The Theory of Relative Dependency: Higher Coupling Concentration in Smaller Modules

A. Güneş Koru, University of Maryland, Baltimore County
Khaled El Emam, University of Ottawa

Our recent research on several large-scale software products has consistently shown that smaller modules are proportionally more defect prone.\textsuperscript{1–3} These findings challenge the common recommendations from the literature suggesting that quality assurance (QA) and quality control (QC) resources should focus on larger modules. Those recommendations are based on the unfounded assumption that a monotonically increasing linear relationship exists between module size and defects.\textsuperscript{1–4} Given that complexity is correlated with size,\textsuperscript{5} following such recommendations ensures that the more complex modules receive greater scrutiny. However, our recent findings imply that, given limited and a fixed amount of resources, focusing on the smaller modules would lead to more effective defect detection.\textsuperscript{1–3} So, it’s important to understand the mechanisms behind those findings to make the case for amending current practices.

Research has shown that higher module dependencies, also called higher coupling, result in more defects by increasing the likelihood of interface defects.\textsuperscript{6–10} “Coupling” as we use it here corresponds to the concept of fan-out (depending on other modules). It doesn’t mean fan-in (being depended upon by other modules) because the literature shows no consistent evidence that a module’s fan-in is related to its defect-proneness. Relying on the evidence from the literature, we formulated a testable hypothesis of relative dependency (RD) for software modules:

\[ H_{RD}: \text{Smaller modules are proportionally more coupled.} \]

\[ H_{RD} \] requires us to study the size-coupling relationship for software modules. So far, the empirical evidence has shown a positive correlation between module size and coupling because of their monotonically increasing relationship.\textsuperscript{11–13} However, those findings aren’t enough to test \[ H_{RD} \] because correlations don’t imply proportionality between the two variables.

To investigate proportionality, we performed an empirical study of data from multiple systems. The results confirmed \[ H_{RD} \] and led us to formulate the theory of relative dependency:
In large-scale software systems, smaller modules will be proportionally more dependent compared to larger ones.

This theory, along with our findings, has important implications for software practice, especially for projects involving considerable refactoring.

### Methods

We performed measurements and tested $H_{RD}$ on 14 open source object-oriented (OO) products from different development teams: the Mozilla 1.0 Web browser and 13 products in KOffice 1.6.3, an office suite. The release dates for Mozilla 1.0 and KOffice 1.6.3 were 5 June 2002 and 7 June 2007, respectively. Table 1 lists the products and their functionality; the list of products covers a wide range of functionalities and measurement characteristics.

### Data Collection

We measured size in physical LOC excluding blank lines and lines including only comments. To measure coupling, we used two OO metrics defined by Shyam Chidamber and his colleagues:

- **Coupling between object classes (CBO)** is the number of other classes whose methods or attributes are used.
- **Depth of inheritance tree (DIT)**

### Table 1

The analyzed products, their functionality, and descriptive statistics for their size and coupling data

<table>
<thead>
<tr>
<th>Product</th>
<th>Functionality</th>
<th>No. of classes</th>
<th>LOC</th>
<th>Coupling between object classes (CBO)*</th>
<th>Depth of inheritance tree (DIT)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mozilla</td>
<td>Web browser</td>
<td>178</td>
<td>284</td>
<td>362</td>
<td>158</td>
</tr>
<tr>
<td>KWord</td>
<td>Word processing</td>
<td>971</td>
<td>124</td>
<td>500</td>
<td>990</td>
</tr>
<tr>
<td>KSpread</td>
<td>Spreadsheet</td>
<td>178</td>
<td>284</td>
<td>362</td>
<td>158</td>
</tr>
<tr>
<td>KPresenter</td>
<td>Presentation</td>
<td>178</td>
<td>284</td>
<td>362</td>
<td>158</td>
</tr>
<tr>
<td>Qoox</td>
<td>Data management</td>
<td>628</td>
<td>130</td>
<td>199</td>
<td>235</td>
</tr>
<tr>
<td>KPlato</td>
<td>Project management</td>
<td>218</td>
<td>57</td>
<td>23</td>
<td>185</td>
</tr>
<tr>
<td>Kivio</td>
<td>Diagramming</td>
<td>130</td>
<td>5</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Krita</td>
<td>Painting and image editing</td>
<td>700</td>
<td>3</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>KChart</td>
<td>Chart drawing</td>
<td>85</td>
<td>9</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Karbon</td>
<td>Scalable-vector drawing</td>
<td>130</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>KFilters</td>
<td>Conversion between file formats</td>
<td>699</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>KO Library</td>
<td>KOOffice library</td>
<td>55</td>
<td>15</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>KFor Formula</td>
<td>Formula editor for KOOffice</td>
<td>17</td>
<td>8</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

*CBO indicates the number of other classes whose methods or attributes are used; DIT indicates the maximum depth of a class in the inheritance tree from the root class.*
Depth of inheritance tree (DIT) is the maximum depth of a class in the inheritance tree from the root class.

We considered only application (product) classes, not those in the C++ libraries. Empirical studies of software engineering have commonly used all three metrics.

In OO programming, classes are considered to be logically cohesive development units, or modules, organized in an inheritance hierarchy. So, we must consider DIT to better understand class dependencies. Indeed, Chidamber and his colleagues stated that DIT is “a measure of how many ancestor classes can potentially affect this class,” which exactly describes the dependencies of a child class to its ancestors.

We performed the measurements at the class level using a mature static-code-analysis tool, Understand for C++ (www.scitools.com). For each class, we took into account its header and implementation files during the size measurements. As a result of our measurements, we created a distinct data set for each product, with the data points corresponding to the C++ classes in the product. Each data point included three numeric values for LOC, CBO, and DIT.

Table 1 includes the descriptive statistics for the data sets for the products in this study.

Data Analysis
For each product, we plotted a concentration curve and produced a concentration index (C) to visualize and quantify the inequality of CBO concentration with respect to module size. We repeated this process for DIT. For more on concentration curves and indices, see the “Using Concentration Curves and Indices” sidebar.

Results
Figure 1 displays the concentration curves for all the products. All the curves except the CBO curve for KFormula are above the equality line, clearly showing that coupling is concentrated in smaller modules. For example, in KChart, the smallest module sizes adding up to 20 percent of the total size had more than 50 percent of the total CBO. In all plots, the inequalities with respect to size were even greater for DIT. For example, in KSpread, the source lines in the smallest modules adding up to 20 percent of the total size had almost 80 percent of the total DIT.

Figure 2 shows bar plots for the concentration indices. The estimate of C is the value on the y-axis corresponding to the bar’s center. A bar’s length shows the 95 percent confidence interval. For ease of reference, the figure also shows the numeric values of C and its confidence intervals.

Figure 2 shows that, for all products except KFormula, not only the point estimates of C but also the 95 percent confidence intervals stay above the zero line (the dashed horizontal line), showing statistical significance. For KFormula, the point estimate for C is also positive; however, its confidence interval reaches below zero. This is understandable because KFormula is a small product providing only 17 data points, which makes it difficult to reject the equality of CBO concentration (C = 0) with enough confidence.

Figure 2 presents compelling quantitative evidence that, in general, coupling is concentrated more in the lines of source code in smaller modules. So, smaller modules are proportionally more coupled than larger modules. This strongly supports $H_{RD}$.

How Code Refactoring Affects Coupling Concentration
Refactoring basically means improving software design without changing its functionality.15 Martin Fowler and his colleagues stress the necessity of refactoring because software design typically decays over time as developers add or change functionality.15 So, refactoring is inherent in software design and maintenance and is particularly favored in agile development.

When discussing “code smells,” Fowler and his colleagues stress that the first three smells in the “stink parade” are duplicated code, long method, and large class.15 So, the expert advice directly suggests focusing refactoring first (or more) on larger modules. Some of the other code smells such as feature envy and shotgun surgery recommend focusing on tightly coupled classes to reduce their dependency, which is in concordance with the traditional advice given to practitioners. Yoshio Kataoka and his colleagues reported that experts’ subjective judgment on refactoring’s effectiveness correlated with coupling reductions.16

Indirectly, more emphasis on highly coupled modules also means more emphasis on larger modules, for two reasons:

- As we mentioned before, a well-established correlation exists between size and coupling.
- Mika Mäntylä and his colleagues examined the correlations between code smells.17 They
found that the Spearman correlation between large class (which is concerned with size) and feature envy (which is concerned with coupling) was 0.59 and significant at $\alpha = 0.01$. We also examined whether refactoring tools' smell detection rules detect larger classes. Borland Together (www.borland.com/us/products/together), a Java IDE, includes a refactoring tool.

Figure 1. Concentration curves for coupling between object classes (CBO) and depth of inheritance tree (DIT), for all products tested. The $x$- and $y$-axes represent the cumulative proportion of total LOC and CBO or DIT, respectively. The curves are plotted for unique class sizes sorted from the smallest to the largest and marked along the $x$-axis from left to right. All the curves except KFormula's CBO curve are above the equality line, showing that coupling is concentrated in smaller modules.
Because the tool supports smell detection for only Java programs, we detected the code smells in JBoss (www.jboss.org), an open source Java application server. (JBoss isn’t in Table 1 because Understand for C++ analyzed C++ source code.) Borland Together detected 7,991 classes in JBoss.

Table 2 shows that, for each smell other than refused bequest, a statistically significant size difference existed between the classes having the smell and those not having it. For the rest of the code smells, the percentiles for median size rankings were very high, directing developers’ attention to larger modules.

All this evidence suggests that larger classes will more likely be refactored.

Fortunately, we were also able to investigate the effect of refactoring on coupling concentration empirically. In the Kivio and Karbon projects, the developers were specifically requested to perform refactoring before an upcoming big release, KOOffice 2.0. (We archived this announcement at www.webcitation.org/5SrjtxRoj.) At the time of writing, the fourth alpha release of KOOffice 2.0 had occurred. When we examined the Kivio sources for this release and compared them with the 1.6.3 release, we saw that Kivio was almost unchanged, showing no signs of refactoring. However, for Karbon, we observed that refactoring had indeed taken place in the alpha release.

The Karbon 1.6.3 release (before refactoring)
Table 2

<table>
<thead>
<tr>
<th>Code smell</th>
<th>Number of classes with this smell</th>
<th>Statistical size comparison with classes not having the smell (p-value from the Wilcoxon rank-sum test)</th>
<th>Median size ranking for classes having this smell (% with 100% being the largest size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duplicated code</td>
<td>64</td>
<td>&lt;2.2e-16</td>
<td>95.95</td>
</tr>
<tr>
<td>Long method</td>
<td>226</td>
<td>&lt;2.2e-16</td>
<td>95.84</td>
</tr>
<tr>
<td>Large class</td>
<td>59</td>
<td>&lt;2.2e-16</td>
<td>98.60</td>
</tr>
<tr>
<td>Shotgun surgery</td>
<td>199</td>
<td>&lt;2.2e-16</td>
<td>84.88</td>
</tr>
<tr>
<td>Feature envy</td>
<td>48</td>
<td>&lt;2.2e-16</td>
<td>94.14</td>
</tr>
<tr>
<td>Message chains</td>
<td>90</td>
<td>4.441e-16</td>
<td>80.23</td>
</tr>
<tr>
<td>Refused bequest</td>
<td>221</td>
<td>0.60</td>
<td>49.59</td>
</tr>
<tr>
<td>Data class</td>
<td>25</td>
<td>6.286e-06</td>
<td>74.57</td>
</tr>
</tbody>
</table>

*We obtained the median ranking after assigning a size rank to each class (we computed an average rank value for equal rankings) and then finding the median-rank value for the sizes having the code smell.

had 183 classes with 26,468 total LOC, with a median class size of 92 LOC. Ninety-six classes were changed, 48 were added, and 66 were deleted. After refactoring, Karbon had 163 classes with 20,617 total LOC, with a median class size of 81 LOC. These numbers show that refactoring affected much of the system. In the resulting alpha version, the newly added classes were significantly smaller than the others (the Wilcoxon rank-sum test gives p = 0.005). No size difference existed be-
correspond to some welfare categories—for example, the income categories for neighborhoods described as very poor, poor, middle-class, rich, and very rich. In the grouped case, the group sums are used to calculate the proportions of people and ill health.

By definition, the concentration index (denoted with \( C \)), which is the measure of inequality, is twice the area between the concentration curve and the equality line. This area is the shaded part of Figure A. The socioeconomic-related inequality of ill health increases as the shaded area increases. \( C \) becomes zero when the concentration curve exactly overlaps with the equality line. It can take any value between \(+1\) and \(-1\). In our research, when the curve rises above the equality line, \( C \) increases toward \(+1\); it decreases toward \(-1\) when the curve drops below the equality line. (In health-related fields, \( C \) usually decreases when the curve rises above the equality line because ill health for the disadvantaged [or poor] is a more common and undesirable phenomenon. We followed the opposite of this convention to simplify the presentation in the main article.)

For the products in our study, we plotted the cumulative proportion of total coupling (on the y-axis) versus the cumulative proportion of total lines of source code (on the x-axis) grouped and ordered by class size. That is, we adopted the grouped case we just explained and used class size (in a way similar to the welfare categories) to group lines of source code. In this scheme, each unique class size formed a size category. Then, we ordered the size categories from the smallest to largest and calculated the cumulative proportions for each category to draw the concentration curves and to calculate \( C \). We calculated the variance and confidence interval for \( C \) using the method detailed by Nanak Kakwani and his colleagues.\(^1\)

When this process is followed, if our hypothesis is true, coupling should be concentrated in the source lines in smaller modules, and the concentration curves obtained should resemble the one in Figure A rising above the equality line. Such a curve’s \( C \) value should be positive, and, if the concentration effect is significant, the 95 percent confidence interval for \( C \) should remain above zero.

To summarize, the concentration curve lets us visualize the nonlinear relationship between module size and coupling. The concentration index, \( C \), quantitatively summarizes the visual information that curve presents. The sign of \( C \) indicates the coupling concentration’s direction (that is, positive when coupling is concentrated in the source code lines in smaller modules, and vice versa). Its absolute value shows the concentration’s strength; its statistical significance is important and should be calculated.

**Reference**


---

Refactoring didn’t affect DIT values significantly. However, among the modified classes, CBO increased proportionally more for smaller classes. The CBO ratio (CBO after refactoring divided by CBO before refactoring) decreased as class size increased, giving a Spearman’s correlation of \(-0.32\), which was highly significant (\( p = 0.002 \)). The CBO concentration index was 0.18 before refactoring and 0.23 after it, showing that coupling was concentrated even more in smaller classes after refactoring.

So, overall, the Karbon analysis shows that refactoring

- produced more classes that were smaller and
- disproportionately increased coupling for smaller classes.

Refactoring activities seem likely to increase the concentration of coupling in smaller modules. Given that refactoring is continuous, especially for agile development projects,\(^1\) this result suggests that systems developed with agile methods will likely face an increasing concentration of coupling in smaller modules. So, giving a higher priority to smaller modules for defect detection will be even more important. As always, in such decisions, developers should also consider other available information about software modules such as their fan-in, business importance, and operational profile, as we stressed in our earlier research.\(^1\)-\(^3\)

**Limitations**

As with any empirical study, our study has limitations. Certainly, it is possible to use or derive some metrics other than LOC, CBO, and DIT. However, software researchers have widely adopted, validated, and accepted these metrics. Future studies could collect additional measures of size and coupling to compare their results with ours.

Our results are limited to the concentration of dependencies over source code lines with respect to module size. Future studies could also investigate the effect of refactoring on the size-coupling relationship in more depth and whether some of
Basili and Perricone’s Work on Module Size and Defect Density

Victor Basili and Barry Perricone studied defect density for different size groups and observed higher defect density for smaller modules.1 However, their analysis suffered from artificial ratio correlations.2,3 When defect density is plotted against size or tabulated against size groups, even randomly generated data will show higher defect density for smaller modules because size is in the denominator of the derived metric, defect density.

Nevertheless, Basili and Perricone concluded that smaller modules are proportionally more problematic; our recent findings4–6 concur with their observation. Then, they speculated that the equal distribution of interface defects over smaller and larger modules could have been responsible for their finding, which implies that smaller modules had proportionally more interface defects. Their definition of interface defects was “those (defects) that were associated with structures existing outside the module’s local environment but which the module used.”1 This definition exactly describes the defects associated with coupling. So, our findings in the main article about a higher concentration of coupling in smaller modules shows that Basili and Perricone’s speculation was plausible.

References

Our study has two implications for practice.

First, our findings provide a plausible mechanism to explain why smaller modules have proportionally more defects as observed in our earlier studies and in Victor Basili and Barry Perricone’s research (see the “Basili and Perricone’s Work on Module Size and Defect Density” sidebar). Coupling is concentrated more in the smaller modules, making them proportionally more defect-prone. This evidence will be useful to practitioners when they seek resources or support to amend or revise their organizations’ QA and QC practices.

Second, refactoring exacerbates the coupling concentration in smaller modules. So, for projects that refactor extensively, focusing defect detection and correction on smaller modules will increase the QA and QC effectiveness even more.

Certainly, we don’t claim that software developers should avoid refactoring. Refactoring is essential for software development and maintenance because of its system-wide benefits. Our findings are only about the effect of refactoring on coupling concentration with respect to module size, which has implications in terms of the effectiveness and efficiency of focused defect detection activities often performed with limited resources.

Empirical studies investigating our findings and their implications will further contribute to the software field’s knowledge and capacity to develop and maintain high-quality software systems effectively and efficiently.

Acknowledgments
We thank the associate editor and the anonymous reviewers for their useful and constructive feedback. We also thank Vic Basili, Dave Card, Barbara Kitchinham, Tim Menzies, and Carolyn Seaman for their feedback on earlier drafts of this article.

References


