Visual Analytics for Inspecting the Evolution of a Graph over Time: Pattern Discovery in a Communication Network

Bruno Schneider Carmela Acevedo Juri Buchmüller Fabian Fischer Daniel A. Keim

Data Analysis and Visualization Group*
University of Konstanz, Germany

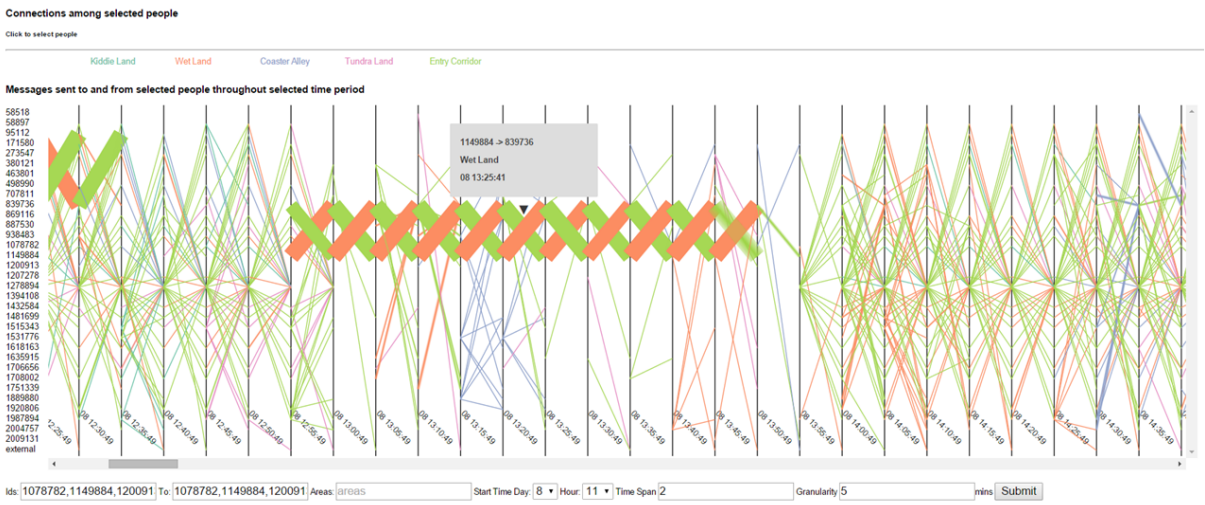

Figure 1: Messages sent and received among a group of visitors in the amusement park. The thick orange and green lines represent high communication activity between two visitors identified with there unique IDs. Each vertical bin corresponds to a time span of 5 minutes.

Abstract

In this paper, we present two approaches developed to visually analyze and find patterns in a communication network. The work was done for the VAST 2015 Mini-Challenge 2 (MC2), featuring a dataset with records of timestamps as well as identification of sender and receiver of text messages. Further information included the location from which a message was sent in the fictional amusement park. In the first approach, we present the data preprocessing pipeline we used for a custom visualization. In the second approach, we present how we used available data preprocessing and visualization software to get a quick and clear overview of the problem, and how we used the generated findings to feed our custom visualizations.

Index Terms: H.2.8 [Database Management]: Database Applications—Data mining; H.5.2 [Information Interfaces]: User Interfaces—Graphical user interfaces (GUI)

1 Introduction

This paper completes our submission for the VAST Challenge 2015, Mini-Challenge 2 (MC2). Visualization and analysis of graphs and their evolution over time is a well-known task in information visualization and visual analytics. In this work, we present two applications based on distinct approaches for tackling the analysis tasks for a communication network data as provided in the challenge.

*e-mail: firstname.lastname@uni-konstanz.de

The available dataset comprised three days of communication data among visitors of a fictional amusement park. This data relates to an app used by most of the visitors to send text messages to other visitors. In the course of three days, 9429 different users were communicating, and 4.153.329 messages were sent in total. Within all users, we identified two outstandingly communicative participants, which were involved in 12.26% of all communications. In comparison, the next most communicative user was involved in only 0.17%. Below, we will call these two major communicators the *broadcasters*. Lots of previous research on the visualization of evolving networks shows the necessity to deal with issues like overplotting, while still keeping the ability to analyze the temporal evolution of networks with high amounts of nodes and edges. Amongst many others, Alencar et al. [2], used small multiples to display several successive static plots with snapshots of a network over time, yet with scalability limitations. Gloor et al. applied animation techniques to represent evolving networks, that come at the risk of experiencing 'memory effects', a phenomenon well described by Nowell et al. [3]. To minimize these problems that are related to the limitation of our memory to retain the most important events to be seen in an animation, techniques like the provision of an auxiliary panel indicating the most important events over time, the usage of time-sliders and the visual encoding of specific and distinguishable temporal changes in graphs for better identification have been applied, as Ahn et al. point out [1].

2 Our Framework

We worked with two approaches for visually inspecting the evolution of a communication network over time. The first one included a data preprocessing stage, where features like the length of stay in

the park, the number of messages sent and received, distinction of external messages and amount of messages sent to distinct users and more were extracted. We applied a k-means clustering algorithm to these extracted features to find subgroups of IDs with similar communication patterns. Besides the idea of finding clusters, this approach helped in filtering the data and selecting subgroups of users (identified by their ID) for visualization, providing better legibility and enabling the usage of visualization techniques that support a limited amount of data objects. After the preprocessing stage, we implemented three visualization techniques to see communication partners and to follow communication patterns over time. The first is graph-based, the second is pixel-based and the third visualization is a custom, parallel coordinates based view on the communication patterns. We also implemented an auxiliary panel for the visualization of the clusters, after applying Principal Component Analysis (PCA) on each cluster.

For our second approach, no special data transformation was done prior to visualization. The switch from overview to detailed visualizations was enabled mainly through interactive filtering of time-intervals, IDs and park location selections.

3 Visual Pattern Recognition

In accordance with the visual analytics approach we have processed the communication dataset to provide a concise visualization aiding the user to form insightful findings. Each message in the dataset comes with four attributes: Time-stamp, ids of the visitor the message was sent to and from and the location of the sending visitor. For each visitor (A) we proceeded to extract several numeric features of interest, namingly the amount of the messages sent to A, the messages sent from A, the distinct visitors that send messages to A, the distinct visitors that receive messages from A, the hours A stayed at the park, the messages sent from A to external ids, the areas visited by A and the difference between distinct senders and receivers of A messages.

Taking these features into account there are three IDs that are clear outliers (the two *broadcasters*, and external messages labeled with a unique common identification) and were therefore not considered to be visitors but other types of entities, e.g. device maintenance servers, thus being disregarded in this section. After these features were extracted we clustered them with a *k-means* approach. These clusters were later visualized in a 2D scatter plot after reducing the dimensionality of the dataset using Principal Component Analysis (PCA).

To further understand the clusters we were dealing with we applied individual PCAs to each of them. We chose the features with the lowest eigenvalues as the ones defining each of them since these are the ones in which the cluster has the lowest variance.

This, however, was not enough to clearly identify different communication patterns among visitors. There were several clusters that seemed to have very similar behavior and we were not aware of the development of these features over time. To alleviate this problem we developed several visual and interactive components.

First of all we provide filtering sliders for each of our features. These, together with the given clusters and their eigenvalues, allow users to reduce the visitors to only a subset that is of interest. We see, for example, that on most clusters the average of messages sent to external IDs is relatively low, but our upper limit for this feature from the given data was 52. Then, we decided to filter users that have sent more than 20 messages to this ID.

The next component we developed was a pixel based visualization that shows the amount of messages sent from each ID over the course of the three days. From that component, we can select the group of visitors that sent the most amount of messages and inspect the connection among this subset in a node-link diagram where two nodes (visitors) are connected if and only if they have communicated among themselves.

Finally, we have developed a visualization that allows us to analyze the progress of communication among a group of visitors. Each visitor is given a vertical position while time is placed horizontally in buckets of the desired size encoded as two vertical lines for its start and end times (Fig. 1). If visitor A communicates with visitor B in time period T then there will be a line connecting t_1 (the start of T) and t_2 (the end of T) from A's to B's row. The width of the line encodes the amount of messages sent between A and B during T and its color is defined by the area in the park A sent the message from.

That visualization allowed us to identify different facts that are not instantly clear from the others. We can see who initializes communication between couples of visitors, which is particularly important to identify and understand the *broadcasters*. Since visitors usually respond within a short period of time, it is also possible to identify where the receiver of a message is at the time a message was sent to them. This was useful to discern between groups that walk together in the park and those that do not.

We developed another visualization with the aid of *Tableau* (Fig. 2) that allowed us to more efficiently identify global communication patterns and to detect unusual activity along the 3 days. Filtering per timespan, Location and IDs was provided. Each park location is represented separately in different columns.

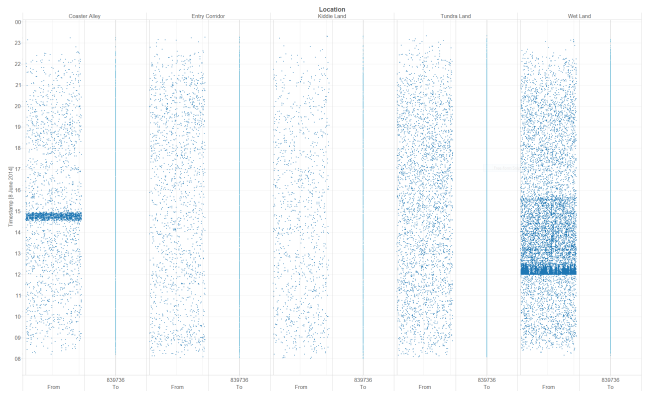

Figure 2: *Tableau* visualization of messages sent from All visitors to one of the *broadcasters*, Sunday (8-23h), All park locations. Time is along Y-axis, and along X-axis each ID with communication activity was plotted, keeping the same horizontal position over time.

4 Conclusion

The visualization developed with *Tableau* (Fig. 2) was very effective to find peaks of massive communication activity. The visualization as seen in Fig. 1 was developed to show who communicated with whom from selected visitors to show local patterns.

In conclusion, the presented visualizations were enabled solving the tasks of the challenge. In addition, the customized visualization (Fig. 1) proved to be helpful for inspecting the evolution over time of data structures that can be modeled as graphs.

References

[1] J.-w. Ahn, M. Taieb-Maimon, A. Sopan, C. Plaisant, and B. Shneiderman. Temporal visualization of social network dynamics: Prototypes for nation of neighbors. In *Social computing, behavioral-cultural modeling and prediction*, pages 309–316. Springer, 2011.

[2] A. B. Alencar, K. Börner, F. V. Paulovich, and M. C. F. de Oliveira. Time-aware visualization of document collections. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 997–1004. ACM, 2012.

[3] L. Nowell, E. Hetzler, and T. Tanasse. Change blindness in information visualization: A case study. In *infovis*, page 15. IEEE, 2001.