Abstract
Static checking tools are useful as an automated software inspection step that can easily be integrated in the development cycle and assist in creating high quality, reliable and secure code. However, an often quoted disadvantage of these tools is that they generate an overly large number of warnings, including many false positives due to their approximate analysis techniques. This information overload effectively limits their usefulness.

In this paper we present ELAN, a technique that helps the user prioritize the information generated by a software inspection tool, based on computing the likelihood that execution reaches the locations for which warnings are reported. This analysis is orthogonal to ranking techniques known from literature, such as severity levels and statistical analysis to filter false positives. We evaluate feasibility of our technique using a number of open source and Philips-specific case studies and assess the quality of our predictions by comparing them to actual values obtained by profiling software embedded in a TV.

1. Introduction
Software inspection [14] is widely recognized as an effective technique to assess and improve software quality and reduce the number of defects [28, 20, 39, 27, 40]. Software inspection involves carefully examining the code, design, and documentation of software and checking them for aspects that are known to be potentially problematic based on past experience.

In this paper, we focus on tools that perform automatic inspection. Such tools are interesting since automatic detection of defects and anomalies allows early (and repeated) detection of project deterioration and helps to ensure software quality, security and reliability. Early feedback enables early corrections, thereby lowering development costs and increasing chances of success.

Most defect detection techniques are built around some form of static analysis of the code. In its simplest form this can be the warnings generated by a compiler set to its pedantic mode. In addition, various dedicated static program analysis tools are available that assist in defect detection and writing reliable and secure software. A well-known example is the C analyzer LINT [23]. These tools form a complementary step in the development cycle and have the ability to check for more sophisticated program properties than can be examined using a normal compiler; moreover, they can often be customized, and as such benefit from specific domain knowledge.

However, such static analyses come with a price: in the case that the algorithm cannot ascertain whether the source code at a given location obeys a desired property or not, it will make the safest approximation and issue a warning, regardless of the correctness. Consequently, this conservative behavior can lead to false positives, incorrectly signaling a problem with the code. Kremenek and Engler [25] observed that program analysis tools typically have false positive rates ranging between 30–100%. In addition, the increased scrutiny with which the code is examined can lead to an explosion in the number of warnings generated, especially when the tool is introduced later in the development process or during maintenance, when a significant code base already exists.

To cope with the large number of warnings, users have to resort to using all kinds of (manual) filtering processes, often based on the perceived impact of the underlying fault. Even worse, our experience indicates that the information overload often results in complete rejection of the tools, especially in cases where the first few defects reported by the tool turn out to be false positives.

Goal In this paper, we set out to help the user of automated code inspection tools to deal with this information overload. Instead of aiming to improve a particular defect detection technique to reduce its false positives, we strive for a generic prioritization approach that can be applied to the results of any software inspection tool and assists the user in selecting
the most relevant warnings.

To this end, we propose ELAN (Execution Likelihood ANalysis), a technique which orders inspection results based purely on a prediction of the likelihood that execution reaches the reported defect. The rationale behind this approach is that the violating code needs to be executed in order to trigger the undesired behavior. As such, the execution likelihood can be considered a contextual measure of severity, rather than the severity based on defect types which is usually reported by software inspection tools. The ELAN algorithm is kept simple on purpose: scalability is an issue, as we want to be able to prioritize inspection results for large industrial systems.

Industrial Context The context of this work is the TRADER project in cooperation with Philips Semiconductors, in which we investigate and develop methods and tools for ensuring reliability of consumer electronics devices.

Modern consumer electronics such as mobile phones, televisions, and audio/video media centers, increasingly rely on embedded software for their operation. In the past, functionality of such devices was mostly implemented in hardware, but nowadays the features of these devices are made easily extendible and adaptable by means of software. As a consequence, the amount of software embedded in consumer electronics has grown tremendously.

For example, a modern television contains several million lines of C code and this amount is growing rapidly with new functionality, such as electronic program guides, increased connectivity with other devices, audio and video processing and enhancements, and support for various video encoding formats. During the development process, this code is routinely inspected using QA-C, one of the leading commercial software inspection tools currently on the market. Nevertheless, the developers have experienced problems handling the information overload mentioned earlier which motivated the research described in this paper.

Overview The remainder of this paper is organized as follows: section 3 presents our prioritization approach, section 4 details the ELAN algorithm. An overview of the experiments can be found in section 5. For a survey of related work, discussion of more advanced algorithms, and an in-depth treatment on the experiments we refer to the full paper.

2. Related Work

Automatic Code Inspection There are a number of tools that perform some sort of automatic code inspection. The most well-known is probably the C analyzer Lint [23] that checks for type violations, portability problems and other anomalies such as flawed pointer arithmetic, memory (de)allocation, null references, and array bounds errors. LClint and splint extend the Lint approach with annotations added by the programmer to enable stronger analy-

ses [12, 13]. Various tools specialize in checking security vulnerabilities. The techniques used range from lightweight lexical analysis [33, 38, 31] to advanced and computationally expensive type analysis [22, 18], constraint checking [35] and model checking [8]. Some techniques deliberately trade formal soundness for performance in order to scale to the analysis of larger systems [11, 6, 17] whereas others focus on proving some specific properties based on more formal verification techniques [2, 7, 9].

Several commercial offerings are available for conducting automated automatic code inspection tasks. Examples include QA-C, K7, CodeSonar, and Prevent. The latter was built upon the MECA/Metal research conducted by Engler et al. [42, 11]. Reasoning provides a defect analysis service that identifies the location of potential crash-causing and data-corrupting errors. Besides providing a detailed description of defects found, they report on defect metrics by measuring a system’s defect density and its relation to industry norms.

Ordering Inspection Results The classic approach most automated code inspection tools use for prioritizing and filtering results is to classify the results based on severity levels. Such levels are (statically) associated with the type of defects detected; they are oblivious of the actual code that is being analyzed and of the location or frequency of a given defect. Therefore, the ordering and filtering that can be achieved using this technique is rather crude. Our approach is based on the idea that this can be refined by taking into account certain properties of the identified defect with respect to the complete source code that was analyzed.

A technique that is more closely related to our approach, is the z-ranking technique by Kremenek and Engler [25]. They share our goals of prioritizing and filtering warnings based on their properties with respect to analyzed code but do so based on the frequency of defects in the results. Their approach aims to determine the likelihood that a given warning is a false positive. It is based on the idea that, typically, the density of defects in source code is low. Thus, when checking source code for a certain problem, there should be a large number of locations where that check is not triggered, and relatively few locations where it is triggered. Conversely, if a check results in many triggered locations and few non-triggered ones, these triggers are more likely to be false positives. This notion is exploited by keeping track of success and failure frequencies, and calculating a numeric score by means of a statistical analysis. The warning reports can then be sorted accordingly.

Besides severity levels and z-ranking, we are not aware of any other work that deals with ordering inspection results.

Static Profiling Static profiling is used in a number of compiler optimizations or worst-case execution time (WCET)
analyses. By analyzing program structure, a prediction is made as to which portions of the program will be most frequently visited during execution. Since this heavily depends upon branching behavior, some means of branch prediction is needed. This can range from simple and computationally cheap heuristics to expensive data flow based analyses such as constant propagation [24, 29, 3], symbolic range propagation [32, 4, 26], or even symbolic evaluation [15]. Although there has been a lot of research on branch prediction, there are only a few approaches that take this a step further and actually compute a complete static profile [41, 36]. For branch prediction, these use heuristics similar to the ones we employ. The difference with respect to our work is that they try to estimate the, computationally expensive, execution frequency, whereas we compute execution likelihood which is less involved. Moreover, in contrast with these approaches, we do analyze the complete program, but perform analysis in a demand-driven manner: i.e., only for the locations associated with the warning reports.

Testability Voas et al. [34] define software testability as “the probability that a piece of software will fail on its next execution during testing if the software includes a fault”. They present a technique, dubbed sensitivity analysis, that analyses execution traces obtained by instrumentation and calculates three probabilities for every location in the program. Together they give an indication of the likelihood that a possible fault in that location will be exposed during testing. The first of these three, execution probability, is similar to our notion of execution likelihood, the chance that a certain location is executed. The other two are the infection probability, i.e. the probability that the fault will corrupt the data state of the program, and the propagation probability, the likelihood that the corrupted data will propagate to output and as such be observable.

Although the concepts involved are very similar to our own, the application and analysis method differ greatly: a location that is unlikely to produce observable changes if it would contain an error should be emphasized during testing, whereas we would consider that location to be one of low priority in our list of results. In addition, Voas approximates these probabilities based on dynamic info whereas we try to make predictions purely statically. Finally, infection- and propagation probability apply to locations that contain faulty code, while the inspection results we are dealing with may also reveal security vulnerabilities and coding standard violations that do not suit these two concepts.

3. Approach

The approach we propose for prioritizing code inspection results is based on the workflow depicted in figure 1. The process consists of the following steps (starting at the top-left node):

1. The source code is analyzed using some code inspection tool, which returns a set of inspection results.

2. The inspection results are normalized to the generic format that is used by our tools. The format is currently very simple and contains the location of the warning in terms of file and line number, and the warning description. We include such a normalization step to achieve independence of code inspection tools.

3. We create a graph (SDG) representation of the source code that is inspected. Nodes in the graph represent program locations and edges model control- and data flow.

4. For every warning generated by the inspection tool, the following steps are taken:

   (a) Based on the reported source location, the analyzer looks for the corresponding vertex in the graph.

   (b) It then proceeds to calculate the execution likelihood of this location/vertex by analyzing the structure of the graph, and annotates the warning with this likelihood.

5. The inspection results are ordered by execution likelihood, possibly incorporating external hints such as severity levels or z-ranking results.

System Dependence Graph Central to our approach is the computation of the execution likelihood of a certain location in the program. In other words, we need to find all possible execution paths of the program that include our location of interest, and make predictions for the conditions (or branches) found along those paths. To this end, we use the program’s System Dependence Graph (SDG) [21], which is a generalization of the Program Dependence Graph (PDG).

In short, the PDG is a directed graph representing control- and data dependences within a single routine of a program (i.e. intraprocedural), and the SDG ties all the PDGs of a program together by modeling the interprocedural control- and
data dependences. A PDG holds vertices for, amongst others, assignment statements, control predicates and call sites. Conditions in the code are represented by one or more control points in the SDG. In addition, there is a special vertex called entry vertex, modeling the start point of control for a function. In the remainder, we will use the terms vertex, program point and location interchangeably. The edges between nodes represent the control- and data dependences. Our approach currently does not consider information from dataflow analysis, so we limit our discussion to control dependences. The most relevant causes for these dependencies are:

- there is a control dependence between a predicate vertex \( v \) and a second vertex \( w \) if the condition at \( v \) determines whether execution reaches \( w \);
- there is a control dependence between a function’s entry point and its top-level statements and conditions;
- there is a control dependence between a call site and its corresponding function entry point.

Clearly, we can find all possible acyclic execution paths by simply traversing the SDG with respect to these control dependences. However, traversing the complete SDG to find all paths to a single point is not very efficient. To better guide our search, we base our traversals on program slicing.

Slicing The slice of a program \( P \) with respect to a certain location \( v \) and a variable \( x \) is the set of statements in \( P \) that may influence the value of variable \( x \) at point \( v \). Although we are not actually interested in dataflow information, this slice must necessarily include all execution paths to \( v \), which is exactly what we are looking for. By restricting ourselves to control flow information, we can rephrase the definition as follows: the control-slice of \( v \) in \( P \) consists of all statements in \( P \) that determine whether execution reaches \( v \). Calculating the execution likelihood of \( v \) is now reduced to traversing all paths within this slice, and this forms the basis for the algorithm discussed next.

4. Execution Likelihood ANalysis

This section will introduce the algorithm calculating the execution likelihood for a single program point. Given the SDG of a project, computation entails traversing the graph in reverse postorder, obtaining probabilities by predicting branch probabilities and combining all the paths found from the main entry point to our point of interest. For simplicity, we assume that the project contains a main function that serves as a starting point of execution, although this is not a strict prerequisite, as we will see later on.

Basic algorithm To calculate the execution likelihood \( e_v \) for a vertex \( v \) in a programs SDG \( P \), we perform the following steps:

1. Let \( B_v \) be the control-slice with respect to \( v \). The result is a subgraph of \( P \) that consists of the vertices that influence whether control reaches \( v \).

2. Starting from the main entry point \( v_s \), perform a depth-first search to \( v \) within \( B_v \), enumerating all the paths (sequences of vertices) to \( v \). This is a recursive process; traversal ends at \( v \), then the transition probabilities are propagated back to \( v_s \), where the transition probability \( p_{w,v} \) is the a posteriori probability that execution reaches \( v \) via \( w \). For any given vertex \( w \) visited within the traversal, this is calculated in the following manner:

   - (a) If \( w \) is \( v \), skip all steps, \( p_{v,v} \) is 1.
   - (b) For every control-dependence successor \( s \) of \( w \) in \( B_v \), determine \( p_{s,v} \).
   - (c) Determine the probability that control is transferred from \( w \) to any of its successors \( s \). We do this by first grouping the probabilities \( p_{w,v} \) by the label of the edge needed to reach \( s \) from \( w \). E.g., when \( w \) represents the condition of an if statement, we group probabilities of the true and false branches of \( w \) together. For every group we determine the probability that at least one of the paths found is taken, we denote this set \( S_w \).

   - (d) If \( w \) is not a control point, there will be just one element in \( S_w \), and its probability is \( p_{w,v} \).
   - (e) If \( w \) is a multiway branch (switch), all its cases are thought to be equally likely, and as such \( p_{w,v} \) can be obtained by adding all probabilities in \( S_w \) and dividing them by the number of cases.

   - (f) If \( w \) represents the condition of an if statement, we consider this a special case of a switch: each of its branches is thought equally likely, so \( p_{w,v} \) is obtained by adding both elements of \( S_w \) and dividing them by 2.

   - (g) If \( w \) is a loop, it is assumed that the loop will be executed at least once. \( S_w \) consists of one element representing the probability of the loop body, and \( p_{w,v} \) is taken to be equal to this value.

3. When recursion returns at our starting point \( v_s \), we have calculated the transition probability from \( v_s \) to \( v \), which is our desired execution likelihood \( e_v \).

As stated earlier, the algorithm is written with the computation of execution likelihood in a project with a single startpoint of execution (i.e. the main function) in mind. If we are dealing with a partially complete project, or we are in any other way interested in the execution likelihood with a different starting point, the approach can be easily modified to suit that purpose. Instead of using program slicing, we use a related operation called program chopping. The chop of a program \( P \) with respect to a source element \( s \) and a target element \( t \) gives us all...
elements of $P$ that can transmit effects from $s$ to $t$. Notably, the chop of $P$ between its main entry point and any other point $v$ is simply the slice of $P$ with respect to $v$. Notice that if we let $B_v$ be the chop with respect to $v$, and $v$, we can take any starting point $v_s$ and end up with the desired conditional execution likelihood.

While traversing the graph, transition probabilities for the paths taken are cached. As such, when traversing towards $v$, for a given $w \in B_v$ we only need to calculate $p_{w,v}$ once. This approach necessarily only works within one traversal, i.e., when computing the execution likelihood of one point, because the prequel to some subpath may differ between traversals. However, when computing the likelihood for multiple locations within one program in a row, it is likely that at least part of the traversal results can be reused. For any point $v$ in a procedure $f$, we can split the transition probabilities into one from $v_s$ to the entry point $s_f$ of $f$, and the transition probability from $s_f$ to $v$. Effectively, this means that for any point in $f$ we only compute $p_{v_s, s_f}$ once.

Another important contributor to performance is that our algorithm is demand-driven: it only computes execution likelihoods for locations of interest, instead of computing results for every location in the program. Given that the number locations with issues reported by an inspection tool will typically be much smaller than the total number of vertices in the graph, this is a sensible choice. Together with our deliberately simple heuristics, it forms the basis for a scalable approach.

**Refined branch prediction heuristics** To gain more insight into the performance/accuracy trade-off, we extend the algorithm with some of the branch prediction heuristics of Ball and Larus [1], used in the manner discussed by Wu and Larus [41]. They tested the heuristics empirically and used the observed accuracy as a prediction for the branch probability. For example, they observed that the value check heuristic predicts 'branch not taken' accurately 84% of the time. Therefore, when encountering a condition applicable to this heuristic, 16 and 84 are used for the 'true' and 'false' branch probabilities, respectively. Whenever more than one heuristic applies to a certain control point, the predictions are combined using the Dempster-Shafer theory of evidence [19], a generalization of Bayesian theory that describes how several independent pieces of information regarding the same event can be combined into a single outcome.

The heuristics replace the simple conventions used in steps (d) through (g) discussed above. The behavior with regard to multiway branches has not been changed, and in cases where none of the heuristics apply the same conventions are used as before. A brief discussion of the application of the different heuristics follows below, their associated branch prediction probabilities can be found in table 2.

<table>
<thead>
<tr>
<th>Project Name</th>
<th>ncKLoC</th>
<th># nodes in SDG</th>
<th># non-global nodes in SDG</th>
<th># CPoints in SDG</th>
<th>C/ln</th>
<th>CPointLoC</th>
<th>ELAN accuracy (w. 5% margin)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>avg</td>
<td>stddev</td>
<td>max</td>
<td>avg</td>
<td>stddev</td>
<td>max</td>
<td>'is executed'</td>
</tr>
<tr>
<td>Lame</td>
<td>3</td>
<td>10368</td>
<td>5022</td>
<td>138</td>
<td>29.3</td>
<td>33.4</td>
<td>1064</td>
</tr>
<tr>
<td>Link</td>
<td>8</td>
<td>30422</td>
<td>10149</td>
<td>769</td>
<td>9.8</td>
<td>18.2</td>
<td>412</td>
</tr>
<tr>
<td>Antword</td>
<td>17</td>
<td>88766</td>
<td>33647</td>
<td>3009</td>
<td>5.6</td>
<td>9.4</td>
<td>358</td>
</tr>
<tr>
<td>Chktex</td>
<td>24</td>
<td>119391</td>
<td>53771</td>
<td>2787</td>
<td>8.0</td>
<td>11.0</td>
<td>599</td>
</tr>
<tr>
<td>Chktex</td>
<td>53</td>
<td>93812.2</td>
<td>38917.5</td>
<td>569</td>
<td>7.9</td>
<td>25.9</td>
<td>934</td>
</tr>
<tr>
<td>Infeasit</td>
<td>67</td>
<td>608975.7</td>
<td>92659.9</td>
<td>7678</td>
<td>5.9</td>
<td>11.1</td>
<td>567</td>
</tr>
</tbody>
</table>

Table 1. Case study programs, their metrics and accuracy of ELAN predictions w.r.t. runtime data

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>Probability</th>
<th>Heuristic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loop branch</td>
<td>0.88</td>
<td>Pointer</td>
<td>0.40</td>
</tr>
<tr>
<td>Value check</td>
<td>0.16</td>
<td>Loop exit</td>
<td>0.20</td>
</tr>
<tr>
<td>Return</td>
<td>0.28</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Heuristics and associated probabilities

Implementation ELAN has been implemented as a plugin for Codesurfer, a program analysis tool that can construct dependence graphs for C and C++ programs. As our approach is

---

6. www.grammatech.com
5. Experiments

This section reports on experiments designed to evaluate the accuracy and performance of our technique. Recall that the execution likelihood predictions are used in ranking different locations; we therefore compare rankings based on predictions with rankings based on measurements in actual program runs. Apart from this ordering, we also evaluate accuracy by checking correspondence of the actual prediction values with their measured counterparts. Both experiments are discussed in section 5.1; in section 5.2 we analyze performance by timing the aforementioned experiments and relate the analysis speed to the size of the SDG involved.

5.1. Correlating predictions with runtime

We will start our discussion by reviewing the experimental setup. Table 1 lists some source code properties for the different case studies, respectively the size in lines of code (LoC, not counting comment or empty lines), the size of the SDG in vertices, the size of the SDG without vertices for global parameters, the total number of control points in the program, the cyclomatic complexity (per function), and the control points related to number of lines in the program (again per function). The apparent discrepancy between the size in LoC and the size of the SDG in vertices can be largely attributed to Codesurfer’s modeling of global variables in the SDG. These are filtered out in the 'non-global' column, which shows a better correspondence with the size in LoC.

The programs were picked such that it would be easy to construct 'typical usage' input sets, and automatically perform a large number of test runs. For every case, at least 20 different test runs were recorded. Every one of the programs was subjected to the following steps:

1. Build the project using Codesurfer. This involves the normal build and building the extra graph representations used by our technique.

2. Build the project using gcc’s profiling options, in order to obtain profiling information after program execution.

3. Run the ELAN algorithm for every control-point vertex in the project. This will give a good indication of analysis behavior distributed throughout the program (since it approximates predictions for every basic block). Moreover, this step was performed twice: with and without branching heuristics.

4. Gather a small dataset representing typical usage for the project, and run the program using this dataset as input. For all the program locations specified in step 3, determine the percentage of runs in which it was visited at least once. This last step uses gcov, which post-processes the profile data gathered by gcc’s instrumentation.

5. Create two sets of program locations, the first sorted by prediction, the second by actual usage, and compare them using Wall’s unweighted matching method [37]. This will give us a correlation score for different sections of the two rankings.

To illustrate this correlation score, consider the following example: suppose we have obtained the two sorted lists of program locations, both having length \( N \), and we want to know the score for the topmost \( m \) locations. Let \( k \) be the size of the intersection of the two lists of \( m \) topmost locations. The correlation score then is \( k/m \), where 1 denotes a perfect score, and the expected score for a random sorting will be \( m/N \).

In our experiments, scores were calculated for the topmost 1%, 2%, 5%, 10%, 20%, 40% and 80%. The correlation scores for the different programs can be found in table 3, where the first number in every cell is the score without the use of heuristics, and the second is obtained using heuristics.

When considering the ranking data in table 3, we will denote the upper \( n \)% of the ranking as block \( n \). We are first and foremost interested in the correlation scores of the top-most blocks of locations. Typically, a randomized ranking would produce correlation scores of 0.01, 0.02 and 0.05 for these top blocks, which is easily outperformed by the ELAN technique. However, there is a problem in using these correlation scores for evaluating the accuracy of the ranking. Consider Antword, for example: in our test series, 3.8% of the locations involved were always executed, which means that the two topmost blocks in our table will consist entirely of locations with measured value 1. There is no way to distinguish between these locations, and as such, no way to compare rankings. Even if predictions were perfect, locations in our predicted block 1 could be anywhere in block 2 of the other ranking. In

<table>
<thead>
<tr>
<th>Portion</th>
<th>Antword</th>
<th>Chktex</th>
<th>Lame</th>
<th>Link</th>
<th>Uni2Ascii</th>
<th>TV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>0.85 .45</td>
<td>0.25 .50</td>
<td>0.14 .14</td>
<td>0.22 .22</td>
<td>1.0 - 1.0</td>
<td>0 .13 - 0.13</td>
</tr>
<tr>
<td>2%</td>
<td>0.44 .44</td>
<td>0.22 .33</td>
<td>0.11 .11</td>
<td>0.11 .11</td>
<td>1.0 - 1.0</td>
<td>1.0 - 1.0</td>
</tr>
<tr>
<td>5%</td>
<td>0.61 .61</td>
<td>0.30 .26</td>
<td>0.25 .24</td>
<td>0.30 .29</td>
<td>0.67 - 0.67</td>
<td>0.15 - 0.15</td>
</tr>
<tr>
<td>10%</td>
<td>0.50 .46</td>
<td>0.34 .33</td>
<td>0.26 .26</td>
<td>0.39 .35</td>
<td>0.29 - 0.29</td>
<td>0.17 - 0.17</td>
</tr>
<tr>
<td>20%</td>
<td>0.32 .45</td>
<td>0.24 .26</td>
<td>0.34 .34</td>
<td>0.57 .57</td>
<td>0.67 - 0.67</td>
<td>0.35 - 0.35</td>
</tr>
<tr>
<td>40%</td>
<td>0.52 .49</td>
<td>0.40 .41</td>
<td>0.46 .46</td>
<td>0.60 .61</td>
<td>0.57 - 0.57</td>
<td>0.63 - 0.64</td>
</tr>
<tr>
<td>80%</td>
<td>0.78 .79</td>
<td>0.76 .76</td>
<td>0.83 .83</td>
<td>0.83 .83</td>
<td>0.79 - 0.79</td>
<td>0.96 - 0.96</td>
</tr>
</tbody>
</table>

Table 3. Correlation scores
Table 4. Average measured execution likelihood

<table>
<thead>
<tr>
<th>Portion</th>
<th>Antiword</th>
<th>Chktex</th>
<th>Lame</th>
<th>Link</th>
<th>Univ2Ascii</th>
<th>TX</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>71.4-75</td>
<td>74.6-80</td>
<td>74-75</td>
<td>87.9-91</td>
<td>1.0-1.6</td>
<td>30-33</td>
</tr>
<tr>
<td>2%</td>
<td>73.8-89</td>
<td>89.9-92</td>
<td>35-47</td>
<td>89-94</td>
<td>1.0-1.6</td>
<td>23-23</td>
</tr>
<tr>
<td>5%</td>
<td>68-83</td>
<td>83.7-78</td>
<td>35-36</td>
<td>89-99</td>
<td>1.0-1.6</td>
<td>16-16</td>
</tr>
<tr>
<td>10%</td>
<td>65-72</td>
<td>72.7-72</td>
<td>40-43</td>
<td>89-99</td>
<td>1.0-1.6</td>
<td>13-14</td>
</tr>
<tr>
<td>20%</td>
<td>49-59</td>
<td>62-61</td>
<td>39-40</td>
<td>81-82</td>
<td>87-87</td>
<td>08-08</td>
</tr>
<tr>
<td>40%</td>
<td>40-40</td>
<td>41-41</td>
<td>30-29</td>
<td>70-72</td>
<td>57-57</td>
<td>06-06</td>
</tr>
<tr>
<td>80%</td>
<td>28.8-47</td>
<td>47-46</td>
<td>35-35</td>
<td>57-56</td>
<td>44-44</td>
<td>09-09</td>
</tr>
</tbody>
</table>

In table 4, the average execution likelihood for all the locations in a certain block are shown; again, the values represent the results without heuristics and with heuristics, respectively. We took the ranking based on our computed likelihood and calculated the average measured execution likelihood. The values shown in table 4 cannot be compared amongst programs, as they depend on the runtime behavior of the individual program. For example, consider the differences between values for Antiword, which has 3.8% of its locations always executed, and Chktex, which has 33.5% of its locations always executed. What does matter, however, is the distribution within one program: we expect the locations ranked higher to have a higher actual execution likelihood, and, with some exceptions, exactly this correlation can be observed here.

Finally, we illustrate overall accuracy by means of the relation between error margins and accurate predictions, using a diagram similar to the ones in [26]. However, instead of the number of branches predicted accurately, the figure displays the number of locations for which the execution likelihood was predicted accurately within the given error margin. Moreover, the set of test programs used for evaluation differs.

5.2. Performance measurements

The benchmark set used in the previous section consists of programs of different size (cf. table 1), which helps to understand the scalability of the approach. Recall our algorithm, which computes a slice, traverses the subgraph obtained, and derives predictions for conditions. This signifies the importance of the size of the SDG, rather than the number of KLoC. This relationship is illustrated in figure 3. Timing measurements were taken for every experiment in section 5.1, and we calculated the average time taken per location. All measurements were performed on a laptop with an Intel Pentium Mobile 1.6Ghz and 512Mb memory running MS Windows XP Pro.

Also, to understand the effect of the caching mechanism introduced in section 4, we performed the analysis for different subsets of the complete set of program points involved in the earlier experiments. Specifically, we took subsets of differing sizes (e.g. 10%, 20% of the original size) by sampling the original set uniformly, and measured the analysis running time for these subsets. The resulting measurements are displayed in figure 4. Typically, we expect a relation like this to be linear, and the fact that the slope of the curve is decreasing shows that the caching mechanism is indeed doing its job.

6. Evaluation

Accuracy When looking at the data of the accuracy experiments, perhaps most striking are the differences between tables 3 and 4. Even though locations with a higher execution likelihood in general seem to be ranked higher, the correlation scores resulting from the comparison with the ranking based on those measured values simply do not measure up. To understand this, we need to look at the locations that will end up high in the ranking: in our test set, the percentage of loca-
There are a number of observations that we can make regarding performance: first of all, the approach seems to scale well to larger software systems, where the version that uses the refined branch prediction heuristics is only slightly outperformed by the simpler one. It goes to show that including more sophisticated analyses can still result in a feasible solution.

Secondly, there is one program that does not conform to this observation: performance on Lame is significantly lower than on any of the others. Manual inspection revealed that the Lame frontend has a function that parses command-line arguments with a great number of short-circuited expressions. Because this occurs early on in the program, it affects many of the locations we are testing. Specifically, this means that computation of 25% of the locations involved requires traversal through this function, and 6.5% of the locations are within this function itself. In a follow-up experiment that left out the suspect function we found that the performance on the rest of the program was according to expectations (30 ms/location). This illustrates the fact that our evaluation experiment is actually somewhat negatively biased, since a relatively large number of locations were taken from a computationally expensive function. However, this kind of biased distribution is unlikely when ordering actual inspection results. In addition, such large short-circuited expressions are atypical for the type of software analysed in the Trader project, leaving little reason for concern at this time.

Finally, note that, because of the caching mechanism, the reported timings are actually worst-case approximations. Because the complete project is covered, adding more warning locations will lead to an efficient use of the cache and therefore a shortened computation time. For instance, one of the benchmarks, involving approximately 2600 points, took 137 seconds. Extending this experiment to an analysis of 13000 points increased the running time to slightly under 300 seconds: just over double the time for a fivefold increase in input, showing that prioritizing a large number of warnings remains feasible.

Orthogonality We remark that our approach is orthogonal to other prioritization and filtering techniques discussed in the related work. However, in combination with these approaches, ELAN can best be applied as final step because the filtering of the earlier stages in combination with our demand driven approach effectively reduces the amount of computations that will need to be done.

Applicability In our test set, we use programs that are one-dimensional in their tasks, i.e. perform one kind of operation on a rather restricted form of input. This limits issues related to the creation of appropriate test inputs, and allows us to focus on evaluating the approach itself. It does mean, however, that we must devote some time to the question how to generalize these results to other kind of programs.

The ELAN approach is based on information implicit in the control structure of the program, and as for the heuristics, in the way humans tend to write programs. This information will always be present in any program. However, there may be
parts of the control structure that are highly dependent on interaction or inputs. Fisher and Freudenberg observed that, in general, varying program input tends to influence which parts of the system will be executed, rather than influencing the behavior of individual branches [16]. This suggests that, typically, there are a number of highly data-dependent branches early on in the program, while the rest of the control structure is rather independent. For example, a command-line tool may have a default operation and some other modi of operation that are triggered by specifying certain command-line arguments. At some point in this program, there will be a switch-like control structure that calls the different operations depending on the command-line arguments specified. This control structure is important as it has a major impact on the rest of the program, and it is also the hardest to predict due to its external data dependence. However, this information (in terms of our example: which operation modi are most likely to be executed) is exactly the type of information possessed by domain experts such as the developers of the program. Therefore, the simple extension of our approach with a means to specify these additional (input) probabilities can further improve applicability to such situations.

7. Concluding Remarks

Contributions We present a method for the prioritization of software inspection results based on statically computing the likelihood that program execution reaches locations for which issues are reported, i.e., we prioritize code inspection results using static profiling. We discuss a novel demand-driven algorithm for computing execution likelihood based on the system dependence graph and we present and evaluate a number of optimizations that further increase accuracy and performance.

We investigate the feasibility of the described approach using a number of case studies in which a prototype tool was developed and applied to several open source software systems and software embedded in a Philips television. We investigate the correlation between our static predictions and actual execution data found by dynamic profiling and we report on the performance of our approach. This empirical validation shows that the approach is capable of correctly prioritizing program locations. In addition, the approach scales well to larger systems.

Future Work Together with Philips Semiconductors, we are preparing a larger case study, in which our technique will be evaluated during development and inspection of software to be embedded in a new digital TV.

Moreover, we want to experiment with a number of ideas to further improve our approach by incorporating more advanced program analysis techniques, such as range propagation [26], that are basically aimed at enabling better estimations of the outcomes of conditions. Also, they can be used to compute execution frequencies, which will benefit the ranking by better distinguishing locations at the top of the ranking. However, since such analyses typically come with additional computational costs, we want to investigate if the improved accuracy actually warrants the expenses involved.

References