Co-citation Analysis of Literature in e-Science and e-Infrastructures

Journal:	*Concurrency and Computation: Practice and Experience*
Manuscript ID	CPE-19-0172.R1
Editor Selection:	Prof. David Walker
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	05-Sep-2019
Complete List of Authors:	Mustafee, Navonil; University of Exeter, University of Exeter Business School Bessis, Nik; EdgeHill University Taylor, Simon; Brunel University, Information Systems and Computing Hou, Jianhua; Sun Yat-sen University, School of Information Management Matthew, Peter; Edge Hill University
Keywords:	Co-citation analysis, e-Science, e-Infrastructure, grid computing, desktop grid computing, cloud computing

Co-citation Analysis of Literature in e-Science and e-Infrastructures

Navonil Mustafee[a], Nik Bessis[b], Simon J. E. Taylor[c], Jianhua Hou[d], Peter Matthew[b]

[a]The Centre for Simulation, Analytics and Modelling (CSAM), University of Exeter, Exeter, EX4 4ST, UK
N.Mustafee@exeter.ac.uk (*correspondig author* – Navonil Mustafee; Tel: +44 (0) 1392 725661)

[b] Department of Computer Science, Edge Hill University, St Helens Rd, Ormskirk L39 4QP, UK
Nik.Bessis@edgehill.ac.uk, Peter.Matthew@edgehill.ac.uk

[c]Department of Computer Science, Brunel University London, Uxbridge, UB8 3PH, UK
Simon.Taylor@brunel.ac.uk

[d]School of Information Management, Sun Yat-sen University, Guangzhou, Guangdong,China
Houjh5@mail.sysu.edu.cn

Abstract

Advances in computer networking, storage technologies and high-performance computing are helping global communities of researchers to address increasingly ambitious problems in Science collaboratively. E-Science is the "science of this age"; it is realized through collaborative scientific enquiry which requires the utilization of non-trivial amounts of computing resources and massive data sets. Core to this is the integrated set of technologies collectively known as e-Infrastructures. In this paper, we explore the e-Science and the e-Infrastructure knowledge base through co-citation analysis of existing literature. The dataset for this analysis is downloaded from the ISI Web of Science and includes over 12,000 articles. We identify prominent articles, authors and articles with citation bursts. The detection of research clusters and the underlying seminal papers provide further insights. Our analysis is an important source of reference for academics, researchers and students starting research in this field.

Keywords

Co-citation analysis, e-Science, e-Infrastructure, grid computing, desktop grid computing, cloud computing

1. Introduction

E-Science can be defined as Science that necessitates the utilization of non-trivial amounts of computing resources and massive data sets to perform scientific enquiry; science that requires access to remote scientific instruments and distributed software repositories; science that generates data that may demand analysis from experts belonging to multiple organizations and specialists in different knowledge domains (virtual research communities (VRCs))[1]. Such science is usually carried out in highly distributed environments by exploiting advanced hardware and software technologies [2]. Core to the growth of e-Science is the integrated set of technologies collectively known as *e-Infrastructures* or *cyberinfrastructures* (terms that emerged concurrently in Europe and North America in the late 2000s) [3]. These include high-speed research communication networks, powerful computational resources (dedicated high-performance computers, clusters, large numbers of commodity PCs), grid and cloud technologies, data infrastructures providing access to data sources, sensors and support for different forms of access (general and specific web-based portals, science gateways and mobile devices). Before the prevalence of the terms e-Infrastructure and cyberinfrastructure, it was already recognized that grid computing provided on-demand access to large amounts of computational power [4] and this focus on providing large-scale computational power made it an enabling technology for e-Science [5].

There are numerous examples of public-funded e-Science projects that are galvanizing VRCs in various disciplines across the world. For instance, a well-known example of a VRC is the international community of physicists engaged in High Energy Physics that are investigating the fundamental properties of the Universe with CERN's *Large Hadron Collider* (LHC). The LHC produces around 650 MBps of data and new scientific equipment planned in the near term will produce 1.25 GBps. This is supported by the *Worldwide LHC Grid* that includes 150 computing and storage sites in 35 countries [3]. In life sciences, one significant VRC is *Biomed*[6] which supports communities with interests in medical imaging (computerized analysis of digital medical images), bioinformatics (gene sequences analysis) and drug discovery (in-silico simulations). Biomed supports the WISDOM malaria initiative, which uses high-speed molecular docking to discover effective drugs against this pervasive disease [7]. HealthGrids are an emerging class of Grid computing that specifically deal with the needs of processing biomedical data and access to medical expertise and devices [8]. There are numerous other examples, and the reader is referred to [2].

Arguably, the majority of e-Science applications in use today are predominantly executed over e-Infrastructures that comprise of dedicated and high-performance clusters (referred to as cluster-based grid computing); however, two other forms of distributed computing, namely, desktop grid computing and cloud

computing, are increasingly being considered as viable alternatives for executing e-Science applications. Furthermore, there is research which focuses explicitly on interoperability amongst these distributed computing technologies thereby making it possible for e-Science applications to be transparently scheduled over the underlying e-Infrastructure technologies comprising of not only grid resources, desktop grid and cloud resources. Notable examples include utilizing *EGEE* (*Enabling Grids for E-sciencE* – now superseded by the *European Grid Infrastructure*; the largest grid infrastructure in the world) for desktop grids [9] and interoperating grids and clouds together [10,11].

In this paper, we explore the e-Science and the e-Infrastructure knowledge domains through co-citation analysis of literature. The purpose of this analysis is to identify the top 20 most significant authors and articles; we refer to them as the turning point articles and authors, and they are identified not merely by the frequency of their citation and co-citation counts, but also by their contribution in shaping the direction of research in this area. A turning point may be identified by analyzing the attributes of a node in a co-citation network. For example, nodes that connect to other nodes from different periods of time, nodes that have rapidly growing number of citations (citation bursts), widely co-cited nodes are all considered as candidates for intellectual turning points [12]. In order to reduce the subjectivity which is inherent in undertaking such an exercise (i.e., creating a ranked list in terms of contribution), we have used the co-citation analysis tool called CiteSpace [12] to base our discussions on objective analysis. However, not all authors publish in journals and conferences that are indexed by the *ISI Web of Science,* which is the source of our dataset, and this is a limitation of this work.

The remainder of the paper is organized as follows. Section 2 presents an overview of e-Infrastructures and associated technologies such as grid computing, desktop grid computing and cloud computing. Section 3 is on co-citation analysis, which is the primary bibliometric method used in this paper. Section 4 describes the dataset and the tool used for analysis. Sections 5 presents our findings on prominent articles (5.1), authors (5.2) and articles with citation burst (5.3). Section 6 is on cluster analysis. Section 7 is the concluding section of our paper.

2. E-Infrastructures

A typical architecture of an e-Infrastructure is shown in Figure 1. A set of high-performance networks (e.g., Internet2 and GÉANT3) support the fast transfer of information between elements of an e-Infrastructure. Above this are various High Performance Computing (HPC) resources such as the Grid computing facilities offered by the European Grid Infrastructure (*EGI.eu*), the supercomputing resources offered by PRACE (*prace-ri.eu*), and distributed computing infrastructures (DCIs) such as institutional desktop grids and cloud computing (*e.g. www.sci-bus.eu*), etc.. Data infrastructures consisting of data repositories (DR) and open

access data repositories (OADR) support the curation of scientific data (e.g., the European project on 'Co-ordination and Harmonization of Advanced e-Infrastructures for Research and Educational Data Sharing'; *chain-project.eu*), sensors and instrumentation are linked together using standard protocols. Access to these resources are governed by an authentication and authorization infrastructure that uses Certification Authorities (CAs) and Identity Federations (IdFs) to facilitate single sign-on (SSO) access to members of the VRCs via Virtual Organizations (VOs). Direct access via standardized middleware such as *gLite*, or web-based general or specific science gateways (e.g. *WS-PGRADE-based portals*) support easy access to these resources and implement a range of applications used by VRCs. Other services such as Eduroam (*eduroam.org*), a secure world-wide roaming service developed for the international research and education community, take advantage of this level of integration and offer further support for collaboration. VOs tend to have a strong computational and engineering focus, although efforts are being made to expand into the social sciences and humanities. Global infrastructural initiatives drive the dissemination of their approaches by providing support for training and outreach. These also have government-level backing and funding (for example via the US National Science Foundation's *Office of Cyberinfrastructure* and the European Commission's *DG-INFSO GÉANT and e-Infrastructure unit*).

<<Figure 1 about here>>

The technology enablers of e-Science and the VRCs are the e-Infrastructures; the computing elements of an e-Infrastructure comprise of grid, desktop grid and cloud computing resources. The majority of the scholarly publications on e-Science and e-Infrastructure refer to these underlying distributed computing technologies and therefore, in the remainder of this section, we present a brief introduction to these technologies. This section will be especially useful for those starting research in this field.

2.1. Grid Computing

The notion of a computational grid was outlined by Ian Foster and Carl Kesselman in their edited book *The Grid: The Blueprint for a New Computing Infrastructure*, as a hardware and software infrastructure that provides access to high-end computational resources [13]. It was further stated that this access should be dependable, consistent, pervasive and inexpensive. It was expected that computational grids would serve not only the scientific communities but also the government. Thus, national-level scenarios, e.g. response to environmental and human-made disasters, climate change simulations, could gain through the use of the nation's fastest computers, data archives and shared intellect. Solving such problems usually necessitates investment in high-end computing resources[14]. However, in concert with such investments, it is judicious that underutilized computation facilities within research organizations and universities be leveraged to derive maximum utilization of existing resources. Desktop grid computing makes this possible.

2.2. Desktop Grid Computing

While much of grid computing is focused on meeting the needs of large VRCs such as academic institutions and R&D centres engaged in e-Science, desktop grid computing, or desktop grids, addresses the potential of harvesting the idle computing resources of desktop PCs [15]. These resources can be part of the same local area network (e.g. a network of PCs in a university) or can be geographically dispersed and connected via a wide area network such as the internet (e.g. computers in different university campuses connected through research and educational networks, like JANET in the UK - *www.ja.net*). Studies have shown that desktop PCs can be underutilized by as much as 75 per cent of the time [16]. Furthermore, spare capacities are available on an hour-to-hour, day-to-day, and month-to-month basis, not only during the evening hours and on weekends, but during the busiest times of normal working hours [17]. This coupled with the widespread availability of desktop computers and the fact that the power of network, storage and computing resources is projected to double every nine, 12 and 18 months, respectively[18], represents an enormous computing resource that can be used for e-Science projects.

Another form of desktop grid computing is Volunteer Computing (VC) wherein the underlying grid infrastructure is composed of computational resources that are donated by the users, e.g., *SETI@Home* project. Although VC was originally conceived for large-scale scientific computation requiring millions of volunteer PCs, more recently, however, VC grids such as the *SZTAKI* Desktop Grid [19] have been used for meeting the computation needs of an organization alone by establishing an institutional VC grid [20].

2.3. Cloud Computing

Cloud computing is defined as a type of "parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers" [21]. The cloud infrastructure usually comprises of commodity hardware and specialized programs for virtualization and resource management. The software components enable dynamic provisioning of hardware resources based on the requirements of the user. Furthermore, through the use of virtualization technology, several instances of virtual machines are started over the provisioned physical resource (every virtual machine has its own copy of the operating system and is sandboxed). Finally, user applications are executed within the confines of the virtual machines. The use of the cloud resources by the user application is usually monitored using accounting applications (e.g. CPU cycles used, data storage utilized, and network bandwidth consumed). The clouds implement a pay-per-use model, and therefore, accounting is a crucial part of this distributed computing technology. Providers of commercial cloud computing platform include Amazon Elastic Compute Cloud EC2 <*aws.amazon.com/ec2/*>, Amazon

Simple Storage Service S3 <*aws.amazon.com/s3/*>) and Eucalyptus Eucalyptus<*www.eucalyptus.com*>. Since cloud computing is a comparatively new distributed computing technology, the uptake of cloud computing as an infrastructure platform for executing e-Science applications has been justifiably low. One notable e-Science project that has used cloud technology is CARMEN – a system that allows neuroscientists to share, integrate and analyze data [22]. EGI.eu is focussing more on cloud computing as being a significant way forward in e-Science, and new interoperable cloud technologies such as *CloudBroker* are being developed (see *www.sci-bus.eu* and *www.cloudsme.eu*).

3. Content Analysis and Co-Citation Analysis

Content analysis is a research method widely used in library and information sciences; it can be applied in qualitative, quantitative and mixed modes for the purposes of systematic and rigorous analysis of documents [23]. Profiling study is a form of content analysis that is usually conducted in relation to a particular journal [24], studies that compare journals [25], or indeed those that aim to methodologically study the contribution of specific research fields and application domains, e.g., analysis of literature in cloud computing [26] and healthcare modelling and simulation[27]. Such studies help to identify currently under-explored research issues, and select theories and methods appropriate to their investigation, all of which are recognized in Information Systems as important issues for conducting fruitful, original and rigorous research [24,28]. Like a profiling study, co-citation analysis can be applied in the context of scholarly publications to identify prominent articles, authors and journals being referenced by the *citing* authors. It identifies co-cited references that occur in the reference list of two or more citing articles, with the resultant co-citation network providing insights into the constituents of a knowledge domain. *Co-citation analysis identifies clusters of "co-cited" references by creating a link between two or more references when they co-occur in the reference lists of citing articles* [29]. Studies that have used co-citation analysis include the study of the Information Science discipline [30-36], the reviews on the intellectual structure of Management Information Systems [37,38], Operations Management [39], strategic management [40], international management [41] and Science in general [42]. However, there is presently no study that has investigated the e-Science and e-Infrastructure knowledge base through co-citation analysis. The co-citation analysis that is presented in this paper uses a visualization-based analysis of bibliographic data downloaded from the *ISI Web of Science*(*http://apps.webofknowledge.com/*) and is an approach similar to that used by [43] - who present a visual survey of agent-based computing; [44]– who visualize research on pervasive and ubiquitous computing; [45] – who used this approach towards visualization of patents and papers in terahertz technology; [46] – who use co-citation analysis for exploring the modelling and simulation knowledge base.

As stated earlier, in this paper, we explore the e-Science and e-Infrastructure knowledge bases through an analysis of co-citations. The resultant co-citation networks provide important insights into knowledge

domains by identifying frequently co-cited papers, authors and journals related to the domains in question, and which could have been overlooked if only conventional citation analysis techniques were used. In a citation-based analysis, the significance of an article is often measured on the basis of the number of citations it has had. However it can be argued that there may exist articles that can be considered high-impact even though the number of citations received are comparatively less (for example, papers that have been cited a few times but across domains; papers that have been cited consistently through the years; papers that have been published recently). Furthermore, it usually takes a few years for a paper to build up its citation count. Thus, complementing traditional citation-based metrics with co-citation analysis is arguably a superior approach for identifying articles that hold promise and which represent the grounded knowledge base of a discipline. Through the investigation of existing e-Science and e-Infrastructure literature, the objective of this paper is to demonstrate this added value of using co-citation analysis (as compared with citation-only analysis) when it comes to undertaking bibliometric research. Furthermore, we have identified prominent authors and articles in this field irrespective of their co-citation frequencies.

4. Methodology

4.1 Data Selection

For this study, we searched the *ISI Web of Science (WOS) Core Collection*. We used the keywords *"e-Science"* OR *"e-Infrastructure"* OR *"cloud computing"* OR *"cyberinfrastructure*" OR *"grid computing"* OR *"desktop grid computing"* to conduct a *topic search* on the article title, abstract, author keywords and ISI keywords plus. Using this search criterion (refer to **Appendix A** for further details), the total number of unique articles retrieved was **12,516**. Next, we downloaded ISI-format meta-data associated with the records. The article meta-data from WOS are tagged using a two-character field tag, for example, the tag TI refers to the article title, AU (authors), SO (publication name), PT (publication type – journal, book), SE (book series title), PY (year of publication), AB (abstract), CR (cited references), BE (editors). For the purposes of co-citation analysis, the important fields used by tools for co-citation analysis (*CiteSpace*, *Bibliometrix* and *VOSviewer*) include, AU, SO, PY, CR.

4.2 Method and Tool

In this study, we mainly use citation analysis method in bibliometrics to conduct co-citation network analysis of research literatures in the fields of e-science and e-infrastructures. Based on structural characteristics and node indicators of the co-citation network, we analyze the classic literatures, turning point literatures and focus literatures in this field, we explore the evolution process of research topics in e-science and e-infrastructures and the research front topics. The use of the tool requires careful selection of a multitude of options (see Figure 2), and an acceptable options' combination frequently requires learning

through “trial and error” as well as knowledge of the underlying research domain. Nodes and links are the building blocks of a co-citation network. The co-citation network constructed by CiteSpace software system is the main method of citation analysis. In the analysis of co-citation network, the co-citation relationship between articles or authors can be abstracted into a network composed of nodes and wires. In which, the nodes representing articles or authors can be classified into different types according to their different positions and importance in the network. These nodes mainly include (1) *landmark nodes* (in the network, it has higher citation rates and the node size is bigger, it is a classic article or classic author in the field), (2) *burst nodes* (in the network, it has high burstness, the node has a drastic increase or decrease of the co-citation frequency at a particular time period) and (3) *pivot nodes* (in the network, it has higher center degree, that is to say, it connects the key nodes of different topic clustering, which act as "Bridges" in the network connection). Therefore, these different types of nodes represent different article or author types. Landmark nodes article represents classic articles in the research field. Burst node article represents the focus article in the research field within a certain period of time, and represents the hot topics in this period of time. Pivot nodes article is the most important turning point article in the research field, which represents the theme differentiation or cross-research topic in the research field. CiteSpace identifies turning points associated with articles and authors irrespective of citation count. This is achieved through the use of the full feature set of CiteSpace, including visual identifications of significant articles and authors through innovative visualization techniques [12].

To ensure the repeatability of this exercise, we now present the specific option values that were selected in CiteSpace. (a) *Time interval of analysis:* 1994-2014 inclusive; (b) *The unit of analysis:* 3 years per time slice (giving us a total of seven time slices, namely, 1994-1996, 1997-1999 .. 2012-2014); (c) *Cited Reference:* our analysis included top 1.0% of most-cited or occurred items from each slice, with the maximum number of selected items per slice being restricted to 100 (this is done to reduce the computation time); (d) *Pruning and merging:* minimum spanning tree [47] is used to prune the merged networks; (e) *Visualization:* A merged network cluster view has been selected. An extensive discussion of these variables is outside the scope of this paper, and the reader is referred to [48].

<<Figure 2 about here>>

In this study, we have used CiteSpace to generate co-citation networks pertaining to cited articles – this is also referred to as *Document Co-citation Network (DCN)* and cited authors - *Author Co-citation Network (ACN)*. The resultant networks have been used for answering questions one and two below. The technologies associated with e-Science and e-Infrastructures are constantly evolving and, therefore, it is important to identify and appreciate the specific areas of research which have been most active in this field of study. With this in mind, questions three and four are introduced.

- *Question One:* Which are the top 20 turning point articles (i.e., articles with high significance, irrespective of the article co-citation count)? We will use DCN to answer this.
- *Question Two:* Who are the top 20 turning point authors (i.e., authors with high impact, irrespective of author co-citation count)? We will use ACN to answer this.
- *Question Three:* Which are the top articles with citation burst? We will use DCN to answer this.
- *Question Four:* What are the prominent groupings of articles, or clusters, associated with DCN?

Section 5 will present the findings related to questions 1-3. Section 6 is on cluster analysis and will present findings related to question 4.

5. Findings

5.1 Turning Point Articles

In this section, we analyse turning point articles (Figure 3, Table 1) and reflect on the importance of these in the e-Science and the e-Infrastructure knowledge base. The time-sliced co-citation networks are distinguished by their colour. The nodes identify the articles; they are connected through links. The links can visually represent various characteristics of the underlying network, for example, the colour of the link represents the year in which a connection between two nodes was first established (in the ACN analysis it is the year in which the two authors were first co-cited), the strength of connection between two nodes is represented by the thickness of the link (in this analysis, the stronger the connection between two nodes, the greater the frequency of co-citations among the authors). The nodes in Figure 3 and Figure 4 are illustrated as multi-coloured circles of varying diameter; the number of citations determines the diameter associated with the nodes. The colours represent the year of citation, and the diameter of the coloured citation rings visually denotes the number of citations in the corresponding year. The text beside the citation rings identifies the article (Figure 3) / author (Figure 4) being represented by the node. For NT "references", the purple rings that surround the citation ring identify turning point articles (Figure 3) and turning point authors (Figure 4).

<<Figure 3 about here>>

From the DCN in Figure 3, we see two distinct clusters of co-citations that have emerged. The first cluster is on grid computing. It identifies several turning point papers by Fosters in the timespan 1997-2004. The second cluster of co-citations is specific to cloud computing. These include the seminal papers that defined this area of research, for example, papers by Armbrust ("Above the Clouds: A Berkeley View of Cloud Computing" and "A View of Cloud Computing"), Buyya ("Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility") and Vaquero ("A break in the clouds:

towards a cloud definition"). Interestingly, all of these papers appear in our citation burst analysis (section 5.3). Figure 3 identifies the work of Dean ("MapReduce: Simplified Data Processing on Large Clusters") as a turning point article. This paper from the 50th-anniversary issue of the *Communications of the ACM* has more than 20,000 citations (November 2017); this is aptly illustrated by the radius of the node, labelled in Fig 3 as Dean J (2008). What is interesting is that, if we compare the DCN characteristic of this paper with the other turning point papers, the former does not have many co-citations (which would have meant a denser network like those identified by clusters repetitive of grid and cloud computing), however, it is linked to seminal papers emerging from both the grid and cloud world. This goes to show the importance of this work.

Table 1:Top twenty turning point articles identified using Document Co-Citation Network(note: the articles are listed based on their relative importance in DCN, column one #Num, and not merely on frequency).

#Num	Freq	Article
1	41	Casanova H, Dongarra J (1997) Netsolve: a Network-Enabled Server for Solving Computational Science Problems. *International Journal of High Performance Computing Applications*, 11(3):212-223.
2	117	Litzkow M, Livny M, Mutka M (1988) Condor-a hunter of idle workstations. In *Proceedings of the 8th International Conference of Distributed Computing Systems*, pp. 104-111.
3	500	Foster I, Kesselman C (1997) Globus: a Metacomputing Infrastructure Toolkit. *International Journal of High Performance Computing Applications,* 11(2):115-128
4	121	Foster I, Kesselman C (1999) *The Grid: Blueprint for a New Computing Infrastructure.* San Francisco, CA: Morgan Kaufmann.
5	66	Buyya R, Abramson D, Venugopal S (2005) The Grid Economy. *Proceedings of the IEEE*, 93(3):698-714.
6	44	Altschul SF, Madden TL, Schäffer AA, et al. (1997) Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research*, 25(17):3389-3402.
7	26	Abramson D, Giddy J, Kotler L (2000) High Performance Parametric Modeling with Nimrod/G: Killer Application for the Global Grid? In *Proceedings of the 2000 International Parallel and Distributed Processing Symposium*, pp. 520-520. IEEE.
8	20	Butler R, Welch V, Engert D, Foster I, et al. (2000). A National-scale Authentication Infrastructure. *Computer*, 33(12): 60-66.
9	11	Johnston W E, Gannon D, Nitzberg B (1999) Grids as Production Computing Environments: The Engineering Aspects of NASA's Information Power Grid. In *Proceedings of the 8th International Symposium on High Performance Distributed Computing*, pp. 197-204. IEEE.
10	592	Foster I, Kesselman C, Tuecke S (2001) The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International Journal of High Performance Computing Applications*, 15(3):200-22.
11	101	Oinn T, Addis M, Ferris J, Marvin D, et al. (2004) Taverna: A Tool for the Composition and Enactment of Bioinformatics Workflows, *Bioinformatics*, 20(17):3045-3054.
12	89	Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool, *Journal of Molecular Biology*, 215(3), 403-410.
13	88	Foster I, Kesselman C, Nick JM, Tuecke S (2002) Grid Services for Distributed System Integration, *Computer*,

		35(6), 37-46.
14	47	Von Laszewski G, Foster I, Gawor J, Lane P (2001) A Java Commodity Grid Kit. *Concurrency and Computation: Practice and Experience*, 13(8-9):645-662.
15	22	Stevens R D, Robinson AJ, Goble CA (2003) myGrid: Personalised Bioinformatics on the Information Grid.*Bioinformatics*, 19(suppl 1), i302-i304.
16	100	Foster I, Kesselman C, Tsudik G, Tuecke S (1998) A Security Architecture for Computational Grids. In *Proceedings of the 5th ACM Conference on Computer and Communications Security*, pp. 83-92. ACM.
17	77	Deelman E, Singh G, Su MH, Blythe J, et al. (2005) Pegasus: A Framework for Mapping Complex Scientific Workflows onto Distributed Systems.*Scientific Programming*, 13(3): 219-237.
18	75	Buyya R, Abramson D, Giddy J, Stockinger H (2002) Economic Models for Resource Management and Scheduling in Grid Computing.*Concurrency and Computation: Practice and Experience*, 14(13-15): 1507-1542.
19	66	Ludäscher B, Altintas I, Berkley C, Higgins D, et al. (2006) Scientific Workflow Management and the Kepler System.*Concurrency and Computation: Practice and Experience*, 18(10):1039-1065.
20	62	Grimshaw AS, Wulf WA, the Legion Team (1997) The Legion vision of a worldwide virtual computer. *Communications of the ACM*, 40(1):39-45.

Our analysis shows, 17 of the papers in the top 20 papers represent the key foundation articles to the general development of e-Infrastructures and e-Science. The other three papers (paper 11, 12 and 15) represent key foundation articles to Bioinformatics, with one article describing an important contribution to both areas (*Taverna* workflow; paper 11). Published from 1997 to 2002, these papers put forward different approaches that have influenced distributed computing and the linking together of computers across the world to work on computational problems: Casanova and Dongarra (1997) (paper 1) ***NETSOLVE***, Grimshaw and Wulf (1997) ***LEGION***, Foster and Kesselman (1997) (paper 3) ***GLOBUS***, Litzkow, Livny and Mutka (1998) (paper 2) ***CONDOR*** and Abramson, Giddy and Kotler (2000) (paper 7) ***NIMROD/G***. Two represent key works that describe the architectural components of "The Grid" or Grid Computing, and today's e-Infrastructure architectures; the book "The Grid: Blueprint for a New Computing Infrastructure" (Foster and Kesselman 1999) (paper 4) and the Grid "Anatomy" paper (Foster, Kesselman and Tuecke 2001) (paper 10). Foster, et al. (1998) (paper 16) is a key article that addresses Grid security when accessing distributed computing facilities in different administrative domains. This theme was further developed in Butler, et al. (2000) (paper 8). Foster, et al. (2002) (paper 13) continue the development of The Grid with the introduction of Grid Services, an important step towards service-oriented computing and the development of Cloud computing. Johnston, Gannon and Nitzberg (1999) (paper 9) is one of the first papers to show the potential of using a Grid to support large-scale research and engineering efforts in a professional (production) environment (at NASA). Buyya, et al. (2002) (paper 18) and Buyya, Abramson and Venugopal (2005) (paper 5) discuss the emerging Grid economy and the relationship between Grid providers and users.

The concept of linking together sequences of tasks, or workflows, to run on a Grid were introduced by Oinn, et al. (2004) (paper 11) (***TAVERNA***), Deelman, et al. (2005) (paper 17) (***PEGASUS***) and Ludäscher, et al. (2006) (paper 19) (***KEPLER***). Although it can be used to support different application domains, *TAVERNA* was developed to support Bioinformatics applications and therefore represents a major contribution in that area. Laszewski, et al. (2001) (paper 14) represents the core paper for the development of Commodity Grids (CoGs) and introduces the concept of Science Portals (which with contemporary Web technologies has developed into today's Science Gateways). Stevens, Robinson and Goble (2003) (paper 15) introduce ***myGrid*** that discusses the development of end-user services for Bioinformatics researchers and represents another important step towards the development of Science Gateways. Altschul, et al. (1990) (paper 12) and Altschul, et al. (1997) (paper 6) are the main reference for the *Basic Local Alignment Search Tool (BLAST)* programs, widely used tools for searching protein and DNA databases for sequence similarities; these tools support Bioinformatics researchers and arguably one of the largest group of end-users of e-Science and e-Infrastructures.

5.2 Turning Point Authors

Our ACN analysis is limited to the first author of the cited reference; this is due to the limitation of CiteSpace, which is our primary tool for analysis. The top 20 turning point authors are shown in Figure 4; Table 2 lists the authors.

<<Figure 4 about here>>

Table 2: Twenty turning point authors identified using Author Co-Citation Network.

Author	Freq	Author	Freq	Author	Freq	Author	Freq
Foster I	3305	Abramson D	248	Grimshaw AS	118	Taylor I	89
Buyya R	1422	Oinn T	184	Pearlman L	106	Altintas I	79
Deelman E	369	Maheswaran M	148	Kwok YK	104	Zhao J	78
Casanova H	314	Braun TD	142	Fahringer T	96	Lorch M	51
Yu J	249	Ludascher B	125	Topcuoglu H	96	Romberg M	36

Each author has made significant contributions to e-Science and e-Infrastructures (particularly Grid computing). *Foster* and *Buyya* are major focal points. *Foster* published many of the early papers that developed the concept of "The Grid" and its architectural components. *Buyya* made important contributions to the economics of Grid computing, the balance between Grid providers and consumers, and distributed resource management in grid computing (including Gridsim and NIMROD/G) and the "vision, hype, and reality" of Cloud computing. *Maheswaran* and *Braun* also published key articles in Grid resource management issues as well as heuristics for mapping tasks onto heterogeneous distributed computing systems, as did *Kwok* and *Topcuoglu* on different aspects of task scheduling on multiprocessors. *Casanova*

published on the development of ***NETSOLVE*** and also heuristics for scheduling parameter sweep applications in grid environments as a way of distributed work across a network of computers. *Grimshaw* contributed the ***LEGION*** Grid "operating system" and later a key article on *Open Grid Services Architecture*. *Fahringer* developed the ***ASKALON*** toolset for cluster and Grid computing. *Lorch* wrote on vital issues of Grid security and technology (such as PRIMA) as did *Pearlman*. Several authors made significant contributions to scientific workflow management: *Ludascher* with *Altintas* ***KEPLER***, *Deelman* ***PEGASUS***, *Taylor* and *Oinn* ***TRIANA,*** and *Yu* on grid discovery and workflow. *Romberg* produced key works relating to the ***UNICORE*** Grid system and widely to the e-Science and Bioinformatics. *Abramson* addressed a wide range of Grid issues and a particularly diverse range of end-user application areas. Contribution of *Zhao* is in relation to the semantic web and provenance, with two key references being [49,50].

5.3 Articles with Citation Burst

A citation burst signifies that a particular publication has received an extraordinary degree of attention from the scientific community, evidenced by the fact that the publication is associated with a surge in the number of citations [51]. The citation burst can last for many years or indeed only one year. Table 3 presents 17 papers from our underlying dataset which has experienced such citation bursts. The papers have a beginning and an end time which corresponds to the period of the citation burst (indicated in columns three and four respectively). The fifth column uses a timeline to depict the relative number of citations received during the aforementioned time periods using shades of grey (the darker the shade of grey, the more the number of citations that have been received). For example, Table 3 shows that the paper by Foster and Kesselman (1999) (paper 2) has experienced a burst in citation soon after its publication and it continued until 2006. Between the 2000-2006 periods, it received the most citations during the mid-point (depicted in black) and has since received a relatively lesser number of citations (shown in dark grey).

Table 3: Top references with citation burst.

Num	References	Begin	End	Begin - End
1	Armbrust M, Fox A, Griffith R, et al. (2010) A View of Cloud Computing. *Communications of the ACM* 53(4): 50-58.	2011	2014	
2	Foster I, Kesselman C (1999) *The Grid: Blueprint for a New Computing Infrastructure*. San Francisco, CA: Morgan Kaufmann.	2000	2006	
3	Foster I, Kesselman C (1997) Globus: a Metacomputing Infrastructure Toolkit. *International Journal of High Performance Computing Applications* 11(2):115-128	2000	2006	

4	Foster I, Kesselman C, Nick J, Tuecke S (2002) The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. *Open Grid Service Infrastructure WG, Global Grid Forum*, June 22, 2002.	2002	2006	
5	Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic, I (2009) Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25(6): 599-616.	2010	2014	
6	Armbrust M, Fox A, Griffith R, et al. (2009) Above the Clouds: A Berkeley View of Cloud Computing. *University of California at Berkeley, Electrical Engineering and Computer Sciences.* Technical Report No. UCB/EECS-2009-28.	2011	2014	
7	Foster I, Kesselman C, Tuecke S (2001) The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International Journal of High Performance Computing Applications* 15(3):200-22.	2003	2007	
8	Dean J, Ghemawat S (2008) MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, 51(1): 107-113.	2011	2014	
9	Vaquero L M, Rodero-Merino L, Caceres J, Lindner, M (2009) A break in the clouds: towards a cloud definition, *ACM SIGCOMM Computer Communication Review*, 39(1): 50-55.	2010	2012	
10	Zhang Q, Cheng L, Boutaba R (2010) Cloud computing: state-of-the-art and research challenges. *Journal of Internet Services and Applications*, 1(1): 7-18.	2012	2014	
11	Sotomayor B, Montero RS, Llorente IM, Foster I (2009) Virtual infrastructure management in private and hybrid clouds, *IEEE Internet Computing*, 13(5): 14-22.	2011	2014	
12	Mell P, Grance T (2011) The NIST definition of cloud computing, *National Institute of Standards and Technology Special Publication*, 800-145.	2012	2014	
13	Calheiros RN, Ranjan R, Beloglazov A, De Rose CAF, Buyya R (2011) CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms. *Software: Practice and Experience*, 41(1):23-50.	2012	2014	
14	Berman F, Fox G, Hey AJG (2003) *Grid Computing: Making the Global Infrastructure a Reality,* Chichester, England: John Wiley and Sons.	2004	2009	
15	Foster I, Zhao Y, Raicu I, Lu S (2008) Cloud Computing and Grid Computing 360-Degree Compared. In *Proceedings of the*	2010	2014	

	2008 Grid Computing Environments Workshop, 12-16 Nov 2008, Austin, TX. IEEE.			
16	Grimshaw AS, Wulf WA, the Legion Team (1997) The Legion vision of a worldwide virtual computer. *Communications of the ACM*, 40(1):39-45.	2001	2006	
17	Mell P, Grance T (2009) NIST definition of cloud computing. *National Institute of Standards and Technology*. October 7, 2009.	2011	2012	

The above list of papers represents a continuum of e-Infrastructure and distributed computing research that has spanned over a decade and shows the emergence of first Grid Computing and then Cloud Computing. Key works on ***GLOBUS*** (Foster and Kesselman 1997) (paper 3), "The Grid" (Foster and Kesselman 1999)(paper 2), Grid "Anatomy" (Foster, Kesselman and Tuecke 2001) (paper 7) and work on Grid service-based infrastructure (Foster, et al. 2002) (paper 4) had significant attention between 2000-2007 and shows their influence on the development of Grid Computing. Grimshaw and Wulf's work on ***LEGION***(paper 16) predated these texts. It could be argued that *LEGION* had a strong influence on Grid development and received similar attention over the 2001-2006 period with similar peak interest. Berman, Fox and Hey's edited 2003 book on Grid Computing (paper 14) took stock of previous Grid texts (including work by Foster, et al.) and included chapters on many contemporary issues and related Grid technologies and represented a key reference text at the time (2004-2009) (Berman, Fox and Hey 2003). What follows are articles that review and attempt to conceptualize the future of the rapidly emerging area of Cloud Computing. The majority of these papers were published in 2009 with significant attention emerging almost immediately. The analysis shows that this attention continues to increase into 2014 and presumably beyond. Foster, et al. (2008) (paper 15) make the link between Grid and Cloud Computing. Buyya, et al. (2009) (paper 5), Vaquero, et al. (2009) (paper 9), Armbrust, et al. (2009) (paper 6), Armbrust, et al. (2010) (paper 1), Zhang Q, Cheng L, Boutaba R (2010) (paper 10) are all influential articles on the state-of-the-art of Cloud Computing and emerging trends. Mell and Grance (2009;2011) (papers 17 and 12) published the NIST definition of Cloud Computing which has helped to provide a standard set of definitions for the area (and has recently been superseded by a larger roadmap standard from NIST). Calheiros, et al. (2011) (paper 13) built on previous work on the modelling and simulation of Grid architectures to develop *CLOUDSIM*, a toolkit to simulate Cloud Computing environment and approaches to resource provision and management. Sotomayor, et al. (2009) (paper 11) make a key contribution to virtual infrastructure management in different cloud provision models. Dean and Ghemawat (2008) (paper 8) describe *MAPREDUCE*, a cluster-based approach to data processing that has a significant link to e-Infrastructures (in particular Cloud Computing).

6. Cluster Analysis

The document co-citation visualization allows us to identify underlying relationships among the cited articles. For example, a thick link between two nodes (denoting high co-citation count among the articles), both of which also have a relatively large diameter (denoting high citation count) and have been consistently over the years would identify two papers that are equally important to a subject matter. But the question is, what is the subject matter? It is possible to infer this from reading the abstracts of the papers with high-citation count. However, this process is time-consuming, and the interpretation is subjective as it is based on a researcher's domain knowledge. A better way to achieve this is to automatically assign meaningful labels to the co-citation clusters that are identified in a co-citation network; CiteSpace "characterizes clusters with candidate labels selected by multiple ranking algorithms from the citers of these clusters and reveals the nature of a cluster in terms of how it has been cited" [33]. CiteSpace presently supports nine different ways of labelling the clusters – allowing selection of candidate terms from three sources (titles, abstracts, and index terms) all of which belong to the citing articles and three ranking algorithms [33]. In our study, we selected title terms and have used the *tf*idf* weighting algorithm[52] for labelling of clusters. We also compared these labels (TFIDF) with those generated by the *log-likelihood ratio (LLR)* algorithm and the *mutual information (MI)* algorithm. According to Gong et al. [53], a combined approach of using LLR and MI is useful since cluster labelling by LLR presents the core concepts of each cluster and includes professional words, whilst the latter provides some common words. CiteSpace has identified a total of ten citation clusters (note that cluster numbering starts from 0); clusters are labelled according to title terms occurring in citing articles (Figure 5). Table 4 lists the six largest clusters and includes cluster ID, cluster labels using TFIDF/LLR/MI (note: TFIDF labels used as cluster names in Figure 5), silhouette value, the number of articles in a cluster and mean cite year. Silhouette is a number between 0 and 1 and is used to evaluate and validate a cluster; the higher the value, the more confident we are of the resulting cluster [53].

Table 4: List of six largest clusters.

ID	*Size*	*Silhouette*	*Mean* **(Year)**	*Label (TFIDF)* **Cluster Name**	*Label (LLR)* **Core Concepts**	*Label (MI)* **Common Usage**
0	36	0.756	2002	systems biology | management	grid computing; cloud computing; job scheduling;	application programming support
1	32	0.894	2009	cloud selection | information technology	cloud computing; grid computing; federated cloud;	outsourcing providerselection
2	30	0.81	2001	grid economy | economicmodel	cloud computing; grid-enabled problem; adaptive computing;	grid task scheduling
3	30	0.738	1999	systems biology	cloud computing; enabling scalable virtual organization; drug design;	task segmentation
4	29	0.903	2008	inter enterprise	cloud computing; data center network; grid computing;	aware scheduling
5	27	0.642	2001	nuclear fusion devices | grid storage system	cloud computing; legion software environment; grid computing;	national model

<<Figure 5 about here>>

6.1 Discussion on papers associated with the identified clusters

Tables 5-10 list some of the seminal papers that are associated with each of the six clusters (numbered from 0 through to 6; see Table 4 above). Every table includes the cluster number, paper number, the reference and the citation count (from Google Scholar); each table lists the most prominent papers from each cluster. A short discussion section is included after every table. Note that, in the discussion section, the articles are identified in accordance with column 1 (Paper Number) in Table 5:

Table 5: List of papers identified in Cluster 1

Cluster Number	Paper Number	References for Cluster 1	Citations (Nov'18)
0	C0P1	Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107-113	~25,000
0	C0P2	White, T. (2012). *Hadoop: The definitive guide*. O'Reilly Media, Inc.	~5600
0	C0P3	Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung. (2003). The Google file system.In Proceedings: *SOSP '03 Proceedings of the nineteenth ACM symposium on Operating systems principles*, 29-43. ACM	~8000
0	C0P4	Kamara, S., & Lauter, K. (2010, January). Cryptographic cloud storage. In *International Conference on Financial Cryptography and Data Security* (pp. 136-149). Springer, Berlin, Heidelberg.	~1472
0	C0P5	Wang, C., Ren, K., Lou, W., & Li, J. (2010). Toward publicly auditable secure cloud data storage services. *IEEE Network*, 24(4):19-24.	~500
0	C0P6	Kalé, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., ... & Schulten, K. (1999). NAMD2: greater scalability for parallel molecular dynamics. *Journal of Computational Physics*, 151(1), 283-312.	~2417

Cluster 0 is associated with the core concepts of grid and cloud computing (LLR label; Table 4) and application programming support (MI label). It, therefore, comes as no surprise that the seminal paper on *MapReduce* by Dean and Ghemawat **[C0P1]** is identified in this cluster. MapReduce, which is a framework and associated implementation for parallelizing computing tasks across large datasets in a distributed computing environment, originated as a proprietary product in Google (the authors are from the same company). This paper has been cited ~25,000 times (November 2018). In terms of e-Science research, one prominent use of the MapReduce has been in the analysis of DNA sequencing data [54]. MapReduce generally relies on having access to large datasets, for example, through the *Hadoop Distributed File System (HDFS)* which stores data on commodity machines. Tom White's book 'Hadoop: The Definitive Guide' (published by is O'Reilly) **[C0P2]**, has been identified as yet another seminal contribution in this cluster. A second paper on filesystems is the one on the *Google File System (GFS)*, authored by Sanjay et al. **[C0P3]**.

Like HDFS, GFS is a scalable distributed file system. Papers **[C0P4]** and **[C0P5]** are on public cloud computing storage and focus on building a secure storage service on public cloud infrastructure **[C0P4]** and implementing an auditable secure cloud storage service **[C0P5]** respectively. Note that Cluster 0 is labelled as 'Systems Biology | Management' (using TFIDF labelling). Most of the papers discussed earlier were to do with the 'Management' of processing (MapReduce), filesystems (HDFS and GFS) and management of cloud services (cryptography and audit trail). There is also one paper that is on 'Systems Biology'. The article presents the NAMD2 program, which is used to simulate the behaviour of biomolecular systems **[C0P6]**. NAMD2 is an e-Science application and can be executed on most parallel machines, including workstation clusters.

Cluster 1 is labelled as 'Cloud Selection | Information Technology', with the core concepts of cloud/grid computing and provider-selection/outsourcing as the common terms (see Table 4). The defining characteristic of this cluster is that the papers (Table 5) are mostly concerned with defining the field of cloud computing, the shift from the conventional PC- based computing to the Clouds and what it would mean to the service providers, developers and users of this technology. Table 6 **[C1P1]** discusses the top 10 obstacles and opportunities for the growth of cloud computing. Some of the challenges include the problem of scalable storage, bottlenecks that may be associated with data transfer, data lock-in, service unpredictability and the problem with software licensing. For each obstacle, the paper identifies an opportunity, for example, data transfer bottlenecks can be eliminated with higher bandwidth switches, and a solution to software licensing issues can be pay-for-use licenses. Indeed, *Software as a Service (SAAS)* can be either subscription-based or usage-based. **[C1P2]** was published in 2008 and, as the author mentions, the future of cloud computing at that time was 'less than clear'. The author, therefore, suggests likely directions based on present practices. Zhange et al. **[C1P3]** present a scholarly review of cloud computing, which includes an overview of cloud computing architecture (IaaS, PaaS, SaaS), business models and underlying distributed computing technologies and commercial services. Vaquero et al.'s study of the definitions of cloud computing **[C1P5]**, whereby more than 20 definitions were considered and which allowed for the extraction of a consensus definition as well as a minimum definition consisting of the essential Cloud characteristics, and *The NIST definition of cloud computing* **[C1P4]** are also identified in this cluster.

Table 6: List of papers identified in Cluster 1

Cluster Number	Paper Number	References for Cluster 1	Citations (Nov'18)
1	C1P1	Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50-58.	~10,000

1	C1P2	Hayes, B. (2008). Cloud computing. *Communications of the ACM*, 51(7), 9-11.	~1548
1	C1P3	Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: state-of-the-art and research challenges. *Journal of internet services and applications*, 1(1), 7-18.	~3300
1	C1P4	Mell, P., & Grance, T. (2011). The NIST definition of cloud computing. *National Institute of Standards and Technology (NIST) Special Publication 800-145*, NIST, Gaithersburg, MD.	~14000 (two sources)
1	C1P5	Vaquero, L. M., Rodero-Merino, L., Caceres, J., & Lindner, M. (2008). A break in the clouds: towards a cloud definition. *ACM SIGCOMM Computer Communication Review*, 39(1), 50-55.	~3800

Cluster 2 is our third cluster. It is labelled as 'Grid Economy | Economic Models', and since these are related to cloud/grid computing, the core concepts remain similar to the other clusters. Grid task scheduling appears as the most common term (using MI labelling). The paper by Foster et al. (2002) **[C2P1]** – see Table 7 - has the highest number of citations in this cluster. Here the authors discuss the Open Grid Services Architecture (OGSA) which serves as a fundamental building block for the grid economy by providing standards for the integration of grid services executing on distributed and heterogeneous resources. **[C2P5]** also relates to OGSA; here, the authors extend the OGSA specifications to include Quality of Service (QoS) provisions and which enables service selection based on QoS assurances. The paper by Buyya et al. (2002) [**C2P2**] is a seminal paper on economic models for grid computing. It proposes a computational grid economy framework for resource allocation and outlines mechanisms that allow for optimizing supply and demand for grid resources. An implementation of the framework is provided through the *Nimrod-G* grid resource broker. **[C2P3]** has relatively less number of citations (~102) but has been identified as an important paper in this cluster. Like [**C2P2**], the article is on the management of resources in a grid computing environment and includes experimental results. The authors introduce an Agent-based Resource Management System (ARMS), which uses performance data from PACE toolkit [55] with a scheduling algorithm that is designed to manage local grid resources. Unlike the previous cluster on 'Cloud Selection | Information Technology', most of the papers here have a strong implementation focus. For example, [**C2P4**] presents the implementation of a new system (*NetSolve*) which allows users to remotely access computational resources that are distributed over a network; the paper on *UNICORE* project **[C2P6]** discusses the implementation of a prototype for uniform access to remote computers using the mechanisms of the World Wide Web and which allows for creation, submission and monitoring of jobs across distributed resources; the work by Livny and Raman (1999) **[C2P7]** on resource management in high-throughput computing systems like *CONDOR* [56].

Table 7: List of papers identified in Cluster 2

Cluster Number	Paper Number	References for Cluster 2	Citations (Nov'18)
2	C2P1	Foster, I., Kesselman, C., Nick, J. M., & Tuecke, S. (2002). Grid services for distributed system integration. *Computer*, 35(6), 37-46.	~2053
2	C2P2	Buyya, R., Abramson, D., Giddy, J., & Stockinger, H. (2002). Economic models for resource management and scheduling in grid computing. *Concurrency and Computation: Practice and Experience*, 14, 1507-1542.	~1100
2	C2P3	Cao, J., Jarvis, S. A., Saini, S., Kerbyson, D. J., & Nudd, G. R. (2002). ARMS: An agent-based resource management system for grid computing. *Scientific Programming*, 10(2), 135-148.	~102
2	C2P4	Casanova, H., & Dongarra, J. (1997). Netsolve: a Network-Enabled Server for Solving Computational Science Problems. *The International Journal of High Performance Computing Applications*, 11(3):212-223.	~900
2	C2P5	Ali, R. J. A., Rana, O. F., Walker, D. W., Jha, S., & Sohail, S. (2012). G-QoSM: Grid service discovery using QoS properties. *Computing and Informatics*, 21(4), 363-382.	~241
2	C2P6	Almond, J., & Snelling, D. (1999). UNICORE: uniform access to supercomputing as an element of electronic commerce. *Future Generation Computer Systems*, 15(5-6), 539-548.	~134
2	C2P7	Livny, M., & Raman, R. (1999). High-throughput resource management. *The Grid: Blueprint for a New Computing Infrastructure*, Foster, I and Kesselman, C (Eds.), Morgan Kaufmann, 311-337.	~158

Cluster 3 is labelled 'Systems Biology' and includes six papers (Table 8). It comes as no surprise that the paper describing the *Taverna Workflow System (TWS)* is included in the list **[C3P1]**. TWS was initially used by bioinformatics scientists to execute complex computer simulations that needed coordinated use of computation and data repositories. The use of TWS has since expanded to other disciplines and is now available as an open-source and domain-independent workflow management system which could be used to design and execute scientific workflows (https://taverna.incubator.apache.org/, last accessed Dec 2018). The second paper is on the *KEPLER* scientific workflow system, which is a community-driven open source project that helps scientist to create, execute, share and analyse models across a wide range of disciplines (https://kepler-project.org, last accessed Dec 2018) **[C3P2]**. The third paper on scientific workflow **[C3P3]** describes the *Pegasus* framework for mapping scientific workflows onto distributed systems. The use of Pegasus in bioinformatics and biology is reported in [57], where it was used to execute BLAST (an important bioinformatics application) "which consists of a set of sequence comparison algorithms that are used to search sequence databases for optimal local alignments to a quest" [57] and a speedup of 5-20 times was achieved. The topic of the next paper **[C3P4]** is Next-generation DNA sequencing. It introduces a new

parallel read-mapping algorithm called *CloudBurst* for mapping data to the human genomes and other reference genomes. It uses the Hadoop implementation of MapReduce (Hadoop and MapReduce discussed in Cluster 0) for parallel execution. The paper by Simmhan et al. (2005) **[C3P5]** develops a taxonomy of data provenance characteristics and focusses mainly on scientific workflow approaches. It is identified in this cluster since bioinformatics and system biology applications frequently use scientific workflow systems for *in silico* experimentation. The final paper in this cluster is on DNA sequencing to the human genome **[C3P6]**. The authors report on an ultrafast, memory-efficient program called *Bowtie,* which is written in C++.

Table 8: List of papers identified in Cluster 3

Cluster Number	Paper Number	References for Cluster 3	Citations (Nov'18)
3	C3P1	Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., ... & Li, P. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17), 3045-3054.	~1700
3	C3P2	Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., and Zhao, Y. (2006). "Scientific workflow management and the Kepler system." *Concurrency and Computation: Practice and Experience* 18(10): 1039-1065.	~2000
3	C3P3	Deelman, E., Singh, G., Su, M. H., Blythe, J., Gil, Y., Kesselman, C., ... & Laity, A. (2005). Pegasus: A framework for mapping complex scientific workflows onto distributed systems. *Scientific Programming*, 13(3), 219-237.	~1300
3	C3P4	Schatz, M. C. (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics*, 25(11), 1363-1369.	~729
3	C3P5	Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *ACM Sigmod Record*, 34(3), 31-36.	~1200
3	C3P6	Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), R25.	~13000

Cluster 4 is labelled 'Inter-enterprise' and it is our fifth cluster. Several of the papers reported in this cluster (Table 9) have also been reported in Cluster 1 (Cloud Selection | Information Technology), for example **[C4P1]**, **[C4P3]** and **[C4P4]**. This is not surprising since cloud computing can have an inter-enterprise dimension (unlike private clouds which exist primarily within enterprise boundaries). Two papers by Rajkumar Buyya have been identified in this cluster **- [C4P2]** and **[C4P7]**. **[C4P2]** is a seminal paper in the field and (arguably) referred to cloud computing as a utility for the very first time. Motivated by the underlying market mechanisms for utilities like water, electricity and gas, the authors outline a proposed

architecture for market-oriented allocation of Cloud resources. Written a decade earlier, the vision and hype of cloud computing back in 2009 (when the paper first appeared) has turned into a reality! The paper also captures the authors' vision of a Cloud exchange for trading services; as such exchanges are likely to be global and inter-organizational, it comes as no surprise that this paper is identified in this cluster. **[C4P7]** is a conference keynote paper which has is a shorted version of the FGCS journal paper **[C4P2]**. The next article is also very relevant to the 'inter-organization' theme of the cluster as it is on private and hybrid clouds **[C4P3]**. It reports on the transition of the Cloud IaaS system from the standard provider-based service to the development of hybrid and private clouds using an organization's existing infrastructure. The paper reports on the *OpenNebula* open-source, virtual infrastructure manager that could be used to deploy virtualized within both the institutional resources and also external provider-based clouds.

Table 9: List of papers identified in Cluster 4

Cluster Number	Paper Number	References for Cluster 4	Citations (Nov'18)
4	C4P1	Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50-58.	~10100
4	C4P2	Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25(6), 599-616.	~6200
4	C4P3	Vaquero, L. M., Rodero-Merino, L., Caceres, J., & Lindner, M. (2008). A break in the clouds: towards a cloud definition. *ACM SIGCOMM Computer Communication Review*, 39(1), 50-55.	~3800
4	C4P4	Mell, P., & Grance, T. (2011). The NIST definition of cloud computing. *National Institute of Standards and Technology (NIST) Special Publication 800-145*, NIST, Gaithersburg, MD.	~14000 (two sources)
4	C4P5	Sotomayor, B., Montero, R. S., Llorente, I. M., & Foster, I. (2009). Virtual infrastructure management in private and hybrid clouds. *IEEE Internet Computing*, 13(5), 14-22.	~1200
4	C4P6	Nurmi, D., Wolski, R., Grzegorczyk, C., Obertelli, G., Soman, S., Youseff, L., & Zagorodnov, D. (2009). The eucalyptus open-source cloud-computing system. In Proceedings of the *9th IEEE/ACM International Symposium on Cluster Computing and the Grid* (CCGRID'09), pp. 124-131. IEEE.	~2300
4	C4P7	Buyya, R., Yeo, C. S., & Venugopal, S. (2008). Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities. In *Proceedings of the 10th International Conference on High Performance Computing and Communications (HPCC'08)*, pp. 5-13. IEEE.	~2700
4	C4P8	Keahey, K., Foster, I., Freeman, T., & Zhang, X. (2005). Virtual workspaces:	~367

		Achieving quality of service and quality of life in the grid. *Scientific Programming*, 13(4), 265-275.	

Cluster 5 is labelled as 'Nuclear Fusion Devices | Grid Storage System' and the papers identified are those listed in Table 10. This cluster includes seminal scholarly work from the pioneers of grid computing **[C5P1]** and **[C5P2]** (article and a book chapter, respectively). These papers include discussion on grid storage, for example **[C5P1]** makes reference to Storage Service Providers (SSPs), and the reference to fusion devices could have been in relation to the application of grid, for example, *FusionGrid* [58] which was funded by the US Department of Energy. **[C5P3]** is on a scalable and distributed monitoring system for cluster computers, grids and planetary-scale systems. The system is called *Ganglia* and is based on a hierarchical design that relies on multicast-based listen/announce protocol to monitor cluster states. **[C5P4]** reports on a grid-enabled computational framework called *Cactus,* which allows for dynamic resource discovery and migration of applications to better resources (upon performance degradation being experienced on resources currently executing the application). The primary selection process to discover available resources is based on CONDOR matchmaking algorithm. The next paper is on the *Legion* middleware project and outlines its vision of a worldwide computer that distributed computation using the World Wide Web **[C5P5]**. **[C5P6]** and **[C5P7]** are on standards. The *Web Services Description Language* **[C5P6]** *(WSDL)*, together with *Simple Object Access Protocol (SOAP)* and *Universal Description, Discovery and Integration (UDDI)* make possible ubiquitous access to grid services through Web services technologies. Grid services (defined by OGSA – refer to cluster 2) are a special type of web services. The Open Grid Services Infrastructure (OGSI) **[C5P7]**, developed by the Global Grid Forum (GGP), provides detailed technical specification on grid services (defined by OGSA) and how they work. The *Globus Toolkit version 3 (GT3)*, which is a toolkit to program grid-based applications, is an implementation of OGSI.

Table 10: List of papers identified in Cluster 5

Cluster Number	Paper Number	Reference	Citations (Nov'18)
5	C5P1	Foster, I., Kesselman, C., & Tuecke, S. (2001). The anatomy of the grid: Enabling scalable virtual organizations. *The International Journal of High Performance Computing Applications*, 15(3), 200-222.	~11800
5	C5P2	Foster, I., Kesselman, C., Nick, J. M., & Tuecke, S. (2003). The physiology of the grid. In *Grid Computing: Making the Global Infrastructure a Reality*, Berman, F., Fox, G.C., and Hey, A.J.G (Eds.). 217-249. Wiley.	~5000
5	C5P3	Massie, M. L., Chun, B. N., & Culler, D. E. (2004). The ganglia distributed monitoring system: design, implementation, and experience. *Parallel Computing*, 30(7), 817-840.	~1500

5	C5P4	Gabrielle, A., Angulo, D., Foster, I., Lanfermann, G., Liu, C., Radke, T., Seidel, E., & Shalf, J. (2001). The Cactus Worm: Experiments with dynamic resource discovery and allocation in a grid environment. *The International Journal of High Performance Computing Applications* 15(4): 345-358.	~200
5	C5P5	Grimshaw, A. S., & Wulf, W. A. (1997). The Legion vision of a worldwide virtual computer. *Communications of the ACM*, 40(1), 39-45.	~1000
5	C5P6	Christensen, E., Curbera, F., Meredith, G., & Weerawarana, S. (2001). *Web services description language (WSDL) 1.1*.W3C.	~3400
5	C5P7	Tuecke, S., Czajkowski, K., Foster, I., Frey, J., Graham, S., Kesselman, C., & Snelling, D. (2009). *Open grid services infrastructure (OGSI)*. Global Grid Forum.	~700

In reference to the papers in cluster 5, it is possible to critique the TFIDF labelling for the cluster as 'Nuclear Fusion Devices | Grid Storage System' (refer to Table 4). However, the LLR labelling for cluster 5 as "Cloud Computing | Legion Software Environment | Grid Computing" seems to be more appropriate considering that the papers in this cluster outlined fundamental grid concepts (anatomy of the grid, physiology of the grid), standards (OGSA, WSDL, OGSI) and grid middleware (Ganglia, Legion, Cactus, GT3). This goes to show that automated cluster identification using scientometric techniques should be accompanied by a reading of the relevant papers. This would enable a better understanding of the field of study.

7. Conclusion

In this paper, we have presented an exploration of the e-Science and the e-Infrastructure knowledge bases using co-citation analysis. We conducted a search on the ISI Web of Science by using a combination of relevant keywords (e.g., grid computing, cloud computing, e-Science, cyberinfrastructures) that appear in an article's title, abstract, author keywords and ISI keywords plus. We used CiteSpace to analyse a dataset of over 12,500 records and identified the top 20 turning point papers and authors, papers with citation bursts and important research clusters. These highlight key works in grid computing, grid-based technologies and cloud technologies that reflect fast exploitation of emerging distributed computing tools and techniques. However, with the advancement of Grid-Desktop, Grid-Cloud interoperability technologies, e.g., [9–11], we also see that novel innovations in Grid to include Cloud and Desktop Grid-based execution of e-Science and e-Infrastructure applications.

Cluster analysis has been an important element of this paper. We report on the papers identified by the six largest clusters (Clusters 0 to 5) and discuss them in the context of the theme of the cluster. The analysis shows the transition of efforts from Grid to cloud to exploit innovations of cloud computing in distributed computing literature (this is recent literature with the mean year of 2009 (cluster 0; Table 4)). It is interesting

to compare this with the mean year of 2001 reported for literature associated with cluster 6 that shows the historical shift of focus from Grid to Cloud (cluster 6 is associated with fundamental concepts of grid computing).

There has been a huge investment of effort in e-Infrastructure and e-Science research. These are complex technological areas that can be difficult to navigate. Our survey and analysis gives an entry point to this area by showing a range of influential works that, for example, show the impact of standards such as OGSA and OGSI and the role of key technologies (e.g. MapReduce and the Google File System), Grid and HTC middleware (e.g. GT3 and CONDOR) and scientific workflows. The survey also shows key application areas in which e-Infrastructures and e-Science have been particularly successful. With continuous innovation in these areas it is important to understand the past in order to avoid repetition and to reflect on the impact of successful research. It is our hope that this analysis of this literature gives a starting point for new researchers in the field, as well as some opportunity for reflection for those only familiar part of this research.

Appendix A – Search Strategy for Compiling the Database of Articles

The search was conducted on the *ISI Core Collection* databases specific to both journals and conferences, e.g., *Science Citation Index Expanded (SCI-EXPANDED) and Conference Proceedings Citation Index-Science (CPCI-S)*. The timespan selected was "All Years". The specific search criterion that was used was as follows: *Topic= (e-Science) OR Topic= (e-Infrastructure) OR Topic= ("cloud computing") OR Topic= ("cyberinfrastructure") OR Topic= ("grid computing") OR Topic= ("desktop grid computing"); Databases= (SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH); Timespan = (All Years)*. The search retrieved a total of **12867**unique articles, of which we only considered journal articles (**4731**), conference proceedings (**8750**), and review papers (**98**); we ignored editorial material, book reviews and news items. The total number of unique articles included in the review was **12516**. It is to be noted that some articles may appear under multiple categories, for example, a review article may also be categorized as a journal paper if it was published in a scholarly source, and as such, the number of unique articles is less than the sum total of the individual categories.

References

[1] N. Mustafee, N. Bessis, S.J.E. Taylor, and S. Sotiriadis. (2013). "Exploring the e-science knowledge base through co-citation analysis," ANT (Ambient Systems, Networks and Technologies) 2013, June 25-28, 2013, Halifax, Nova Scotia, Canada. Procedia Computer Science, 19 (2013), pp.586 – 593.

[2] N. Bessis, Grid technology for maximizing collaborative decision management and support: advancing effective virtual organizations, Hershey, New York: Information Science Reference, 2009.

[3] I. Bird, B. Jones, K.F. Kee, The Organization and Management of Grid Infrastructures, Computer (Long. Beach. Calif). 42 (2009) 36–46. doi:10.1109/MC.2009.28.

[4] M. Baker, R. Buyya, D. Laforenza, Grids and Grid technologies for wide-area distributed computing, Softw. Pract. Exp. 32 (2002) 1437–1466. doi:10.1002/spe.488.

[5] T. Hey, A.E. Trefethen, The UK e-Science Core Programme and the Grid, *Futur. Gener. Comput. Syst. 18* (2002) 1017–1031. doi:10.1016/S0167-739X(02)00082-1.

[6] LSGC, Life Science Grid Community, (n.d.). *http://lsgc.org* (November, 2018).

[7] V. Breton, N. Jacq, V. Kasam, M. Hofmann-Apitius, Grid-added value to address malaria., IEEE *Trans. Inf. Technol. Biomed*. 12 (2008) 173–81. doi:10.1109/TITB.2007.895930.

[8] A. Naseer, L.K. Stergioulas, Web-services-based resource discovery model and service deployment on HealthGrids., *IEEE Trans. Inf. Technol. Biomed.* 14 (2010) 838–45. doi:10.1109/TITB.2010.2040482.

[9] E. Urbah, P. Kacsuk, Z. Farkas, G. Fedak, G. Kecskemeti, O. Lodygensky, et al., EDGeS: Bridging EGEE to BOINC and XtremWeb, *J. Grid Comput*. 7 (2009) 335–354. doi:10.1007/s10723-009-9137-0.

[10] A. Anjum, R. Hill, R. McClatchey, N. Bessis, A. Branson, Glueing grids and clouds together: a service-oriented approach, *Int. J. Web Grid Serv*. 8 (2012) 248–265.

[11] Y. Huang, N. Bessis, P. Norrington, P. Kuonen, B. Hirsbrunner, Exploring decentralized dynamic scheduling for grids and clouds using the community-aware scheduling algorithm, *Futur. Gener. Comput. Syst.* 29 (2013) 402–415. doi:10.1016/j.future.2011.05.006.

[12] C. Chen, Searching for intellectual turning points: progressive knowledge domain visualization., Proc. Natl. Acad. Sci. U. S. A. 101 Suppl (2004) 5303–10. doi:10.1073/pnas.0307513100.

[13] I. Foster, C. Kesselman, Computational grids, in: I. Foster, C. Kesselman (Eds.), Grid Bluepr. a Futur. Comput. Infrastruct., San Francisco, CA: Morgan Kaufmann, 1998.

[14] E. Asimakopoulou, N. Bessis, Advanced ICTs for Disaster Management and Threat Detection: Collaborative and Distributed Frameworks, Hershey, New York: Information Science Reference, 2010.

[15] S. Choi, M. Baik, C. Hwang, J. Gil, H. Yu, Volunteer availability based fault tolerant scheduling mechanism in desktop grid computing environment, in: Third IEEE Int. Symp. Netw. Comput. Appl. 2004. (NCA 2004). Proceedings., IEEE, 2004: pp. 366–371. doi:10.1109/NCA.2004.1347802.

[16] M.W. Mutka, Estimating capacity for sharing in a privately owned workstation environment, IEEE Trans. Softw. Eng. 18 (1992) 319–328. doi:10.1109/32.129220.

[17] M.W. Mutka, M. Ivny, The available capacity of a privately owned workstation environment, Perform. Eval. 12 (1991) 269–284. doi:10.1016/0166-5316(91)90005-N.

[18] H. Casanova, Distributed computing research issues in grid computing, ACM SIGACT News. 33 (2002) 50. doi:10.1145/582475.582486.

[19] P. Kacsuk, J. Kovacs, Z. Farkas, A.C. Marosi, G. Gombas, Z. Balaton, SZTAKI Desktop Grid (SZDG): A Flexible and Scalable Desktop Grid System, J. Grid Comput. 7 (2009) 439–461. doi:10.1007/s10723-009-9139-y.

[20] S.J.E. Taylor, M. Ghorbani, N. Mustafee, S.J. Turner, T. Kiss, D. Farkas, et al., Distributed computing and modeling & simulation: Speeding up simulations and creating large models, in: Proc. 2011 Winter Simul. Conf., IEEE, Phoenix, Arizona, 2011: pp. 161–175. doi:10.1109/WSC.2011.6147748.

[21] R. Buyya, C.S. Yeo, S. Venugopal, Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities, in: 2008 10th IEEE Int. Conf. High Perform. Comput. Commun., IEEE, 2008: pp. 5–13. doi:10.1109/HPCC.2008.172.

[22] P. Watson, P. Lord, F. Gibson, P. Periorellis, G. Pitsilis, Cloud Computing for e-Science with CARMEN, in: Proc. 2nd Iber. Grid Infrastruct. Conf. Proc., Porto, 2008: pp. 3–14. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.154.5773 (accessed November 8, 2017).

[23] M.D. White, E.E. Marsh, Content Analysis: A Flexible Methodology, Libr. Trends. 55 (2006) 22–45. https://www.ideals.illinois.edu/handle/2142/3670 (accessed Septemeber 15, 2017).

[24] P. Palvia, P. Pinjani, E.H. Sibley, A profile of information systems research published in Information & Management, Inf. Manag. 44 (2007) 1–11. doi:10.1016/j.im.2006.10.002.

[25] E. Claver, R. González, J. Llopis, An analysis of research in information systems (1981–1997), Inf. Manag. 37 (2000) 181–195. doi:10.1016/S0378-7206(99)00043-9.

[26] L. Heilig, S. Voss, A Scientometric Analysis of Cloud Computing Literature, IEEE Trans. Cloud Comput. accepted ((2014). doi:10.1109/TCC.2014.2321168.

[27] N. Mustafee, K. Katsaliaki, S.J.E. Taylor, Profiling Literature in Healthcare Simulation, Simulation. 86 (2010) 543–558. doi:10.1177/0037549709359090.

[28] R.D. Galliers, E.A. Whitley, R.J. Paul, The European Information Systems Academy, Eur. J. Inf. Syst. 16 (2007) 3–4. doi:10.1057/palgrave.ejis.3000669.

[29] S. Raghuram, P. Tuertscher, R. Garud, Research Note —Mapping the Field of Virtual Work: A Cocitation Analysis, Inf. Syst. Res. 21 (2010) 983–999. doi:10.1287/isre.1080.0227.

[30] H.D. White, K.W. McCain, Visualizing a discipline: An author co-citation analysis of information science, 1972–1995, J. Am. Soc. Inf. Sci. 49 (1998) 327–355. doi:10.1002/(SICI)1097-4571(19980401)49:4<327::AID-ASI4>3.0.CO;2-4.

[31] Astrom,,F. (2007). Changes in the LIS research front: Time-sliced co-citation analyses of LIS journal articles, 1990-2004. *Journal of the American Society for Information Science and Technology*, 58(7): 947-957.

[32] Chang, Y.W, Huang, M.H., Lin, C.W. (2015). Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses. *Scientometrics*, 105(3):2071-2087.

[33] Chen, C.M., Ibekwe-SanJuan, F., Hou, J.H. (2010). The structure and dynamics of cocitation clusters: A multiple-perspective co-citation analysis. *Journal of the American Society for Information Science and Technology*, 61(7):1386-1409.

[34] Yang, S., Han, R., Wolfram, D., & Zhao, Y. (2016). Visualizing the intellectual structure of information science (2006–2015): Introducing author keyword coupling analysis. *Journal of Informetrics*, 10(1):132-150.

[35] Zhao, D.Z., Strotmann, A. (2008). Evolution of Research Activities and Intellectual Influences in Information Science 1996-2005: Introducing Author Bibliographic-Coupling Analysis, *Journal of the American Society for Information Science and Technology*,59(13): 2070-2086.

[36] Zhao, D.Z., Strotmann, A. (2014). The Knowledge Base and Research Front of Information Science 2006-2010: An Author Cocitation and Bibliographic Coupling Analysis. *Journal of the Association for Information Science and Technology*, 65(5): 995-1006.

[37] M.J. Culnan. (1986). The Intellectual Development of Management Information Systems, 1972–1982: A Co-Citation Analysis, *Management Science*, 32(2):156–172.

[38] N. Mustafee. (2011). Evolution of IS research based on literature published in two leading IS journals - EJIS and MISQ. In Proc. *19th European Conference on Informatin Systems,* Association for Information Systems, Helsinki, Finland, 2011: p. 228.

[39] A. Pilkington, J. Meredith. (2009). The evolution of the intellectual structure of operations management—1980–2006: A citation/co-citation analysis, *Journal of Operations Management*, 27(3):185–202. doi:10.1016/j.jom.2008.08.001.

[40] S.P. Nerur, A.A. Rasheed, V. Natarajan. (2007)The intellectual structure of the strategic management field: an author co-citation analysis, *Strategic Management Journal*, 29 (3): 319–336. doi:10.1002/smj.659.

[41] F.J. Acedo, J.C. Casillas. (2005). Current paradigms in the international management field: An author co-citation analysis, *International Business Review*, 14 (5): 619–639. doi:10.1016/j.ibusrev.2005.05.003.

[42] M. Kas, K.M. Carley, L.R. Carley. (2012). Trends in science networks: understanding structures and statistics of scientific networks, *Social Network Anaysis and Mining*, 2(2):169–187. doi:10.1007/s13278-011-0044-6.

[43] M. Niazi, A. Hussain, (2011). Agent-based computing from multi-agent systems to agent-based models: a visual survey, *Scientometrics*, 89 (2011) 479–499. doi:10.1007/s11192-011-0468-9.

[44] R. Zhao, J. Wang. (2011). Visualizing the research on pervasive and ubiquitous computing, *Scientometrics*, 86 (2011) 593–612. doi:10.1007/s11192-010-0283-8.

[45] G. Liu. (2013). Visualization of patents and papers in terahertz technology: a comparative study, *Scientometrics*. 94 (2013) 1037–1056. doi:10.1007/s11192-012-0782-x.

[46] N. Mustafee, K. Katsaliaki, P. Fishwick (2014). Exploring the modelling and simulation knowledge base through journal co-citation analysis, *Scientometrics* 98 (2014) 2145–2159. doi:10.1007/s11192-013-1136-z.

[47] C. Chen, S. Morris, Visualizing evolving networks: minimum spanning trees versus pathfinder networks, in: Proc. *Ninth Annu. IEEE Conf. Inf. Vis., IEEE Computer Society*, Seattle, Washington, 2003: pp. 67–74. http://dl.acm.org/citation.cfm?id=1947368.1947384.

[48] C. Chen. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature, *J. Am. Soc. Inf. Sci. Technol*. 57 (2006) 359–377. doi:10.1002/asi.20317.

[49] J. Zhao, C. Goble, R. Stevens, D. Turi, Mining Taverna's semantic web of provenance, Concurr. Comput. Pract. Exp. 20 (2008) 463–472. doi:10.1002/cpe.v20:5.

[50] J. Zhao, S.S. Sahoo, P. Missier, A. Sheth, C. Goble, Extending Semantic Provenance into the Web of Data, IEEE Internet Comput. 15 (2011) 40–48. doi:10.1109/MIC.2011.7.

[51] C. Chen, The CiteSpace Manual - version 0.65 (last updated April 12, 2014), (2014). *http://cluster.ischool.drexel.edu/~cchen/citespace* (accessed November, 2018).

[52] G. Salton, A. Wong, C.S. Yang, (1975). A vector space model for automatic indexing, Commun. ACM. 18 (1975) 613–620. doi:10.1145/361219.361220.

[53] Gong, X., Jiang, L., Yang, H., & Wei, F. (2011). Mapping Intellectual Structure: A Co-citation Analysis of Food Safety in CiteSpace II. Gene, 412. http://cluster.ischool.drexel.edu/~cchen/courses/INFO633/12-13/gong.pdf (accessed November 2018).

[54] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297-303.

[55] Cao, J., Kerbyson, D.J., Papaefstathiou, E. and Nudd, G.R. (2000). Performance modeling of parallel and distributed computing using PACE. In *Proceedings of the 19th IEEE International Performance, Computing and Communication Conference*, Phoenix, USA, pp. 485–492.

[56] Litzkow, M. J., Livny, M., and Mutka, M. W. (1988). Condor—A hunter of idle workstations. In *Proceedings of the 1988 Conference on Distributed Computing Systems*, pp. 104–111. IEEE: New York.

[57] Deelman, E., Blythe, J., Gil, Y., Kesselman, C., Mehta, G., Patil, S., Su, M.H., Vahi, K. and Livny, M. (2004). Pegasus: Mapping scientific workflows onto the grid. In Grid Computing (pp. 11-20). Springer, Berlin, Heidelberg.

[58] Schissel, D. P., J. R. Burruss, A. Finkelstein, S. M. Flanagan, I. T. Foster, T. W. Fredian, M. J. Greenwald et al. (2004). Building the us national fusion grid: Results from the national fusion collaboratory project. *Fusion Engineering and Design*, 71(1-4): 245-250.

Figures

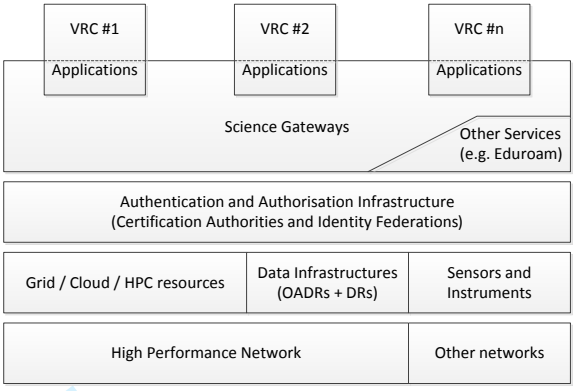

Figure 1: A Typical e-Infrastructure Architecture (*eI4Africa.eu*).

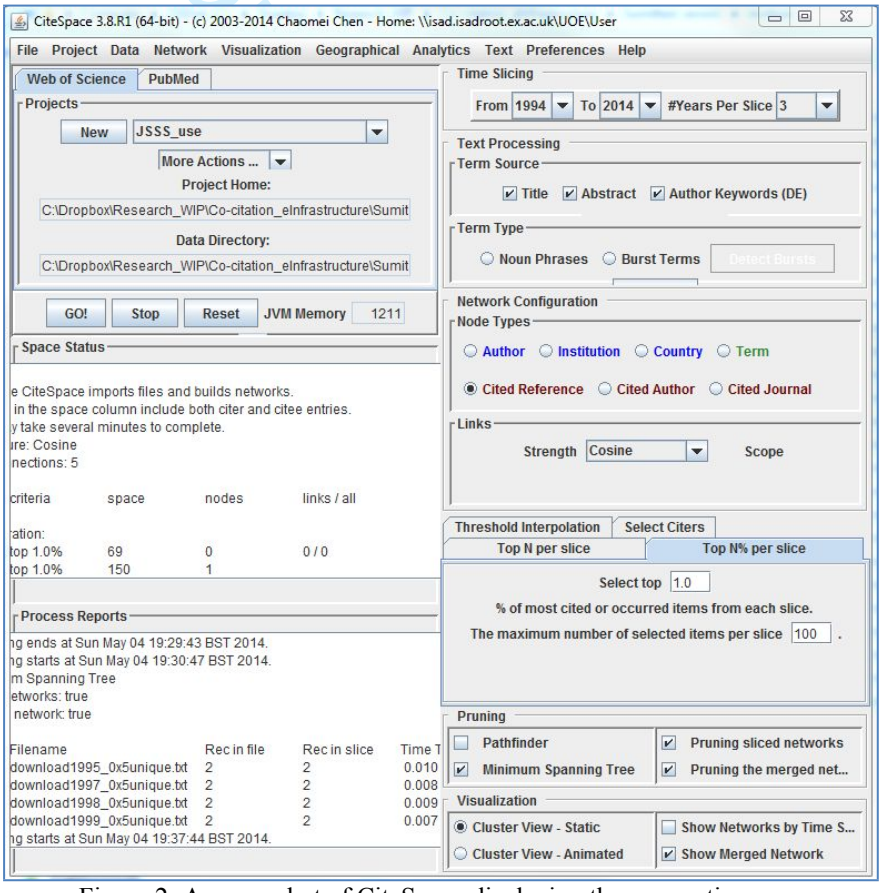

Figure 2: A screenshot of CiteSpace displaying the user options.

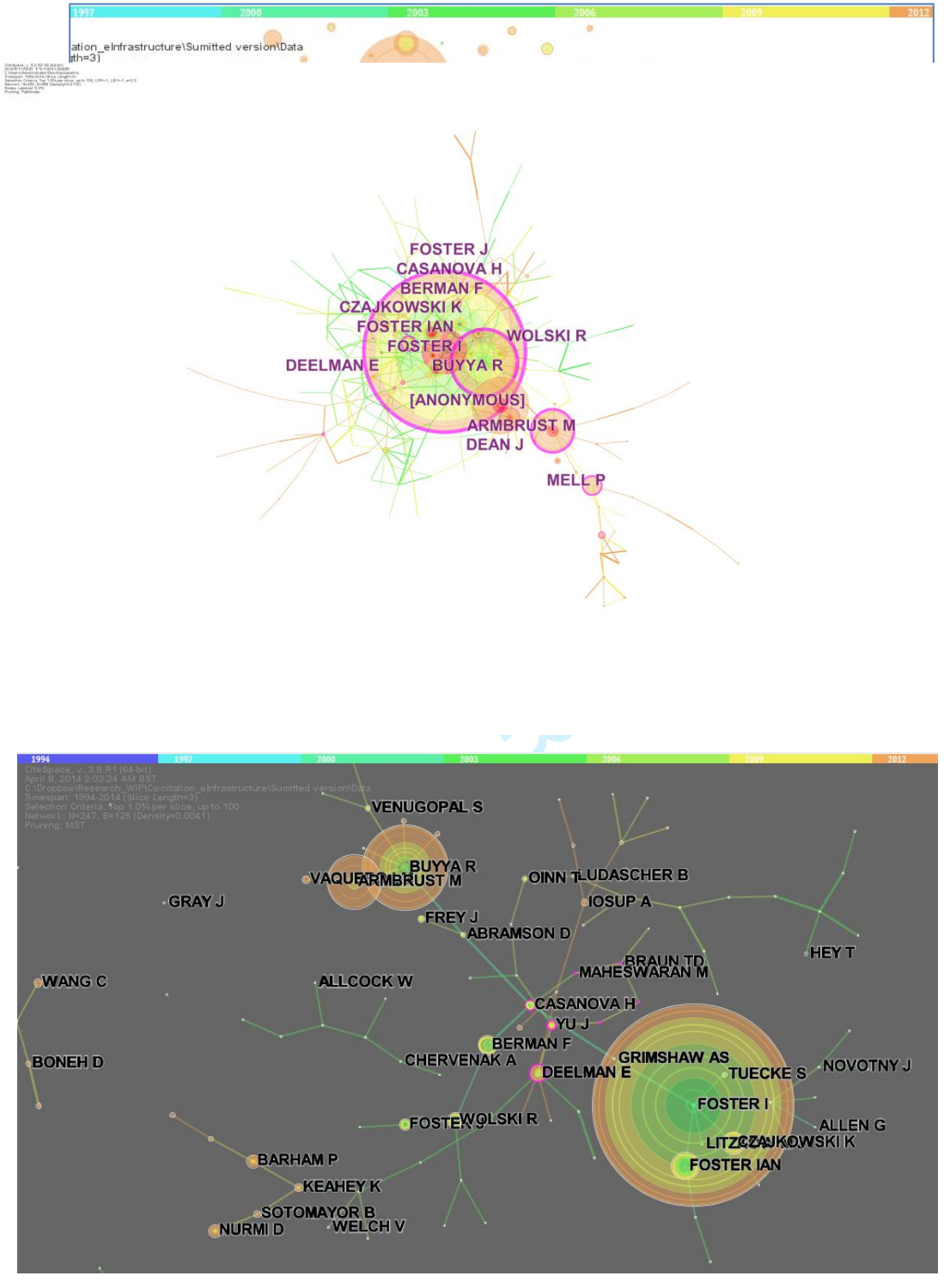

Figure 4: Author Co-Citation Network (ACN) generated using ISI Web of Science data and CiteSpace.

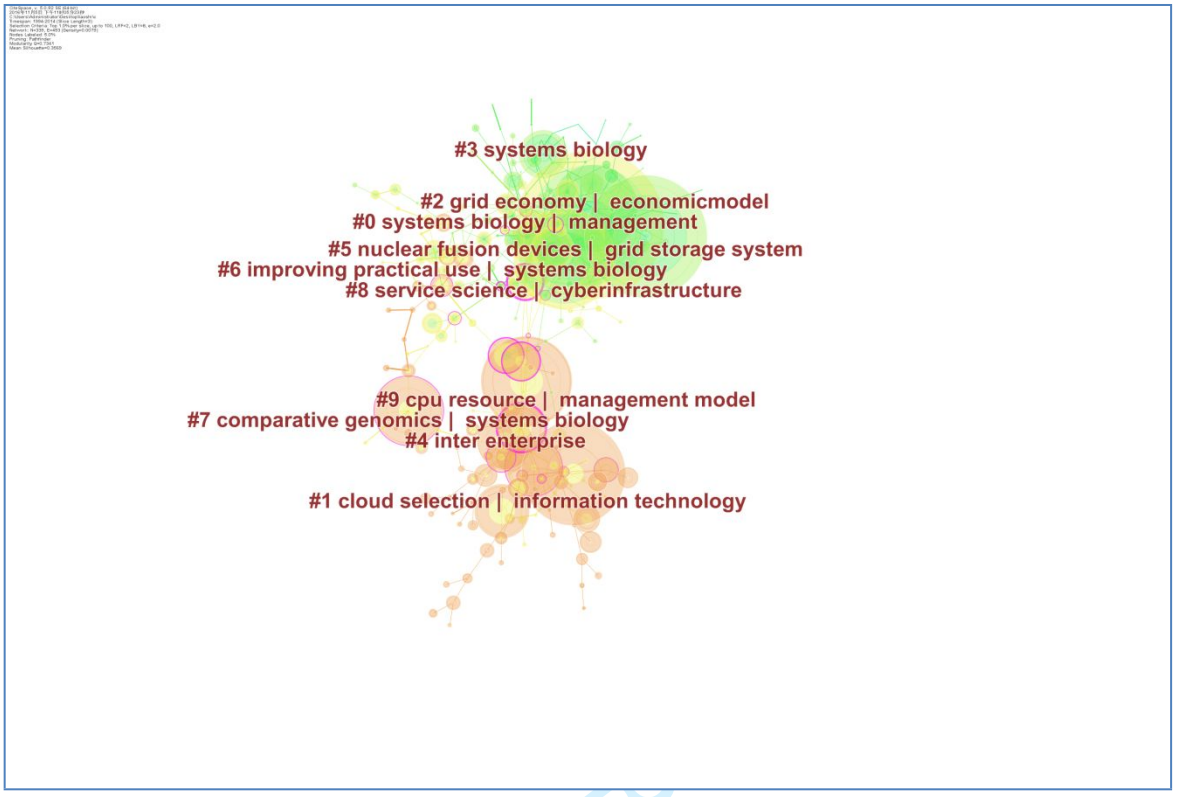

Figure 5: Clusters identified in the DCN solution space and named using candidate labels.