Serums

Project no. 826278

SERUMS

Research & Innovation Action (RIA)
SECURING MEDICAL DATA IN SMART-PATIENT HEALTHCARE SYSTEMS

Report on Final Data Masking, Data Fabrication and Semantic-Preserving Encryption D4.3

Due date of deliverable: 31st March 2022

Start date of project: 1st January 2019

Type: Deliverable
WP number: WP4

Responsible Institution: IBM
Editor and editor's address: Michael Vinov (vinov@il.ibm.com)
Partners Contributing: UCL, USTAN, SCCH, ZMC, UCY, FCRB

Approved by:

Version 1.0

Project co-founded by the European Commission within the Horizon H2020 Programme		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Release History

Release No.	Date	Author(s)	Release Description/Changes made
V0.1	14/03/23	Michael Vinov (IBM)	Initial draft
V0.2	29/03/2022	Mohit Kumar (SCCH)	Updated Section 4
V0.3	29/03/2022	Eduard Baranov (UCL)	Updated Section 3
V0.4	31/03/2022	Michael Vinov (IBM)	Review ready version
V1.0	03/04/2022	Michael Vinov	Submission ready version

SERUMS Consortium

Partner 1	**University of St Andrews**
Contact Person	Name: Juliana Bowles Email: jkfb@st-andrews.ac.uk
Partner 2	**Zuyderland Medisch Centrum**
Contact Person	Name: Cindy Wings Email: c.wings@zuyderland.nl
Partner 3	**Accenture B.V.**
Contact Person	Name: Bram Elshof, Wanting Huang Email: bram.elshof@accenture.com, wanting.huang@accenture.com
Partner 4	**IBM Israel Science & Technology Ltd.**
Contact Person	Name: Michael Vinov Email: vinov@il.ibm.com
Partner 5	**Sopra-Steria**
Contact Person	Name: Andre Vermeulen Email: andreas.vermeulen@soprasteria.com
Partner 6	**Université Catholique de Louvain**
Contact Person	Name: Axel Legay Email: axel.legay@uclouvian.be
Partner 7	**Software Competence Centre Hagenberg**
Contact Person	Name: Michael Rossbory Email: michael.rossbory@scch.at
Partner 8	**University of Cyprus**
Contact Person	Andreas Pitsillides Email: andreas.pitsillides@ucy.ac.cy
Partner 9	**Fundació Clínic per a la Recerca Biomèdica**
Contact Person	Name: Santiago Iriso Email: siriso@clinic.cat
Partner 10	**University of Dundee**
Contact Person	Name: Vladimir Janjic Email: vjanjic001@dundee.ac.uk

Table of Contents

Executive Summary

Securing Medical Data in Smart Patient-Centric Healthcare Systems (SERUMS) is a research project supported by the European Commission (EC) under the Horizon 2020 program. This document is the third final deliverable of Work Package 4: “Secure and Privacy-Preserving Data Communication”. The leader of this work package is IBM, with involvement from the following partners: UCL, USTAN, SCCH, ZMC, UCY and FCRB. The goal of this work package is to explore and develop techniques and mechanisms to ensure the security and protection of the personal medical data that is shared as part of a coherent smart healthcare system. The objectives of WP4 are to:

- develop advanced data masking and synthetic data fabrication technologies to enable sharing of personal medical data between components of the Smart Health Centre system developed in WP6;
- develop metrics and techniques to verify both the security and the functional properties of the advanced data analytics and the Serums patient-centric Smart Health Centre system;
- explore and develop technology for encrypting information while preserving certain required semantics, in order to enable advanced data analytics while adhering to privacy regulations.

This deliverable entitled “Report on Final Data Masking, Data Fabrication and Semantic-Preserving Encryption” is the third and final deliverable of the WP4. It describes enhanced versions of the data masking and data fabrication technologies that are used in the project to enable sharing of personal healthcare data between the project partners and development of the Smart Health Centre. The deliverable report also describes an enhanced version of the technology to verify the quality of fabricated synthetic data on final versions of the data analytics and authentication tools and an enhanced version of the semantic-preserving data encryption technology to enable and facilitate the application of the Serums advanced data analytics on personal medical data, while fully adhering to necessary privacy regulations.

1 Introduction

1.1 Role of the Deliverable

The aim of this deliverable is to report and describe the design and development of enhanced versions of the data masking, data fabrication, data quality verification and semantic-preserving data encryption technologies. All these technologies are used to explore and develop techniques and mechanisms to ensure the security and protection of the personal medical data that is shared as part of a coherent smart health-care system and to enable and facilitate the application of the Serums advanced data analytics on personal medical data, while fully adhering to necessary privacy regulations.

1.2 Relationship to Other SERUMS Deliverables

Tasks 4.1 and 4.2 of WP4 are closely related to the work done in WP2 – "Smart Patient Record Construction". Masked data and synthetic fabricated data of WP4 is formatted based on the Smart Patient Record format definition developed in WP2. T4.3 of WP4 is closely related to WP2 and WP5. The data technology developed in T4.3 will be applied to verify quality of fabricated medical data and its usage for data analytics and authentication tools of WP2 and WP5. In addition, the output of T4.4 will be used by the WP2 of the project.

1.3 Structure of this Document

This document is structured as follows: *Chapter 2* describes the enhanced version of IBM's Data Fabrication Technology and its usage for fabrication of the project synthetic medical data. *Chapter 3* describes the methodology that is used for verification of the fabricated data quality and its usage for development and testing of the project advanced data analytics and user authentication tools. *Chapter 4* provides a description of the semantic- and privacy-preserving encryption methodology that is used to enable and facilitate the application of the project advanced data analytics on personal medical data, while fully adhering to necessary privacy regulations. *Chapter 5* concludes the deliverable.

2 Data Masking and Synthetic Data Fabrication

2.1 Introduction to DFP

IBM's Data Fabrication Platform (DFP) [1][2] is a web based central platform for generating high-quality data for testing, development, and training. The platform provides a consistent and organizational wide methodology for creating test data. The methodology used is termed "Rule Guided Fabrication".
The primary DFP use case for fabricating synthetic data contains two actors: a user (initiator) and Database/File (participator). This use case includes two sub-use cases: data requirements modelling and data generation. The data requirements use case includes three sub-use cases: resources and structure definitions, constraint rules definitions and fabrication configuration definitions. The data structure for databases (schema, tables, columns, etc.) is automatically imported, however structural hierarchy of data elements (structs, arrays, tables, fields, types) need to be manually defined by the user. The constraint rules are required to construct a model of the data and thus enable creation of meaningful realistic data vales. Input and output resources are standard relational databases (e.g., DB2, Oracle, PostgreSQL, SQLite) and standard file formats (e.g., Flat file, XLS, CSV, XML, JSON).

More detailed description of the DFP tool is available in the D4.1 and D4.2 documents of the project.

2.2 DFP Enhancements

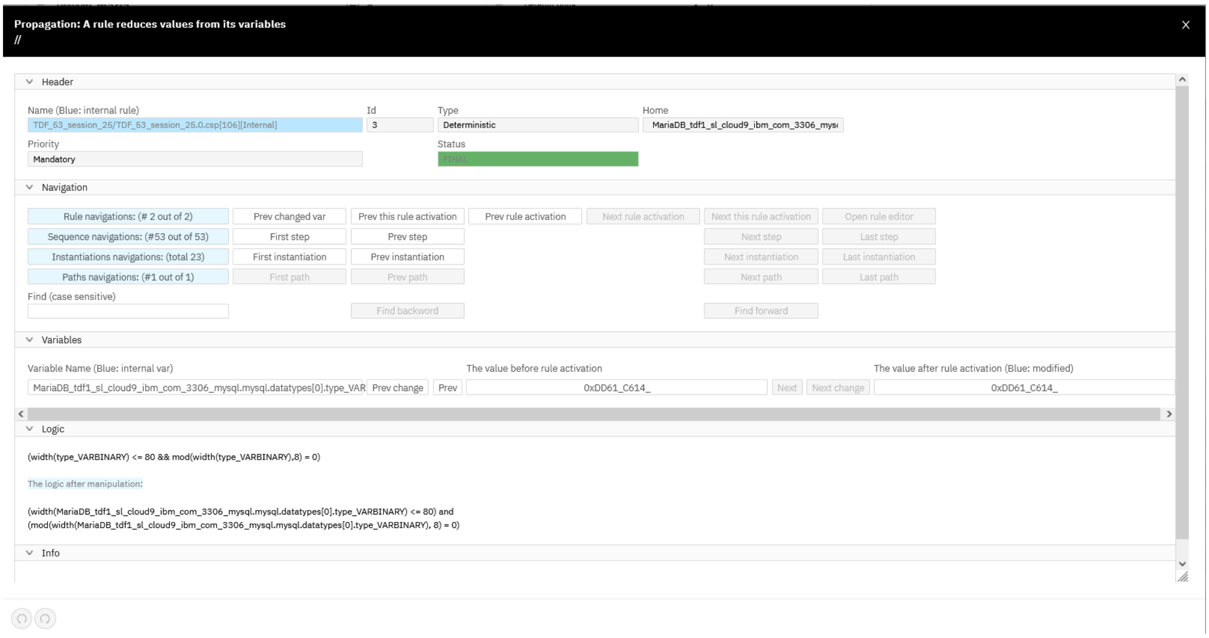
During the third year of the SERUMS project the Data Fabrication Platform technology has been significantly enhanced to enable improved user experience and fabrication of more complex synthetic data. The major enhancements of the tool include:

1. Failure analysis tool (a debugger)
2. New rule operators
3. Support for multi-ratio rules
4. Fabrication statistics dashboard
5. Support for SSL connection to Database servers
6. Usability enhancements

Below is a more detailed description of the above improvements.

Failure analysis tool (a debugger)

The aim of the new debugger is to enable a user to easily find a potential data modelling problem. It enables to run the internal Constraints Satisfaction Problem (CSP) solver step-by-step and thus figure out an exact order of the fabrication rules invocation and data variable values reduction. Moreover, the debugger enables to navigate the CSP solver forward and backwards. All this helps a user to easily find which one out of tens or hundreds of defined fabrication rules is responsible for potentially illegal variable value or fabrication failure. Below is a sample GUI window of the debugger.



Propagation: A rule reduces values from its variables
//

Header

Name (Blue: internal rule): TDF_53_session_25/TDF_53_session_25.0.csp[106][Internal]
Id: 3
Type: Deterministic
Home: MariaDB_tdf1_sl_cloud9_ibm_com_3306_mys
Priority: Mandatory
Status: FINAL

Navigation

Find (case sensitive)

Variables

Variable Name (Blue: internal var): MariaDB_tdf1_sl_cloud9_ibm_com_3306_mysql.mysql.datatypes[0].type_VAR
The value before rule activation: 0xDD61_C614_
The value after rule activation (Blue: modified): 0xDD61_C614_

Logic

```
(width(type_VARBINARY) <= 80 && mod(width(type_VARBINARY),8) = 0)

The logic after manipulation:

(width(MariaDB_tdf1_sl_cloud9_ibm_com_3306_mysql.mysql.datatypes[0].type_VARBINARY) <= 80) and
(mod(width(MariaDB_tdf1_sl_cloud9_ibm_com_3306_mysql.mysql.datatypes[0].type_VARBINARY), 8) = 0)
```

Info

New rule operators

The following new operators have been added to the fabrication rules modelling language:

- Padding
  The *pad* operator adds spaces to the right of the first parameter to have in total the number of characters specified by the second parameter. The number of bytes should be as specified in the optional third parameter.

- Trim
  The *trim* operator returns a string without its heading and tailing spaces and tabs.

- Correlation
  The *correlation* operator defines a statistical relationship between random variables. It commonly refers to the degree to which a pair of variables are linearly related. The first three parameters define the first variable, its mean and its standard deviation, the next three parameters define these properties to the second variable. The last parameter is the correlation coefficient.

- New Distribution types (Exponential, Beta, Gamma, Rayleigh, Log, Norm, Weibull, Poisson)

Support for multi-ratio rules

Based on specific requirement of the USTAN use case we've added a more sophisticated support for multi-ratio fabrication rules. More specifically a user can define complex definition of ratios between parent-child tables. For example, in the USTAN used case the new rules have been defined based on the new support:

- For each Patient there should be several Intentions created

- For each Intention there should be several Regimes
- For each Regime there should be several Cycle (based on some predefined distribution)

Fabrication statistics dashboard

The new dashboard provides a user information about the fabrication session progress at run-time. It includes information about a number of solved problems, number of fabrication retries, the fabrication task complexity (estimated), number of parallel solvers involved, number of active database connections, number of fabricated records per minute, estimated fabrication completion time.

Support for SSL connection to Database servers

This new feature enables the TDF tool to connect and update Database servers through the secure SSL connection.

2.3 Data Masking

Data masking is a well-known method of creating a structurally similar but inauthentic version of an organization's data that can be used for purposes such as software testing, software development and user training. The purpose is to protect the actual personal or sensitive data while having a functional substitute for occasions when the real data is not required. In data masking, the format of data remains the same, only the values are changed. The data may be altered in several ways, including encryption, character shuffling, and character or word substitution.

It was the SERUMS consortium decision that most of the data used for the development and testing of the SERUMS data analytics, user authentication technologies and its patient-centric healthcare system will be synthetic data fabricated by IBM’s Data Fabrication Technology described in Section 2.1 above. Moreover, usage of synthetic realistic data solves a known weakness of the data masking approach – its reversibility and a need for the real data access. In case synthetic fabricated data will not be sufficient or “good enough” for the development and testing requirements of the project, we will consider applying the same IBM’s DFP tool to produce masked data from the project use-cases real data.

3 Verification of Fabricated Data Quality

Testing and verification activities in SERUMS require input data. The utilization of real patients' data has high privacy risks. On the other hand, fabricated data can be used without any restrictions: it does not contain any information related to any person or organization. The use of fabricated data imposes requirements on its quality. Fabricated data shall be "realistic" for testing and verification results being practical. For example, machine learning models cannot be trained on pure random data, data patterns and hidden intrinsic dependencies existing in real data must be present in fabricated data.

To address this challenge, we have proposed and implemented an approach based on ML as reported in the deliverables D4.1 and D4.2. The core idea of the approach is to use machine learning algorithms to attempt to distinguish real and fabricated data. Given two datasets, real and fabricated, if at least one of the used algorithms succeeds, i.e. its classification accuracy is high, then the data has some dependencies intrinsic to only one dataset. This indicates that the fabrication ruleset needs to be updated. The diagram below illustrates the overall process.

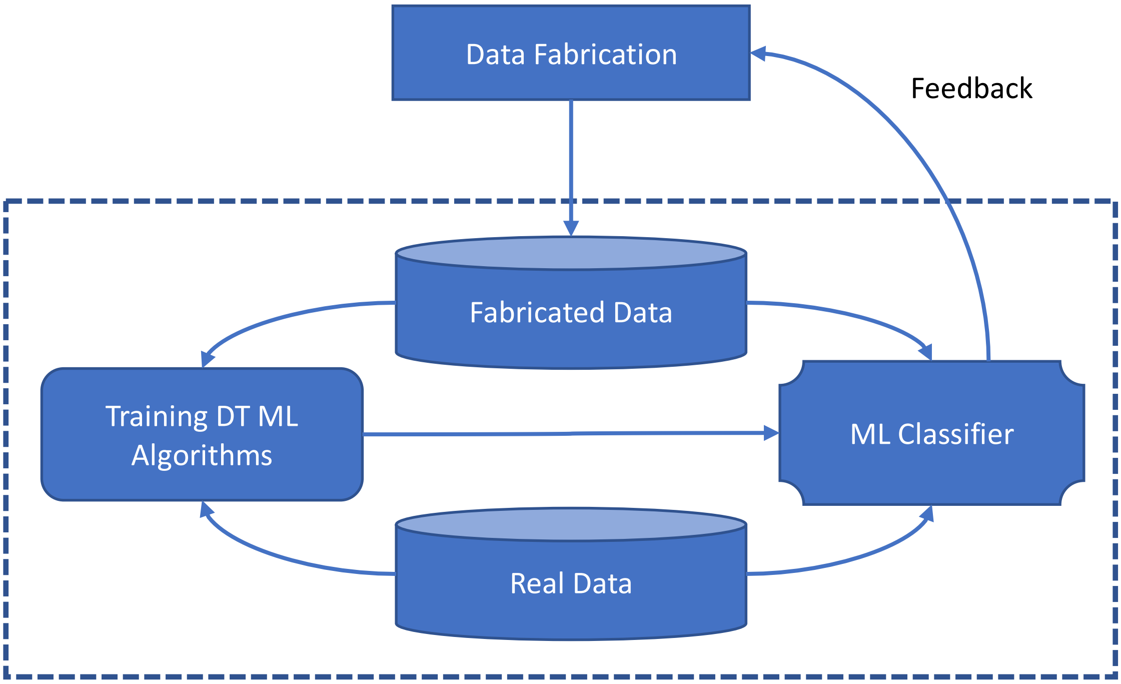


Figure 1ML-Distinguisher diagram

At the first step of the approach, the initial fabrication ruleset is used to generate a fabricated dataset. The real and fabricated datasets are then given to the ML algorithm that trains a classifier. We use Decision-Tree (DT) based ML algorithms to facilitate the improvement of the fabrication ruleset. The classifier attempts to distinguish real and fabricated data and its accuracy is recorded. High accuracy indicates the presence of features allowing the classifier to distinguish the data. DT based algorithms provide the information on such features and range them according to their significance. Exploration of the significant features provides the indication on fabrication rules requiring modification. After the update of the fabrication ruleset, the process is repeated until the classification accuracy becomes low.

We have implemented the approach in a tool called ML-Distinguisher with the following DT based ML algorithms: AdaBoost, Gini, Entropy, and Random Forest. All of them provide the details of the classification decision process and facilitate the search for the rules requiring modification.

The latest enhancements of the ML-Distinguisher include several quality-of-life updates as well as taking sequential data into consideration. The updates include semi-automation of the data preparation for the ML training and an option to ignore specified columns in the data (e.g. timestamps that could be different in real and fabricated data which shall not be a decisive feature for the classification). For the sequential data, we consider the following types of sequences: 1) multiple columns in data entries forming a sequence (e.g. patient's pulse being measured multiple times at the beginning of a visit and after some exercises) 2) multiple data entries with the same value in a specified column (e.g. a sequence of patient’s visits detected by the same PatientID). We added an additional preprocessing mechanism to find such sequences and to add the findings to the training dataset.

The ML-Distinguisher has been used during the ruleset creation for datasets generation. It has been run in the local private environment of the partners ensuring privacy and isolation of the real medical data.

The results of USTAN data fabrication have been reported in D4.2; here we recap them briefly. USTAN used ML-Distinguisher for the fabrication of data for 4 tables: Demographics, Diagnosis, smr01s, smr06s. The results for the initial fabrication ruleset are shown in Table 1.

Table 1. USTAN results after Iteration 1

Table	Gini	Entropy	AdaBoost	Random Forest
Demographics	0.8109	0.8098	0.8301	0.8252
Diagnosis	0.9020	0.8997	0.9038	0.8982
smr01s	0.9846	0.9869	0.9887	0.9883
smr06s	0.9313	0.9346	0.9476	0.9356

The feedback of the ML-Distinguisher has been used to update the fabrication ruleset. After 3 iterations the accuracy of the classification has dropped as shown in Table 2.

Table 2. USTAN results after Iteration 3

Table	Gini	Entropy	AdaBoost	Random Forest
Demographics	0.6306	0.6295	0.5958	0.6313
Diagnosis	0.5951	0.5948	0.5826	0.6811
smr01s	0.7164	0.7204	0.7227	0.7432
smr06s	0.8483	0.8333	0.803	0.8618

ZMC created a set of rules for the fabrication of data from wearable devices. The validation of the initial ruleset showed the classification accuracy of >0.99. Based on the feedback of the ML-Distinguisher, the ruleset has been updated. This resulted in a reduction to ~0.89. Further inspection in the most significant features allowed ZMC to find the difference in the datasets creation. Due to manufacturing issues, the majority of real sensors used to create the real dataset had a fixed amount of time they were able to measure. This had the effect that the duration of each measurement could be grouped based on the battery life of sensors. Consequently, it affected other aspects of the resulted dataset, such as total percentage of active time during the day and amount of steps taken. The ruleset for data fabrication however was designed with a normal distribution in mind without taking the above mentioned grouping into account. After this discovery, ZMC re-collected the real dataset with the grouping issue removed and adjusted the ruleset. The latest accuracy has been reduced to 0.83. The accuracy results are combined in Table 3.

Table 3. ZMC results

Data	Initial accuracy	Iteration 1 accuracy	Iteration 2 accuracy
WearableDevices	0.99	0.89	0.83

FCRB is using the tool for their ruleset, the results are not yet available.

4 Semantic-preserving Data Encryption

The emergence of cloud infrastructure not only raises the concern of protecting data in storage, but also requires an ability of performing computations on data while preserving the data privacy. Fully homomorphic encryption (FHE) being capable of directly performing an unbounded number of operations on encrypted data forms a solution to the privacy concerns in the cloud computing scenario. For a FHE scheme, the bootstrapping procedure is introduced to reduce the noise contained in a ciphertext for allowing arbitrary computations. The bootstrapping operation is performed on a ciphertext via evaluating the decryption function homomorphically using the bootstrapping key (which is the encryption of the private decryption key under the public encryption key). Bootstrapping is the computationally most expensive part of a homomorphic encryption scheme. Despite the recent advances in the bootstrapping procedure, the bootstrapped bit operations are still several times slower than their plaintext equivalents. Thus for an efficient secure machine learning scenario in practice, the large computational overhead issue of fully homomorphic encryption must be addressed.

Within the framework of SERUMS, aiming at the practical secure privacy-preserving distributed machine (deep) learning, an architecture and a methodology are suggested for mitigating the impracticality issue of fully homomorphic encryption (arising from large computational overhead) via very fast gate-by-gate bootstrapping and introducing a learning scheme that requires homomorphic computation of only efficient-to-evaluate functions. The developed methodology is validated on a practical biomedical application example related to detection of individual stress using heart rate variability analysis.

5 Conclusions

The aim of this deliverable D4.3 is to report and describe the design and development of final versions of the project data masking, data fabrication, data quality verification and semantic-preserving data encryption technologies. All these technologies are used to explore and develop techniques and mechanisms to ensure the security and protection of the personal medical data that is shared as part of a coherent smart health-care system and to enable and facilitate the application of the Serums advanced data analytics on personal medical data, while fully adhering to necessary privacy regulations.

First, the document describes IBM's Data Fabrication Technology and the tool enhancements implemented during the third project year to improve user experience and enable modeling and fabrication of more complex medical data for the project use-cases. Further, the document describes an extended version of our approach to estimating the quality of fabricated synthetic data to ensure that all data analytics and user authentication tools developed by Serums consortium will be fully applicable for real medical data in the future. The document also describes our approach to Fully Homomorphic Encryption to be able to apply the advanced data analytics and machine-learning algorithms for analyzing encrypted personal data.

This document is the third deliverable of Work Package 4: "Secure and Privacy-Preserving Data Communication".

References

[1] “Create high-quality test data while minimizing the risks of using sensitive production data.” *IBM InfoSphere Optim Test Data Fabrication*, IBM, 2017, https://www.ibm.com/il-en/marketplace/infosphere-optim-test-data-fabrication.

[2] “Test Data Fabrication.” *Security and Data Fabrication*, IBM Research, 2011, https://www.research.ibm.com/haifa/dept/vst/eqt_tdf.shtml.

[3] “Constraint Satisfaction.” IBM Haifa Research, IBM, 2002, https://www.research.ibm.com/haifa/dept/vst/csp.shtml.

[4] Y. Richter, Y. Naveh, D. L. Gresh, and D. P. Connors (2007), "Optimatch: Applying Constraint Programming to Workforce Management of Highly-skilled Employees", International Journal of Services Operations and Informatics (IJSOI), Vol 3, No. 3/4, pp. 258 - 270.

[5] Y. Naveh, Y. Richter, Y. Altshuler, D. Gresh, and D. Connors (2007), "Workforce Optimization: Identification and Assignment of Professional Workers Using Constraint Programming", IBM J. R&D.

[6] Y. Naveh, M. Rimon, I. Jaeger, Y. Katz, M. Vinov, E. Marcus, and G. Shurek (2006), "Constraint-Based Random Stimuli Generation for Hardware Verification", AI magazine Vol 28 Number 3.

[7] E. Bin, R. Emek, G. Shurek, and A. Ziv (2002). "Using a constraint satisfaction formulation and solution techniques for random test program generation", IBM Systems Journal, 2002.

[8] Merav Aharoni, Odellia Boni, Ari Freund, Lidor Goren, Wesam Ibraheem, Tamir Segev (2015), "Rectangle Placement for VLSI Testing", CPAIOR 2015: 18-30

[9] O. Boni, F. Fournier, N. Mashkif, Y. Naveh, A. Sela, U. Shani, Z. Lando, A. Modai (2012) "Applying Constraint Programming to Incorporate Engineering Methodologies into the Design Process of Complex Systems" Proceedings of the Twenty-Fourth Conference on Innovative Applications of Artificial Intelligence, Toronto, Ontario, Canada. AAAI 2012.

[10] Y. Ben-Haim, A. Ivrii, O. Margalit and A. Matsliah (2012) "Perfect Hashing and CNF Encodings of Cardinality Constraints", SAT 2012, Trento, Italy.

[11] E. Bin, O. Biran, O. Boni, E. Hadad, E. K. Kolodner, Y. Moatti, D. H. Lorenz (2011), "Guaranteeing High Availability Goals for Virtual Machine Placement", ICDCS 2011.

[12] Jeonghee Shin, John A Darringer, Guojie Luo, Merav Aharoni, Alexey Y Lvov, G Nam, Michael B Healy (2011), "Floorplanning challenges in early chip planning", SOCC Conference, 2011 IEEE International, pp. 388—393

[13] Y. Naveh (2010). "The Big Deal, Applying Constraint Satisfaction Technologies Where it Makes the Difference". Proceedings of the Thirteenth International Conference on Theory and Applications of Satisfiability Testing (SAT'10).

[14] S. Asaf, H. Eran, Y. Richter, D. P Connors, D. L. Gresh, J. Ortega, M. J. Mcinnis (2010). "Applying Constraint Programming to Identification and Assignment of Service Professionals". Accepted for presentation in The 16th International Conference on Principles and Practice of Constraint Programming (CP2010). The paper received the Best Application Paper Award.

[15] B. Dubrov, H. Eran, A. Freund, E. F. Mark, S. Ramji, and T. A. Schell, (2009). "Pin Assignment Using Stochastic Local Search Constraint Programming" in Proceedings of the 15th International Conference on Priniciples and Practice of Constraint Programming (CP'09), Edited by Ian P. Gent, pp 35-49.

[16] Y. Richter, Y. Naveh, D. L. Gresh, and D. P. Connors (2007), "Optimatch: Applying Constraint Programming to Workforce Management of Highly-skilled Employees", IEEE/INFORMS International Conference on Service Operations and Logistics, and Informatics (SOLI), Philadelphia, pp. 173-178.

[17] S. Sabato and Y. Naveh (2007), "Preprocessing Expression-based Constraint Satisfaction Problems for Stochastic Local Search", Proceedings of The Fourth International Conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (CP-AI-OR).

[18] Y. Naveh, M. Rimon, I. Jaeger, Y. Katz, M. Vinov, E. Marcus, and G. Shurek (2006), "Constraint-Based Random Stimuli Generation for Hardware Verification", IAAI 2006.

[19] Y. Richter, A. Freund, and Y. Naveh (2006), "Generalizing AllDifferent: The SomeDifferent constraint", Proceedings of the 12 International Conference on Principles and Practice of Constraint Programming - CP 2006, Lecture Notes in Computer Science, Volume 4204, pages 468-483.

[20] Y. Naveh and R. Emek (2006). "Random stimuli generation for functional hardware verification as a CP application - a demo", IAAI 2006.

[21] Y. Naveh (2005). "Stochastic solver for constraint satisfaction problems with learning of high-level characteristics of the problem topography" CP 2005

[22] F. Geller and M. Veksler (2005), "Assumption-based pruning in conditional CSP", in van Beek, P., ed., CP, "Principles and Practice of Constraint Programming - CP 2005" of Lecture Notes in Computer Science (3709), 241-255 Springer.

[23] R. Dechter, K. Kask, E. Bin, and R. Emek (2002). "Generating random solutions for constraint satisfaction problems", AAAI 2002.

[24] D. Lewin, L. Fournier, M. Levinger, E. Roytman, G. Shurek (1995). "Constraint Satisfaction for Test Program Generation", Internat. Phoenix Conf. on Computers and Communications, March 1995.

[25] Juvekar, C., Vaikuntanathan, V., Chandrakasan, A.: Gazelle: A low latency framework for secure neural network inference. arXiv preprint arXiv:1801.05507 (2018).