1 Assessment Outcomes

The tables below present the outcomes of our assessments on the different cases. Note that we present the assessments of three cases here, while in the paper we only discuss two of our cases.

Table 1: Coupling metrics of M1 and M2.

refactor	**service A**	**service B**	**CC**	**$CC_{A \to B}$**	**$CC_{B \to A}$**	**support**	**SC**
pre M1	S2.2.1	S2.2.2	0.67	0.82	0.58	258	-
pre M2	S2.3.1	S2.3.3	0.74	0.68	0.81	78	-
pre M2	S2.3.2	S2.3.3	0.48	0.76	0.35	34	-
pre M2	S2.3.2	S2.3.1	0.45	0.8	0.32	36	-

Table 2: Average metric values for entire application (M1 and M2)

refactor	**avg CC**	**avg SC**	**avg LOC**	**avg WSIC**	**avg SIDC**	**avg CF**
pre M1	0.31	-	34540.90	-	-	27.34
pre M2	0.31	-	34540.90	-	-	27.34

Table 3: Cohesion and size metrics of M1 and M2

refactor	**service**	**LOC**	**WSIC**	**SIDC**	**CF**
pre M1	S2.2.1	36745	24	0.0054	46.36
pre M1	S2.2.2	175942	-	-	87.21
pre M2	S2.3.1	14021	-	-	15.93
pre M2	S2.3.2	1155	2.00	0.5	4.52
pre M2	S2.3.3	2167	-	-	11.37

Table 4: Coupling metrics of D1 and D2

refactor	**service A**	**service B**	**CC**	**$CC_{A \to B}$**	**$CC_{B \to A}$**	**support**	**SC**
pre D1	-	-	-	-	-	-	-
post D1	S1.1.3	S1.1.2	0,60	0,93	0,45	14	0,75
pre D2	-	-	-	-	-	-	-
post D2	S2.4.2	S2.4.3	0,3	0,52	0,21	12	-
post D2	S2.4.1	S2.4.3	0,25	0,56	0,16	9	-
post D2	S2.4.2	S2.4.1	0,15	0,13	0,19	3	-

Table 5: Average metric values for the entire application (D1 and D2)

refactor	avg CC	avg SC	avg LOC	avg WSIC	avg SIDC	avg CF
pre D1	1.00	-	632.50	-	-	1.00
post D1	0.60	-	399.00	-	-	1.93
pre D2	0.31	-	37696.39	-	-	34.24
post D2	-	-	-	-	-	-

Table 6: Cohesion and size metrics of D1 and D2

refactor	service	LOC	WSIC	SIDC	CF
pre D1	S1.1.1	417.00	4.00	0.5	9
post D1	S1.1.3	93.00	-	-	1.07
post D1	S1.1.2	279.00	4.00	0.5	2.8
pre D2	S2.4.1	166981.00	-	-	154.36
post D2	S2.4.1	172167.00	-	-	84
post D2	S2.4.2	4396.00	-	-	82
post D2	S2.4.3	19417	-	-	201

Table 7: Coupling metrics for H1, R2.1 and H2

refactor	service A	service B	CC	$CC_{A \to B}$	$CC_{B \to A}$	support	SC
pre H1	S1.2.2	S1.2.1	0.66	0.64	0.7	7	0
pre/post R2.1	see GitHub page	see GitHub page	see GitHub page	see GitHub page	see GitHub page	see GitHub page	-
pre H2	S2.5.1	S2.5.2	0.89	0.91	0.89	201	-
post H2	S2.5.1	S2.5.2	0.92	0.86	1	6	-
post H2	S2.5.2	S2.5.3	0.92	1	0.86	6	-
post H2	S2.5.1	S2.5.3	0.85	0.86	0.86	6	-

Table 8: Average metric values for entire application (H1, R2.1 and H2)

refactor	avg CC	avg SC	avg LOC	avg WSIC	avg SIDC	avg CF
pre H1	0.66	0	951.50	-	-	2.05
pre R2.1	0.14	-	37.349.75	-	-	24.86
post R2.1	0.31	-	34.540.90	-	-	27.34
pre H2	0.31	-	34.540.90	-	-	27.34
post H2	-	-	-	-	-	-

Table 9: Cohesion and size metrics for H1, R2.1 and H2

refactor	**service**	**LOC**	**WSIC**	**SIDC**	**CF**
pre H1	S1.2.2	1189.00	-	-	1.3
pre H1	S1.2.1	714.00	4.00	0.5	2.8
pre R2.1	S2.1.1	1510.00	-	-	6.61
pre R2.1	S2.1.2	38910.00	-	-	69.83
pre R2.1	S2.1.3	1.584.00	-	-	2.33
pre R2.1	S2.1.4	13.241.00	6	0.5	26.44
pre R2.1	S2.1.5	319	-	-	0.44
pre R2.1	S2.1.6	11345	8	0	1.11
pre R2.1	S2.1.7	74673	-	-	47.28
pre R2.1	S2.1.8	157216	-	-	44.83
post R2.1	S2.1.9	175942	-	-	87.21
post R2.1	S2.1.1	2544	-	-	26.5
post R2.1	S2.1.2	3370	-	-	91
post R2.1	S2.1.3	2640	18	0.5621	14.44
post R2.1	S2.1.10	2231	19	0.1023	16.25
post R2.1	S2.1.6	3533	24	0.0054	46.36
post R2.1	S2.1.7	2497	-	-	10.04
post R2.1	S2.1.8	3611	16	0.55	37.5
pre H2	S2.5.1	75592	29	0.2875	172
pre H2	S2.5.2	176040	-	-	170.73
post H2	S2.5.1	75943	29	0.2875	332.14
post H2	S2.5.2	171476	-	-	159.43
post H2	S2.5.3	383006	6	0.5	469.43

Table 10: Coupling metrics for a selection of the service pairs of Spinnaker

service A	**service B**	**CC**	**$CC_{A \to B}$**	**$CC_{B \to A}$**	**support**
S3.0.4	S3.0.1	0,79	0,82	0,77	1582
S3.0.1	S3.0.3	0,77	0,69	0,87	1424
S3.0.4	S3.0.3	0,76	0,70	0,83	1355
S3.0.10	S3.0.2	0,42	0,50	0,36	288
S3.0.6	S3.0.9	0,42	0,38	0,47	223

Table 11: Cohesion and size metrics for the services S3.0.4, S3.0.1 and S3.0.3

service	**LOC**	**WSIC**	**SIDC**	**CF**
S3.0.4	302131.00	-	-	67.13
S3.0.1	277895.00	-	-	143.72
S3.0.3	158256.00	-	-	63.44

Table 12: Average metric values for entire Spinnaker application

avg CC	**avg SC**	**avg LOC**	**avg WSIC**	**avg SIDC**	**avg CF**
35.89	-	96424.36	-	-	33.85

2 Validation

The tables below present the alignment between our assessment outcomes and the expected outcomes based on the observations from the experts.

Table 13: Framework supporting the interpretation of the metric values in different refactor contexts.

Metric	Merge	Decomposition
CC	pre: if the CC value between two services was 0.66 or more, this was regarded as evidence in favour of merging these services.	post: a CC value was considered to suggest that the decomposition of service A into services B and C had been beneficial for maintainability if the CC value between service B and C was 0.33 or less.
SC	pre: due to the lack of thresholds proposed in the literature to base the classification of SC values on, we can only reason about the alignment of the evolution of SC during a refactor and the evolution in maintainability experienced by the expert. In the case of a merge, one would expect the average SC to decrease as the number of entities to contribute to coupling decreases. An SC which is higher than the system average was regarded as evidence in favour of merging the involved services.	post: due to the lack of thresholds proposed in the literature to base the classification of SC values on, we can only reason about the alignment of the evolution of SC during a refactor and the evolution in maintainability experienced by the expert. In case of a decomposition, if the resulting services had an under-average SC, the refactor was regarded as beneficial for the maintainability of the system.
WSIC	pre: a WSIC was considered to contradict the suggestion of merging services A and B to be beneficial for maintainability if the WSIC of either service A or B fell within the lower 50% intervals as proposed by [**?**], i.e., if the WSIC was higher than 15.	pre: considering the thresholds calculated by [**?**], we regarded WSICs higher than 15 as supporting evidence for decomposing a service. post: services resulting from a decomposition were expected to have lower WSICs than their ancestor.

Table 13 continued from previous page

SIDC	pre: a SIDC value was considered to contradict the suggestion of merging services A and B to be beneficial for maintainability if the SIDC of either service A or B fell within the lower 50% intervals as proposed by [**?**], i.e., if the SIDC was lower than 0.64.	pre: considering the thresholds calculated by [**?**], we regarded SIDC values lower than 0.64 as supporting evidence for decomposing a service. post: services resulting from a decomposition were expected to have higher SIDCs than their ancestor.
LOC	pre: a LOC value was considered to contradict the suggestion of merging services A and B to be beneficial for maintainability if the LOC value of either service A or B was higher than the average LOC value of all services in the system.	pre: a LOC value of a service which was higher than the LOC of an average service in the system was considered as support for the decomposition of that service. post: services resulting from a decomposition were expected to have lower LOC values than their ancestor.
CF	pre: a CF was considered to contradict the suggestion of merging services A and B to be beneficial for maintainability if the CF of either service A or B was higher than the average CF of all services in the system.	pre: if the CF of a service was higher than the average CF of all services in the system, this was considered as support for the decomposition of that service. post: services resulting from a decomposition were expected to have lower CFs than their ancestor.

Table 14: Relation between the assessment outcomes and the expert's observations for the analysed merges.

Metric	**Assessment observations**	**Expectations based on expert's experiences**
CC	Pre-M2: two of the services had a strong bidirectional change coupling, while the service pairs which included the third service had a lower change coupling which we classified as of regular strength.	Strong bidirectional change couplings were expected between the services.

Table 14 continued from previous page

	Pre-M1: a strong change coupling was observed for the service pair involved in this refactor.	
SC	Pre-M2: SC could not be measured due to the event-driven nature of the system. Pre-M1: SC could not be measured due to the event-driven nature of the system.	Above-average SC values were expected between the services.
WSIC	Pre-M2: only S2.3.2 offered an interface, which had a WSIC of 2. Pre-M1: only S2.2.1 offered an interface, which had a WSIC of 24.	The WSIC of each service was expected to be lower than 15.
SIDC	Pre-M2: only S2.3.2 offered an interface, which had a SIDC of 0.5. Pre-M1: only S2.2.1 offered an interface, which had a SIDC of 0.0054.	The SIDC value of each service was expected to be higher than 0.64.
LOC	Pre-M2: each service had an under-average LOC value. Pre-M1: both services had a LOC value which was higher than the average of the system.	The LOC value of each service was expected to be under average.
CF	Pre-M2: each service had an under-average CF. Pre-M1: both services had a CF which was higher than the average of the system.	The CF of each service was expected to be under average.

Table 15: Relation between the assessment outcomes and the expert's observations for the analysed decompositions.

Metric	**Assessment observations**	**Expectations based on expert's experiences**
CC	Post-D1: the CC value of the service pair resulting from the decomposition is 0.6, which we classified as of regular strength rather than weak. Post-D2: the CC values of the service pairs resulting from the decomposition could all be classified as weak.	Post-refactor, weak bidirectional change couplings were expected between the services.
SC	Post-D1: the SC between the resulting service pair was 0.75, but as we lacked averages due to this being the only service pair in the system, this value was not conclusive. Post-D2: SC could not be measured due to the event-driven nature of the system.	Post-refactor, under-average SC values were expected between the services.
WSIC	Pre-D1: the service had a WSIC of 4. Post-D1: only one of the services has an interface, also with a WSIC of 4. Pre- and post-D2: neither of the involved services offered an interface, so no WSIC could be determined.	Pre-refactor, a WSIC higher than 15 was considered an indication for decomposing. Post-refactor, the WSICs of the services resulting from the refactor were expected to be lower than the WSIC of the pre-refactor service.
SIDC	Pre-D1: the service had a SIDC of 0.5. Post-D1: only one of the services has an interface, also with a SIDC of 0.5. Pre- and post-D2: neither of the involved services offered an interface, so no SIDC could be determined.	Pre-refactor, a SIDC lower than 0.64 was considered an indication for decomposing. Post-refactor, the SIDC values of the services resulting from the refactor were expected to be higher than the SIDC of the pre-refactor service.

Table 15 continued from previous page

LOC	Pre-D1: the service had an under-average LOC. Post-D1: the LOC values of the services resulting from the decomposition were lower than the LOC of the pre-refactor service. Pre-D2: the service had a LOC value which was above average. Post-D2: the LOCs of the services resulting from the decomposition were lower than the LOC of the pre-refactor service, except for the LOC of S2.4.1, which is only lower when only considering lines of Java.	Pre-refactor, an above-average LOC was considered an indication for decomposing. Post-refactor, the LOC values of the services resulting from the refactor were expected to be lower than the LOC value of the pre-refactor service.
CF	Pre-D1: the service had an under-average CF. Post-D1: the CFs of the services resulting from the decomposition were lower than the CF of the pre-refactor service. Pre-D2: the service had a CF which was above average. Post-D2: the CFs of the services resulting from the decomposition were lower than the CF of the pre-refactor service, except for the CF of the source-docs-policy.	Pre-refactor, an above-average CF was considered an indication for decomposing. Post-refactor, the CFs of the services resulting from the refactor were expected to be lower than the CF of their ancestor.

Table 16: Relation between the assessment outcomes and the expert's observations for the analysed hybrid refactors.

Metric	**Assessment observations**		**Expectations based on expert's experiences**
CC	Pre-H1: the two involved services exhibited strong bidirectional change couplings. Pre-H2: a strong change coupling was observed for the service pair involved in this refactor.	Post-H1: this measurement was not feasible as H1 was never implemented. Post-H2: strong change coupling values, higher than the change coupling of the pre-refactor service-pair, were observed between the services.	Pre-refactor, regular to strong bidirectional change couplings were expected between the services. Post-refactor, the bidirectional change couplings between the services were expected to decrease.
SC	Pre-H1: the SC of 0 of the service-pair is equal to the average SC, as the average is only based on this single service pair. Pre-H2: SC could not be measured due to the event-driven nature of the system.	Post-H1: - Post-H2: -	Pre-refactor, above-average SC values were expected between the services. . Post-refactor, the SC values between the services were expected to decrease.
WSIC	Pre-H1: the single service that did offer an interface, had a WSIC of 4. Pre-H2: only S2.5.1 offered an interface, which had a WSIC of 29. Post-H2: S2.5.1 still had a WSIC of 29, and the new service (S2.5.3) had a WSIC of 6.	Post-H1: - Post-H2: the S2.5.1 still had a WSIC of 29, and the new service (S2.5.3) had a WSIC of 6.	Pre-refactor, a WSIC higher than 15 was considered an indication for decomposing. Post-refactor, the WSIC of each service was expected to decrease.

Table 16 continued from previous page

SIDC	Pre-H1: the single service that did offer an interface, had a SIDC of 0,5. Pre-H2: only the S2.5.1 offered an interface, which had a SIDC of 0,2875.	Post-H1: - Post-H2: the S2.5.1 still had a SIDC of 0,2875, and the new service (S2.5.3) had a SIDC of 0,5.	Pre-refactor, a SIDC lower than 0,64 was considered an indication for decomposing. Post-refactor, the SIDC value of each service was expected to increase.
LOC	H1: one service had an under-average LOC and one an above-average LOC. Pre-H2: both services had a LOC value which was higher than the average of the system.	Post-H1: - Post-H2: only the LOC of S2.5.2 decreased, while the LOC of S2.5.1 increased. The new service (S2.5.3) had a LOC larger than each of the pre-refactor services.	Pre-refactor, an above-average LOC was considered an indication for decomposing. Post-refactor, the LOC value of each service was expected to decrease.
CF	Pre-H1: as the system-average is calculated over the two involved services, naturally one service had an under-average CF while the other one had a CF higher than the average. Pre-H2: both services had a CF which was higher than the average of the system.	Post-H1: - Post-H2: the CF of S2.5.2 decreased, while the CF of S2.5.1 increased. The new service (S2.5.3) exhibited a CF higher than each of the pre-refactor services.	Pre-refactor, an above-average CF was considered an indication for decomposing. Post-refactor, the LOC value of each service was expected to decrease.