VAST Challenge 2020
Mini-Challenge 1

Miembros del equipo:
Subotovsky N., Lewinger A., Viegener A., Knebel J., Pecina L., Véliz F., Dell'Era D., Escudero S., Copa V., Iguaran J.J., Burastero O.

Los investigadores de CGCS (Centro Global Cyber Strategy) utilizaron los datos donados por los grupos de hackers "sombrero blanco" que se usaron para crear perfiles anónimos de ciertos grupos (personas?). Uno de estos grupos ha sido identificado por los psicólogos sociales de CGCS como los que tienen más probabilidades de parecerse a la estructura del grupo que causó accidentalmente este corte de internet. Se le ha pedido que examine los registros de CGCS e identifique aquellos grupos que se parecen más al perfil identificado.

Preguntas:

1. **Utilizando el análisis visual, compare el subgráfico de la plantilla con los posibles candidatos proporcionados. Muestre dónde están de acuerdo y en desacuerdo los dos gráficos. Use su herramienta para responder las siguientes preguntas:**
    a. **Compare los cinco subgrafos candidatos con la plantilla provista. Muestre dónde están de acuerdo y en desacuerdo los dos gráficos. ¿Qué subgrafo coincide mejor con la plantilla?**

Para resolver esta pregunta, se recurrió a una visualización utilizando grafos de fuerza en donde se pueden visualizar todos los nodos y sus respectivas conexiones dejando de lado las relaciones demográficas:

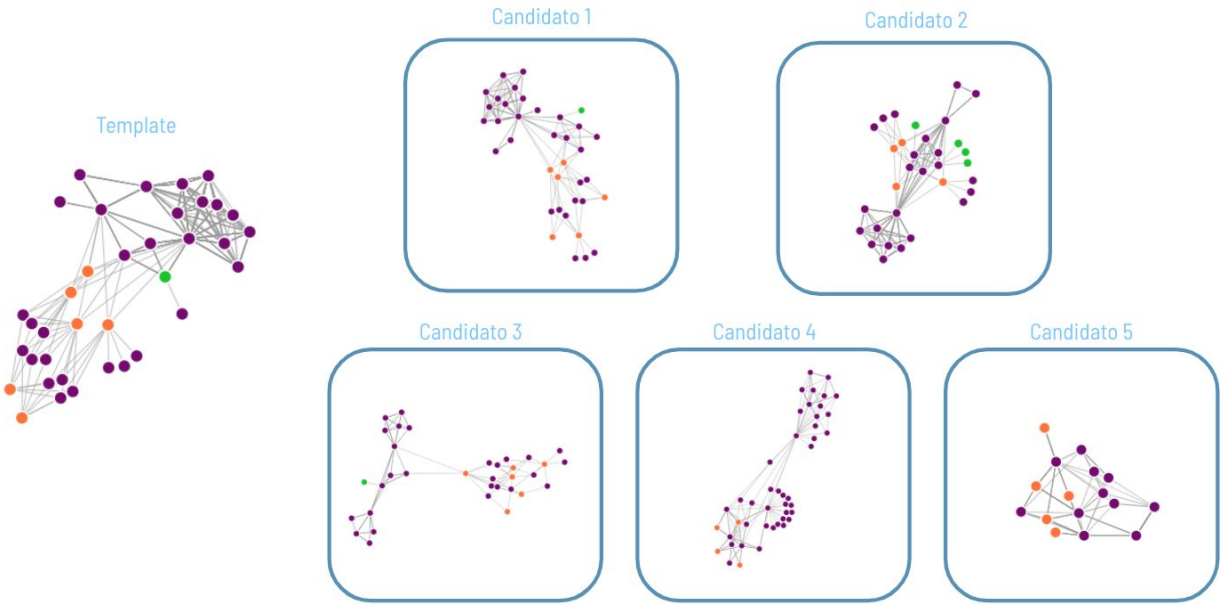

Figura 1 Grafo de fuerza

En los grafos anteriores los nodos morados representan personas, los nodos naranja países de color y los verdes documentos. Analizando el grafo template se pueden apreciar los siguientes patrones: la existencia de 2 grupos bien diferentes: aquellas personas densamente interconectadas por comunicaciones (e-mails/llamadas telefónicas) y otro grupo sin conexiones entre sí, pero conectados por lugares comunes a los que hicieron viajes. Existe también un documento del cual una persona es autora y a través de esa persona se establece una conexión con tres personas más que a su vez conectan con el resto de la red. Estos patrones se consideraron claves en la búsqueda de aquel grafo similar entre los candidatos.

A continuación se analizan los grafos candidatos comenzando por los considerados más diferentes. El número 5 no posee esta estructura de separación entre dos subgrupos por lo que se descarta de inmediato. En el caso del grafo 4, a pesar de poseer esta separación, se observa que el grupo de nodos conectados por viajes están también conectados entre sí, patrón que no puede apreciarse en el grafo template. El candidato 3 presenta una subdivisión entre los nodos que no presentan viajes y este patrón tampoco se encuentra en el template. En cambio los grafos 1 y 2 poseen todos los patrones identificados en el template, y, de ellos, el candidato 1 es el más semejante en base a estas primeras visualizaciones.

b. ¿Qué partes clave de la mejor coincidencia ayudan a distinguirla de las otras posibles coincidencias?

En orden de explorar más detalladamente los patrones relacionados con los tipos de nodo, se procedió a visualizar al template y los candidatos en forma de grafos paralelos:

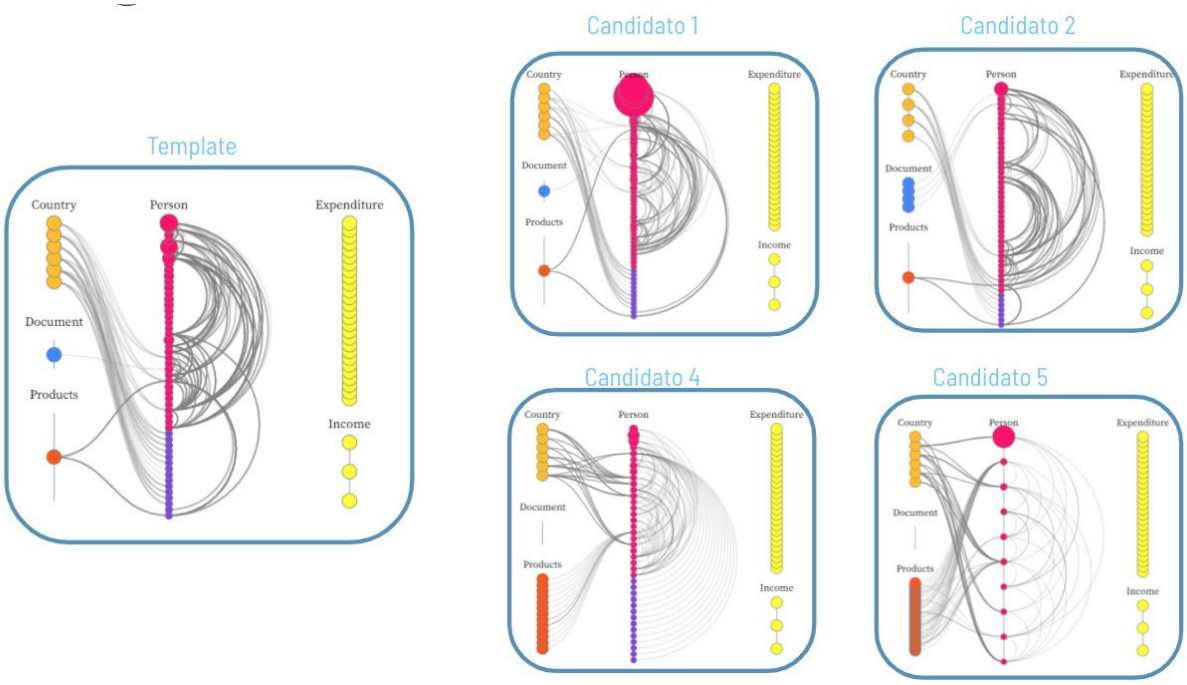

Figura 2 *Grafo paralelo*

En la anterior visualización, se reafirma lo anteriormente mencionado que los grafos 4 y 5 (el grafo 3 no se muestra ya que presenta patrones muy diferentes y quedó descartado de antemano) poseen patrones de conexión muy diferentes. Se puede apreciar que están más centrados en productos y conexiones entre personas. Por otro lado, en lo que respecta a los candidatos 1 y 2, podemos observar que la división entre las personas viajeras y aquellas interconectadas mediante algún medio se encuentra más claramente definida. El candidato 2 tiene a su vez más documentos y las personas interconectadas entre sí no tienen una densidad tan alta como el candidato 1. En base a lo último expuesto, el grafo 1 sigue siendo el mejor candidato de todos.

Un aspecto importante a explorar es el temporal y como ocurren las relaciones en base a ello, tanto las comunicaciones como los viajes o compras y ventas.

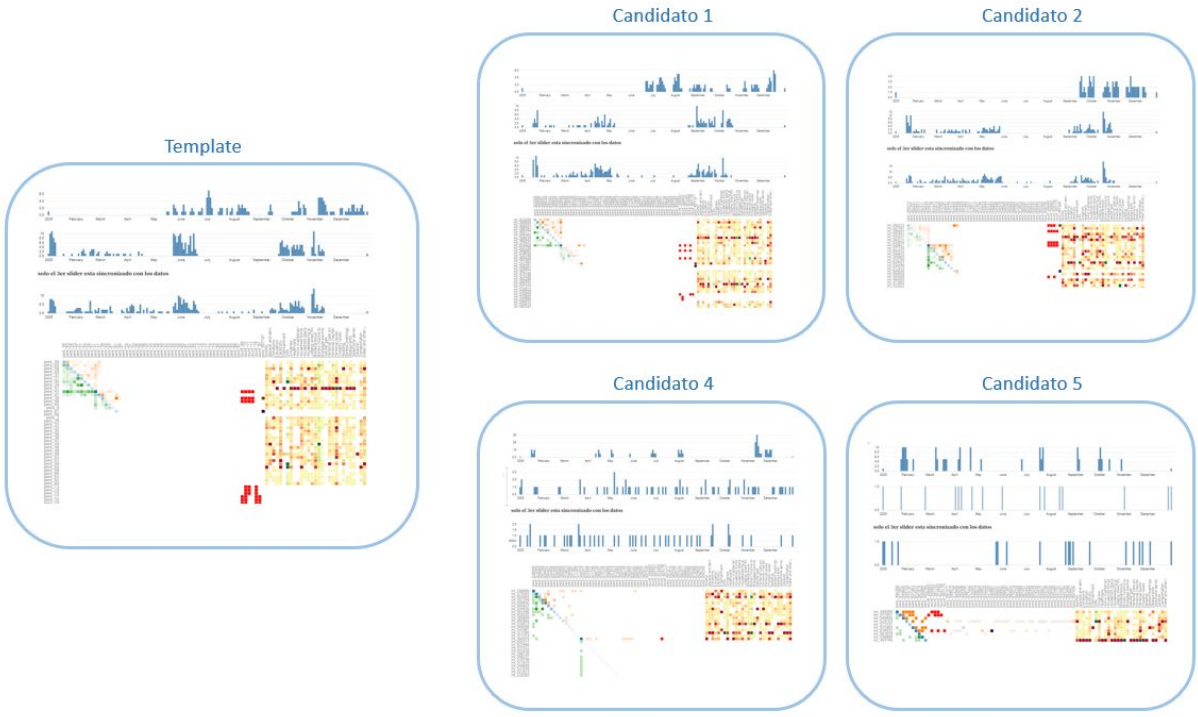

Figura 3 Matriz de adyacencia

En esta visualización presentan de forma expandida las conexiones y los viajes de las diferentes personas a lo largo del tiempo. En la diagonal superior de color naranja, figuran los emails; en la diagonal inferior, en verde, las llamadas telefónicas, y en azul las conexiones totales para cada nodo. Además se agregaron a la derecha de la misma: en rojo los viajes realizados a los diferentes países, luego con una escala divergente las compra-ventas de productos, y a continuación con otra escala divergente las finanzas. La opacidad de la primera parte de las celdas de la matriz representa la cantidad de conexiones, relativa al máximo visualizado. Mediante 'mouseover' se puede visualizar la cantidad absoluta de dichas conexiones, o del balance de finanzas / productos. También se incorpora un filtro temporal que permite ver la actividad en el lapso seleccionado, donde además el filtro está expresado como un gráfico de barras temporal que muestra la actividad en el periodo seleccionado. Los 3 filtros / gráficos temporales representan viajes, llamados telefónicos y mails (el filtro activo es solo el de los emails; no se pudo sincronizar la selección entre los diferentes filtros).

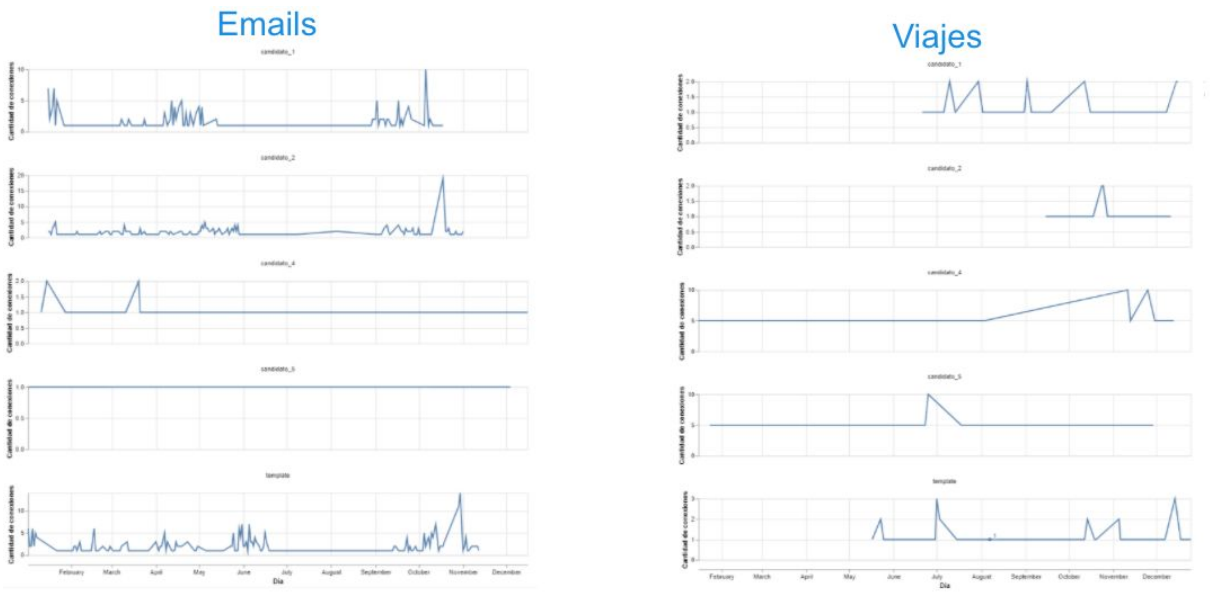

Figura 4 Comparación temporal

En el grafo template, las comunicaciones del grupo interconectado por algún canal de comunicación ocurren de la siguiente manera: en intervalos cortos de tiempo, seguidos de un periodo largo sin comunicaciones y finalizando con un pico antes de la caída del internet. En cuanto a los viajes, éstos ocurren solamente a partir de la segunda mitad del año y son realizados por las personas pertenecientes al grupo restante (que no tienen comunicación entre ellos).

Observando los candidatos y sus patrones de comunicación se puede apreciar fácilmente que los candidatos 1 y 2 siguen el mismo patrón, en especial el grafo 1 que tiene sigue de manera muy similar los picos y caídas respecto al template. En cambio, en los demás posibles candidatos no se observa nada similar.

En cuanto a los viajes, se puede ver para todos los candidatos que éstos ocurrieron en la segunda mitad del año. Sin embargo el patrón con diversos picos se observa más semejante en el candidato 1, por lo que confirmamos que este es el candidato que más se parece al template.

2. **CGCS tiene un conjunto de IDs "semilla" que pueden ser miembros de otras redes potenciales que podrían haber estado involucradas. Echa un vistazo al gráfico muy grande. ¿Puede determinar si esas ID conducen a otras redes que coinciden con el template?**

Las 3 semillas provistas consisten en un arco de algún tipo de relación en el grafo grande. Hay dos semillas de relación persona-documento (S1 y S2) y una semilla de relación persona-producto (S3). La idea general en este punto es encontrar en el grafo completo un recorte del mismo que sea lo suficientemente similar al grafo template.

Lo primero que se realizó fue definir un perfil para cualquier nodo de tipo persona. Para eso se eligió el perfil demográfico y perfil viajero.

Perfil demográfico: se generó como un vector de dimensión igual a las categorías demográficas existentes. Para cada categoría se le asigna el atributo monto, o cero en caso de no contar con tal categoría. Ej: 3 categorías (A, B y C) y una persona gastó 22 en A y recibió 15 por C, entonces su vector perfil es de la forma: [22, 0, 15].
Perfil viajero: se generó como un vector con todos los países origen y países destino posibles, y se le asigna la cantidad de veces que viajó desde o hacia un país. Ej: Existen 3 países (A, B y C), una persona salió 2 veces de A y arribó 1 a B y 1 a C,entonces su vector perfil es de la forma: [2, 0, 0, 0, 1, 1].

Antes de comenzar con la búsqueda de la red dentro del grafo completo se tuvo que definir una medida de comparación entre los nodos y sus perfiles. La medida de comparación elegida fue la suma de la distancia coseno de los dos tipos de perfiles entre el de los nodos de tipo persona.

En el punto anterior se identificó que existen dos grupos separados de nodos y que solo existen 3 nodos que conectan dichos grupos; estos los identificamos como **nodos hubs**. Éstas personas son importantes ya que proveen la conexión necesaria para indicarle al algoritmo en qué momento debe buscar el grupo de viajeros.

El primer paso del algoritmo es identificar manualmente la semilla con algún nodo en el grafo template e identificar los nodos hubs en template (ids: 39, 40, 41). Luego comienza un proceso iterativo y voraz (greedy) que consiste en cada paso buscar los nodos vecinos del grafo completo que sean más similares a los vecinos que se encuentran en el grafo template, utilizando solo las conexiones de comunicaciones. En esta primera etapa solo se identificó al grupo que no realizó viajes. En una segunda etapa se parte de los nodos hubs, tanto del template como en el grafo completo, en busca de los países visitados y, a partir de éstos, comienza la búsqueda de los nodos más parecidos que visitaron dicho país.

A partir de las semillas 1 y 3 se obtuvieron sub-grafos que pueden ser utilizados para comparar contra el grafo template. A partir de la semilla 2 no se logró identificar ningún sub-grafo y tampoco se analizaron otras posibilidades de cómo iniciar una nueva búsqueda a través de esta última semilla.

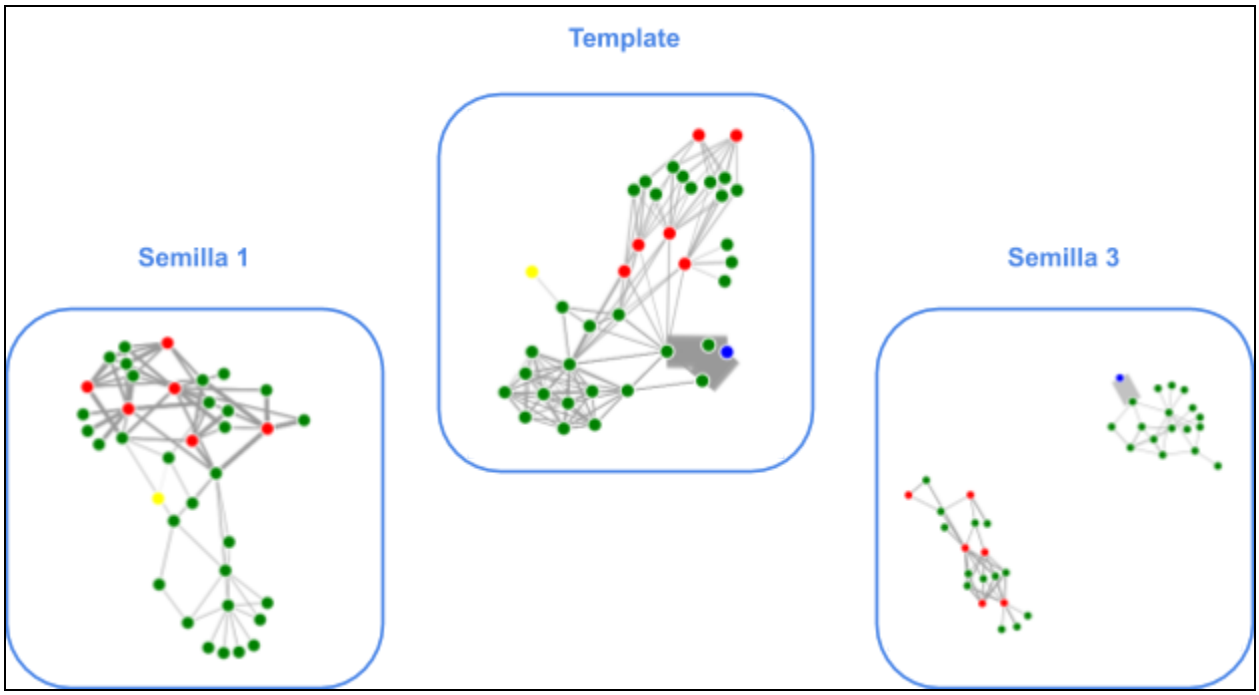

Figura 5 *Grafos a partir de semillas*

Con los mismos métodos explicados en el punto 1, se puede visualizar el subgrafo obtenido del grafo completo y verificar si tiene alguna similitud con el grafo template. También se podría agregar una funcionalidad cuando se hace hover sobre algún nodo persona del template, que destaque su equivalente en el subgrafo y muestre en un panel lateral el detalle de los perfiles y el valor de similaridad entre ambas personas.

4. En base a sus respuestas, identifique el grupo de personas que cree que es responsable por el corte. ¿Cuál es su razonamiento?

Cada visualización tiene sus ventajas en aspectos particulares. Por ejemplo, el grafo de fuerza permite visualizar patrones de comportamiento en el grafo template a nivel macro, de forma tal que puedan contrastarse con los patrones presentes en los demás grafos. Por otro lado, el grafo paralelo permite distinguir, de forma visual, los diferentes tipos de nodos, pudiéndose de esta manera apreciar otro tipo de patrones. A su vez, el gráfico temporal hace foco, tal como su nombre lo indica, en patrones relacionados con la dimensión “tiempo”, comparando cada candidato contra el template, y filtrando por tipo de conexión. Por último, a partir de la matriz de adyacencia es posible apreciar, con una forma alternativa, patrones de comportamiento y conexión entre los nodos, a partir de diferentes gamas de colores (una por cada tipo de conexión).

Entendemos que el grafo 1 presenta los patrones de comportamiento más similares al template. Si bien hay otros grafos que presentan algunas similaridades específicas, el grafo 1 es el que más consistentemente se asemeja al template. Fundamentalmente observamos:

- La existencia de 2 grupos bien diferentes: aquellas personas densamente interconectadas por comunicaciones (e-mails/llamadas telefónicas) y otro grupo sin conexiones entre sí, pero conectados por lugares comunes a los que hicieron viajes. Existe también un documento del cual una persona es autora y a través de esa persona se establece una conexión con tres personas más que a su vez conectan con el resto de la red.
- El patrón de comunicaciones (llamadas y e-mails) tiene picos y caídas temporales semejantes al del template, con una caída abrupta antes del corte de Internet.
- No hay viajes en la primera mitad del año, y luego se incrementan paulatinamente.

5. ¿Cuál fue el mayor desafío que tuvo al trabajar con el grafo completo? ¿Cómo superó esa dificultad? ¿Qué podría hacer más fácil trabajar con ese tipo de datos?

Por el gran volumen del grafo completo resulta imposible visualizar patrones sin antes aplicar algún criterio de recorte. Las semillas fueron de ayuda como un buen punto de partida para la exploración, y a ello se sumó la comparación de perfiles para poder seleccionar personas similares a las del template.