Analysis of cell phone calls history

Z. Konyha, K. Matković, W. Freiler*
VRVis Research Center, Austria

R. Miklin, T. Lipić, M. Berić†
University of Zagreb, Croatia

D. Gračanin‡
Virginia Tech, USA

ABSTRACT

This paper provides an overview of the tools and techniques used in our analysis of a ten day long cell phone call history data set of the VAST 2008 Challenge[1]. Contestants in this challenge are expected to identify persons and temporal/geographical patterns in the data set. We have combined the evaluation of network graph properties and visualized the call history and aggregate data computed from it to accomplish this task.

Index Terms: I.3.0 [Computer Graphics]: General—; I.3.6 [Computer Graphics]: Methodology and Techniques—(Interaction techniques); J.4.1 [Social and Behavioral Sciences]: Sociology—

1 ANALYSIS TOOLS

In our analysis we used ComVis[2], an interactive visualization application based on the concept of multiple, linked views and composite brushing. ComVis offers histograms, scatter plots, parallel coordinates and function graph views [3] that can display time series data. Brushes in the same or in different views can be composited using sequences of AND, OR and SUB operations. This dynamic filtering is a key feature for interactive analysis [1].

ComVis is a single-user application, but it can capture the state of the visual analysis session (data, brushes, views, etc.) and store it in a single file. Collaborators can exchange such files to provide the common context and framework for visual analysis. This makes off-line collaboration possible. This funcionality allowed our team members located in Vienna (Austria), Zagreb (Croatia) and Blacksburg (USA) to better leverage regular communication channels (audio/video conferences, workshops) using the common visual analysis context captured in a `.cvv` file.

We also used a Python script to compute aggregate data and a Python package called NetworkX[3] to compute various graph properties.

2 ANOMALIES IN THE DATA SET

We started by looking for suspicious items in the data. Some of the towers in the provided map appear to be located in the ocean. This is clearly strange, but the description of the challenge indicates that the map is not accurate.

We found four records with negative call duration. We discarded them. We could not see any common features or patterns in those four calls (traces of a malfunctioning tower, for example) which would make us consider discarding further records.

Based on the locations of the towers and the times of the outgoing calls one can compute the average speed the person was traveling at between subsequent calls. There are several records which indicate incredible speeds, for example, covering more than 30 miles in only four minutes. Again, we found no common pattern in those records. We did not discard them and at the end of our analysis we learned that they have no impact on our results.

*e-mail: {konyha, matkovic, freiler}@vrvis.at
†e-mail: r.miklin@gmail.com, {tomislav.lipic, marko.beric}@fer.hr
‡e-mail: gracnain@vt.edu

[1]http://www.cs.umd.edu/hcil/VASTchallenge08/
[2]http://www.comvis.at
[3]https://networkx.lanl.gov/

3 ANALYSIS PROCEDURES

Our analysis involved several iterations of studying quantitative graph properties, looking at visualizations of the detailed call history and finding patterns in the aggregate data computed from the call history.

3.1 Network Graph Properties

We created a Python script that uses NetworkX to gather information about the network graph described by the call history.

The instructions of the challenge suggest that ID200 is Ferdinando Catalano. We verified this assumption. Node 200 has six neighbors, 1, 2, 3, 5, 97 and 137. Those six nodes, in turn, have many. This supports the assumption that ID 200 is the leader of the network. He has few contacts, but those are in touch with a considerable part of the network.

We learned that the graph is connected, there are no isolated nodes or subgraphs. We tried to find central nodes in the network. Closeness centrality is considered to be a good centrality measure in social network analysis [2]. Informally, large node centrality means that the node is connected to many others over short paths. We found that nodes 1, 5, 0, 309 and 306 have large closeness centrality. This gave us the idea that some of those nodes can be associated with some of the names in the challenge. We started an interactive, visual analysis procedure for more details.

3.2 Visual Analysis and Aggregates

In Figure 1 we brushed calls from ID200 to each of his partners and learned that he called ID5 most often. ID5 also called ID200 most often. We concluded that ID5 is Estaban Catalano. However, we did not know how to assign the remaining numbers to names.

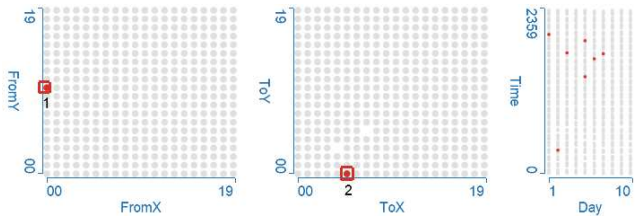

Figure 1: Each point in the 20x20 matrix represents a caller (left) or callee (middle). Caller 200 and callee 5 are brushed. The scatter plot on the right shows the days and times when ID200 called ID5.

We took a different approach and computed and analyzed aggregate data instead of looking into details. For each person, the aggregate data includes the number of incoming and outgoing calls, number of IDs called by the person, number of IDs calling the person. We also created a time series indicating the number calls aggregated over one hour long periods. In other words, $f_p(t) = \{$number of calls to or from person p in the time frame $[t, t+1)$ hours, where $0 \leq t < 240\}$. We created separate time series for incoming and outgoing calls, too. For each person, there is one function graph of 240

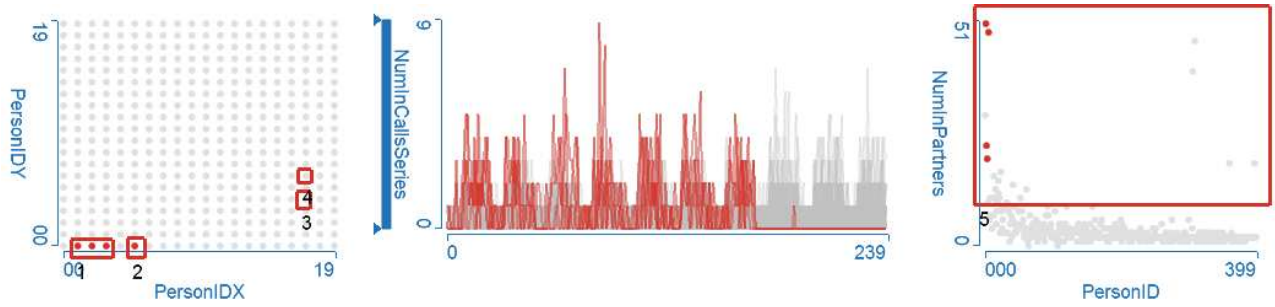

Figure 2: The logical OR of the four brushes in the matrix select contacts of Ferdinando Catalano. The scatter plot on the right shows the number of people calling each person. Brush 5 narrows the focus to persons who were called by many people. The function graph view shows that they made and received many calls in the first seven days, but only few afterwards.

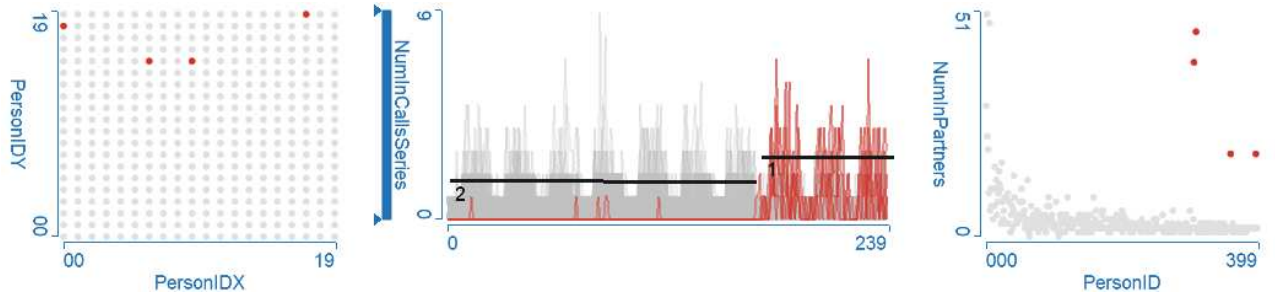

Figure 3: IDs that received many calls in the last three days, but not in the first seven are selected by a combination of two brushes. The brushes are the black horizontal lines in the function graph view. The scatter plot on te right shows that they received calls from many people.

sample points in Figure 2. This display is an alpha-blended composition of 400 individual function graphs. The peaks and valleys corresponding to daytime and nights are immediately visible.

We brushed Ferdinando Catalano's contacts in the matrix. In the scatter plot on the right, we narrowed the focus to persons who received calls from many people. We learned that about 50 people called IDs 1 and 5, and about 20 called IDs 2 and 3. We assume that 1 and 5 are important coordinators. We already know ID5, so we concluded that ID1 is David Vidro. IDs 2 and 3 belong to Juan and Jorge Vidro, but we cannot decide which is which.

The function graph view reveals that they talk often in the first seven days, but not in the last three. There are only very few nodes with this pattern. We tried to find people with the exact opposite pattern: many received calls in the last three days only. Brush 1 in Figure 3 selects graphs with many received calls in the last three days. Brush 2 subtracts those with many received calls in the first seven days. The matrix shows that there are only four persons with this pattern: 306, 309, 363 and 397. Interestingly, they are also called by many people.

In a Python script, we compared the sets of people calling IDs 1, 2, 3 and 5 to those who called IDs 306, 309, 360, 397. We found that most of the people calling ID5 (Estaban Catalano) in the first seven days were calling ID306 in the last three. IDs (1,309), (2,397) and (3,360) constitute similar pairs. We concluded that those pairs of numbers belong to the same persons. They started using different numbers after day seven. We also found that the only common contacts of IDs 1, 2, 3 and 5 are 0 and 200. ID0 is active throughout the ten day period. It is an important node in the network but we do not know the associated name. The only common contacts of IDs 306, 309, 360 and 397 is 300. ID300 also becomes active on the last three days while ID200 starts talking to different people than he did before. Therefore, we assume that Ferdinando Catalano started using ID300 after day seven.

In a linked scatter plot, we studied the locations of towers those phones were calling from to get an idea of the geographical extents of the movement. We found that in the first seven days they stayed in the middle of the island or in the north in cities near towers 11 and 30. However, in the last three days some of them moved to the south of the island. Significantly, on day 10 Estaban Catalano is near tower 12 and David Vidro moves to tower 22. On day 9 and 10 Ferdinando Catalano is near tower 17.

While computing aggregates, we discovered that there are records where the person initiates or receives a call while another one is in progress or calls more than one person at the same time. For example, ID1 called IDs 2, 3 and 5 at 10:00 on the first day, and at 10:03 on day five. We assume that those were conference calls.

4 Conclusion

We have not used Comvis for the analysis of such networks before and we were positively surprised by how well basic linked scatter plots with advanced composite brushing can perform in this context. The exploration of the aggregates was a crucial step for the success of the analysis. The information computed in our Python application was also valuable and matched the results of the visual analysis. We believe that this kind of semi-automated information retrieval can be quite productive and accurate for larger data sets.

References

[1] C. Ahlberg and B. Shneiderman. Visual information seeking: tight coupling of dynamic query filters with starfield displays. In *CHI '94: Conference companion on Human factors in computing systems*, pages 313–321, New York, NY, USA, 1994. ACM Press.

[2] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1:215–239, 1978/79.

[3] Z. Konyha, K. Matković, D. Gračanin, M. Jelović, and H. Hauser. Interactive visual analysis of families of function graphs. *IEEE Transaction on Visualization and Computer Graphics*, 12(6):1373–1385, Nov.-Dec. 2006.