**Ing. Martin Flusser – Doctoral Thesis – Mathematical Modeling of Anomalies in Large-Scale Vector and Structural Data**

Supervisor Review
RNDr. Petr Somol, Ph.D., 14th May 2023

Martin Flusser had been known to me as a clever student versed in mathematical thinking even before I agreed to act as supervisor of his doctoral thesis effort. Later on I got to know his involvement in Cisco Systems as an intern, which showed his ability to apply formal thinking to complex problems dictated by the technological foundations of our digital world. His combined experience with both the scientific fundamentals and the applied environment later helped outline the goals for work on what has now become his thesis.

The thesis explores a core problem of anomaly detection from multiple novel and complementary angles. In this it touches a topic of great current interest and impact, and explores it both from theoretical and applicational standpoint. The close relation of the topic to cybersecurity problems Mr. Flusser observed as part of his Cisco internship helps ensure that any positive results will have the potential of significant applied impact.

In terms of Anomaly Detection state of the art the thesis addresses particularly those settings that are less commonly addressed - world-wide extent of data implying extreme demands on usability under capped computational resources, maintainability of applied method over time under such conditions, and then the utilisation of information content in structural data which is prevalent in digital world but rarely addressed in theoretical literature. All these topics by far can not be considered solved in prior art and provide broad space for innovative ideas and meaningful progress. The goals of the thesis have been set accordingly with the ambition to provide well founded and impactful advance.

Among all the presented results the key idea revolves around the construction of auxiliary representation of the anomaly space to enable training a surrogate model for any existing anomaly detector. The idea is general and can be applied with virtually any pair of source and target anomaly detectors. The author covers the problem in general to reasonable extent, however, then correctly focuses more on one particular target model - the neural networks - to take advantage of the neural models´ speed of inference, the enormous current supportive knowledge base as well as software and hardware support. On the side of source anomaly detectors the author explores k-nearest neighbour as the prime example of non-parametric anomaly detection models. The special reason for such a choice is the excellent accuracy that k-nearest neighbour models achieve in industrial applications, while suffering from problematic computational and space complexity under

extreme problem size. The core idea presented in the thesis is then explored and shown to bring crucial advantages. The key advantage is the possibility to replicate the great accuracy of k-nn in a surrogate model that scales significantly better. As a significant second effect the author found out that surrogate models may even improve on accuracy, through the regularising effect of surrogate learning. Accordingly, this significant result has been accepted for publication in an impacted journal. The core idea later proved strong enough to open up new avenues of further expansion as well as solving problems as of yet considered unsolvable in practice. The extent of these problems addressed in the thesis forms a coherent and logically well connected study enabling practical application of the idea in multiple forms.

The student worked on the development of all respective ideas independently and with determination until he reached at least a partial goal with respect to success metrics, either based on accuracy or computational complexity. The key ideas presented in individual chapters of the thesis are his original ideas. Apart from the idea of surrogate models he worked out the online version of surrogate model training and eventually worked out the basic set of data generation strategies for the multiple-instance-learning form of surrogate method. All experimental evaluation, from the design of experiments to the reasoning and actual measurements have been proposed and performed by the student. In terms of achieved advances – the biggest value is in applied research results with extensive, mostly empirical evaluation. All proposed methods are well suitable to address real world problems and close gaps in existing prior art portfolios. The impact of these results is for me without doubt, I see direct applicability of at least some of the results in the digital industry, particularly in cybersecurity where the scale of application can reach dozens of millions of users.

To comment on the author's growth areas and thesis flaws: 1. the author occasionally struggles with formulating ideas well. Important ideas occasionally receive only brief treatment, followed by lengthy description of some side-aspect of lesser importance. Occasionally even trivial ideas receive disproportionate space. Their presence is usually well meant to introduce context for a more complex topic, the problem is in the proportion. A related problem may be the occasional repetition of the same reasoning without adding much novelty. 2. I would recommend a bit more care with respect to formalism, introduced notion should be always followed with absolute consistency and should not later get changed. 3. The student always strives to explore prior art sources in depth and always cites extensively. It was therefore surprising to witness one or two cases of omission, when an idea (though of of lesser importance) had not been subjected to sufficient prior art scrutiny. I should stress that such cases have later been identified and corrected when the student discovered how much the context of prior art can change the very core of one's assumptions. I believe this proved to be a valuable lesson for the student as it cost him a notable additional time.

From the three publications I co-authored with the student I can claim to have contributed none of the crucial ideas or worked-out methods or experiments. My impact in all cases have been in questioning the author's assumptions, helping to assess the importance of some of the research paths in consideration, and consulting evaluation methodologies. On the level of the thesis as a whole, only the topic of the last chapter has been recommended by me to the student, I thus gave the lead towards an open problem that would be great to address. This has been accepted by the student and developed with little further input from my side. Throughout the texts I then gave some advice regarding structure, readability, and general scientific presentation discipline.

Summary

All important ideas and results in the thesis have been proposed, researched and evaluated by the student and thus are his own contributions. The thesis covers three connected research areas in which prior art leaves significant gaps. Some of them have been successfully closed by this thesis. More than one of the non-trivial presented results have clear practical value in industrial applications (in terms of scalability up to world-wide scale). I consider the goals of the student's research work fulfilled. For all these reasons I recommend this thesis to be accepted as a doctoral thesis.

Petr Somol
Institute of Information Theory and Automation, Czech Academy of Sciences
and Gen Digital
somol@utia.cas.cz, petr.somol@gmail.com