Experiments with Coupling and Cohesion Metrics in a Large System*

Timothy C. Lethbridge and Nicolas Anquetil

School of Information Technology and Engineering
150 Louis Pasteur, University of Ottawa
K1N 6N6 Canada
tcl@site.uottawa.ca, anquetil@csi.uottawa.ca

Abstract

We apply coupling and cohesion metrics to various decompositions of a large telecommunications system. Several findings emerge: 1) There is a baseline level of coupling and cohesion that represents the average connectedness of any pair of files in the system. 2) As would be expected, measures of cohesion are greater than this baseline in well-defined subsystems. 3) Interestingly, and contrary to intuition, measures of coupling for such subsystems tend also to be greater than the baseline and to increase as cohesion increases. 4) It is possible to easily calculate upper bounds for coupling and cohesion that no subsystem in a given system can exceed. 5) Measuring quality by subtracting coupling from cohesion, as has been proposed in the literature, gives anomalous results since coupling and cohesion are frequently not on the same scale. We propose improved coupling, cohesion and quality metrics that normalize for the baseline and ceiling levels in a given system. Using our proposed metrics it should be possible to compare different systems, something that the current metrics do not permit.

1. Introduction

The general objective of our research is to assist software engineers to understand very large software systems that are typically poorly documented and organized. Part of our approach to this problem involves automatically extracting subsystems–Software engineers will use the extracted subsystems (also known as 'clusters') to guide their exploration of an unfamiliar system. For related research see Lakhotia (1997) and Müller et al (1993).

In order to validate our subsystem extraction methods, we wish to evaluate them for quality. We hope to show that the extracted subsystems in some sense encapsulate some useful aspect of the architecture. One approach to such evaluation is to look at coupling and cohesion. In this paper we discuss the use of coupling and cohesion metrics from the literature. We explain several problems with them, particularly the fact that calcuated values do not fall within the same range from system to system.

The next section describes methods we have been using for calculating cohesion and coupling, and a derived measure of quality that is the difference between cupling and cohesion. The subsequent section describes experiments we have performed using these metrics. The remainder of the paper describes problems with the conventional metrics, and proposes improved, normalized, metrics.

* This work is supported by NSERC and Mitel Corporation and sponsored by the Consortium for Software Engineering Research (CSER).

2. Background

2.1 General method of calculating coupling and cohesion

We use the method for calculating coupling and cohesion described in Patel et al (1992) and Kunz and Black (1995). The only difference is that in our case we are clustering files, whereas Kunz and Black were clustering 'processes'. The basis for this method is a function $Sim_W(X,Y)$ that takes two files X and Y, computes vectors X_W and Y_W from these (described below), and computes a value between 0 and 1:

$$Sim_W(X,Y) = \frac{X_W{}^T Y_W}{\|X_W\| \times \|Y_W\|}$$

Several different types W of vectors can be used in the similarity functions; these are described in the next section.

The more similar two characteristic vectors, the more closely related their files are likely to be. We would expect a cohesive subsystem to have all its files closely related. Thus Kunz and Black describe their metric for cohesion as the average similarity of all distinct pairs of files in a cluster P:

$$Coh_W(P) = \frac{\sum_{i>j} Sim_W(p_i, p_j)}{m(m-1)/2}$$

Here, m is the number of files in the cluster P, and each p is a member of cluster P.

Coupling is correspondingly defined as the average similarity of all pairs of files in the system, such that one file in each pair is in the cluster P, and the other file is outside the cluster:

$$Coup_W(P) = \frac{\sum_{i,j} Sim_W(p_i, q_j)}{m \times n}$$

Here, m is the number of files in cluster P, and n is the number of files not in cluster P.

A good subsystem is expected to exhibit high cohesion and low coupling. However, because there are two separate metrics, comparison between subsystems is difficult. If one subsystem has higher coupling (better) and higher cohesion (worse) than the other, it is not clear which one is the best. Kunz and Black defined a metric for the quality of a system as simply the difference between the cohesion and coupling:

$$Q_W(P) = Coh_W(P) - Coup_W(P)$$

2.2 Methods of calculating the similarity

For the method described above, the similarity measure $Sim_W(X,Y)$, depends on how the characteristic vectors are constituted. For our experiments, we used three different methods which we are diffentiating using the W subscript in the above equations:

TR: A characteristic vector counts the references to particular user-defined types within the file it describes. This is the method used by Kunz and Black (1995).

DR: A vector represents the use of named variables in the two files being compared, as defined in Patel et al (1992).

RC: A vector represents the calling of routines in one file by routines in the other and vice-versa. We introduced this method to see whether different ways of computing cohesion and coupling affect our results.

3. Experiments with coupling and cohesion

We performed our experiments on a telecommunications system consisting of about 4500 files. We used the 11 techniques described in table 1 to cluster these files into subsystems; our original objective was to evaluate the effectiveness of these techniques at extracting highly cohesive subsystems which were as loosly coupled with each other as possible. However, in the context of this paper, we became more interested in evaluating the cohesion and coupling metrics themselves.

Description of clustering technique	**Clustering technique number** (where 1 means least cohesive and 11 means most, according to figure 1)
Clusters created manually	
Clusters selected by experts (did not include all files)	6
Clusters recorded in the configuration management system	3
Similar to 3, but only considering clusters containing at least one file from the experts' partition (6)	4
Clusters generated automatically	
.. using similarity based on data references	
Automatic data-reference clustering	7
Similar to 7, but only considering clusters containing at least one file from the experts' partition (6)	11
... using similarity based on routine calls	
Automatic routine-call clustering	8
Similar to 8, but only considering clusters containing at least one file from the experts' partition (6)	10
... using similarity based on abbreviations in file names (Anquetil and Lethbridge 1998)	
Automatic file name clustering	1
Similar to 1, but only considering clusters containing at least one file from the experts' partition (6)	5
Similar to 1, but only considering first abbreviation in each file name	2
Similar to 2, but only considering clusters containing at least one file from the experts' partition (6)	9

Table 1: Eleven clustering techniques used in experiments.

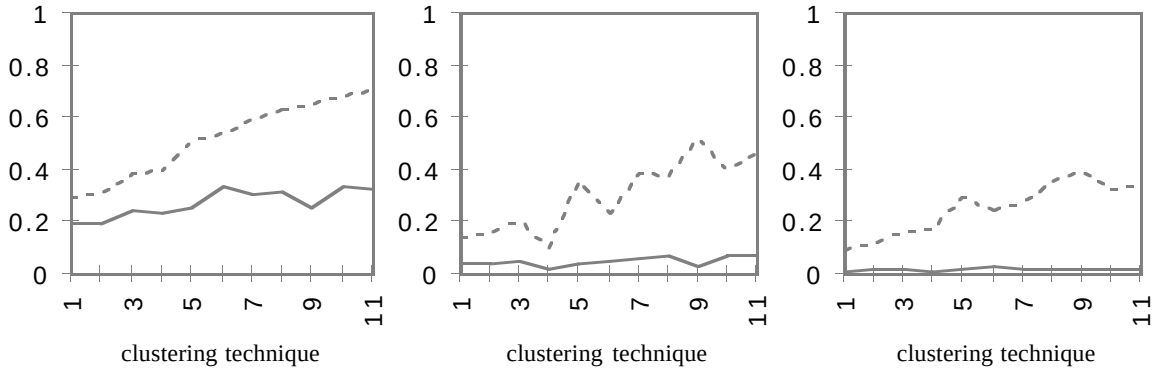

Figure 1: Un-normalized cohesion (dashed line) and coupling (solid line) for similarity methods TR, DR and RC respectively. Clustering techniques from table 1 are shown on the x axis.

Similarity method	Coefficient of linear correlation between Coh() and Coup()
TR	0.86
DR	0.57
RC	0.53

Table 2: Correlations between cohesion and coupling for the similarity methods

Figure 1 shows coupling and cohesion calculated according to the method described in section 2, for each of the 11 clustering techniques and three similarity methods. The following observations can be made about the figure:

- Coupling is roughly correlated with cohesion; in other words, coupling and cohesion tend to increase together. Table 2 shows the correlation coefficients for each of the three similarity methods. These observations might seem non-intuitive – one might imagine that as one improves design, or improves partitioning of a system, the cohesion should increase and the coupling should decrease.

  However, the evidence to the contrary can be explained as follows: Imagine a subsystem S containing a file that is highly coupled with many other files throughout the system (perhaps because it is a utility). Then the presence of this file in any subsystem would tend to increase both coupling and cohesion. A similar argument can be made explaining how an individual file with very low coupling to any other file (perhaps it performs an isolated function) would tend to decrease both the overall coupling and cohesion of any cluster in which it is placed.

- The three similarity methods give roughly similar indications about which clustering techniques have the highest and lowest cohesion and coupling . In all three cases, clustering techniques on the left have lower values for the metrics, and the metrics generally increase towards the right. This observation implies that we do not need to be concerned about which similarity method we use.

- The distance between the two curves in figure 1 corresponds to the quality metric Q; this is shown separately in figure 2. Q is supposed to measure quality by combining cohesion and coupling metrics additively (where coupling is negated). However, since coupling is always less than cohesion, yet is measured on the same zero-to-one scale, it always contributes less to the overall Q metric. Furthermore, since coupling increases less quickly than cohesion, for higher values of Coh, Q becomes ever more dependent

solely on Coh and thus Coup becomes irrelevant. This anomaly in the Q metric is especially clear in the case of similarity method RC, where Coup values are always rather low. Table 3 shows the extent to which the Q values are correlated with Coh, and Coup. in fact, it is rather striking that Q is *positively* correlated with coupling and seees to be saying that as coupling increases, so does quality.

We will rectify these problem with the Q metric when we derive a new metric in section 4.

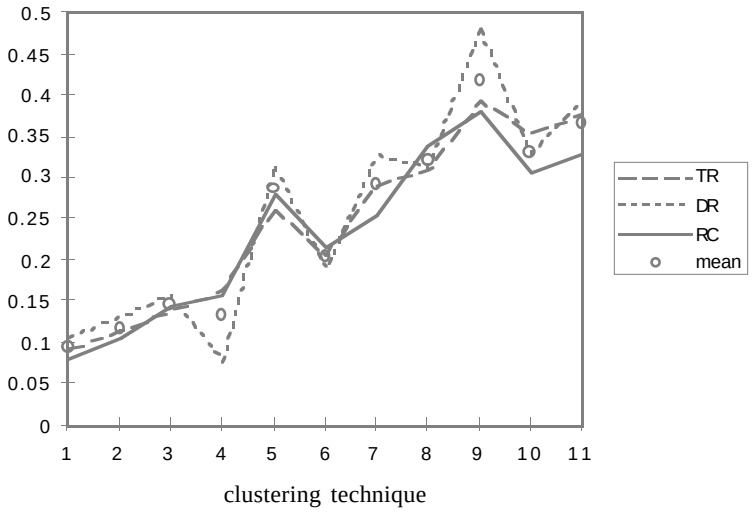

Figure 2: The quality metric Q for the three similarity methods.

Similarity method	Correlation of Q with Coh	Correlation of Q with Coup
TR	0.97	0.71
DR	0.99	0.48
RC	1.00	0.50

Table 3: Illustration that the quality metric Q() is very tightly correlated with cohesion.

4. Deriving improved metrics

4.1 Baseline values for coupling and cohesion

Given the methods of calculating coupling and cohesion described above, it is apparent that, for each similarity method, there is a baseline value for the coupling and cohesion metrics:

$$B_W = Coh_W(S)$$

where $Coh_W(S)$ is the cohesion of the entire system, i.e. the average of the similarity between all pairs of files. Random subsystems will tend to have:

$$Coh_W(P) = Coup_W(P) = B_W$$

Column 2 of table 4 shows the value of B for each similarity method. The B values can also be seen in figure 1 where the coupling curves are slightly above the baseline in all three cases. Note that it is possible for Coh and Coup to have values below the baseline. If Coh

were below the baseline, it would be indicative of an utterly uncohesive subsystem – worse than in the random case. It would clearly show good design if Coup were to be below the baseline in a subsystem, however in our system it rarely occurs that the average subsystem has this property.

We will use the B values to help derive a new metric in section 4.3.

Similarity Method	B	C_{coh}	C_{coup}
TR	0.208	1.000	0.443
DR	0.038	1.000	0.152
RC	0.016	1.000	0.081

Table 4: Baseline and ceiling values for coupling and cohesion in the system under study

4.2 Ceiling values for coupling and cohesion

For any given system and similarity method there will be maxima that the cohesion and coupling of no subsystem can exceed.

We can calculate the cohesion ceiling thus:

$$C_{coh_w} = \underset{i>j}{Max}\left(Sim_w(q_i, q_j)\right)$$

In other words, the highest similarity between any pair of files in the entire system. It should be clear that no average of similarities, and hence no cohesion metric for a cluster, could ever exceed this. In the system we studied, C_{coh_w} always equals 1, since there is always at least one pair of files that have identical similarity vectors. This is shown in column 3 of table 4.

The cohesion ceiling is also an upper bound on the coupling, however an improved upper bound can be defined as:

$$C_{coup_w} = \underset{1,n}{Max}\left(Coup_W(P1)\right)$$

This is the largest coupling value for all one-file clusters. No larger cluster could have a larger coupling since the calculation of Coup for such a larger cluster would necessarily include file clusters that have a lower Coup than that which was found to have C_{coup_w}.

Values of C_{coup_w} for our experimental system are shown in column 4 of table 4. Note that , especially for similarity method RC, they are very low. Note also that the graphs in figure 1, approach these calculated upper bound figures.

As with the baseline metric, we will use the C_{coh_w} and C_{coup_w} values in the derivation of new metrics in the next section.

4.3 Normalizing coupling and cohesion

The following summarizes problems identified so far with the metrics described in the literature::

- The baseline and ceiling values, B, C_{coh} and C_{coup}, and hence the ranges of the metrics, will differ between systems and between similarity methods. Thus, if using these original metrics it is important not to compare values from one system to another. It is

also not possible to compare one's system to systems whose cohesion and coupling values are quoted in the literature.

- The Q metric (Cohesion minus coupling) tends to be highly correlated with cohesion and positively correlated with coupling, which is misleading. These anomalies occur because the scales on which values of cohesion and coupling fall are quite different – coupling usually being a much smaller value.

We can solve both of the above problems by normalizing the cohesion and coupling metrics so that the zero point corresponds to the baseline B, and the 1-point corresponds to the appropriate ceiling C. Thus, the normalized cohesion and coupling metrics can be calculated as follows:

$$NCoh_W(P) = \frac{Coh_W(P) - B_W}{C_{coh_W} - B_W}$$

$$NCoup_W(P) = \frac{Coup_W(P) - B_W}{C_{coup_W} - B_W}$$

Both these metrics have an upper bound of 1, and a theoretical lower bound of -B/(C-B). For cohesion, the practical lower bound will be zero: if the cohesion were below zero then we have an extraordinarily bad subsystem that is less cohesive than a random system. For coupling it is possible to have values that are slightly below zero, but well above -1.[1]

Figure 3 shows three graphs of normalized coupling and cohesion in our experimental system, one for each of the similarity methods.

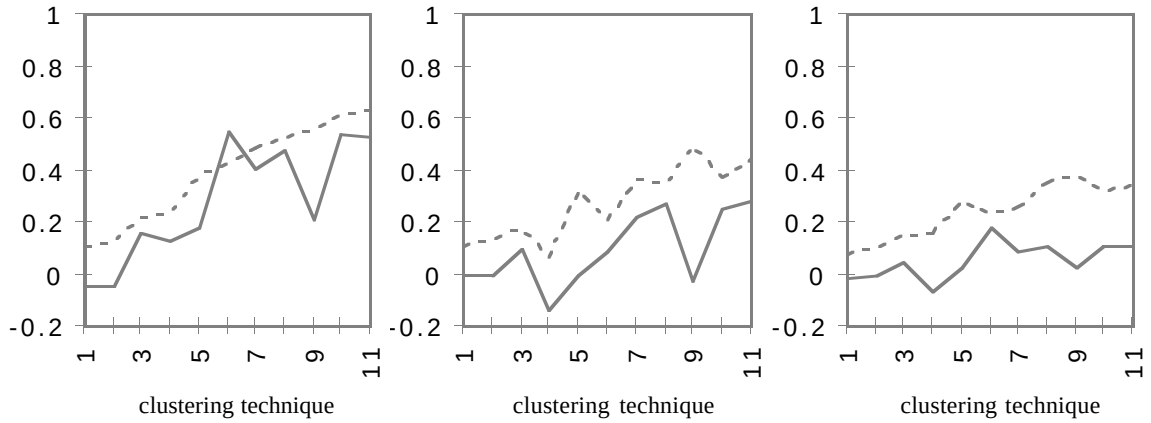

Figure 3: Normalized cohesion (dashed line) and coupling (solid line) for similarity methods TR, DR and RC respectively. Compare this with figure 1.

Compared with the non-normalized metrics in figure 1, the cohesion curves are visually very similar. The coupling curves, however, yield greater information since they cover much more of the range of the graph.

[1] Actually, in a rather odd system where the similarity between all pairs of files is almost equal, the lower bounds for our normalized metrics approach negative infinity. However we cannot envisage a realistic system where the lower bound would be less than about -1. If it were, the system would be so trivial that we wouldn't waste our time calculating these metrics.

The effect of normalization becomes even more prominent when we calculate the new quality metric thus:

$$NQ_W(P) = NCoh_W(P) - NCoup_W(P)$$

Experimental values of NQ are shown in figure 4. When compared with figure 2, two things become apparent: The two curves have common features, with high points at clustering technique 5 and 9, plus low points at clustering techniques 6 and 10. When one looks at the values, however, there is a dramatic change from what Q considered to be the best and worst quality versus what NQ considers best and worst: For example, Q considered technique 1 to be the worst, followed by technique 2; NQ considers technique 6 to be the worst, followed by technique 3.

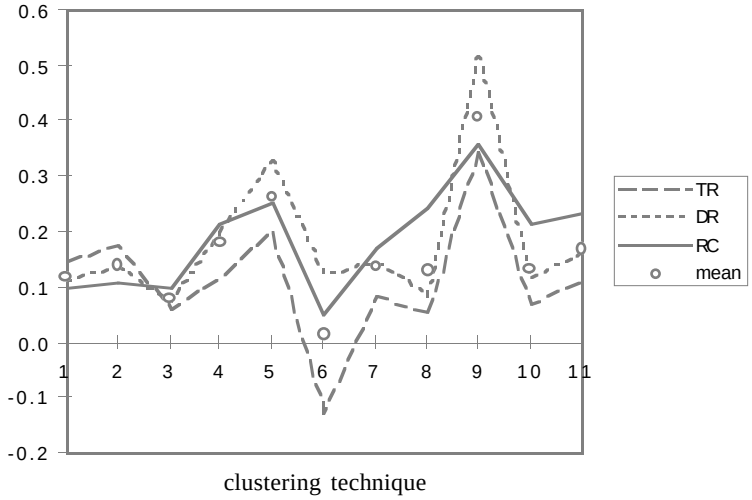

Figure 4: The quality metric NQ for the three similarity methods.

These changes have occurred because in NQ, coupling is being taken into account with equal weight to cohesion. In Q, the weighting of coupling was arbitrary and negligible. To further illustrate the effect of normalization, note that in table 5 NQ is no longer unreasonably correlated with cohesion. In fact, as would be expected, it is either uncorrelated or positively correlated with cohesion, and uncorrelated or negatively correlated with coupling.

Similarity method	Correlation of NQ with NCoh or Coh	Correlation of NQ with NCoup or Coup
TR	-0.05	-0.55
DR	0.48	-0.45
RC	0.75	-0.15

Table 5: Illustration that the normalized quality metric NQ is, in general, positively correlated with cohesion and negatively correlated with coupling.

5. Conclusions and future work

We can draw the following conclusions from the experiments and new metrics presented in this paper, to the extent that the large system with which we experimented is characteristic of systems in general:

- No matter which of three alternative methods we use to compute similarity, we obtain roughly the same information about the coupling and cohesion of subsystems.

- Cohesion in systems tends to be correlated with coupling. Thus, these metrics should not be thought of as completely independent indicators of quality.

- The design quality metric Q, which is the difference between cohesion and coupling, tends to be dominated by cohesion, and can therefore not be considered an independent metric.

- Standard cohesion and coupling metrics produce values that are typically in a narrow range which depends on the system being studied. We can efficiently compute, for the entire system an upper bound(C) for each of coupling and cohesion, and a common baseline (B) for both metrics. No subsystem will have its cohesion or coupling above the respective C value, whereas the baseline value represents random couplng and cohesion in the system.

- By mapping the B and C values to zero and 1 repectively, we can derive normalized metrics whose values fall within a consistent range (with occasional negative values for coupling) no matter which system is being compared or which method is being used to compare files. These normalized metrics should allow researchers and project managers to compare systems more effectively.

- The normalized design quality metric NQ (the difference between normalized cohesion and coupling) is much more reasonable than Q. The original Q metric is not independent from the coupling metric. The new NQ metric, on the other hand, tends to be positively correlated with cohesion and negatively correlated with coupling.

Now that we have presented these results for one software system, the most important future work will be to replicate them with other systems. Another possible research direction might be to experimentally modify the weight of NCoup and NCoh (from the current equal weighting) to discover a quality metric that corresponds best with software engineers' subjective opinions about quality.

References

Anquetil, N and T. Lethbridge, "Design Quality of Subsystems Extracted from File Names", submitted to Working Conference on Reverse Engineering, 1998a.

Anquetil, N and T. Lethbridge, "Extracting Concepts from File Names: A New File Clustering Criterion", Int. Conf. Software Engineering, Japan, 1998b.

Kunz, T and J. Black, "Using Automated Process Clustering for Design Recovery and Distributed Debugging", IEEE Trans. Software Engineering, 21, 6, June 1995, pp. 515-527.

Lakhotia, A. "A Unified Framework for Expressing Software Subsystem Classification Techniques", J. Systems and Software, 36, March 1997, pp. 211-231.

Müller, H.A., M. Orgun, S. Tilley, and J. Uhl, "A Reverse-Engineering Approach to Subsystem Structure Identification", Software Maintenance: Research and Practice, 5 pp 181-204, 1993

Patel, S., W. Chu and R. Baxter, "A Measure for Composite Module Cohesion", Proc 14th int. Conf. on Software Engineering, Melbourne, Australia, May 1992, pp. 38-48.