Bedřichov 2017

Smart Grids & Data Analysis Research

Bruno Rossi

brossi@mail.muni.cz
Department of Computer Systems and Communications,
Lasaris (Lab of Software Architectures and Information Systems)
Masaryk University, Brno

lasaris
www.lasaris.cz

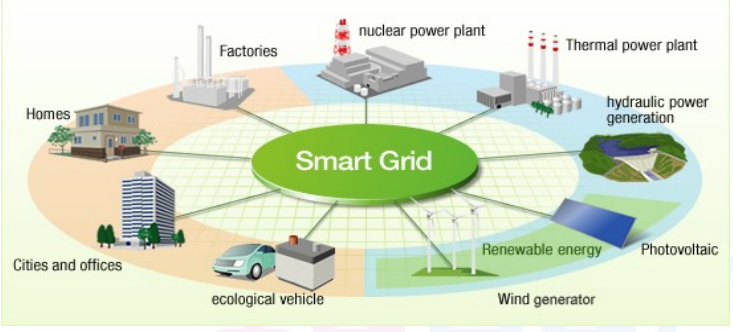
Smart Grids Research at Lasaris

Research Topics

+ Monitoring & Control
  - stream data processing
  - big data anomaly detection
+Architecture design
  - smart infrastructure design
  - quality optimization (reliability, latency, security)
+ Modelling and simulation
  - Support for Smart Grids testing/simulation/experiments
  - Conceptual models for Smart Grids

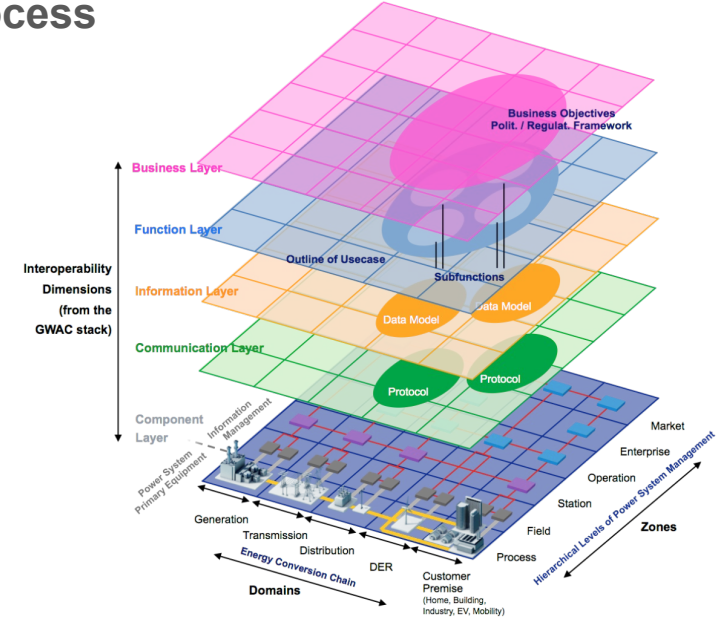
People involved

+ Tomáš Pitner
+ Bruno Rossi
+ Stanislav Chren
+ Jan Herman
+ Katarína Hrabovská
+ Martin Škop
+ Martin Schvarcbacher



Smart Grids Testing Laboratory

- The main project aim is to support a **Smart Grids Testing Laboratory** with a software infrastructure to **enhance / automate testing and simulations** of smart meters configurations and communication between smart meters and data concentrators. The final goal is to provide **processes** that can be **scientifically repeatable / replicable**, together with constant **monitoring and reporting activities** for the whole **testing process**



Smart Grids Laboratories

A Smart Grid Laboratory provides an infrastructure supporting R&D, demonstrations, experimentations, simulations & testing by considering a set of Smart Grids related use cases

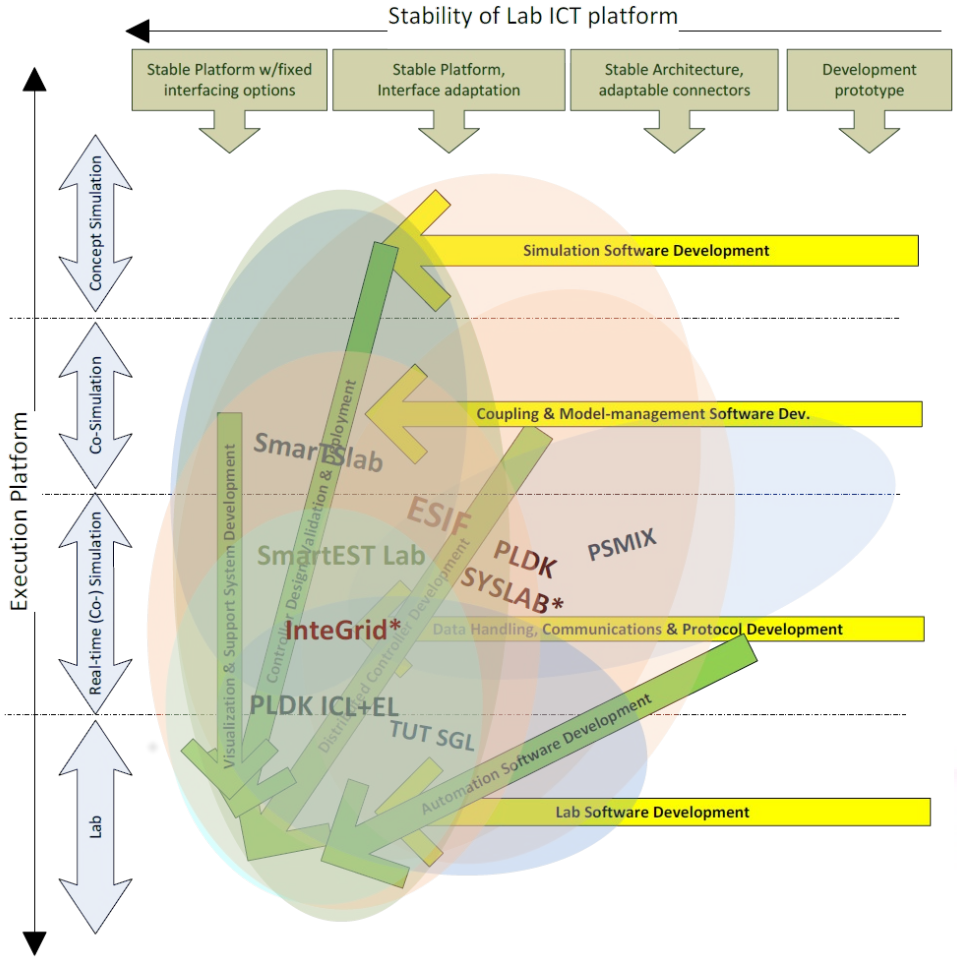
Some sample use cases / domain areas:

- Voltage stability evaluation
- Emulation of load scenarios
- Energy storage
- Monitoring and control systems validations
- Communication technologies evaluation
- Information security & privacy
- Smart Meters interoperability evaluation
- Smart Grids and home automation
- Big data analytics in Smart Grids

Smart Grids Typical Actors & Needs

- **Grid Operator**: Roll-out of new components, impact on grid operations?
- **Producer of ICT-Components/Systems**: What are the communication (bandwidth, topology, latency, …) requirements? What about monitoring and control, scheduling, ...?
- **Producer of Grid-Components**: How do I get the products “Smart Grids ready”?
- **Virtual Power Plant Operator, Energy Supplier**: new business possibilities supported by Smart Grids? Is a business case viable? Optimize revenues?
- **Policy and Regulation**: define legal and economical frameworks (e.g. regulations etc.) to reach socio-politic/economical/technical goals?

Smart Grids Laboratories



- **Emulation** (integrated or co-simulated): emulated component mimics the the real world hardware counterpart
- **Co-simulation**: orchestrate simulations running by different means
- **Real-time simulations**: the real time expectation that the simulator needs to fulfill to interact with external components (hardware or software)
- **Hardware in the loop (HIL)**: used to develop complex real-time embedded systems in which some components are real hardware, whereas others are simulated

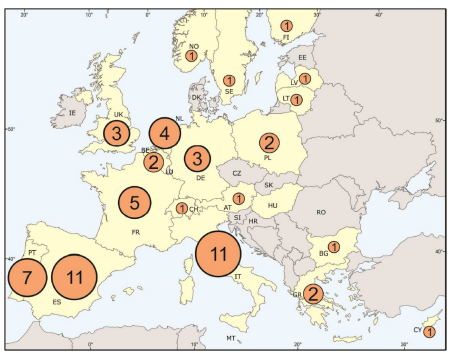
Heussen, Kai, and Oliver Gehrke. State of the Art Smart Grid Laboratories-A Survey about Software Use: RTLabOS D1. 2. Technical University of Denmark, Department of Electrical Engineering, 2014.

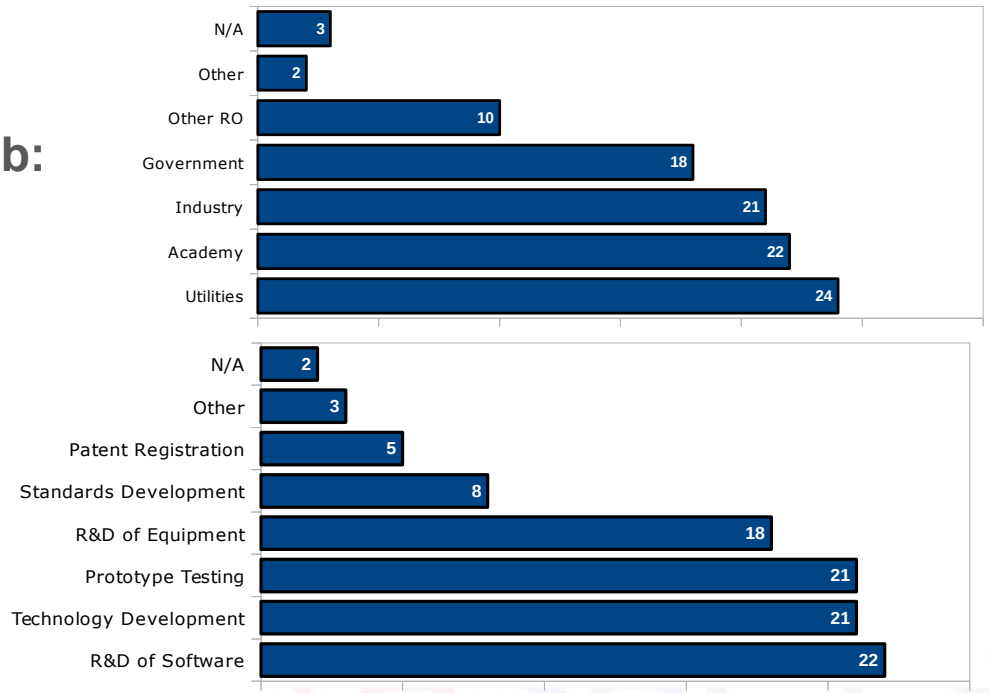
Setting-up Smart Grids Labs is expensive

- According to the Smart Grid Laboratories Inventory (2nd edition in 2016) by the EU Joint Research Centre (JRC)
- 24 Labs in 2015, 69 + 31 in 2016 (including also US)

→ **Initial budget for a standard lab: around €2M**

→ **For a large one: €30M**





Andreadou, N., Olariaga Guardiola, M., Papaioannou, I., & Prettico, G. (2016). Smart Grid Laboratories Inventory 2016. Online website: http://ses.jrc.ec.europa.eu/smart-grid-laboratories-inventory

Smart Grids Labs Main Challenges

- Multi-domain approach for analysis & validation at the system level
- Methods to analyze Smart Grids Components (e.g smart meters)
- Simulations, real world design and testing methods
- Training of researchers/engineers for cyber-physical systems

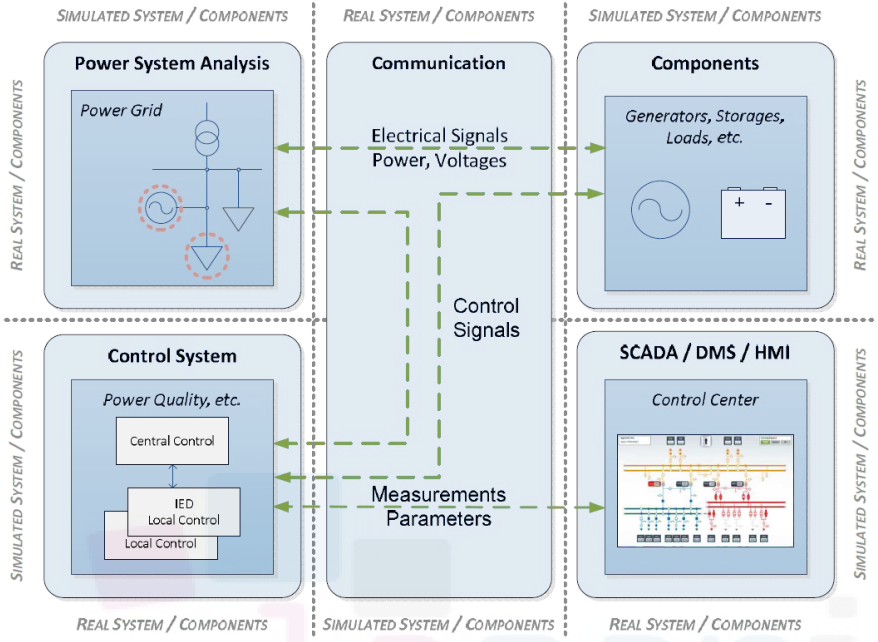
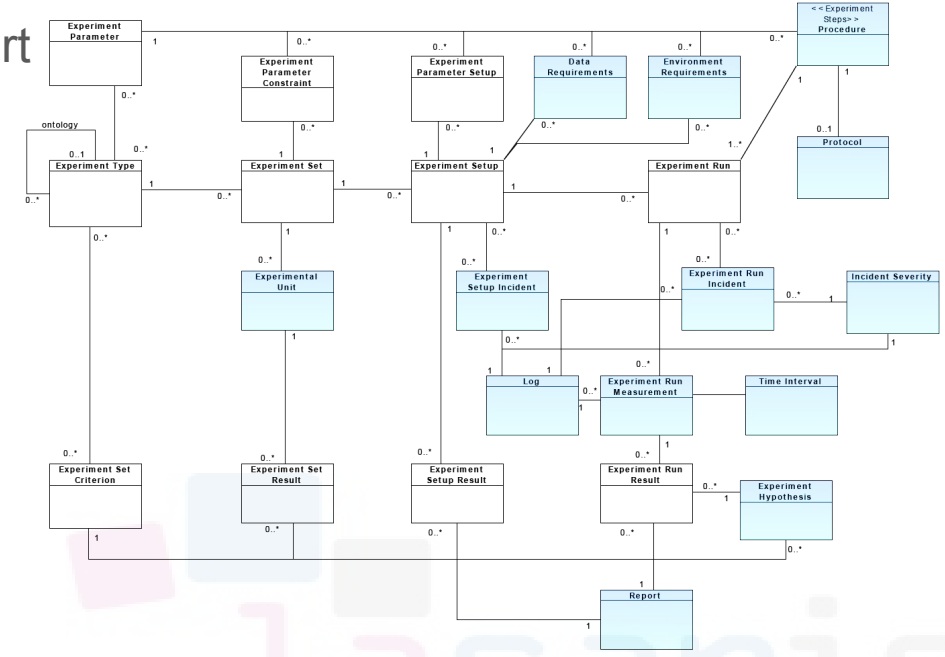


Image source: Lauss, G., et al. “Smart Grid Research Infrastructures in Austria”, Technical Report, 2015.

Main Results (1/2)

- Conceptual Models for Smart Grids Testing based on ISO/IEC/IEEE 29119 Software Testing Standard & ISO/IEC 15939:2007 Systems and software engineering - Measurement process
- Several reviews on the state of the art (e.g. Smart Grids laboratories)
- Ongoing implementation of the testing / simulation support



Main Results (2/2)

Smart Grids Co-Simulations with Low-Cost Hardware

Martin Schvarcbacher and Bruno Rossi
Faculty of Informatics
Masaryk University, Brno, Czech Republic
Email: {schvarc,brossi}@mail.muni.cz

***Abstract*—Smart Grids have nowadays gained wide diffusion and relevance. Due to the complexity of the grid, many Smart Grids laboratories have emerged over the years to provide partially virtualized environments for testing and co-simulation testbeds for the modern grid. However, the costs for setting-up Smart Grids laboratories are substantial, representing a barrier for newcomers and for educational purposes. In this paper, we propose an hardware-in-the-loop (HIL) architectural solution based on Arduino and Raspberry PI boards, supported by the Mosaik framework to simulate different Smart Grids scenarios on a small and cost-effective scale. We highlight the educational benefits that the solution can bring for understanding simulations and HIL in an affordable & effective way in an easy-to-deploy environment.**

***Keywords*-Smart Grids, Smart Meters, Hardware in the Loop, Co-Simulations, Cyber-Physical Systems.**

I. Introduction

A Smart Grid has been defined as a form of electricity network that enables "intelligent" integration of all the actions and behaviours of the connected actors, to efficiently deliver sustainable, economic and secure electricity supplies [1], [2].

While modern Smart Grids have ambitious aims, they also pose several challenges, mainly due to the multidisciplinary nature of the area, ranging from power equipment to needs in terms of data analysis to increase the "smartness" of the power grid. As such, communication between the different involved roles is fundamental, to the point that the education of students to several aspects of the grid is seen as one of the main challenges in the area [3]. For this reason, a recent trend is the emergence of Smart Grids laboratories that can serve not only to test Smart Grids software and devices, but also to educate students to the real needs of large-scale Smart Grids in a controlled environment [4].

However, average costs of setting-up a Smart Grid laboratory are in the order of €2M, reaching €30M for larger laboratories [4]. Such values represents a serious barrier for setting-up new laboratories for educational purposes. For this reason, in this paper we propose a virtualized and low-cost environment that students can use to test and validate different Smart Grids scenarios. Such environment can be a first step for looking into hardware-in-the-loop (HIL) and co-simulation environments—environments that are focused on orchestrating several simulations running on different devices, combining also software simulations [3], [5]. The importance of simulations is relevant in the area of Smart Grids due to the complexity of the different layers and sub-systems involved. Simulations can help in tackling away some of the complexity, by having simulation models that run in their own runtime environment.

The proposed solution is based on the Raspberry Pi and Arduino platforms that can be used to test a co-simulation environment for Smart Grids. We focus in particular on two scenarios that can be relevant for the presented prototype: i) sunlight level scenario, to simulate sunlight levels, and ii) a load scenario to predict power usage over time.

The paper is structured as follows. Section II presents the related works, in terms of other low-cost hardware-software solutions that have been proposed to simulate the Smart Grids infrastructure. Section III presents the details of building the prototype integrating hardware devices (Raspberry Pi and Arduino), software components (Mosaik) and the proposed architecture. Section IV proposes examples of usages with two main scenarios. Section V presents evaluation and discussion and Section VI concludes the paper.

II. Related Work

The usage of low-cost hardware/software for Smart Grids testing and validation has acquired more interest in recent years, mainly due to the availability of cheap devices that can be used for the purpose.

Commodity Hardware for Smart Grids—typically Raspberry Pis and Arduinos—has been used for a wide range of applications, for example to enable Smart Meters readings (voltage and users' power consumption) to be remotely transmitted [6], to test self-healing capabilities for multi-agent based approaches [7], or for network reconfiguration in secondary substations [8].

Aurilio et al. (2014) deployed a Raspberry Pi as data concentrator in a low-cost solution for the management and control of a power network based on power meters, monitoring connected loads by communicating with a data concentrator (Raspberry Pi) via Power Line bus [9].

However, while commodity hardware has been used in different parts of the Smart Grids infrastructure, our work is more focused in the area of co-simulations.

Armendariz et al. (2014) developed a platform for co-simulation based on a real-time power system simulator (OpalRT), and a communication network emulator (OPNET). Rasp-

Smart Grids Co-Simulations with Low-Cost Hardware

Martin Schvarcbacher and Bruno Rossi

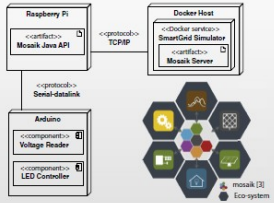
Project Context

We showcase a low-cost environment that students can use to test and validate different Smart Grids scenarios. Such environment can be a first step for looking into hardware-in-the-loop (HIL) and co-simulation environments - environments that are focused on orchestrating several simulations running on different devices, combining also software simulations. This solution can be used for understanding simulations and HIL in an affordable and effective way in an easy to deploy environment [1].

Goals

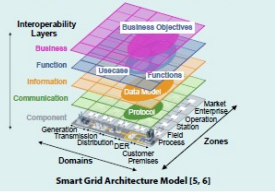
- Education of students in co-simulation concepts using easily accessible and hands-on training in Smart Grids technologies
- Creating ways for cheaper hardware prototyping of Smart Grids by having low-cost simulation nodes

Architecture



Smart Grids & Lasaris

- The Smart Grid can be regarded as an electricity network that benefits both from two-way cyber-secure communication technologies and computational intelligence for electricity generation, transmission, substations integration and consumption to reach the goals of a safe, secure, reliable, resilient, efficient, and sustainable infrastructure [4].
- Lasaris is involved in research on Smart Grids with industrial partners:
  - Supporting Smart Grids testing/simulation infrastructure
  - Data analysis for Smart Grids (load control, anomalies detection)



Smart Grid Architecture Model [5, 6]

Scenarios

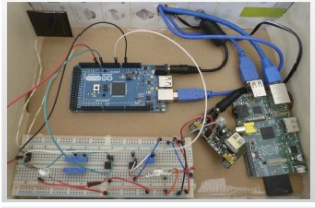
a) Sunlight Levels for a Location

- Uses past weather data to estimate sunlight levels
- Allows evaluating different PV panel deployments
- Each node represents a PV power station in the grid

b) Power Grid Load

- Knowing whether the power grid can meet the current or near future requirements becomes necessary as more intermittently available renewable resources are added to the power grid [2]
- The amount of power produced is compared to expected grid load to determine power deficiencies when using only renewable resources
- Used to determine when power plants need to be switched on to supplement renewable energy sources
- Predicting the required production capacity can be beneficial for the Smart Grid stability

Simulation Node



Project Results

- We created a platform for Smart Grid deployment prototyping
- Students can easily setup their own Smart Grid environments and test them under various changing conditions
- Test cases include: Smart Grid deployment, interoperability, stability
- Our future goal is a full power grid simulation using only commodity hardware

Hardware Node Components

- Photo-voltaic (PV) panel:
  - Produces power proportional to the illumination levels
- LED array:
  - Generates multiple illumination levels
- Arduino Mega:
  - Controls the LED array and reads the voltage level from a PV panel
  - Sends measured data and receives control commands
- Raspberry Pi:
  - Data collection and network communications

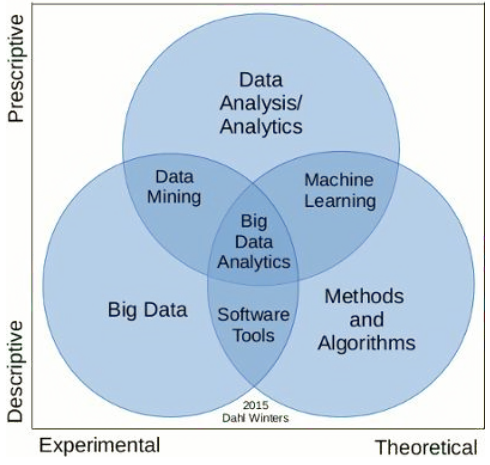
References

[1] M. Schvarcbacher and B. Rossi, "Smart Grids Co-Simulations with Low-Cost Hardware," 43rd Euromicro Conference on Software Engineering and Advanced Applications, Vienna, 2017 (to appear).
[2] S. Chren, B. Rossi and T. Pitner, "Smart grids deployments within EU projects: The role of smart meters," 2016 Smart Cities Symposium Prague (SCSP), Prague, 2016, pp. 1-5.
[3] G. Nannen et al., "Low-cost integration of hardware-components into co-simulation for future power and energy systems," IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society, Yokohama, 2015, pp. 5304-5309.
[4] B. Rossi, S. Chren, B. Buhnova and T. Pitner, "Anomaly detection in Smart Grid data: An experience report," 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, 2016, pp. 2313-2318.
[5] CEN-CENELEC-ETSI, Smart Grid Coordination Group, "Smart Grid Reference Architecture," 2012.
[6] M. Uslar et al, "Standardization in smart grids: Introduction to IT-related methodologies, architectures and standards," Springer Science & Business Media, 2012.

Big Data Group

Research Interests

+ Big Data Quality
+ Predictive Analytics in Big Data
+ Big Data & Smart Cities / Smart Grids
+ Scaling-up Machine Learning techniques
+ Visualization of spatio-temporal data



People involved

+ Barbora Bühnová
+ Tomáš Rebok
+ Mouzhi Ge
+ Bruno Rossi
+ Adam Kučera
+ Jan Herman
+ Martin Macák

Big Data Group – my interests so far

Anomaly Detection in Smart Grid Data: An Experience Report

Bruno Rossi, Stanislav Chren, Barbora Buhnova and Tomas Pitner
Faculty of Informatics
Masaryk University, Brno, Czech Republic
Email: {brossi,chren,buhnova,tomp}@mail.muni.cz

***Abstract*—In recent years, we have been witnessing profound transformation of energy distribution systems fueled by Information and Communication Technologies (ICT), towards the so called Smart Grid. However, while the Smart Grid design strategies have been studied by academia, only anecdotal guidance is provided to the industry with respect to increasing the level of grid intelligence. In this paper, we report on a successful project in assisting the industry in this way, via conducting a large anomaly-detection study on the data of one of the power distribution companies in the Czech Republic. In the study, we move away from the concept of single events identified as anomaly to the concept of collective anomaly, that is itemsets of events that may be anomalous based on their patterns of appearance. This can assist the operators of the distribution system in the transformation of their grid to a smarter grid. By analyzing Smart Meters data streams, we used frequent itemset mining and categorical clustering with clustering silhouette thresholding to detect anomalous behaviour. As the main result, we provided to stakeholders both a visual representation of the candidate anomalies and the identification of the top-10 anomalies for a subset of Smart Meters.**

***Index Terms*—Smart Grids, Smart Meters, Anomaly Detection, Clustering, Frequent Itemset Mining.**

I. Introduction

The Smart Grid can be regarded as an electricity network that benefits both from two-way cyber-secure communication technologies and computational intelligence for electricity generation, transmission, substations integration, distribution and consumption to reach the goals of a clean, safe, secure, reliable, resilient, efficient, and sustainable infrastructure [1].

The investment into large-scale Smart Grid deployment can be very risky, as confirmed for instance by investment losses during the Xcel Energys SmartGridCity project [1], [2].

There is a recent trend in adding more *"smartness"* in the Smart Grid infrastructure, so that the large amount of information that can be mined from normal usage can be used to drive the decision-making process and optimize the overall infrastructure management [3], [4]. This effect is enhanced by the two-way nature of the more modern infrastructures that allow operators to fine-tune parameters remotely based on the knowledge acquired from the operating conditions.

In the this paper, we deal with anomaly detection from Smart Grid data, that is looking for specific patterns in Smart Meter's data streams that do not conform to expected behaviour. In general terms, anomaly detection is a broad concept that has been applied to different fields, ranging from systems intrusion detection to fraud detection, with varying definitions of expected behaviour [5]. Based on real data from one of the power distribution companies in the Czech Republic, we propose an approach for the detection of the anomalies in the Smart Metering infrastructure that could be useful to promptly intervene to investigate the cause of unexpected behaviour. Based on this analysis, we report also about the insights acquired in terms of extensions of the approach that would allow us to implement such online system within the Smart Grid infrastructure.

The proposed approach is based on frequent itemset mining by encoding the different event types streamed from Smart Meters, applying segmentation of the itemsets and using categorical clustering for the evaluation of the itemsets and detection of unexpected patterns. The proposed approach is based on the analysis of event types from the Smart Meters. It allows us to detect anomalies that might have impact on the Smart Grid security, reliability or maintenance—for example suspicious manipulation with Smart Meter casing, under/over-voltage in specific locations or failure to switch remotely controlled appliances.

The paper is structured as follows. Section II overviews related work in the area of anomaly detection within Smart Grids. Section III then discusses the context of the study and provides descriptive information about the dataset. The anomaly detection approach is described in Section IV together with the rationale for its derivation. Section V presents the application within the Smart Grid domain according to the contextual information provided. The main evaluation and discussion from the experimental part is presented in Section VI, while Section VII brings up the conclusions.

II. Related Work

As the Smart Grid implementation is a strategic act for many countries, extensive attention has been paid to the study of smart infrastructures in recent years [1], [6]. Fang et al. [1] divide the smart infrastructure into three subsystems: (1) the *smart energy subsystem*, concerned with power generation, transmission, and distribution, (2) the *smart information subsystem*, concerned with information metering, measurement, and management, and (3) the *smart communication subsystem*,

978-1-5090-1897-0/16/$31.00 ©2016 IEEE

Cost-sensitive Strategies for Data Imbalance in Bug Severity Classification: Experimental Results

Nivir Kanti Singha Roy
Ericsson AB
417 56, Gothenburg, Sweden
nivir.kanti.singha.roy@ericsson.com

Bruno Rossi
Faculty of Informatics
Masaryk University, Brno, Czech Republic
brossi@mail.muni.cz

***Abstract*—*Context:* Software Bug Severity Classification can help to improve the software bug triaging process. However, severity levels present a high-level of data imbalance that needs to be taken into account. *Aim:* We investigate cost-sensitive strategies in multi-class bug severity classification to counteract data imbalance. *Method:* We transform datasets from three severity classification papers to a common format, totaling 17 projects. We test different cost sensitive strategies to penalize majority classes. We adopt a Support Vector Machine (SVM) classifier that we also compare to a baseline "majority class" classifier. *Results:* A model weighting classes based on the inverse of instance frequencies yields a statistically significant improvement (low effect size) over the standard unweighted SVM model in the assembled dataset. *Conclusions:* Data imbalance should be taken more into consideration in future severity classification research papers.**

***Keywords*-Software Bug Severity; Supervised Classification; Data Imbalance.**

I. Introduction

A bug (issue) report contains the natural language description of a problem or enhancement for a system under development, plus more structured information, such as the assignee, the reporter, expected time of resolution, severity level, and other set of features useful to characterize the issue. Depending on the development model in use, issue trackers may replace completely requirements documents, so that all requirements pass through the issue tracker. Bug reporting systems are very important in the current software development context, and their importance has risen with the application of more collaborative development methodologies for software development, for example *pull request* mechanisms that are supported by modern issue trackers.

It is not surprising that issue trackers constitute a central point of focus in current software engineering empirical research (e.g., [1], [2]). In this paper, we deal with the classification of severity of bug reports—that is providing evaluation of models to classify issues into severity levels, based on different set of features considered. The whole area derived from the research from Menzies and Marcus published in 2008, when a model based on a rule learner supported by entropy and information gain was used to classify severity levels for NASA projects [3].

There is, however, one pitfall in the classification of bug severity (e.g., [4]): imbalance across different severity levels can bring issues for correctly classifying new instances. It is not unusual in the area to consider only some of the classes available, or to report very different classification performance results for the different classes [4]. Furthermore, data imbalance has already been found to impact the results of software defect prediction performance, and needs to be taken into account [5].

In the current paper, we are looking at ways in which we can improve classification results based on such initial unbalanced situations. The contribution of the current paper is twofold:

- provide an aggregated dataset from some of the most relevant previous papers. Namely, we converted datasets in [1], [6], [3] to a common format helpful for uniform classification and future comparison[1];
- evaluation of different cost-sensitive strategies to hamper data imbalance;

The paper is structured as follows: Section II provides the related research. Section III discusses the proposed empirical approach to deal with data imbalance. Section IV presents the data analysis results and Section V provides the conclusions.

II. Related Research

Severity Classification. Over the years, a large amount of empirical knowledge has been derived in the area of severity classification. What we know so far in the area is that, in general, the Multimodal Naïve Bayes (MNB) and Support Vector Machines (SVM) classification models outperform the Naïve Bayes (NB) model and the k-Nearest Neighbour (k-NN) classification models in most of the studies (e.g. [7]). Most of the papers consider NASA PITS data from PROMISE repository, as it was used in the seminal paper from from Menzies and Marcus (2008) [3]. Among open source projects, majority of articles use Firefox and Mozilla data (see [7], [6]).

The models are generally based on textual analysis of the issues (e.g., [1]), however it is recent evolution to consider more semantically-aware models [4], [8]—suggesting to use topic models to reduce the number of features for the classifier to improve the classification performance. However, there are also studies that consider other aspects for severity, for example stack traces, reports length, attachments, and steps to reproduce [9].

[1]dataset is available at www.unlimited2.com/pages/SEAA2017.html

Big Data Group – my interests so far

Anomaly Detection in Smart Grid Data: An Experience Report

Bruno Rossi, Stanislav Chren, Barbora Buhnova and Tomas Pitner
Faculty of Informatics
Masaryk University, Brno, Czech Republic
Email: {brossi,chren,buhnova,tomp}@mail.muni.cz

***Abstract*—In recent years, we have been witnessing profound transformation of energy distribution systems fueled by Information and Communication Technologies (ICT), towards the so called Smart Grid. However, while the Smart Grid design strategies have been studied by academia, only anecdotal guidance is provided to the industry with respect to increasing the level of grid intelligence. In this paper, we report on a successful project in assisting the industry in this way, via conducting a large anomaly-detection study on the data of one of the power distribution companies in the Czech Republic. In the study, we move away from the concept of single events identified as anomaly to the concept of collective anomaly, that is itemsets of events that may be anomalous based on their patterns of appearance. This can assist the operators of the distribution system in the transformation of their grid to a smarter grid. By analyzing Smart Meters data streams, we used frequent itemset mining and categorical clustering with clustering silhouette thresholding to detect anomalous behaviour. As the main result, we provided to stakeholders both a visual representation of the candidate anomalies and the identification of the top-10 anomalies for a subset of Smart Meters.**

***Index Terms*—Smart Grids, Smart Meters, Anomaly Detection, Clustering, Frequent Itemset Mining.**

that has been applied to different fields, ranging from systems intrusion detection to fraud detection, with varying definitions of expected behaviour [5]. Based on real data from one of the power distribution companies in the Czech Republic, we propose an approach for the detection of the anomalies in the Smart Metering infrastructure that could be useful to promptly intervene to investigate the cause of unexpected behaviour. Based on this analysis, we report also about the insights acquired in terms of extensions of the approach that would allow us to implement such online system within the Smart Grid infrastructure.

The proposed approach is based on frequent itemset mining by encoding the different event types streamed from Smart Meters, applying segmentation of the itemsets and using categorical clustering for the evaluation of the itemsets and detection of unexpected patterns. The proposed approach is based on the analysis of event types from the S… ers. It allows us to detect anomal… the Smart G… ble

…ection III then discusses the context of the study and provides descriptive information about the dataset. The anomaly detection approach is described in Section IV together with the rationale for its derivation. Section V presents the application within the Smart Grid domain according to the contextual information provided. The main evaluation and discussion from the experimental part is presented in Section VI, while Section VII brings up the conclusions.

Future work: scaling-up (online models, alternative algorithms)

…ent, and sustainable infrastructure [1].

The investment into large-scale Smart Grid deployment can be very risky, as confirmed for instance by investment losses during the Xcel Energys SmartGridCity project [1], [2].

There is a recent trend in adding more *"smartness"* in the Smart Grid infrastructure, so that the large amount of information that can be mined from normal usage can be used to drive the decision-making process and optimize the overall infrastructure management [3], [4]. This effect is enhanced by the two-way nature of the more modern infrastructures that allow operators to fine-tune parameters remotely based on the knowledge acquired from the operating conditions.

In the this paper, we deal with anomaly detection from Smart Grid data, that is looking for specific patterns in Smart Meter's data streams that do not conform to expected behaviour. In general terms, anomaly detection is a broad concept

II. Related Work

As the Smart Grid implementation is a strategic act for many countries, extensive attention has been paid to the study of smart infrastructures in recent years [1], [6]. Fang et al. [1] divide the smart infrastructure into three subsystems: (1) the *smart energy subsystem*, concerned with power generation, transmission, and distribution, (2) the *smart information subsystem*, concerned with information metering, measurement, and management, and (3) the *smart communication subsystem*,

978-1-5090-1897-0/16/$31.00 ©2016 IEEE

Cost-sensitive Strategies for Data Imbalance in Bug Severity Classification: Experimental Results

Nivir Kanti Singha Roy
Ericsson AB
417 56, Gothenburg, Sweden
nivir.kanti.singha.roy@ericsson.com

Bruno Rossi
Faculty of Informatics
Masaryk University, Brno, Czech Republic
brossi@mail.muni.cz

***Abstract*—*Context:* Software Bug Severity Classification can help to improve the software bug triaging process. However, severity levels present a high-level of data imbalance that needs to be taken into account. *Aim:* We investigate cost-sensitive strategies in multi-class bug severity classification to counteract data imbalance. *Method:* We transform datasets from three severity classification papers to a common format, totaling 17 projects. We test different cost sensitive strategies to penalize majority classes. We adopt a Support Vector Machine (SVM) classifier that we also compare to a baseline "majority class" classifier. *Results:* A model weighting classes based on the inverse of instance frequencies yields a statistically significant improvement (low effect size) over the standard unweighted SVM model in the assembled dataset. *Conclusions:* Data imbalance should be taken more into consideration in future severity classification research papers.**

***Keywords*-Software Bug Severity; Supervised Classification; Data Imbalance.**

I. Introduction

A bug (issue) report contains the natural language description of a problem or enhancement for a system under development, plus more structured information, such as the assignee, the reporter, expected time of resolution, … level, and other set of features useful … Depending on …

…gies for software …pported by modern issue trackers.

It is not surprising that issue trackers constitute a central point of focus in current software engineering empirical research (e.g., [1], [2]). In this paper, we deal with the classification of severity of bug reports—that is providing evaluation of models to classify issues into severity levels, based on different set of features considered. The whole area derived from the research from Menzies and Marcus published in 2008, when a model based on a rule learner supported by entropy and information gain was used to classify severity levels for NASA projects [3].

There is, however, one pitfall in the classification of bug severity (e.g., [4]): imbalance across different severity levels can bring issues for correctly classifying new instances. It is not unusual in the area to consider only some of the classes available, or to report very different classification performance results for the different classes [4]. Furthermore, data imbalance has already been found to impact the results of software defect prediction performance, and needs to be taken into account [5].

In the current paper, we are looking at ways in which we can improve classification results based on such initial unbalanced situations. The contribution of the current paper is twofold:

- provide an aggregated dataset from some of the most relevant previous papers. Namely, we converted datasets in [1], [6], [3] to a common format helpful for uniform classification and future comparison[1];
- evaluation of different cost-sensitive strategies to hamper data imbalance;

The paper is structured as follows: Section II provides the related research. Section III discusses the proposed empirical approach to deal with data imbalance. Sect… data analysis results …

Future work: comparison of large number of cost-sensitive strategies

…the area is that, in …ynomial Naïve Bayes (MNB) and Support …vector Machines (SVM) classification models outperform the Naïve Bayes (NB) model and the k-Nearest Neighbour (k-NN) classification models in most of the studies (e.g. [7]). Most of the papers consider NASA PITS data from PROMISE repository, as it was used in the seminal paper from from Menzies and Marcus (2008) [3]. Among open source projects, majority of articles use Firefox and Mozilla data (see [7], [6]).

The models are generally based on textual analysis of the issues (e.g., [1]), however it is recent evolution to consider more semantically-aware models [4], [8]—suggesting to use topic models to reduce the number of features for the classifier to improve the classification performance. However, there are also studies that consider other aspects for severity, for example stack traces, reports length, attachments, and steps to reproduce [9].

[1]dataset is available at www.unlimited2.com/pages/SEAA2017.html

Future Work

- Publish work done on the Smart Grids area & continue research
  - Testing models for Smart Grids
  - Review of existing laboratories
  - Co-simulation implementation supported by a testing framework
- Consolidate the Big Data group → find direction after the initial exploratory phase

DSD/SEAA 2018 in Prague

http://dsd-seaa2018.fit.cvut.cz

About Euromicro DSD/SEAA

Established in 1973, Euromicro is an international scientific, engineering and educational organization dedicated to advancing the arts, sciences and applications of Information Technology and Microelectronics.

The Euromicro **conferences and journals are known worldwide** for their scientific quality and are a live testimony of how the commitment **established 44 years ago** is being fulfilled.

Prague, the capital of the Czech Republic, is **one of the most attractive and** most **visited cities in Europe**.

Call For Papers

DSD 2018

Euromicro Conference on Digital System Design

- Paper Submission Deadline: 1st April 2018
- Notification of Acceptance: 15th May 2018
- Camera-Ready Papers: 15th June 2018

SEAA 2018

Euromicro Conference on Software Engineering and Advanced Applications

- Paper Submission Deadline: 1st March 2018
- Notification of Acceptance: 15th May 2018
- Camera-Ready Papers: 15th June 2018