What characterizes academic research about cloud computing? – a research profiling approach

Contribution to workshop on Cloud Computing,
IFI Blindern 22. October 2012

Copyright © 1997-2008 Search Technology, Inc. All rights reserved. This program is protected by U.S. and international copyright laws as described in Help About.

Jarle Moss Hildrum & Ben Eaton
Center for Technology, Innovation and Culture, University of Oslo
Institute for Informatics, University of Oslo

‘*Since its emergence around 2007 the topic (cloud computing) has exploded in interest within academic and technical literatures (…) It is difficult to fully make sense of this diverse set of publications.’*

(Venter and Whitley 2012, p. 2)

Objective of the presentation

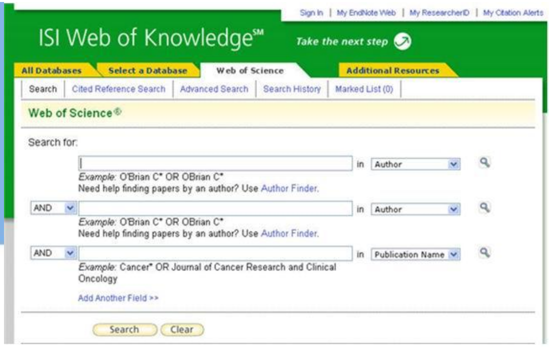
– Present a large-scale **research profile** of cloud computing published on the ISI – web of science between 2006 and 2011

– Examine the degree to which current concerns of corporate users of cloud computing (as presented by Venters and Whitley 2012) are also becoming the concerns of academic researchers

Traditional Literature Review	Research Profiling
Micro focus (paper-by-paper)	Macro focus (patterns in the literature as a body) using search engines (ISI, Scopus) and text-mining software
Narrow range (~20 references)	Wide range (~20 – 20,000 references)
Tightly restricted to the topic	Encompassing the topic + related areas
Text discussion	Text, numerical, and graphical depiction
How, why	Who, what, when, where

Porter et al. (2002, p. 353)

Case study: What characterizes research on cloud computing published on the ISI Web of Science?


Sign In | My EndNote Web | My ResearcherID | My Citation Alerts
ISI Web of Knowledge℠
Take the next step
All Databases
Select a Database
Web of Science
Additional Resources
Search | Cited Reference Search | Advanced Search | Search History | Marked List (0)
Add Another Field >>
Search
Clear

Web of Science®

Search for:

in Author

Example: O'Brian C* OR OBrian C*
Need help finding papers by an author? Use Author Finder.

AND in Author

Example: O'Brian C* OR OBrian C*
Need help finding papers by an author? Use Author Finder.

AND in Publication Name

Example: Cancer* OR Journal of Cancer Research and Clinical Oncology

Copyright © 1997-2008 Search Technology, Inc. All rights reserved. This program is protected by U.S. and international copyright laws as described in Help About.

Search engine: ISI Web of Knowledge

- Contains Science Citation Index (SCI), Social Science Citation Index (SSCI) and Arts & Humanities Citation Index (A&HCI)
- The ISI Database is a gold standard by which national governments (e.g. USA, UK, Australia) evaluate national R&D performance.
- SCI offers dense coverage of most areas of science, and is cleanly and uniformly structured.

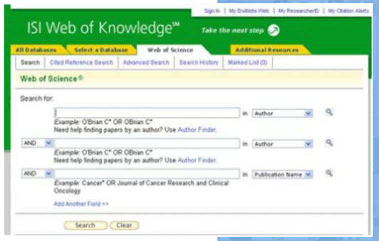
Limitations: no books (486 cloud books on Amazon), some important papers (Armbruster et al 2010) not registered, stringent citation counts

Source: Porter & Cunningham (2005, p. 356)

Search procedure in ISI WoS

- Initial search string for finding publications about cloud computing: cloud computing*, Cloud comput*, cloud-based*, cloud service*, PaaS*, IaaS*, SaaS*, cloud*, the cloud*
- Reduced search string after iterative testing and elimination of inefficient search terms: cloud computing*, cloud-based*, cloud service*, PaaS*, SaaS*, IaaS*
- Search restricted to publication titles
- Restriction to largest ISI categories: Computer science, engineering, telecommunications & Business economics

➢ Number of publications after initial search: 1224

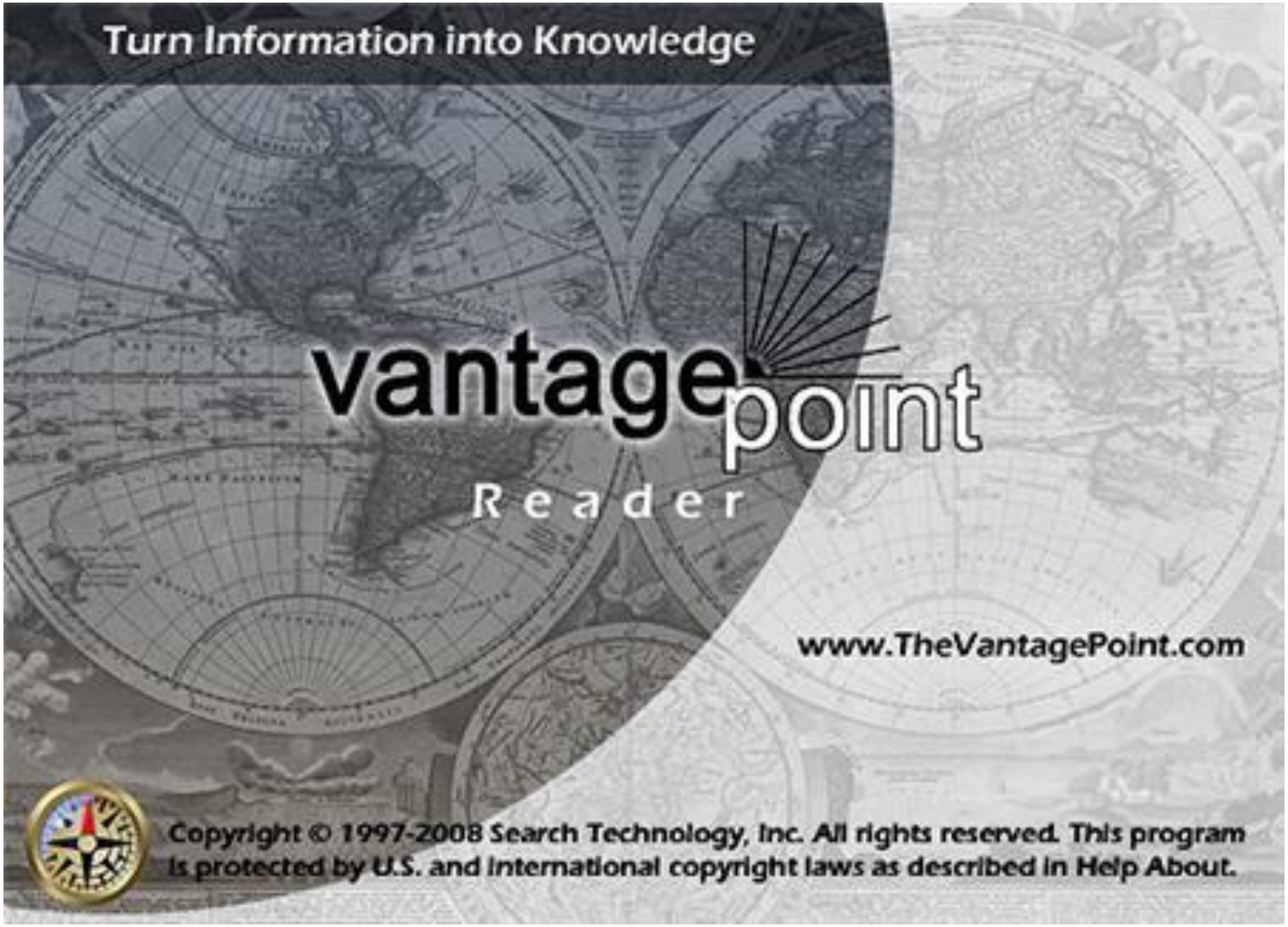


Example: O'Brian C* OR OBrian C*
Need help finding papers by an author? Use Author Finder.

Example: O'Brian C* OR OBrian C*
Need help finding papers by an author? Use Author Finder.

Example: Cancer* OR Journal of Cancer Research and Clinical Oncology

The text mining tool



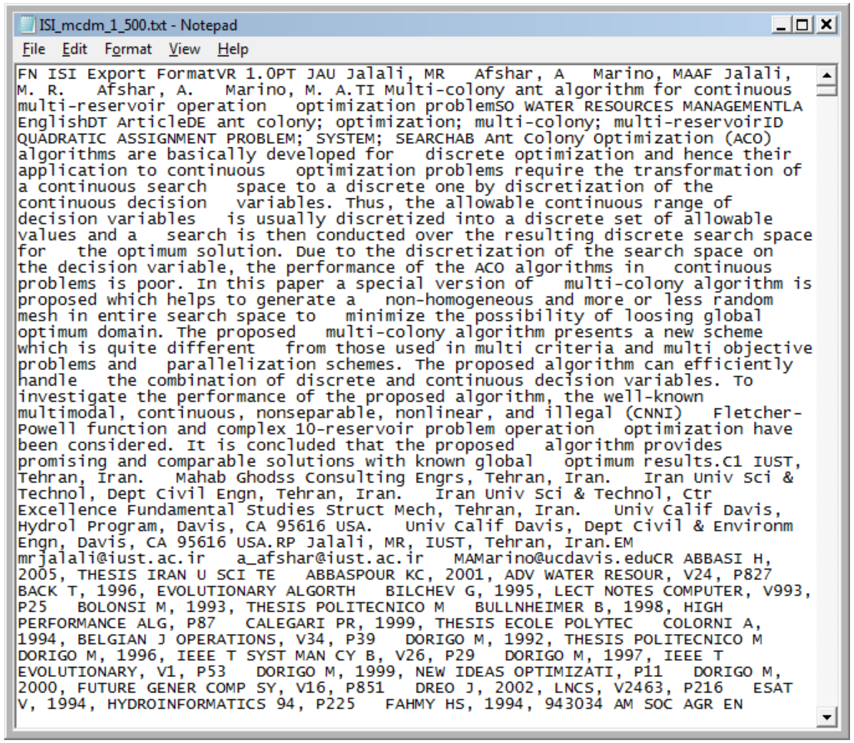
Turn Information into Knowledge

vantagepoint

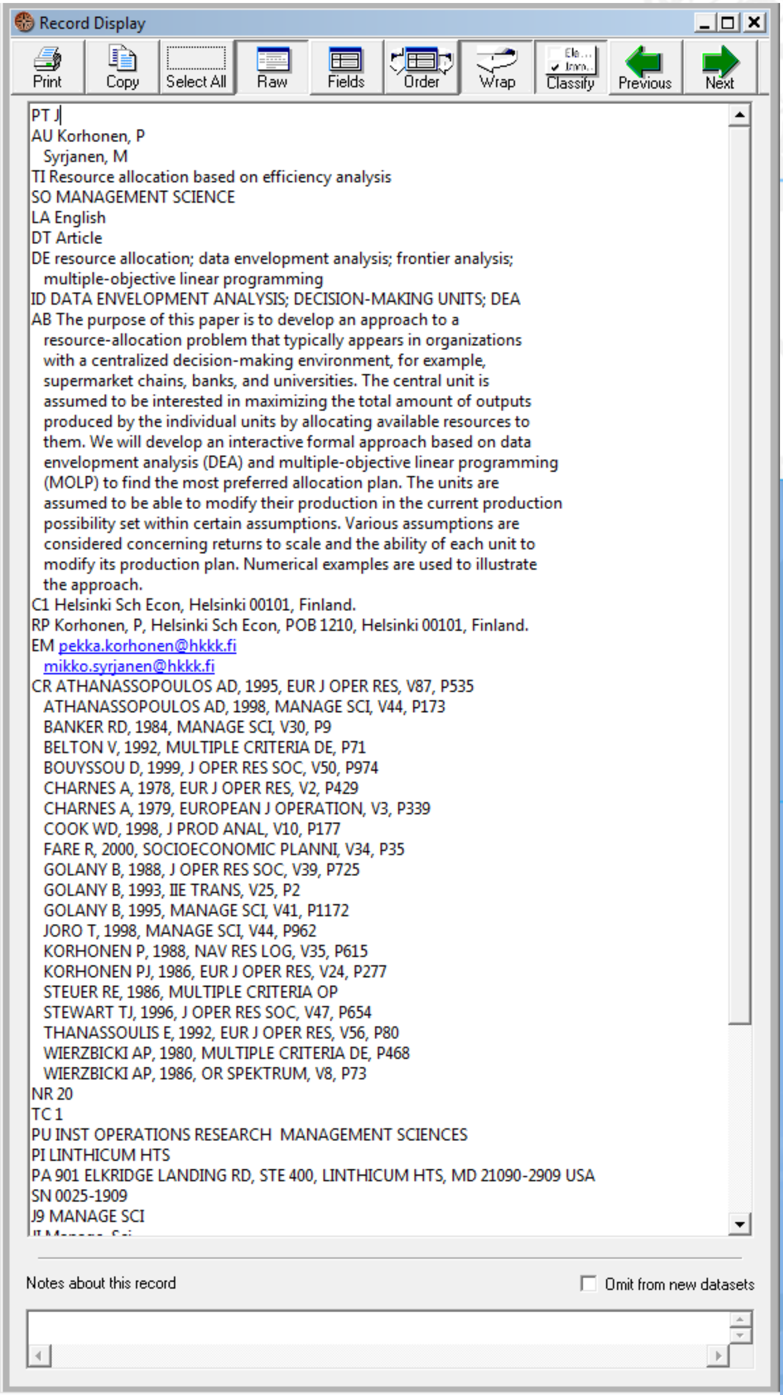
Reader

www.TheVantagePoint.com

The search result from ISI is saved and imported to Vantagepoint software
The bibliographic information of each article is nicely “fielded” in the text-mining tool



```
FN ISI Export FormatVR 1.0PT JAU Jalali, MR   Afshar, A   Marino, MAAF Jalali,
M. R.   Afshar, A.   Marino, M. A.TI Multi-colony ant algorithm for continuous
multi-reservoir operation   optimization problemSO WATER RESOURCES MANAGEMENTLA
EnglishDT ArticleDE ant colony; optimization; multi-colony; multi-reservoirID
QUADRATIC ASSIGNMENT PROBLEM; SYSTEM; SEARCHAB Ant Colony Optimization (ACO)
algorithms are basically developed for   discrete optimization and hence their
application to continuous   optimization problems require the transformation of
a continuous search   space to a discrete one by discretization of the
continuous decision   variables. Thus, the allowable continuous range of
decision variables   is usually discretized into a discrete set of allowable
values and a   search is then conducted over the resulting discrete search space
for   the optimum solution. Due to the discretization of the search space on
the decision variable, the performance of the ACO algorithms in   continuous
problems is poor. In this paper a special version of   multi-colony algorithm is
proposed which helps to generate a   non-homogeneous and more or less random
mesh in entire search space to   minimize the possibility of loosing global
optimum domain. The proposed   multi-colony algorithm presents a new scheme
which is quite different   from those used in multi criteria and multi objective
problems and   parallelization schemes. The proposed algorithm can efficiently
handle   the combination of discrete and continuous decision variables. To
investigate the performance of the proposed algorithm, the well-known
multimodal, continuous, nonseparable, nonlinear, and illegal (CNNI)   Fletcher-
Powell function and complex 10-reservoir problem operation   optimization have
been considered. It is concluded that the proposed   algorithm provides
promising and comparable solutions with known global   optimum results.C1 IUST,
Tehran, Iran.   Mahab Ghodss Consulting Engrs, Tehran, Iran.   Iran Univ Sci &
Technol, Dept Civil Engn, Tehran, Iran.   Iran Univ Sci & Technol, Ctr
Excellence Fundamental Studies Struct Mech, Tehran, Iran.   Univ Calif Davis,
Hydrol Program, Davis, CA 95616 USA.   Univ Calif Davis, Dept Civil & Environm
Engn, Davis, CA 95616 USA.RP Jalali, MR, IUST, Tehran, Iran.EM
mrjalali@iust.ac.ir   a_afshar@iust.ac.ir   MAMarino@ucdavis.eduCR ABBASI H,
2005, THESIS IRAN U SCI TE   ABBASPOUR KC, 2001, ADV WATER RESOUR, V24, P827
BACK T, 1996, EVOLUTIONARY ALGORTH   BILCHEV G, 1995, LECT NOTES COMPUTER, V993,
P25   BOLONSI M, 1993, THESIS POLITECNICO M   BULLNHEIMER B, 1998, HIGH
PERFORMANCE ALG, P87   CALEGARI PR, 1999, THESIS ECOLE POLYTEC   COLORNI A,
1994, BELGIAN J OPERATIONS, V34, P39   DORIGO M, 1992, THESIS POLITECNICO M
DORIGO M, 1996, IEEE T SYST MAN CY B, V26, P29   DORIGO M, 1997, IEEE T
EVOLUTIONARY, V1, P53   DORIGO M, 1999, NEW IDEAS OPTIMIZATI, P11   DORIGO M,
2000, FUTURE GENER COMP SY, V16, P851   DREO J, 2002, LNCS, V2463, P216   ESAT
V, 1994, HYDROINFORMATICS 94, P225   FAHMY HS, 1994, 943034 AM SOC AGR EN
```



```
PT J
AU Korhonen, P
   Syrjanen, M
TI Resource allocation based on efficiency analysis
SO MANAGEMENT SCIENCE
LA English
DT Article
DE resource allocation; data envelopment analysis; frontier analysis;
   multiple-objective linear programming
ID DATA ENVELOPMENT ANALYSIS; DECISION-MAKING UNITS; DEA
AB The purpose of this paper is to develop an approach to a
   resource-allocation problem that typically appears in organizations
   with a centralized decision-making environment, for example,
   supermarket chains, banks, and universities. The central unit is
   assumed to be interested in maximizing the total amount of outputs
   produced by the individual units by allocating available resources to
   them. We will develop an interactive formal approach based on data
   envelopment analysis (DEA) and multiple-objective linear programming
   (MOLP) to find the most preferred allocation plan. The units are
   assumed to be able to modify their production in the current production
   possibility set within certain assumptions. Various assumptions are
   considered concerning returns to scale and the ability of each unit to
   modify its production plan. Numerical examples are used to illustrate
   the approach.
C1 Helsinki Sch Econ, Helsinki 00101, Finland.
RP Korhonen, P, Helsinki Sch Econ, POB 1210, Helsinki 00101, Finland.
EM pekka.korhonen@hkkk.fi
   mikko.syrjanen@hkkk.fi
CR ATHANASSOPOULOS AD, 1995, EUR J OPER RES, V87, P535
   ATHANASSOPOULOS AD, 1998, MANAGE SCI, V44, P173
   BANKER RD, 1984, MANAGE SCI, V30, P9
   BELTON V, 1992, MULTIPLE CRITERIA DE, P71
   BOUYSSOU D, 1999, J OPER RES SOC, V50, P974
   CHARNES A, 1978, EUR J OPER RES, V2, P429
   CHARNES A, 1979, EUROPEAN J OPERATION, V3, P339
   COOK WD, 1998, J PROD ANAL, V10, P177
   FARE R, 2000, SOCIOECONOMIC PLANNI, V34, P35
   GOLANY B, 1988, J OPER RES SOC, V39, P725
   GOLANY B, 1993, IIE TRANS, V25, P2
   GOLANY B, 1995, MANAGE SCI, V41, P1172
   JORO T, 1998, MANAGE SCI, V44, P962
   KORHONEN P, 1988, NAV RES LOG, V35, P615
   KORHONEN PJ, 1986, EUR J OPER RES, V24, P277
   STEUER RE, 1986, MULTIPLE CRITERIA OP
   STEWART TJ, 1996, J OPER RES SOC, V47, P654
   THANASSOULIS E, 1992, EUR J OPER RES, V56, P80
   WIERZBICKI AP, 1980, MULTIPLE CRITERIA DE, P468
   WIERZBICKI AP, 1986, OR SPEKTRUM, V8, P73
NR 20
TC 1
PU INST OPERATIONS RESEARCH  MANAGEMENT SCIENCES
PI LINTHICUM HTS
PA 901 ELKRIDGE LANDING RD, STE 400, LINTHICUM HTS, MD 21090-2909 USA
SN 0025-1909
J9 MANAGE SCI
JI Manage Sci
```

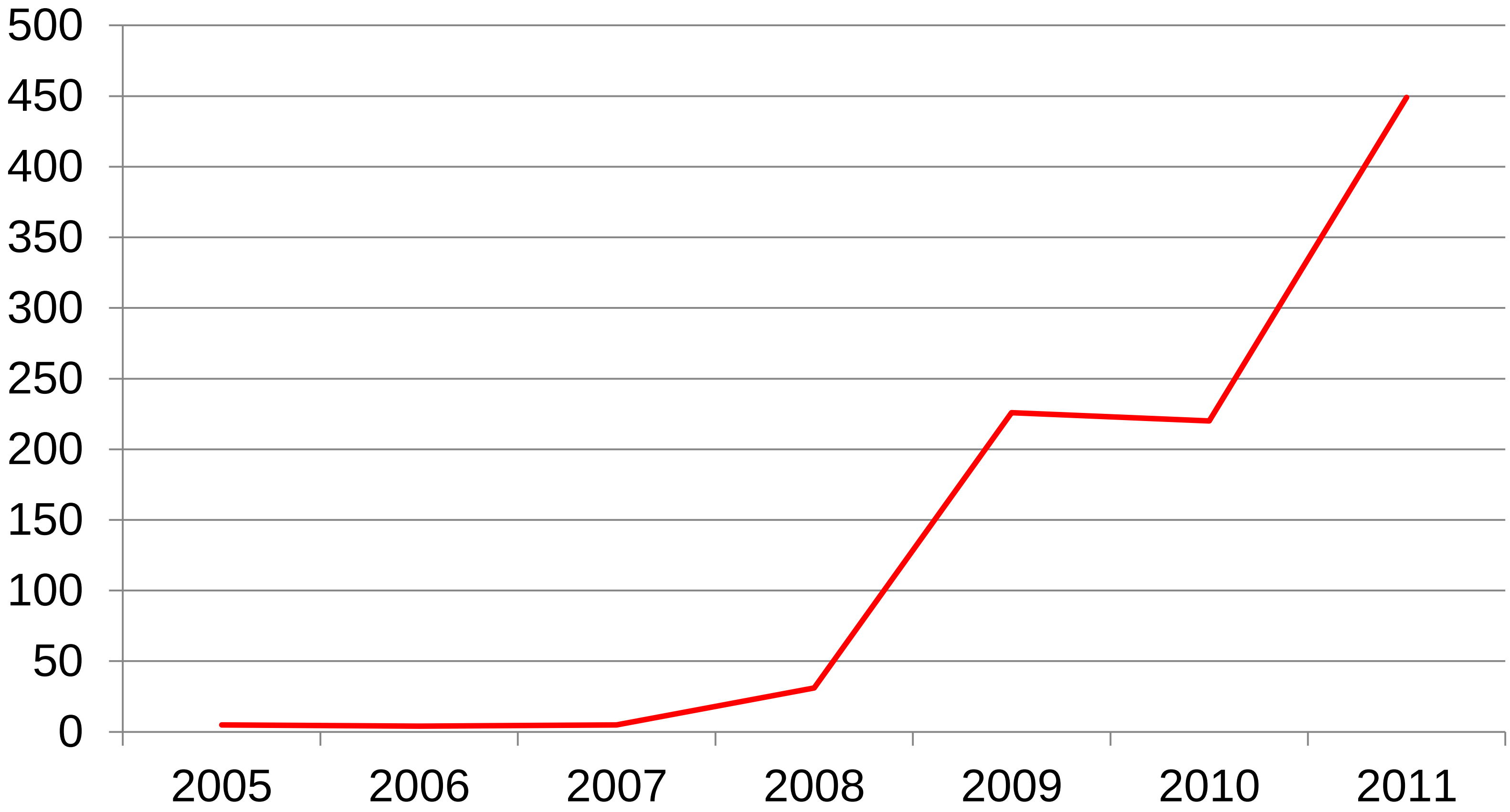
Analysis – part 1

- How is the field growing?

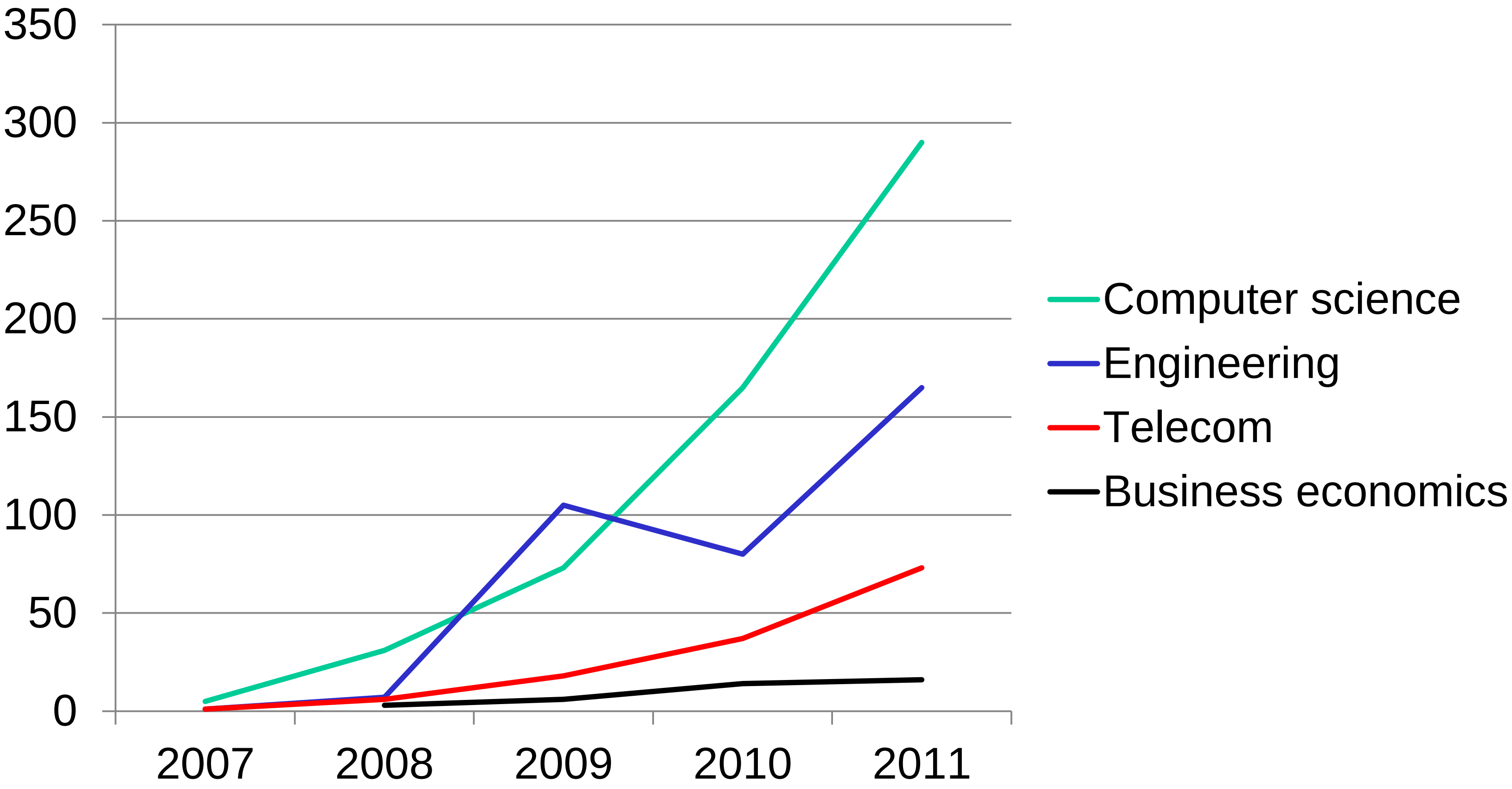
- What disciplines, instututions & authors are the most prolific contributors?

-Are there dominant research groups and if so where are they?

Publications with ‘cloud computing’ or equivalent in title - raw count



Disciplinary fields and their development



Publication outlets with papers about cloud computing (*average no. of papers per outlet 1,8)*

	# outlets	% of total
# outlets with 1 paper	**415**	**68**
# outlets with 2 papers	101	17
# outlets with 3 papers	33	5,4
# outlets with 4-10 papers	48	7,8
# of outlets with 11-20 papers	6	0,9
# of outlets with >20 papers	1	0,2
Total	608	100%

Distribution of ISI-WOS citations

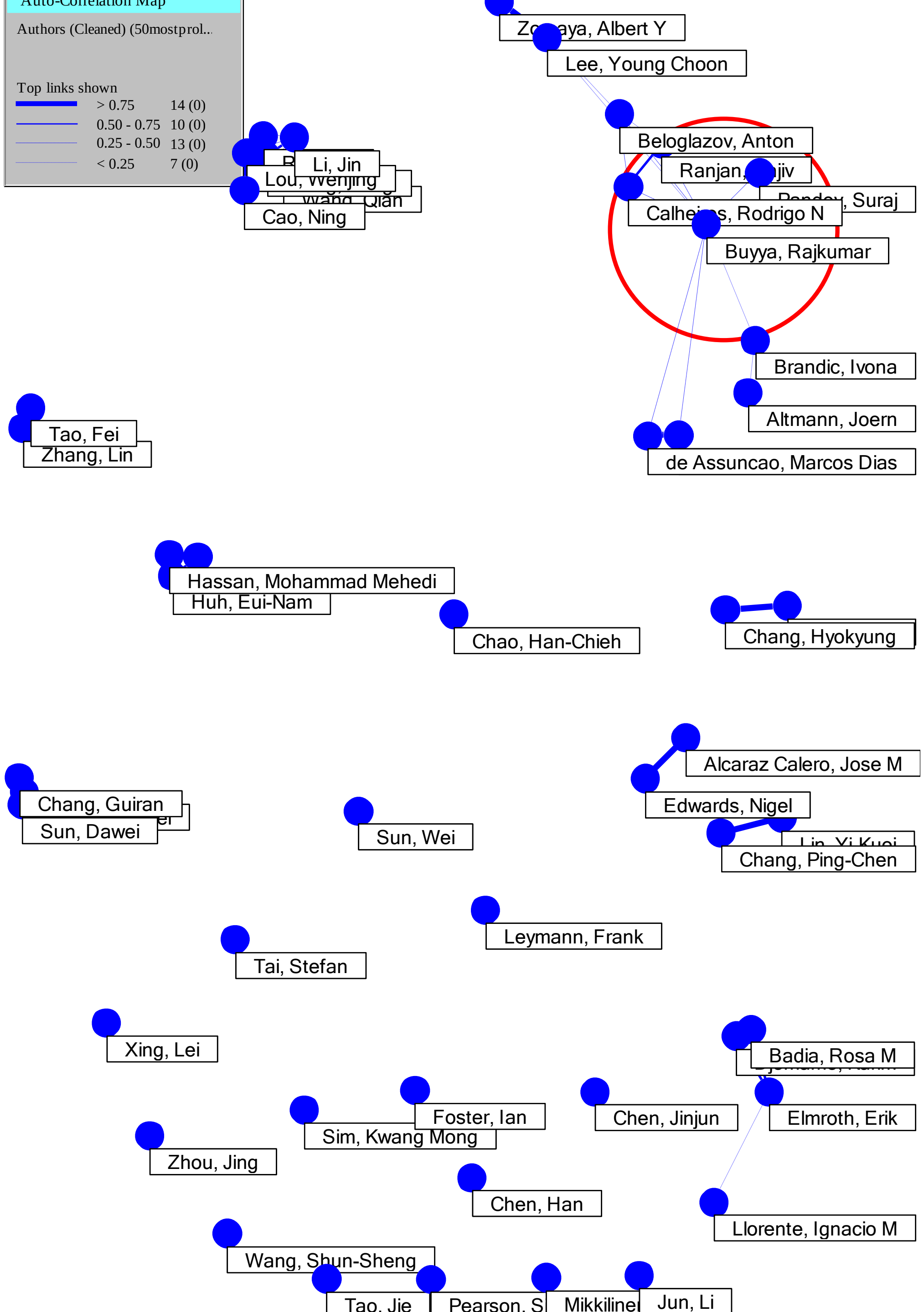
# Citations	# papers	% of total
0	854	76
1	115	10
2	52	5
3-10	75	7
11-20	11	1
21-50	6	0,6
194	1	0,08
Total	**1108**	**100%**

10 most prolific authors and institutions

Author	# records		Institution	# records
Buyya, R. (Univ. Melbourne)	25		Univ. Melbourne	27
Lou, W. (Worcester Polytech, USA)	11		Tsingua Univ.	18
Ren, K. (Guangzhou Univ.)	11		Univ. Elect Sci & Tech. China	18
Wang, Cong (Worcester Polytech)	10		Fujitsu Ltd	16
Hassan, M. M. (Kyung Hee Univ, S. Korea)	8		IBM Corp	15
Huh, E. (Kyung Hee Univ, S. Korea)	8		Wuhan Univ.	14
Wang, Quian (Guangzhou Univ)	7		Kyung Lee Univ.	11
Brandic, I. (Vienna Univ Technol)	6		IIT	10
Li, J. (IIT)	6		Microsoft Corp	10
Mikkelini, R. (Kawa Objects Inc)	6		Beijing Univ Posts & Telecommun	9
% of total publications (1108)	8,8%		Sum	13,5%

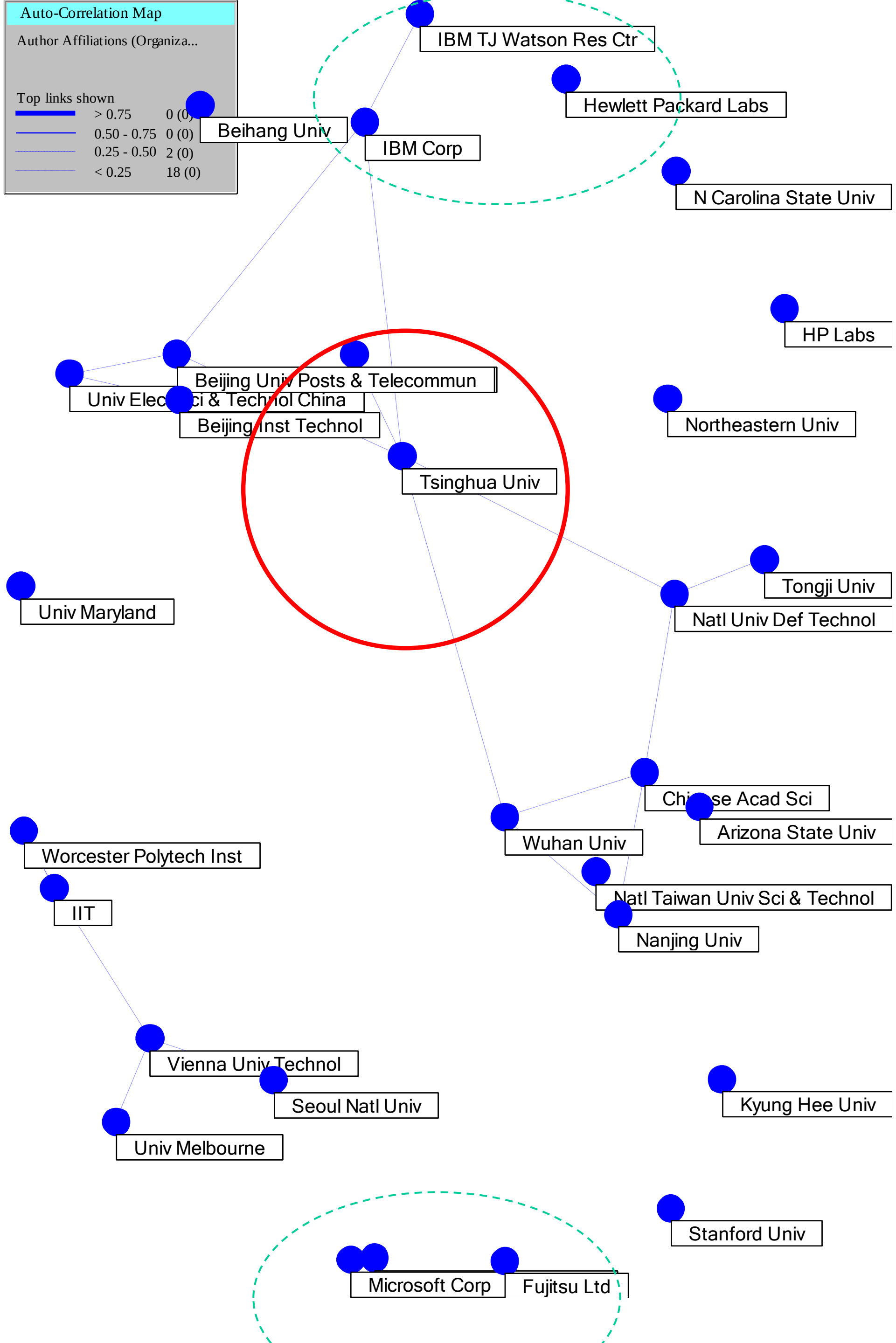
Collaboration map of 50 most prolific authors (>3 records

- Lines and distance between nodes indicate degree to which authors occur as authors in the same article abstracts



Collaboration map of 50 most prolific institutions (>3 records

- Lines and distance between nodes indicate degree to which institutions occur as contributors in the same article abstracts



Analysis part 2: a text mining approach

Which are the key academic research dimensions in cloud computing?

Do these reflect Venter & Whitleys 2012s seven dimensions of ‘cloud desires’ (which are based on interviews from the corporate sector)

What are the trends right now?

Keyword extraction and data cleaning of ISI-WoS dataset

1. Words in abstracts, keywords and titles were extracted from all 1108 records in the cleaned dataset. **Total: 19977 words**

   – Words from abstracts and titles extracted by VP NLP function - Word separation, removal of stop words (the, that, who, etc)

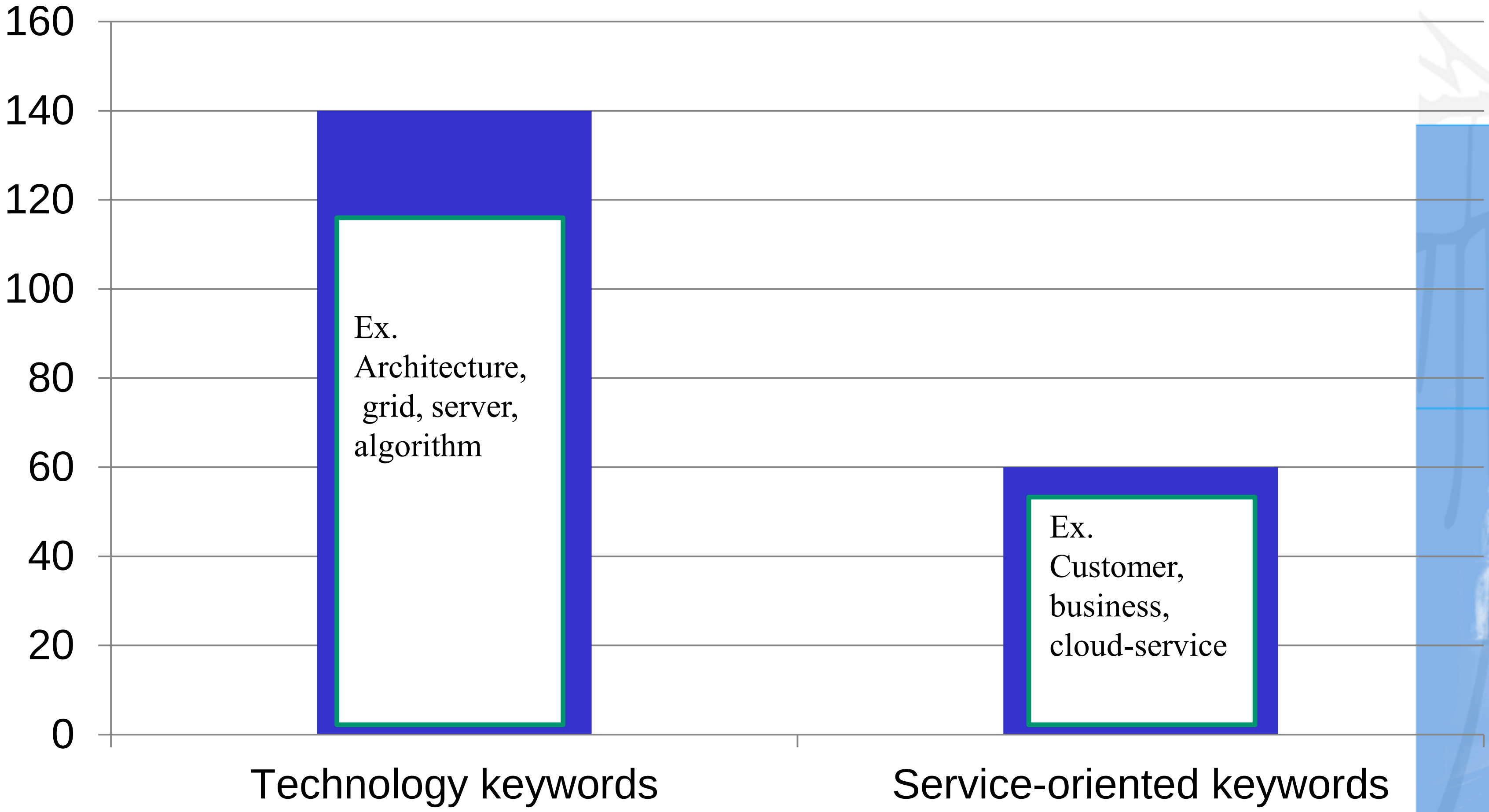
2. Combination of equivalent terms (*platform* and *platforms)* – using word stemming filters in VP. Manual removal of remaining equivalents **total: 18031**

3. Removal of main search words, names of countries & companies + trivial words. **Total: 18009**

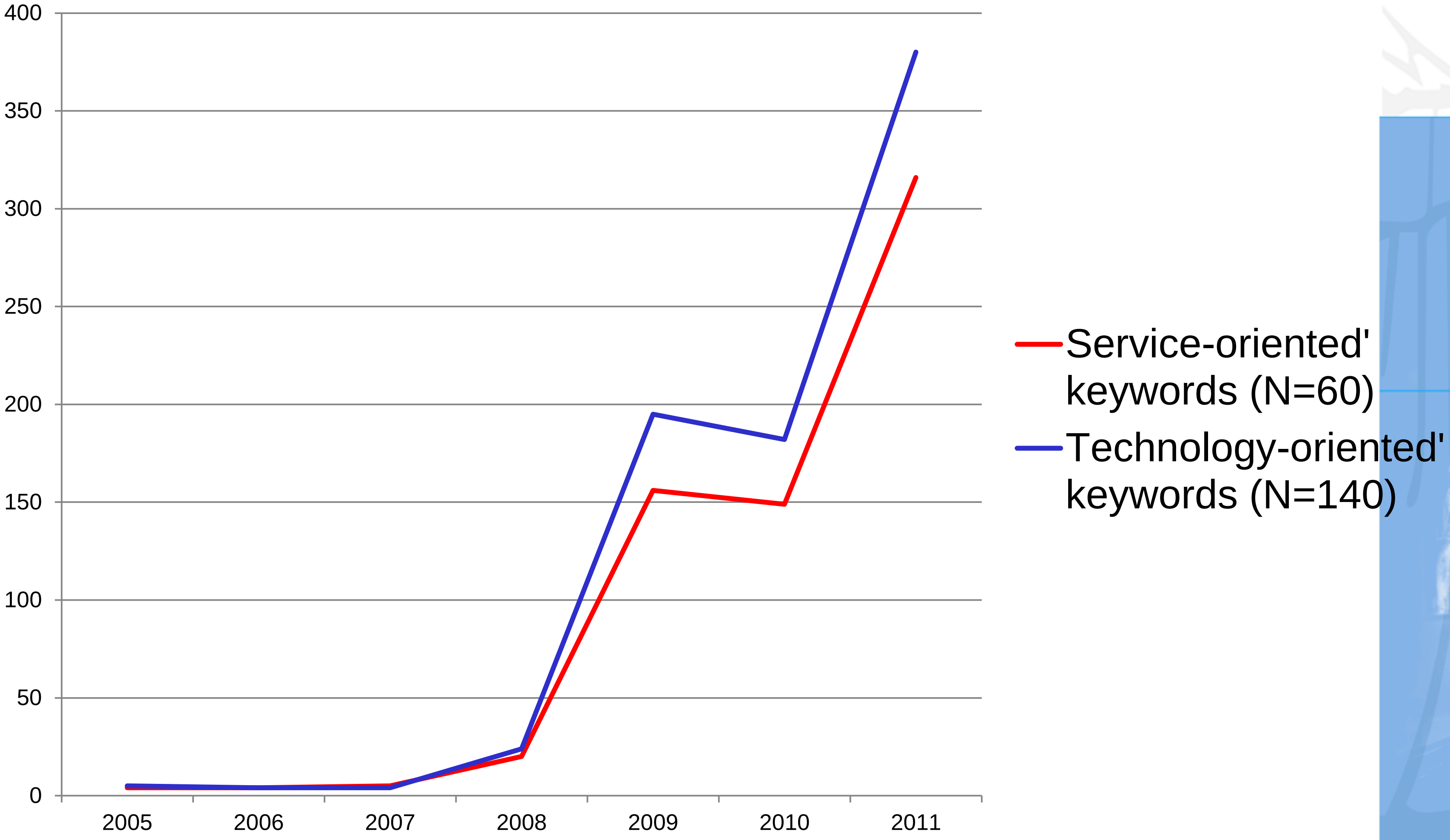
Most frequent keywords in cleaned list

Top 5 Keywords	**#Records**	**# Instances**
Service	372	591
User	187	275
Environment	168	249
Secure	123	177
Infrastructure	109	149

Distribution of ‘technology keywords’ and ‘service keywords’ – among 200 top occurring keywords (>15 records) in ISI-WoS dataset



Occurrence of ‘service-oriented’ and ‘technology-oriented’ keywords in publications about cloud computing 2005-2011



Center for Technology, Innovation and Culture, University of Oslo. Web: http//www.tik.uio.no

Claims about cloud computing based on insights from corporate sector (Venters and Whitley 2012, p. 3)

The technological dimensions of cloud desire	
Equivalence	Receive technical services that are equivalent to locally running services
Variety	Receive service that provides variety with respect to relevant use
Abstraction	Receive technical services that abstracts away unnecessary complexity
Scalability	Receive service which is scalable to demand
The service dimension of cloud desire	
Efficiency	Receive service that help users be more economically efficient
Creativity	Receive services which aid innovation and creativity
Simplicity	Receive service which is simple to understand and use

ISI-WoS keywords reflecting Venter & Whitleys technological and service desires

The technological dimensions of cloud desire	Keywords	# records	# instances
Equivalence	Equivalence	2	3
	Equivalent	5	5
Variety	Variety	33	35
	Variation	8	8
Abstraction	Abstraction	13	13
	Abstracting	12	15
Scalability	Scalability	**34**	**34**
	Scalable	**97**	**113**
Average hits per keyword		26	28
The service dimension of cloud desire	Keywords	# records	# instances
Efficiency	Efficiency	**42**	**44**
	Efficient	**139**	**146**
Creativity	Creativity	2	2
	Creative	1	1
Simplicity	Simplicity	0	0
	Simple	31	32
Average hits per keyword		**36**	**38**

Principal components analysis (PCA) of 1108 documents about cloud computing

Basic form of factor analysis which linearly transforms a original set of observed variables to a substantially smaller set of artificial variables that represents most of the information in the original set (Dunteman 1989)

Typically used on survey data with 10-30 variables for each artificial dimension

In the context of this analysis, the cases are the 1108 documents extracted from ISI – WoS and the variables are 122 keywords (binary 1-0) that are common across these documents

Tradeoff between keyword coverage and explained variance

Principal components analysis procedure

1. Elimination of keywords that appeared in fewer than 15 records, resulting in 122 unique and very frequently cited words. These cover 89 % of the 1108 records in the dataset

2. Principal components analysis in Vantagepoint reduced 112 keywords-variables into 14 principal components.

3. **Can be interpreted as theoretical constructs if validated against theory and knowledge about the research domain (Venters & Whitley 2012)**

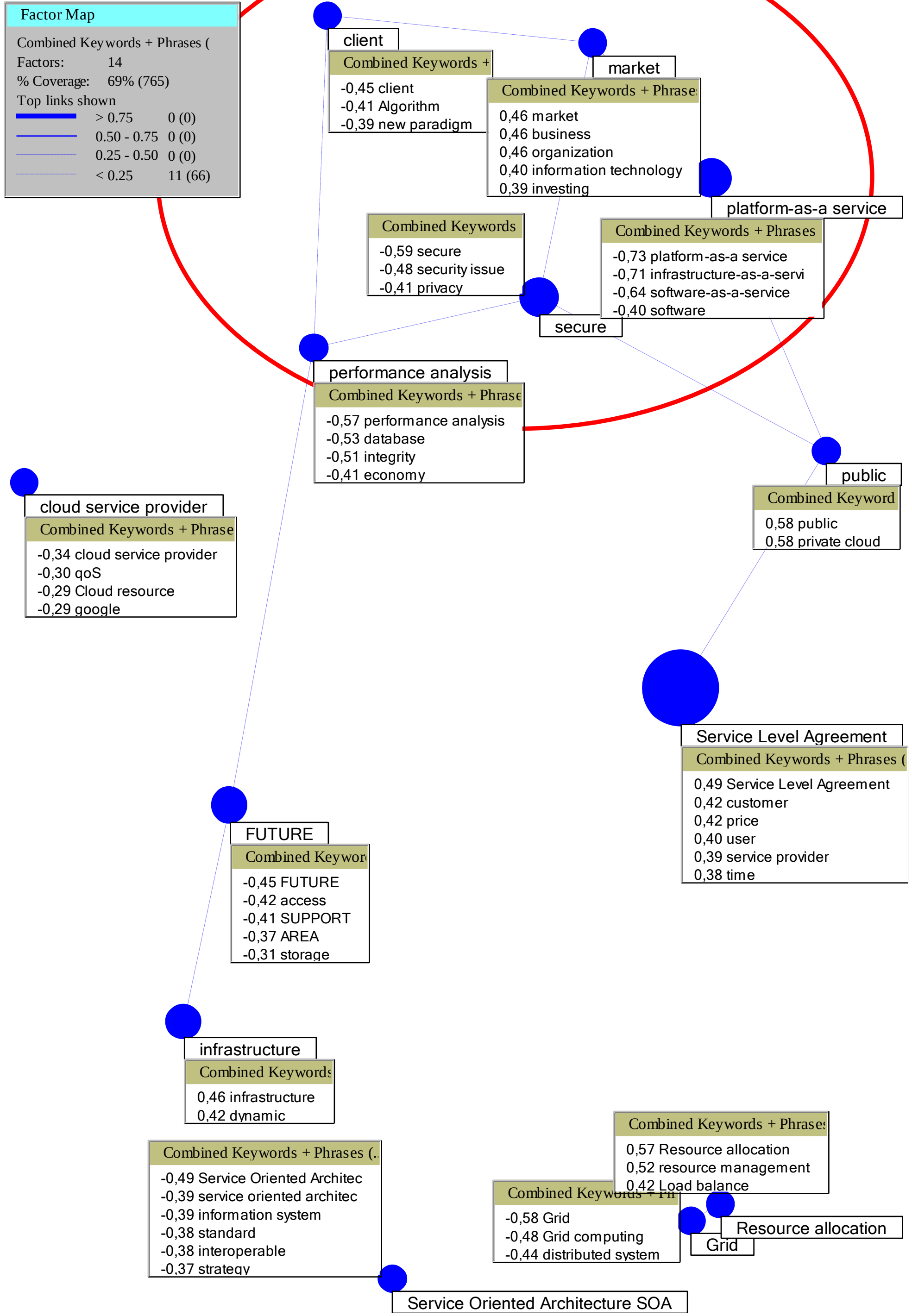
Will show the results in the VP software here….

Map of 14 principal components

- Decomposes keywords lists into a set of discrete clusters

- High factor loadings (> 0,5 / -0,5) indicate that the keywords occur frequently together in the same article abstracts

- Lines and distance between components indicate degree to which keywords of different components co-occur in the same article abstracts



Factor Map

Combined Keywords + Phrases (

Factors: 14

% Coverage: 69% (765)

Top links shown

▬	> 0.75	0 (0)
—	0.50 - 0.75	0 (0)
—	0.25 - 0.50	0 (0)
—	< 0.25	11 (66)

client

Combined Keywords +

-0,45 client
-0,41 Algorithm
-0,39 new paradigm

market

Combined Keywords + Phrase

0,46 market
0,46 business
0,46 organization
0,40 information technology
0,39 investing

platform-as-a service

Combined Keywords + Phrases

-0,73 platform-as-a service
-0,71 infrastructure-as-a-servi
-0,64 software-as-a-service
-0,40 software

secure

Combined Keywords

-0,59 secure
-0,48 security issue
-0,41 privacy

performance analysis

Combined Keywords + Phrase

-0,57 performance analysis
-0,53 database
-0,51 integrity
-0,41 economy

cloud service provider

Combined Keywords + Phrase

-0,34 cloud service provider
-0,30 qoS
-0,29 Cloud resource
-0,29 google

public

Combined Keyword

0,58 public
0,58 private cloud

Service Level Agreement

Combined Keywords + Phrases (

0,49 Service Level Agreement
0,42 customer
0,42 price
0,40 user
0,39 service provider
0,38 time

FUTURE

Combined Keywor

-0,45 FUTURE
-0,42 access
-0,41 SUPPORT
-0,37 AREA
-0,31 storage

infrastructure

Combined Keywords

0,46 infrastructure
0,42 dynamic

Combined Keywords + Phrases (..

-0,49 Service Oriented Architec
-0,39 service oriented architec
-0,39 information system
-0,38 standard
-0,38 interoperable
-0,37 strategy

Service Oriented Architecture SOA

Combined Keywords + Phrases

0,57 Resource allocation
0,52 resource management
0,42 Load balance

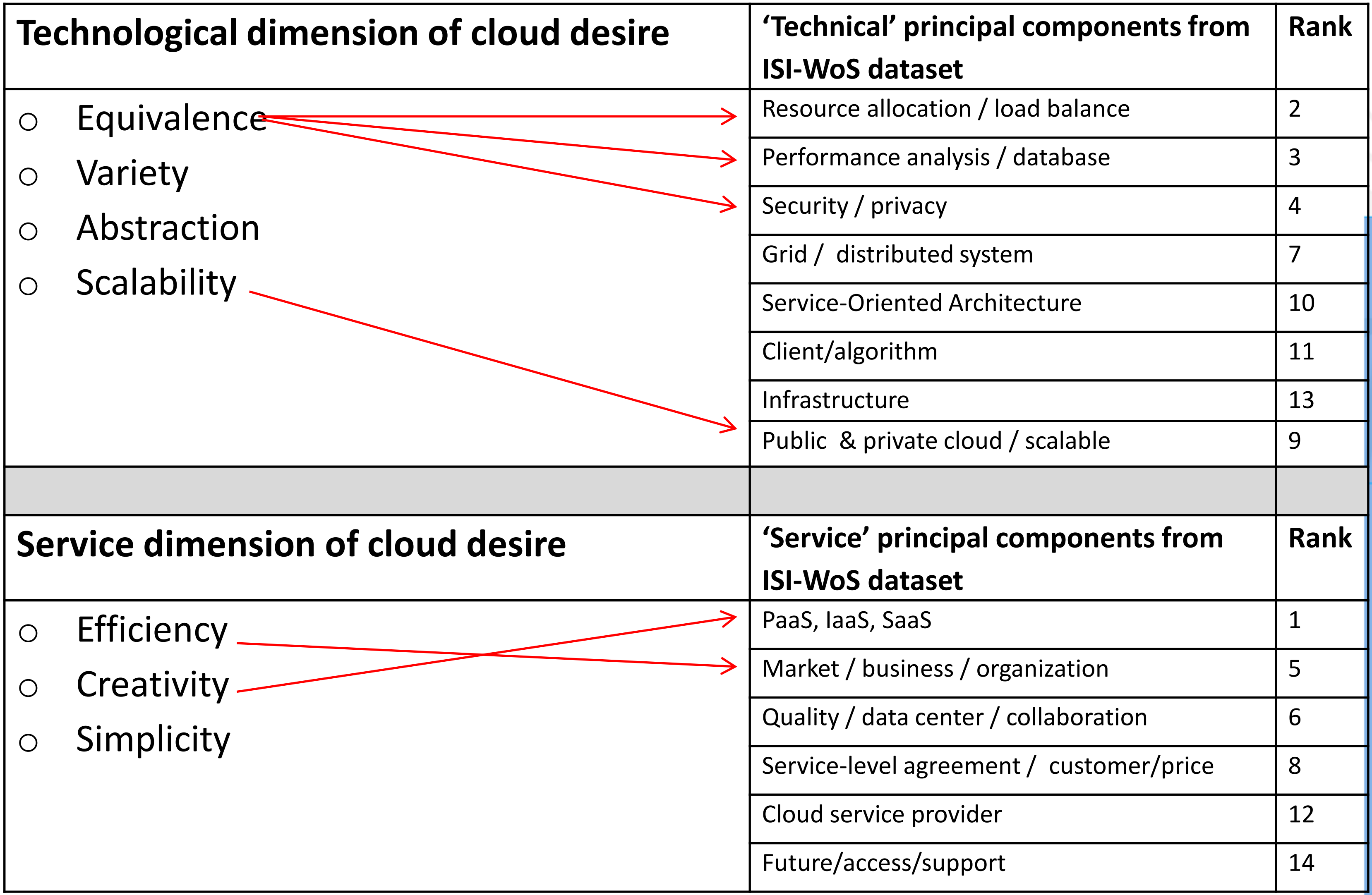
Resource allocation

Combined Keywords + Ph

-0,58 Grid
-0,48 Grid computing
-0,44 distributed system

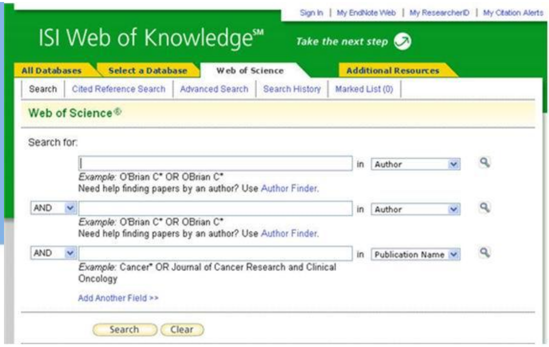
Grid

Claims about cloud desires vs principal components from ISI-WoS data



Technological dimension of cloud desire	‘Technical’ principal components from ISI-WoS dataset	Rank
○ Equivalence ○ Variety ○ Abstraction ○ Scalability	Resource allocation / load balance	2
	Performance analysis / database	3
	Security / privacy	4
	Grid / distributed system	7
	Service-Oriented Architecture	10
	Client/algorithm	11
	Infrastructure	13
	Public & private cloud / scalable	9
Service dimension of cloud desire	**‘Service’ principal components from ISI-WoS dataset**	**Rank**
○ Efficiency ○ Creativity ○ Simplicity	PaaS, IaaS, SaaS	1
	Market / business / organization	5
	Quality / data center / collaboration	6
	Service-level agreement / customer/price	8
	Cloud service provider	12
	Future/access/support	14

Which are the most important trends right now?



Web of Science®

Search for:

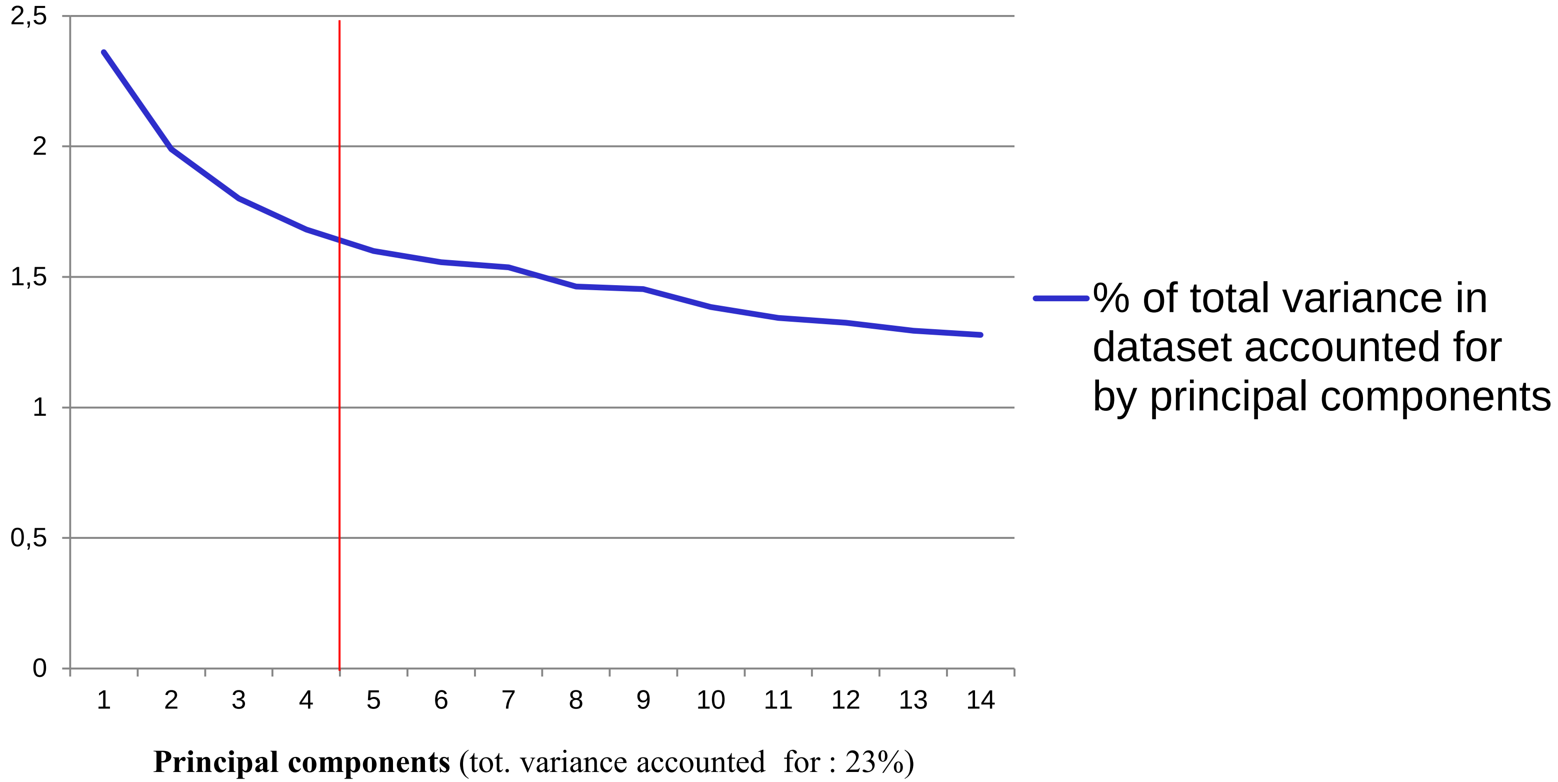
Example: O'Brian C* OR OBrian C*
Need help finding papers by an author? Use Author Finder.

Example: O'Brian C* OR OBrian C*
Need help finding papers by an author? Use Author Finder.

Example: Cancer* OR Journal of Cancer Research and Clinical Oncology

Copyright © 1997-2008 Search Technology, Inc. All rights reserved. This program is protected by U.S. and International copyright laws as described in Help About.

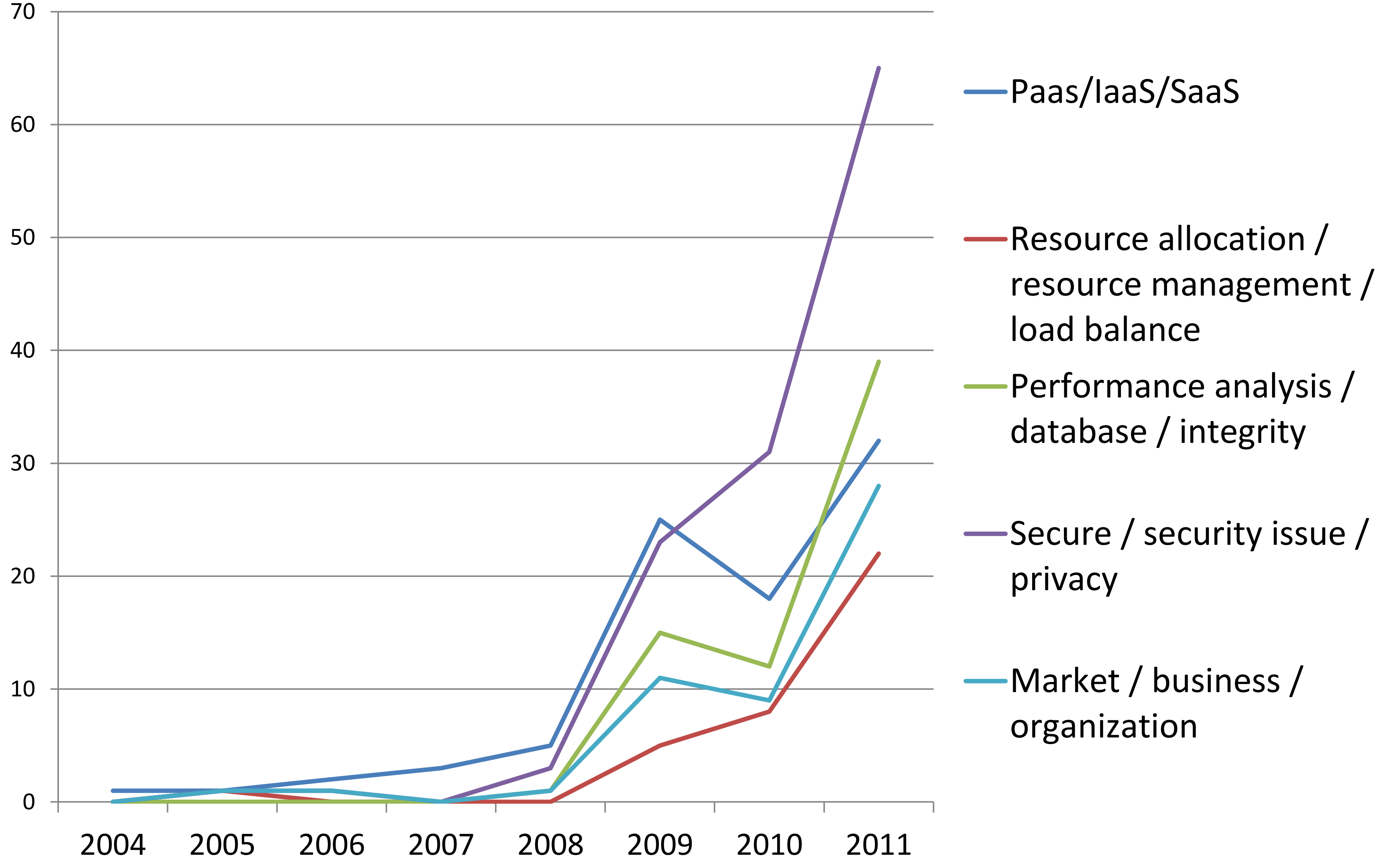
Scree plot of 14 principal components



Top 5 principal components

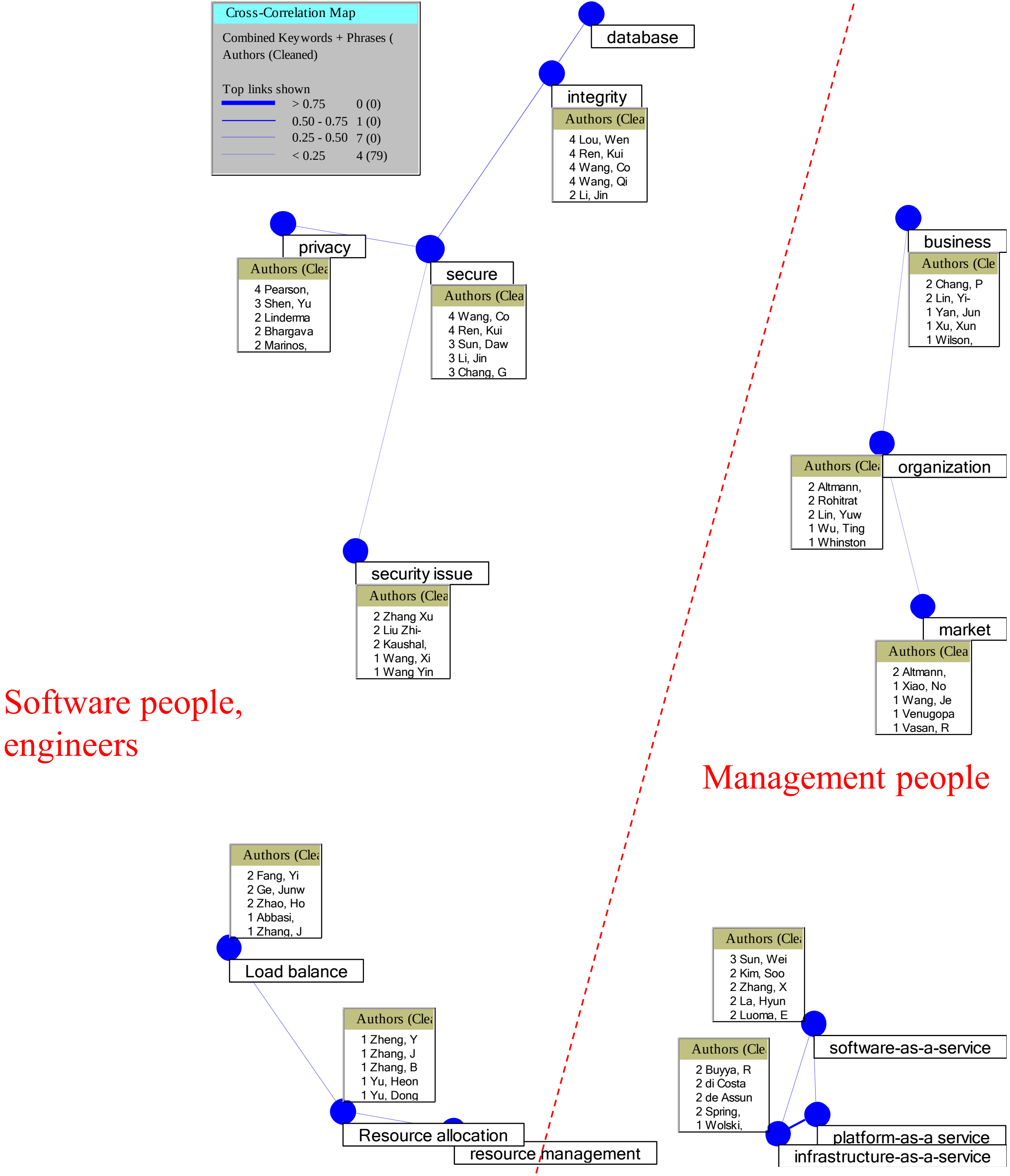
Top 5 principal components		Cumulative variance	Keywords	Eigenvalue loadings (+1 to -1)
1	Platform as a service	*2,4 %*	Platform as a service	-0,73
			Infrastructure as a service	-0,71
			Software as a service	-0,64
2	Resource allocation	4,4	Resource allocation	0,57
			Resource management	0,52
			Load balance	0,42
3	Performance analysis	6,2 %	Performance analysis	-0,59
			Database	-0,52
			Integrity	-0,50
4	Secure	7,8 %	Secure	-0,59
			Security issue	-0,47
			Privacy	-0,40
5	Market	9,4 %	Market	0,47
			Organization	0,46
			Business	0,46

Evolution of PCA constructs based on raw count of top3 loading keywords



Cross-correlation map of authors writing about all 5 PCA constructs

- Authors x top 3 keywords in PC
- Show groups of people who write about the same things
- Lines and distance between components indicate degree to which the same keywords occur in the article abstracts of different authors



Cross-Correlation Map

Combined Keywords + Phrases (
Authors (Cleaned)

Top links shown

▬	> 0.75	0 (0)
—	0.50 - 0.75	1 (0)
—	0.25 - 0.50	7 (0)
—	< 0.25	4 (79)

database

integrity

Authors (Clea
4 Lou, Wen
4 Ren, Kui
4 Wang, Co
4 Wang, Qi
2 Li, Jin

privacy

Authors (Clea
4 Pearson,
3 Shen, Yu
2 Linderma
2 Bhargava
2 Marinos,

secure

Authors (Clea
4 Wang, Co
4 Ren, Kui
3 Sun, Daw
3 Li, Jin
3 Chang, G

security issue

Authors (Clea
2 Zhang Xu
2 Liu Zhi-
2 Kaushal,
1 Wang, Xi
1 Wang Yin

business

Authors (Cle
2 Chang, P
2 Lin, Yi-
1 Yan, Jun
1 Xu, Xun
1 Wilson,

organization

Authors (Clea
2 Altmann,
2 Rohitrat
2 Lin, Yuw
1 Wu, Ting
1 Whinston

market

Authors (Clea
2 Altmann,
1 Xiao, No
1 Wang, Je
1 Venugopa
1 Vasan, R

Software people, engineers

Management people

Authors (Clea
2 Fang, Yi
2 Ge, Junw
2 Zhao, Ho
1 Abbasi,
1 Zhang, J

Load balance

Authors (Clea
1 Zheng, Y
1 Zhang, J
1 Zhang, B
1 Yu, Heon
1 Yu, Dong

Resource allocation

resource management

Authors (Clea
3 Sun, Wei
2 Kim, Soo
2 Zhang, X
2 La, Hyun
2 Luoma, E

software-as-a-service

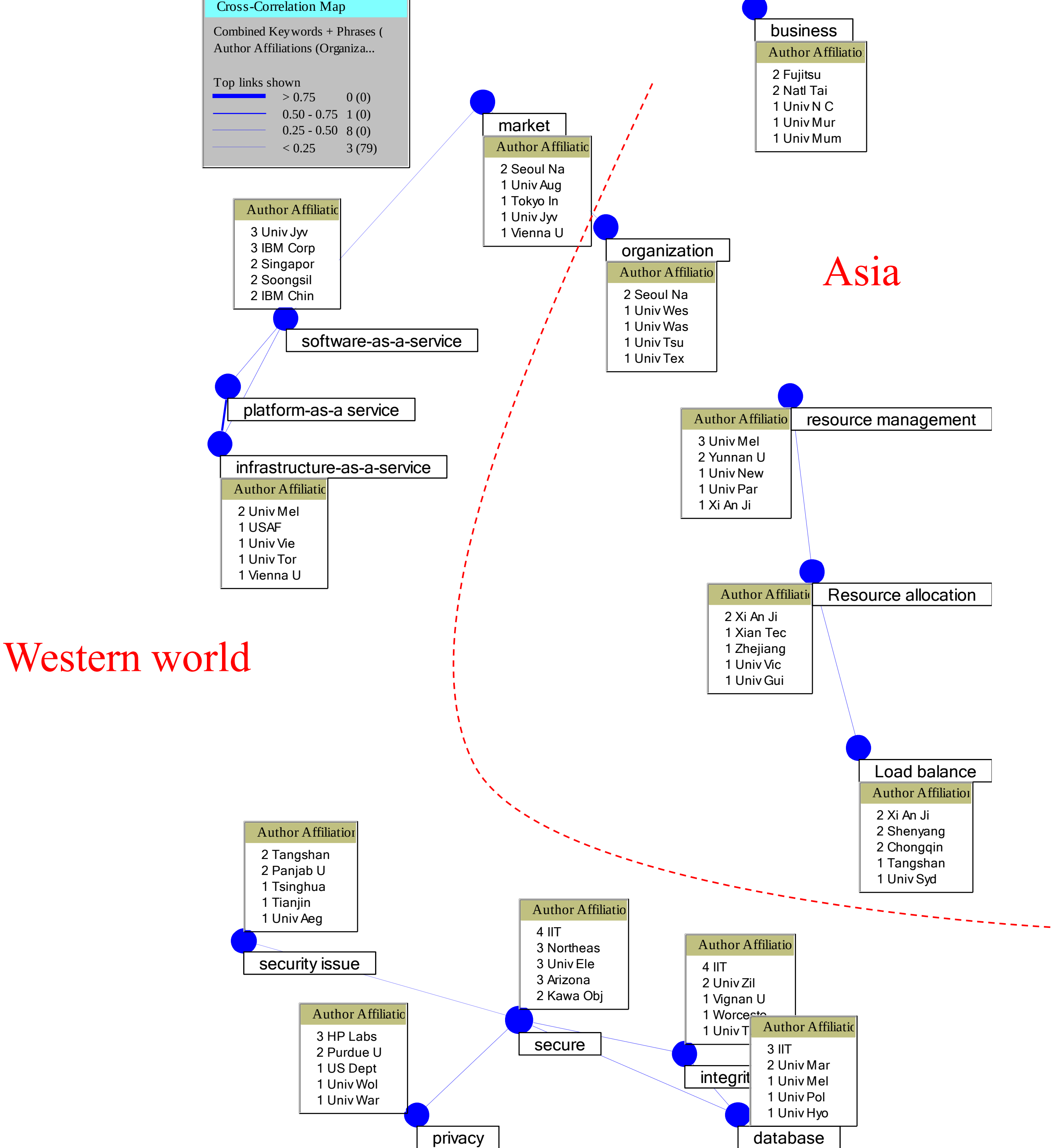
Authors (Cle
2 Buyya, R
2 di Costa
2 de Assun
2 Spring,
1 Wolski,

platform-as-a service

infrastructure-as-a-service

Cross-correlation map of institutions publishing about all 5 PCA constructs

- Authors x top 3 keywords in PC
- Show groups of institutions that publish about the same things
- Lines and distance between components indicate degree to which the same keywords occur in article abstracts of authors from different institutions



Conclusions 1

- Cloud computing is a fast-evolving but highly scattered research area

  -Many isolated research communities with little contact and few overlapping research topics

- ‘Not an integrated research field, but should rather understood as a phenomenon that is the object of research of many different fields

Conclusions 2

- ISI- WoS keyword analysis and principal components analysis seem to mirror Venters and Whitleys claims about ‘technological and service-oriented’ cloud desires.
  - One exception is the ‘creativity’ dimension which is not strongly reflected in the academic literature
- We need to put more time and effort into comparing and linking these two analyses (a function)

Conclusions 3

- Realtively equal balance between technical and service-oriented topics, but emphasis seems to be moving towards services (a sign of early maturation)
- Security, performance analysis and PaaS/SaaS/IaaS seem to be the most important current trends
- Chinese institutions seem well positioned to dominate the research area in the future

.