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.

SPEEDD will develop a prototype for **proactive event-driven decision-making**: decisions will be triggered by forecasting events — whether they correspond to problems or opportunities — instead of reacting to them once they happen. The motivation for proactive computing stems from social and economic factors, and is based on the fact that **prevention is often more effective than cure**. The decisions will be in **real-time**, in the sense that they will be taken under tight time constraints, and require **on-the-fly processing of Big Data**, that is, **extremely large amounts of noisy data flooding in from different geographical locations, as well as historical data**. In credit card fraud management, for example, it is necessary to forecast and act upon fraudulent activity in a matter of milliseconds, given tens of thousands of transactions per second taking place all over the world (a typical payment processing company has to process transactions coming from more than 20 thousand points-of-sale around the world), as well as several months of historical data. Moreover, data streams are often incomplete and contain erroneous data (several of the data fields could be left empty or contain incorrect information due to terminal misconfiguration).

The SPEEDD methodology for proactive event-based decision-making comprises the following steps. First, Big Data is continuously acquired and aggregated from various types of sensor. The aggregated data is analysed and fused in order to **recognise**, in real-time, events and situations of special significance. To allow for sub-second recognition, SPEEDD minimizes communication volume by moving as little data as possible from one place to another. Second, the events recognised are correlated with historical information to **forecast** problems and opportunities that may actually take place in the near future.

Third, the forecast events along with the recognised events are leveraged for real-time operational **decision-making**. Fourth, visual analytics tools prioritise and **explain** possible proactive actions, enabling human operators to reach and execute the correct decision. These tools support user-interaction in order to facilitate decision-making.

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 first year of the project, we have developed the following pieces of innovative event processing technology. First, a tool for incremental learning of event patterns has been devised. The tool uses inductive logic programming techniques to produce the event patterns, and abductive logic programming techniques to complete any missing data annotation. Moreover, the tool takes advantage of Big Data using an incremental learning technique, in which at most one pass through historical data is necessary when revising a set of event patterns.

Second, weight learning techniques for estimating the confidence values of event patterns have been implemented. These techniques operate on the probabilistic graphical models produced by Markov Logic. The output of weight learning is a set of event patterns that may be used for complex event recognition and forecasting under uncertainty.

Third, the state-of-the-art IBM Proactive Technology Online (Proton) CEP engine has been significantly extended to deal with the inherent uncertainty of Big Data. Various extensions to the event metadata, operand types, built-in functions and mathematical operations have been implemented. Consequently, Proton can readily perform robust complex event recognition in the presence of uncertainty.

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

To meet this objective, we worked on worst-case, stochastic and randomized decision-making techniques. First, the theoretical conditions to determine when an equilibrium of a traffic flow model is optimal were established. The resulting conditions considerably improve on the state-of-the-art understanding of this problem in the literature. Based on these conditions, it was shown that Alinea, a fully distributed decision-making algorithm proposed in the literature for highway traffic, leads to flow optimal equilibria. This theoretical study motivated the implementation of Alinea as the main decision-making algorithm in the first integrated SPEEDD prototype.

Second, for proactive credit card fraud management, distributionally robust stochastic programming methods were adapted to the classification of fraudulent transactions.

Third, to make randomized optimization methods applicable to large-scale and uncertain decision-making problems, like the ones encountered in SPEEDD, we extended their theoretical foundations. Techniques for relaxing the convexity assumptions, typically imposed on the decision-making problem to limit the number of samples necessary for a good decision with high confidence, were considered. The performance of randomized optimization algorithms was significantly improved for certain classes of problems, by exploiting structure in the dependence on the uncertainty. Finally, a connection between such randomized optimization methods and results in machine learning was established, leading to an overarching framework that allows us to unify the treatment of many existing randomized decision-making algorithms and develop new ones.

To produce an integrated prototype, the developed proactive decision-making algorithms were linked with the complex events recognized by the advanced event processing technology of SPEEDD.

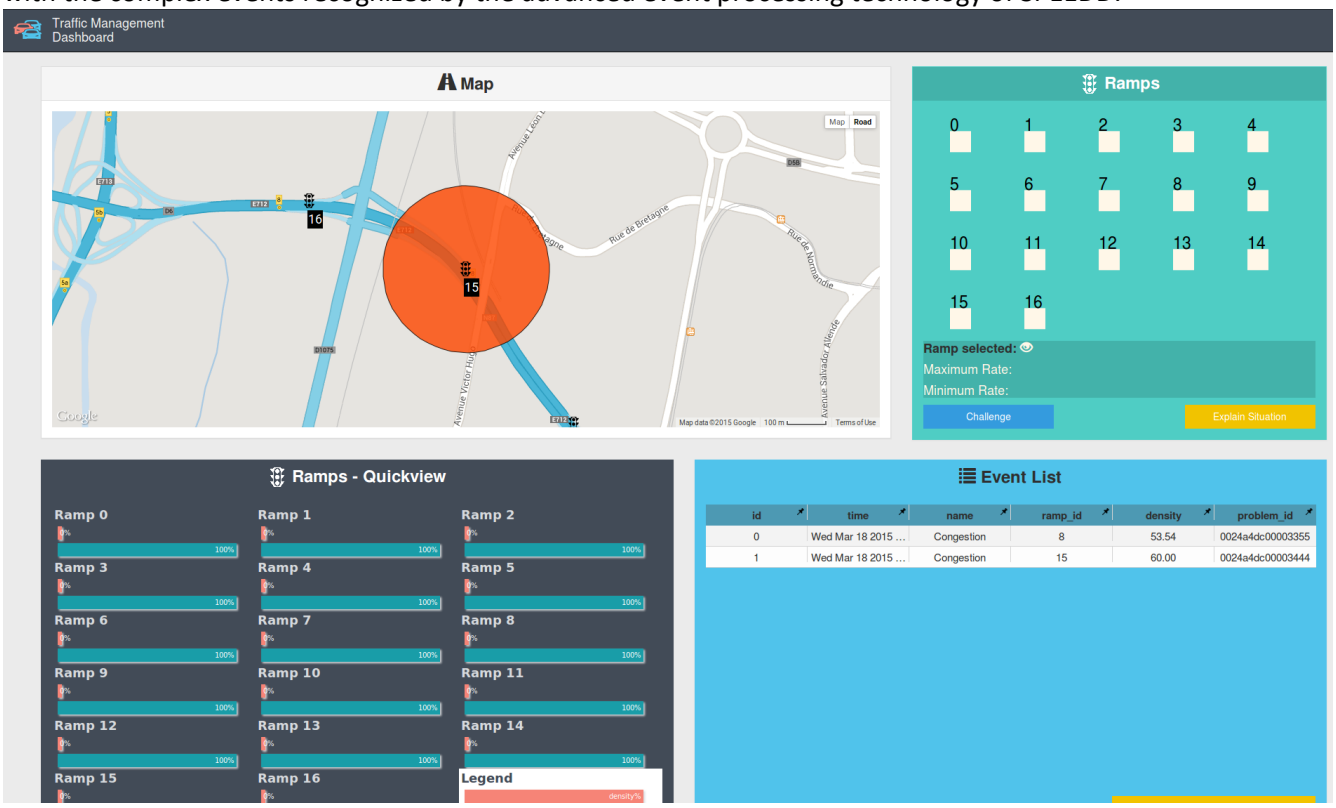

Figure 1: User Interface of the SPEEDD prototype for proactive traffic management

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. Therefore, we applied Cognitive Work Analysis to provide a systems view of decision-making in the use cases. For proactive credit card fraud management, we have conducted interviews with personnel in banks to develop insight into fraud management practice and conducted a detailed literature review to help better understand the decision making challenges. For proactive traffic management, we visited the DIR-CE in Grenoble to undertake a field study of controller activity. This involved the use of eye-tracking

equipment, the development of a Hierarchical Task Analysis of traffic management and a discussion with operators and managers as to the primary goals and objectives for traffic management.

The Hierarchical Task Analysis defines the goal and task structure for current activity. We used this to consider how the introduction of SPEEDD solutions might change these structures. Eye-tracking data (collected in the field and the laboratory) indicates strategies for information search and sampling which we relate to goal structure. Rational decision models use these data to predict optimal strategies under different goal and information conditions. We have focused on the definition of the relationship between eye-tracking and goal structure, and on defining the rational decision model.

The User Interface designs developed in the first year of the project are compatible with the user requirements. Moreover, the developed designs draw on the notion of a Model View Controller and are fully operational in the integrated SPEEEDD prototype (see Figure 1 and Figure 2).

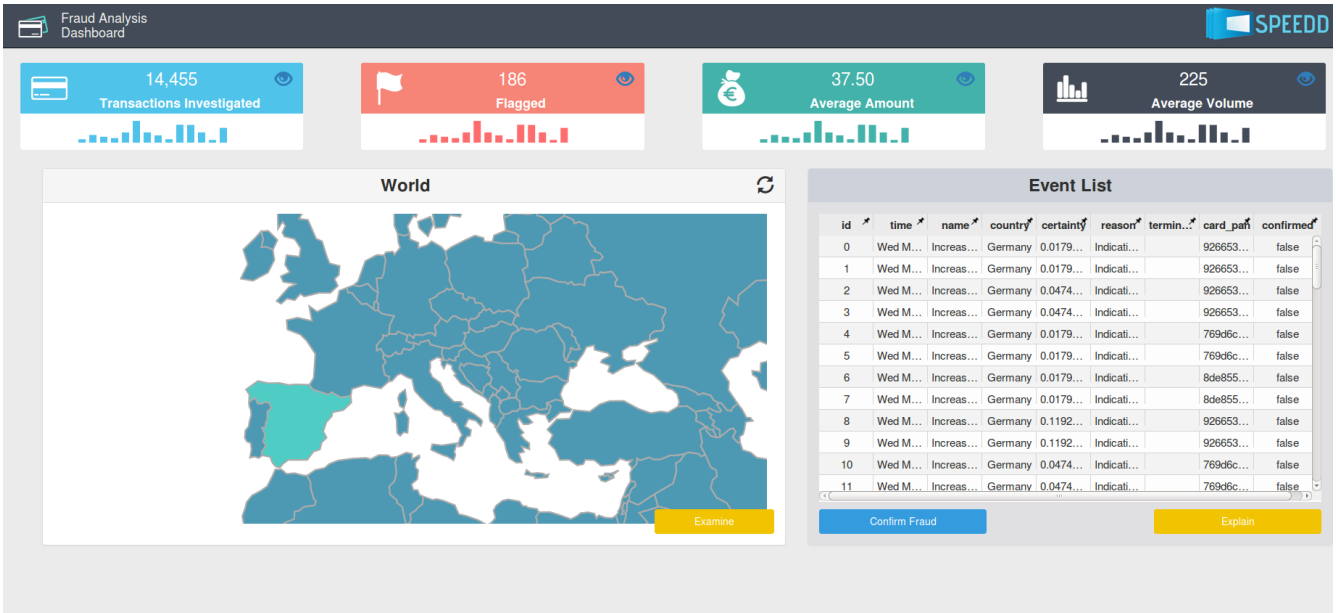

Figure 2: User Interface of the SPEEDD prototype for proactive credit card fraud management

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 proposed a chain topology for Non-Deterministic Finite State Automata (NFA) in order to detect event sequences in a lazy manner. 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. In addition, we proposed a tree topology NFA that does not require selectivity order to be defined in advance. We formally showed that this tree-structured NFA is at least as efficient as the chain-structured NFA arranged in the best performing selectivity order. Our experimental evaluation demonstrates a performance gain of two orders of magnitude, while requiring only a small fraction of the memory resources.

With respect to communication-scalable algorithms, we proposed a method to *efficiently* monitor 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. Ordinary

Least Squares regression, a well-known and very common regression model, is useful both for predicting new values given old ones, but also for understanding behavior through discovered coefficients. Our monitoring approach deals with concept drift in complex event patterns, and is efficient both in terms of communication between nodes and in terms of local computation at each node. Our experiments on the SPEEDD traffic data showed orders-of-magnitude reduction in communication.

Real-world demonstrations

The pieces of SPEEDD technology were integrated into a prototype for proactive event-driven decision-support. In order to achieve high performance, 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.

The large-scale demonstration of SPEEDD in the traffic management use case incorporates a further achievement of the project: the development of a traffic micro-simulator of the Grenoble area (see Figure 3). The simulator is essential to SPEEDD for two reasons. First, it provides synthetic data for the urban area where sensor data are unavailable. Second, it enables testing proactive decision-making by allowing us to close the loop and see in real time the effects of decisions. Such tests can only be performed in simulation, in order to avoid perturbing real traffic. The first version of the simulator includes the Grenoble South Ring and an area in the city center which is particularly crucial for traffic congestions.

Figure 3: Traffic Simulator

SPEEDD' s expected impact

SPEEDD aims at innovative businesses, which realise 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 particular, they want to enable businesses do things that they currently can't, thus reducing costs and generating further business opportunities. In the context of SPEEDD, this is translated into making sense of Big Data, through the detection of complex events, forecasting events that have not happened yet, taking these into account to proactively make decisions and providing innovative visual analytics methods to close the decision-making loop.

Towards its goals, SPEEDD integrates a number of technologies from different research areas. This multi-faceted nature of SPEEDD facilitates potential economic impact on multiple markets, ranging from business analytics, to security and finance. The common denominator in all of these is the need for real-time proactive decision-making, based on the recognition and forecasting of events on distributed Big Data.

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.

Plans for the rest of the project

In the second year of the project the SPEEDD partners aim to move full-steam towards the innovations required for achieving the project's vision: scalable and robust event recognition, forecasting and decision-making. These innovations will be realised in the second versions of the event processing, decision-making, visual analytics and data monitoring technologies. Additionally, the seamless integration of the individual technologies in the SPEEDD prototype will progress to provide the second integrated prototype. Research and development will be guided by the initial user assessment.

Additionally, we will continue the dissemination of the project's results in scientific and business conferences, as well as through our Web site and social media. So stay tuned!

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/