# Homework VII

Ajay Shaan Shanmugam and Siddharth Chandrasekaran
2008 Vast Challenge MC - II

November 9, 2017

**Problem Statement:**

For this homework, we have chosen the VAST 2008 mini challenge 2 which involves identifying the social network of specific persons of interest using visual analytics. The call record dataset provided includes information on the caller, receiver, the duration of call, the corresponding timestamp and the cell tower through which the call was made. Please note that the caller and receiver records do not have useful identifiers. i.e They use numbers instead of actual names. Using provided bits of information on these suspects, we undertake the task of identifying them using a network graph, clustering the dataset using unsupervised learning techniques and gaining insights from their call patterns over the 10 day period. We aren't using the cell tower information for this mini challenge since the geo location data is relevant only to solve the grand challenge.

**Techniques used and challenges faced for the visualization:**

Since Bokeh doesn't have default support to make graphs (it has an option to render networkx graphs but it is not possible to add interactive visualizations and widgets to the dataset). We had to improvise and create a graph using the basic glyphs of bokeh such as a circle for the node and a line for edges (with thickness based on number of emails, importance of the email and the length of the message body).
Since we needed the graph to be well aligned and evenly spaced out we decided to optimize the modularity of the graph using the louvain technique. This ensures that the graph gets divided into "communities" based on the density of the communications and that these communities are evenly spaced out. We found a python module called python-louvain which can do this for us and used it to get the coordinates of the edges and the nodes which can be plotted in bokeh.
We added a widget to give the viewer the option to color code nodes based on clusters or to display all the nodes in a single color (to look at the data devoid of the bias clustering algorithms give us). This graph helps us identify patterns in the network, inspect clusters, individuals (using hover).

We used three clustering algorithms, namely
1. Triangle based clustering
2. Clustering based on geometric average of the subgraph edge weights
3. Square clustering

and displayed all of them one after the other (bokeh doesnt have the support to visualize them side by side on different tabs). We didnt want to add a drop down box since we wanted to compare and contrast the algorithms at the same time. These three visualizations helped us understand the overall social network and trends between individual clusters (one of the color coded clusters is the Paraiso network).

**Clustering Algorithms in detail :**

- Triangles:
  This clustering algorithm is based on triplets of nodes. A triplet consists of three connected nodes. A

triangle therefore includes three closed triplets, one centered on each of the nodes ( this means the three triplets in a triangle come from overlapping selections of nodes). The triangle clustering coefficient is the number of closed triplets (or 3 x triangles) over the total number of triplets (both open and closed). This measure gives an indication of the clustering in the whole network (global), and can be applied to both undirected and directed networks.

$$C = \frac{3\times\text{number of triangles}}{\text{number of connected triplets of vertices}} = \frac{\text{number of closed triplets}}{\text{number of connected triplets of vertices}}.$$

- Clustering based on geometric means of subgraphs
  For every node, the clustering coefficient is determined as the geometric mean of edge weights of the subgraphs that involve that node. Normalization of the edge weights with the maximum weight in the entire graph is done before taking the geometric mean.

$$c_u = \frac{1}{deg(u)(deg(u)-1))} \sum_{uv} (\hat{w}_{uv}\hat{w}_{uw}\hat{w}_{vw})^{1/3}$$

- Square:
  This technique computes the squares clustering coefficient for nodes. For each node we return the fraction of possible squares that exist at the node. Here, for every node, the probability that two neighbors of a node share a common neighbor that's different from itself is determined. i.e the fraction of possible squares that have the node as a vertex is calculated for every node. These values are then used to cluster the nodes.

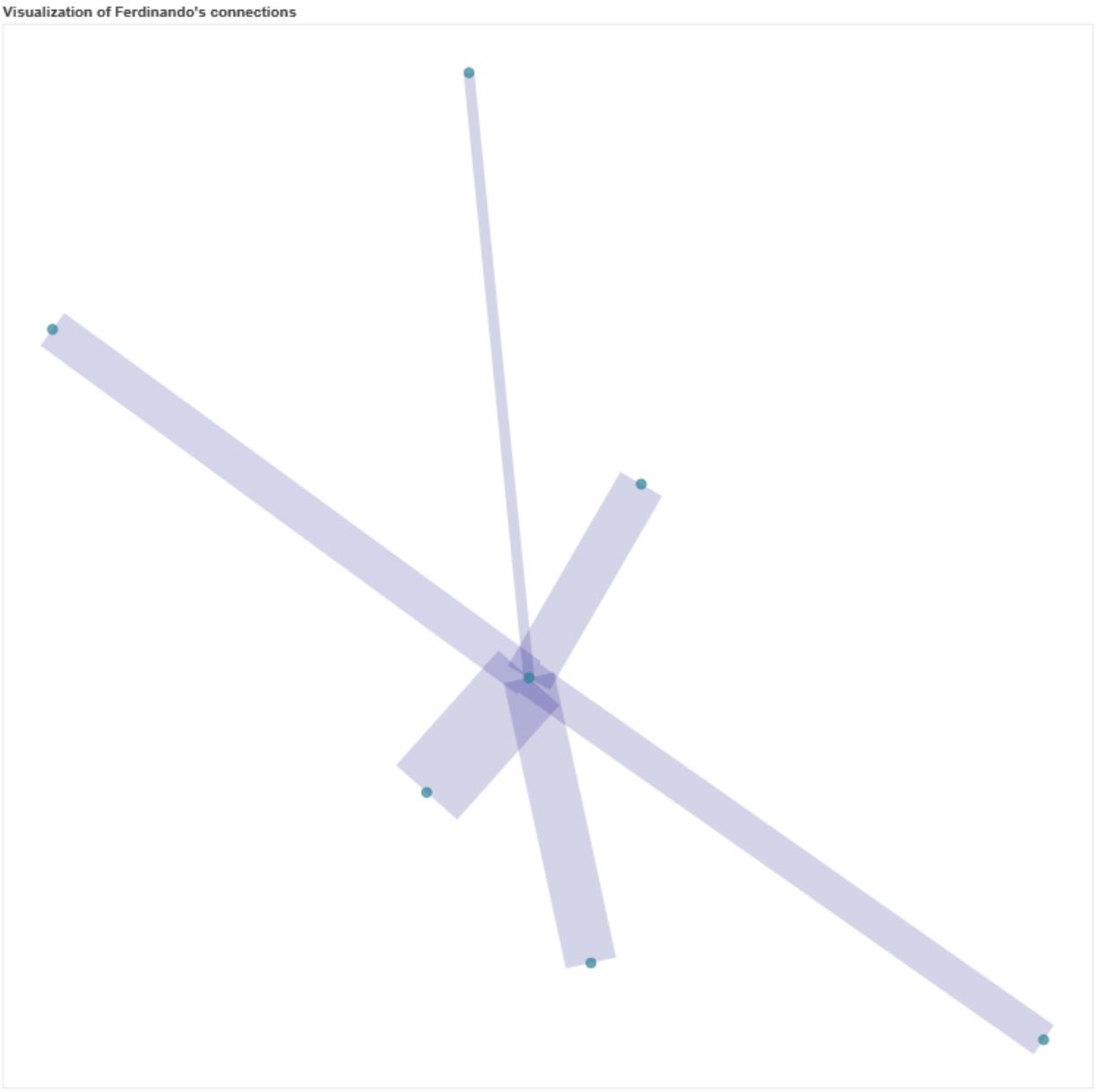
$$C_4(v) = \frac{\sum_{u=1}^{k_v}\sum_{w=u+1}^{k_v} q_v(u,w)}{\sum_{u=1}^{k_v}\sum_{w=u+1}^{k_v}[a_v(u,w)+q_v(u,w)]},$$

**Layout used for the visualizations:**

- Spring layout:
  In this layout, the nodes are positioned according to Fruchterman-Reingold force-directed algorithm. This algorithm aims at minimizing crossing edges and maximizing number of equal length edges. Nodes that are not connected by an edge tend to be farther apart than connected nodes. Also nodes of same cluster tend to have shorter coils of spring between them than nodes of varying clusters. The force on these springs are inversely proportional to the edge weights, and therefore as the graph reaches equilibrium after many iterations, the nodes connected by thinner (lighter) edges are found to have sprung out further than the ones with heavier edges.

**Results:**

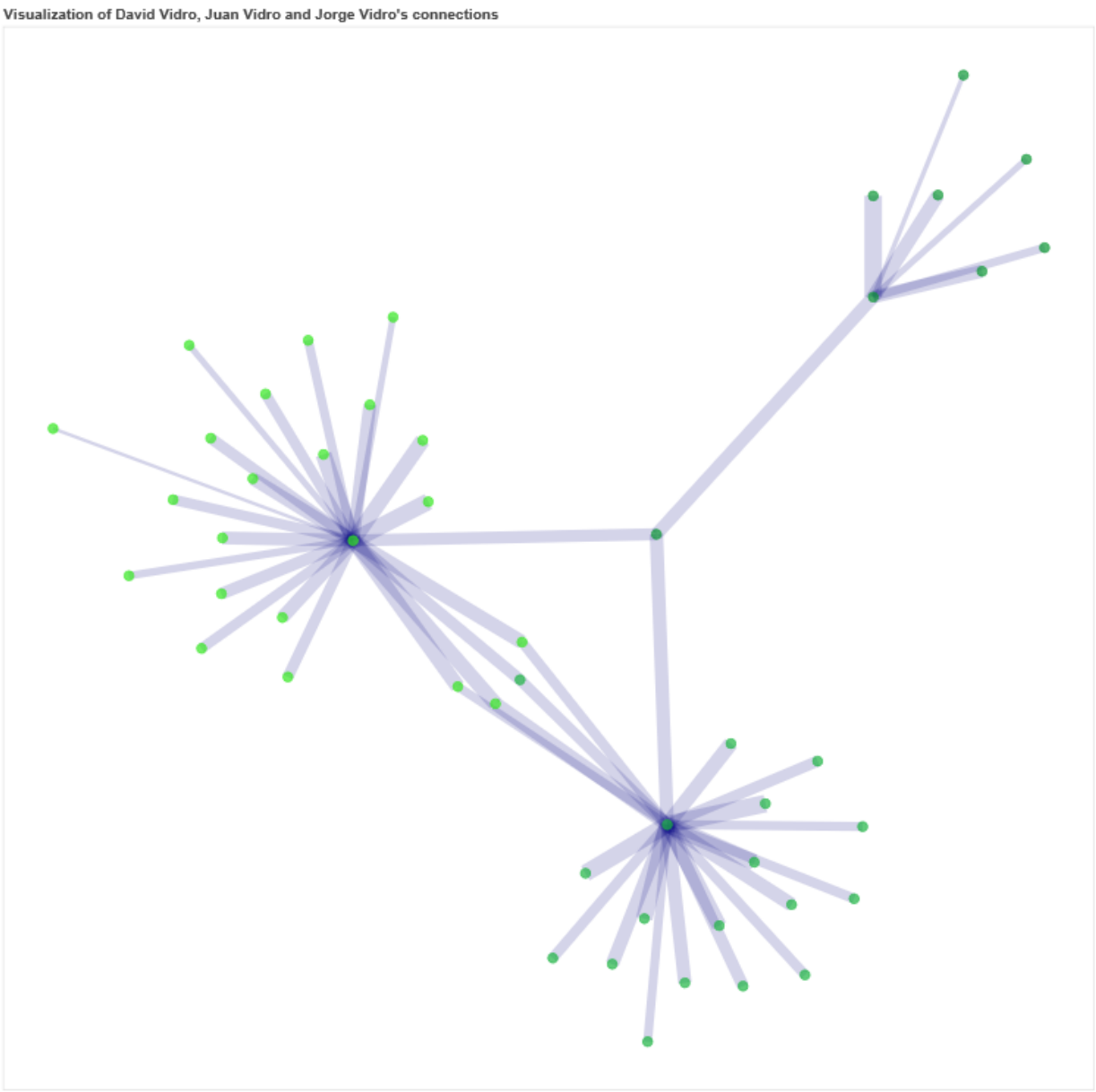
The fourth plot delineates Ferdinando's connections based on the call log dataset. Here, the weight aka thickness of an edge is determined by both the frequency of calls made between the nodes connecting the edge and the duration of each call made. This gives the best representation of the magnitude of talktime between our nodes of interest.




Given that identifier 200 is Ferdinando and based on the provided hint that Ferdinando talks to his brother Estaban most frequently, we identify node 5 (The node connected to node 200 by the thickest edge) to be Estaban.

We have specific information that David organizes a lot of communication among the network he can be identified from plotting the Paraiso network as show below

Another clue that's been provided to us is that close relatives and associates that Ferdinando would be calling include David Vidro, Juan Vidro and Jorge Vidro, in addition to his brother Estaban. Therefore, out of the other 5 nodes, the nodes connected to Ferdinando by the next 3 thickest edges would be David Vidro, Juan Vidro and Jorge Vidro.

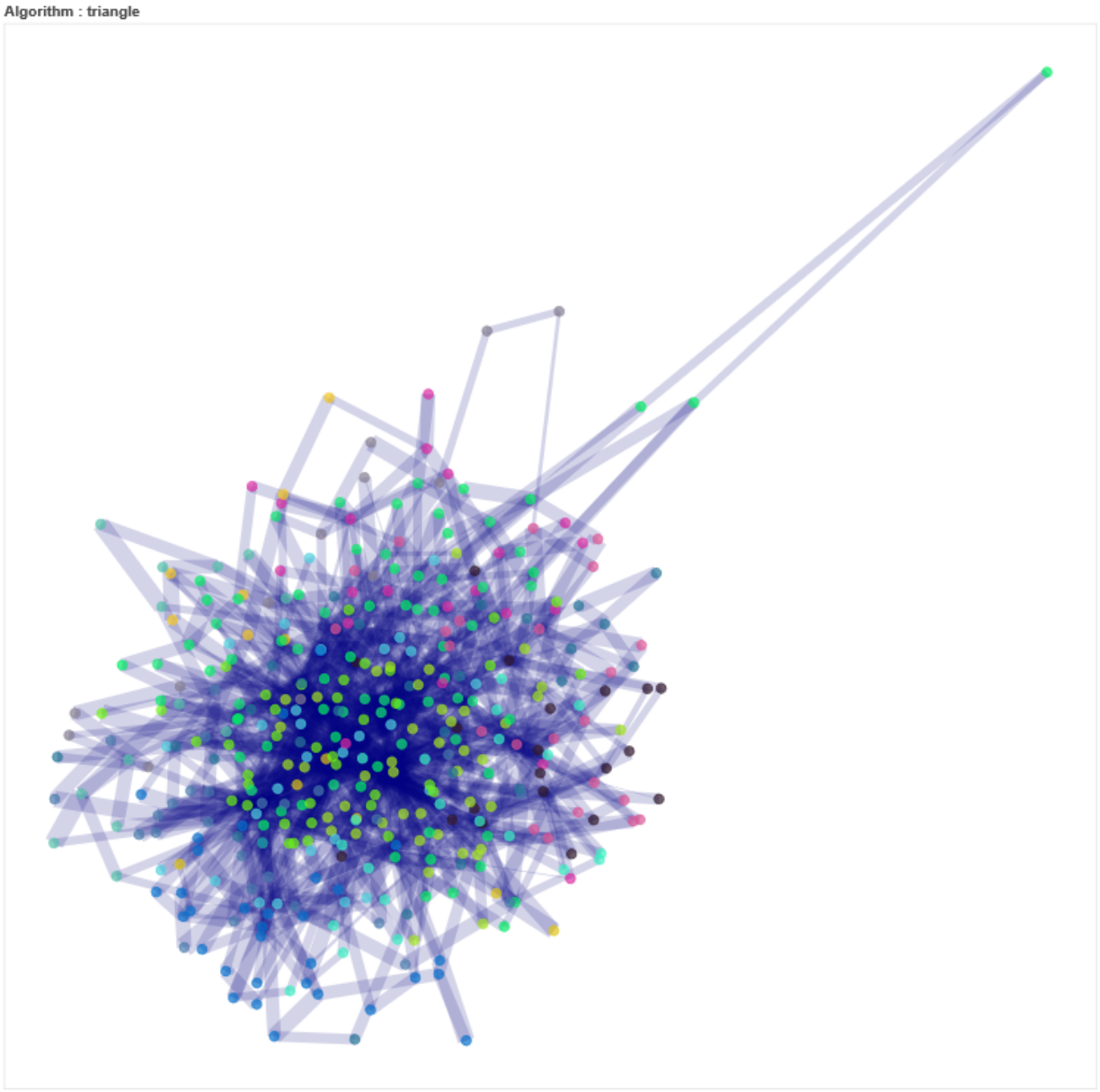



The final hint that we've been given is that David organizes most of the Paraiso activities. This gives us a strong reason to believe that he must have the most number of connections among the three. Therefore, we plot a new graph (5th graph) outlining all connections of David, Juan and Jorge. From this, we infer that node 2 has the highest degree (the most number of connections) and therefore we identify him as one of the masterminds behind the Paraiso activities, David!

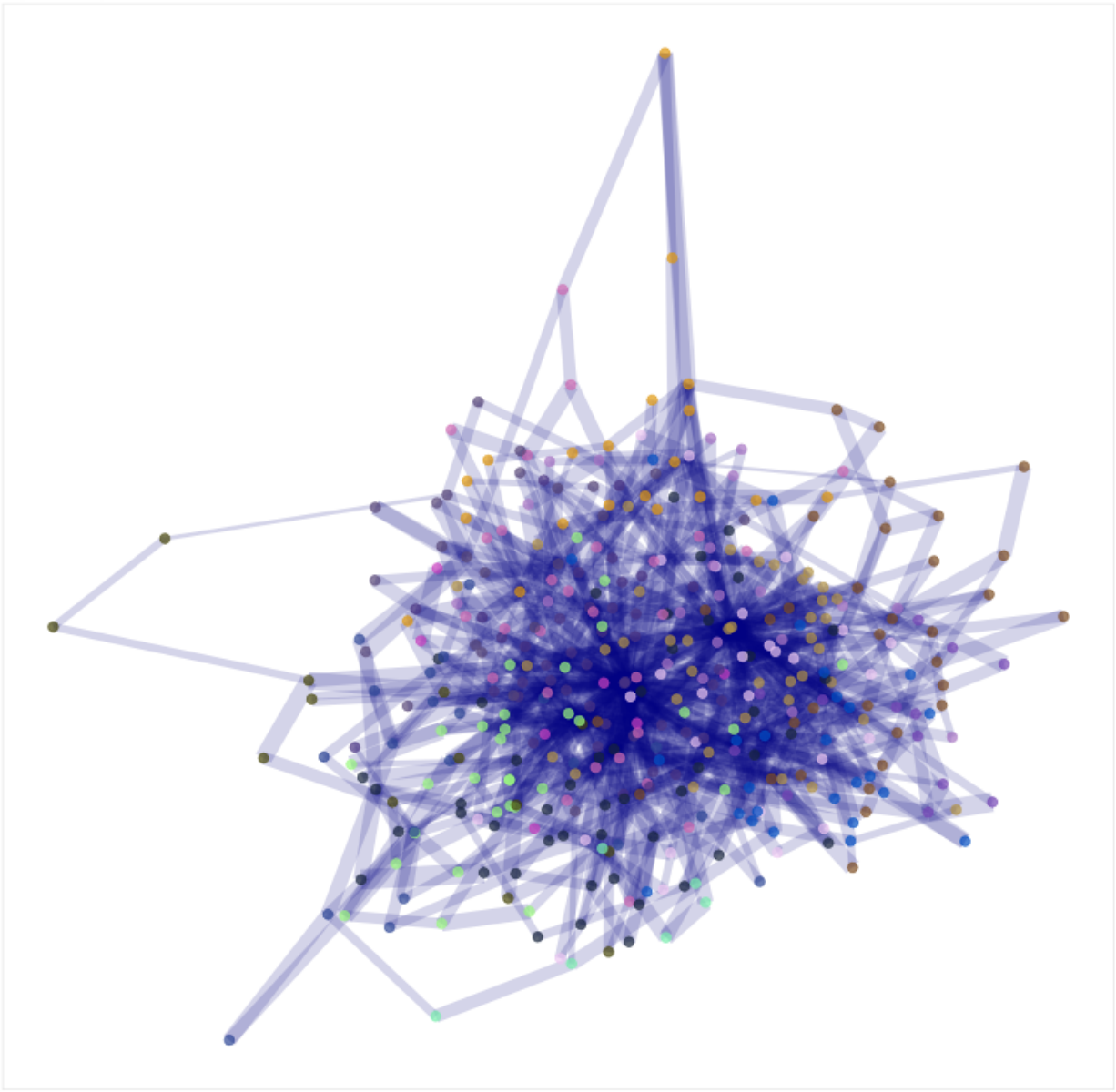
As for distinguishing between Juan and Jorge, we don't have enough information in the problem statement or in the dataset to make a judgement on their identities.

Another interesting inference from the fifth graph is that nodes 0, 1, 34, 27 are mutual connections of David and Juan/Jorge. Since they're confirmed to be David's associates, we can infer that Juan/Jorge, David, Estaban, and the aforementioned nodes are definitely a part of the Paraiso activities.

**Screenshots from the overall social network visualizations with the various clustering algorithms:**



Algorithm : square



Algorithm : clustering