Validity Issues of Digital Trace Data for Platform as a Service: A Network Science Perspective

Mehmet N. Aydin[1](✉), Dzordana Kariniauskaite[1],
and N. Ziya Perdahci[2]

[1] Department of Management Information Systems,
Kadir Has University, Istanbul, Turkey
{mehmet.aydin, dzordana}@khas.edu.tr
[2] Department of Informatics,
Mimar Sinan Fine Arts University, Istanbul, Turkey
nazim.ziya.perdahci@msgsu.edu.tr

Abstract. Data validity becomes a prominent research area in the context of data science driven research in the past years. In this study, we consider an application development on a cloud computing platform as a promising research area to examine digital trace data belonging to records of development activity undertaken. Trace data display such characteristics as found data that is not especially produced for research, event-based, and longitudinal, i.e., occurring over a period of time. Having these characteristics underlies many validity issues. We employ two application development trace data to articulate validity issues along with an iterative 4-phase research cycle. We demonstrate that when working with digital trace data, data validity issues must be addressed; otherwise it can lead to awry results of the research.

Keywords: Validity · Digital trace data · Cloud computing · Metadata actions
Network science

1 Introduction

Data validity becomes a prominent research area in the past years. One of the reasons for that is the growing amount of digital trace data, especially in (social) network analysis [5]. [8] describe digital trace data as records of activity (trace data) undertaken through an online information system (thus, digital). Trace data displays such characteristics: it is found data, which means that it is not especially produced for research, it is event-based and it is longitudinal, occurring over a period of time. Having these characteristics underlies many validity issues in network analysis. Thus, when working with digital trace data, validity issues must be addressed; otherwise it can lead to awry results of the research. One of the promising domains to better understand validity issues in network

The original version of the book was revised: Misspelt second author name has been corrected and the chapter title has been updated. The erratum to the book is available at https://doi.org/10.1007/978-3-319-77703-0_116

Á. Rocha et al. (Eds.): WorldCIST'18 2018, AISC 745, pp. 654–664, 2018.
https://doi.org/10.1007/978-3-319-77703-0_65

analysis is to examine digital trace data in the context of application development on cloud computing platforms like Salesforce Platform, Mendix, and Imona. Noticeably, digital trace data is the very existence of these platforms and they are considered as game changers in the software development world [15]. As shall be discussed in detail, they provide novice developers with a novel way of developing software applications with the premise of fast and easy development, and high productivity.

In this study, we employ two software applications development trace data to articulate validity issues along with an iterative 4-phase research cycle. We demonstrate that when working with digital trace data in network science, data validity issues must be addressed; otherwise it can lead to misleading results of the research.

[8] propose a model (Fig. 1) for addressing the validity issues while carrying out a study based on digital trace data. The model basically introduces ten issues, which should be solved in six main phases (Information system, Event data, Nodes and links, Network, Measure and Construct) of network analysis. Although being comprehensible and easy to follow, the model proposed by [6] does not provide an example with real-world data. Overall, such type of studies is limited and needs to be refined further, especially from an information systems perspective. Thus, the purpose of this paper is to provide a fine-grained guideline for addressing validity issues in digital trace data driven network science research [2]. To this end, we take into account a real-world case for addressing validity issues while working with digital trace data. The data used in this research is metadata actions of two software applications developed on an application platform as a service (aPaaS). The study presents a 4-phase research cycle for addressing the validity issues and insights in dealing with the issues.

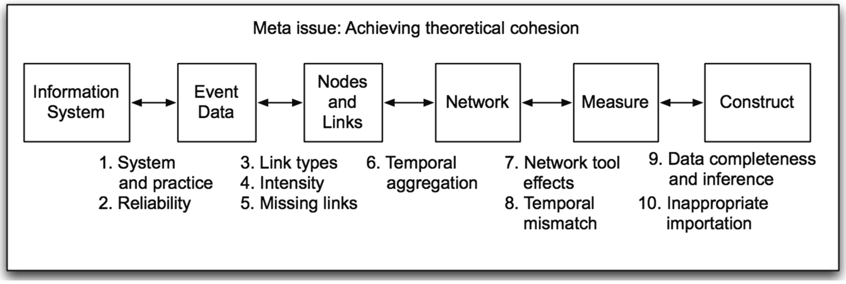

Fig. 1. Howison et al.'s proposed approach to address validity issues of digital trace data [8]

This paper begins by shortly providing a relevant research background, followed by a description of the method used in this study and a real-world example of how to apply this method. In the next sections findings, discussion and conclusions are provided.

2 Background and Method

For the purpose of this study we take into account three relevant research points: articulations of validity issues in social network analysis for digital trace data, application platform as a service as digital trace data, and network science as a promising approach to better understand validity issues and rethinking software development as a complex system.

Regarding the first research point that is essentially validity issue, the model (Fig. 1), proposed by [8] is composed of six elements connected by five links raising ten issues. First of all, when working with a digital trace data it should be looked into the information system producing that data. The misuse of the system can cause the misinterpretations of the collected data. Another issue that should be considered in this phase is the reliability of the data generated. In the next step, the complication of converting digital trace data into nodes and links should be solved. In order to do that, the researcher should make one of the most crucial decisions, which is determining the type and intensity of the links, as well as deciding on the missing links.

Creating a network where the order of the events matter can raise a problem of temporal aggregation. In another step, while using a network to obtain some measures, a researcher should address network tool effects and temporal mismatch. There is a large selection of software tools available for social network analysis (SNA) [12, 14] thus choosing proper software for an analysis is an important step, since these tools can help researchers with avoiding errors as well as at the same time can threaten the validity of their use. The temporal mismatch issue can be addressed by deciding the period of time over which measures derived from that network will be measured. The last issues a researcher should consider emerge when aligning a measure and a construct are data completeness and inference and inappropriate importation.

Regarding the second research point that is application platform as a service as digital trace data, one may need to look into the fundamentals of cloud computing that challenge conventional approaches and tools for software development, distribution, use, and pricing [1]. A particular research and practitioner effort in cloud computing, which is of interest to this research is Platform as a Service (PaaS). PaaS focus has widened its scope in recent years, touching upon not only the runtime environment of a SaaS application, but also the tools and methods of developing apps in the cloud. PaaS in this context can be defined as a “complete application platform as a service” that offers independent software vendors (ISVs) as well as individuals the means to produce “multitenant SaaS solutions or various application integration solutions” in a fast and efficient way [7]. Thus, some of the PaaS offerings comprise a pre-determined development environment with all tools for development, testing, and deployment cycles of an application [4].

Regarding the third research point that is network science, one can argue that all metadata actions on aPaaS can be considered as entangled interconnectedness actions or entities that can be represented as a complex system. This calls for a network science approach as a novel way of examining digital trace data as a complex system [2]. Interconnectedness naturally leads to formation of things and their relationships as a network. In the last two decades, there has been a significant interest in better understanding of real-world entities and their relations as complex systems. So-called network science has been promoted to investigate the very nature of these complex systems in various contexts such as social, technological, health sciences, and political. Complex systems in contexts are revisited and represented as networks.

Even though the problems and associated contexts are different so the networks are labeled differently, a number of common distinguishing network characteristics have been at the center of attention in various research domains including social and management sciences, natural sciences and engineering, and life sciences.

Thus, scholars have attempted to address challenges with finding appropriate theoretical accounts, statistical network science and scientific approach to better understanding of real-world complex systems as networks [13]. In the following we shall discuss how metadata actions generating digital trace data are considered as a network data and validity issues are addressed in two real-world cases.

The data examined in this research is the metadata actions of mobile management of *SheepHerd* application and *GradeBoo*k application developed on Imona.com (www.imona.com). Imona.com is an application development platform (Application Platform as Service: aPaaS) where developers can not only create new applications, but it also offers the possibility of extending the functionality of any application already placed in its marketplace [1]. Imona.com is a particular type of aPaaS, called metadata aPaaS [1]. At a high abstraction level, metadata aPaaS provides visual tools to customize data models, application logic, workflow, and user interface. The underlying metadata model for this cloud platform is essential to this research as it provides us a reference conceptual model (see Fig. 2) to reflect on network models of digital trace data to be discussed later on.

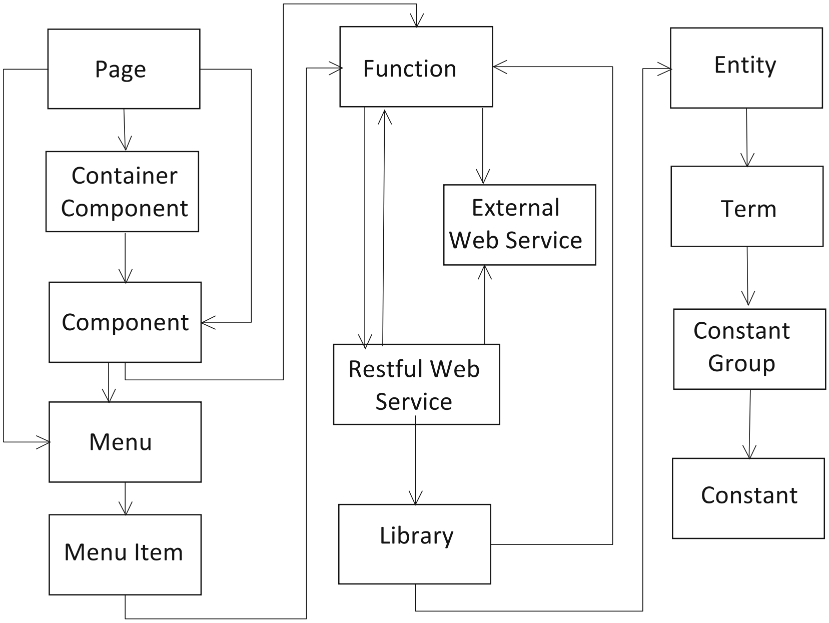

Fig. 2. Conceptual model for metadata aPaaS of ImonaCloud.com

The dataset is composed of metadata actions that were created while developing an application over a period of time. For each metadata created the timestamp is available. Data preparation is made with Excel. Description of the network data and visual analysis of network diagram is produced with Gephi [3], which is an open-source and free visualization and exploration platform for SNA.

As [6] writes: “In practice, the process of achieving alignment between a theoretical context and the chain of reasoning underlying valid measurement is an iterative one, most likely involving multiple adjustments and decisions and revisiting these to achieve a cohesive logic”. Thus, applying the model (Fig. 1) is suggested to be an iterative process. Indeed, the cases presented in this study contain not only iterative phases, but also phases that are iterative in themselves.

3 Findings and Discussion

3.1 Phase 1: Exploratory Network Analysis Without Metadata-Based Context Understanding Leads to Nowhere

One network scientist performed the first phase. The only information he knew was what SQL queries were used to fetch the digital trace data pertaining to the metadata actions and a few details about the aPaaS application platform itself. Several years ago, when the system was in its private beta version, the network scientist studied it by self-learning.

Believing that the network scientist has enough knowledge about the platform an exploratory data analysis was conducted. The data used at this stage was raw data and it was considered as reliable. The researcher decided by himself what to consider a node and a link to be in a network where metadata actions are best described as well as link types and intensity, which were chosen to be undirected and unweighted. Particularly for this case, missing link concept can have two meanings: first, no metadata action has been taken, and second, metadata action has been taken, but the corresponding software artifact created was removed afterwards by another metadata action. In the next step, several network models and visualization of the networks were created, however, the discussions between one software engineering scientist and the network scientist lead to the conclusion that the network model could not capture truly the nature of the metadata actions. The primary reason for the initial attempt being a failure was because the researcher did not have the data dictionary, and for that reason the digital trace data could not have been understood which led to awry results. The results were also communicated with the practitioner who also concurred that they failed to make sense.

In summary, in this phase, based on past experience the researcher completely ignores first two data validity issues and performs exploratory data analysis (Force Atlas layout algorithm for visualization, degree distribution, component analysis) addressing 3, 4, 5 and 6 issues. The results of the first attempt lead to nowhere, however, it shows which data validity issues must be addressed.

3.2 Phase 2: Shared Understanding of MetaData Context and Network Model by Practitioner and Network Scientist

After the experiences gained in the first phase a practitioner shared a metadata model of the system (Fig. 2), which is referred to as a conceptual model. In the next step, the raw data was reconsidered for creating a new network model and conducting network analysis (Degree Distribution, Component Analysis, network model visualization for node types) along with the conceptual model. The network scientist thought that there had to be a one-to-one correspondence between the conceptual model and the (found) raw data. It turned out that this was not the case as there was only a partial match between the two. For example, the raw data and the conceptual model both include an entity element. On the other hand, the scientist was not able to relate the rest of the elements of the Model part to the conceptual model (Fig. 2). After asking the practitioner about this lack of one-to-one correspondence between the conceptual model and the raw data issue, it was found out that there was a notational difference between the elements of the conceptual model and the digital trace data and indeed there was a one-to-one correspondence between all elements of the Model part and the raw data. To the surprise of the network scientist, the metadata actions of the rest of the conceptual model, however, did not have any recorded digital trace in the raw data set. Moreover, contrary to the comments of the practitioner it was noticed that some metadata actions did not have unique identifier numbers assigned. When inquired about these system issues, the platform developer reported that assigning the same "unique id" to concurring metadata actions that are atomically created was a design decision, and recording digital traces pertaining to Controller and View metadata actions never took place. Notice that researchers in software engineering have extensively examined Model-View-Controller approach [9].

In summary, the network scientist addresses system and practice issues and reliability issues properly. Having resolved these two issues makes it possible to address issue 3, 4, 5, 6, 7, and 8.

3.3 Phase 3: Network Analytics Addressing Completeness and Inference Validity Issues

At this stage, the researcher created a kind of new data dictionary, making a conclusion that the real data only contains a complete Model part (the rightmost column) (Fig. 2) and lacks View and Controller parts. In the next step, the following graph model is incorporated into the complete understanding of the digital trace to realize a network of metadata actions: Nodes will represent each of the elements of the conceptual model and the metadata actions will provide the edges. The graph model is undirected the edges have no weights attached. It is a simple graph. Figure 3 depicts the graph model, having four metadata actions. To be more specific, the graph tells us that: According to the digital trace the metadata action ADD_CONSTANT_TO_CONSTANT_GROUP is executed four times consecutively to add the two ListItems (constant_name) with developer named pgf2 and gnrh names to the corresponding Lists (cons_group_name) with developer named animal_extra_homon_type and extra_hormon_type.

The visualization of the network was created according to this graph model (see Fig. 4). The size of the nodes is scaled according to their degree of connectivity; node colors distinguish network fragments. Hubs, nodes having considerably larger connections than the rest of the network, are depicted by larger symbols. The network is mostly fragmented, having 21 fragments, and the mean degree of connectivity is 1.868, and average path length of 2.95, just shy of three. There is no clustering, that is, triadic connectivity is totally absent in the network.

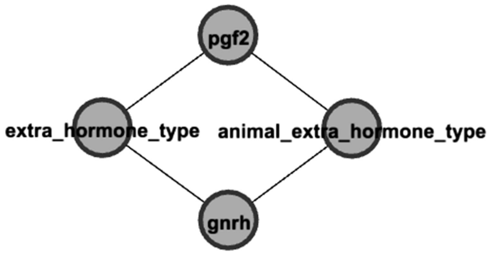

Fig. 3. The network model illustrated by a 4-node component.

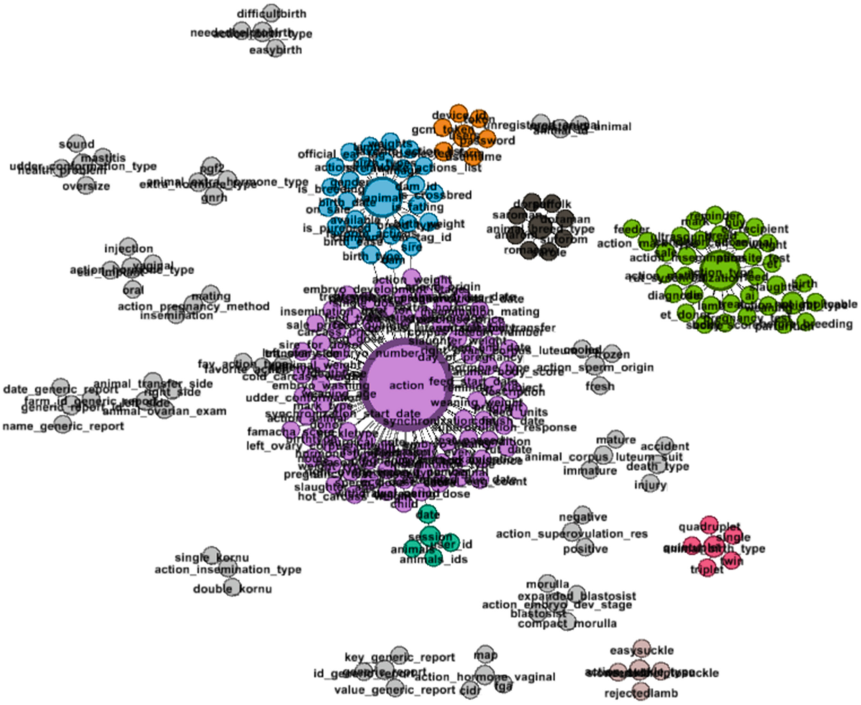

Fig. 4. Network of the temporal aggregation of the metadata actions in the end of the application development project

The appearance of square sub graphs (see Fig. 3) is especially noteworthy as it is commonly seen in real-world complex systems [6] This might be a signature of the network model and deserves to be investigated further.

In summary, finally, the network scientist, the software engineering scientist and the practitioner are in agreement that the network model captures the true nature of the meta data actions and it is time to address the two remaining issues, namely data completeness and inappropriate importations (Table 1).

Table 1. Validity issues in the case examined

Phase	Validity issues	Case findings and network analysis
1	3. Linked types 4. Intensity 5. Missing links 6. Temporal aggregation	Based on past experience the researcher completely ignores first two data validity issues and performs exploratory data analysis addressing 3, 4, 5 and 6 issues. The results of the first attempt lead to nowhere, however it shows which data validity issues must be addressed. The network analysis performed includes Force Atlas layout algorithm for visualization, degree distribution, component analysis
2	1. System and practice 2. Reliability 3. Linked types 4. Intensity 5. Missing links 6. Temporal aggregation 7. Network tool effects 8. Temporal mismatch	The network scientist addresses system and practice issues and reliability issues properly. Having resolved these two issues makes it possible to address issue 3, 4, 5, 6, 7, and 8. The network analysis performed includes network model visualization for node types, giant connected component, hub-spoke phenomenon
3	1. System and practice 2. Reliability 3. Linked types 4. Intensity 5. Missing links 6. Temporal aggregation 7. Network tool effects 8. Temporal mismatch 9. Data completeness and inference	The network scientist, the software engineering scientist and the practitioner are in agreement that the network model captures the true nature of the metadata actions and it is time to address the two remaining issues, namely data completeness and inappropriate importations. The network analysis performed includes structural network analysis, clustering coefficients
4	1. System and Practice 2. Reliability 3. Linked types 4. Intensity 5. Missing links 6. Temporal aggregation 7. Network tool effects 8. Temporal mismatch 9. Data completeness and inference	While developing a new application in an iterative manner, the network scientist kept track of every new record of digital trace data in order to see whether all steps are being captured. Data completeness is achieved by fixing the last exception. At last, to some extent importations of essential network constructs are articulated. The network analysis performed includes all network metrics mentioned at three phases above

3.4 Phase 4: Cross Check for All Validity Issues in a New Metadata Action Context

The aim of the fourth phase was to address the two remaining issues, namely data completeness and inappropriate importations. In order to properly address remaining issues, the data validity steps had been iterated once again, since the data validation is not a linear process, as double-sided arrow symbols indicate in the model (Fig. 1). To address all validity issues better, this time it was decided for a network scientist to develop an application on the platform. The two reasons for doing this was to use as many software development features the platform provides as possible and to see whether they all are being recorded. The system and practice issues in digital trace data validation process are one of the difficulties a scientist should pass and the way to do that is to use the system alone, as well as interview the experts of the system. Since the network scientist has studied the system when it was in private beta state (now the platform has changed), the application was developed with the help of a practitioner. The application *GradeBook*, which, basically, is the tool for the teachers allowing to create new courses, add new students to those courses and grades for different assignments, as well as to see students' average grades for a particular course (to develop this application was chosen without any particular reason).

The application was developed two times: first time, it was developed with the help of the practitioner and it took around 5–6 h. The second time the network scientist developed the exact application without any additional help from a platform expert and it took less than one hour. While developing the application the second time, the network scientist kept track of every new record of digital trace data in order to see whether all steps are being captured. This process showed that all metadata actions have been recorded except one, which is setting the term *coursename* as a unique key to the entity named *course*. This deficit was reported to the practitioner. In the next step, the visualization of the toy network was created (Fig. 5).

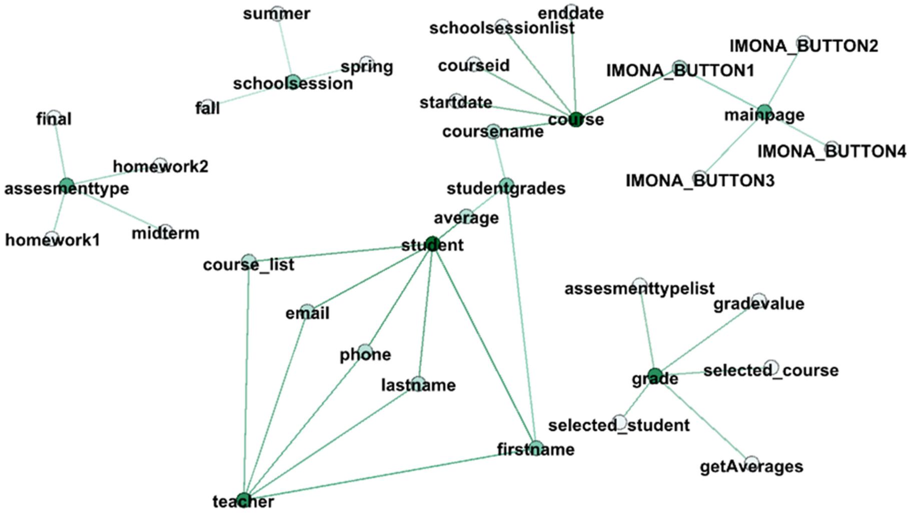

Fig. 5. Network of the GradeBook application as the outcome and an indication of experiencing with all validity issues

4 Conclusions

We contend that although [8] model (Fig. 1) is a major step forward in coping with the validity issues, it is certainly not the end of story. The cases at hand clearly demonstrates that we are still far from a fully operational framework for addressing all validity issues. There are several reasons for this.

First of all, not all digital trace data have similar origins and a rigorous set of measures are needed to assess and improve the proposed validation approach in network science research. This is an open issue as the network science is yet to be fully employed in new research domains including digital trace driven data in cloud computing.

In particular, to the best of our knowledge, cloud aPaaS platforms are still so new that digital trace of actions has never been studied before, even though the importance of metadata in social network analysis is acknowledged [10]. Research teams face severe hurdles even in the first two steps of Howison et al.'s model, thanks to the diverse group of people who contribute to these platforms namely system and practice issues. Should we contact the practitioner, the platform developer, or both to initiate the research in the first place? Or, should we just take the raw data, do the analysis first on our own, explore the data and return to the field. The case shows that while following the steps in the model it is not just enough to answer yes/no to the questions raised by the issues existing in the model. We hope that this case will have a positive effect on research groups who will face similar hurdles.

From the network science perspective, the current scene is more or less in accordance with our expectations. The network displays some of the characteristics common to real-world complex systems such as the existence of rather short distances (average path length of nearly three) between nodes (small-world phenomenon) and hub-and-spoke connectivity structure, and existence of 4-node cycles. The absence of triadic cycles, however, is in contrast with the real-world networks where high clustering coefficient is a signature. The work is currently underway to understand this seemingly problematic outcome.

As for the practitioner, the observation of 4-phase research cycle has been a welcome surprise. The practitioner's comment on the existence of these was that it was a sign of software reusability on the developer's site that could not have been surfaced without the incorporation of theory of networks. This brings up potential and challenging research agenda in the software engineering domain [11]. Even if the study lacks the complete list of metadata actions, simply because they were not recorded, aPaaS platform owners are fully aware of the value of this research, agreed that they will take all of the necessary measures to include the rest of the metadata actions into the digital trace and support the research further. Network science approach to developer behavior analysis utilizing digital trace data may eventually help us to radically change the way of thinking and working for software development.

Acknowledgments. We are indebted to CEO of Imonacloud for providing us with an opportunity to employ the digital trace data on Imonacloud platform.

References

1. Aydin, M.N., Perdahci, N.Z., Odevci, B.: Cloud-based development environments: PaaS. In: Encyclopedia of Cloud Computing, p. 62 (2016)
2. Barabási, A.L.: Network Science. Cambridge University Press, Cambridge (2016)
3. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In: The Proceedings of the Third International ICWSM Conference ICWSM, San Jose, California, pp. 361–362. AAAI Press, Menlo Park (2009)
4. Beimborn, D., Miletzki, T., Wenzel, S.: Platform as a service (PaaS). Bus. Inf. Syst. Eng. **3**(6), 381–384 (2011)
5. Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G.: Network analysis in the social sciences. Science **323**(5916), 892–895 (2009)
6. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: structure and dynamics. Phys. Rep. **424**(4), 175–308 (2006)
7. Fylaktopoulos, G., Skolarikis, M., Papadopoulos, I., Goumas, G., Sotiropoulos, A., Maglogiannis, I.: A distributed modular platform for the development of cloud based applications. Future Gener. Comput. Syst. **78**, 127–141 (2017)
8. Howison, J., Wiggins, A., Crowston, K.: Validity issues in the use of social network analysis with digital trace data. J. Assoc. Inf. Syst. **12**(12), 767 (2011)
9. Karsai, G., Sztipanovits, J., Ledeczi, A., Bapty, T.: Model-integrated development of embedded software. Proc. IEEE **91**(1), 145–164 (2003)
10. Kumar, R., Novak, J., Tomkins, A.: Structure and evolution of online social networks. In: Link Mining: Models, Algorithms, and Applications, pp. 337–357. Springer, New York (2010)
11. Paz, F., Pow-Sang, J.A.: A systematic mapping review of usability evaluation methods for software development process. Int. J. Softw. Eng. Appl. **10**(1), 165–178 (2016)
12. Tichy, N.M., Tushman, M.L., Fombrun, C.: Social network analysis for organizations. Acad. Manag. Rev. **4**(4), 507–519 (1979)
13. Vinciotti, V., Wit, E.: Preface to the themed issue on ‘statistical network science and its applications’. J. Roy. Stat. Soc. Ser. C (Appl. Stat.) **66**(3), 451–453 (2017)
14. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications, vol. 8. Cambridge University Press (1994)
15. Lima, S., Rocha, Á., Roque, L.: An overview of OpenStack architecture: a message queuing services node. Cluster Comput. 1–12 (2017)