PREPRINT. Theoretical Computer Science, 106:391-393, 1992.

Letter to the Editor: Comments on "On the Balance Property of Patricia Tries: External Path Length Viewpoint"

Arne Andersson
Department of Computer Science
Lund University
Sweden
email: arne@dna.lth.se

The method of analyzing data structures and algorithms by assuming that the treated items are randomly and independently drawn from some kind of distribution, known or unknown, is widely used among computer scientists. Although real-world data are very seldom generated this way, such analysis gives a lot of help in understanding the behaviour of the algorithms and, therefore, it is well motivated. However, some care has to be taken when conclusions are drawn from the analysis.

In the paper "On the Balance Property of Patricia Tries: External Path Length Viewpoint" by Kirschenhofer, Prodinger, and Szpankowski, *Theoretical Computer Science* Volume 68, Number 1, it is argued that from a practical point of view the average case behaviour of a data structure is often more interesting than the worst case behaviour. The authors state that "the algorithmic design has often to be targeted at coping with quite unrealistic, if not pathological, inputs and the possibility is neglected that a simpler algorithm might perform just as well, or even better, in practice." Motivated by this argument they present an elegant average case analysis of the variance of the external path length of Patricia tries under the model of random input drawn from the uniform distribution. From this analysis the authors conclude that the external path length in a Particia trie "is asymptotically equal to $n \log_2 n$ *with probability one* (i.e. almost surely)". These results are used to show that "from the practical (average) viewpoint, the Patricia tree does not need to be *reconstructed* in order to keep it balanced", and, hence, "the Patricia is a very well-balanced tree, and in most practical

cases it does not need to be additionally rebalanced."

To me, it seems that when using reasonable meanings of the terms "pathological inputs" and "most practical cases", the conclusions drawn by Kirschenhofer et al. are not correct. Note though that I do not claim that their mathematical analysis is wrong. Indeed, the analysis is interesting and with more modest claims it would have been a very nice paper.

First, I agree with the statement that a good average behaviour is often sufficient in practical applications. However, when studying the practical behaviour of a data structure, what input could be more "pathological" than randomly chosen, uniformly distributed data? The analysis in the paper is heavily dependent on this model, and the authors do not mention how their analysis relate to practical situations.

A scientific investigation of a real-world phenomenon is very often performed by examining a *model*, either theoretically or by laboratory experiments. From the behaviour of the model we draw conclusions about the behaviour of nature. Examples are:

1. We can draw conclusions about how the *model* behaves, and how nature behaves in that particular case when the conditions of the model are exactly fulfilled.

2. We can draw conclusions about how nature behaves in cases when the conditions of the model are *approximately* fulfilled. We may argue that in those cases nature will behave approximately in the same way as our model.

3. Finally, we can draw the conclusion that nature will *always* (or almost always) behave like our model.

The conclusion drawn by Kirschenhofer et al. is of the third category.

Which type of conclusion *should* have been drawn?

The first type is quite modest and always true if the investigation is properly made, which seems to be the case in the paper referred to.

The second type is often realistic, but some care has to be taken. From the analysis of the expected external path length in a Particia trie for random independent uniform data we can probably draw some conclusions about which external path length we may expect when input is roughly evenly distributed. However, it is not clear in what way we can apply the *variance* for the random uniform model to such inputs. This must be clarified and motivated.

The third type is very strong and in order to draw such a conclusion we must show that our model properly reflects nature. This is certainly not made in the paper referred to.

Thus, the first type of conclusion is the one that could have been drawn by the authors. Probably, the second type could also have been made, but this would require a careful discussion. The third type of conclusion should definitely *not* have been drawn. Indeed, my personal experience is that in most practical applications the external path lengths of Patricia tries are significantly larger than $n \log_2 n$, in contrast to what is claimed by Kirschenhofer, Prodinger, and Szpankowski.

To sum up, it is of course important that scientific investigations — empirical and theoretical — are properly performed. However, in my opinion, even more important is that the *conclusions* are properly related to the results of the investigations. The cited paper is an example where this is not the case.