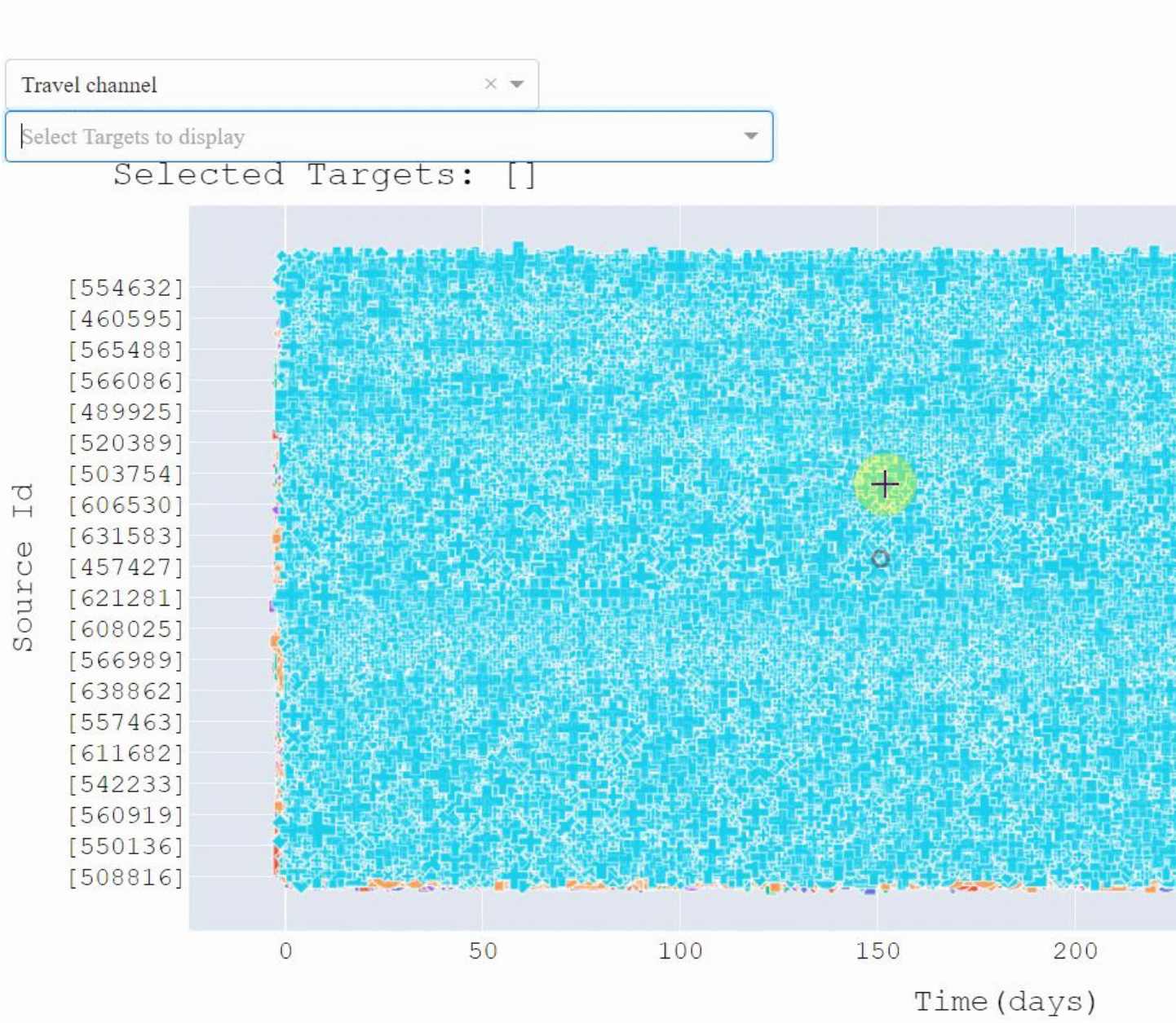

Figure 23. The Travel Channel for the seeds from Stage 1 analysis was extracted and the Sources were put into a list, List 2.

List 1 and List 2 (see Figure 23) were then compared to find common nodes which were present in various clusters from the Travel Channel of the large graph and also belong to the three graphs built from the level 1 seeds. The common nodes were the ones which were finally kept to build travel channel.

Stage 3: The final results were two extended Networks with Communication, Demographic and the Co-Authorship channel data added for all of these nodes.

Stage 4: The extended graphs contained sources which had contacted others only once or twice. These sources were removed and based on the distribution of frequencies of calls and emails in the template, size of the graphs was reduced till we had number of nodes in the same range as the template.

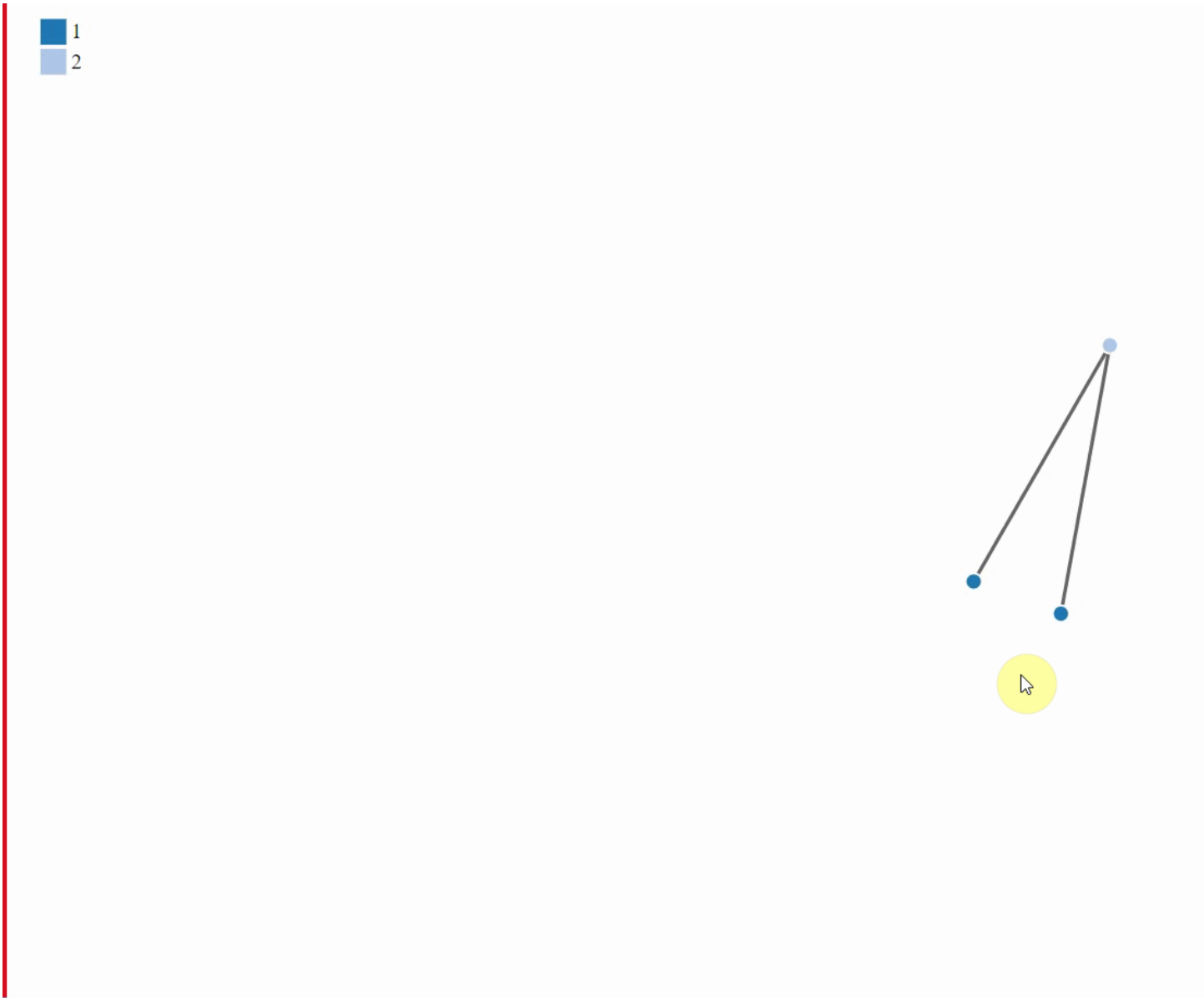

Figure 24. Animated creation of Q3-Graph1.

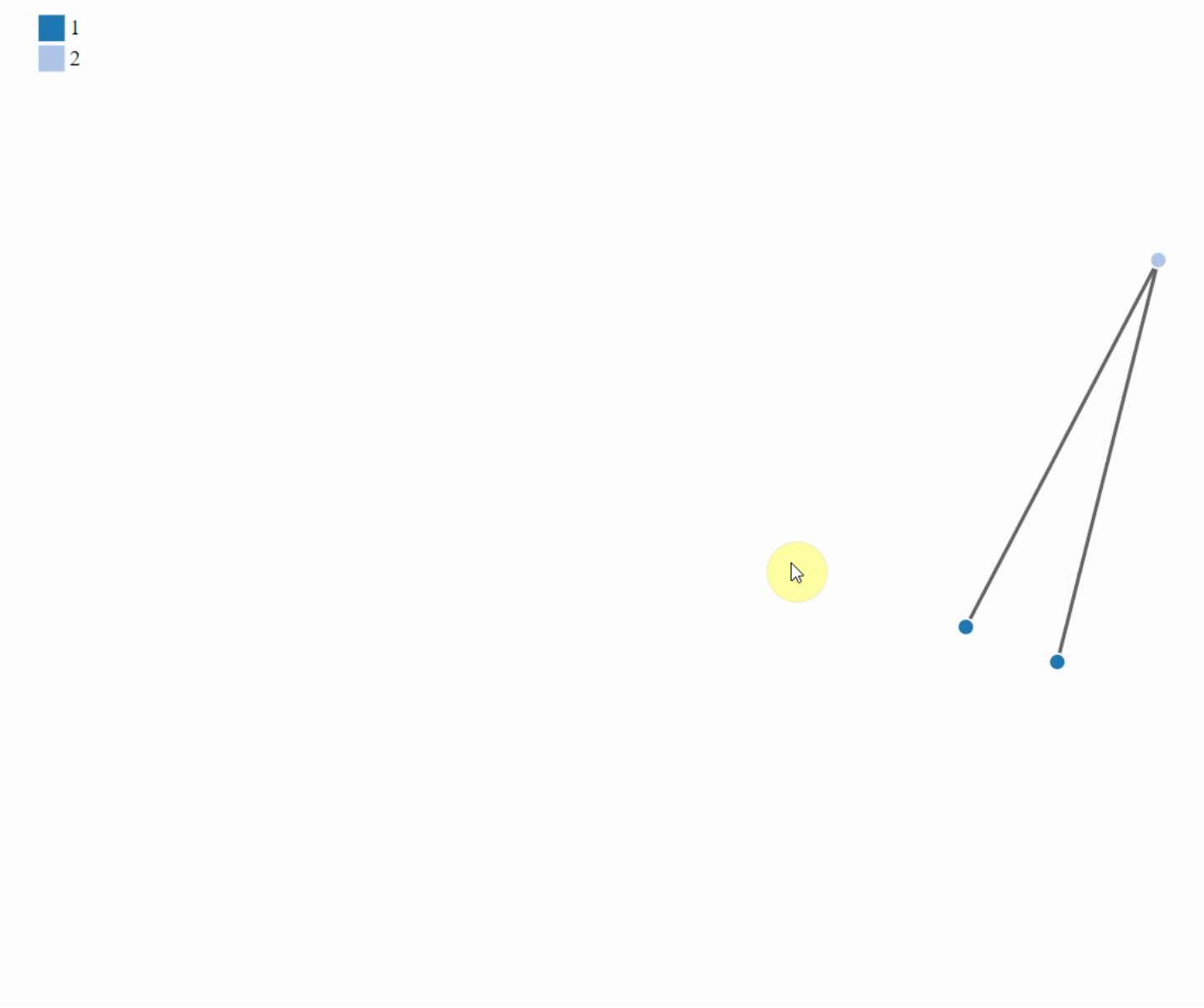

Figure 25. Animated creation of Q3-Graph2.

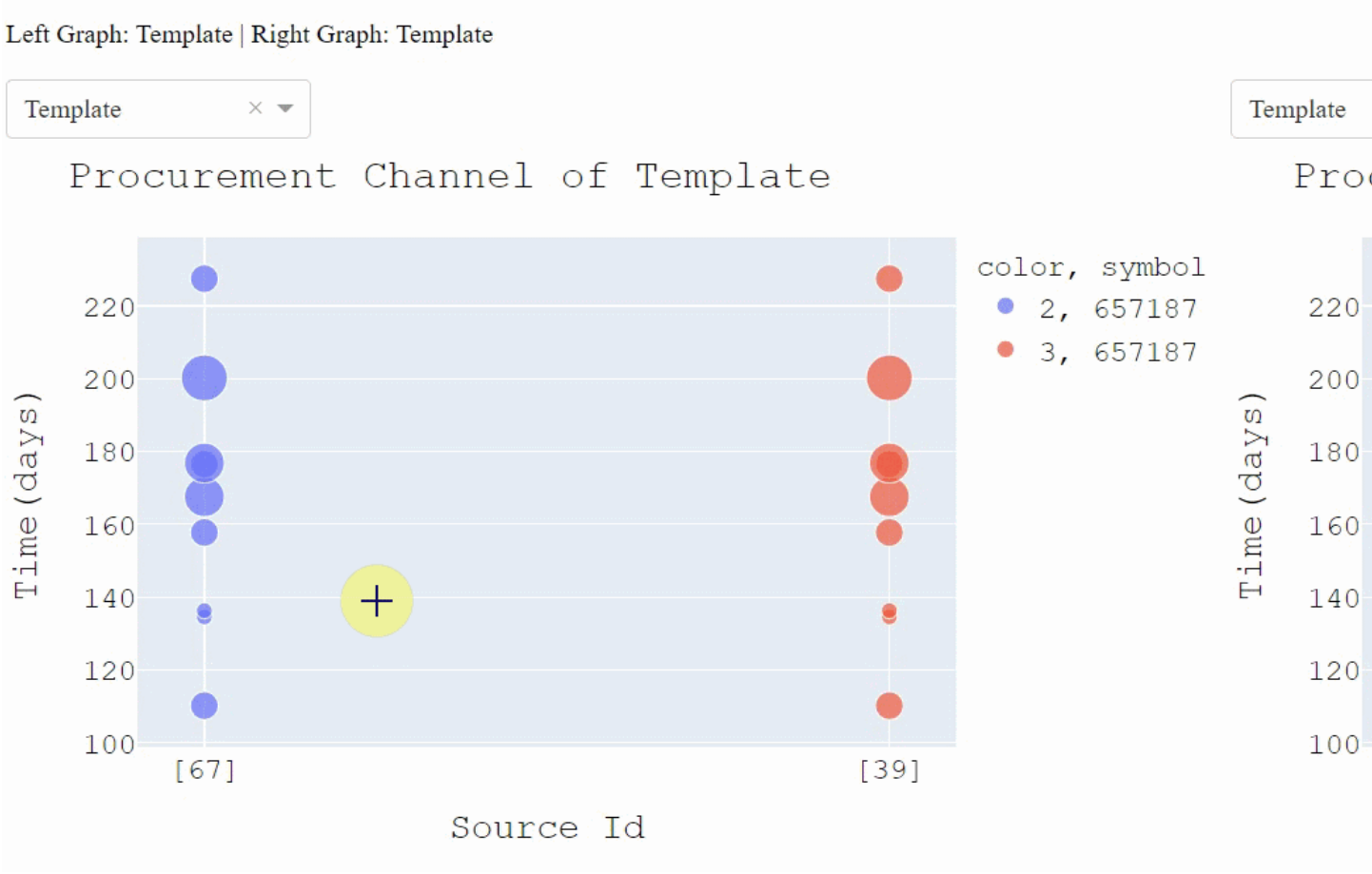

Figure 26. ***Procurement channel*** *analysis shows transactions between two Sources for one Target.*

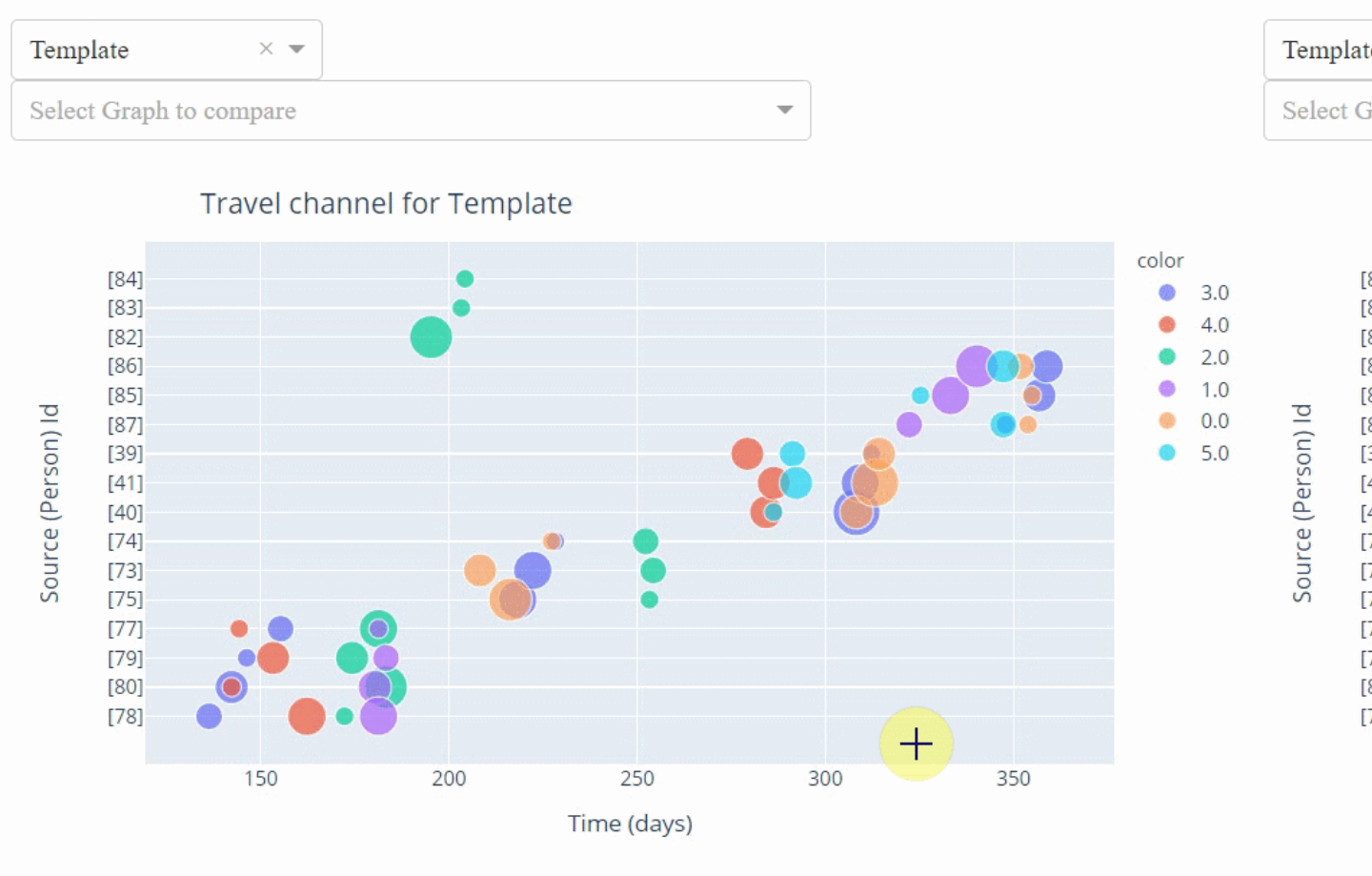

*Figure 27. **Travel channel** analysis shows clusters forming for different Target Locations.*

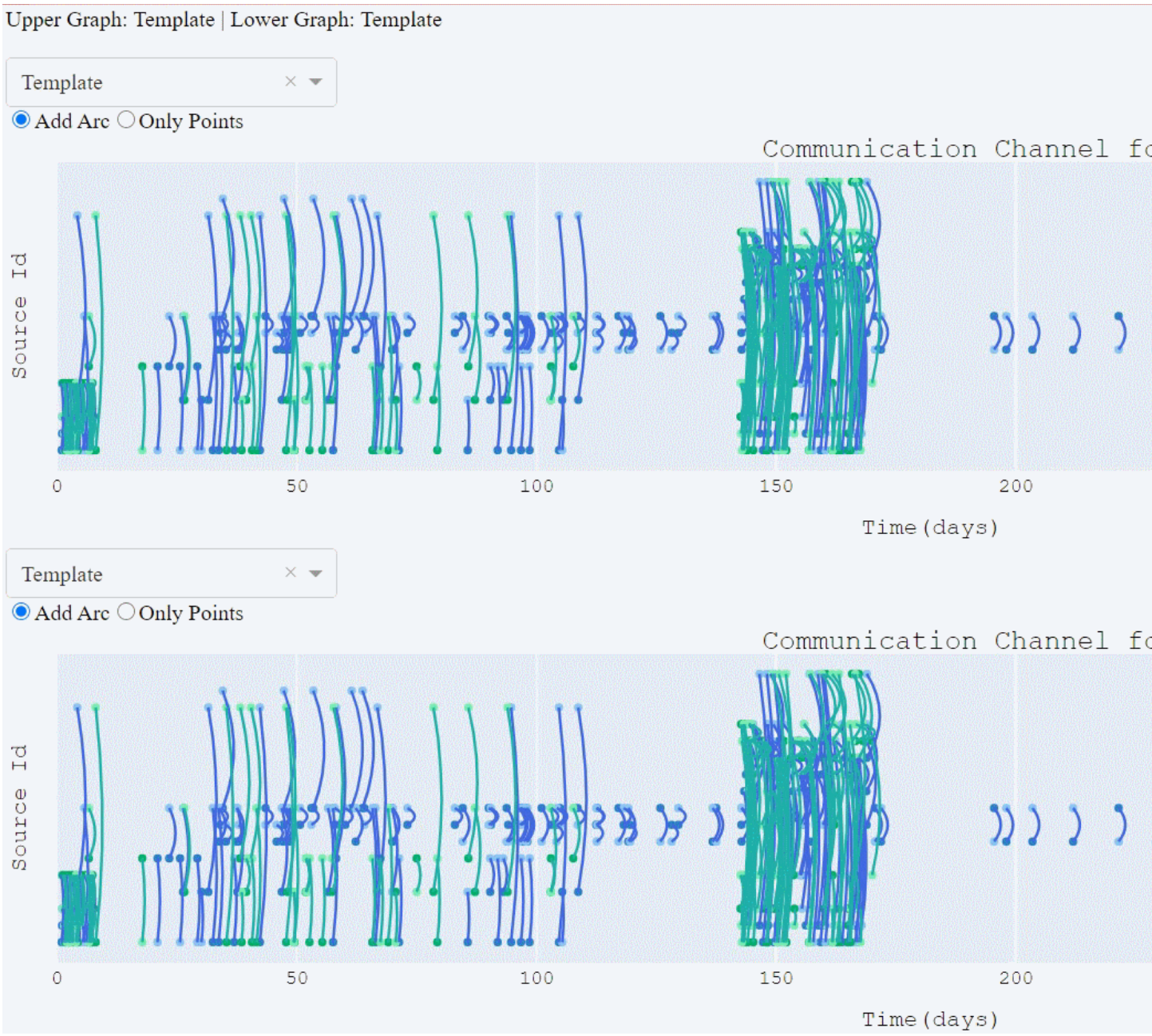

Figure 28. ***Communication channel*** *analysis for the reduced network showed the frequencies of communication similar to the template*

*Figure 29. **Demographic channel** analysis for the reduced network showed the similar trends as template.*

Hence, we have Q3-Graph1 and Q3-Graph2 from this analysis which resemble the template subgraph.

4 — Based on your answers to the question above, identify the group of people that you think is responsible for the outage. What is your rationale? Please limit your response to 5 images and 300 words.

Answer:

To finally conclude which group had caused the outage, a comparison is made between subgraph 2 from question 1 (Q1-Graph2), seed 1 graph (Q2-Graph1), seed 3 graph (Q2-Graph3) from question 2, the two graphs retrieved from question 3 (Q3-Graph1 and Q3-Graph2) and the template subgraph. The same procedure as question 1 part a is repeated.

Figure 30. PC to find trends from the candidate subgraphs which resemble the template. Subgraph 2 from question 1(Q1-Graph2) and seed 1 graph (Q2-Graph1) resemble the template highest.

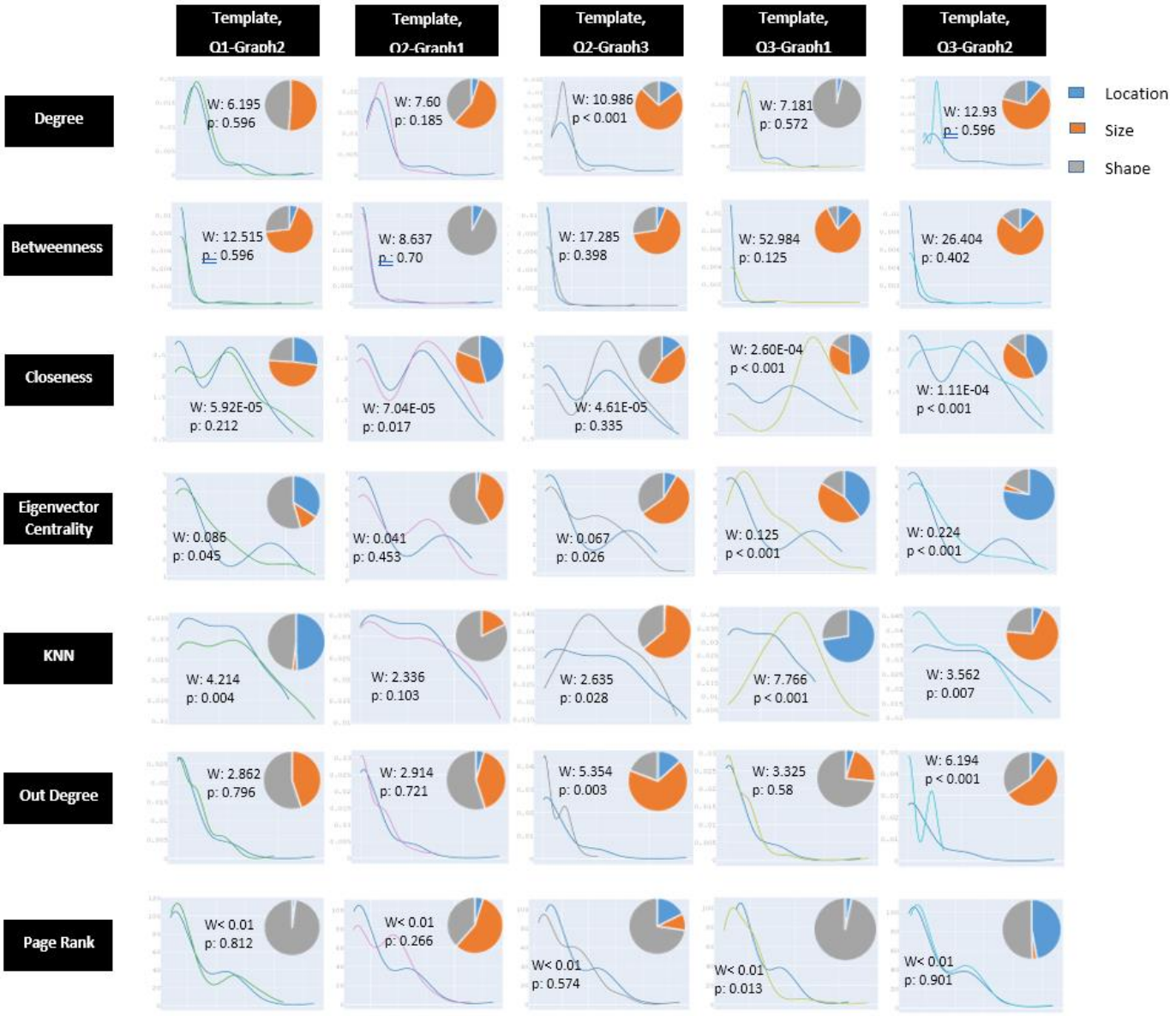

Figure 31. Matrix comparison for distributions of seven similarity measures with Wasserstein metric (W), p-value (p) and relative contribution of three factors (source, shape and location) towards the difference among the candidate subgraphs and template.

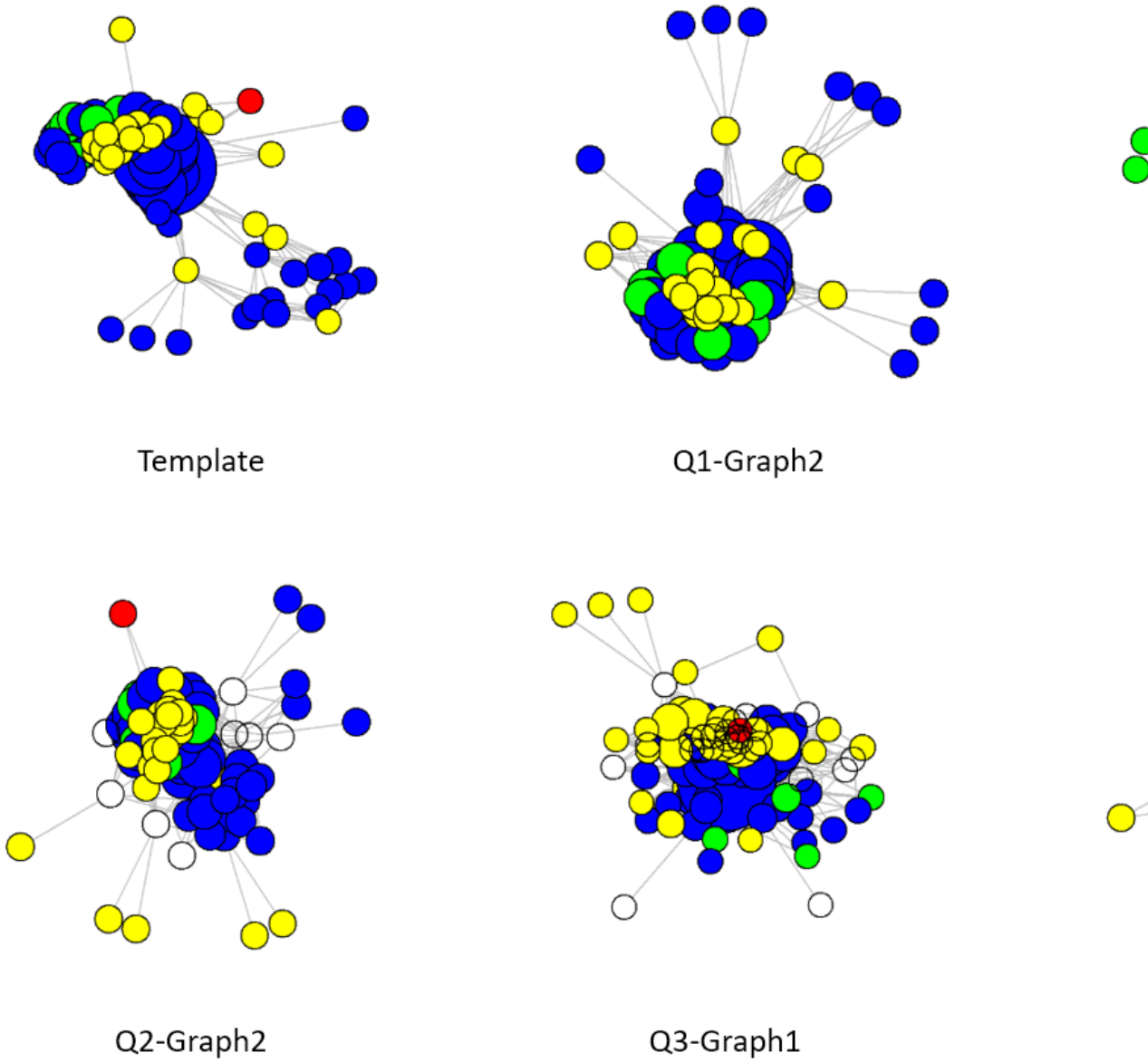

Figure 32. ***Out-degree:*** *Network graphs with node sizes proportional to the magnitude of the similarity measure. Subgraph 2 from question 1 (Q1-Graph2), seed 1 graph (Q2-Graph1) and seed 3 graph (Q2-Graph3) are the most similar to the template.*

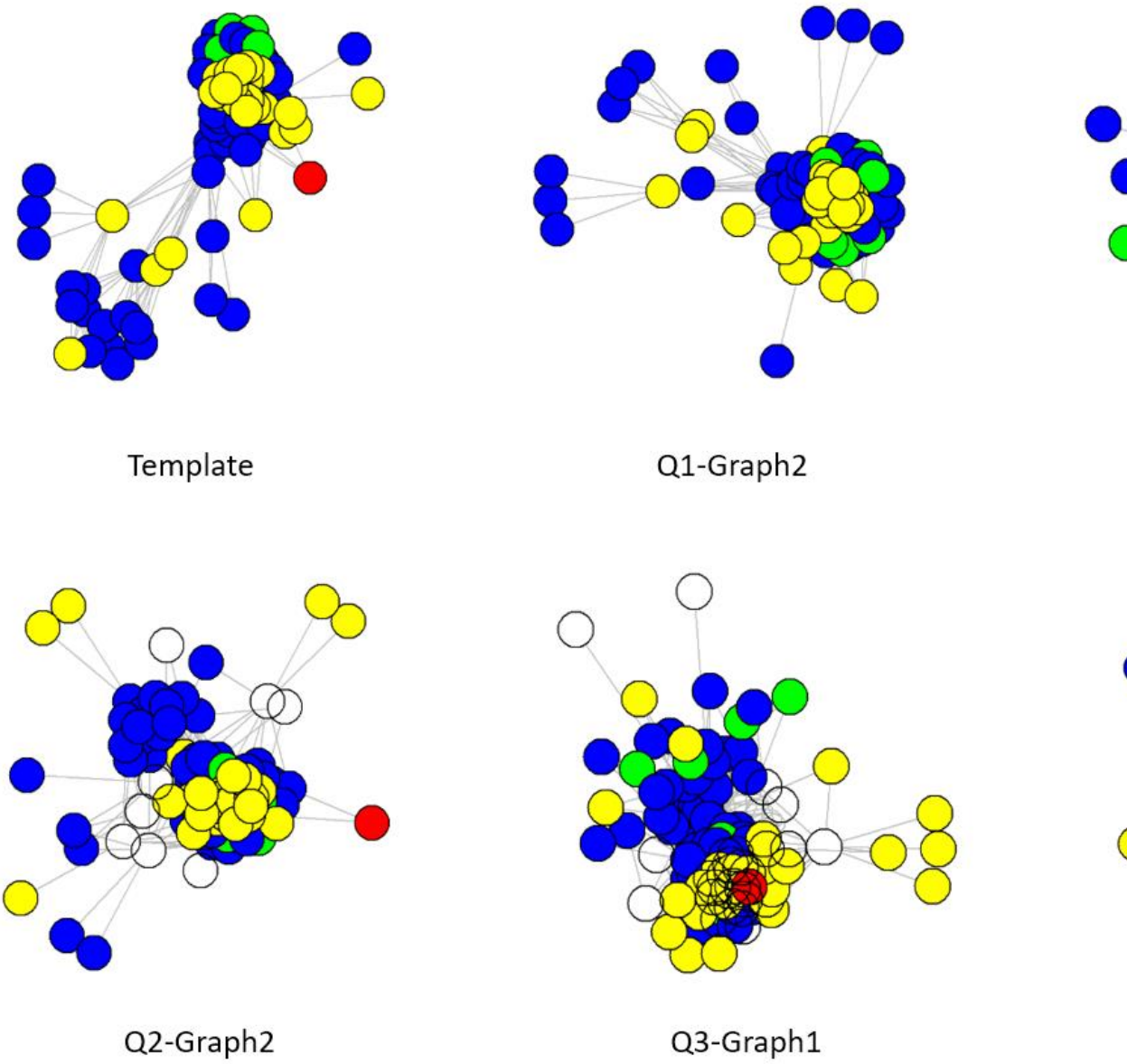

Figure 33. ***Eigenvector****: Network graphs with node sizes proportional to the magnitude of the similarity measure. Subgraph 2 from question 1 (Q1-Graph2), seed 1 graph (Q2-Graph1) and seed 3 graph (Q2-Graph3) are the most similar to the template.*

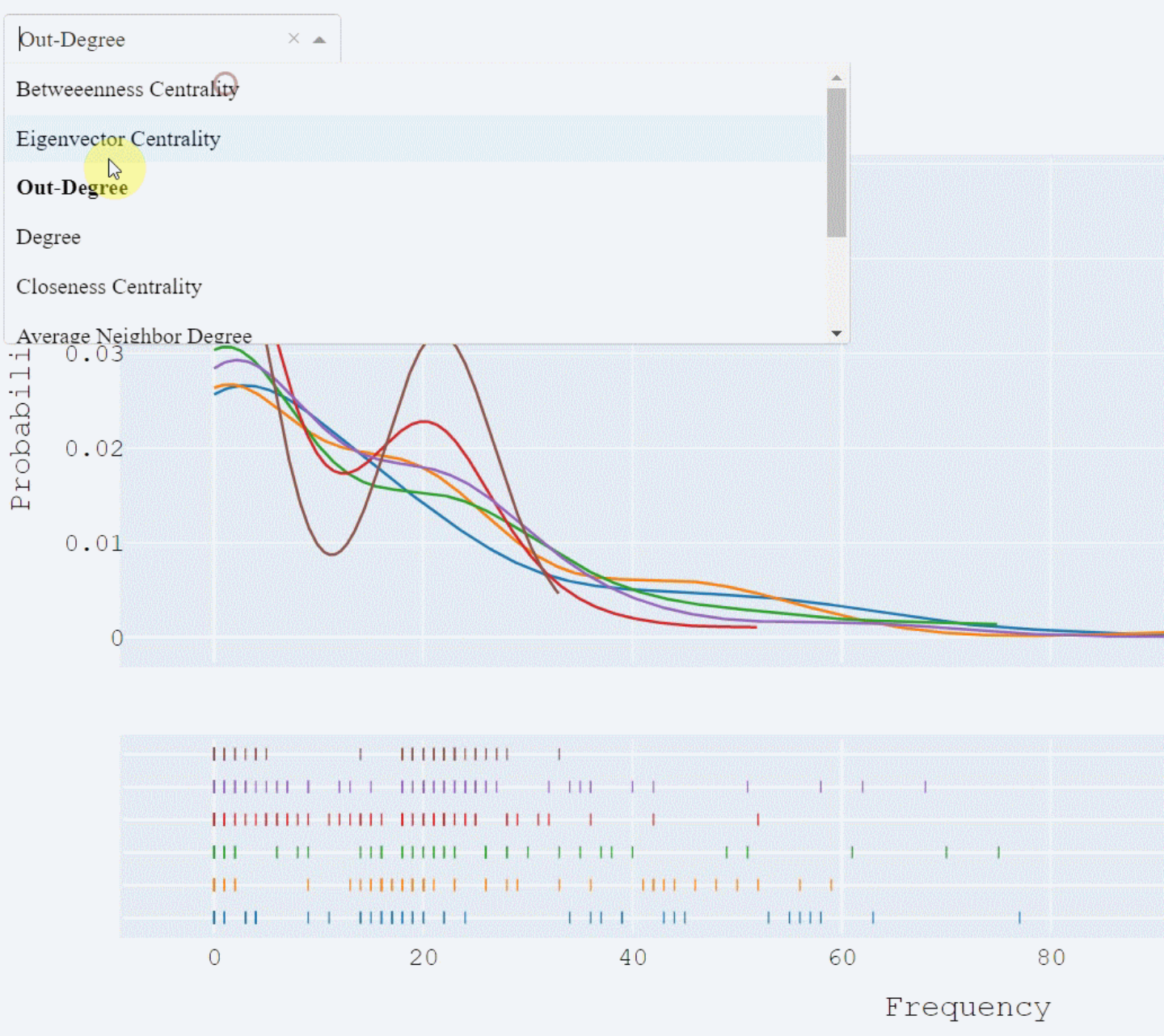

Figure 34. Visualization of the estimated probability density and individual values as "rugs" for out-degree, betweenness centrality and eigenvector centrality.

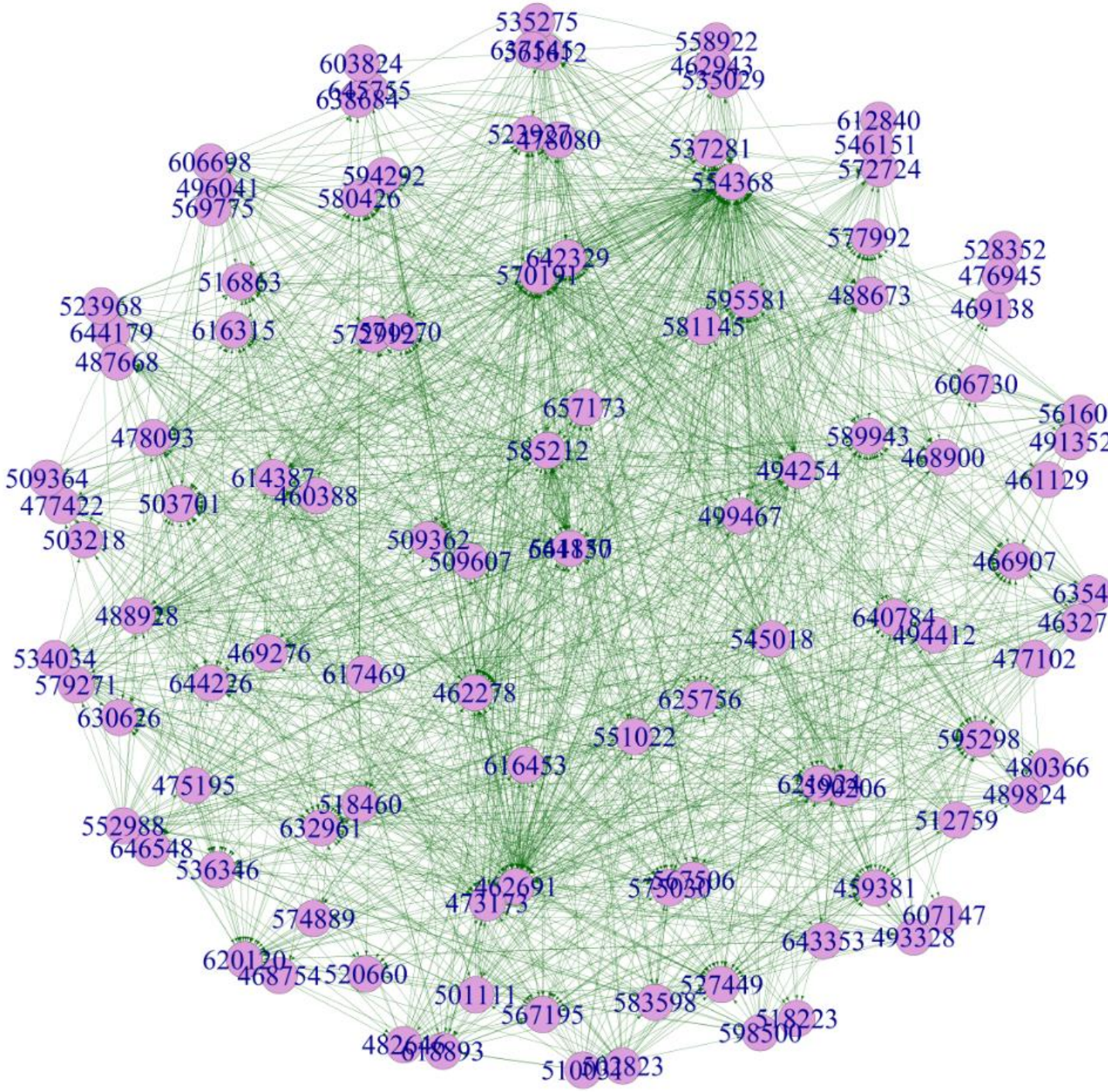

Figure 35. Network graph displaying person's ID of Q3-Graph1.

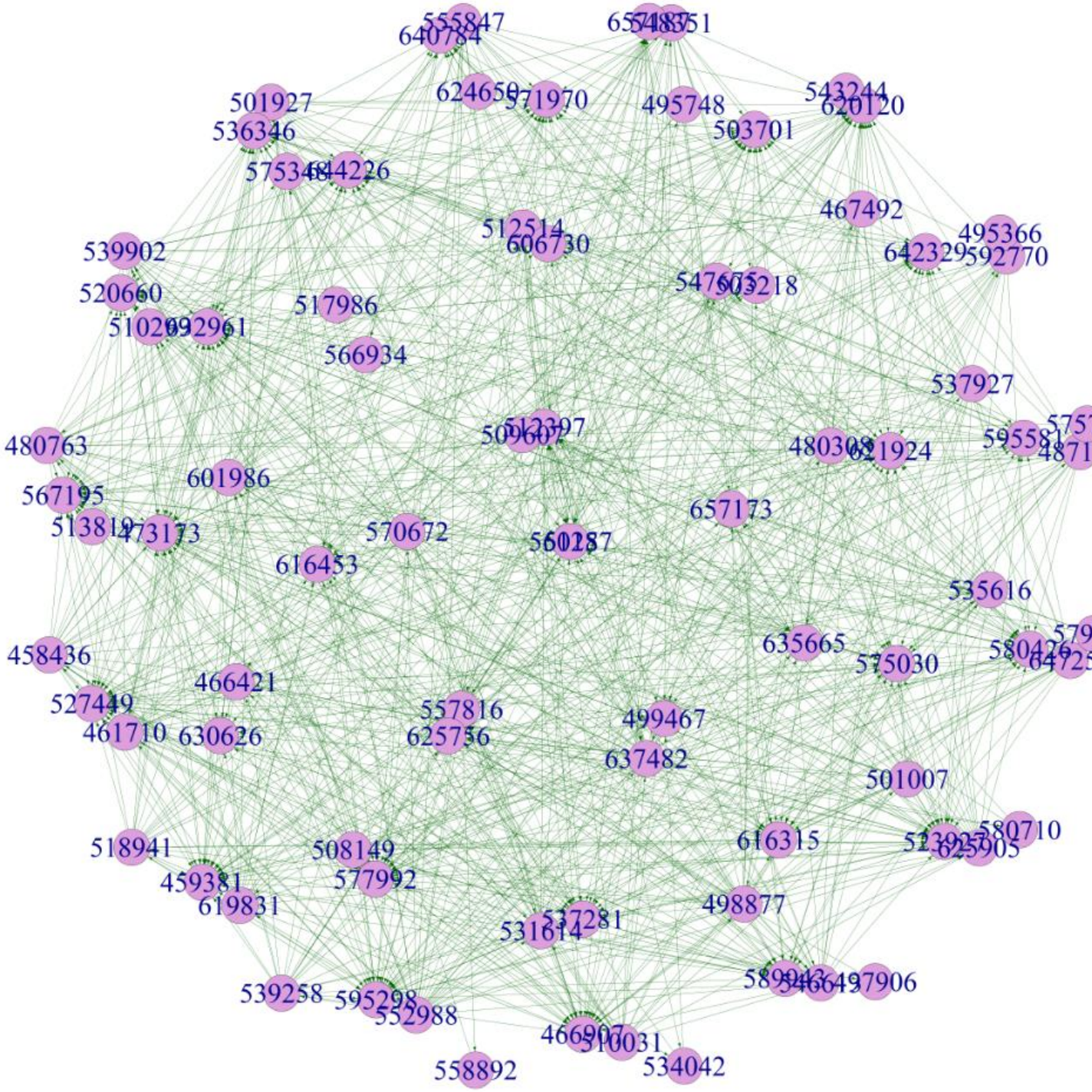

Figure 36. Network graph displaying person's ID of Q3-Graph1.

Conclusion: Q1-Graph2 and the Q2-Graph1 show very high resemblance to the template. However, analyzing the density curves more closely, it is very evident that Q2-Graph1 resembles the template more than Q1-Graph2 in terms of eigenvector centrality and betweenness. As these measures represent how important/central nodes in a network are connected and how information flows within the network, ***the probability of Q2-Graph1, from seed 1, being the group of people responsible for the outage is maximum***.

Note: This group can be visualized in Figure_20.

5 — What was the greatest challenge you had when working with the large graph data? How did you overcome that difficulty? What could make it easier to work with this kind of data?

The difficulties faced with the large graph data were as follows:

- **Loading the entire dataset**: Not many libraries are recognized which are able to load this huge amount of data efficiently.
- **Memory**: The dataset was too large to store in the RAM and needed to be saved in the hard-drive which costed computational time.
- **Computational Complexity**: The dataset was too large and complex and required a large amount of computational time.
- **Complexity in Interpretation**: The dataset was too large and complex to visualize or interpret in one go and needed to be divided into parts for better interpretability. Also, defining the whole dataset as a graph "dataframe" and finding subgraphs or any other graph related characteristics was computationally impossible.

Solutions:

- R libraries like "fread" and "ff package" were able to load the large dataset quickly.

- Using online tools like “Colab” helped optimize the use of RAM and gave us more computational power.
- Dividing the dataset into Channels and working on the patterns, helped segment the data for a better interpretation.

What could make it easier to work with this kind of data?

- Usage of larger RAMs.
- Libraries or tools which can do the computations in the GPU.