Journal-first presentation at IEEE/ACM 42nd International Conference
on Software Engineering (ICSE '20), Seoul, South Korea (2020)

An Extended Abstract of "Metamorphic Robustness Testing: Exposing Hidden Defects in Citation Statistics and Journal Impact Factors"

Zhi Quan Zhou
School of Computing & IT
University of Wollongong
Wollongong, NSW 2522, Australia
zhiquan@uow.edu.au

T. H. Tse
Department of Computer Science
The University of Hong Kong
Pokfulam, Hong Kong
thtse@cs.hku.hk

Matt Witheridge
School of Computing & IT
University of Wollongong
Wollongong, NSW 2522, Australia
mw204@uowmail.edu.au

ACM Reference Format:
Zhi Quan Zhou, T. H. Tse, and Matt Witheridge. 2020. An Extended Abstract of "Metamorphic Robustness Testing: Exposing Hidden Defects in Citation Statistics and Journal Impact Factors".

There has been an increasing concern from both industry and the research community about *software robustness*: the ability of dealing with erroneous input or unexpected situations. To assess robustness, the software under test (SUT) needs to be tested with invalid or erroneous input, and a major approach for this purpose is fuzzing, where random or semi-random input is used to test the SUT. Although fuzzing can generate unexpected test cases, it may not necessarily cover all types of real-life erroneous input, and the tester may not be able to fuzz the environment. For example, when testing a Web search engine in a real-life operational environment, it is straightforward to apply a fuzzer (fuzz testing tool) to generate random query terms, but it is difficult to change the environment (which is the real-world Internet), unless the testing is conducted in a constrained environment with mock databases. Another limitation of fuzzing is that, due to the *oracle problem*, it is hard for a fuzzer to detect logic errors (which do not crash the SUT, but instead produce incorrect output values).

A growing body of research has examined the concept of metamorphic testing (MT), and proven it to be a highly effective approach to the oracle problem. The present research [2] extends MT for robustness testing beyond fuzzing, in the context of testing big data applications. Our objective is to assess the SUT's robustness in terms of producing logically correct or reasonable output for erroneous input that does not crash the system. In this research, the subject software under consideration is automatic indexing systems, which provide fundamental IT infrastructure for the present-day knowledge society. Such systems are difficult to test due to the sheer volume of data that they process. The contributions of this research are summarized as follows:

ICSE'20

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(1) We propose a *metamorphic robustness testing* approach, which tests the robustness of software for erroneous inputs in the absence of an oracle. We identify the following three *metamorphic relations* to test citation database systems: (*i*) Given two large sets of publications without systematic differences in factors related to potential impacts, their mean citation counts should generally have little systematic difference. (*ii*) Older publications should generally have higher citation counts than newer publications. And (*iii*) Given two large sets of publications without systematic differences in factors related to potential impacts, when time progresses, the mean citation counts of the two sets should increase at a similar rate.

(2) We report a surprising finding that the inclusion of hyphens in paper titles impedes citation counts, and that this is a result of the lack of robustness of the Scopus and Web of Science citation database systems in handling hyphenated paper titles—this finding is obtained through large-scale empirical studies using metamorphic robustness testing. We show that our results are valid for the entire literature as well as for individual fields such as chemistry.

(3) We go on to investigate the impact of hyphens in paper titles at the journal level, and report a further surprising finding that there is a strong and significant negative correlation between the journal impact factor (JIF) of *IEEE Transactions on Software Engineering* (TSE) and the percentage of hyphenated paper titles published in TSE. A similar finding is made for *ACM Transactions on Software Engineering and Methodology*. A software engineering field-wide study reveals that the higher JIF-ranked journals are publishing a lower percentage of papers with hyphenated titles.

We provide a careful analysis of the validity of this research to avoid falling into the trap of equating correlation with causation. Our results challenge the common belief that citation counts are a reliable measure of the impact of papers. Our findings have been reported by international news media. On the Internet, the implications of our findings have been discussed from diverse perspectives, some of which extend beyond the conventional computing and citation domains (for example, see [1]). On 7 August 2019, our work was reported by *Nature* (https://go.nature.com/2TdWoM6).

REFERENCES

[1] Chris Burns. 2019. Brave New World predictions are coming true with hyphens. https://www.slashgear.com/brave-new-world-predictions-are-coming-true-with-hyphens-31578773/

[2] Zhi Quan Zhou, T. H. Tse, and Matt Witheridge. DOI: https://doi.org/10.1109/TSE.2019.2915065. Metamorphic robustness testing: Exposing hidden defects in citation statistics and journal impact factors. *IEEE Transactions on Software Engineering* (DOI: https://doi.org/10.1109/TSE.2019.2915065).