Normalised Squeeziness and Failed Error Propagation

David Clark[a], Robert M. Hierons[b], Krishna Patel[b]

[a]*Department of Computer Science, University College London, London, WC1E 6BT United Kingdom*
[b]*Department of Computer Science, The University of Sheffield, Sheffield, S1 4DP, United Kingdom*

Abstract

Failed Error Propagation (FEP) can reduce test effectiveness and recent work proposed an information theoretic measure, Squeeziness, as the theoretical basis for avoiding FEP. This paper demonstrates that Squeeziness is not suitable for comparing programs with different input domains. We then extend Squeeziness to Normalised Squeeziness and demonstrate that this is more generally useful.

Keywords: Software engineering; formal methods; software testing; fault masking

1. Introduction

This paper is motivated by the need to understand and predict *Failed Error Propagation (FEP)*, a situation arising in software testing in which the program state is corrupted, due to a faulty statement being executed, but the expected/correct output is still produced. Clearly, FEP has the potential to reduce the effectiveness of testing and empirical studies have confirmed that FEP is an issue in practice e.g. Masri et al. found that most programs in their corpus were affected by coincidental correctness (including FEP); 60% of tests were affected for 13% of the programs [5, 4, 2]. In what follows, we distinguish between coincidental correctness, a general term meaning an error is executed but the test oracle still observes the correct output, and FEP which is a particular type of coincidental correctness.

In order to reason about FEP we make some strong, simplifying assumptions [1]. First, we assume that there is only one error in the program. Second, we assume that

Email addresses: `david.clark@ucl.ac.uk` (David Clark), `r.hierons@sheffield.ac.uk` (Robert M. Hierons), `krishna.patel@sheffield.ac.uk` (Krishna Patel)

the error is not embedded in the body of a loop. The consequence of these assumptions is that, if we examine the error program and its fixed version, there is a program point immediately after the error in the error program, the alignment point, after which the subprograms of the two programs are the same. We can then model FEP as the behaviour of a code fragment on a pair of different initial states for the fragment, i.e. the corrupted and uncorrupted states produced at the alignment point by the two programs. FEP occurs when the same output is observed. Thus, in our model, a prerequisite for FEP is collisions (two different program states produce the same output), or many to one behaviour by a code fragment. This FEP model and its assumptions have been statistically validated by Androutsopoulos et al. who report a rank correlation of over 0.95 between variations on an information theoretic measure and the probability of FEP in numerical C programs with seeded errors [1].

This measure, Squeeziness, has been shown to be strongly associated with the probability of collisions [3]. In this current paper we examine an information theoretic measure closely related to Squeeziness, called Normalised Squeeziness. This latter is designed to compare different programs with different input domains. We show that it has an improved correlation with the probability of collisions.

To simplify both the presentation and experimentation we assume in what follows that the probability distribution is discrete and uniform. However, it is not difficult to parametrise the reasoning with an arbitrary probability distribution. We examine correlations between the probability of collisions, and both Squeeziness and Normalised Squeeziness. Studying purely abstract functions, we experimentally demonstrate that Normalised Squeeziness is the better estimation measure for the probability of collisions.

2. Squeeziness

Definition 1. *Entropy of a random variable.*

$$\mathcal{H}(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

Definition 2. *Squeeziness [3]. Let I and O be random variables and f be a total function, $f : I \rightarrow O$. The* Squeeziness *of f on I, $Sq(f, I)$, is defined as the loss of* *information after applying f to I*

$$Sq(f) = \mathcal{H}(I) - \mathcal{H}(O)$$

The entropy of a random variable satisfies a partition property [6, 4, 2]. Consider random variable A. We overload A to mean both the r.v. and its set of events. Consider a partition of A with n parts. Let B be the r.v. with event space $B_1, \ldots, B_n$ (the parts of the partition) and name the associated probability distributions as follows:

$$A \sim \sigma \tag{1}$$

$$B \sim \sigma' \text{ where } \sigma'(B_i) = \sum_{a \in B_i} \sigma(a) \tag{2}$$

$$B_i \sim \sigma_i \text{ where } 1 \leq i \leq n \text{ and for } a \in B_i, \sigma_i(a) = \sigma(a)/\sigma'(B_i) \tag{3}$$

We have the following property, used to construct the example motivating Normalised Squeeziness..

Property 1. *Partition Property of Entropy*

$$\mathcal{H}(A) = \mathcal{H}(B) + \sum_{1 \leq i \leq n} \sigma'(B_i).\mathcal{H}(B_i)$$

This can be interpreted as follows. Given a partition on a random variable, the entropy of the random variable is equal to the expected value of the entropy within each part plus the entropy between the parts.

A collision is said to have occurred when two different inputs map to the same output. Let $f : I \to O$ be the input-output semantics of a program and I and O be overloaded to represent random variables in the inputs and outputs respectively. Let r.v. $O \sim \sigma_O$. Let $f^{-1}o$ be the inverse image of $o \in O$. Then $I = \bigcup_{o \in O} I_o$ where $I_o = f^{-1}o$ and $\bigcap_{o \in O} I_o = \emptyset$, i.e. $\{I_o \mid o \in O\}$ partitions I. Collisions have the potential to cause FEP, if an infected state collides with an uninfected state. We compare Squeeziness with the probability of collisions due to the execution of a program.

Definition 3. *Function Collision. A pair of inputs to function f, $t, t' \in I$,* collide *if $t \neq t' \ \wedge \ ft = ft'$.*

Let us suppose that $O = \{o_1, \ldots, o_n\}$ and I_i denotes set $f^{-1}o_i$ with size m_i. Further, let $d = \sum_{i=1}^{n} m_i$ denote the input space size. Given a uniform distribution on inputs, the probability of collision is the following (sampling without replacement) [3].

$$PColl(f) = \sum_{i=1}^{n} \frac{m_i * (m_i - 1)}{d * (d-1)}$$

We can approximate this by the following function (sampling with replacement).

$$p_{coll}(f, I) = \sum_{o \in O} \sigma_O^2(o)$$

The two measures converge as the size of the input space gets larger.

3. A Squeeziness Limitation

In this section we examine the suitability of Squeeziness for comparing functions with different input domains.

Example 1. *Consider a function f from I to O and a function g_1 formed by taking the disjoint union of $k > 1$ copies of f as follows: g_1 has input domain $I' = \{x^i | x \in I \wedge 1 \leq i \leq k\}$ and output domain $O' = \{y^i | y \in O \wedge 1 \leq i \leq k\}$ (I' has k copies of I and O' has k copies of O); and for all $x \in I$ and $1 \leq i \leq k$, if $f(x) = y$ then $g_1(x^i) = y^i$.*

I' is partitioned into k parts, $I_1, \ldots, I_k$, and each I_i is a copy of I. Similar is true for O', the O_i, and O. The probability distribution over the parts in each case is the uniform distribution with probability $1/k$. Let J be the random variable in the parts of I' and let σ_I be the probability distribution on the events of random variable I. Then $I \sim \sigma_I$, $I' \sim \sigma$ (where for $0 \leq i \leq k, x \in I, \sigma(x^i) = 1/k \, . \, \sigma_I(x)$), and $J \sim \mathcal{U}_k$ (the uniform probability distribution on k discrete items).

We can similarly consider the output space for g_1 where J is again the random variable in the parts of the partition and σ_O is the probability distribution on O ($O \sim \sigma_O$, $O' \sim \sigma$ where for $0 \leq i \leq k, x \in O, \sigma(x^i) = 1/k \, . \, \sigma_O(x)$, and $J \sim \mathcal{U}_k$).

Now consider the Squeeziness of f and g_1. We have that $Sq(f, I) = \mathcal{H}(I) - \mathcal{H}(O)$ and $Sq(g_1, I') = \mathcal{H}(I') - \mathcal{H}(O')$. By the partition property

$$\mathcal{H}(I') = \mathcal{H}(J) + \sum_{0 \le i \le k} 1/k \, . \, \mathcal{H}(I_i) = log_2 k + \mathcal{H}(I)$$

and by the same argument $\mathcal{H}(O') = log_2 k + \mathcal{H}(O)$*, so*

$$Sq(g_1, I') = \mathcal{H}(I') - \mathcal{H}(O') = \mathcal{H}(I) - \mathcal{H}(O) = Sq(f, I)$$

Calculating the probability of collisions for $f : I \to O$ and $g_1 : I' \to O'$ we have:

$$p_{coll}(f, I) = \sum_{o \in O} \sigma_O^2(o)$$

$$p_{coll}(g_1, I') = \sum_{o \in O'} [1/k \, . \, \sigma_O(o)]^2 = 1/k^2 \, . \sum_{o \in O'} \sigma_O^2(o)$$

In the above, f and g_1 have the same Squeeziness. However, the probability of collisions is higher with f than g_1 since the inverse images of outputs have the same sizes but the input domain size is larger for g_1 than f (i.e. for f one is more likely to choose two inputs mapped to a given output).

Thus, as Squeeziness uses the random variable in inputs as a parameter, it may not be a good basis for comparing two programs if they have different input domains. An alternative, to using the loss of information, is to measure the *proportion* of information lost.

Definition 4. *Normalised Squeeziness. The* Normalised Squeeziness *of total function* $f : I \to O$*,* $NSq(f)$*, is defined as the proportion of information lost after applying* f *to* I

$$NSq(f) = \frac{\mathcal{H}(I) - \mathcal{H}(O)}{\mathcal{H}(I)}$$

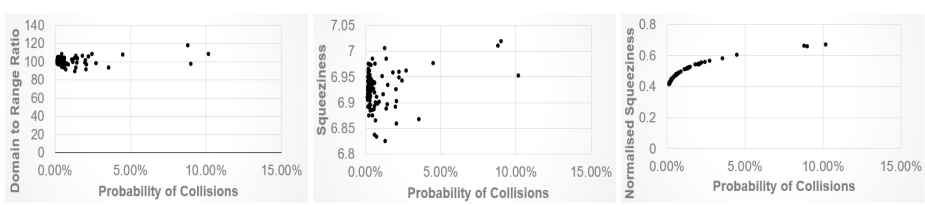
In the following section we report on the results of experiments that evaluated Normalised Squeeziness by comparing it with Squeeziness and what appears to be the first method of estimating the likelihood of FEP; the *Domain To Range Ration (DTRR)* [3]. Given a function (or program) p, the DTRR of p is simply the size of the output domain divided by size of the input domain. The Domain to Range Ratio provides a rough indication of the number of inputs that map on to a single output, and thus, an estimate of the probability of collisions, and by implication, the likelihood of FEP.

4. Simulation Study

This section describes a study that evaluated the DTRR, Squeeziness, and Normalised Squeeziness metrics by comparing them with a fourth metric (probability of collisions). We initially replicated the previous work, using fixed input domain size; the results are not described since they were consistent with earlier work (Squeeziness correlates strongly with probability of collision and more strongly than DTRR). In the following we describe two new experiments. All of the metrics of interest can be computed if we know the sizes of the inverse images of outputs. Thus, we followed the previously used approach of randomly[1] choosing inverse image sizes [3], which we now describe.

4.1. Variable Input Domain Size and Fixed Maximum Inverse Image Size

Let us suppose that we have chosen maximum input domain size Md and also we have chosen a maximum Mm on the size of the inverse image of outputs. In the experiments described in this section, we fixed Md and Mm and each experimental subject was generated as follows. First, the size of the input domain d was randomly chosen from the range 1 to Md. Second, we iteratively generated inverse image sizes in the range from 1 to Mm until the sum of these was at least Md; if the sum exceeded Md then we reduced the size of the last value chosen so that the sum was Md. This resulted in a value of d and values of the m_i. We then computed the four metrics of interest. This process was repeated 100 times for each Md/Mm pair used.



(a) $(10^5, 200)$ (b) $(10^5, 200)$ (c) $(10^5, 200)$

Figure 1: Datasets used to derive the statistical test results reported in the first row of Table 1.

[1]Note that randomisation was facilitated by Python's inbuilt random module.

Parameters		**DRR**		**Sq**		**NSq**	
Md	Mm	$\rho_{\mathbf{DRR}}$	**P(2-tailed)**	$\rho_{\mathbf{Sq}}$	**P(2-tailed)**	$\rho_{\mathbf{NSq}}$	**P(2-tailed)**
10^5	200	0.004	0.969	0.024	0.812	0.999	4.81E-127
10^5	200	-0.073	0.468	-0.068	0.501	0.997	6.05E-112
10^5	2000	-0.148	0.141	-0.088	0.386	0.995	3.26E-100
10^5	2000	-0.113	0.265	-0.055	0.589	0.996	3.99E-104
10^6	200	-0.306	0.002	-0.201	0.045	1.000	1.46E-160
10^6	200	-0.091	0.367	-0.199	0.047	0.999	1.15E-147

Table 1: Correlations between Prob. of Collisions and Squeeziness, Normalised Squeeziness, and DTRR

We used the following pairs of values (Md, Mm): $(10^5, 200)$, $(10^5, 2000)$, $(10^6, 200)$, using each twice to explore sampling effects. The results are shown in Table 1. Figures 1a to 1c show the distributions of values for the three experiments that are presented in the first row of Table 1 (the distribution of values are similar for the other experiments). The order of the scatterplots corresponds to the arrangement in Table 1.

Interestingly, in several cases the correlation between Squeeziness and the Probability of Collisions is negative but in others the correlation is positive. It thus appears that Squeeziness is not a good basis for predicting Probability of Collisions. A similar observation can be made with respect to DTRR. In contrast, Normalised Squeeziness consistently has a strong, positive correlation with the Probability of Collisions, obtaining a correlation coefficient that is always above 0.995.

4.2. Variable Input Domain Size and Variable Maximum Inverse Image Size

The experiments described above used fixed Mm. To allow greater variety, we ran experiments in which Mm was randomly chosen (for each experimental subject) in the range 1 to MMm for some MMm.

We used two pairs of values for (Md, MMm), $(10^5, 200)$, $(10^5, 2000)$, again using each pair twice. The results (Table 2) indicate that again only Normalised Squeeziness is correlated with Probability of Collisions. Interestingly, the correlation is marginally weaker than before but we still obtain a positive, strong, and consistent correlation co-

Parameters		DRR		Sq		NSq	
Md	MMm	ρ_{DRR}	**P(2-tailed)**	ρ_{Sq}	**(2-tailed)**	ρ_{NSq}	**P(2-tailed)**
10^5	200	0.559	1.49E-09	0.559	1.45E-09	0.921	7.37E-42
10^5	200	0.610	1.57E-11	0.606	2.29E-11	0.925	7.03E-43
10^5	2000	0.562	1.19E-09	0.570	6.22E-10	0.975	6.24E-66
10^5	2000	0.672	2.04E-14	0.674	1.54E-14	0.959	2.29E-55

Table 2: Experiments based on Md and MMm.

efficient.

4.3. Threats to Validity

There are several threats to validity. The threats to internal validity relate to the possibility that the tools used in the experiments were faulty. We addressed these in two main ways. First, we carefully tested the tool we developed for the simulations and the components it was built from. In addition, we ran a replication study, obtaining results that were similar to those produced in previous work. For the statistical analysis we utilised a widely used tool (SPSS)

Threats to construct validity refer to the possibility that we did not measure properties of interest. As previously explained, we concentrated on the probability of collisions. There is a clear connection between collisions and FEP (for a sub-program following the alignment point) and this observation has been supported by previous work. However, experiments that looked at actual FEP would have value.

Threats to external validity relate to our ability to generalise from any results given. There are always such threats and in this case the main threat comes from the way in which the simulations were generated: it is possible that the inverse images of outputs are very differently distributed in real programs. To address this, we require experiments using a range of real programs along with faulty versions. This is an issue that will be addressed by future work.

5. Conclusions

Previous work introduced the measure called Squeeziness, which can be used to estimate the likelihood of FEP [3]. In this paper, we showed that Squeeziness is less appropriate for comparing programs that have different input domains and Normalised Squeeziness resolves this issue. Specifically, we found that, when the input domain size varied, there was little correlation between the Probability of Collisions and the two previous measures (Squeeziness and the Domain to Range Ratio). However, there was a correlation between Normalised Squeeziness and the Probability of Collisions.

Acknowledgements

Patel and Hierons were partly supported by EPSRC grant EP/P006116/1. Clark was partially supported by EPSRC grant EP/P005888/1.

[1] K. Androutsopoulos, D. Clark, H. Dan, R. M. Hierons, and M. Harman. An analysis of the relationship between conditional entropy and failed error propagation in software testing. In *Proceedings of the 36th International Conference on Software Engineering*, pages 573–583, New York, USA, 2014. ACM.

[2] R. A. Assi, C. T. M. Maalouf, and W. Masri. Does the testing level affect the prevalence of coincidental correctness? *CoRR*, abs/1808.09233, 2018.

[3] D. Clark and R. M. Hierons. Squeeziness: An information theoretic measure for avoiding fault masking. *Information Processing Letters*, 112(8-9):335340, 2012.

[4] W. Masri and R. A. Assi. Prevalence of coincidental correctness and mitigation of its impact on fault localization. *ACM Transactions on Software Engineering and Methodology*, 23(1):1–28, 2014.

[5] W. Masri, R. A. Assi, M. El-Ghali, and N. Al-Fatairi. An Empirical Study of the Factors That Reduce the Effectiveness of Coverage-based Fault Localization. In *Proceedings of the 2nd International Workshop on Defects in Large Software Systems*, pages 1–5, NY, USA, 2009. ACM.

[6] A. Renyi. *Probability Theory*. Dover Publications Inc, 1970.