*First publ.in: Ieee Symposium On Visual Analytics And Technology (vast 2009) Challenge Awards (2009)*

# Integrative Visual Analytics for Suspicious Behavior Detection

**VAST 2009 Challenge Awards:**
**Traffic Mini Challenge: Excellent Analytical Technique Featuring Integration of Data Mining and Visual Analytics**
**Flitter Mini Challenge: Good Analytic Debrief**
**Grand Challenge: Outstanding Integration of Mini Challenge Results into Debrief**

Peter Bak* University of Konstanz
Christian Rohrdantz† University of Konstanz
Svenja Leifert‡ University of Konstanz
Christoph Granacher§ University of Konstanz
Stefan Koch¶ University of Konstanz
Simon Butscher‖ University of Konstanz
Patrick Jungk** University of Konstanz
Daniel A. Keim†† University of Konstanz

**ABSTRACT**

In the VAST Challenge 2009 suspicious behavior had to be detected applying visual analytics to heterogeneous data, such as network traffic, social network enriched with geo-spatial attributes, and finally video surveillance data. This paper describes some of the awarded parts from our solution entry.

**Index Terms:** H.5.2 [Information Interfaces & Presentations]: User Interfaces – Graphical User Interfaces (GUI); I.3.6 [Methodology and Techniques]: Interaction Techniques.

## 1 INTRODUCTION

For the VAST Challenge 2009 we focused on a tight integration of automatic data analysis steps with interactive visual analytics. The application of visualizations was twofold: They were used to generate new hypotheses as well as exploring and confirming the output of automatic algorithms. In order to extract the required information from the provided data, it was crucial to introduce visualization techniques at all stages of the knowledge discovery process, rather than to design more sophisticated and innovative ways of visualizing the same data. In conclusion, the contribution of the current work is in its methodology combining automatic algorithms and visualization, rather than in its techniques. For many insights standard visualizations were sufficient to generate new and interesting insights. For our analyses we made use of different open source programs like the data mining tool Konstanz Information Miner (KNIME) ([1], [3]) and the network analysis tool Pajek [2]. Nevertheless, the development of innovative techniques to analyze heterogeneous data sources should be encouraged by our work, in order to make it more efficient and applicable to larger datasets.

## 2 FEATURING INTEGRATION OF AUTOMATIC AND VISUAL DATA ANALYSIS FOR INVESTIGATING NETWORK TRAFFIC

For the solution of the Traffic Mini Challenge an analysis pipeline was designed that combined automatic analyses and interactive visual explorations in a process loop (see Figure 1). Different approaches led to partly overlapping and partly supplementing intermediate results, from which useful knowledge could easily be derived. The results of the first analysis steps triggered further investigations. The interactive loop between generating new insights and creating new sources of information was indispensable for the success of this process.

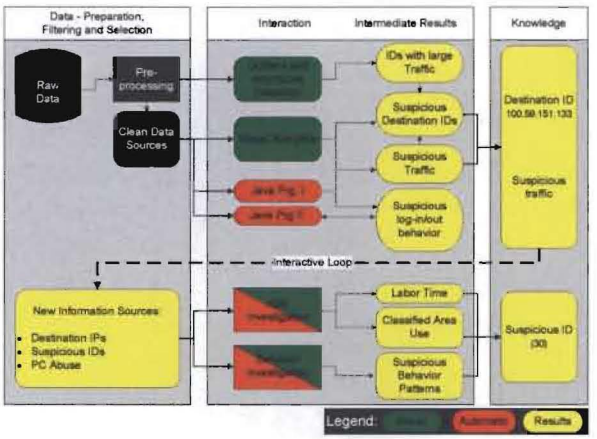


Figure 1: Pipeline used to extract information in the traffic Mini Challenge. Automatic analytic tools and visualization techniques are tightly coupled and introduced at all stages of the knowledge discovery process. The feedback loop – from generating new insights to creating new sources of information through user interaction – led finally to the desired results.

At the end any suspicious peculiarity had been checked and verified by multiple different approaches. We attained different independent evidences for the guilt of one specific person, so that we could identify the culprit free of doubt. Analysts were supported by automatic algorithms, so that their attention could be focused on what is considered the strength of humans: reasoning about visible patterns and relations, as well as deriving conclusions and planning further analyses.

The results of our analysis first led to the discovery of the IP address, to which classified information was sent. This was achieved through visual investigation of anomalies and outliers of Internet traffic. Consecutively, we extracted all the traffic to this IP address from all internal sources, which was done automatically. The automatic analysis was carried out using KNIME [1], an information mining application. Knowing these facts, our analysis further focused on the analysis of workers' behavior and alibi for the times suspicious traffic occurred. Finally, the worker that undoubtedly

*e-mail: bak@dbvis.inf.uni-konstanz.de
†e-mail: christian.rohrdantz@uni-konstanz.de
‡e-mail: svenja.leifert@uni-konstanz.de
§e-mail: christoph.granacher@uni-konstanz.de
¶e-mail: stefanmoritzkoch@googlemail.com
‖e-mail: simon.butscher@uni-konstanz.de
**e-mail: patrick.jungk@uni-konstanz.de
††e-mail: keim@dbvis.inf.uni-konstanz.de

sent classified information to a criminal organization could be identified.

## 3 Analytic Process for Analyzing Social Networks

In order to identify and extract the sub-network showing the interrelation between the suspicious employee and the criminal organization, we conducted and iterative approach as shown in Figure 2. For this purpose, first a program was written that automatically extracted interesting sub-networks. The set of possible candidate networks was then iteratively narrowed. Every extracted sub-network was visually explored and evaluated with Pajek [2]. Finally, a self implemented visualization tool was used to display its geographic implications.

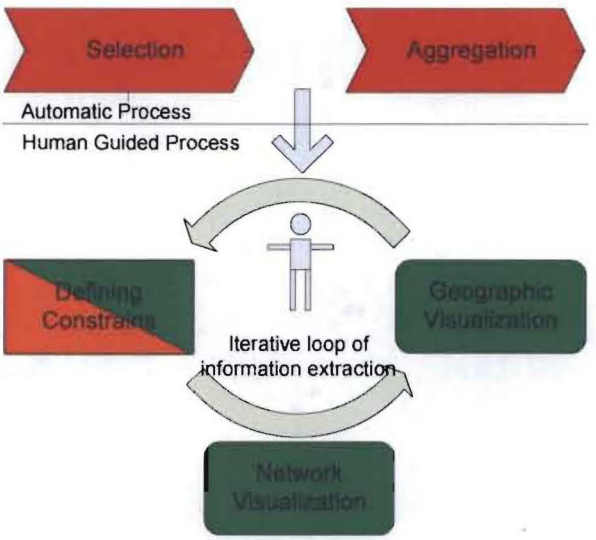


Figure 2: Analytic process of investigating social network data. The iterative loop of information extraction enabled the user to guide the knowledge discovery process.

The amount of sub-networks to be visually explored was fairly small, because even with relaxed constraints, only few candidates were returned. As a result, one sub-network was identified as an exact match to the given scenario, which is shown in Figure 3.

## 4 Integrating Results

The major task of taking part in the grand challenge was to provide a solution describing not only who the employee is that leaks information and how he/she could be identified, but also to characterize his/her behavior. This task implies the combination of the findings of the three mini challenges. The third mini challenge was also involved in this process, assessing the suspicious events from the surveillance video. Our great challenge, to integrate the results of these mini challenges into a final debrief, was to grasp the behavioral profile of the suspicious employee. This required a detailed pattern analysis of all components of network traffic, geographic locations, social connections and working hours, which finally led to an understanding of her/his criminal activity.

Our work was mainly based on generating and testing hypotheses. In this process visualization was indispensable, in order to gain access to the information beneath the data. Our approach was dominated by a top-down progress. In this, we first generated a complete list of hypotheses and step-by-step disregarded those that did

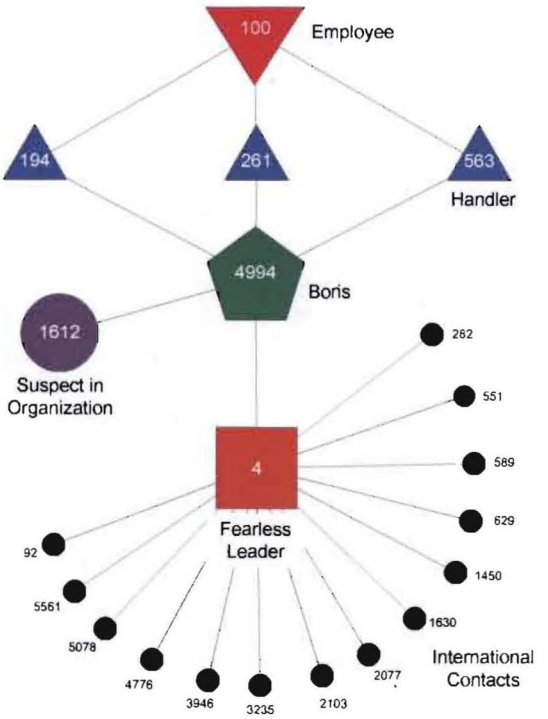


Figure 3: Results of the analytic process describe the scenario given by the challenge.

not fulfill the criteria, which were continuously narrowed, until one solution could be found.

## 5 Lessons Learned

Our approach to solve the given challenge was confirmed by our results. Nevertheless, the advantages and drawbacks of our approach should not be left undiscussed. As a guiding principle, we always aimed at involving the human as early as possible in the analytic process. This involvement was mostly supported by intuitive and easy to understand visualization techniques. Also, we tried to support this process by automatic analysis techniques. Their combination is, in our opinion, the most successful approach.

However, the data, task descriptions and background information provided by the challenge committee were very helpful and made the results possible. The provided data was easy to clean due to its artificialness and also easy to handle due to its limited size. Scalability and appropriateness to specific tasks and data sources still remains a research challenge in visual analytics.

## Acknowledgements

The authors wish to thank Hangzai Luo for his initial support.

## References

[1] Knime - konstanz information miner: www.knime.org.

[2] Pajek - network analysis and visualization tool: http://pajek.imfm.si/doku.php.

[3] M. R. Berthold, N. Cebron, F. Dill, G. D. Fatta, T. R. Gabriel, F. Georg, T. Meinl, P. Ohl, C. Sieb, and B. Wiswedel. Knime: the konstanz information miner. In *Proceedings 4th Annual Industrial Simulation Conference, Workshop on Multi-Agent Systems and Simulation (ISC 2006)*, 2006.