The CGCS has collected data from the community of cyber researchers over the last year. There are 123,892,863 records. They have compiled the data into a graph where each person is identified by an anonymous ID number. There are 6 different channels of data, all of which are represented as a transaction between two nodes. The graph data is presented in a compact edge list format. In the table below, we have provided some basic graph properties:

Channel	Channel Description	Node Types	eType	Directed (Y/N)	Bipartite (Y/N)	Temporal (Y/N)
1	Phone	1	0	Y	N	Y
2	Email	1	1	Y	N	Y
3	Procurement	1,2	2,3	Y	Y	Y
4	Co-authorship	1,3	4	Y	Y	Y
5	Demographics	1,4	5	Y	Y	N
6	Travel	1,5	6	Y	Y	Y

There are 5 different node types in each data set and 7 different edge types. Node type 1, which represents persons, serves as a unifying type throughout the data set. In the multi-type graphs, the edge directionality reinforces the node type assignment (edges always go from node type 1 to some other node type). Type 1 nodes are the only nodes with a spatial location assigned.

Node types are as follows:

1. Person (used in all channels)
2. Product category (for the procurement channel, eType = 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)

A file has been provided for you to easily identify the node type of any unique identifier in the data. See CGCS-GraphData-NodeTypes.csv for node types in the large graph and subgraphs that have been extracted from it. See CGCS-Template-NodeTypes.csv for node types in the template.

Edge types are as follows:

0. Email
1. Phone
2. Sell (procurement)
3. Buy (procurement)
4. Author-of
5. Financial (income or expenditure, depending on direction)

6. Travels-to

All graph files contain the following columns:

- Source
- eType (edge type)
- Target
- Time

Time is in seconds from 12:00 AM Jan. 1, 2025 (like Unix timestamps, except that time=0 is the year 2025, not 1970).
Many of the channels also include:

- Weight
- SourceLocation
- TargetLocation
- SourceLatitude
- SourceLongitude
- TargetLatitude
- TargetLongitude

The SourceLocation and TargetLocation are integer values between 0 and 5, representing countries. The Latitude and Longitude columns are locations within the country associated with the person. There may be a significant amount of noise associated with this location. (Locations are purposefully placed in remote places on Earth and do not correspond to real countries).

Communications channels (eType 0 and 1)
Sample:

Source	eType	Target	Time	Weight	SourceLocation	TargetLocation	SourceLatitude	SourceLongitude	TargetLatitude	TargetLongitude
599956	1	635665	1423376	1	0	0	34.2958	-39.026	29.3296	-37.8076
599956	0	589639	1437564	1						

Communications channels represent phone and email transactions between two people. Some records have location information, and some do not. The weight for communications is always 1, representing 1 call or email.

Procurement channels (eType 2 and 3)
Sample:

Source	eType	Target	Time	Weight
550287	2	657187	10250285	170

512397	3	657187	10250285	170

The procurement channels represent sale (eType 2) and purchase (eType 3) of an item, which is always listed in the Target column. Whenever there is a sale of an item in channel 2 there will be a corresponding entry in channel 3 at the same time. The two people can be linked via the item they are both connected to. There may also be communications between the two people. The weight for procurements represents the value of the item. Procurements do not have location information.

Co-authorship channel (eType 4)
Sample:

Source	eType	Target	Time	Weight
616050	4	590502	-662041253	0.166667

The co-authorship channel represents publication of scientific or technical articles. The source column contains the person who is the author, and the target column contains a unique identifier for the publication. Multiple authors may be connected to the same publication. The time for publications occurs before all other records in the data so the values are negative. Times are still relative to Jan. 1, 2025. The weight column indicates the fraction of the authors for the given publication. For example, the person listed in the table above is one of 6 authors, so the weight is 1/6 = 0.166667. Authorship does not have location information.

Demographics channel (eType 5)
Sample:

Source	eType	Target	Time	Weight
608827	5	630626	31536000	21699.3
552988	5	608827	31536000	143858
608827	5	567195	31536000	15088.4
608827	5	527449	31536000	456.71
608827	5	459381	31536000	2378.6

The demographics channel is a graph representation of the spending characteristics of each person in up to 30 categories. This channel may also be thought of as attributes of each person. The person is listed in the source column when money is spent in a category and listed in the target column when money is received in that category (such as income or gifts). The categories are listed in the file DemographicCategories.csv. A person may not be connected to each category, such as the case where a homeowner does not have rent expenses. The time for all records in this channel is 31536000, which is the 365th day, and comes after all other record types. The weight channel shows how much is spent (or received) in a given category. Be careful! This channel creates many edges that do not represent person-to-person connections in the same way as the other channels. Demographic records do not have location information.

Travel channel (eType 6)
Sample:

Source	eType	Target	Time	Weight	SourceLocation	TargetLocation	SourceLatitude	SourceLongitude	TargetLatitude	TargetLongitude
649553	6	509607	26595547	4	5	4	22	156	1	-165
570284	6	509607	26681947	1	5	4	22	156	1	-165

The travel channel connects people (always listed in the source column) with locations (listed in the target column). Time represents the start of a trip and weight represents the length of the trip in days. All location columns should have data for each record. The SourceLocation and TargetLocation columns have identifiers for countries of the origin and destination of each trip. More specific latitude and longitude values are also provided. However, these values may not be perfectly accurate.

The template:
A template file (CGCS-Template.csv) is provided in the same edge list graph format as the large graph data. The template is a profile of activities that CGCS has built to represent suspicious activity associated with the hack. CGCS researchers hope that the group responsible will match, or partially match, this graph pattern. No specific location information (latitude and longitude) is provided in the template. However, location information for country may be present. The co-authorship channel in the template includes values replaced by -99. This indicates that any publication would satisfy a match at any time and with any given weight. The node ID value of -99 is listed as a publication type in the CGCS-Template-NodeTypes.csv file. It is not a publication when listed in the time or weight columns. The earliest time in the template is listed as 86400, which is the end of the first day. Relative times between transactions may be important. You may expect to not find an exact match to the template in the graph data.

Candidate Subgraphs:
Five subgraphs are provided for comparison to the template. They are:

- Q1-Graph1.csv
- Q1-Graph2.csv
- Q1-Graph3.csv
- Q1-Graph4.csv
- Q1-Graph5.csv

These graphs are portions of the large graph that have been extracted.

Seed Graphs:

Three seed graphs are supplied as starting points for your search in question 2. The seed files are:

- Q2-Seed1.csv
- Q2-Seed2.csv
- Q2-Seed3.csv

Each seed graph has a single record from which you will attempt to build a graph that matches the template.

The BIG graph:
All records collected by CGCS are contained in a single file (CGCS-GraphData.csv). There are 123,892,863 records in this file. The uncompressed size is 6.2 GB. The data has been compressed and partitioned into a multi-part zip archive. This archive can be unzipped using free utilities such as WinZip or 7zip. Please contact the VAST challenge committee (vast_challenge@ieeevis.org) if you have difficulty with this file or unzipping.