Data Quality and Trust : A Perception from Shared Data in IoT

John Byabazaire*, Gregory O'Hare*, Declan Delaney†
*School of Computer Science, University College Dublin, Ireland
†School of Electrical and Electronic Engineering, University College Dublin, Ireland
john.byabazaire@ucdconnect.ie, { declan.delaney, gregory.ohare}@ucd.ie

***Abstract*—Internet of Things devices and data sources are seeing increased use in various application areas. The proliferation of cheaper sensor hardware has allowed for wider scale data collection deployments. With increased numbers of deployed sensors and the use of heterogeneous sensor types there is increased scope for collecting erroneous, inaccurate or inconsistent data. This in turn may lead to inaccurate models built from this data. It is important to evaluate this data as it is collected to determine its validity. This paper presents an analysis of data quality as it is represented in Internet of Things (IoT) systems and some of the limitations of this representation. The paper discusses the use of trust as a heuristic to drive data quality measurements. Trust is a well-established metric that has been used to determine the validity of a piece or source of data in crowd sourced or other unreliable data collection techniques. The analysis extends to detail an appropriate framework for representing data quality effectively within the big data model and why a trust backed framework is important especially in heterogeneously sourced IoT data streams.**

I. Introduction

The increase in the deployment of sensors comes with the increase in the volumes of data collected. It is estimated that by 2020 40 Zettabytes (or 40 trillion gigabytes) of data will have been generated and consumed [1]. Large volumes is one of the characteristic of Big Data. Other characteristics include; data velocity, variety, veracity and value. This data is collected and goes through a series of stages until it is used to inform decision making, control processes or visualisation. These stages form the big data model.

The big data model is a series of stages that data goes through from the time it is created to when it is visualized. Each of these stages is critical for the success of the other. Figure 1 shows the various stages of the big data model. Data collection, data pre-processing, data processing, and data use comprise the stages of the big data model. It is important to note that for each stage, data quality can have different features and representations. This is equally true for different data users and applications within the IoT ecosystem.

The IoT ecosystem has grown to incorporate everything in our surrounding from smart homes, smart cities, manufacturing to environmental sensing. Each of these application areas generate and consume data. Currently both research and industry are harnessing the opportunities of sharing and consuming data across various domains of the IoT ecosystem in what we refer to as shared IoT. In Fig 2, we show how for example, a smart city application generates and consumes its own data but can also benefit from data fusion from other IoT applications for better insights. To benefit from this, it is important to ensure that the shared data conforms to certain quality standards and can be trusted by the consuming application.

As data is at the center of inferring new insights, it is important to assess the quality of the data from which decisions are made. Poor understanding of the quality of the data can lead to poor decisions. Quantifying, understanding and making these data quality issues visible throughout the big data model is essential for effective insight. A tangible link between data quality, data quality types and their effect on the data through the stages in the big data model is however, yet to be defined.

In this paper, we present a data quality assessment framework based on trust that can be used to assess the quality of data in cases where there is no reference data to compare to. The framework also shows how we can incorporate the big data model in the data quality assessment process and provide visibility of data quality throughout.

II. Background and Related Work

A. Data Quality

Data is readily available to most companies. This is being used to drive decisions, create new products and expand markets. It therefore becomes very important to determine the quality of data that is being used to foster such decisions and actions. Data quality has widely been studied in database management [2], [3], [4] and also in the big data context [5]. Problems resulting from poor data quality can have bad implications on business decisions [6]. In an IoT context, various factors are responsible for the degrading data quality [7]. Some of these include; deployment scale, resources constraints, fail-dirty, security vulnerability, privacy preservation processing.

Data quality is subjective making it dependant on the use case and domain area. It is understood differently in academia and industry [8]. Sidi et. al. [9] defines data quality as the appropriateness for use or meeting user needs. According to Heravizadeh et. al. [10], quality means the totality of the characteristics of an entity (data) that bear on its ability to satisfy stated and implied needs.

The quality of data is highly dependant on the intended use. This is a multidimensional concept that is hard to assess because every user defines their own quality attributes. These

978-1-7281-7440-2/20/$31.00 ©2020 IEEE

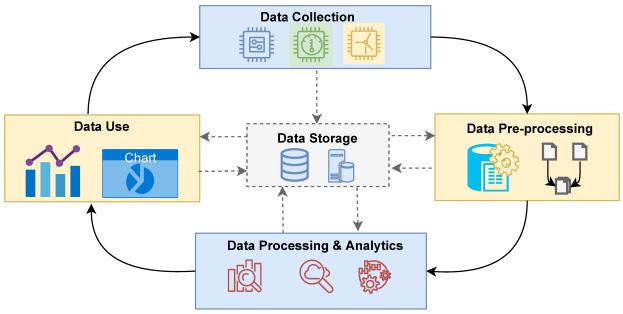

Fig. 1: The Big Data Model

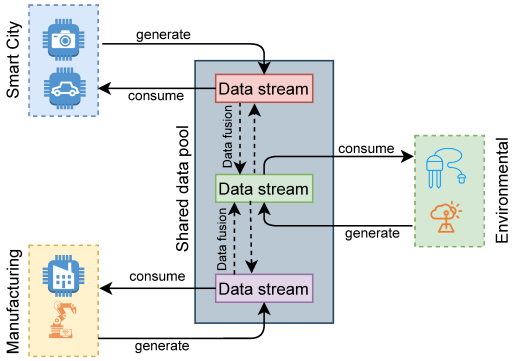

Fig. 2: Shared IoT Ecosystem

quality attributes are collectively know as data quality dimensions [11]. Examples of these include but not to the following; accuracy, accessibility, timeliness, believably, relevancy. Each of these and others have been fully defined in [11].

B. Data Quality Dimensions

Data Quality Dimensions (DQD) provide an acceptable way to measure data quality. Several authors have defined different DQD and each with an associated metric [8]. A DQD is a characteristic or feature of information for classifying information and data requirements. As such, it offers a way for measuring and managing data quality as well as information [9]. It is important to note that there is no standard definition of DQD that is acceptable as domain independent [12]. Some of these could be task independent, therefore not restrained by the context of application while others are task dependent [13]. Lee et. al. [11], studied many of these and summarised them into four main categories as show in the table I.

TABLE I: Data Quality Dimension Categories

Data Quality category	Data Quality dimensions
Intrinsic	Accuracy, Objectivity, Believability, Reputation
Accessibility	Accessibility, Access security
Contextual	Relevancy, Value-added,Timeliness, Completeness, Amount of data
Representational	Interpretability, Ease of understanding, Concise representation, Consistent representation

C. Trust

Trust has been widely studied in many disciplines from sociology [14] to computer science [15]. Each of these considers and defines it differently. In general, trust is a measure of confidence that an entity will behave in an expected manner, despite the lack of ability to monitor or control the environment in which it operates [16]. Different users have different requirements they consider to trust a data source. Examples of these include reliability, competence, credentials, reputation.

In the era where information is widely available, users are tasked to find a way of gauging the quality of such. These data source build a reputation and over time become trustworthy. Trust in itself is a process, therefore trust can be formed, improved and also lost. Some data sources have built trust over time and now they are more trusted than others.

Quality assessment requires us to compare something (data) to something we consider a gold standard. For example the performance of a certain sensor has to be compared to another to determine if it is working well. On other hand, trust offers us an opportunity to estimate data quality when no other information is available to use as a gold standard.

D. Data Quality as a Measure of Trust

In a wider sense, trust has been used as a measure of quality especially in information systems. It is assumed that if more people trust a product or service, the better the quality of such and vice versa. This same principle has been used widely in information search on the internet and more recently in recommender engines.

Like trust, quality is an iterative process that must constantly be re-assessed. To achieve a certain level of trustworthiness different trust attributes (reliability, competence) must be evaluated at every stage and how these contribute to each other.

E. State of the Art

In the IoT ecosystem, several devices (sensors and actuators) are deployed to collect and act on data. Based on this data, predictive models are built to automate decision making across different domains. For these autonomous processes to achieve the desired outcomes, the data from these devices should be trusted to be of a certain quality threshold depending on the purpose for which it is intended.

Several approaches in literature have been proposed to ensure that data retains its quality across the big data model. In these, some have advanced a data centric approach by trying to mitigate the errors in the data itself [17], [18], [19]. For example detecting of anomalous data points, automatic detection of faulty sensors. Others have proposed a process centered approach where they check the data collection process [9], [20], [21], [22]. For example, the experiment procedure, the kind of equipment used. From the network perspective, a framework was proposed that looks at QoS for multiple

applications in IoT systems by harnessing machine learning techniques [23]. In this section will group these based on the data quality dimension categories proposed by [11].

1) ***Intrinsic****:* This category looks at quality properties in the data itself, for example accuracy, believability. Efrat et. al. [18] looked at a multivariate anomaly detection technique for ensuring data quality of dendrometer sensor networks. The anomalous sensors are identified statistically by comparing a sensor's readings to an expected reading from a similar, healthy sensors network. As a gold standard, companys experts used the system on a regular basis to verify the classifications created by the anomaly-detection algorithm.

Tsai et. al. [17] proposed an abnormal sensor detection architecture that leverages machine learning techniques. They trained a Bayesian model that can predict values of sensor nodes via other correlated sensors. Their results show they can detect abnormal sensors in real-time. They also analyze sensor data patterns and the Bayesian model's estimate log to classify the error type of sensors. By error type, they then derive the formula to recover the faulty sensor reading, so that it can increase the reliability of sensing system.

2) ***Representational and Accessibility****:* This addresses the computer systems that store and provide access to information. Such systems must ensure that the data is easy to understand and easy to manipulate. Fatimah et. al. [9] proposes an efficient data quality evaluation scheme by applying sampling strategies on big data sets. The results showed that the mean quality score of samples is representative for the original data and illustrate the importance of sampling to reduce computing costs when Big data quality evaluation is concerned. Most of the work in this category has widely been studied in database management system [2], [3], [4].

3) ***Contextual****:* This looks at quality properties that must be considered within the context of the task at hand. For example, it must be relevant, timely, and appropriate in terms of amount. In small data enterprises, data is collected based on the problem at hand. In such scenarios it is not important to add context to the data because it will be constrained to that problem. In large data enterprises were different fusion sources are combined and the data has to used to solve several problems, adding context to such data becomes paramount. Faniel et. al. [20] emphasise the importance of context of the data. To the best of our knowledge no solutions have considered the inclusion of context while assessing data quality.

Current solutions take a decentralised approach where by they try and solve data quality at a given stage of the big data model and most of them have shown promising results. The problem here is that ensuring data quality should be a process that should be re-evaluated at every stage of the big data model. This would also help show how different big data model stages would contribute to the overall quality score of a given dataset.

III. Existing Challenges

Our approach to data quality assessment is to investigate how each stage in the big data model is affected and how this affects the other stage. To understand how data quality measures proliferate and affect data use cases, we must first understand the relationship between data quality and the big data model. Thus far the literature does not consider data quality a fundamental aspect of the big data model. A challenge exists with regard to structuring data quality dimensions within the big data model so that the effect of data quality is visible throughout the model. We call these structure related challenges.

With a given data quality structure in mind, considerations on how data quality measurements from one stage of the big data model can and should affect data quality at other stages in the big data model. This may involve combining or weighting quality measurements for a given stage or use case. A number of challenges exist in this space which we call challenges of methods.

A. Structure related challenges

- How do we define a data quality assessment framework where all the data quality factors present in the data cycle are represented. Data quality is not an isolated aspect affecting only a certain stage in the big data model. Taleb et. al. [5] also conclude that there is a need for assessment of quality as earlier as possible (data inception stage) and an end-to-end quality assessment model that combines all the big data model phases. This paper addresses this challenge by proposing such a model.

B. Method related challenges

- Data quality can be highly subjective. A single data point or source can have varying quality depending on the use case context. How might data quality be represented in a general manner throughout the big data model yet allow subjective handling.
- Data quality is measured and represented in different ways depending on the stage and context within the big data model. How can these data quality measures be combined across the big data model stages to infer a quality metric which is useful for use case quality determination.

IV. Motivation

This section serves and provides the motivation for introducing trust as a driver for data quality measurement and for incorporation data quality into the big data model.

Quality is a complex, multidimensional and continuous process [24]. Furthermore, it is important for the users of data to have visibility of the quality factors from the initial stages in the Big Data Model (collection phase) [20]. This motivates the need for a structured approach to handling data quality within the big data model. We will discuss trust and data quality mapping within the big data model individually.

A. Trusted data:

Whilst data quality metrics are important, they still present challengs; firstly there is a non exhaustive set of features to

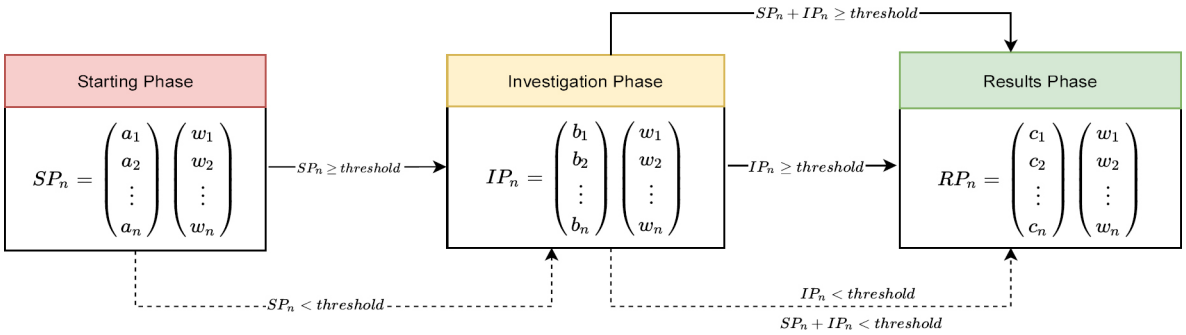

Fig. 3: Proposed Framework that is based on Trust formation stage

define data quality for any use case. Secondly, data quality features are difficulty to measure in the field without an explicit gold standard measurement. In this case the use of trust can be used as an indicator for measuring data quality. A trust process can build a good quality reference measure over time without the need for a gold standard. This is typically the case in most IoT deployments where sensors are placed in an adhoc manner and are expected to function correctly without oversight.

Trust has been used previously as a measure of data quality. Keßler et. al.[25] studied how the quality of geographic information can be estimated through the notion of trust as a proxy measure. To evaluate the trust measure, its result were compared to results obtained from a quality field survey based on three data quality dimensions; accuracy, consistency and completeness. Their results showed that data quality can be estimated using a trust model based on data provenance.

B. Data quality mapping:

There is no effective means to bring data quality through the big data model. It is important for data quality from all stages of the big data model to be reflected effectively in the end use. There is currently no quality assessment model that can achieve this. Such a model needs to characterise data quality at each individual stage but longitudinally through the entire big data model.

This paper presents a formal data quality assessment framework for use within the big data model. This allows us to determine a quality metric at each stage of the big data model. This would also enable us to model the effect of this metric on the next stage of the big data model. The addition of weights to each parameter in the model allows each effect to tuned as per use case.

V. Proposed Framework

Quality and data trust are continuous processes which must be assessed through out the data life cycle, from when the data is created to when it is consumed. Quality like trust can be improved over time. Having the best equipment setup does not guarantee quality data. Also the exclusion of the context of the data as a factor that affect data quality. This is especially important in application areas like precision agriculture. For example while applying fertilizers to a field, it is more important to determine how this was done rather than how much was applied. The quality of such a figure (quality applied) can only be assessed if its context is taken into consideration.

Based on the natural norm of trust that it can be improved over time and result into good quality data whose quality threshold is acceptable for a certain use case, we propose a data quality assessment framework shown in Fig 3.

A. Trust Formation

Trust has been widely used as a measure of quality in information sciences, in cloud computing based on Quality of Service (Qos) [26] and the internet in general. User have been assessing the quality of information based of some kind of intrinsic trust evaluation. Trust itself is a continuous assessment process. Along this process, trust can be formed, improved or lost. While users are assessing information, they go through a series of steps and the overall decision is based on the incremental evaluation from the various stages. To this end we defined three (3) trust stages that a user goes through to when they are evaluating data. This was also informed by the work done in [20]. Fig 4 shows how these relate to the big data model.

- Initial Trust: this is trust that derived without looking at the data itself but rather the context of the data. Here things like the source, equipment/sensors used, documentation are assessed.
- Investigation Trust: If the user is satisfied with the previous stage, they then go ahead and assess the data itself. Here, they see if the data has any error. For example, missing values are checked and initial the pre-processing is done.
- Result Driven Trust: if the data has been used in any other models/analytics, then the user will go ahead and look at these.

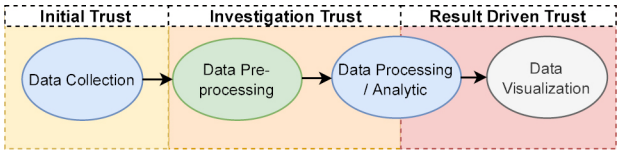

Fig. 4: Flow diagram showing the relationship between Big Data Model and Trust Formation Stages

B. Framework Phases

We define three phases that are inspired by the user trust formation. Naturally users have been assessing the quality of data. First the user will evaluate the context of the data by looking at things like; where did the data come from (reputation of the source), documentation about the data. Secondly, if the user is satisfied with the evaluation above, then they examine the data itself. Here the user looks at any errors that could be in the data. Here things like missing values are examined, pre-prosessing is also done at this stage. Finally, the user then look for any results in form of analysis or models that has come from this data.

Similarly we define three distinct phases. At each of the phases, we define parameters a_1 to a_n that are contained at each phase. Table II list a non exhaustive list of some of the parameters that can be considered.

TABLE II: Parameters at each Phase

Phase	Parameters
Starting Phase	Equipment setup, Documentation, Author of the data
Investigation Phase	Missing values, Number of parameters, Data size
Results Phase	Successful model/analytics, Publications, Visualizations

C. Determining Weights

Once parameters have been determined at all phases, then weights to each of parameters are calculated. Depending on the use case, each of the parameters is weighted differently. Overall, with time the goal is to have a learning model to automatically determine the weights at each phase. Once these have been determined, a Phase Index (PI) is calculated. For example SP_n for the starting phase. If the PI is greater than the threshold started for a given use case, then we proceed to the next phase. However if the PI is less than the stated threshold, the process can be terminated or this PI can be added to the PI in the next phase as way of improving trust at the level.

D. Why the Trust Framework is Important

Focusing on the challenges of the current data quality assessment methods in IoT highlighted above, and motivated by the trustworthiness evaluation techniques in social media [27], [28], recommender systems [29] and multi-agent systems [30], this work has proposed an innovative end-to-end trust based framework that can be used to assess the quality of heterogeneous IoT data streams. Various considerations have been made to reflect the complexity and uncertainty characteristics of trust. We therefore believe that this framework can be able to offer the following key innovative features to the IoT data quality eco-system:

- Continuous and end-to-end data quality assessment: Quality assessment is an iterative process which should involve continuous and onset evaluations from the start to the end of the process and how various stages affect the overall quality assessments. To the best on our knowledge, current data quality assessment methods only look at quality issues that are inherent to the data itself and ignore the effects other stages of the data life cycle have on the overall quality assessment.
- Incorporate meta-data in the assessment of IoT data quality: Most systems today are built with the ability of tagging data with meta-data. Unfortunately none of the current IoT data quality assessment techniques have considered the impart of such on the overall quality assessment.

VI. Conclusion and Future Work

Today data is driving manufacturing, healthy, agriculture and other business decisions. It is significant to assess the quality of data that goes into these processes as erroneous data could lead to catastrophic outcomes. In this paper we have proposed a framework that can be used to estimate the quality of data.

This is based on the big data model and the three trust formation stages highlighted above. This among other thing would enable one to estimate data quality in cases where there is no gold standard to compare to. The other advantage is that one would be able to represent data quality in a general manner throughout the data life cycle.

This paper listed a number of challenges. Those that are related to the structure and those that are related to methods. We explored how the structural challenges can be addressed by proposing a trust assessment framework where all the data quality factors present in the big data model can be represented and how this framework can be implemented in a shared IoT deployment. This left us with important questions for example; how do we effectively determine the weights for each of the parameters at each stage, how to effectively score trust across the phases of the big data model. These and other challenges highlighted above forms part of our future work.

Acknowledgment

This research is funded under the SFI Strategic Partnership Programme (16/SPP/3296) and is co-funded by Origin Enterprises plc

References

[1] U. Sivarajah, M. M. Kamal, Z. Irani, V. Weerakkody, Critical analysis of Big Data challenges and analytical methods, Journal of Business Research.

[2] P. Z. Yeh, C. A. Puri, An efficient and robust approach for discovering data quality rules, in: Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 2010.

[3] F. Chiang, R. J. Miller, Discovering data quality rules, Proceedings of the VLDB Endowment.

[4] W. Fan, Data quality: Theory and practice, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2012.

[5] I. Taleb, M. A. Serhani, R. Dssouli, Big Data Quality: A Survey, in: 2018 IEEE International Congress on Big Data (BigData Congress), IEEE, 2018, pp. 166–173.

[6] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. Van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, P. Buono, Research directions in data wrangling: Visualizations and transformations for usable and credible data (2011).

[7] A. Karkouch, H. Mousannif, H. Al Moatassime, T. Noel, Data quality in internet of things: A state-of-the-art survey (2016).

[8] M. Chen, M. Song, J. Han, E. Haihong, Survey on data quality, in: Proceedings of the 2012 World Congress on Information and Communication Technologies, WICT 2012, 2012.

[9] F. Sidi, P. H. Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, A. Mustapha, Data quality: A survey of data quality dimensions, in: Proceedings - 2012 International Conference on Information Retrieval and Knowledge Management, CAMP'12, 2012.

[10] M. Heravizadeh, J. Mendling, M. Rosemann, Dimensions of business processes quality (QoBP), in: Lecture Notes in Business Information Processing, 2009.

[11] Y. W. Lee, D. M. Strong, B. K. Kahn, R. Y. Wang, AIMQ: A methodology for information quality assessment, Information and Management.

[12] H. Baqa, N. B. Truong, N. Crespi, G. M. Lee, F. Le Gall, Quality of Information as an indicator of Trust in the Internet of Things, in: 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), IEEE, 2018, pp. 204–211.

[13] S. Juddoo, Overview of data quality challenges in the context of Big Data, in: 2015 International Conference on Computing, Communication and Security, ICCCS 2015, 2016.

[14] D. Helbing, A mathematical model for the behavior of individuals in a social field, The Journal of Mathematical Sociology-doi:10.1080/0022250X.1994.9990143.

[15] M. Maheswaran, C. T. Hon, A. Ghunaim, Towards a gravity-based trust model for social networking systems, in: Proceedings - International Conference on Distributed Computing Systems, 2007. doi:10.1109/ICDCSW.2007.82.

[16] Sarbjeet Singh, S. Bawa, Privacy, trust and policy based authorization framework for services in distributed environments, INTERNATIONAL JOURNAL OF COMPUTER SCIENCE 2 (2).

[17] F.-K. Tsai, C.-C. Chen, T.-F. Chen, T.-J. Lin, Sensor Abnormal Detection and Recovery Using Machine Learning for IoT Sensing Systems, in: 2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA), IEEE, 2019, pp. 501–505.

[18] E. Vilenski, P. Bak, J. D. Rosenblatt, Multivariate anomaly detection for ensuring data quality of dendrometer sensor networks, Computers and Electronics in Agriculture.

[19] N. Javed, T. Wolf, Automated sensor verification using outlier detection in the Internet of things, in: Proceedings - 32nd IEEE International Conference on Distributed Computing Systems Workshops, ICDCSW 2012, 2012.

[20] I. M. Faniel, T. E. Jacobsen, Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data, Computer Supported Cooperative Work.

[21] A. Yoon, Data reusers' trust development, Journal of the Association for Information Science and Technology.

[22] Y. Kim, A. Yoon, Scientists' data reuse behaviors: A multilevel analysis, Journal of the Association for Information Science and Technology.

[23] D. Delaney, G. OHare, A Framework to Implement IoT Network Performance Modelling Techniques for Network Solution Selection, Sensors 16 (12) (2016) 2038. doi:10.3390/s16122038.

[24] I. Taleb, H. T. Kassabi, M. A. Serhani, R. Dssouli, C. Bouhaddioui, Big Data Quality: A Quality Dimensions Evaluation, in: Proceedings - 13th IEEE International Conference on Ubiquitous Intelligence and Computing, 13th IEEE International Conference on Advanced and Trusted Computing, 16th IEEE International Conference on Scalable Computing and Communications, IEEE Internationa, 2017.

[25] C. Keßler, R. T. A. De Groot, Trust as a proxy measure for the quality of volunteered geographic information in the case of openstreetmap, in: Lecture Notes in Geoinformation and Cartography, 2013.

[26] P. Manuel, A trust model of cloud computing based on Quality of Service, Annals of Operations Research.

[27] L. Zhao, T. Hua, C. T. Lu, I. R. Chen, A topic-focused trust model for Twitter, Computer Communicationsdoi:10.1016/j.comcom.2015.08.001.

[28] W. Jiang, J. Wu, F. Li, G. Wang, H. Zheng, Trust evaluation in online social networks using generalized network flow, IEEE Transactions on Computersdoi:10.1109/TC.2015.2435785.

[29] P. Verma, A. Mathuria, S. Dasgupta, Item-Based Privacy-Preserving Recommender System with Offline Users and Reduced Trust Requirements, 2019, pp. 218–238. doi:10.1007/978-3-030-36945-3_12.

[30] M. Sievers, A. M. Madni, P. Pouya, R. Minnichelli, Trust and Reputation in Multi-Agent Resilient Systems*, in: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), IEEE, 2019, pp. 741–747.