Towards the definition of an information quality metric for information fusion models

Horacio Paggi[1,*], Javier Soriano[1], Juan A. Lara[2], Ernesto Damiani[3,4]

[1]Universidad Politécnica de Madrid, Campus de Montegancedo, 28660, Boadilla del Monte, Madrid, Spain

[2]Madrid Open University, UDIMA, Carretera de La Coruña km 38,5, Vía de Servicio, nº 15, 28400, Collado Villalba, Madrid, Spain

[3]Centre on Cyber–Physical Systems, Khalifa University of Science and Technology, Hadbat Al Zaafran, 127788, Abu Dhabi, United Arab Emirates

[4]SESAR Lab, Università degli Studi di Milano, Via Bramante 65, 26013, Crema, Italy

*Corresponding author: horacio.paggi@gmail.com

ABSTRACT

Managing information quality has become important in cyber-physical systems dealing with big data. In this regard, different models have been proposed, mainly in flat peer-to-peer networks, in which exchanging information efficiently is a key aspect due to scarce resources. However, little research has been conducted on information quality metrics for cyber-physical scenarios. In this paper, we propose an information quality metric and show its application to an information fusion model. It is a "model-oriented quality metric" since it allows non-predefined variants on its configuration depending on the application domain. The model was tested on several simulations using open datasets. The results obtained in the performance of the model confirm the validity of the information quality metric, proposed in this paper, on which the model is based. The model may have a wide variety of applications such as mobile recommendation or decision making in critical environments (emergencies, war, and so on).

Keywords: Adaptive Peer-to-Peer systems; Information fusion; uncertain information handling; information quality metric.

1. Introduction

The spread of interconnected devices is a common phenomenon in nowadays networks. In addition, currently, people tend to interact with intelligent devices with more frequency. In this scenario, some issues related to equipment battery care, security or availability of communication services are starting to arise. Moreover, interconnections can occur in any possible order and the information exchanged may be very diverse in terms of the quality that information has. This particularly applies to human beings that take part in communications, since they tend to naturally inject subjectivity and uncertainty in the information they deal with. For this reason, current cyber-physical systems need to count on Information Quality (IQ) metrics which reflects characteristics of human communication and allows, through the local maximization of the exchanged data's IQ, the improvement of the network's performance, irrespective of its components (human or, most commonly, devices).

In this research, we propose an IQ metric that supports an Information Fusion (IF) model based on a flat Peer-to-Peer (P2P) network (the connections among peers are formed arbitrarily and without any kind of hierarchy, that is, all the peers act as equals [1]) with no dedicated elements beforehand (servers, switches, hubs, and so on) [2]. The IF model works in a way that a certain external agent (denoted Ω) queries the P2P network, composed of different agents, in search for a certain information and then, the agents (denoted α) in the network collaborate for providing an answer as accurate as possible. Given that people can take part in the network (specifically in what we called intelligent spaces [3]), one of our design goals was to reflect dimensions of IQ that are common in human communication, for example, the closeness of the source of information, the uncertainty it has about its data, its vagueness, etc. Regarding the closeness of the source of information, it is important to clarify that it is measured for a certain agent α as the amount of hops that it is neccessary to take in order to get from α to the source of information according to the topology of the network. In the proposed model we only consider the closesness with respect to the external querying agent Ω. The model tends to increase the information quality for each peer of the network. However, it does not necessarily guarantee the achievement of the maximum theoretical information quality at a certain time, due to the fact that resources are limited and random.

The model was designed to be relatively simple to compute, considering the usually limited computational resources of the network components. The main computation process to be conducted by an agent α of the network consist on calculating tue quality of a certain answer (as we will explain later in equation 3). To do so, the agent needs to calculate a quotient if there is to need to query other peers of the networks. Otherwise, a series of basic mathematical operations (quotient, maximum, multiplication and powers) involving a few scalar values need to be performed (in our case,the name of operations will be equal to the number of peers queried by the agent α, that is, between 3 and 5 in our experiments). Considering the computing capacity of modern devices, authors believe that the calculation of such operations is not an issue. A different matter is the calculation of some of the parameters used in our approach that are described later (such as quality variation, vagueness or uncertainty), for which the available hardware needs to be considered.

The main two contributions of this paper are:

a) It depicts a type of IQ metric (it is a type or model rather than a specific metric because several options are allowed in its computation) that can be understood and computed locally by machines or humans in a rather direct way. That is, it can be applied in cyber-physical scenarios, where IQ is crucial.

b) It shows that the defined metric and the model's behaviors lead to an increase in the performance of the entire network.

The rest of the paper is organized as follows. The state of the art regarding IQ metrics and IQ representation formats is described in Section 2. Section 3 presents the main features of the model proposed in this paper, including the definition of the IQ metric, and it includes a description of the dynamics of the model. The results obtained after implementing the proposed model based on our IQ metric are presented in Section 4. Finally, the conclusions and future lines of research are included in Section 5.

2. State of the art

2.1 Information quality

First of all, we would like to point out that we will use data quality and information quality as equivalent terms, which follows the line started by many other relevant authors [4]. It is possible to consider data quality as a series of dimensions describing the quality of the information produced by an information system; that is, a measure of the success of the system producing this information. Therefore, information can be considered as a product with a certain quality.

From the entire set of feasible dimensions, only a few are used to explain the data quality. The most common are timeliness, accuracy, consistency, completeness, fitness for use and relevance [5]. Other alternative dimensions have been also considered depending on the different frameworks used for evaluation and the application domains; a description of twelve of such dimensions is provided in [4]. The importance of IQ lies in the fact that it can influence the decision making process quality ([6]).

In this work, two of the dimensions of IQ defined in the literature are considered significant: uncertainty and vagueness, which can be seen as two different aspects of indeterminacy, as stated by Novák [7]. Indeterminacy is a phenomenon that refers to how much it is known about the consequences of receiving a message. Therefore, in our research, it is important to focus on vagueness and uncertainty since they can seriously affect the peers' responses, as stated in [8]. Uncertainty is mainly related to the error or imprecision associated with data while vagueness is a inherent issue of natural language (for example, in the sentence "long book", how many pages does "long" mean?). A deeper analysis of the different aspects of indeterminacy has been carried out by Bossè et al. [9]. For example, in the case of information expressed as text, they distinguish between uncertainty within the text (imprecision, vagueness, polysemy) and uncertainty about the content (for example, citing their example, as in "Sally gave Mary her book"). Additionally, they cite the probability (and its expression) that the idea expressed in the text is actually happening.

From the different possible ways of modeling vagueness, in this work, we selected Lofti Zadeh's fuzzy sets theory [10], which is the basis of many frameworks for uncertainty representation (see for example [11]). One of the components of vague information is its vagueness itself, its *fuzziness* (dispersion); In other words, the greater the dispersion we have, the poorer the quality will be. This idea is supported, for example, by Bardossy *et al*. [12], who state that the best of two fuzzy numbers is the less vague. The dispersion H of an answer expressed as a fuzzy number *R* can be measured as the area under the graph of its membership function $\mu = \mu(x)$, which is formally defined in Equation (1).

$$\begin{cases} H(R) = \int_{\infty}^{\infty} \mu R(x) dx \quad (continuous) \\ H(R) = \sum_{x \in X} \mu R(x) \quad (discrete) \end{cases} \qquad (1)$$

Other ways of measuring this dispersion have been defined. For example, Yager (cited by [13]) proposes (when *R* is finite) the formalization expressed in Equation (2).

$$H(R) = \left|X\right| - \sum_{x \in X} \mu R(x) - \mu R^{-}(x) \qquad (2)$$

where R^{-} is the complementary set of R and X is the universal set where x varies. This measures the lack of distinction between a fuzzy set and its complement, i.e., how fuzzy it is.

It is crucial for decision-makers to count on IQ metrics that reflect the effects on the information of imperfections such as uncertainty and vagueness. IQ is usually represented by metadata with values inside the interval [0, 1], which help decision-makers to measure and compare data quality in the context of decision-making tasks [14].

We did not specifically consider human usability; rather, we looked for the simplest human- and machine-readable representation that was at the same time easy to calculate and did not generate a major computational overload. Decision-makers may suffer from information overload as a result of the use of such metadata within the decision-making process. On this ground, the system uses a set of preferred agents. The tags are usually designed to be created and used by people and not by machines. This means that the right tags are not available for fast-growing fields like the IoT (Internet of Things).

2.2 Use of information fusion to reduce information imperfections

Information fusion (IF) techniques have been used to reduce indeterminacy and other data “imperfections” and to improve the results obtained from them, reducing the negative impact of such imperfections in the decision-making process [15]. In other words, the receiver should make the best decisions considering the semantic or pragmatic content assigned to the received message. Referring to this, Foo et al. state that some of the advantages of using IF are: improvement in data accuracy, reduction of the data uncertainty and ambiguity, improvement of the situation awareness (SAW) and in the inferences which lead to a better decision making [16].

One of the advantages of the IF approach is that every agent is allowed to reason and to handle vagueness and uncertainty in a wide spectrum of ways. Therefore it is especially interesting for open systems where, with the sustained increase of the connectivity between systems due to IoT, the fusion of heterogeneous data has been spreading in the last years. In a certain sense, the approach proposed in this work is bio-inspired because it is based on the “natural” method a person would use to manage vagueness and uncertainty.

To fuse data coming from the queried peers several methods can be applied. Bayesian methods are a formalism that can be applied when uncertainty is represented by means of probabilities.

Dempster-Shafer’s theory [17] generalizes Bayesian methods and is able to represent incomplete knowledge, and update beliefs as new information comes in. In addition, it is able to handle uncertainty explicitly. Dezert-Smarandache’s theory [18] is an improvement over Dempster-Shaffer`s one because it is able to formally combine any type of independent data although it is mainly focused on the fusion of data, being quantitative or qualitative, uncertain and imprecise and highly conflictive. The work on this theory is vast, see for example [19] for a review of its applications.

Semantic methods use, not only data but also semantic information given by different sources. The fusion of vague data can be applied to the aggregation of vague expert opinions or to the aggregation of opinions using continuous fuzzy sets [20].

In P2P systems, IF for the improvement of data quality is discussed in [21]. The authors present a model for data integration oriented to their quality. They propose a multilayer structure for P2P systems designed to record data sources and manage the information metadata in a global form. They build on this a model of data quality evaluation based on the data schema and the data's statistical properties. They divide the P2P network into two layers: one for schema management and another for the data sources. Every management node keeps track of a group of sources nodes storing their schemata and metadata. Management nodes communicate to share the information they hold, so the overall structure is a hierarchy. The model of data quality evaluation divided into two: a schema of the data source quality and quality of the service of data quality. To select the data source heuristics rules are applied, for example: choose the source with more records or more integrity schemata.

The literature points out that other ad-hoc techniques can be used, for instance, taking arithmetic averages of data (if they are qualitative) as it is done in [22] or their mode: the fusion to be performed depends on the application. In this work, we compare the effects of choosing the information to be fused in different ways. Consequently, the proposed algorithms that will be described in Section 3 refer to just "perform the fusion", where "fusion" is considered a generic function.

Uncertainty may appear at different levels in a system: in the data, in the fields or sources to be queried, etc. It is important to note that in this work we only addressed the reduction of the intrinsic data uncertainty by using IF to do so. It was also assumed that there is no uncertainty (doubt) about which fields to consult or use in order to get the value of another field.

2.3 Intelligent spaces

We define an Intelligent Space (IS) as a system that supports (helps) people with information in a physical way, for instance, by using robots and monitors [3].

A collaborative object is "a collection of sensors, actuators, controllers and other collaborative objects that communicate between them to reach, more or less autonomously, a common goal" [23] and are part of an IS. Collaborative objects are associated with applications of very diverse kinds, from robots working as a team to wireless sensors networks. Here this notion will be extended and will define a collaborative agent as a collection of sensors (hard or soft, real or virtual), actuators, controllers, persons or other collaborative agents that communicate between them to reach, more or less autonomously, a common goal. It can be seen that a person can be considered a collaborative agent. Using the term "agent" instead of "object" reinforces the idea of "more or less autonomously" through the concept of agency, characteristic of agents.

Lee's definition of intelligent space is met whenever components forming a collaborative agent have a physical existence in a certain space and such existence impacts other components of the space. The Spaces (physical or virtual) populated by collaborative agents will be called here as Generalized

Intelligent Spaces (GIS). For example, when a set of software agents interact to give the best possible answer to an external user who queried them, these agents form a virtual generalized intelligent space

By definition, a collaborative agent will also have properties similar to the ones that physical spaces need to have in order to be considered intelligent:

a) To be designed for people and to ease the ordinary human activities.

b) The information about the status of the agent's components and the events that occurs inside him is caught automatically by the components of the agent.

c) The essential functions of an IS defined in [24] can be generalized to the case of agents and sensors defined here as follows:

   - Observe the space (physical or virtual) in which the agent acts, using distributed sensors of the proper kind (physical, software or virtual).
   - Extract useful information from these obtained data and provide different services.

These generalizations are also valid for GIS. Additional properties of the IS are [24] :

a) Have a function that will carry out the fusion of the information acquired by each sensor and that is shared with other sensors efficiently.

b) Be flexible and scalable.

In these generalized collaborative spaces, it is important to count on an IQ metric of the information provided by each component, which is also easily interpretable and computable be by persons or by machines and which can be used to numeric or symbolic information. In this paper, we will define a metric (or a family of them) for the IQ with these characteristics, of application in a GIS.

3. Model description

In this section, we provide a description of our IQ metric model altogether several properties of it and of the dynamics of the IQ-organized network.

We made two basic assumptions in the definition of the IQ model. First, the agents or peers of the network are heterogeneous, which means that they can by either machines or humans. Second, there is not a special-purpose agent's organization but they can communicate with no restrictions other than the relative to resource availability. We made these two assumptions since it is very common to find real scenarios in which they are represented such as military maneuvers in a hostile land, disaster management or just communities where members collaborate in order to get the best possible answer to a query received from outside. In all those cases, resources are usually limited.

The model is composed of a specification of the computation of the IQ metric and a minimal set of behaviors that all the components need to have in order to make feasible this calculus. A comprehensive description of the network and its behavior when using this IQ model metric can be found in [25].

3.1 IQ metric used

3.1.1 IQ Dimensions considered

Dimensions are aspects of data that can be measured and through which data quality can be described and quantified. They include completeness, validity, timeliness, consistency, and integrity.

In this research, IQ is computed as a single number (avoiding multi-criteria decision making) that is calculated by using a series of formulas that will be described next in this document. We discarded multi-criteria decisión making because it implies to calculate the representativeness (weight) of each criteria and the relative importance of the different combinations, which may lead to a combinatorial explosion that would overload the system with no clear performance gains. In particular, when an agent α queries another agent of the network, the IQ is computed considering the next dimensions:

- The (relative) importance of each part of the information item.
- The importance that is given by agents to the fact that information is first hand, second hand, etc.
- The quality variation obtained in the responses obtained from different agents.
- The amount of queried agents and the number of answers received.
- The uncertainty and vagueness of the answer provided by an agent.

3.1.2 Calculus of IQ

Let α be an agent receiving from another agent denoted Ω a message m with a certain quality which depends on the agent and the field considered. There is a probability that α queries other agents for sub-fields m_i of m (where it can occur that $m_i \equiv m$, that is, α may query for the whole message). The queried agents can, in turn, repeat the process. Upon receiving all answers (or after the corresponding Time Outs – denoted as TOs), the agent α will perform the fusion of the obtained answers. Agent α also receives the number of times the query for m_t was forwarded until reaching α (this number is called here the 'level' of α for that query) and knows the number of consulted agents by it.

In this work, whenever we refer to a sub-field we mean a direct sub-field, reachable in one decomposition step (that is, if A is a sub-field of B and B is sub-field of C, A is not a sub-field of C). Field decomposition in sub-fields is proper of each knowledge domain, of each application, and is assumed unique (common) across all agents.

The quality of field m_t for the agent α is denoted as $Q(\alpha,m_t)$ and it is computed as follows. Let Z be the parameter indicating how important is that an answer is first hand or not. If it does not have importance, it is set to 0 ($Z = 0$); otherwise, $Z > 0$ is taken. The greater the value of Z, the greater the importance of having first-hand data.

Let us denote $\overline{Q}$ as the quality of α when predicting m_t without querying anyone. The value of the quality of α predicting (estimating) m_t, denoted as Q, is computed as a function of the vagueness and uncertainty for m_t in α if α does not query anybody; or as a function considering the vagueness and uncertainties of α, the queried agents which answered, the number of peers queried directly by α, and the value of Z. Analytically, Q is computed as described next.

If *α* does not query any other peer or if no peer answered it, it holds what Equation (3) expresses.

$$Q(\alpha,m_t)\equiv\overline{Q}(\alpha,m_t)=\left(\frac{1}{\left|1+H(\alpha,m_t)\right|(1+U(\alpha,m_t))}\right) \tag{3}$$

Otherwise, Equation (4) applies.

$$Q(\alpha,m_t)\equiv\left[\max\left\{\overline{Q}(\alpha,m_t),q\right\}\right]^{(z+1)/R} \tag{4}$$

Where *q* is calculated according to Equation (5).

$$q=\frac{1}{1+V(\alpha,m_t)}\max_{\beta_j\in T}\left\{\frac{Q(\beta_i,m_t)}{L(\beta_i,m_t)^Z}\right\},\beta_i\in T \tag{5}$$

And *R* is calculated as shown in Equation (6).

$$R=\frac{|T|^2}{S} \tag{6}$$

In previous formulas:

a) L is the number of times that the query was forwarded, calculated as shown in Equation (7). In other words, it increases 1 unit after each forwarding carried out from the original query performed by Ω (starting in 1).

$$L(\beta_k,m_j)=L(\alpha,m_t)+1,\forall k,j,t \tag{7}$$

b) *S* is the number of agents queried by α, no matter if they answer or not.

c) *T* is the set of peers which answered to α about m_t and |*T*| is its cardinal.

d) *V* is a measure of the qualities variation of the received answers by α about m_t, calculated in Equation (8).

$$V=\max_T\{Q(\beta_i,m_t)\}-\min_T\{Q(\beta_i,m_t)\} \tag{8}$$

e) $U(\alpha,m_t)\in[0,1]$ is a measure of the own uncertainty of α about m_t, for example, the percentage of error of α when computing m_t (typically, because of faulty or imprecise hardware).

f) $H(\alpha,m_t)$ is a measure of the datum vagueness.

Next, the quality of field *m* is computed according to α from the qualities of its components using Equation (9), in which *g* is composed of $\{g_1, g_2, \ldots, g_j, \ldots\}$.

$$Q(a,g) = \sum_j W_j(g_j) Q(\alpha, g_j) \qquad (9)$$

where $W_j(g_j)$ represents the weight (the importance) given by the agent to g_j in order to form g.

The parameters of these formulas correspond to the dimensions considered in 3.1.1 as summarized in Table 1.

Table 1. Description of parameters appearing in IQ metric formulas

Parameter	Equation(s)	Description
W	(9)	The relative relevance of each field
Z	(5)	The importance attributed to the fact that the information is of first, second … hand
V	(8)	The variation in the qualities of the answers
S	(6)	The number of queried agents
T	(5), (6)	The number of answering agents
H	(3)	The vagueness proper of α
U	(3)	The uncertainty proper of α
L	(5)	The number of forwards of the query

The importance of information coming first hand, second hand, … is expressed by the use of the variable *L* in the formulas, and the vagueness and uncertainty in the answer given to α by an agent are implicit in the *Q* value. The quality of a composed field is computed by α from the qualities of its components, e.g. as a linear combination of the qualities of the component fields. Also, note that the IQ metric can be applied for both symbolic and numerical data.

3.1.3 Properties of the defined metric

Many authors have sketched the desirable properties of a quality metric, considering some of them the case of information quality. We have designed our IQ metric in order to satisfy the properties mentioned next. In [26] it is stated that a metric should have:

- **Orthogonality**: the ability to represent different aspects of the measured system. The different parameters and variables used in the computation of the metric defined here express the system's characteristics: for example, variable *L* expresses whether queries or parts of them are forwarded to other agents.
- **Formality**: the specification of the metric is precise, objective and not ambiguous.
- **Implementability**: it is independent of the technology used to implement it. This can be seen by analyzing the quality calculus (see 3.1.2).
- **Interpretability**: the ease with which the user can understand, analyze and use properly the results of the metric. In this case, the result obtained is a number in (0,1] being 0 the worst and 1 the best qualities.

Referring to IQ, in [27] the authors mention as desirable properties of the metric:

- **Clarity of definition**: in this case, the definition is clear due to the way it is expressed

- **Measurability**: the metric should be quantifiable in a discreet range ([0,1] in this paper).
- **Representation**: the value of the metric should be able of being represented in a concise and meaningful. A real value between 0 and 1 (as is used here) is the paradigm of it.
- **Drill-down capability**: to be able to show the data that generated the present value of the metric. In our case, this is always possible because every agent knows the received answers and their qualities, the agents that did not answer it, its own quality and its answer, etc.

Finally, other well known properties of quality metrics are:

- **Comprehensible** and **interpretable**. This is valid in this case because the quality is computed in the same way by all the elements of the network.
- **Reproducible**: the measure must be the same if the measurement is repeated in similar conditions. In our case, while there are enough resources for all, the same agents to be queried are chosen and there is no learning in them, the measure will have the same value.
- **Owner of a goal**: there must be a reason to perform the measurement. As it is shown with the different case studies the metric is used here for decision making.

Note that the nature of the above properties is clearly qualitative. In this stage of our research, we have considered these dimensions to be included while designing the metrics but further experimentation needs to be conducted to formally demonstrate their fulfillment. As we will see later, the validation procedure included in this paper aims to prove the validity of the intelligent system using the proposed metric.

3.2 Behaviors

In this section, the main behaviors of the agents are described. Basically, agents are reactive: their actions are described as caused by events. Only the main IQ metric-related behaviors are presented in this section, while a deeper explanation of the rest of the behaviors can be found at [25]. Note that when we talk about the intelligent mode in this paper we are referring to the mode in which the proposed model and metric are used, while simple mode refers to a naïve behavior of the networks with no intelligence incorporated (no specific intelligent criteria or heuristic is used for selecting the peers to query). Also note that a time-out is an event that occurs when an agent does not receive back an answer for a certain query within a certain lapse of time. This may occur to any agent in the network or even to the external querying agent **Ω**. In this case, we may say that the system caused a time-out. As we will see later, this indicator will be used to evaluate the performance of our proposal.

A. **A QUERY REACHES AGENT *Y***: It runs when an agent *Y* receives a query from an agent *Z* which asks about a certain part of the original message (field *m*). Then, *Y* decides whether or not to ask other peers in a random way, for several sub-fields of *m*. These sub-fields (if any) are considered as selected randomly (although the agent can select them in a deterministic way). For this selection, the quality that the agent *Y* has for the sub-field is considered: the lower the quality that an agent has, the greater the probability of consulting. *Y* controls that not too many queries are sent to the same agent to avoid surcharges of agents. If *Y* did not query for any sub-field of *m*, *Y* would respond *Z* with its own estimation of *m*. This behavior is described in Table 2.

Table 2. Definition of behavior "A query reaches agent *Y*".

Name	A query reaches agent *Y*
Description	A query reaches *Y* from agent *Z* asking about field *m* for a message identified as *id.message*
Pseudocode	**Begin** **For** several composed fields of *m* chosen randomly (including the proper whole *m*) **do** Query, with a probability inversely proportional to the quality of the approximation of Y to this subfield, a set of $\hat{N}$ agents different to Z //*It is controlled that there are not too many queries to the same agent.* **End for** **If** there are no queries generated by *Y* pending for sub-fields of *m* and the *id.message* **then** Answer to *Z* about *m* and for *id.message* **End if** **End**

B. **AN ANSWER REACHES *Y***: It runs when an agent *Y* receives an answer (identified as a message *id.message*) from an agent *Z* responding about a certain part of the original message (field *m*). Then *Y* checks if the message arrived before the expiration of the available time (i.e., it is not a time-out). If it was not a time-out and if all the agents queried by *Y* for that field have already answered, *Y* will answer to all the agents who queried *Y* for the sub-field (or a field including it). If it was a time-out, the information coming in the answer is discarded and the occurrence of the time-out is counted. If the number of time-outs is greater than the limit of allowed time-outs, *Y* starts working in the intelligent mode, if not already on it. This behavior is described in Table 3.

Table 3. Definition of behavior "An answer reaches *Y*".

Name	An answer reaches *Y*
Description	An answer comes from agent *Z* to *Y* for the field m_t and the message *id.message*
Pseudocode	**Begin** **If** the answer arrived before the TIMEOUT **then** **If** agents are working in intelligent mode **then** **Count** the number of answers obtained for m_t **End if** **else** **Count** the TIME OUT **If** the maximum number of TIMEOUTS was exceeded **then** **Set** Y to work in intelligent mode **End if** **Let** answer=NULL **End if** **If** all of the agents have already answered for that m_t and *id.message* or it was TIME OUT **then** **Answer** **End if** **End**

C. ***Y* PROVIDES AN ANSWER**: It runs when an agent *Y* has received all the answers for all the queries done for the sub-fields of *m* (or a time-out happened) and computes the value of *m* making a fusion of its own value with the received answers and calculates the quality of the result. In turn, the computed values (value and quality) are sent to those who queried *Y* considering that: a) if *Y* working on intelligent mode, the answer is given always if the quality of *Y* is greater than the querying one; b) otherwise, there is a probability of answering $p= (1-Q)*LR$, being *Q* the quality of *Y* and *LR* the percentage of left resources (typically, messages) of *Y*; i.e., as long as fewer messages are left in *Y*, for a constant quality, it is less probable that *Y* answers when its quality is less than the consultant's one. If no intelligent mode is being used, *Y* always answers. Recursively the agents having pending queries about fields containing *m* respond. These fields containing *m* are called "parents" in the explanation of this behavior shown in Table 4.

Table 4. Definition of behavior "*Y* Provides an answer".

Name	*Y* provides an answer
Description	*Y* responds forming the field *m* which can be calculated now

Pseudocode	**Begin**
	Perform fusion of the answers and the computed value by *Y* *//If it was a TIME OUT the answer is the empty set*
	For each parent of m_t which can be computed now and which was queried by some *Z*
	    Compute the value *V* of the father
	    Compute the quality *Q* of *V* according to *Y*
	    If *Y* is working in intelligent mode **then**
	        If *Q* > querying agent's quality **then**
	            Answer to Z with value *V* and quality *Q*
	        End if
	    else
	        With a probability proportional to (1-Q)*LR, answer to Z with value V and quality Q *//LR is the percentage of resources (typically messages) remaining; LR = 1 if the agent is not working in intelligent mode*
	    End if
	End for
	End

3.3 Dynamics of a system with this architecture

Overall, the system operates when an external agent Ω queries the system about the value (*V*) of a certain information field (*m*), for which that agent already has some knowledge with a certain quality q_Ω. The query is received and managed by the system, being forwarded to the agents of the network $(a_1, a_2, ..., a_n)$ who collaboratively work in order to provide an answer *V* with a certain quality *Q*, as long as the response time is lower than a certain threshold (no time-out) and the provided quality is higher than the one that the external agent had beforehand ($Q > q_\Omega$). Figure 1 depicts this process.

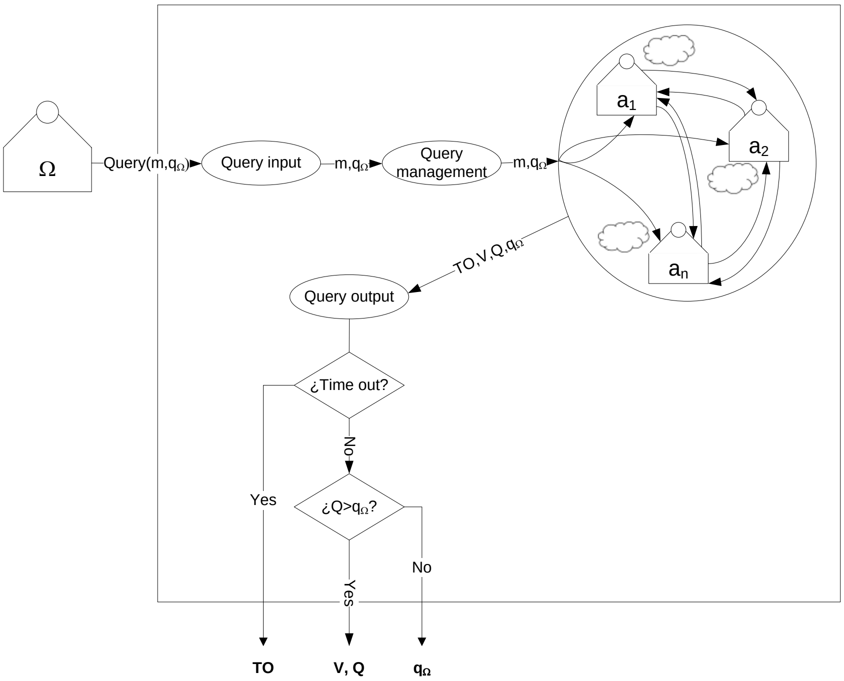


Figure 1. Overall system behavior.

Figure 1 shows that there is a group of agents that process a query (or a part of a query) and provide with an answer (*V*) with a certain quality value (*Q*) in a lapse of time. If an answer is not provided after that time, a Time-Out (TO) is considered to have occurred in the system. The absecnce of an answer may be due to the lack of resources or due to the fact that the answer's quality is lower that the one owned by the querying agent **Ω** before querying the system. In other words, the system does not provide an answer with a quality lower than que original one. Note that this is the behavior in permanent scheme, while stages such as training and peer selection is not depicted in this figure (they are explained next and shown in Figure 2).

The system can be also described by means of a series of stages, as depicted in Figure 2. In the earliest stage, the agents' predictive models are trained in order to be able to provide an answer (classification typically) when queried. After the training stage, the system is ready to receive queries. When the system receives a query, it is presented to the agents of the network who collaborate in a way that there are no preferred partners for each agent. As they start to collaborate and collect knowledge about the mean quality provided by the rest of peers, agents become selective and start to work only with preferred peers, in an attempt to optimize resources. This process repeats as long as there are queries to process. Otherwise, the system will stop.

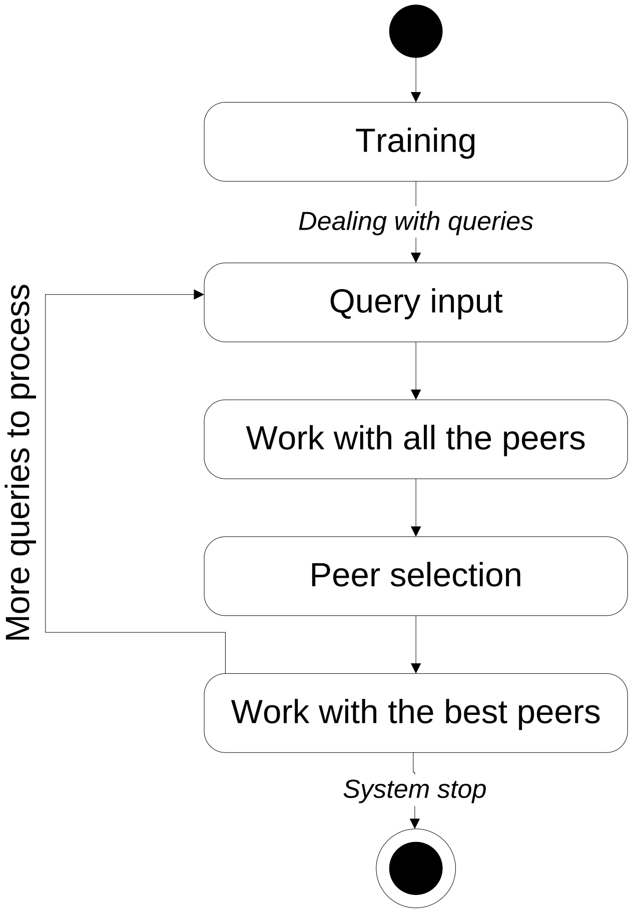


Figure 2. System main stages.

It is important to clarify that we do not consider the notion of a topological neighborhood and therefore we do not take into account how fast, difficult or costly it is for a certain agent to communicate with the rest of agents. Nevertheless, the proposed model has potential since it can deal with distributed data processing, where each agent can focus on a certain data subset, which let the system handle variety and a high volume of data in a reliable and efficient way.

4. Quantitative results

4.1 Performance evaluation

In this stage of our research, we tested our model and the IQ metric proposed in 3 different domains. The selected study cases were: A) the categorization of a certain day as a high-ozone-level day; B) the identification of phishing websites; and C) the decision of buying a car depending on car features such as the price, type of engine, number of doors, safety equipment, and so on [15].

A very basic research question (RQ) was studied for case A: **Given a P2P network with constrained resources in which their agents behave following the communications intelligent model proposed, does the performance (measured using successes obtained and consumed messages) improve?.**

Considering the research question, the following metrics were used for processes evaluation:

- Success rate: in the context of P2P networks, success is achieved when a resource needed to satisfy a requirement is found. In our rearch, we have adapted this indicator so it considers the possibility of answering a query with a quality higher than the original one. In our case, success will occur when achieving an answer with quality greater or equal than the quality initially given by Ω. A success is called strict when the quality of the answer is strictly greater than the one of Ω, and it is called broad otherwise. Strict successes are interesting because they correspond to cases in which the work of the network was really useful. In the results shown below strict successes are denoted as “success” for brevity.
- Mean traffic: is the average of the messages used by the agents (which can be related to the cost of the query) for a message received from Ω or of the used messages needed to give an answer with a minimum quality.
- Number of time-outs (TOs): number of times in which the network does not answer to Ω in time. The number of TOs have a direct impact on the quality of service, as stated in [28].
- Effect of the agents’ quantity and available messages in the network on the success rates. The number of agents is related to the scalabity of the network, an issue deeply analyzed in the literature, as stated for example in [29]. Those works also analyze how the networks behave when increasing the number of available resources (messages, for instance), similarly to what we have done in our experiments and will describe next.
- Effect of the agents’ number and available messages on the TOs.

In cases B and C, due to the nature of data (wider and with more variety), the research question was split into two sub-research questions, to check whether: a) the intelligent mode requires fewer messages for Ω to get an answer of a given quality and b) the intelligent mode has a greater success rate.

4.2 General results

In this section, we present our overall results. Our model was instantiated and applied to the three case studies mentioned, whose data were obtained from the University of California in Irvine public repository [30]. The simulation of the proposed model was programmed in Java and Python, considering from 5 up to 1000 agents in the different runs. We conducted simulations due to the unavailabity of the massive hardware resources neccessary to support hundreds of agents running concurrently.

The null hypothesis "H_0: the system using the model described here performs equal or worse than a system which does not consider information quality to optimize communications" was rejected in most experiments for all case studies (see Table 5). Note that it was not always possible to get a 100% of rejections in all experiments of the three cases, possibly due to the fact that the three cases used the same values of the model parameters (better results could have been obtained tuning the model separately with an adequate set of parameters values for each case).

Table 5. Overall results obtained

Case of study	H_0 Rejection (%)	Number of groups of experiments
A – Ozone levels	80	3
B – (Web)Sites phishing	87.5	12
C – Vehicle recommendation	87.5	12

For case A, with a simpler and smaller dataset, we conducted 3 experiments to measure how the number of successes and the number of consumed messages evolve with different values of the number of agents and minimum expected quality, comparing the simple (not taking any performance consideration when choosing the best peers to be queried) against the intelligent mode. In cases B and C, more complex and larger, we could conduct more comprehensive experiments (12 in total) in which we measured the evolution of different performance indicators (TOs, successes, messages employed) for various combinations of different factors (minimum number of messages, initial number of messages, maximum number of messages, number of agents, minimum expected quality). This experimental setup has already been validated and used by the authors as shown in [25], where more details about it can be found. The values presented to describe the behaviour of the system in both simple and intelligent modes have not been taken anywhere else from outside but they have been originally obtained in our research.

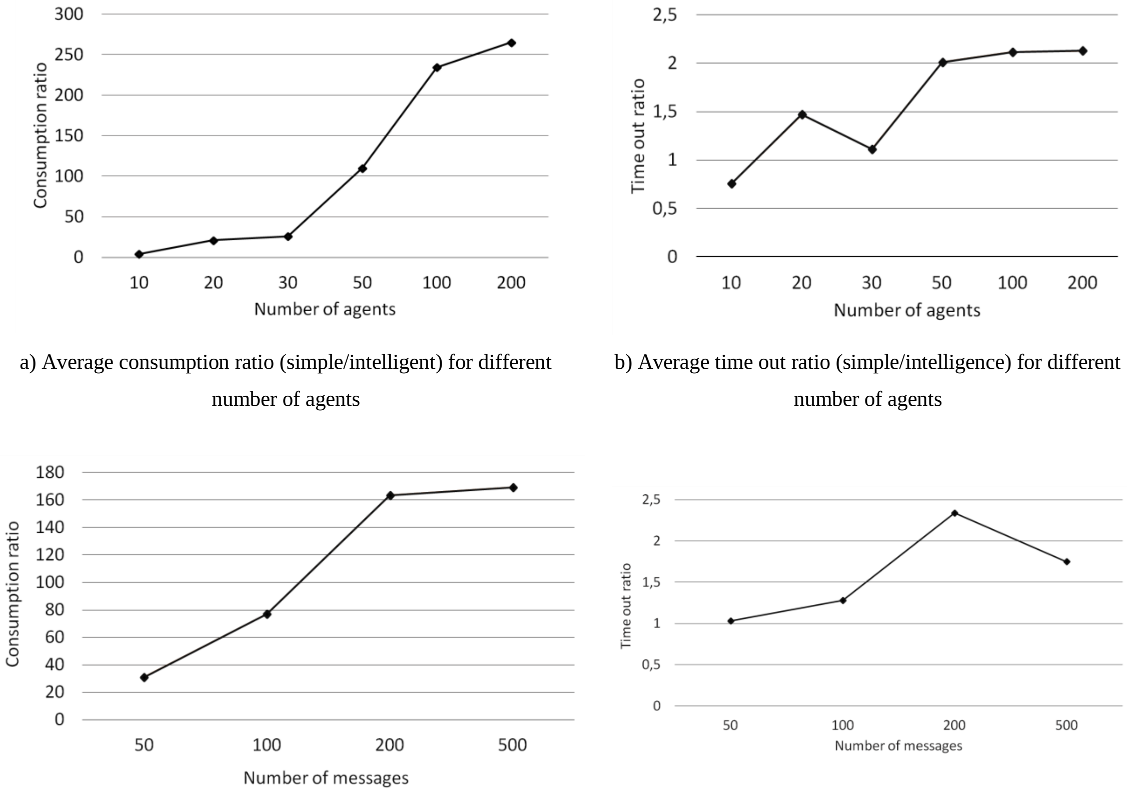
4.3 Detailed results

In this section, the results for each case study are discussed. We omit the details on the statistical tests made for the sake of conciseness.

4.3.1 Case A

The number of required messages to answer a set of queries coming from Ω was studied to check: a) whether or not fewer messages are required to get an answer of a certain quality when using the intelligent model, and b) whether or not a higher success rate is obtained when using the intelligent model. Each test was performed for 10, 20, 30, 50, 100 and 200 agents using 50, 100, 200, and 500 messages. The results are shown in Figure 3. It is important to focus on the performance of the system when the number of agents or initially available messages increases. The graphics show the quotient of the values Simple/Intelligent. Figures 3-a and 3-c (consumption, this is the number of messages used for strict success in the *y*-axis versus the number of agents or messages in the *x*-axis) highlight that the intelligent mode has advantages over the simple one when the number of agents or available messages increases because the tendency of the quotient has a positive slope. Similarly, Figures 3-b and 3-d (average number of TOs versus the number of agents or available messages) show how the TOs increase with the augment in the number of agents or the available messages in the system.

This research question is also related to the amount of queries that are not answered due to TOs. The intelligent mode reduces systematically the number of TOs (messages that are not answered because the preset time to answer the query expires), averaging on the different number of messages initially available; symmetrically, processing using intelligent mode always reduce the quantity of TOs averaged on all the agents. Therefore the results were positive, confirming the RQ for this case study.



a) Average consumption ratio (simple/intelligent) for different number of agents

b) Average time out ratio (simple/intelligence) for different number of agents

c) Average consumption ratio (simple/intelligent) for different number of messages

d) Average time out (simple/intelligent) for different number of messages

Figure 3. The behavior of the system when the agents or messages increase (Case A).

4.3.2 Case B

The experiments performed in this case of study have been described in section 4.2. The following aspects were studied:

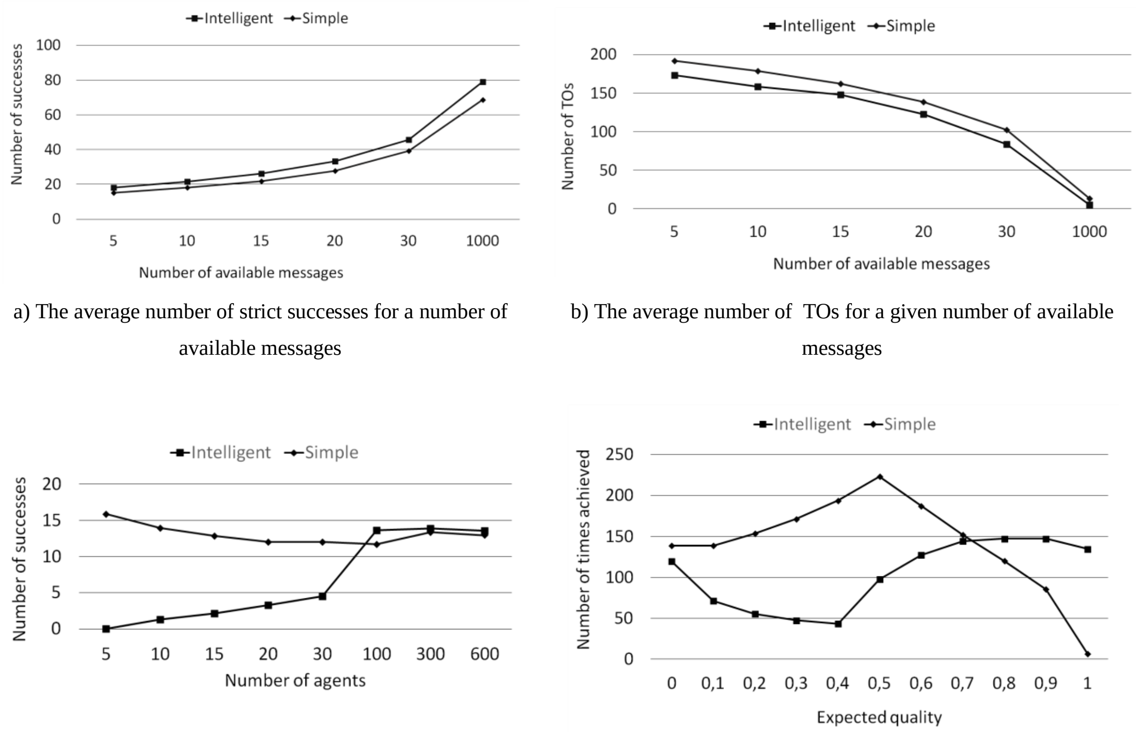
- The average number of strict successes for a given number of messages (Figure 4-a).

- The average number of TOs occurring for a given initial quantity of messages (Figure 4-b).
- The average number of successes for a certain number of agents (Figure 4-c).
- The frequency of appearance of a quality greater or equal to α, for a given α (Figure 4-d).

Although our tests show that the system gives results which are strict successes more frequently when it works in intelligent mode, this is true only for the lower qualities.

Referring to the number of queries which are not answered due to TOs, note that the intelligent mode reduces the number of TOs considering these TOs averaged on the different numbers of available messages; symmetrically, the operation in intelligent mode always reduce the number of TOs averaging over all the numbers of agents. It was found that the number of strict successes that can be obtained by changing the number of agents in case B is always greater in mode intelligent.

The statistical experiments summarized in Figures 4-a to 4-d, which shows part of our findings, confirm the advantages of using the intelligent mode and the validity of the IQ metric proposed in this paper.



a) The average number of strict successes for a number of available messages

b) The average number of TOs for a given number of available messages

c) The average number of successes for a given number of agents

d) Frequency of obtaining a minimum quality

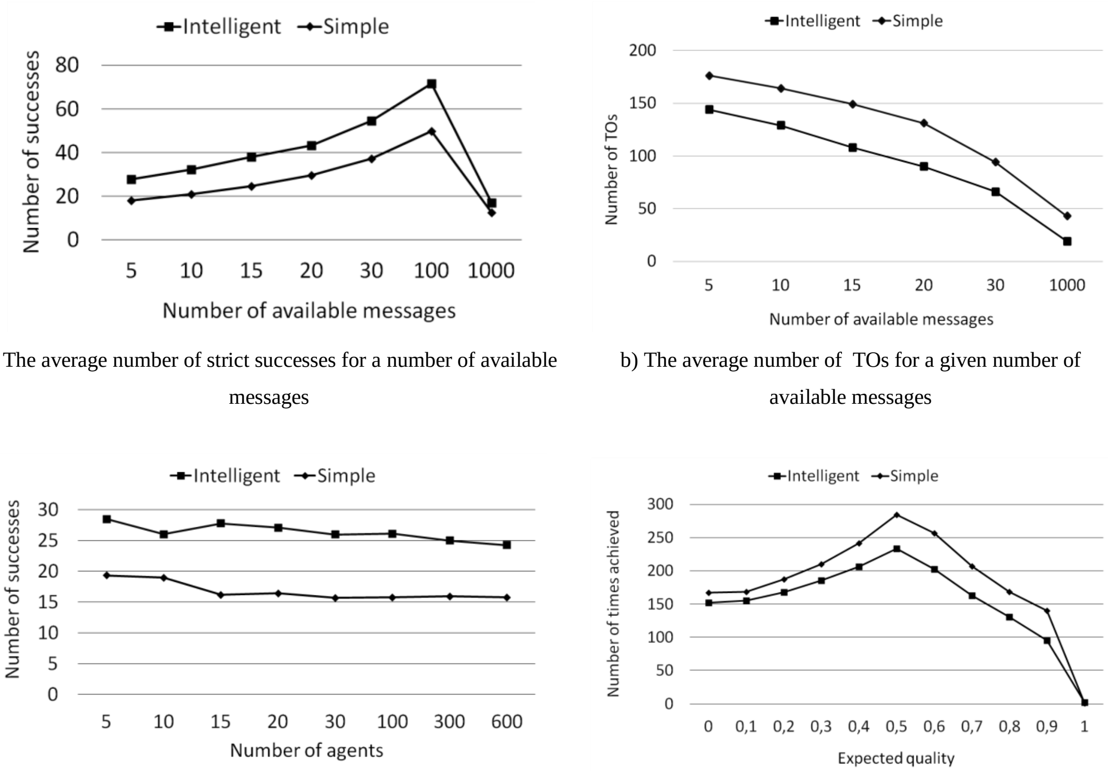
Figure 4. Detailed results obtained for case B.

Figure 4.a shows that the number of successes is always higher in the intelligent mode for any of the scenarios considered when varying the number of available messages. Figure 4.b confirms that the higher the number of available messages, the higher the system's performance is (the lower number of TOs that will occur). We found that the intelligent mode provides better results in terms of success rate when there are more than 100 agents working on the system (see Figure 4.c) but we could not confirm this hypothesis with the statistical tests performed (not included for the sake of simplicity). Finally, Figure 4.d shows that the intelligent mode obtains high qualities (0.7 or above) more frequently than then simple one.

4.3.3 Case C

The behavior of the proposed model proved to be good in this case, too. A similar analysis to the one conducted in case B was performed. From the results, it is noticeable that the intelligent mode produces high qualities (of 0.7 or more) more frequently than the model simple.

In this case, the intelligent mode looks preferable when there are more of 100 agents in the system (network) although this could not be proved statistically. The next graphs (Figures 5a-5d) suggest that the question is also valid in this case C provided there are certain constraints in the number of agents.



a) The average number of strict successes for a number of available messages

b) The average number of TOs for a given number of available messages

c) The average number of successes for a given number of agents

d) Frequency of obtaining a minimum quality

Figure 5. Detailed results obtained for case C.

5. Conclusions and Future Work

This paper proposed an IQ metric that can be used in distributed models based on flat P2P networks whose elements perform information fusion and are organized taking into account the quality of the exchanged information. It is understandable that the purpose of the proposed model is aligned with sense-making theory since the quality metric proposed in this paper has been designed in a way that considers both information's uncertainty and vagueness.

Our proposal yields positive results in different aspects. The conducted experiments have demonstrated to outperform performance (successes and time-outs, mainly) when using the IQ metric proposed in comparison with a similar flat model with no use of the IQ and the intelligent mechanisms proposed in this paper.

Future research lines are listed below:

- To evolve the proposed IQ formulas so agents' learning in time is reflected. This way, the calculation would be a dynamic process.
- In our model, agents are unaware of what happens with their responses, whether they are actually useful, whether they reach the destination in time, and so on. It would be interesting to count on a feedback mechanism so agents can adapt their behavior depending on the received feedback after providing responses.
- In our research, we assumed no uncertainty about the composition of information fields. It would be interesting to apply information fusion techniques to automatically discover the information field structure and possible changes in it occurring during the time the system is working.
- To apply the proposed model and its IQ metric to other application domains in order to confirm the good results obtained.

References

[1] Park, H., R. Izhak-Ratzin, M. Schaar, C. Zhu, Y.-n. Li, and X.-m. Niu, Peer-to-Peer Networks - Protocols, Cooperation and Competition. 2010.

[2] Meer, H.D. and C. Koppen, Self-organization in Peer-to-Peer Systems, R. Steinmetz and K. Wehrle, Editors. 2005, Springer.

[3] Gadeo-Martos, M.A., J.A. Fernandez-Prieto, J. Canada-Bago, and J.R. Velasco, An Architecture for Performance Optimization in a Collaborative Knowledge-Based Approach for Wireless Sensor Networks. Sensors, 2011. 11(10): p. 9136.

[4] Knight, S.-a. and J. Burn, Developing a Framework for Assessing Information Quality on the World Wide Web. Informing Science, 2005.

[5] Sevtiyuni, P. E., Oktadini, N. R., and Bardadi, A. Information Risk Assessment Model of Accuracy and Timeliness Dimensions. in Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019). 2020. Atlantis Press.

[6] Janssen, M., H. van der Voort, and A. Wahyudi, Factors influencing big data decision-making quality. Journal of Business Research, 2017. 70(C): p. 338-345.

[7] Novák, V., Are Fuzzy Sets a Reasonable Tool for Modeling Vague Phenomena? Fuzzy Sets and Systems., 2005. 156. pp. 341-348.

[8] Paggi, H. and M. Cochez, Indeterminacy Reduction in Agent Communication Using a Semantic Language. WSEAS Transactions on Systems, 2015. 14: p. 77-89.

[9] Bossé, É. and B. Solaiman, Information Fusion and Analytics for Big Data and IoT. 2016: Artech House Publishers.

[10] Zadeh, L., Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. SMC, 1973. 3(1): p. 28-44.

[11] Ekel, P., I. Kokshenev, R. Parreiras, W. Pedrycz, and J. Pereira Jr, Multiobjective and Multiattribute Decision Making in a Fuzzy Environment and Their Power Engineering Applications. Inf. Sci., 2016. 361(C): p. 100-119.

[12] Hsu, H.-M. and C.-T. Chen, Aggregation of Fuzzy Opinions Under Group Decision Making. Fuzzy Sets Syst., 1996. 79(3): p. 279-285.

[13] Ayyub, B.M. and G.J. Klir, Uncertainty Modeling and Analysis in Engineering and the Sciences. 2006: CRC Press.

[14] Akbar, R.M.I., R.A. Sularso, and K. Indraningrat, The Effect of Price, Ease of Transaction, Information Quality, Safety, and Trust on Online Purchase Decision. e-Journal Ekonomi Bisnis dan Akuntansi, 2020. 7(1): p. 77-81.

[15] Nakamura, E.F., A.A.F. Loureiro, A. Boukerche, and A.Y. Zomaya, Localized algorithms for information fusion in resource constrained networks. Inf. Fusion, 2014. 15: p. 2-4.

[16] Foo, P.H. and G.W. Ng, High-level Information Fusion: An Overview. J. of Adv. in Inf. Fusion, 2013. 8(1).

[17] Fine, T.L., Review: Glenn Shafer, A mathematical theory of evidence. Bull. Amer. Math. Soc., 1977. 83(4): p. 667-672.

[18] Smarandache, F. and J. Dezert, The Combination of Paradoxical, Uncertain, and Imprecise Sources of Information based on DSmT and Neutro-Fuzzy Inference. CoRR, 2004. abs/cs/0412091.

[19] Florentin Smarandache, J.D., Advances and Applications of DSmT for Information Fusion, Vol. IV: Collected Works. 2015.

[20] Bai, Y.-T., B.-H. Zhang, X.-Y. Wang, X.-B. Jin, J.-P. Xu, T.-L. Su, and Z.-Y. Wang, A Novel Group Decision-Making Method Based on Sensor Data and Fuzzy Information. Sensors (Basel, Switzerland), 2016. 16(11): p. 1799.

[21] Zhao, Z. Data Quality-Oriented Data Integration in Peer-to-Peer System. in 2009 Ninth International Conference on Hybrid Intelligent Systems. 2009.

[22] V K, S. and C. Tharini, An Energy Efficient Routing and Fault Tolerant Data Aggregation (EERFTDA) algorithm for wireless sensor networks. Vol. 23. 2017. 15-32.

[23] Michel Banatre, P.J.M., Anibal Ollero , Adam Wolisz Cooperating Embedded Systems and Wireless Sensor Networks. 2008: Wiley. 384.

[24] Sasaki, T., D. Brscic, and H. Hashimoto. Implementation of Distributed Sensor Network for Intelligent Space. in 2007 IEEE International Conference on Mechatronics. 2007.

[25] Paggi, H., J. Soriano, and J.A. Lara, A multi-agent system for minimizing information indeterminacy within information fusion scenarios in peer-to-peer networks with limited resources. Information Sciences, 2018. 451-452: p. 271-294.

[26] Mustafa, P.K. and P.R. Khan, Quality Metric Development Framework (qMDF). Journal of Computer Science, 2005. 1.

[27] Inc., K.I. Developing Information Quality Metrics. 2005; Available from: http://www.knowledge-integrity.com/columns/dmr200505.htm.

[28] Kapur, A., N. Gautam, R. Brooks, and S. Rai, Design, performance and dependability of a peer-to-peer network supporting qos for mobile code applications. Small, 2002. 10(20): p. 30.

[29] Ge, Z., D.R. Figueiredo, S. Jaiswal, J. Kurose, and D. Towsley. Modeling peer-peer file sharing systems. in IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428). 2003. IEEE.

[30] Lichman, M., UCI Machine Learning Repository. 2013.

Horacio Paggi holds a Ph.D. from the Polytechnic University of Madrid, Spain. He also holds a Master’s in Computers Engineering and a Master’s in Mathematical Engineering from the University of the Republic (Uruguay). His actual research field is multi-agent systems and information quality. He works at the Innovation and Knowledge Management Department of the Uruguayan State oil&gas company (ANCAP).

Javier Soriano is Associate Professor at Universidad Politécnica de Madrid, UPM, Spain. He is Director of the Computer Networks and Web Technologies Laboratory. He holds a Ph.D. with Honors in Computer Science from UPM. He has coauthored more than 60 papers published in international impact journals, research books and conferences. His research focuses on distributed systems and future Internet technologies.

Juan A. Lara is Associate Professor at Madrid Open University, UDIMA, Spain. He is member of the Department of Computer Science. He holds a Ph.D. in Computer Science. He is author of more thirty papers published in international impact journals. His research interests in computer science include data mining, knowledge discovery in databases, data fusion, artificial intelligence and e-learning.

Ernesto Damiani is Full Professor at Università degli Studi di Milano, Senior Director of Artificial Intelligence and Intelligent Systems Institute, Khalifa University, and President of the Consortium of Italian Computer Science Universities (CINI). He has authored more than 130 journal papers. His areas of interest include Artificial Intelligence, Machine Learning, Big Data Analytics, Edge/Cloud security and performance, and cyber-physical systems.