Adaptive Anomalous Behavior Identification in Large-Scale Distributed Systems

Javier Álvarez Cid-Fuentes

School of Computer Science

The University of Adelaide

This dissertation is submitted for the degree of

Doctor of Philosophy

Supervisors: Dr. Claudia Szabo and Prof. Katrina Falkner

August 2017

A mi familia.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968. The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Javier Álvarez Cid-Fuentes

August 2017

Acknowledgements

First of all I would like to thank my two supervisors, Dr. Claudia Szabo and Prof. Katrina Falkner, for their support and guidance during these three years (and a bit). Thank you for your patience, help, insight, and above all, thank you for giving me the opportunity to work with you and for teaching me everything about research.

I would also like to thank the people who, besides my supervisors, helped me to carry out my studies in Adelaide. Francesc, Cruz, Rosa, and Wamberto, this thesis would not have been possible without your support.

I am also very thankful to the university staff and students for creating a really enjoyable atmosphere that makes things much easier. I am especially grateful to the people who, in one way or another, helped me during my studies. Thanks to Dan and Marianne for their immense technical support; to Yongrui and Ali for our fruitful talks; to Dídac for his brief but really insightful assistance; to Yuval for his help in the area of security; and to Lachlan and Jamal for the good times spent together. Finally, I would like to give special thanks to Pádraig and Zhigang for their academic and personal support. These years would not have been the same without you.

Por otro lado, también me gustaría darle las gracias a todas las personas que, a nivel personal, han estado a mi lado durante estos años. En primer lugar, muchísimas gracias a mis padres, que han trabajado durísimo toda su vida para darnos a mí y a mi hermana unas oportunidades que ellos no tuvieron. Sin vosotros no sería quien soy, ni estaría donde estoy. Gracias también a mi hermana, por estar siempre ahí, y a mis

tíos, Maribel y José Luís, por haber hecho esta experiencia más llevadera aun estando a miles de kilómetros de aquí.

Gracias a todos los amigos que están lejos pero están, gracias a Esteve por el soporte y por haberme traído un trocito de casa contigo en los días en los que estuviste aquí, y gracias a Juan por aquel fin de semana en Montreal que se pareció a los míticos jueves. Agradecérselo también a todas esas *aves de paso* que han hecho de mi estancia en Australia una experiencia inolvidable: Carles, Vicent, Zulia, Sandra, Nacho, Mehdi, Luís, Tania, Andrea, Carmen, Gorka, Rafa, Cristian, Jonathan, Alex y Jorge. A Alfonso y Gerard por los buenísimos ratos y las risas. Y en especial a Pablo y Cassia por su gran ayuda y las muchas tardes y noches que hemos compartido.

Por último, *last but not least*, dar las gracias a Ana, una de las personas más importantes de mi vida, por haber compartido conmigo este viaje que no ha sido fácil, por la cantidad y la calidad de los momentos, y por haber estado a mi lado incondicionalmente.

Abstract

Distributed systems have become pervasive in current society. From laptops and mobile phones, to servers and data centers, most computers communicate and coordinate their activities through some kind of network. Moreover, many economic and commercial activities of today's society rely on distributed systems. Examples range from widely used large-scale web services such as Google or Facebook, to enterprise networks and banking systems. However, as distributed systems become larger, more complex, and more pervasive, the probability of failures or malicious activities also increases, to the point that some system designers consider failures to be the norm rather than the exception.

The negative effects of failures in distributed systems range from economic losses, to sensitive information leaks. As an example, reports show that the the cost of downtime in industry ranges from $100K to $540K per hour on average. These undesired consequences can be avoided with better monitoring tools that can inform system administrators of the presence of anomalies in the system in a timely manner. However, key challenges remain, such as the difficulty in processing large amounts of information, the huge variety of anomalies that can appear, and the difficulty in characterizing these anomalies.

This thesis contributes a novel framework for the online detection and identification of anomalies in large-scale distributed systems that addresses these challenges. Our framework periodically collects system performance metrics, and builds a behavior characterization from these metrics in a way that maximizes the distance between nor-

mal and anomalous behaviors. Our framework then uses machine learning techniques to detect previously unseen anomalies, and to identify the type of known anomalies with high accuracy, while overcoming key limitations of existing works in the area. Our framework does not require historical data, can be employed in a *plug-and-play* manner, adapts to changes in the system behavior, and allows for a flexible deployment that can be tailored to numerous scenarios with different architectures and requirements.

In this thesis, we employ our framework in three anomaly detection application domains: distributed systems, large-scale systems, and malicious traffic detection. Extensive experimental studies in these three domains show that our framework is able to detect several types of anomalies with 0.80 *Recall* on average, and 0.68 mean *Precision* or 0.082 mean FPR depending on the domain. Moreover, our framework achieves over 0.80 accuracy in the identification of various types of complex anomalous behaviors. These results significantly improve similar works in the three explored research areas.

Most importantly, our approach achieves these detection and identification rates with significant advantages over existing works. Specifically, our framework does not rely on historical anomalous data or on assumptions on the characteristics of the anomalies that can make anomaly detection easier. Moreover, our framework provides a flexible and highly scalable design, and an adaptive method that can incorporate new system information at run time.

Table of contents

List of figures

List of tables