1. Publishable Summary

Scalable Data Analytics – Scalable Algorithms, Software Frameworks and Visualisation ICT-2013.4.2a

European project analyses Big Data to anticipate upcoming problems and helps mitigate them proactively

Vision and methodology

Eliminating or mitigating an anticipated problem, or capitalizing on a forecast opportunity, can substantially improve our quality of life, and prevent environmental and economic damage. Changing traffic light priority and speed limits to avoid traffic congestions, for example, will reduce carbon emissions, optimize transportation and increase the productivity of commuters. Similarly, adding credit cards to watch lists as a result of forecasting fraud will reduce the cost inflicted by fraudulent activities on payment processing companies, banks, insurance companies and merchants, and consequently lower credit card rates. Crucially, at the business level, making smart decisions ahead of time can become a differentiator leading to significant competitive advantage.

During its first two years, SPEEDD brought together leaders in Academia and Industry to develop technology that foresees and automatically responds to significant events. These teams have worked on state-of-the-art Event Processing components, proactive Decision Making modules and innovative Visualization tools to assist operators and organizations with the demanding task of event forecasting and corrective decision-making. These tasks are especially challenging in the era of Big Data, where the rate of incoming events to such systems ranges from the tens of thousands to millions of events per second.

To achieve its goals, SPEEDD has met several milestones. First, to better understand the needs of traffic and bank operators, the team conducted field studies with the control room staff in the DIR-CE traffic centre in Grenoble, France, and consulted bank personnel and Feedzai experts to develop insights into common fraud management practices. Second, through the use of state-of-the-art open-source streaming and messaging platforms, IBM's PROTON Event Processing engine was integrated with the Decision Making algorithms implemented by ETHZ and the Visual Analytics platform designed by the University of Birmingham. Third, advanced Machine Learning algorithms were developed by NCSR "Demokritos" in order to provide human-digestible event patterns. Fourth, Technion devised advanced distributed algorithms for monitoring the deviation of a global model towards anomalous states. Last but not least, the team has been able to "close the loop" and evaluate the quality of decision making by incorporating a traffic micro-simulator developed by CNRS, based on the commercial product AIMSUN.

Use cases

The proposed approach is applicable to a wide range of domains where proactivity is necessary. During the project, the SPEEDD technology is being tested in two such domains:

- *Proactive traffic management,* aiming to forecast traffic congestions and, as a result, act in order to attenuate them.
- *Proactive credit card fraud management*, aiming to significantly improve fraud detection accuracy, without compromising detection efficiency, and forecast various types of fraudulent activity, which are constantly evolving, in order to mitigate the effects.

The data streams in both use cases are highly complex and uncertain: data often convey erroneous information, there are delays in data transmission, crucial information is often missing and the corresponding event patterns are imprecise. Currently there are no solutions that are able to sufficiently deal with data streams of this size, complexity and uncertainty. State-of-the-art on-line event forecasting techniques can handle neither the size of these data streams nor their uncertainty.

Progress so far

Objective 1: Advanced event processing technology

In the second year of the project, we developed the following pieces of innovative event processing technology. First, we produced a tool for incremental weight learning using Markov Logic Networks. Our method computes a weight value for each event pattern, indicating a degree of confidence. Big Data streams are handled by means of a windowing operator.

Second, we developed a tool for incremental structure learning under uncertainty. The tool optimizes simultaneously the weight and structure of event patterns, producing event patterns from scratch, or refining existing ones. To deal with Big Data streams, our method makes uses of background knowledge to prune the search space of event patterns, and employs a windowing operator. Our empirical comparison with state-of the-art techniques confirmed the effectiveness of the method on event pattern construction.

Third, the IBM Proactive Technology Online (Proton) CEP engine has been extended, incorporating, among others, refined implementations of the sigmoid function and trend operators. Consequently, Proton can effectively perform complex event recognition and forecasting under uncertainty.

Objective 2: Innovative proactive event-driven decision-making tools

To meet this objective, we continued working on worst-case, stochastic and randomized decision-making techniques. First, we developed a distributed ramp metering algorithm, for the traffic use case, that achieves almost globally optimal performance despite relying exclusively on local information. We established, for monotonic traffic models, the theoretical conditions for global optimality using distributed control. The insights gained from this work have also shed light on when coordination between local controllers can result in an improvement in performance. Moreover, we constructed a signalized traffic model for inner-city traffic control and developed efficient control algorithms using decomposition techniques that support real-time performance.

Second, being motivated by the credit card fraud and traffic use cases, we have started exploring the use of inverse optimization to identify an expected value cost function used in the rational decision model of human “expert” decision makers.

Third, we presented a distributed model-free randomized algorithm that learns optimal policies for a multi-agent system based only on the pay-offs received when playing the policies on-line. Uncertainty in this case is encoded through uncertain pay-offs the agents receive for their actions; the optimization in policy space relies on randomized steps to deal with this uncertainty. The resulting method was applied to the traffic use case to learn the optimal coordination pattern between ramps for freeway traffic control in the presence of uncertainty due to varying traffic demands. We designed a distributed learning algorithm and established theoretically that the algorithm can learn Pareto-efficient solutions in the presence of disturbances from a finite set, despite having no knowledge of the system model or utility functions.

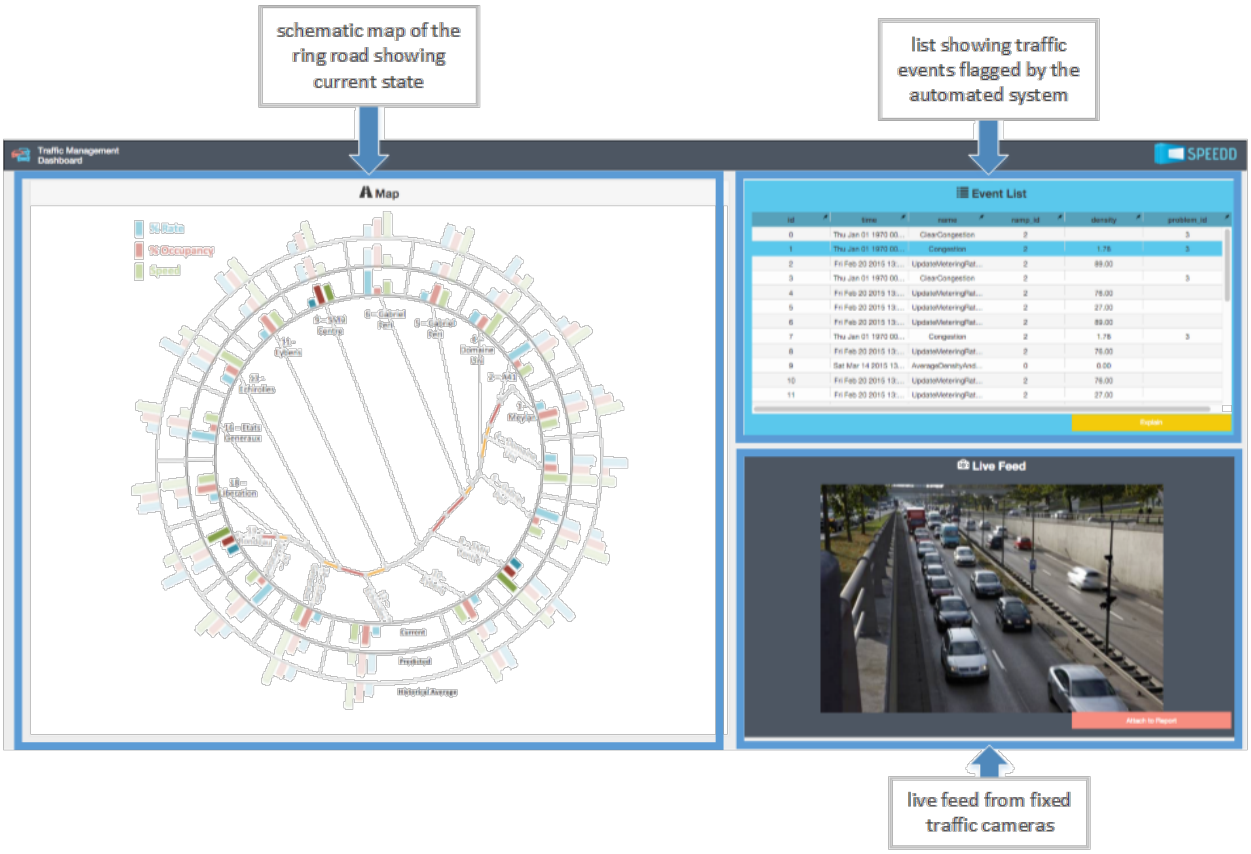

Figure 1: User Interface of the SPEEDD prototype for proactive traffic management

Objective 3: Visual analytics for real-time interaction with Big Data and proactive decision-support

Our design philosophy is to develop User Interfaces based on a detailed understanding of user requirements which, in turn, are based on a theory of work and decision making. For the fraud use case, we have been developing a model of decision making that corresponds to information search for different patterns of fraud. The results of the model correspond with those from trials involving human participants (both experienced fraud analysts and inexperienced students). This work fulfills the plan for integrating eye-tracking with rational decision models, for predicting the impact of user interface design on decision making.

For the traffic management use case, we have been exploring the impact of automation of decision making. This work provides an indication of both the manner in which reliability of automation can

influence ‘automation bias’ and a set of baseline parameters against which operator performance can be considered.

We have been developing and implementing forms of visualization which are novel for the operators and which extend underlying ideas of ‘ecological interface design’ (see Figure 1 and Figure 2). In both use cases, laboratory studies have been conducted to evaluate the impact of the designs on performance, and visits to operators have been made in order to evaluate the user interface design concepts.

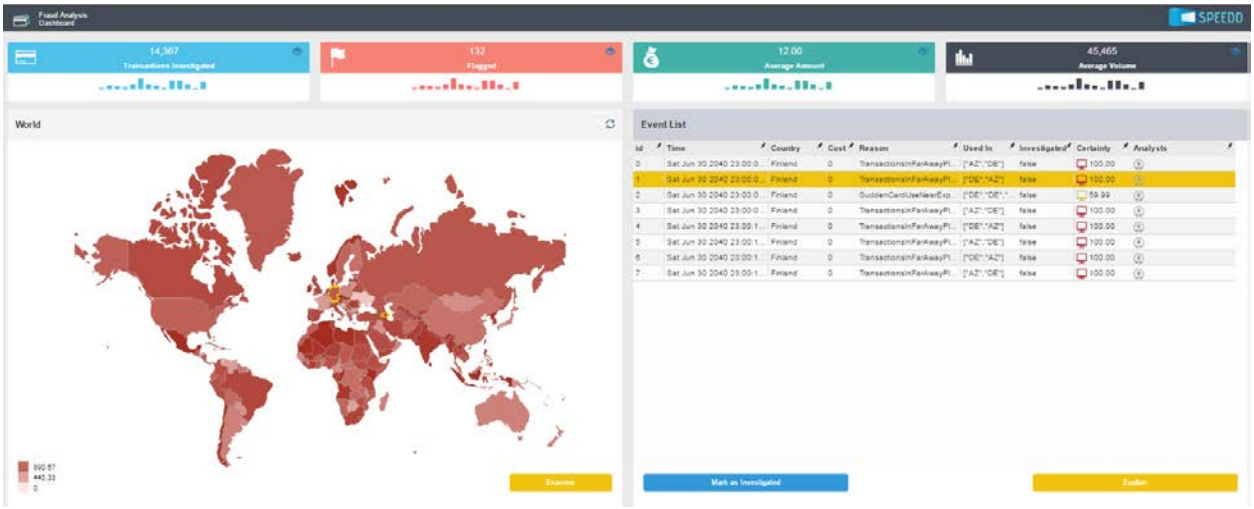

Figure 2: User Interface of the SPEEDD prototype for proactive credit card fraud management

Objective 4: Highly scalable data monitoring

To produce a highly scalable proactive event-driven prototype, we have advanced the state-of-the-art by developing novel computation-scalable and communication-scalable algorithms. In the former case, we have been working on a chain topology for Non-Deterministic Finite State Automata (NFA) that supports lazy complex event recognition. This lazy mechanism waits until the most selective event in a sequence arrives, and then adds events to partial matches according to a predetermined order of selectivity. The method is extensible to all event operators such as negation, disjunction and conjunction.

With respect to communication-scalable algorithms, we have been working on *efficiently* monitoring the on-going deviation of a learned model, expressing a complex event pattern, from a hypothetical true global model. Following the predictive nature of SPEEDD, we started with linear regression. We extended this paradigm to more machine learning models, such as Linear Discriminant Analysis. The problem here goes beyond regression, because there is a need to deal with matrix inversion in a distributed setup.

Real-world demonstrations

The second versions of the SPEEDD technologies have been integrated into the prototype for proactive event-driven decision-support. The SPEEDD runtime platform is based on state-of-the-art stream processing technology and messaging infrastructure. The integrated prototype has been instantiated in the proactive traffic management and the credit card fraud management use cases (see Figure 1 and Figure 2). These two real-world cases demonstrate end-to-end flow of the selected scenarios using the SPEEDD runtime platform.

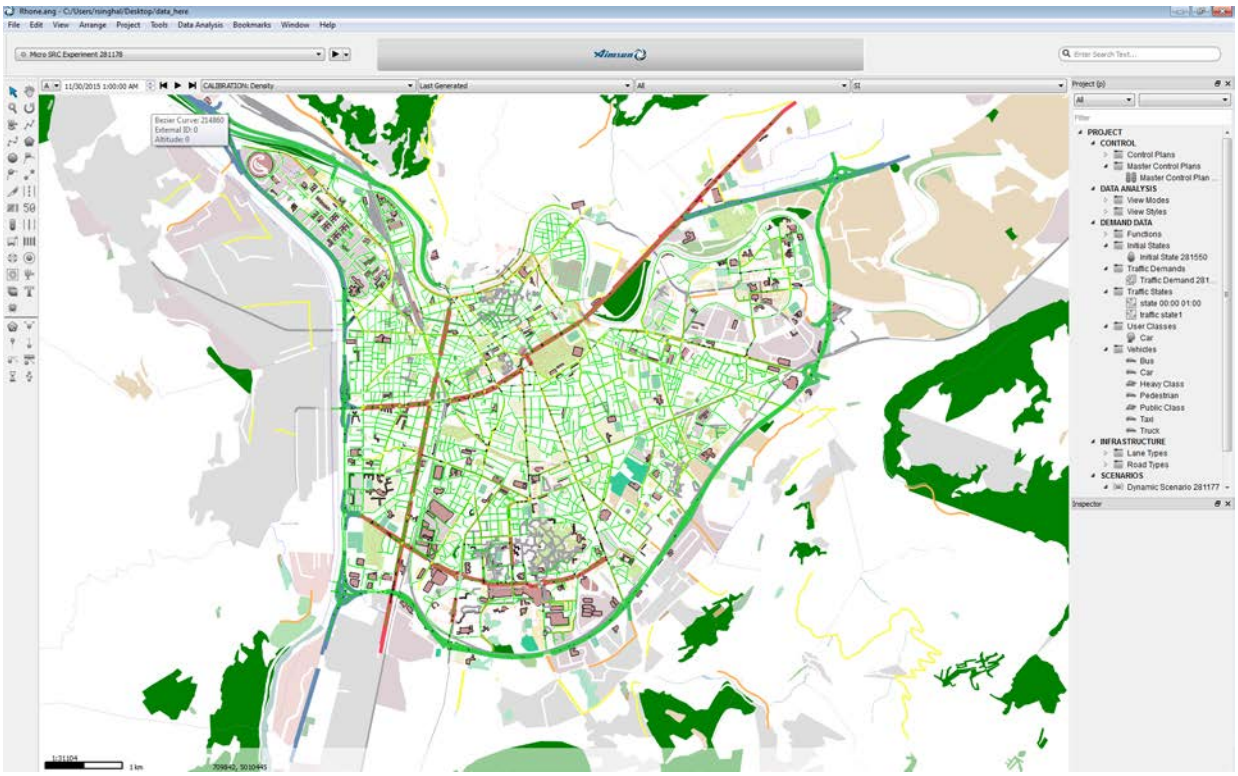

Figure 3: Traffic simulator

The large-scale demonstration of SPEEDD on traffic management incorporates a further key achievement of the project: the development of a traffic micro-simulator. The simulator is essential because it provides synthetic data for thorough experimentation, and enables testing proactive decision-making by allowing us to close the loop and see in real time the effects of decisions. In the second year of the project, the micro-simulator has been finalized, and includes a model of the entire city of Grenoble (see Figure 3).

SPEEDD' s impact

SPEEDD aims at innovative businesses, which realize that it is not sufficient to produce technology for processing Big Data. They want to produce intelligent technology that will allow them to make the most of the data, thus maximizing their impact to business and their corresponding market share. In order to bring the SPEEDD prototype to the market, we developed a business plan, which includes analyses of the market, the competitive landscape, business value, risk management, targeted domains and tangible assets of the project. All partners are moving together to empower SPEEDD and gain advantage in both the industrial and academic areas.

The volume, variety and uncertainty of Big Data Streams are both a challenge and an opportunity. Event-driven proactive decision-making is made possible by the combination of innovative technologies from the areas of complex event processing, control and human-computer interaction. It is the potential of this combination that makes SPEEDD such an exciting endeavor. Building on SPEEDD technologies, European companies, including those of the SPEEDD consortium, will be able to provide technology and services that add significant value to the business of their customers, thus positioning them well ahead of existing and foreseen competition in the Big Data market.

Dissemination Highlights

The SPEEDD researchers continued to publish their work in top journals and conferences, including the Machine Learning Journal, Automatica, IEEE Transactions on Automatic Control, the IEEE Conference on Decision and Control, the ACM International Conference on Distributed Event-Based Systems, the ACM International Conference on Knowledge Discovery and Data Mining, and the ACM International Conference on Human Factors in Computing Systems. Furthermore, two publications received **best paper awards** at the **ACM International Conference on Distributed Event-Based Systems** and the **International Conference on Ergonomics & Human Factors**.

Plans for the final year

For its final year, SPEEDD aims to break even more new ground. Some of the project's goals include the automated learning of event processing rules from raw data, designing distributed and event-driven decision making algorithms for the traffic use case, evaluating decision making in the credit card use case and integrating collected eye-tracking data to refine the Visual Analytics component. Above all, SPEEDD aims to provide a prototype of a system that will allow businesses to make smart decisions ahead of time, by analyzing, filtering and reacting to highly heterogeneous and massive amounts of noisy and uncertain data, while outputting human-understandable solutions in real-time.

Further information:

http://speedd-project.eu/

http://twitter.com/speedd_project

http://www.linkedin.com/groups?home=&gid=8238655&trk=anet_ug_hm

Project Co-ordinator: Georgios Paliouras, NCSR “DEMOKRITOS”

paliourg@iit.demokritos.gr http://www.iit.demokritos.gr/~paliourg/