# Trust and Big Data: A Roadmap for Research

Johannes Sänger, Christian Richthammer, Sabri Hassan, Günther Pernul
Department of Information Systems
University of Regensburg
Regensburg, Germany
Email: firsname.lastname@wiwi.uni-regensburg.de

***Abstract*—We are currently living in the age of Big Data coming along with the challenge to grasp the golden opportunities at hand. This mixed blessing also dominates the relation between Big Data and trust. On the one side, large amounts of trust-related data can be utilized to establish innovative data-driven approaches for reputation-based trust management. On the other side, this is intrinsically tied to the trust we can put in the origins and quality of the underlying data. In this paper, we address both sides of trust and Big Data by structuring the problem domain and presenting current research directions and interdependencies. Based on this, we define focal issues which serve as future research directions for the track to our vision of Next Generation Online Trust within the FORSEC project.**

***Keywords*—*Big Data, trust, reputation***

## I. Introduction

We are currently situated in a time where cheap storage, communication on the Internet, Online Social Networks (OSNs) or new sensor technologies have led to a data explosion. The term having recently emerged to catch this phenomenon is Big Data. Related topics addressed in research include techniques and models, algorithms and systems, machine learning, hardware infrastructure, Big Data analytics or even security and privacy, just to name a few. Application areas encompass various fields such as healthcare, medicine, finance, business, law, education, transportation or telecommunication. This list can easily be continued.

It is all the more astonishing that the notion of trust has not yet been payed much or nearly no attention. The research in trust, particularly in trust and reputation systems, has already become relatively mature [1]. However, the huge amount and diversity of data and data sources provides lots of new opportunities as well as it poses many challenges for online trust. In this paper, we provide a list of issues regarding trust and Big Data that need to be answered.

We distinguish two branches for the research in trust and Big Data, namely *Big Data for trust* and *trust in Big Data*. The former addresses questions of how to use Big Data for trust assessment whereas the latter discusses possibilities and challenges to create trust in Big Data. Based on this general segmentation, the rest of this paper is organized as follows: In Section II we give a short overview of the problem context, the relevance and motivation of this work. Thereby, the notions of Big Data and trust are shortly expounded. We discuss the question of how Big Data can be used for trust assessment and computation in Section III. We thereto provide a list of research questions that particularly address the application of Big Data in trust and reputation systems. Subsequently, Section IV elaborates on the challenges to create trust in Big Data. Here, we name several issues that need to be clarified in future research. Finally, we sum up the most important questions identified in our work and point out our future directions in Section V.

## II. Background

While the era of Big Data began quite recently, trust is a topic that has been discussed in research for decades. To impart a common understanding, we provide a short description of the notions of *Big Data* and *trust*.

### A. Big Data

Whereas the term Big Data has been around for quite a while, in the last two years Big Data seems to be everywhere. Besides the dimension of data *volume* which might be the most self-evident characteristic, Big Data is usually described utilizing the further dimensions of *variety* and *velocity* [2]. These 3 dimensions make up the so-called *3V model* which has already been introduced in the year 2001 [3], albeit with a slightly different meaning. Taming Big Data, it is not only necessary to master the mass quantities of data but also to tackle the variety and multitude of heterogeneous data types, formats (structured, semi-structured and unstructured data) and data sources. Further challenges include the increasing velocity of data creation, processing and analysis [2]. Therefore, Big Data can be concisely defined as *"high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making"* [4]. Extensions of the 3V model encompass further aspects of Big Data, e.g. the need to turn the processed and stored data into *value* [5] at the end of the day or the challenges that come along with the inherent uncertainty of Big Data directly affecting its *veracity* [2]. Thereby, veracity connotes to the uncertainty that comes a long with imprecise data types and varying degrees of data quality affecting both the reliability and predictability of data [2]. The issues of trust in conclusions drawn on data with uncertain levels of data quality are discussed in more detail within Section IV.

As mentioned before, Big Data is utilized in manifold contexts and practical use cases ranging from traffic flow optimization to the deep personalization of services and advertisements on the Internet. With regard to science, Big Data research is also somewhat eclectic. It comprises various disciplines using Big Data for boosting their research efforts. Moreover, it includes research that is straightly directed to the advancement of the underlying core technologies of Big Data processing, e.g. optimizing the performance of MapReduce. Regardless of the research discipline, privacy and security of

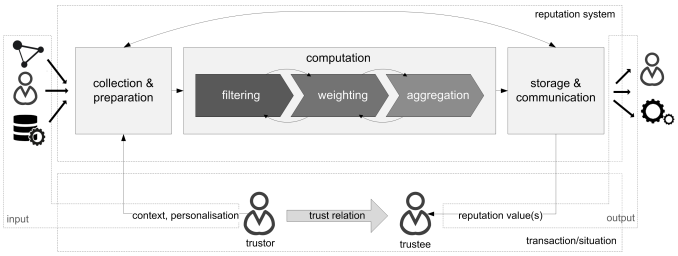


Fig. 1. Generic process of a reputation system, inspired by [9]

information are two areas being intensively studied. Among others, trust is one important topic at this juncture.

### B. The notion of trust

Although trust has been intensively studied in multiple fields, it still lacks a uniform and generally accepted definition. Reasons are the multifaceted terms trust is associated with as well as the multidimensionality of trust as an abstract concept. The definition often cited in literature regarding trust and reputation online was proposed by Gambetta in 1988 [6] and is referred to as reliability trust: "*Trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action (or independently of his capacity ever to be able to monitor it) and in a context in which it affects his own action.*" Multiple authors furthermore comprise security and risk which can lead to more complex definitions.

In the recent decade, various trust models have been developed to establish trust. Thereby, two common ways can be distinguished, namely policy-based and reputation-based trust establishment [7]. Policy-based trust is based on the exchange of hard evidence (e.g. credentials). Reputation-based trust, in contrast, is derived from the history of interactions. Hence, it can be seen as an estimation of trustworthiness. In Section III we focus on Big Data for reputation-based trust. Reputation is defined as follows: "*Reputation is what is generally said or believed about a person's or thing's character or standing*" [8]. It is based on referrals, ratings or reviews from members of a community and can, therefore, be considered as a collective measure of trustworthiness [8].

## III. Big Data for Trust

The huge amount, variety and velocity of data being created on a daily basis provides lots of opportunities for trust and reputation systems. In this section, we have a close look at the generic process of a reputation system. We thereby discuss the consequences for the single process steps when using Big Data sources as input. We furthermore derive several questions that need to be answered in this context. These will be focal points for our future research.

The generic process of reputation systems can be divided into three steps as depicted in Figure 1: (1) *collection & preparation*, (2) *computation* and (3) *storage & communication*.

### A. Collection & Preparation

In the collection and preparation phase, information about the past behavior of a trustee are gathered and prepared for subsequent computing. The vast number of web applications such as eCommerce platforms, Online Social Networks or content communities has led to huge amounts of reputation data being created. To receive a comprehensive picture, multiple sources should be comprised. We thereto identified the following questions:

1) How can data of different sources be collected and integrated?
   A first step toward an integration of several reputation systems to a cross-community reputation system has already been made by [10] and [11]. The sharing of reputation information will become even more important, when reputation profiles have been growing over a period of years. Particularly users who have invested a lot of time in shaping their good reputation may be greatly interested in transferring their reputation profiles to different platforms. While [12] only discussed the need for interoperable reputation systems, we want to go one step further and integrate both explicit and implicit information. Explicit information is particularly created for the rating of a trustee such as seller ratings on ebay. Implicit trust information, in contrast, are information that can be derived from data not explicitly created for trust evaluation like the structure of an OSN. Therefore, potential data sources should not only be interoperable reputation systems, but also every application containing any information about an actor. This leads to the second question:
2) How can implicit reputation information be extracted from the data provided in different settings?
   Variety is one important property of Big Data. Most of the data created today is semi- or unstructured. Since the share of data being created in a structured manner will decrease in the future years, methods to extract reputation information from unstructured data (mostly implicit) will become interesting. Natural language processing and machine learning already provide some incipient approaches.

Once the preparation is completed, the reputation data serves as input for the computation phase.

### B. Computation

The computation phase is the central part of every reputation system. It takes the reputation information collected as input and generates a reputation value as output. The logical process of this phase can be divided into the three process-steps *filtering, weighting* and *aggregation*. The purpose of these steps is obvious: The first question to be answered is *which* information is useful for further processing (filtering). The second process-step concerns the question of *how relevant* the information is for the specific situation (weighting). Particularly the filtering step will become crucial for large amounts of data, since the computation can be a very complex and CPU-intensive process. Finally, the reputation values are aggregated to calculate one or several reputation scores. Questions that need to be answered in this context are as follows:

3) How can Big Data be reduced to those information useful for trust assessment?

Volume is one significant characteristic of Big Data. In order to make these quantities of data manageable and prepare them for aggregation, filtering is necessary. However, it has not become clear yet, how to make out the pieces of data useful for further computation. Possible attributes could be the significance, data quality or trustworthiness of the sources (see Section IV.

4) How can we differentiate between "fast moving" data and information important in the long term?
As mentioned above, Big Data is characterized by high velocity. This means that the currency of reputation data may become even more important to keep the information up-to-date. However, different types of information may have different life cycles. Quality of service, for instance, may change overnight. Positions in an OSN, in contrast, are more robust. Therefore, we need currency metrics that allow a differentiated view and adoption to various settings.
5) How can the computation process be designed in an incremental manner?
As new reputation data is gradually created, an incremental computation process is needed that allows to add new information to already calculated reputation values. Since the computation process can be quite expensive, incremental aggregation algorithms should avoid a multiple calculation.

### C. Storage & Communication

After reputation scores are calculated, they are either stored locally, in a public storage or both. Common reputation systems do not only provide the reputation scores but also offer extra information to support the end users in understanding the meaning of a score-value. They should furthermore reveal the computation process to accomplish transparency. Here, we identified the following issues:

6) How can computation outcomes be stored so that they are reusable?
As mentioned above, computation can be a very complex and expensive process. Thus, the outcomes of calculations that were made once should be stored in a way that they are reusable comprising enough information for further computation.
7) How can the computation process be transparently communicated?
Since data of various sources and data types are evaluated using partly complex techniques, it is difficult to communicate the outcomes in a transparent way. Here, visualizations could be a proper way of representing the results. Visual analytics methods could furthermore encourage the end user to participate in the evaluation process.

## IV. Trust in Big Data

### A. Trust in data quality

The developments in the area of Big Data enable the utilization of more data sources than ever before. However, this implies the need for verifying their trustworthiness and especially for evaluating the quality of the input data. With respect to data quality, we identified the following issues:

8) How can manipulations of input data be detected?
Adversaries have several options to compromise the quality of input data [13], especially in connection with mobile sensor devices. Several researchers realised the need for technology that attests the authenticity of sensor readings and proposed the utilization of Trusted Platform Modules to address this. A modern application domain of sensed data is particapatory sensing, which is likely to receive a lot of attention in research.
9) How can the quality of input data be improved?
Apart from adversarial data manipulations, input data can be innately sparse, uncertain and incomplete. Wu et al. [14] name several data mining approaches with need for research to address these issues, including: Feature selection and unsupervised learning methods for sparsity, error-aware data mining for uncertainty, and data imputation methods for incompleteness. Another research direction in connection with the quality of input data is the generation of authentic synthetic data, which are needed to benchmark different Big Data solutions, for instance [15].
10) Do Big Data lead to better results?
Using input data of bad quality may lead to untrustworthy conclusions. If informed decisions are not possible on these grounds, Big Data approaches are at least able to slightly overcome the lack of data quality with data quantity and enable them at all. Tien [16] argues that this approach may even be more realistic than the optimality focus of traditional methods. However, we suggest a critical investigation of this fundamental question and warn that improved research findings should not be automatically ascribed to Big Data.
11) How can trust in result data be increased?
An interesting approach is the concept of result provenance, which enables organisations to observe which input data and which intermediary inferences have led to a final outcome. Glavic [17] introduces the term Big Provenance and calls this a field that is still largely unexplored. He provides an up-to-date survey of existing Big Provenance approaches and introduces some future challenges.

### B. Measuring trust in Big Data

12) How can trust in Big Data be measured?
Finding ways to quantify trust would enable the comparability of data sources and help to make the notion of trust more tangible. Recently, different proposals have emerged. Lukoianova and Rubin [18] focus on veracity as a critical quality factor and introduce a Big Data veracity index that combines the three dimensions of subjectivity, deception and implausibility. Belov et al. [19] propose a system for automated trust assessment of online open media that is based on both the assessment of the source and the content. And Albanese [20] tries to quantify the trustworthiness of both the data source and the data

items by developing a solution similar to Google's PageRank method.

### C. Trust in nodes

Even with ensured quality of input data, Big Data processing can only be trusted to the extent to which trust in the underlying systems is established.

13) How can the trustworthiness of nodes be ensured?
In [13], the idea of trust establishment is brought up but not specified any further. Using Mandatory Access Control is suggested as an alternative. Huang et al. [21] consider nodes that are manipulated so that they either do not process their tasks completely or so that they intentionally return wrong results. They design mechanisms for detecting compromised nodes based on watermark injection and random sampling methods.
14) How can the trustworthiness of nodes in the cloud be ensured?
The problem of untrusted nodes is exacerbated when Big Data applications are moved into the cloud. Promising directions for solutions include drawing on accountability [22], combining the benefits of the public and the private cloud, and integrations with Trusted Computing concepts.

### D. Trust in Cloud Service Providers (CSPs)

Apart from the correct functionality of the CSP's hardware, the CSP itself has to be trusted to use the data only in appropriate and agreed ways. Threats include roll-back attacks [13] and the concealment of data loss incidents [23].

15) How can these threats be mitigated?
Untrusted CSPs motivate the research field of integrity verification, which is also referred to as data auditing or third-party auditing if the verification is conducted by a third party. A recent proposal, additionally addressing privacy-preserving techniques in connection with third-party auditing, can be found in [23].
16) How can trust in CSPs be evaluated?
One recent approach, which addresses providers of database services in paricular, is introduced in [24]. Here, a measuring mechanism is proposed to select the best CSP based on user-selected relative and direct factors introduced in the paper.

### E. Trust in information sharing

Trust also plays a major role in the area of intelligence-driven security. This Big Data domain has gained increased importance because of the advent of Advanced Persistent Threats (APTs). To address these new kind of threats, several authors expressed the need for information sharing among researchers and organisations. Two directions of trust need to be considered in connection with information sharing.

17) Can partnering organisations be trusted to provide data in an adequate quality?
This question is related to our remarks regarding trust in data quality and measuring trust in Big Data. However, the context of information sharing lays the focus more on the quality of data that has already been processed by another organisation rather than raw data.
18) Can partnering organisations be trusted to handle the data only in appropriate and agreed ways?
On the one hand, there are data that an organisation may not want to become public, such as data about particular incidents. And on the other hand, there are sensitive or personally identifiable data from customers and users whose sharing may be regulated by law and therefore not trivially manageable. The latter has attracted the attention of researchers for years and brought forth the research field of de-identification. The advent of Big Data further strenghtens the need for provably effective and efficient solutions.

## V. Conclusion

"Big Data" has been a hot topic in the last years and interest in it is supposed to increase even further. The several research areas of Big Data are relevant to academia as well as to industry and the ordinary citizen, which makes the arising issues numerous and diverse. Consequently, a considerable amount of research has already been carried out in the particular fields. To the best of our knowledge, however, the notion of trust in connection with Big Data has received nearly no attention so far. To address this shortcoming, we proposed two research branches emerging from the combination of the terms of interest, namely *Big Data for Trust* and *Trust in Big Data*. In each of these branches, we raised several research questions and provided information on current research efforts that are promising for answering them. These questions will hopefully initiate fruitful discussions among researchers in order to solve some fundamental problems in the area of Big Data and help to turn the hype into real benefits. Although we dedicated separate sections to the two research branches, we believe that they should not be treated strictly on their own. On the one hand, with the computation of trust being reliant on trustworthy data, trust in Big Data is a prerequisite for the pursuant trust assessment. On the other hand, the assessment of trust in Big Data relies on correct trust computation mechanisms.

In our future work, we will keep paying attention to the developments regarding the research questions raised in this paper. At the moment, we are also examining which of them are particularly suitable for generating synergies with other research efforts of ourselves. For example, we plan to harness the novel possibilities provided by Big Data for Next Generation Online Trust in the context of the FORSEC research project.

## Acknowledgment

The research conducted by Christian Richthammer and Johannes Sänger leading to these results was supported by "Bavarian State Ministry of Education, Science and the Arts" as part of the FORSEC research association (http://www.bayforsec.de/).

## REFERENCES


[1] A. Jøsang, "Robustness of Trust and Reputation Systems: Does It Matter?" in *Trust Management VI*, ser. IFIP advances in information and communication technology, T. Dimitrakos, R. Moona, D. Patel, and D. McKnight, Eds. Springer Berlin Heidelberg, 2012, vol. 374, pp. 253–262.

[2] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, "Analytics: The real-world use of big data: How innovative enterprises extract value from uncertain data," New York, 2012.

[3] D. Laney, "3-D Data Management: Controlling Data Volume, Velocity and Variety." Stamford, 2001.

[4] Gartner, "IT Glossary.Big Data." [Online]. Available: http://www.gartner.com/it-glossary/big-data/, Accessed on: 03/13/14.

[5] G. Geethakumari and A. Srivatsava, "Big Data Analysis for Implementation of Enterprise Data Security," *IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS)*, vol. 2, no. 4, pp. 742–746, 2012.

[6] D. Gambetta, "Can We Trust Trust?" in *Trust: Making and Breaking Cooperative Relations*, D. Gambetta, Ed. Oxford: Basil Blackwell, 1988, pp. 213–237.

[7] D. Artz and Y. Gil, "A survey of trust in computer science and the Semantic Web," *Web Semantics*, vol. 5, no. 2, pp. 58–71, 2007.

[8] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision Support Systems*, vol. 43, no. 2, pp. 618–644, 2007.

[9] G. Swamynathan, K. C. Almeroth, and B. Y. Zhao, "The design of a reliable reputation system," *Electronic Commerce Research*, vol. 10, no. 3-4, pp. 239–270, 2010.

[10] F. Pingel and S. Steinbrecher, "Multilateral Secure Cross-Community Reputation Systems for Internet Communities," in *Trust, Privacy and Security in Digital Business*, ser. Lecture notes in computer science, S. Furnell, S. Katsikas, and A. Lioy, Eds. Springer Berlin Heidelberg, 2008, vol. 5185, pp. 69–78.

[11] N. Gal-Oz, T. Grinshpoun, and E. Gudes, "Privacy Issues with Sharing Reputation Across Virtual Communities," in *Proceedings of the 4th International Workshop on Privacy and Anonymity in the Information Society*, ser. PAIS '11. New York and NY and USA: ACM, 2011, pp. 3:1–3:5.

[12] S. Steinbrecher, "The Need for Interoperable Reputation Systems," in *Open Research Problems in Network Security*, ser. Lecture notes in computer science, J. Camenisch, V. Kisimov, and M. Dubovitskaya, Eds. Springer Berlin Heidelberg, 2011, vol. 6555, pp. 159–169.

[13] Cloud Security Alliance, "Expanded top ten big data security and privacy challenges," 2013.

[14] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[15] Z. Ming, C. Luo, W. Gao, R. Han, Q. Yang, L. Wang, and J. Zhan, "Bdgs: A scalable big data generator suite in big data benchmarking," *CoRR (Computing Research Repository)*, 2014.

[16] J. M. Tien, "Big data: Unleashing information," *Journal of Systems Science and Systems Engineering*, vol. 22, no. 2, pp. 127–151, 2013.

[17] B. Glavic, "Big data provenance: Challenges and implications for benchmarking," in *Specifying Big Data Benchmarks*, ser. Lecture Notes in Computer Science, D. Hutchison *et al.*, Eds. Berlin and Heidelberg: Springer Berlin Heidelberg, 2014, vol. 8163, pp. 72–80.

[18] T. Lukoianova and V. L. Rubin, "Veracity roadmap: Is big data objective, truthful and credible?" *Advances In Classification Research Online*, vol. 24, no. 1, pp. 4–15, 2014.

[19] N. Belov, J. Schlachter, C. Buntain, and J. Golbeck, "Computational trust assessment of open media data," in *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2013, pp. 1–6.

[20] M. Albanese, "Measuring trust in big data," in *Algorithms and Architectures for Parallel Processing*, ser. Lecture Notes in Computer Science, D. Hutchison *et al.*, Eds. Cham: Springer International Publishing, 2013, vol. 8286, pp. 241–248.

[21] C. Huang, S. Zhu, and D. Wu, "Towards trusted services: Result verification schemes for mapreduce," in *2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, 2012, pp. 41–48.

[22] Z. Xiao and Y. Xiao, "Achieving accountable mapreduce in cloud computing," *Future Generation Computer Systems*, vol. 30, pp. 1–13, 2014.

[23] C. Wang, S. S. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for secure cloud storage," *IEEE Transactions on Computers*, vol. 62, no. 2, pp. 362–375, 2013.

[24] Priyadarshani, W. P. Eureka, G. N. Wikramanayake, and Ekanayake, E. M. Piyal, "Measuring trust and selecting cloud database services," *ACSIJ (Advances in Computer Science: An International Journal)*, vol. 2, no. 5, pp. 114–120, 2013.