Editorial: Towards convergence of big data, semantics and cloud: research advances

Beniamino Di Martino*

Department of Ingegneria Industriale e dell'Informazione,
Second University of Naples,
Via Roma, 29 – 81031, Aversa (CE), Italy
Fax: +39-081-5037042
Email: beniamino.dimartino@unina.it
*Corresponding author

Flora Amato

Department of Ingegneria Elettrica e delle Tecnologie dell'informazione,
University Federico II of Naples,
Via Claudio, 21 – 80125, Napoli, Italy
Fax: +39-081-7683816
Email: flora.amato@unina.it

Biographical notes: Beniamino Di Martino is a Full Professor of Information Systems at the Second University of Naples and Vice Director of the Department of Industrial and Information Engineering. He is the author of nine international books and more than 200 publications in international journals and conferences. He is a Project Coordinator of EU funded FP7-ICT Project mOSAIC, and has been participating in various research projects supported by EC, national and international organisations. He is the Editor/Associate Editor of four international journals and editorial board member of several international journals. He is a member of several IEEE and European Commission Working Group on Cloud Computing.

Flora Amato is an Assistant Professor at the Department of Electrical Engineering and Information Technology of the University of Napoli Federico II, where she carries out her research activities since 2006. She received her PhD in Computer and Control Engineering in 2009. She was also a member of several programme and organisation committees of international conferences. Her research activities mainly concern the following issues: semantic processing in information retrieval and extraction, formal modelling and verification techniques, and knowledge management. She is a member of the GIRPR the Italian Chapter of the International Association for Pattern Recognition (IAPR).

1 Introduction

Dealing with big data realms implies to face, more than ever, two main kinds of problems:

1 efficient processing of huge, continually produced, various amount of data

2 effective extraction and association of relevant information from data.

While data processing requires more and more advanced technologies in order to improve performances and manage large scale data, on the other hand the need for innovative methods to retrieve information from data is growing up. Cloud computing technologies are one of the most effective solutions to big data processing, and semantic processing procedures usually rely on natural language processing methodologies that are effectively used to capture the meanings of data contents and correlations among data. These elements are the main topics of this special issue: 'Big data, semantics and the cloud'. Big data, semantics and cloud represent the frontiers of the modern data engineering: we believe that a real enhancement in data engineering must go towards both efficient management of data and semantic-based, meaningful use of data.

The digital era facilitated the information production and changed completely the storage mean of documents from paper to electronic. The increasingly expensiveness and complexity of traditional business applications and systems due to the quantity and variety of hardware and software necessary for their execution is leading to find alternative solutions. Due to the ease of data production within the internet era, knowledge workers are increasingly overwhelmed by information from multiple information sources: e-mails, intranets, the web, microblogs, etc., and yet still find it hard to navigate and search for accessing the specific information required for the task at hand. This implies that knowledge worker productivity is reduced and that organisations may be making decisions on the basis of incomplete knowledge. Furthermore, an inability to access key information can lead to compliance failure (Tiropanis et al., 2009).

The technical and scientific issues related to the data booming, designated as the ‘big data’ challenges, have been identified as highly strategic by major research agencies (Xhafa and Barolli, 2014; Zhiling et al., 2013). Within this scenario, search engines represent a key means to navigate and access to the data (Di Martino et al., 2014; Amato et al., 2008) enabling the user to find the right information (Amato et al., 2009; Amato et al., 2014).

Most search engines in use today strongly rely on keywords matching and on the ability of the user in the query expression. This leads to the retrieval of a large amount of irrelevant information with a direct impact on the user that spends a lot of time in browsing the results and/or to construct more complex queries to refine the search output (Amato et al., 2010).

Most definitions of big data refer on the so-called five Vs: volume, variety, velocity, verification and value. In order to address the big data issues related to the huge dimension of data-sets, the great variety of sources (including structured, semi-structured or unstructured data) and the frequency of data generation and need for elaboration, new computational architectures are needed.

Nowadays the traditional business applications focused for specific tasks are increasingly expensive and complex. The quantity and variety of hardware and software necessary for their execution are overwhelming.

Moreover the management of the traditional business applications includes installing, configuring, testing, executing, protecting and updating them and consequently it is required to dispose of a whole team of experts in order to fulfil these activities. In many cases, companies would not risk their business for software development and maintenance preferring to follow a delivery computing model, choosing a solution from the internet cloud. In fact, cloud computing obviate these problems providing the hardware and software to end users: a specialist provider, such as Amazon, Google and Microsoft, provides the necessary software as a service (SaaS), platform as a service (PaaS) or infrastructure as a service (IaaS). The shared hardware and software provides functionality similar to public services: the user pays only the required features, updates are automatic and scaling it up or down is simple.

To deal with big data analysis, innovative forms of data processing are needed in order to extracting information and discovering knowledge from the continuous and increasing data growing. The volume and variety of data directly impact on computational load. So the data increasing implies that the computational power should increase in order to reduce latencies providing actionable intelligence at the right time. Indeed, at the state of the art, innovative architectural solutions have been proposed. They range from classical data warehouse techniques to innovative cloud-based architectures that provide potentially infinite resource and power computation.

2 Related literature

New approaches of big data processing have been proposed to extract the useful information from large datasets or streams of data. These new methods are based on algorithms, tools and architectures to deal with the large amount of data in a very short time (very often real time). Some examples of big data applications are the following: costumer personalisation, churn detection, mining DNA of each person, to discover, monitor and improve health aspects of every one and clinical decision systems to support the physician diagnosis, etc.

Furthermore to process ‘big data’ the most common approaches are based on the graph algorithms, parallel and distributed architecture. For example, some big data infrastructure deals with Apache Hadoop[1] software for data-intensive distributed applications, based in the MapReduce programming model and a distributed file system called Hadoop distributed file system (HDFS). Hadoop allows designing applications which rapidly process large amounts of data in parallel on large clusters of compute nodes. A MapReduce job splits the input dataset into independent subsets that are dealt with map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to compute the final result of the job. More specific to big graph mining some open source tools exist, such as for example, Pegasus (Kang et al., 2012), a big graph mining system built on top of MapReduce. It allows to find patterns and anomalies in massive real-world graphs. Then GraphLab (Low et al., 2010) is another tool based on graph structure, high-level graph-parallel system built without using MapReduce. GraphLab computes over dependent records which are stored as vertices in a large distributed data-graph. Algorithms in GraphLab are expressed as vertex programs which are executed in parallel on each vertex and can interact with neighbouring vertices.

In big data mining, there are many open source initiatives. One most popular is the Apache Mahout[2] that is a scalable machine learning and data mining open source software based mainly on Hadoop. It contains implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining, features selection methods and so on. In Canny and Zhao (2013), a collection of hardware, software and design patterns has been introduced which manage very fast large-scale data at very low cost. The main components of this suite are: the data engines, a hardware design pattern that balances storage, CPU and GPU acceleration for typical data mining workloads, BIDMat which is an interactive matrix library that integrates CPU and GPU acceleration and novel computational kernels, BIDMach that is machine learning system that includes very efficient model optimisers, Buttery mixing, a communication strategy that hides the latency of frequent model updates needed by fast optimisers and design patterns to improve performance of iterative up-date algorithms. In Han et al. (2013), a truly parallel graph engine, called TurboGraph, has been presented that employs full parallelism including multicore parallelism and FlashSSD IO parallelism and full overlap of CPU processing and I/O processing as much as possible. In details, a novel parallel execution model, namely pin-and-slide has been introduced. Extensive experimental results with large real datasets show that TurboGraph consistently and significantly outperforms other state-of-art approaches. The TurboGraph implementation is available at http://wshan.net/turbograph as executable files. Big data Pipelines split complex analyses of large datasets into a series of simpler tasks, with independently tuned components for each task. Moreover in Raman et al. (2013) a novel model for reasoning across components of big data pipelines has been proposed in a probabilistically well-founded manner. This model is based on the interaction of components as dependencies on an underlying graphical model. Different message passing schemes on this graphical model provide various inference algorithms to tradeoff end-to-end performance and computational cost. Their framework with an efficient beam search algorithm has been instantiated. Two Big data pipelines, parsing and relation extraction, have been exploited to demonstrate the framework efficiency. The MapReduce framework has become the de-facto framework for large-scale data analysis and data mining. In Gupta and Fegaras (2013), a new design pattern for a family of iterative graph algorithms for the MapReduce framework. The method is to separate the immutable graph topology from the graph analysis results. Each MapReduce node participating in the graph analysis task reads the same graph partition at each iteration step, which is made local to the node, but it also reads all the current analysis results from the distributed file system (DFS). These results are correlated with the local graph partition using a merge-join and the new improved analysis results associated with only the nodes in the graph partition are generated and dumped to the DFS. The algorithm provides one phase for pre-processing the graph and the repetition of one map-based MapReduce job for the actual analysis. Luo et al. (2013) describes the current big data computational model DOT and DOTA, and is proposed a more general and practical model p-phases DOT (p-DOT). p-DOT is not a simple extension, but with profound significance: for general aspects, any big data analytics job execution expressed in DOT model or BSP model can be represented by it; for practical aspects, it considers I/O behaviour to evaluate performance overhead. Furthermore, describes a cost function implying that the optimal number of machines is near-linear to the square root of input size for a fixed algorithm and workload, and demonstrate the effectiveness of the function through several experiments. Sun et al. (2013) introduces a service selection model with the service location considered. The service location is introduced as a new feature helping to select the optimal services with lowest transmission cost. The huge size of data makes the transmission time been a majority cost. The distance of any two services is introduced to represent the transmission speed between these two services. The evaluation function of service selection is the simple additive weighting (SAW) of each distance between two nearby services in the sequence process. In order to solve this optimisation problem, the original data are converted into a graph and the vertices invoked in the shortest path from the data reading service to data writing service, are the optimal selection of the original optimisation problem, which is also the optimal selection (Sanders et al., 2013). Big data applications often store or obtain their data distributed over many computers connected by a network. Technical contribution are algorithms with sublinear communication volume, technical contribution are several related results on distributed bloom filter replacements, duplicate detection, and database join, single shot Bloom filters as a useful tool for reducing communication volume in distributed settings, often allowing sublinear communication volume. Kumar et al. (2013) describes the single linkage clustering based on minimal spanning tree built with the Filter-Kruskal method to the proposed clusiVAT algorithm, which is based on sampling the data, imaging the sample to estimate the number of clusters, followed by non-iterative extension of the labels to the rest of the big data with the nearest prototype rule. Numerical experiments with both synthetic and real data confirm the theory that clusiVAT produces true single linkage clusters in compact, separated data.

In the domain of cloud computing, a number of efforts are representing cloud resources, services and in general cloud concepts in OWL producing so called *cloud ontologies*. Particularly, Singh et al. (2012) proposed an ontology which is focused on the technologies involved in the cloud phenomenon and describes the different layers of cloud computing, the relationships between them and the users of each cloud layer while Moscato et al. (2011) proposed an ontology, built upon existing standards, developed to improve interoperability among existing cloud solutions, platforms and services, both from end-user and developer side. Di Martino et al. (2014a) described an unified OWL ontology of cloud resources at PaaS and SaaS level which focuses on the classification and categorisation, based on a functional analysis, of cloud services and virtual appliances. In Di Martino et al. (2014b, 2014c), the description of functional and non-functional characteristics of some specific cloud services is proposed, alongside with information related to exchanged parameters, and collaboration between services (Moscato et al., 2012a, 2014).

While cloud patterns can be extremely useful to model cloud solutions and applications and, therefore, can convey meaningful information to support software porting to the cloud and services' interoperation, they can be hampered by the lack of a shared machine readable formalism for their representation. Works aiming at defining a semantic-based formalism for the accurate description of both static and behavioural aspects of cloud patterns can be found in Di Martino and Esposito (2013) and Di Martino et al. (2013). Here cloud patterns' components are described using an OWL ontology, while the orchestration among such components is obtained through OWL-S.

A project aimed at developing an open-source platform based on cloud services is mOSAIC (Dana et al., 2013; Di Martino et al., 2011). The project enables application developers to select cloud services according to their application needs. Using the *cloud ontology*, the *semantic engine* and the *semantic service discovery*, the vendor-agnostic API and various tools, the application developers are able to specify their service requirements and communicate them to the platform. The selection process is based on the multi-agent brokering performed by the cloud agency that search for services matching the applications' request. By using mOSAIC approach and software cloud-application developers and maintainers are able to postpone their decision on the procurement of cloud services from design time until run-time, while end-user applications are able to find best-fitting cloud services to their actual needs and efficiently outsource computations and storage.

Targeting the application developer, an entire set of tools was built for an easy design of the cloud applications. In particular the semantic engine (Cretella and Di Martino, 2014; Cretella et al., 2012; Cretella and Di Martino, 2012) and dynamic semantic discovery service (Cretella and Di Martino, 2013)) support the user in discovering the resources and services offered by mOSAIC and various cloud providers, based on application and cloud patterns, and perform their semiautomatic integration in the mOSAIC API. A machine readable (OWL) cloud ontology (Moscato et al., 2011, 2012b) been defined at these purposes, which is being included in the IEEE Intercloud Standard. The selection of the cloud service to be consumed is semi-automated in mOSAIC by a unique cloud agency (Venticinque et al., 2011; Petcu et al., 2011), a multi-agent systems capable to broker and negotiate the resources and to establish the service-level-agreements with the selected cloud(s) according to the needs of the applications, and to monitor and possibly dynamically reconfigure the resources provided; six cloud commercial cloud providers and six open-source and deployable infrastructure(-as-a-)services are currently connected.

3 The selected works

We have selected six papers presenting innovative contents about big data, semantics and the cloud. In the following we report a short description for them. The work 'Heterogeneity-aware scheduler for stream processing frameworks' is focused on the problems of the stream processing application scheduling on heterogeneous clusters. It presents an overview of current state of the art of the stream processing on heterogeneous clusters with focus on the resource allocation and scheduling. Common scheduling approaches used with stream processing frameworks are discussed and their disadvantages in heterogeneous environment are demonstrated on a simple stream application. Finally, the work proposes a novel heterogeneity-aware scheduler for stream processing framework based on design-time knowledge as well as benchmarking techniques, which lead to a near optimal resource-aware deployment over the cluster nodes and thus better utilisation of the cluster itself.

The work 'Robust fingerprinting codes for database using non-adaptive group testing' proposes a tool for identifying distributors by database fingerprinting. Authors face two basic problem in fingerprinting database: designing the fingerprint and embedding it. For the first problem, they proved that non-adaptive group testing, which is used to identify specific items in a large population, can be used for fingerprinting and that it is secure against collusion attack efficiently. For the second problem, they developed a solution that supports up to 262,144 fingerprints for 4,032 attributes, and that is secure against three types of attacks: attribute, collusion and complimentary; identifying, with the hardware at their disposal, illegal distributor within 0.15 seconds.

The work 'A semantic cloud infrastructure for data-intensive medical research' describe a data infrastructure defined for a European project. In particular authors presented a platform called VPH-Share. The platform is designed for understanding physiological processes in the human body in terms of anatomical structure and biophysical mechanisms. Besides storing, sharing, integrating and linking a wide variety of heterogeneous bio-medical datasets relevant to the VPH community, the project envisions the facilitation of a secure data infrastructure, as well as search and exploration facilities using semantic technologies. The data infrastructure and management platform is built on top of a hybrid cloud environment. The platform offers tools that cover the whole life-cycle of datasets including integration, selection, semantic annotation and publishing datasets as a service. A comprehensive user interface enables end-users to search and explore bio-medical data with the support of semantic technologies, concealing the complexity of the underlying service environment.

In the work 'Why rank-level fusion? And what is the impact of image quality?' authors analyse changes of the rank assigned to the genuine identity in multimodal scenarios in presence of low quality data. The goal of a biometric identification system is to determine the identity of the input probe. In order to accomplish this, a classical biometric system uses a matcher to compare the input probe data against each labelled biometric data present in the gallery database. The output is a set of similarity scores that are sorted in decreasing order and ranked. The identity of the gallery entry corresponding to the highest similarity score (or lowest rank) is associated with that of the probe. In multibiometric systems, the outputs of multiple biometric classifiers are consolidated. Such a fusion can be accomplished at the score-level or rank-level (apart from other levels of fusion). Recent research has established benefits of rank-level fusion in identification systems; however, these studies have not compared the advantages, if any, of rank-level fusion schemes over classical score-level fusion schemes. In the presence of low quality biometric data, the genuine match score is claimed to be low and expected to be an unreliable individual output. Conversely, the rank assigned to that genuine identity is believed to remain stable even when using low quality biometric data. The contribution of the paper is two-fold:

1 investigating the rank stability in both unimodal and multimodal biometric systems

2 comparing the identification performance of rank-level and score-level fusion in the presence of low quality data.

The performance is evaluated using two datasets:

1 the first dataset is a subset of the database face and ocular challenge series (FOCS) collection (the good, bad and ugly database), composed of three frontal faces per subject for 407 subjects

2 the second dataset was collected at West Virginia University, composed of rolled fingerprints for 494 subjects (70 of these 494 are low quality).

Results show that a variant of the highest rank fusion scheme, which is robust to ties, performs better than the other non-learning-based rank-level fusion methods explored in this work. However, experiments demonstrate that score-level fusion yields better identification accuracy than existing rank-level fusion schemes.

In the work 'Malicious traffic analysis on mobile devices: a hardware solution' authors propose a hybrid computing architecture which enables the communication between the Android OS and a traffic analysis hardware accelerator, coexisting on the same chip. The security of smartphone devices, in fact, is jeopardised by viruses, intrusion attempts and trojans and most of them come from the internet traffic. Due to huge volume and complex nature, those threats are difficult to discover and immunise. The mobile devices cannot adopt classical approaches to improve security, such as the traffic analysis, because they are mobile, so resource constrained and without a power supply. Furthermore, the most widespread mobile operating systems, such as Android, do not provide facilities for this kind of task. Novel approaches have been presented in the literature, in which traffic analysis is executed in hardware using the Decision Tree classification algorithm. In order to show the feasibility of the approach, in terms of throughput, latency and energy consumption, the proposed architecture is hosted by new FPGA chip family, the Xilinx Zynq architecture, a SoPC based on dual-core ARM.

In the work 'A platform for big data analytics on distributed scale-out storage system' authors propose an innovative big data platform consisting of big data storage and big data processing. Big data analytics is the process of examining large amounts of data of a variety of types to uncover hidden patterns, unknown correlations and other useful information. Existing big data platforms such as data warehouses cannot scale to big data volumes, cannot handle mixed workloads, and cannot respond to queries quickly. Hadoop-based platform based on distributed scale-out storage system emerges to deal with big data. In Hadoop a name node is used to store metadata in a single system's memory, which is a performance bottleneck for scale-out. The Gluster file system has no performance bottlenecks related to metadata because it uses an elastic hashing algorithm to place data across the nodes and it runs across all of those nodes. In order to achieve massive performance, scalability and fault tolerance for big data analytics, a big data platform is proposed. For big data storage, Gluster file system is used and Hadoop MapReduce is applied for big data processing. The Hadoop big data platform and the proposed big data platform are implemented on commodity Linux virtual machines clusters and performance evaluations are conducted. According to the evaluation analysis, the proposed big data platform provides better scalability, fault tolerance, and faster query response time than the Hadoop platform.

References

Amato, F., Casola, V., Mazzeo, A. and Romano, S. (2010) 'A semantic based methodology to classify and protect sensitive data in medical records', in *Information Assurance and Security (IAS), 2010 Sixth International Conference on*, IEEE, August, pp.240–246.

Amato, F., Mazzeo, A., Moscato, V. and Picariello, A. (2009) 'Semantic management of multimedia documents for e-government activity', in *Complex, Intelligent and Software Intensive Systems, CISIS '09, International Conference on*, IEEE, pp.1193–1198.

Amato, F., Mazzeo, A., Moscato, V. and Picariello, A. (2014) 'Exploiting Cloud Technologies and context information for recommending touristic paths', in *Intelligent Distributed Computing VII*, Vol. 511, pp.281–287, Springer International Publishing.

Amato, F.M., Penta, A. and Picariello, A.A. (2008) 'Building RDF ontologies from semistructured legal documents', *Complex, Intelligent and Software Intensive Systems, CISIS, International Conference on.*

Canny, J. and Zhao, H. (2013) 'Big data analytics with small footprint: squaring the cloud', in *Proceeding of KDD*, pp.95–103.

Cretella, G. and Di Martino, B. (2012) 'Towards a semantic engine for cloud applications development support', *Proc. of CISIS: The Sixth International Conference on Complex*, *Intelligent*, *and Software Intensive Systems*, IEEE CS Press, Palermo, Italy, July 4–6, pp.198–203, ISBN: 978-0-7695-4687-2, DOI 10.1109/CISIS.2012.159.

Cretella, G. and Di Martino, B. (2013) 'Semantic and matchmaking technologies for discovering, mapping and aligning cloud providers' services', *Proc. of 15th International Conference on Information Integration and Web-based Applications and Services (iiWAS2013)*, ACM Press, Vienna, Austria, pp.380–384, 2–4 December, ISBN 978-1-4503-2113-6.

Cretella, G. and Di Martino, B. (2014) 'A semantic engine for porting applications to the cloud and among clouds', *Software – Practice and Experience (SPE)*, Wiley Online Library, doi: 10.1002/spe.2304.

Cretella, G., Di Martino, B. and Stankovski, V. (2012) 'Using the mOSAIC's semantic engine to design and develop civil engineering cloud applications', *Proc. of 14th International Conference on Information Integration and Web-based Applications & Services (iiWAS)*, IEEE CS Press, December 2–4, pp.378–386, ISBN 978-1-4503-1306-3.

Dana, P., Di Martino, B., Salvatore, V., Massimiliano, R., TamÃ¡s, M., Gorka, E.L., Fabrice, B., Roberto, C., Miha, S., Svatopluk, S. and Vlado, S. (2013) 'Experiences in building a mOSAIC of clouds', *Journal of Cloud Computing: Advances, Systems and Applications*, May, Vol. 2, No. 12, p.22, Springer, DOI: 10.1186/2192-113X-2-12.

Di Martino, B. and Esposito, A. (2013) 'Towards a common semantic representation of design and cloud patterns', *Proc. of 15th International Conference on Information Integration and Web-based Applications & Services (iiWAS)*, ACM Press, ISBN: 978-1-4503-2113-6, pp.385–390.

Di Martino, B., Cretella, G. and Esposito, A. (2013) 'Semantic and agnostic representation of cloud patterns for cloud interoperability and portability', *Proc. of IEEE Fifth International Conference on Cloud Computing Technology and Science (CloudCom)*, IEEE CS Press, Vienna, ISBN: 978-0-7695-5095-4 doi 10.1109/CloudCom.2013.123, 2–5 December, pp.182–187.

Di Martino, B., Aversa, R., Cretella, G., Esposito, A. and Kolodziej, J. (2014a) 'Big data (lost) in the cloud', *International Journal of Big Data Intelligence (IJBDI)*, Vol. 1, Nos. 1/2, pp.3–17, Inderscience.

Di Martino, B., Cretella, G. and Esposito, A. (2014b) 'Towards an unified OWL ontology of cloud vendors' appliances and services at PaaS and SaaS level', *Proc. of 8th International Conference on Computational Intelligence in Security for Information Systems (CISIS)*, IEEE CS Press, ISBN 978-1-4799-4325-8/14, pp.570–575.

Di Martino, B., Cretella, G., Esposito, A. and Sperandeo, R. (2014c) 'Semantic representation of cloud services: a case study for Microsoft windows azure', *Proc. of 6th International Conference on Intelligent Networking and Collaborative Systems (INCoS)*, IEEE CS Press, ISBN 978-1-4799-6387-4/14, pp.647–652.

Di Martino, B., Cretella, G., Esposito, A. and Carta, G. (2014d) 'Semantic representation of cloud services: a case study for openstack', in G. Fortino et al. (Eds.): *Internet and Distributed Computing Systems*, *Lecture Notes in Computer Science*, No. 8729, pp.39–50, Springer, Heidelberg, ISBN 978-3-319-11691-4.

Di Martino, B., Petcu, D., Cossu, R., Goncalves, P., Mahr, T. and Loichate, M. (2011) 'Building a mOSAIC of clouds', in M.R. Guarracino et al. (Eds.): *Euro-Par Workshops*, *Lecture Notes in Computer Science,* No. 6586, pp.571–578, Springer, Heidelberg, ISBN 978-3-642-21877-4.

Gupta, U. and Fegaras, L. (2013) 'Map-based graph analysis on MapReduce', in *Proceeding of IEEE 20th International Conference on Web Services.*

Han, W.S., Lee, S., Park, K., Lee, J.H., Kim, M.S., Kim, J. and Yu, H. (2013) 'TurboGraph: a fast parallel graph engine handling billion-scale graphs in a single PC', in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM*, August, pp.77–85.

Kang, U., Chau, D.H. and Faloutsos, C. (2012) 'PEGASUS: mining billion-scale graphs in the cloud', *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.5341–5344, DOI: 10.1109/ICASSP.2012.6289127.

Kumar, D., Palaniswami, M., Rajasegarar, S., Leckie, C., Bezdek, J.C. and Havens, T.C. (2013) 'clusiVAT: a mixed visual/numerical clustering algorithm for big data', in *Proceeding of IEEE 20th International Conference on Web Services*.

Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C. and Hellerstein, J.M. (2010) 'Graphlab: a new parallel framework for machine learning', in *Conference on Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, California, July.

Luo, T., Liao, Y., Chen, G. and Zhang, Y. (2013) 'P-DOT: a model of computation for big data', *2013 IEEE International Conference on Big Data*, pp.31–37.

Moscato, F., Amato, F., Amato, A. and Aversa, R. (2014) 'Model-driven engineering of cloud components in MetaMORP(h) OSY', *International Journal of Grid and Utility Computing*, Vol. 5, No. 2, pp.107–122.

Moscato, F., Aversa, R. and Amato, A. (2012a) 'Describing cloud use case in MetaMORP(h)OSY', *Proceedings 6th International Conference on Complex, Intelligent, and Software Intensive Systems*, *CISIS*, pp.793–798.

Moscato, F., Di Martino, B. and Aversa, R. (2012b) 'Enabling model driven engineering of cloud services by using mOSAIC ontology', *Scalable Computing*, March, Vol. 13, No. 1, pp.29–44, ISSN: 1895-739X.

Moscato, F., Aversa, R., Di Martino, B., Petcu, D., Rak, M. and Venticinque, S. (2011) 'An ontology for the cloud in mOSAIC', in: Lizhe Wang, Rajiv Ranjan, Jinjun Chen and Boualem Benatallah (Eds.): *Cloud Computing: Methodology*, *System*, *and Applications*, pp.467–487, CRC Press, Taylor & Francis Group, ISBN 978-1-4398-5641-3.

Petcu, D., Craciun, C., Neagul, M., Panica, S., Di Martino, B., Venticinque, S., Rak, M. and Aversa, R. (2011) 'Architecturing a sky computing platform', in M. Cezon and Y. Wolfsthal (Eds.): *Towards a Service based Internet –Service Wave*, *Lecture Notes in Computer Science*, Springer, Heidelberg, No. 6569, pp.1–13, ISBN 978-3-642-22759-2.

Raman, K. et al. (2013) 'Beyond myopic inference in big data pipelines', *Proceedings of the 19th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp.86–94.

Sanders, P., Schlag, S. and Muller, I. (2013) 'Communication efficient algorithms for fundamental big data problems', in *Proceeding of IEEE 20th International Conference on Web Services*.

Singh, G., Kaur, N. and Kaur, M. (2012) 'Toward a unified ontology of cloud computing', *International Journal of Computers and Technology*, Vol. 3, No. 2, pp.1–10, ISSN 22773061.

Sun, R. et al. (2013) *System and Method for Transmission Point (TP) Association and Beamforming Assignment in Heterogeneous Networks*, US Patent Application 13/757,303.

Tiropanis, T., Davis, H., Millard, D. and Weal, M. (2009) 'Semantic technologies for learning and teaching in the Web 2.0 era', *Intelligent Systems*, IEEE, Vol. 24, No. 6, pp.49–53.

Venticinque, S., Aversa, R., Di Martino, B., Rak, M. and Petcu, D. (2011) 'A cloud agency for SLA negotiation and management', in M.R. Guarracino et al. (Eds.): *Euro-Par Workshops*, *Lecture Notes in Computer Science*, Springer, Heidelberg, No. 6586, pp.587–594, ISBN 978-3-642-21877-4.

Xhafa, F. and Barolli, L. (2014) 'Semantics, intelligent processing and services for big data', *Future Generation Computer Systems*, Vol. 37, pp.201–202, doi:10.1016/j.future.2014.02.004.

Zhiling, L., Ying, L. and Jianwei, Y. (2013) 'Location: a feature for service selection in the era of big data', in *Proceeding of IEEE 20th International Conference on Web Services*.