VAST 2008 Challenge: Social network dynamics using cell phone call patterns

Romain Bourqui* LaBRI
Frederic Gilbert† INRIA Bordeaux
Umang Sharan‡ INRIA Bordeaux
Paolo Simonetto§ LaBRI
Faraz Zaidi¶ LaBRI

ABSTRACT

This paper is a summary of the contest entry submitted to the VAST contest 2008 mini challenge 3. The primary task of the mini challenge was to characterize the Catalno/Vidro social network based on the cell phone call data provided and to characterize the temporal changes in the social structure over the ten day period. This paper summarizes the data analysis performed on the synthetic data set provided, describes the visualization algorithms and tools employed and the key observations from our analysis. We use Tulip [1] for exploring the data set. Tulip is a scalable and flexible framework for visualizing large graphs providing the user an easy platform for exploring and manipulating large networks.

Keywords: Visual analytics,Social networks,Spatio-Temporal visualization,VAST contest

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces;

1 INTRODUCTION

The goal of the Mini Challenge 3: Cell Phone Calls (Social Network) is to identify the social network and the temporal dynamics associated with the evolving network structure over a period of 10 days based on a set of cell phone call records. The data set consists of information about 9834 calls between 400 cellphones over a 10 day period in June 2006 in the Isla Del Sueno. The records are expected to provide critical information about the Catalano social network structure. Given the quantity and nature of the data involved, we decided to employ Tulip to help analyze the evolution of the call structure over time. Tulip facilitates easy visualization of large graphs—such as *call graphs* and *location graphs* in this case. The rest of the paper discusses the design of the Tulip and its role in the data analysis task followed by the conclusion.

2 THE TULIP FRAMEWORK

The Tulip graph visualization framework provides an efficient and modular environment in which tools can be experimented for the purpose of information visualization research. Tulip enables the user to draw and display huge graphs, enable seamless navigation through different parts of the graph, employ different geometric operations and algorithms like subgraph extraction easily through plugins. One of the most challenging aspects of the mini challenge was the data analysis. Scripts were written to extract the caller and callee ids, time stamp, call duration and the cell tower location from the raw cell phone call logs. Thereafter, we imported the induced call graph and location graph in Tulip to facilitate better visual analysis.

We define a call graph $G_C = (V, E_C)$ as an undirected graph where nodes V correspond to caller ids and edges E_C correspond to calls between two nodes in the network. Similarly, we define a location graph $G_L = (V, E_L)$ as an undirected graph on nodes V where $e = (u, v) \in E_L$ iff nodes u and v used the same cell tower to make a call at least once. Further, we define a weighing function $f : E_L \rightarrow \mathbb{N}$ on the edge set E_L where, $f(e) = n$ and $e(u, v) \in E_L$ implies that nodes u and v used the same cell tower *at least n* times in the phone call logs. We also define a temporal sequence of snapshots for the call graphs and location graphs to analyze the temporal dynamics at a finer granularity. G_C^t and G_L^t $\forall\, t \in \{1,2,...,10\}$ represents the temporal snapshots of G_C and G_L respectively, one for each day. Figures 1 shows G_C in the center and G_C^1 and G_C^5 in the right column in Tulip.

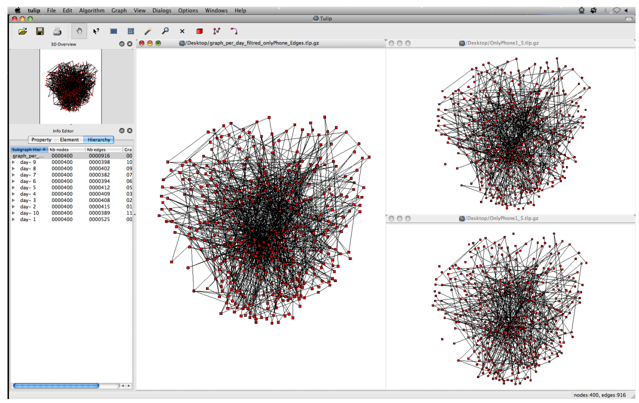

Figure 1: Call graphs G_C, G_C^1 and G_C^5

Similarly, Figure 2 shows G_L, G_L^1 and G_L^5 in Tulip. Tulip provides several views for the best visualization of the data set—a central view for visualizing the graph as a node-link structure with zooming and panning tools for quick navigation and a list view containing node and link attributes in the sidebar enabling quick filtering and attribute-based pruning.

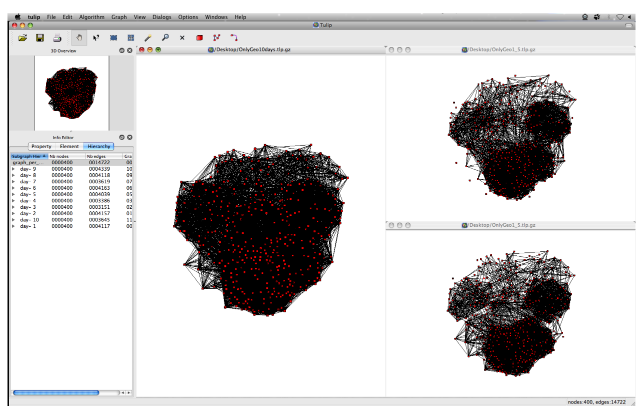

Figure 2: Location graphs G_L, G_L^1 and G_L^5

*e-mail: bourqui@labri.fr
†e-mail: frederic.pierre.gilbert@inria.fr
‡e-mail: umangsh@gmail.com
§e-mail: paolo.simonetto@labri.fr
¶e-mail: faraz.zaidi@labri.fr

3 Analysis Process

3.1 Setting up Tulip

We began our investigation by importing the entire cell phone call logs into Tulip, and exploring the call and location graphs using the multi-view interface supported by the tool. Temporal snapshots were created for both G_C and G_L. We further created call filter graphs for each temporal snapshot to differentiate between strong and weak relationships: $G^t_{L,i} = (V,E')$ where $e \in E'$ iff $e \in E^t_L$ and $f(e) \geq i \; \forall \; i \in \{1,2,...,5\}$.

3.2 Inferring Social Hierarchy

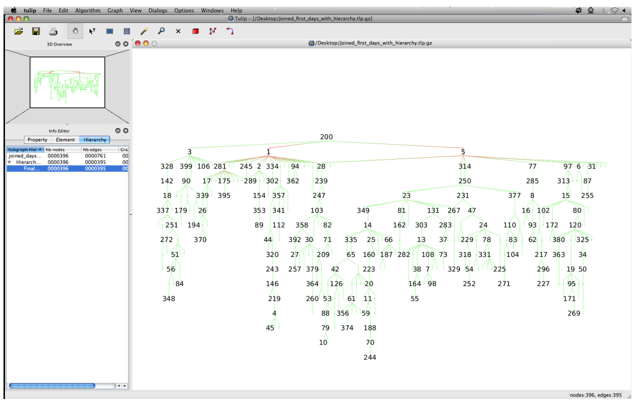

Figure 3: Catalano social network structure

We began our investigation into the Catalano social network by exploring relationships between different nodes based on the call graph and the location graph. We explored both G_C and G_L as we believed that both the call relationships and geographically proximity would play an important role in determining the social structure. We used several metrics like the delta-efficiency metric on $G^t_{L,i}$ and G^t_C to establish the social network hierarchy in the input data set. This metric was introduced by Latora et. al. and extended by Memon [4, 3]. We construct the 'importance' hierarchy on this social network by applying Kruskal minimum spanning tree algorithm on the delta-efficiency score matrix. Figure 3 shows the Catalano social network structure inferred from the data set provided. We infer that node 200 is the main boss while nodes 1,3,5 are important nodes in the Catalano hierarchy. We note that the call patterns change drastically on days 8,9,10 whereby the social hierarchy shifts and previously unimportant nodes—306 and 309—emerge into prominence (for more details, please refer to our answer page). To discount for such effects, we inferred the hierarchy by analyzing call patterns from days before day 8.

3.3 Temporal changes in Catalano social structure

We continued our analysis for characterizing the temporal changes in the Catalano social network in Tulip. We summarized the community structure using a clustering algorithm based on [2] on $G^t_{L,i}$ and G^t_C. Figure 4 shows that the clustering algorithm we applied to infer the Catalano social structure. The subfigure on bottom left visualizes the cluster proximity coefficient on $G^t_{L,i}$. We further explored the evolution pattern in the community structure with time. We analyzed the communities (clusters)—after filtering isolated nodes—that are isomorphic to each other at different time steps. Tulip provided an excellent interface to compute the cluster proximity coefficient matrix for each time step. Thereafter, we computed the cluster overlap measure for each pair of isomorphic clusters to infer the communities in the Catalano social structure. Figure 5 shows the Catalano social network at the end of the 10 day period. The blue edges denote strong intra-cluster relationships while the yellow edges represent the weaker inter-cluster edges.

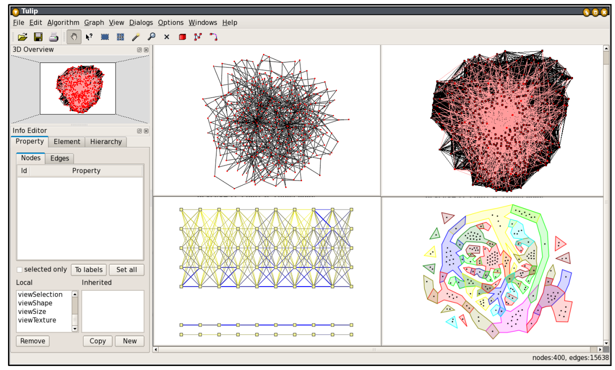

Figure 4: Change in network layout in last two days

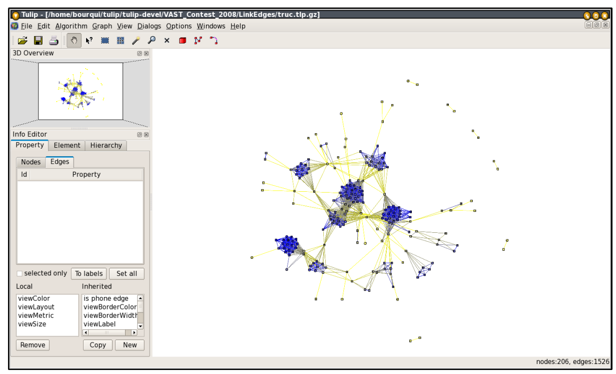

Figure 5: Catalano community structure

4 Conclusion

We use existing visualization tools like Tulip to explore relationships within the cellphone calls data set. Tulip provides the users with an overview of the entire data set facilitating exploratory data analysis. The extracted relationships modeled as the call graph and the location graph with temporal snapshots, serve as a concrete visual analytic technique to assist investigators in analyzing the Catalano social network structure. The strength of our approach lies in the simplicity in exploring and navigating through the data set using Tulip—easily switching between different views to extract the maximum information.

References

[1] D. Auber. Tulip : A huge graph visualisation framework. In P. Mutzel and M. Jünger, editors, *Graph Drawing Softwares*, Mathematics and Visualization, pages 105–126. Springer-Verlag, 2003.

[2] D. Auber, Y. Chiricota, F. Jourdan, and G. Melancon. 9th ieee symposium on information visualization (infovis 2003), 20-21 october 2003, seattle, wa, usa. In *INFOVIS*. IEEE Computer Society, 2003.

[3] N. Memon, D. L. Hicks, and H. L. Larsen. How investigative data mining can help intelligence agencies to discover dependence of nodes in terrorist networks. In *ADMA*, pages 430–441, 2007.

[4] N. Memon and H. L. Larsen. Practical approaches for analysis, visualization and destabilizing terrorist networks. In *ARES '06: Proceedings of the First International Conference on Availability, Reliability and Security*, pages 906–913, Washington, DC, USA, 2006. IEEE Computer Society.