From Measures to Conclusions using Analytic Hierarchy Process in Dependability Benchmarking

Miquel Martínez, David de Andrés, Juan-Carlos Ruiz
STF-ITACA
Universitat Politècnica de València
Campus de Vera s/n, 46022, Spain
Email: {mimarra2, ddandres, jcruizg}@disca.upv.es

Jesús Friginal
LAAS-CNRS
7 avenue du Colonel Roche, F-31077 Toulouse, France
Email: jesus.friginal@laas.fr

***Abstract*—The goal of dependability benchmarks is to provide guidelines for comparison and selection of alternatives in critical application domains. However, and despite its intrinsic complexity, the analysis of benchmark measures remains today handmade and it is rarely considered in dependability benchmarks specifications. As a result, benchmark conclusions may vary from one benchmark user to another, and sometimes they are even difficult to relate to reported benchmark measures. To mitigate such problems, this paper proposes the use of the Analytic Hierarchy Process (AHP) technique to make explicit and repeatable the measures analysis process followed by benchmark users. In addition, an Assisted Pairwise Comparison Approach is proposed to reduce the sources of uncertainties existing in AHP. A wireless sensor network example shows the level of repeatability, consistency and objectivity the proposal promotes in the dependability benchmarks analysis process.**

I. Introduction

Conventional benchmarks characterise computer-based systems attending to different criteria such as performance, power consumption and cost. The aim of any benchmark is enabling the comparison among alternative systems, according to the established criteria, to take a well-based decision. Dependability benchmarks extend this concept to characterise those systems not only in the absence, but also in the presence, of accidental faults and attacks [1]. Accordingly, considered criteria must also encompass dependability and security characteristics [2], like the system robustness against considered perturbations.

In order to be useful, dependability benchmarks must satisfy a number of properties, like scalability, portability, non-intrusiveness, and representativeness. Among them, repeatability and reproducibility are of prime importance. On the one hand, *repeatability*, as defined by the Dependability Benchmarking project [1], guarantees statistically equivalent results when the benchmark is run more than once in the same environment and the same prototype. Without repeatability no one would be able to trust the results obtained from benchmarking experiments. On the other hand, *reproducibility* guarantees that another party obtains statistically equivalent results when the benchmark is implemented from the same specifications and is used to benchmark the same system. Reproducibility is strongly related to the amount of details given in the specifications.

The need to satisfy those properties made that, specially in the early days of dependability benchmarking, lots of works primarily focused on the definition of experimental procedures to benchmark a wide range of application domains [3], like web servers [4], on-line database transactional systems [5], or automotive systems [6]. All these works place a great emphasis in precisely i) describing the experimental set up, for third parties to be able to reproduce the same experimentation, ii) defining repeatable experimental procedures, including non-intrusive and controllable fault and attack injection techniques, and iii) identifying the set of measures to be considered and how they can be computed from obtained measurements.

It must be noted that little attention is paid to properly characterise measurement systems and express measurement results according to measurement theory [7]. For instance, in the dependability benchmarking domain the terms *measure* and *measurement*, as they will be used throughout this paper, make reference to what is understood as *mesurand* and *measurement result* in the metrology domain [**?**]. This may negatively affect the repeatability and reproducibility of the experimental procedure due to low quality measurements resulting from incomplete or ambiguous specifications. This problem was addressed in [7] by clearly determining existing sources of uncertainty in dependability measurements for distributed systems, whereas other works, like [8] and [9], focused on improving the quality of dependability measurements.

All these works have greatly contributed to improve dependability benchmarks properties, from specification to experimentation and monitoring. However, the last and also critical stage of the whole process, the comparison of benchmarked alternatives according to obtained measures to make an informed decision is still barely addressed. In most cases, the analysis process is very ambiguous or not documented at all, making the comparison among different results quite difficult. Repeating the same analysis after benchmarking new alternatives, modifying different parameters from the benchmark set up, or just to check the correctness of the previous assessment could lead to very different and even contradictory results due to ill defined analysis processes. In the same way, third parties trying to reproduce the same kind of analysis on their own systems may find it frustrating and meaningless. This lack of explicit criteria to compare alternatives greatly compromises the repeatability and reproducibility of the dependability benchmark process as a whole.

Although the arithmetic and geometric mean are sometimes used to compare alternatives, they are rather simplistic techniques that fail to grasp all the complex relationships existing among selected criteria. For instance, improving one criterion, like *throughput*, may negatively affect another criterion, like *power consumption*. This kind of problem involving conflicting criteria to reach a decision is addressed by *multi-criteria decision making* (MCDM) techniques in the field of operational research [10]. First attempts of using MCDM techniques to make explicit the comparison and selection process in dependability benchmarks were proposed in [11] [12].

This paper takes a step forward in this direction to improve the repeatability and reproducibility of the decision making process of dependability benchmarks by means of MCDM techniques. In concrete, the Analytic Hierarchy Process (AHP) [13], which allows to mathematically express the subjective and personal preferences of an individual or a group when making decisions, has been selected for this study. On the one hand, it has became a widely used technique to solve decision making problems in many areas like business [14], education [15], or engineering [16]. On the other hand, it allows for the hierarchical decomposition of the requirements of the analysis process, which appeals both industry (commonly interested in obtaining the right answer to the problem) and academia (more interested in analysing the problem from different perspectives and levels of detail). Although using a formal method to specify the decision making process, thus making explicit how the comparison and selection process should be performed, AHP requires a number of judgemental decisions relying on the expertise of the evaluator or group of evaluators. Accordingly, the selected alternative may vary depending on several factors that may negatively affect the properties of the dependability benchmark. To prevent this problem, this paper makes a deep analysis of the different elements that may affect the properties of the dependability benchmark and proposes a novel approach, that complements AHP, and that ensures the coherence, consistency, repeatability, and reproducibility of the decision making process.

The rest of the paper is structured as follows. Section II describes the basis of AHP which are required to understand how they can both benefit and harm the properties of the benchmark. How to integrate AHP into a dependability benchmark is presented in Section III by means of case study, focusing on wireless mesh networks, which will be used throughout the paper. The different problems deriving from the judgemental decisions taken when applying AHP are identified in Section IV. A novel Assisted Pairwise Comparison Approach (APCA) is defined in Section V to prevent the previously identified problems from affecting the benchmark properties. Finally, Section VI presents conclusions and future work.

II. The Analytic Hierarchy Process

The AHP is a technique that enables the decomposition of complex decision-making problems into smaller and easier to solve sub problems, by grouping the different considered criteria into more general criteria. The result is a hierarchical representation of the requirements of the analysis, being the top level criterion (root) the goal of the analysis, and the lowest level criteria (leaves) those defined by the measures to be analysed. Each hierarchy level can be seen as a different level of abstraction of the problem.

This hierarchy will be later used to compute a priority for each considered alternative reflecting its contribution to the goals optimisation. Thus, the contribution of each subcriterion to the upper level criterion is defined through its relative priority. These priorities are obtained by means of the pairwise comparison of all the subcriteria contributing to a given criterion, which eases the task of the evaluator (*law of comparative judgement* [17]). Those comparisons are assigned a number (intensity) stating how many times more important or dominant one criterion is over another regarding the criterion with respect to which they are compared. Table I lists the different numerical scale (from 1 to 9) [13] denoting the intensity of the importance of criterion A with respect to criterion B. The contribution of each alternative to the lowest level criteria is defined following the very same procedure.

TABLE I. The fundamental scale of absolute numbers for pairwise comparison

Definition	Description	Intensity[a]
Equal	A and B are equally important	1
Moderate	A is somewhat more important than B	3
Strong	A is much more important than B	5
Very strong	A is very much more important than B	7
Extreme	A is absolutely more important than B	9

[a] Intensities of 2, 4, 6 and 8 can be used to express intermediate values. Very close importance values can be represented with 1.1–1.9.

The pairwise comparison of N elements (criteria or alternatives) contributing to a given criterion is represented in a $N \times N$ matrix known as *pairwise comparison matrix*. As this matrix is reciprocal if the intensity of element A with respect to element B is $I_{AB} = X$, then the intensity of element B with respect to element A is $I_{BA} = 1/X$. Hence, $\forall i, j \in N : I_{ij} \times I_{ji} = 1$.

Those matrices should be consistent, i.e. if element A is more important than element B, and B is more important than element C, then I_{AC} should be greater than I_{BC}. A consistency ratio (CR) can be computed, as detailed in [18], to help evaluators to check that intensities representing the relative importance between elements are consistent. Matrices with $CR < 0.1$ are considered consistent because their small level of inconsistency is due to subjective appreciations [19].

Priorities are derived from the pairwise comparison matrices by computing the principal right eigenvector of the matrix. However, a more straightforward procedure can be used: i) compute the geometric mean (GM) for each row of the matrix, ii) sum up the geometric mean value obtained for each row, and iii) divide each geometric mean by the total sum. Priorities must be understood at two levels. Those directly obtained from the matrix are known as *local* priorities, and reflect a element's contribution to the immediate upper level criterion. The contribution of an element to the overall goal (*global* priority) is obtained by multiplying the local priority of the element by its upper level criterion's global priority. When a criterion is an immediate descendant of the goal, its local and global priorities are the same.

Finally, the priority of each alternative, i.e. its contribution to the system's goal, is the result of adding all the global priorities obtained for the lowest level criteria. These priorities are the ones defining the final alternatives ranking.

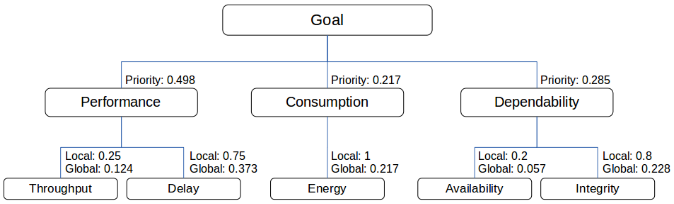

Fig. 1. AHP hierarchy tree making explicit the analysis criteria

III. AHP WITHIN DEPENDABILITY BENCHMARKING: WIRELESS MESH NETWORKS AS A CASE STUDY

Wireless Mesh Networks (WMNs) are a particular type of ad hoc networks which is currently being used, among other things, to provide cheaper and more flexible access to Internet than their wired counterparts to isolated or remote areas. As these networks rely on a wireless medium with no predefined communication infrastructure to communicate mobile and often performance- and power-constrained nodes, they may be subjected to a wide range of perturbations (both accidental faults and malicious attacks). Hence, in this context, dependability benchmarks aim at assisting network administrators to select the best ad hoc routing protocol for a given deployment, determine the main weaknesses of the selected routing protocol against particular perturbations, and fine tune that protocol accordingly, among other things. What is more, MCDM techniques in general, and AHP in particular, appear as suitable mechanisms to greatly improve the repeatability and reproducibility of the decision making process after including new fault-/attack-tolerance strategies, tuning the selected routing protocol, or injecting a new kind of fault/attack, among other possible uses. Accordingly, the classical stages any dependability benchmark follows have been enriched to integrate the AHP decision making process.

The set of measures defined to characterise the behaviour of the network in presence of perturbations consists of five different measures: i) the average amount of traffic effectively received during experimentation (*throughput*), ii) the average packets delay in milliseconds (*delay*), iii) the percentage of time routes are available for inter nodes communication (*availability*), iv) the percentage of packets whose data remain unaltered (*integrity*), and v) the average energy consumed by nodes (*energy*). The worst case threshold for each of the considered measures in this case study has been defined as i) 120 Kbps for *throughput*, ii) 300 ms for *delay*, iii) 60 % for *availability*, iv) 70 % for *integrity*, and v) 20 J for *energy*.

At this stage, those measures should be grouped together into higher level criteria to define the required AHP hierarchy.

Pairwise Matrix $\qquad\qquad\qquad\qquad$ Priority$(^{GM_x}/_T)$

$$\begin{array}{cc} & \begin{array}{ccc}P & D & C\end{array}\\ \begin{array}{c}P\\D\\C\end{array} & \begin{pmatrix}1&2&2\\ ^1/_2 & 1 & 1.5\\ ^1/_2 & ^1/_{1.5} & 1\end{pmatrix}\end{array} \rightarrow \begin{array}{l} GM_P(\,1 \quad 2 \quad 2\,) = 1.5874 \\ GM_D(^1/_2 \quad 1 \quad 1.5) = 0.9085 \\ GM_C(^1/_2 \quad ^1/_{1.5} \quad 1\,) = 0.6934 \\ \hline T \quad \rightarrow \qquad\qquad + 3.1893\end{array} \rightarrow \begin{array}{c}P\\D\\C\end{array}\begin{pmatrix}0.49\\0.28\\0.22\end{pmatrix}$$

$a)$ $\qquad\qquad\qquad\qquad$ $b)$ $\qquad\qquad\qquad\qquad$ $c)$

Fig. 2. Pairwise comparison matrix (a), geometric means (b), and local priorities (c) for performance (P), dependability (D), and consumption (C)

TABLE II. EXPERIMENTAL SET UP

Feature	Description
Nodes	10× Linksys WRT54GL routers 6× HP 520 laptops
Routing protocol	Optimized Link State Routing olsrd v 0.5.6
Traffic	UDP Constant Bit Rate of 200 Kbps extracted from daily observations [20]
Perturbations	(A) Ambient noise (S) Selective Forward attack* (J) Jellyfish attack* (T) Tampering attack* (F) Flooding attack*
Target	3-hop communication between nodes A and F in Figure 3
Number of experiments	15 per perturbation
Duration	9 minutes per experiment

* Node M in Figure 3 plays the role of Malicious node

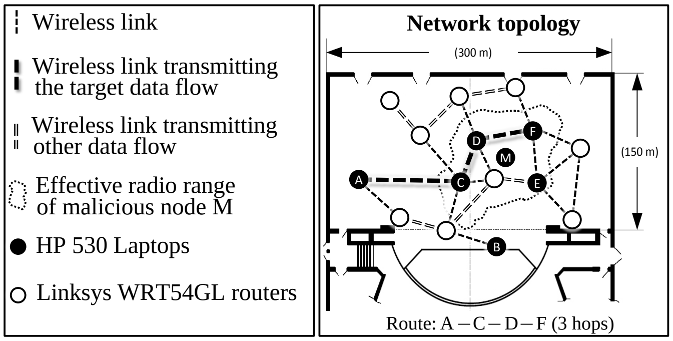
Wireless link
Wireless link transmitting the target data flow
Wireless link transmitting other data flow
Effective radio range of malicious node M
HP 530 Laptops
Linksys WRT54GL routers
Network topology
(300 m)
(150 m)
A
B
C
D
E
F
M
Route: A – C – D – F (3 hops)

Fig. 3. Wireless mesh network topology

As shown in Figure 1, this benchmark considers three upper level criteria: *performance*, *dependability*, and *consumption*. The particular contribution of each criterion to the immediate upper level criterion and the goal is computed by means of pairwise comparison matrices. Being the main aim of the network to enable the communication among nodes, the benchmark user has decided that a good performance should be of prime importance. Furthermore, as targeting WMNs, power consumption cannot be considered negligible for mobile devices, and it should be just a little less important than dependability. This is translated into the matrix depicted in Figure 2, which also illustrates the process followed to compute the related local priorities. Figure 1 shows the resulting local and global priorities for all the considered criteria after the benchmark user has built the required matrices. This makes explicit the relationship among criteria, so any ulterior analysis could be carried out following exactly the very same directives, thus enhancing its repeatability and reproducibility. Furthermore, defining these relationships *before* the experimental procedure takes place ensures that priorities are not biased by prior knowledge of obtained measures.

The particular experimental set up for this case study is listed in Table II. This set up defines five different scenarios in which the target network is subjected to one of the five most harmful perturbations in the WMNs domain [21]. The

TABLE III. EXPERIMENTAL RESULTS FOR EACH SCENARIO

Scenario	Throughput (Kbps)	Delay (ms)	Availability) (%)	Integrity (%)	Energy (J)
(A)mbient noise	145.2	48.2	73.6	92.12	8.2
(S)elective forwarding	121	42	91.2	97.53	8
(J)ellyfish	184.8	1086.5	88.7	98.54	10.3
(T)ampering	183.6	39.7	93.1	5.2	10.6
(F)looding	149	62.9	72.1	97.56	15.4

$$Ev_1 \quad \begin{array}{c|ccccc} & A & S & J & T & F \\ \hline A & 1 & 1 & 2 & 2 & 5 \\ S & 1 & 1 & 2 & 2 & 5 \\ J & 1/2 & 1/2 & 1 & 1 & 2 \\ T & 1/2 & 1/2 & 1 & 1 & 2 \\ F & 1/5 & 1/5 & 1/2 & 1/2 & 1 \end{array}$$

$$Ev_2 \quad \begin{array}{ccccc} A & S & J & T & F \\ \hline 1 & 1/1.5 & 2 & 2 & 3 \\ 1.5 & 1 & 4 & 4 & 4 \\ 1/2 & 1/4 & 1 & 1 & 3 \\ 1/2 & 1/4 & 1 & 1 & 3 \\ 1/3 & 1/4 & 1/3 & 1/3 & 1 \end{array}$$

$$Ev_3 \quad \begin{array}{ccccc} A & S & J & T & F \\ \hline 1 & 1/3 & 2 & 2 & 6 \\ 3 & 1 & 4 & 4 & 8 \\ 1/2 & 1/4 & 1 & 1 & 4 \\ 1/2 & 1/4 & 1 & 1 & 4 \\ 1/6 & 1/8 & 1/4 & 1/4 & 1 \end{array}$$

$$Ev_4 \quad \begin{array}{ccccc} A & S & J & T & F \\ \hline 1 & 1/4 & 5 & 5 & 7 \\ 4 & 1 & 6 & 6 & 8 \\ 1/5 & 1/6 & 1 & 1 & 6 \\ 1/5 & 1/6 & 1 & 1 & 6 \\ 1/7 & 1/8 & 1/6 & 1/6 & 1 \end{array}$$

$$Ev_5 \quad \begin{array}{ccccc} A & S & J & T & F \\ \hline 1 & 1/2 & 4 & 4 & 8 \\ 2 & 1 & 5 & 5 & 9 \\ 1/4 & 1/5 & 1 & 1 & 7 \\ 1/4 & 1/5 & 1 & 1 & 7 \\ 1/8 & 1/9 & 1/7 & 1/7 & 1 \end{array}$$

Fig. 5. Pairwise comparison matrices for energy as defined by all 5 evaluators

goal of this experimentation is to compare the behaviour of the network in presence of these five faults and determine which are the best and worst scenarios for the selected routing protocol. In this way, specific configurations and countermeasures could be deployed to face those weaknesses. The detailed experimental procedure, including how measurements are taken and how they are processed to obtain the required measures can be found in [22].

The average value of each measure for each scenario is presented in Table III. At this point, obtained measures should be analysed according to the previously defined hierarchy tree and the stated thresholds to rank the considered alternatives. It is now when pairwise comparison matrices for the lowest level criteria (measures) are built. Next section studies in detail a number of threats to the reliability, reproducibility, and repeatability of the analysis process supported by these matrices.

IV. Pairwise comparison of alternatives: threats to dependability benchmarking properties

The integration of AHP into dependability benchmarks specification presents clear benefits to the repeatability and reproducibility of the results analysis process. Making explicit the hierarchical aggregation of criteria and their particular contribution to the system's goal ensures that all evaluators will address the decision making problem following the very same guidelines. For instance, i) a given evaluator may benchmark a number of alternatives and apply the same guidelines later when new alternatives are available (new perturbations, new routing protocols, new fault tolerance mechanisms, etc.), ii) another evaluator may later repeat that analysis process on the same data to check the correctness of the procedure and understand the reasoning behind the obtained ranking, and iii) different evaluators may perform new experiments on similar scenarios and they can now follow the same decision making process to compare drawn conclusions.

The application of AHP in the last stage of the dependability benchmarking process (analysis of results), involves the pairwise comparison of alternatives with respect to the lowest level criteria (measures) to compute their local priorities. As this relies on the experience and judgement of evaluators, they are usually selected among experts in the field. Even though alternatives are compared two by two to ease the task of evaluators, and they are experts in their field, this paradoxically poses a number of threats to the reliability, repeatability, and reproducibility of the decision making process. For instance, in case of considering a large number of alternatives and measures, the huge amount of different comparisons to be made can wear out the evaluator, who can become careless in the following pairwise comparisons. In the same way, if a large number of comparisons is required, they will probably be performed in successive days, which may induced small variations in the evaluator's judgement.

$$\begin{array}{c|ccccc} & 73.4 & 88.6 & 88.7 & 90.5 & 71.9 \\ \hline 73.4 & 1 & 1/3 & 1/3 & 1/5 & 1 \\ 88.6 & 3 & 1 & 1 & 1/2 & 3 \\ 88.7 & 3 & 1 & 1 & 1/2 & 3 \\ 90.5 & \textcircled{5} & 2 & 2 & 1 & \textcircled{5} \\ 71.9 & 1 & 1/3 & 1/3 & 1/5 & 1 \end{array}$$

Fig. 4. Pairwise comparison matrix for availability as defined by evaluator 1

As an example, Figure 4 depicts the pairwise comparison matrix defined by evaluator 1 (Ev_1), in this case study, for the availability of the network. Although the consistency ratio of this matrix indicates that pairwise comparisons are *consistent* ($0.0014 < 0.1$), that does not mean that they are *coherent*. Scenarios S and J with availability 88.6% and 88.7%, respectively, have been considered *somewhat more important* than scenarios A and F, with availability 73.4% and 71.9%, respectively. It may seem coherent to evaluate with the same intensity (3) a difference of 15.2 and 15.3 percentage points (pp) with respect to scenario A, and 16.7 and $16.8pp$ with respect to scenario F (differences in a $1.4 - 1.6pp$ range). However, scenario T, with availability 90.5%, have been considered as *much more important* than scenarios A and F. This is clearly not coherent with previous comparisons, as differences of 17.1 and $18.6pp$, in a $0.3 - 1.8pp$ range with respect to previous ones, have been assigned a much higher intensity (5).

A common, although time consuming and costly approach to minimise all these problems derived from the judgemental nature of pairwise comparisons is inviting a set of experts to take part in the evaluation process. Even though the computed consistency ratio may prove matrices to be consistent, and assuming that they are all coherent, the internal (judgemental) guidelines and comparison scales used by each evaluator renders pairwise matrices very different among evaluators. Figure 5, which depicts the matrices built by all five evaluators (Ev_1 to Ev_5) for the energy consumed by network nodes, illustrates this fact.

For instance, the intensity of scenario A with respect to scenario F (I_{AF}) has been defined as 5, 3, 6, 7, and 8 by the evaluators. The same can be said about I_{SF} $(5, 4, 8, 8, 9)$ and I_{JF} $(2, 3, 4, 6, 7)$. The dispersion of these intensities is so large (from equal/moderate importance to very strong importance) that, although the reasoning of a given evaluator can be easily followed, it is very difficult to find a common line of reasoning among all the evaluators for the target system as a whole. In fact, as shown in Figure 6 the local priorities obtained for the

Throughput

$$\begin{array}{c|ccccc} & Ev1 & Ev2 & Ev3 & Ev4 & Ev5 \\ \hline A & 0.096 & 0.085 & 0.136 & 0.130 & 0.092 \\ S & 0.029 & 0.029 & 0.041 & 0.040 & 0.028 \\ J & 0.390 & 0.383 & 0.350 & 0.350 & 0.400 \\ T & 0.390 & 0.383 & 0.350 & 0.350 & 0.400 \\ F & 0.096 & 0.119 & 0.123 & 0.130 & 0.080 \end{array}$$

Delay

$$\begin{array}{c|ccccc} & Ev1 & Ev2 & Ev3 & Ev4 & Ev5 \\ \hline A & 0.201 & 0.180 & 0.218 & 0.194 & 0.172 \\ S & 0.201 & 0.180 & 0.218 & 0.194 & 0.227 \\ J & 0.024 & 0.025 & 0.026 & 0.023 & 0.024 \\ T & 0.447 & 0.434 & 0.380 & 0.493 & 0.459 \\ F & 0.125 & 0.180 & 0.158 & 0.096 & 0.118 \end{array}$$

Availability

$$\begin{array}{c|ccccc} & Ev1 & Ev2 & Ev3 & Ev4 & Ev5 \\ \hline A & 0.076 & 0.107 & 0.065 & 0.094 & 0.051 \\ S & 0.220 & 0.164 & 0.269 & 0.226 & 0.264 \\ J & 0.220 & 0.202 & 0.269 & 0.226 & 0.230 \\ T & 0.409 & 0.360 & 0.328 & 0.384 & 0.420 \\ F & 0.076 & 0.167 & 0.068 & 0.071 & 0.035 \end{array}$$

Integrity

$$\begin{array}{c|ccccc} & Ev1 & Ev2 & Ev3 & Ev4 & Ev5 \\ \hline A & 0.118 & 0.226 & 0.110 & 0.087 & 0.079 \\ S & 0.285 & 0.248 & 0.269 & 0.253 & 0.285 \\ J & 0.285 & 0.248 & 0.331 & 0.445 & 0.423 \\ T & 0.025 & 0.027 & 0.025 & 0.023 & 0.023 \\ F & 0.285 & 0.252 & 0.264 & 0.192 & 0.190 \end{array}$$

Energy

$$\begin{array}{c|ccccc} & Ev1 & Ev2 & Ev3 & Ev4 & Ev5 \\ \hline A & 0.313 & 0.251 & 0.223 & 0.270 & 0.306 \\ S & 0.313 & 0.412 & 0.483 & 0.519 & 0.452 \\ J & 0.150 & 0.136 & 0.128 & 0.092 & 0.108 \\ T & 0.150 & 0.136 & 0.128 & 0.092 & 0.108 \\ F & 0.075 & 0.065 & 0.039 & 0.028 & 0.026 \end{array}$$

Fig. 6. Local priorities for evaluators' decision matrices

lowest level criteria are very different among all the evaluators.

It is to note that, for this case study, the particular ranking for each alternative with respect to a given criterion is nearly the same for all the evaluators. They all agree that the best scenarios for throughput, delay, availability, integrity, and energy are J/T, T, T, J, and S, respectively. Likewise, they all agree that the worst scenarios are S, J, A/F, T, and F, respectively. However, the local priorities are so different that, once applied to the hierarchy tree previously defined, the global priority of each alternative, and thus the final ranking is quite different among evaluators (see Table IV).

This problem is usually addressed by means of group decision techniques, like the aggregation of individual judgements (AIJ) or the aggregation of individual priorities (AIP) [23]. These techniques try to find a consensus among evaluators depending on whether they want to act together as a single unit (AIJ) or as separate individuals (AIP). AIJ builds pairwise comparison matrices by computing the geometric mean of the individual intensities assigned by each evaluator, whereas AIP obtains the global priority of each alternative by computing the geometric mean of the individual priorities computed by each evaluator [24]. These methods lead to more reliable conclusions as they are obtained from a consensus reached among evaluations performed by a group of experts in the field. The ranking obtained by consensus using the AIJ method is listed in Table IV.

Nevertheless, involving a set of experts to increase the confidence that can be placed on the results provided by the analysis process also poses some problems. First, the economic impact of hiring a set of experts could be very high, specially when they must be contacted for pairwise comparison again and again after making any change in the system or the experimental set up, like considering different routing protocols, nodes' speed, mobility pattern, traffic, perturbations), or fault tolerance mechanisms. Second, the time required to build the required pairwise matrices can be quite long, as experts will not surely be fully dedicated to just this task. Third, the reproducibility and repeatability of the process may also be affected as judgements may variate along time. Although applying group decision techniques tends to mitigate this effect, when judgements from several experts fluctuate in opposite directions different rankings may be obtained from the same set of data.

TABLE IV. RESULTING RANKING AFTER COMPUTING THE PRIORITY FOR EACH ALTERNATIVE

Evaluator	**Ranking (from best to worst)**				
	1	**2**	**3**	**4**	**5**
Ev_1	T (0.2628)	J (0.2587)	S (0.1814)	A (0.1601)	F (0.1370)
Ev_2	T (0.2533)	J (0.2436)	S (0.1888)	A (0.1662)	F (0.1481)
Ev_3	J (0.2525)	T (0.2302)	S (0.2241)	A (0.1552)	F (0.1380)
Ev_4	J (0.2678)	T (0.2391)	S (0.2225)	A (0.1563)	F (0.1142)
Ev_5	J (0.2852)	T (0.2592)	S (0.2168)	A (0.1432)	F (0.0956)
AIJ	J (0.2622)	T (0.2500)	S (0.2081)	A (0.1546)	F (0.1252)

Next section presents the proposed approach to increase the reliability, repeatability, and reproducibility of the decision making process in dependability benchmarking without requiring any set of experts, thus also reducing the cost associated to its participation.

V. Assisted Pairwise Comparison Approach

Figure 7 depicts the relationship between the priority obtained by the most important criterion in a pairwise comparison matrix of just two elements and the intensity defined in that comparison. As priorities resulting from pairwise comparisons matrices of two elements are complementary, if the priority of the most important element is p, the priority of the other element is $1-p$. It must be noted that small variations for low intensities result in higher priority variations than in the case of considering high intensities. For example, by changing the intensity of the pairwise comparison between criteria A and B from 2 to 3, the priority of A increases (and thus the priority of B decreases) $8.3pp$. However, when increasing the intensity from 8 to 9, the priority of A only increases (and that of B decreases) $1pp$. Fluctuations for very small intensities (from 1.1 to 1.9) may imply great differences in the resulting priority. That is why, small variations of the judgement made by evaluators when retaking the comparison process may lead to very different results and greatly affect the expected properties of the dependability benchmarking analysis process.

The aim of the *Assisted Pairwise Comparison Approach* (APCA) is to automate the pairwise comparison process, thus preventing judgemental decisions from interfering with

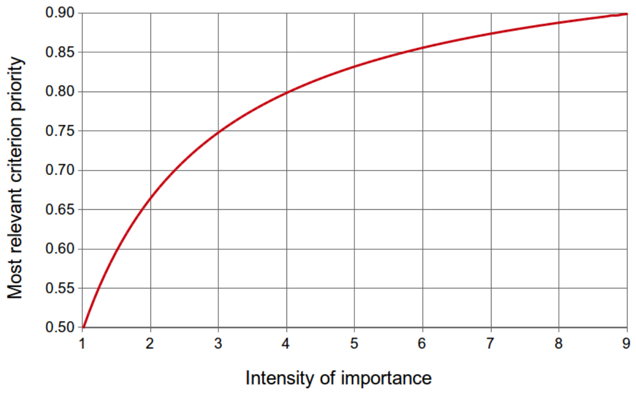

Fig. 7. Resultant priority from a pairwise comparison depending on the fundamental scale value

dependability benchmarking attributes. In such a way, the decision making process becomes completely repeatable and reproducible, as successive applications of this approach always render the very same results. Likewise, it also improves the confidence on the provided rankings, as computed pairwise comparison matrices are always consistent and coherent.

In order to automate the comparison process it is necessary to define a method to unify the interpretation of the relevance of one alternative against another with respect to a given criterion. The first problem is that the values obtained for each measure present very different ranges and hence determining their relative relevance is not so obvious. This issue is usually addressed by normalising those values in a 0 to 100 scale, which states the quality of this alternative with respect to a given measure according to acceptance values. These acceptance values, the upper and lower thresholds for a given measure, can be obtained by means of experimentation, through literature, or expertise. They must be defined in the dependability benchmark specification, so they could be known beforehand and help any evaluator in understanding and repeating the decision making process. Table V defines the minimum and maximum thresholds beyond which the quality of any alternative with respect to the selected measure is either maximised (100) or minimised (0).

To ensure that the normalisation process is known and applied in the same way, normalisation functions should be also defined in the dependability benchmark specification. This case study makes use of Eq. 1 for the linear normalisation of *the higher the better* measures (benefit normalisation function), whereas Eq. 2 is the linear normalisation function for *the lower the better* measures (cost normalisation function). These functions compute the quality (q_i) of the value obtained by an alternative (m_i) for a given measure (i) according to its acceptance values (T_{max_i} and T_{min_i}). Table VI lists the quality of the different alternatives for all the considered measures after the normalisation process. Obviously, different normalisation functions (exponential, logarithmic, discrete, etc.) could be defined according to the requirements of the system.

TABLE V. ACCEPTANCE VALUES DETERMINING THE REQUIRED BOUNDARIES OF THE CONSIDERED MEASURES

Acceptance value	Throughput	Delay	Availability	Integrity	Energy
T_{min}	120Kbps	40ms	60%	70%	5J
T_{max}	190Kbps	300ms	95%	99%	20J

$$q_i = \begin{cases} 0, & m_i \leq T_{min_i} \\ \frac{m_i - T_{min_i}}{T_{max_i} - T_{min_i}}, & T_{min_i} < m_i < T_{max_i} \\ 1, & m_i \geq T_{max_i} \end{cases} \quad (1)$$

$$q_i = \begin{cases} 1, & m_i \leq T_{min_i} \\ \frac{T_{max_i} - m_i}{T_{max_i} - T_{min_i}}, & T_{min_i} < m_i < T_{max_i} \\ 0, & m_i \geq T_{max_i} \end{cases} \quad (2)$$

After normalisation, the pairwise comparison process can be easily automated, as the difference between qualities is always expressed in pp. The minimum difference that can be found when comparing two qualities is $0pp$, which means that they are exactly equal. Accordingly, it should be associated with an intensity of 1. Likewise, the maximum difference between qualities ($100pp$) can be obtained when one alternative completely satisfies the requirements and the other alternative completely fails to do so. This case should be assigned the highest possible intensity (9). From this analysis is easy to determine that automatically computing the intensities of pairwise comparisons is just a matter of mapping the difference in quality between alternatives to the fundamental scale of comparison. For this case study, APCA considers a uniform distribution of quality difference (0 to 100) along the intensity range (1 to 9). Hence, if Q is the quality difference between two alternatives, being A more important than B, $I_{AB} = (Q \times (9-1) \div (100-0)) + 1$. This means that to increase the intensity in one unit, the difference in quality between alternatives must be of $12.5pp$. Obviously, other distributions can be used according to particular characteristics of the defined dependability benchmark. Algorithm 1 shows the algorithm used to implement APCA and build pairwise decision matrices.

The consistency of the pairwise comparison matrices generated by APCA has been experimentally verified for scenarios with an increasing number of alternatives (from 3 to 8). For each scenario, the consistency ratio (CR) of all the

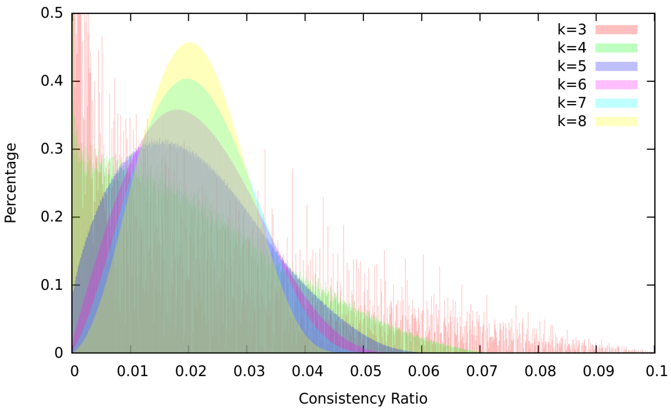

Fig. 8. Consistency ratio for all possible pairwise comparison matrices with a number of alternatives between 3 and 8

Algorithm 1 APCA

```
**n** *normalized elements to compare*
for $i$ from 1 to $n$ do
    for $j$ from $i$ to $n$ do
        $difference = n_i - n_j$
        if $difference < 0$ then {$n_j$ is greater than $n_i$}
            $a_{ji} = 1 + (n_j - n_i) \times 0.08$
            $a_{ij} = 1/a_{ji}$ {reciprocity}
        else if $difference > 0$ then {$n_i$ is greater than $n_j$}
            $a_{ij} = 1 + difference \times 0.08$
            $a_{ji} = 1/a_{ij}$ {reciprocity}
        else {$n_i$ is equal to $n_j$}
            $a_{ij} = 1$
            $a_{ji} = 1$
        end if
    end for
end for
```

TABLE VI. QUALITY OF THE DIFFERENT ALTERNATIVES FOR ALL THE CONSIDERED MEASURES AFTER NORMALISATION

Scenario	Throughput	Delay	Availability	Integrity	Energy
A	36	96.85	38.86	76.27	78.67
S	1.43	99.23	89.14	94.93	80
J	92.57	0	82	98.41	64.67
T	90.86	100	94.57	0	62.67
F	41.43	91.19	34.57	95.03	30.67

possible matrices that could be generated with normalised values between 0 and 100 was computed. Figure 8 depicts the distribution of the obtained consistency ratio with increasing number of alternatives (k). Matrices were consistent in all cases ($CR < 0.1$).

The application of Algorithm 1 to the normalised values listed in Table VI sets the intensities for the pairwise decision matrices comparing all the alternatives for each of the five alternatives. The local and global priorities for each alternative and the finally obtained ranking are shown in Table VII.

It must be noted that this ranking is exactly the same obtained by means of a group of experts using AIJ, which validates the proposed approach. APCA not only provides a deterministic, and thus reproducible and repeatible, analysis process, but also eliminates the costs associated to group decision making techniques as no experts are required for its application.

VI. CONCLUSIONS AND FUTURE WORK

The comparison of alternatives with respect to defined criteria to reach informed decisions has not been formally and rigorously addressed so far in dependability benchmarking. Ambiguity and lack of documentation in the specification of the analysis stage are quite common, which negatively affect the repeatability and reproducibility of the decision making process, and simplistic approaches like arithmetic or geometric mean can barely handle the complex relationships among criteria. MCDM techniques have proved their feasibility to solve decision problems involving conflicting criteria in other application fields, so it just seemed natural to determine whether they could also be applied to dependability benchmarking.

TABLE VII. LOCAL/GLOBAL PRIORITIES AND RANKING OBTAINED BY MEANS OF APCA

Scenario	Global priorities					Goal	Ranking
	Throughput	Delay	Availability)	Integrity	Energy		
A	0.036	0.025	0.004	0.027	0.068	0.160	4
S	0.011	0.025	0.013	0.065	0.068	0.181	3
J	0.146	0.003	0.013	0.065	0.033	0.259	2
T	0.146	0.056	0.023	0.006	0.033	0.263	1
F	0.036	0.016	0.004	0.065	0.016	0.137	5

This paper has studied in depth the possible integration of AHP, one of the most widely used MCDM techniques, to support the decision making process in dependability benchmarks. The analysis of the effects of different subjective and judgemental components of AHP has shown that they can lead to small variations in pairwise comparisons for a given criteria. The combined influence of all these small fluctuations may result in totally different conclusions among different evaluators, or even among different assessments performed by the same evaluator. Accordingly, dependability benchmarks will barely satisfy the required repeatability and reproducibility properties in such cases.

In order to prevent these judgemental elements from affecting the analysis process, this paper proposed a novel Assisted Pairwise Comparison Approach (APCA) to be used with AHP. Unlike common group decision techniques, which require a set of experts to take part in the decision making process, APCA automates the pairwise comparison process thus reducing its associated time and cost. The combined use of AHP and APCA provides a decision making methodology for dependability benchmarks that not only allows to consider complex relationships among criteria, but also assures the coherence, consistency, repeatability and reproducibility of the comparison and selection process.

Complementing AHP with APCA is just a first step towards the definition of a proper methodology for the analysis of dependability benchmarking results. The analysis process can be very sensitive to a given number of factors, like changes in the influence of certain criteria, or variation in the input data [25]. A deep study of input data ranges with respect to possible hierarchy trees may enable the definition of guidelines to build hierarchy trees leading to highly robust rankings. Likewise, these guidelines may reduce the judgemental decisions taken to build hierarchy trees, which depend, up to now, on the expertise of the person defining the benchmark. These are all promising ideas for future research.

ACKNOWLEDGEMENT

Work partially supported by the Spanish project ARENES (TIN2012-38308-C02-01).

REFERENCES

[1] DBench, "Dependability Benchmarking," IST Programme, European Commission, IST 2000-25425, [Online]. Available: http://www.laas.fr/DBench, 2013.

[2] A. Avizienis *et al.*, "Basic concepts and taxonomy of dependable and secure computing," *IEEE Transactions on Dependable and Secure Computing*, vol. 1, no. 1, pp. 11–33, Jan–March 2004.

[3] K. Kanoun and L. Spainhower, *Dependability benchmarking for computer systems*. John Wiley & Sons, 2008, vol. 72.

[4] J. Dures, M. Vieira, and H. Madeira, “Dependability benchmarking of web-servers,” in *Computer Safety, Reliability, and Security*, ser. Lecture Notes in Computer Science, M. Heisel, P. Liggesmeyer, and S. Wittmann, Eds. Springer Berlin Heidelberg, 2004, vol. 3219, pp. 297–310.

[5] M. Vieira and H. Madeira, “A dependability benchmark for oltp application environments,” in *Proceedings of the 29th international conference on Very large data bases - Volume 29*, ser. VLDB ’03. VLDB Endowment, 2003, pp. 742–753.

[6] J.-C. Ruiz, P. Yuste, P. Gil, and L. Lemus, “On benchmarking the dependability of automotive engine control applications,” in *Dependable Systems and Networks, 2004 International Conference on*, 2004, pp. 857–866.

[7] A. Bondavalli, A. Ceccarelli, L. Falai, and M. Vadursi, “A new approach and a related tool for dependability measurements on distributed systems,” *Instrumentation and Measurement, IEEE Transactions on*, vol. 59, no. 4, pp. 820–831, 2010.

[8] L. Angrisani and M. Vadursi, “Cross-layer measurements for a comprehensive characterization of wireless networks in the presence of interference,” *Instrumentation and Measurement, IEEE Transactions on*, vol. 56, no. 4, pp. 1148–1156, 2007.

[9] M. Choi, N. Park, V. Piuri, and F. Lombardi, “Reliability measurement of mass storage system for onboard instrumentation,” *Instrumentation and Measurement, IEEE Transactions on*, vol. 54, no. 6, pp. 2297–2304, 2005.

[10] M. Koksalan, J. Wallenius, and S. Zionts, *Multiple Criteria Decision Making: From Early History to the 21st Century*. World Scientific Publishing Company; 1 edition (June 6, 2011), 2012.

[11] M. Martínez, D. de Andrés, J.-C. Ruiz, and J. Friginal, “Analysis of results in dependability benchmarking: Can we do better?” *M&N 2013, International Workshop on Measurements and Networking*, pp. 127–131, 2013.

[12] M. Martinez, D. D. Andres, and J.-C. Ruiz, “Gaining confidence on dependability benchmarks’ conclusions through “back-to-back” testing (practical experience report),” in *Dependable Computing Conference (EDCC), 2014 Tenth European*, May 2014, pp. 130–137.

[13] T. Saaty, “Decision making with the analytic hierarchy process,” *International Journal of Services Sciences*, vol. 1, no. 1, pp. 83–98, 2008.

[14] M.-K. Chen and S.-C. Wang, “The critical factors of success for information service industry in developing international market: Using analytic hierarchy process (ahp) approach,” *Expert Systems with Applications*, vol. 37, no. 1, pp. 694–704, 2010.

[15] X. Sun *et al.*, “Construction and operation of analytic hierarchy process about moral education evaluation in colleges and universities.” *Advances in Information Sciences & Service Sciences*, vol. 3, no. 11, 2011.

[16] İ. Ertuğrul and N. Karakaşoğlu, “Performance evaluation of turkish cement firms with fuzzy analytic hierarchy process and topsis methods,” *Expert Systems with Applications*, vol. 36, no. 1, pp. 702–715, 2009.

[17] L. L. Thurstone, “A law of comparative judgment.” *Psychological review*, vol. 34, no. 4, p. 273, 1927.

[18] J. A. ALONSO and M. T. LAMATA, “Consistency in the analytic hierarchy process: A new approach,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 14, no. 04, pp. 445–459, 2006. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S0218488506004114

[19] T. L. Saaty, “Decision-making with the ahp: Why is the principal eigenvector necessary,” *European Journal of Operational Research*, vol. 145, no. 1, pp. 85 – 91, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0377221702002278

[20] “Hillsdale WMN,” Online: http://dashboard.open-mesh.com/overview2.php?id=Hillsdale, 2012.

[21] I. F. Akyildiz and X. Wang, “A survey on wireless mesh networks,” *IEEE Communications Magazine*, vol. 43, no. 9, pp. S23–S30, 2005.

[22] D. de Andrés, J. Friginal, J. C. Ruiz, and P. Gil, “An Attack Injection Approach to Evaluate the Robustness of Ad Hoc Networks,” in *IEEE 15th Pacific Rim International Symposium on Dependable Computing*, 2009, pp. 228–233.

[23] E. Forman and K. Peniwati, “Aggregating individual judgments and priorities with the analytic hierarchy process,” *European journal of operational research*, vol. 108, no. 1, pp. 165–169, 1998.

[24] B. Srdjevic, Z. Srdjevic, T. Zoranovic, and K. Suvocarev, “Group decision-making in selecting nanotechnology supplier,” in *Nanomaterials: Risks and Benefits*, ser. NATO Science for Peace and Security Series C: Environmental Security, I. Linkov and J. Steevens, Eds. Springer Netherlands, 2009, pp. 409–422. [Online]. Available: http://dx.doi.org/10.1007/978-1-4020-9491-0_32

[25] W. Wolters and B. Mareschal, “Novel types of sensitivity analysis for additive MCDM methods,” *European Journal of Operational Research*, vol. 81, no. 2, pp. 281–290, 1995.