Atomic Architectural Component Recovery
for Program Understanding and Evolution

R. Koschke
University of Stuttgart, Breitwiesenstr. 20-22, D-70565 Stuttgart, Germany
koschke@informatik.uni-stuttgart.de

Abstract

Component recovery and remodularization is a means to get back control on large and complex legacy systems suffering from ad-hoc changes by recovering logical components and restructuring the physical components accordingly to decrease coupling among components and increase cohesion of components. This thesis is on unifying, quantitatively and qualitatively evaluating, improving, and integrating automatic and semi-automatic methods for component recovery.

1. Component recovery

Over time, ad-hoc changes deteriorate the system structure. These changes make program understanding increasingly more difficult; they decrease cohesion within components and increase coupling among components aggravating subsequent changes. Ad-hoc changes are rarely well documented making architectural descriptions obsolete. All these negative effects lead to increased maintenance costs. To get back control on large and complex legacy systems suffering from ad-hoc changes, it is necessary to recover the software architecture as built, to improve the architecture, and to restructure the system accordingly. Component recovery is a means to this end. Its purpose is to retrieve logical components from existing systems. Component are structural entities used to group elements. Types of components considered in this thesis are modules and subsystems. As opposed to the actual physical components of a system, i.e., the files, packages, or source directories of the system, logical components may only contain related elements that jointly contribute to a common purpose. This thesis is on automatic and semi-automatic methods for component recovery.

2. Historic perspective - the impact of this dissertation

This dissertation was published two years ago - too little time to fully develop its impact. Yet, a look back in the history of component recovery and the recent reception of this dissertation allows at least a glimpse on the impact of this dissertation on research in that area.

2.1. Period of pioneering feat

Component recovery and remodularization techniques have been studied since the early eighties. Belady and Evangelisti in 1982 [9] and Hutchens and Basili in 1985 [10] were among the first researchers in that area. Since then, component recovery has been a very active research area. At the time when the research to this dissertation started, more than 20 research papers had been published on fully automatic and semi-automatic techniques for component recovery, and their number is still growing [1-32].

These papers proposed many new interesting techniques. However, many of these techniques were not really evaluated quantitatively. If authors did evaluate their techniques, the evaluation was not comparable to evaluations of other authors because different evaluation techniques were used for different subject systems. Therefore, it was not clear which techniques one should use under which circumstances; it was not even clear whether these automatic techniques compare to manual remodularization at all.

The abundance of published methods called for frameworks to unify, classify, and compare them in order to make informed selections of appropriate component recovery techniques.

2.2. The dissertation’s contribution in the period of consolidation

The dissertation is the first contribution to a consolidation in component recovery. Its contributions are explained in the following.
Qualitative comparison and categorization. This thesis introduces a classification of component recovery techniques based on a unification of 23 published techniques. Focusing on structural techniques, 16 fully automatic techniques are classified into connection-, metric-, graph-, and concept-based subcategories and the commonalities and variabilities of these techniques are discussed in depth. The classification has been adopted by several authors. The by-product of the categorization is a generic description of the underlying algorithms for whole categories, which allows to quickly instantiate new members of a certain category by only adding the technique-specific parts.

Quantitative evaluation. Lakhotia and Gravely have pointed out the need for a systematic quantitative comparison of component recovery techniques in 1995 [17]. They developed ways to measure the difference between two single components produced by different techniques. Yet, they never actually applied their comparison metric. Neither did they describe how two sets of components should be compared using their metric - which is necessary if one wants to compare all candidate components of a technique to candidates of other techniques or to reference components of a human oracle, respectively: Obviously, it would not make sense to apply the metric to all pairs of components of different clusterings - but then, which two components should be compared to each other?

In joint work with Jean-Francois Girard and Georg Schied, we developed a framework that allows to measure recall and precision of component recovery techniques with respect to a set of reference components ascertained by software engineers. This framework solved the above mentioned problems. Moreover, we actually applied this framework to compare five different widely used techniques successfully in 1997 [4].

In the dissertation, I further refine the framework and apply it to almost all established connection-, metric-, and graph-based techniques. Altogether, twelve techniques are quantitatively assessed. As an oracle for this quantitative comparison, four C systems in the range of 30-55 KLOC are used (altogether 136 KLOC). Five independent software engineers and ten computer science students (undergraduate and graduate level) detected components in these systems manually. The manually detected components are used as reference components to which the candidate components of the techniques are compared. Measures were taken to assure the quality of the reference components.

The overall result of the evaluation is a major contribution to the research as it compares most existing component recovery techniques and points out their problems: They fall short of precision and recall with respect to human remodularization. With respect to the human oracle, only about 40% of the reference components can be found by the automatic techniques on average. As a matter of fact, most of the reference components are only found partially, i.e., several elements are missing or wrong elements are added.

About 50% of the correctly proposed candidate components that have a one-to-one relationship to a reference component are proposed by only one technique, i.e. the distinctive contribution of techniques is quite high. Nevertheless, even if one unites all candidate components, between 35% and 50% of the reference components of the four systems still go undetected (i.e., are either detected only partially or not at all).

Moreover, many false positives, i.e., components not present in the oracle, are proposed by the techniques. Yet, an investigation of the false positives shows that about 40% of the reference components classified as false positives could actually be considered components. They are either too small to be considered by the engineers, simply overlooked by the engineers, or represent alternative groupings (e.g., procedural versus object-oriented groupings). Thus, automatic techniques may in fact be supportive for a complete analysis or may be used to produce alternative views.

Similarity Clustering. Before creating a completely new technique, one should investigate whether an existing technique can be improved. I chose Schwaneke’s metric-based hierarchical clustering technique because it offers most flexibility [25]. The extensions to Schwaneke’s technique are manifold. One improvement of Similarity Clustering, the new technique, over Schwaneke’s original technique is the extension of the similarity metric to all kinds of global declarations (Schwanke considered only routines). Beyond entities, relationships and roles are also considered by Similarity Clustering leading to a more semantic-based clustering. Schwaneke neither considers relationships nor roles. Moreover, considering name similarities among identifiers and filenames additionally leverages pragmatic information. The thesis also shows how to calibrate the parameters automatically based on clustering samples, which is necessary to overcome the practical difficulty with metric-based techniques to adjust their many parameters. Experiments show that 20% of the reference component are sufficient to calibrate the parameters.

Similarity Clustering is more flexible and adaptable than other techniques, and, according to the quantitative evaluation, Similarity Clustering is among the best techniques for all systems. Its disadvantage is the higher number of false positives.

Integration of automatic techniques and enhancement for incremental use. In order to overcome weaknesses of the existing fully-automatic techniques, the thesis describes ways to integrate and combine the existing techniques in a semi-automatic method, in which computer and
maintainer collaborate to detect components. The integration is eased by the above mentioned unification and categorization of automatic component recovery techniques. In this framework, the automatic techniques can be run successively and their results be validated by the user. To this end, all the techniques are enhanced to work incrementally. The unification of the automatic techniques makes it possible to implement incremental variants for whole classes of techniques.

The results of the techniques can be combined by high-level operators modeled on intersection, union, and difference for fuzzy sets. An alternative way of integration is offered by a voting approach that summarizes the individual agreement of automatic techniques.

The integration is open to new techniques. Additional techniques can be integrated with very little effort thanks to the generic operators used to combine the technique.

**Controlled experiments for quantitative evaluations of semi-automatic and manual approaches.** The comparison framework is also further extended in the dissertation to cope with approaches that involve human intervention, such as the semi-automatic method proposed in this thesis. A controlled experiment is described that compares the semi-automatic method with a purely manual approach. The layout and evaluation scheme of the controlled experiment can be re-used for similar evaluations.

The outcome of the controlled experiment itself points out the dominating importance of human judgement and the shortcomings of the existing fully-automatic techniques: The experiment does not show any statistically significant superiority of the semi-automatic method over a purely manual approach. One possible explanation for the result of the experiment is that the semi-automatic method inherits the weaknesses of the used fully automatic techniques.

**Consequences for future research.** To overcome the weaknesses of automatic techniques, future research should experiment with more precise information gained from dataflow analyses and more domain-oriented information. The generic integration and evaluation frameworks developed in this thesis will make it easy to integrate new techniques and to measure their progress. Since all methods will have to cope with the vagueness and subjectivity of the grouping criteria for components, more attention should also be paid to human interaction, retrieval, and visualization techniques in order to better integrate human knowledge.

### 2.3. The time after the publication of the dissertation: its reception

After the dissertation was published, the evaluation framework was successfully used in two other comparisons. The concept-based techniques were evaluated by Alexander Porrmann using the same framework [29]. Additionally, in joint work with Gerardo Canfora and Jörg Czeranski, we improved the Delta-IC method [7] (originally developed by Canfora and others [11]). The difference between the old and new version of Delta-IC was measured using the evaluation framework.

The evaluation scheme originally developed by Girard, Schied, and me and further refined in the dissertation has gathered momentum and opened ways for quantitative comparisons of clustering techniques. Today, other research groups have proposed alternative, yet similar evaluation schemes (Tzerpos and Holt [31]; Mitchel and Mancoridis [30]; Girard and Würthner [32]). An agreement on a common evaluation scheme has not been reached yet, but is currently actively debated.

I hope that this dissertation serves as a stepping stone toward an established classification, integration, and comparison framework that is routinely used by all researchers in the area of component recovery in order to truly measure progress in this field. Beyond honest hopes for the future, I can safely state at least one thing about this dissertation: The dissertation still stands out as today’s broadest and most detailed qualitative and quantitative evaluation of existing component recovery techniques.

### 2.4. The author’s publications in the context of the dissertation

An electronic version of this thesis is available at http://www.informatik.uni-stuttgart.de/sfi/ps/rainer/thesis


References


