



UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

ANALYSIS OF SCIENTIFIC VIRTUAL COMMUNITIES OF PRACTICE

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL
INDUSTRIAL

JACQUELINE PAZ ARAYA REBOLLEDO

PROFESOR GUÍA:
SEBASTIÁN RÍOS PÉREZ

MIEMBROS DE LA COMISIÓN:
JAVIER BUSTOS JIMÉNEZ
FELIPE AGUILERA VALENZUELA

SANTIAGO DE CHILE
2015

ANÁLISIS DE COMUNIDADES VIRTUALES DE PRÁCTICA

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE: Ingeniero Civil Industrial
POR: Jacqueline Paz Araya Rebolledo
FECHA: Enero 2015
PROFESOR GUÍA: Sebastián Ríos Pérez

Las diferentes redes sociales han surgido a partir del sentido común y natural de los humanos por reunirse en torno a un tema, sintiendo que pertenecen a una *Comunidad*, la cual es representada por una red de relaciones complejas entre las unidades que cambia con el tiempo. Una *Comunidad* es un grupo de vértices que comparten propiedades comunes y desempeñan un papel similar dentro del grupo, las cuales pueden ser clasificadas como *Comunidades de interés*, en el que los miembros comparten un interés particular, y *Comunidades de práctica*, donde los miembros comparten inquietudes, participan y desarrollan un tema volviéndose expertos. Si estas interacciones ocurren sobre plataformas en línea, son llamadas *Comunidades virtuales de interés* (VCoI) y *Comunidades virtuales de práctica* (VCoP).

El estudio de las *Comunidades virtuales* (VC) no sólo ayuda a entender su estructura interna, sino que también a descubrir cómo el conocimiento es compartido, los principales miembros, proporcionar herramientas a los administradores para mejorar la participación y asegurar la estabilidad de la comunidad en el tiempo. El área de Análisis de Redes Sociales y de Minería de Datos han estudiado el problema, pero ninguno toma en cuenta el significado del contenido que los miembros de una comunidad generan.

Por lo tanto, la principal contribución de este trabajo es tomar en cuenta la semántica de los contenidos creados por los miembros de dos VCoP, así como las propiedades estructurales de las redes que forman, para estudiar la existencia de otros miembros claves, buscar los principales temas de investigación, y estudiar las propiedades de las nuevas redes creadas con contenido. Se utilizó una VCoP científica del área de computación ubicua, y otra del área Web Semántica, considerando como data los autores de los papers aceptados en las conferencias de las comunidades y su contenido.

Este trabajo propone dos métodos, el primero, busca representar cada artículo escrito por los miembros por sus Keywords, y el segundo, busca extraer los temas subyacentes de cada paper con el modelo probabilístico LDA. Con el resultado de estos métodos, las interacciones entre autores pueden ser construidas basándose en el contenido en lugar de sólo la relación de coautoría (red base para comparar los métodos). La metodología propuesta es un proceso híbrido llamado SNA-KDD que incluye la extracción y procesamiento de datos de texto, para su posterior análisis con SNA para descubrir nueva información, utilizando teoría de grafos, algoritmos de clasificación (HITS y PageRank) y diferentes medidas estructurales para redes.

Los resultados muestran que las redes científicas en estudio pueden ser modeladas como VCoPs usando la metodología SNA-KDD usando teoría de grafos. Esto queda evidenciado en los resultados de la métrica *Modularidad*, obteniendo valores sobre 0,9 en la mayoría de las redes, lo que indica una estructura de comunidad.

Además, los métodos propuestos para introducir el contenido generado por sus miembros, Keywords y Modelo de Tópicos LDA, permite reducir la densidad de todas las redes, eliminando relaciones no relevantes. En la red de Computación Ubicua, con 1920 nodos, se redujo de 5.452 arcos a 1.866 arcos para método de Keywords y a 2.913 arcos para modelo LDA; mientras que en la red de Web Semántica permitió reducir de 20.332 arcos a 13.897 arcos y 8.502 arcos, respectivamente.

La detección de miembros claves se realizó contra una comparación de los autores más prominentes del área según las citaciones en *Google Scholar*. Los resultados indican que la mejor recuperación de miembros claves se da en el método de tópicos por LDA con HITS para el primer dataset, para el segundo se da en Keywords, tanto en métricas de *Recall* como en *Precision*.

ANALYSIS OF SCIENTIFIC VIRTUAL COMMUNITIES OF PRACTICE

The different social networks have been emerging as a common and natural sense of humans to gather around a specific subject, with the feeling of belonging to a *Community*. With the actual technology, it is possible to express these feeling through the Internet over online platforms such as blogs, forums, social networks, chats and thousand of others. This online gathering is understand as a *Community*, a network that represent complex relationships between units that changes over time. A *Community* is a group of vertices that share common properties and plays a similar role within the group. These communities can be classified as communities of interest in which members share and discuss a particular interest, and communities of practice, where members share concerns, participate and develop something they do and seek to become better at that. If these interactions occur over online platforms, these are called virtual communities of interest and practice (VCoI - VCoP).

The study of Virtual Communities (VC) not only helps to understand their internal structure, but it also helps to understand and discovery how the knowledge is sharing, who are the key members, providing better tools for the administrators to improve participation ensuring the stability of the community over time. The field of Social Network Analysis (SNA) and Data Mining has address this challenge, but neither of them take into account the meaning of the content that members of a community generate, which may contain useful information and reveal new knowledge about members and their interactions.

Therefore, the main contribution of this thesis is to take into account the semantic of the content created by the members of two Virtual Communities of Practice as well as the structural properties of the networks, to study the existence of other key members, search for the major research topics, and study the properties of the new networks created with content, in a scientific VCoP of the Ubiquitous Computing area and a scientific VCoP of the Web Semantic area.

This work proposes two methods, one is to represent each article wrote by each member by its keywords, and another to extract the underlying topics of each paper with the LDA probabilistic model. With this, new interactions can be built based on the content instead of just relationship of coauthorship (network used as a base to compare the methods proposed). The information extracted of the papers include its Keywords and the Title, abstract and body content to use LDA. The base network was built using the relations of authorship with an author to another, i.e., an author has an edge with another if they wrote together.

The methodology proposed is an hybrid process called SNA-KDD that includes extracting and processing text data to later analysis with SNA to discover new knowledge, using graph theory, ranking algorithms (HITS and PageRank) and different structural metrics for networks.

The results show that scientific study networks can be modeled as Virtual communities of practice using the SNA-KDD methodology using graph theory. This is evidenced by the results of the metric *Modularity*, obtaining values of 0.9 in most networks, indicating a structure of community.

Furthermore, the methods proposed to introduce the content generated by its members, Keywords and topic LDA model, can reduce the density of all networks, eliminating irrelevant relations. In Ubiquitous Computing network with 1,920 nodes, decreased from 5,452 to 1,866 arcs for Keywords method and 2,913 arcs for LDA model; while the network of Semantic Web enabled reduction of 20,332 to 13,897 arcs and 8,502 arcs, respectively.

Detection of key members was conducted against a comparison of the most prominent authors of the area as citations in *Google Scholar*. The results indicate that the best recovery key members is given in the method of topics for LDA with HITS for the first dataset, for the second occurs in Keywords in both metric of *Recall* and *Precision* .

Wisest is she who knows she does not know

Agradecimientos

Cuando el día acaba y todo es oscuridad y silencio, siempre ahí está mi compañero de vida. Gracias a la incondicionalidad, amor y apoyo incommensurable de Nicolás, mi esposo, es que logré superar los numerosos desafíos que me hicieron llegar hasta esta instancia de mi vida.

Quisiera agradecer a mis padres, Sergio y Ruth, por alentarme en mis estudios y tener una palabra de ánimo en los momentos de flaqueza. A mis hermanos, Daniela y Pablo, por ser compañeros en los buenos y malos momentos de la vida, animandome a crecer no solo como estudiante si no como persona.

Más que un jefe y profesor co-guía, quisiera agradecer al profesor Javier Bustos, por su amistad y fe ciega en mis capacidades y trabajo. El desarrollo de esta memoria se debe en gran parte a su apoyo, creatividad y aliento día a día, permitiendo desarrollarme laboralmente en Nic Chile Research Labs.

Quisiera agradecer a mi profesor guía, Sebastián Ríos por su ayuda y guía en la realización de esta memoria, al igual que al miembro de comisión Felipe Aguilera, por sus consejos y disponibilidad para discutir los temas tratados en este trabajo. También a los miembros del Centro de inteligencia de negocios, CEINE, por su ayuda en el seguimiento y organización de esta memoria.

A los amigos que juntos superamos el largo y desafiante camino universitario, Camila, Joaquín, Pablo, Javier y muchos otros. Y a mis queridas amigas, Karina, Camila y Victoria por alegrar cada paso de mi vida.

A todos mis compañeros de trabajo en Nic Chile Research Labs, en especial a Camila, María Grazia, Nabelka y Gabriel por brindarme apoyo, comprensión y una amistad sincera.

Finalmente, quisiera agradecer a todas las mujeres que, anónimamente, forjaron la historia, haciendo posible que yo y tantas otras, pudiésemos entrar en campos tan desafiantes como lo es la ingeniería.

Contents

Agradecimientos	iv
1 Introduction	1
1.1 The community and key-member detection problem	1
1.2 Objectives	3
1.2.1 General Objectives	3
1.2.2 Specific Objectives	4
1.3 Expected Results	4
1.4 Methodology	4
1.5 Thesis Structure	6
2 Related Work	7
2.1 Communities and Social Networks	7
2.1.1 Online Social Networks	7
2.1.2 Virtual Communities	8
2.2 Social Network Analysis	9
2.2.1 Graph Theory to represent Networks	10
2.2.2 Metrics in SNA	12
2.3 Community detection	14
2.3.1 Modularity	14
2.4 Key members in Virtual Communities	15
2.4.1 Discovery Techniques	15
2.5 Keywords Based Text Mining	18
2.6 Topic model: Latent Dirichlet Allocation	18
3 Methodology	21
3.1 Data Selection	21
3.2 Data Cleaning and Preprocessing	22
3.2.1 Stop Words	22
3.2.2 Stemming	22
3.3 Data Reduction	23
3.3.1 Keywords Based Network	23
3.3.2 Topic Based Network	24
3.4 Network Configuration	24
3.4.1 Network Construction	27
3.5 Characterization and Analysis of communities	27

4	Community and Key members discovery on real scientific VCoPs	29
4.1	Ubiquitous Computing	30
4.2	International Conferences on Pervasive and Ubiquitous Computing	31
4.3	The International Semantic Web and European Semantic Web Conference series	32
5	Results and Discussion	36
5.1	Topic Analysis	36
5.1.1	Topic Analysis Ubiquitous and Pervasive Computing Conferences . .	37
5.1.2	Topic Analysis Semantic Web Conference Series	42
5.2	Ubiquitous and Pervasive Computing Conferences Networks	48
5.2.1	Original Network - Ubiquitous and Pervasive Computing Conferences	48
5.2.2	Keywords Based - Ubiquitous and Pervasive Computing Conferences	50
5.2.3	Topic Based - Ubiquitous and Pervasive Computing Conferences . . .	53
5.3	Semantic Web Conference Series Networks	58
5.3.1	Original Network - Semantic Web Conference Series	58
5.3.2	Keywords Based - Semantic Web Conference Series	60
5.3.3	Topic Based - Semantic Web Conference Series	63
5.4	Key Members discovery	68
5.4.1	Key Members discovery in Ubiquitous and Pervasive Computing Conferences	70
5.4.2	Key Members discovery in Semantic Web Conference Series	78
6	Conclusions and Future Work	81
6.1	Future Work	83
	Bibliography	84
	Appendix	88
A	Original Network HITS - PageRank plots Ubiquitous and Pervasive Computing Conferences	88
B	Keywords Based Network HITS - PageRank plots Ubiquitous and Pervasive Computing Conferences	90
C	Topic Based, 15 topics Network HITS - PageRank plots Ubiquitous and Pervasive Computing Conferences	91
D	Key Members discovery in Ubiquitous and Pervasive Computing Conferences for 25 Topics	92
E	Key Members discovery in Ubiquitous and Pervasive Computing Conferences for 50 Topics	94
F	Key Members discovery in Semantic Web Conference Series for Keywords Based network	97
G	Key Members discovery in Semantic Web Conference Series for 15 Topics . .	99
H	Key Members discovery in Semantic Web Conference Series for 25 Topics . .	102
I	Key Members discovery in Semantic Web Conference Series for 50 Topics . .	104
J	Topic Analysis for 25 topics in Ubiquitous and Pervasive Computing Conferences	107
K	Topic Analysis for 50 topics in Ubiquitous and Pervasive Computing Conferences	108
L	Topic Analysis for 25 topics in Semantic Web Conference Series	110
M	Topic Analysis for 50 topics in Semantic Web Conference Series	111

List of Tables

5.1	Display of 10 words for each Topic	37
5.2	Probabilities of words in 15 Topics	37
5.3	Topics grouped under the concept Location Data	38
5.4	Top 10 members by PageRank score over the complete network and Topic based network for all topics and for concept “Location Data”	39
5.5	Manually given names for each Topic	40
5.6	Display of 10 words for each Topic	43
5.7	Probabilities of words in Topics	43
5.8	Topics grouped under the concept “Topics models”	43
5.9	Top 10 members by PageRank score over the complete network and Topic based network for all topics and for topic “Topics Models”	44
5.10	Manually given names for each Topic	45
5.11	Original Network Statistics - Ubiquitous and Pervasive Computing Conferences	48
5.12	Communities and modularity for Original Network	49
5.13	Random graph of same size as original network	50
5.14	Keywords Based Network Statistics - Ubiquitous and Pervasive Computing Conferences	50
5.15	Communities and modularity for Keywords Based Network	51
5.16	Random graph of same size as keyword based network	52
5.17	15 Topic Based Network Statistics - Ubiquitous and Pervasive Computing Conferences	54
5.18	Communities and modularity for 15 Topic Based Network	54
5.19	Random graph of same size as topic based network	55
5.20	25 Topic Based Network Statistics - Ubiquitous and Pervasive Computing Conferences	55
5.21	Communities and modularity for 25 Topic Based Network	56
5.22	50 Topic Based Network Statistics - Ubiquitous and Pervasive Computing Conferences	56
5.23	Communities and modularity for 50 Topic Based Network	56
5.24	Comparison of structural properties for Original network, Keywords Based network and Topic Based network	57
5.25	Original Network Statistics - Semantic Web Conference Series	58
5.26	Communities and modularity for Original Based Network	59
5.27	Random graph of same size as original network	60
5.28	Keywords Based Network Statistics - Semantic Web Conference Series	60
5.29	Communities and modularity for Keywords Based Network	61
5.30	Random graph of same size as keyword based network	62

5.31	15 Topic Based Network Statistics - Semantic Web Conference Series	63
5.32	Communities and modularity for 15 Topic Based Network	64
5.33	Random graph of same size as topic based network	65
5.34	25 Topic Based Network Statistics - Semantic Web Conference Series	65
5.35	Communities and modularity for 25 Topic Based Network	66
5.36	50 Topic Based Network Statistics - Semantic Web Conference Series	66
5.37	Communities and modularity for 50 Topic Based Network	66
5.38	Comparison of structural properties for Original network, Keywords Based network and Topic Based network	67
5.39	Original Network - Ubiquitous and Pervasive Computing Conferences	70
5.40	Keyword Based Network - Ubiquitous and Pervasive Computing Conferences	71
5.41	Topic Based Network - 15 Topics - Ubiquitous and Pervasive Computing Conferences	71
5.42	Key Members discovery recovery measures	72
5.43	Original Network - Semantic Web Conference Series	79
5.44	Keyword Based Network - Semantic Web Conference Series	79
5.45	Topic Based Network - 15 Topics - Semantic Web Conference Series	79
5.46	Key Members discovery recovery measures	80
6.1	Topic Based Network - 25 Topics - Ubiquitous and Pervasive Computing Conferences	92
6.2	Topic Based Network - 50 Topics - Ubiquitous and Pervasive Computing Conferences	94
6.3	Keyword Based Network - Semantic Web Conference Series	97
6.4	Topic Based Network - 15 Topics - Semantic Web Conference Series	99
6.5	Topic Based Network - 25 Topics - Semantic Web Conference Series	102
6.6	Topic Based Network - 50 Topics - Semantic Web Conference Series	104
6.7	Display of 10 words for each 25 Topics	107
6.8	Probabilities of words in 25 Topics	107
6.9	Display of 10 words for each 50 Topics	108
6.10	Probabilities of words in 50 Topics	109
6.11	Display of 10 words for each 25 Topics	110
6.12	Probabilities of words in 25 Topics	110
6.13	Display of 10 words for each 50 Topics	111
6.14	Probabilities of words in 50 Topics	112

List of Figures

1.1	Methodology according to SNA-KDD	5
2.1	Directed Graph	11
2.2	Undirected Graph	11
2.3	Exmample of Adjacency Matrix representation of a graph	11
3.1	SNA-KDD Methodology	21
4.1	Number of Papers per Year in Conferences on Pervasive and Ubiquitous Computing	33
4.2	Number of Papers per Year in Semantic Web Conference series	35
5.1	Social Network Visualization (a) Network with no filter, (b) Topic based network and (c) Topic based network filter with Topic concept “Location Data”	39
5.2	Topic Probability over the years in Ubiquitous and Pervasive Computing Conferences	41
5.3	Topic Probability over the years in Ubiquitous and Pervasive Computing Conferences	42
5.4	Social Network Visualization (a) Network with no filter, (b) Topic based network and (c) Topic based network filter with Topic concept “Topics models”	44
5.5	Topic Probabilities over the years in Semantic Web Conference series	46
5.6	Topic Probability over the years in Semantic Web Conference series	47
5.7	Topic Probability over the years in Semantic Web Conference series	47
5.8	Original Network with average degree display	49
5.9	Keyword Based Network with average degree display	51
5.10	Keywords Cloud based on word occurrences	53
5.11	15 Topic Based Network with average degree display	54
5.12	Ubiquitous and Pervasive Computing Conferences	57
5.13	Original Network with average degree display	59
5.14	Keywords Based Network with average degree display	61
5.15	Keywords Cloud based on word occurrences	63
5.16	15 Topic Based Network with average degree display	64
5.17	Semantic Web Conference Series Networks	67
5.18	Search of an author in Google Scholar	69
5.19	Profile of an author in Google Scholar	69
5.20	Search for investigation topics in Google Scholar	70
5.21	Number of Papers of top 10 authors found by HITS and PageRank for Keyword Based network	73

5.22	Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for Keyword Based network	74
5.23	Number of Papers of top 10 authors found by PageRank compared to first top 10 authors with more papers for Keyword Based network	75
5.24	Number of Papers of top 10 authors found by HITS and PageRank for 15 Topics	76
5.25	Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 15 Topics	77
5.26	Number of Papers of top 10 authors found by PageRank compared to first top 10 authors with more papers for 15 Topics	78
6.1	Histogram of HITS for Original Network	88
6.2	Histogram of PageRank for Original Network	89
6.3	Histogram of HITS for Keyword Based Network	90
6.4	Histogram of PageRank for Keyword Based Network	90
6.5	Histogram of HITS for Topic Based Network	91
6.6	Histogram of PageRank for Topic Based Network	91
6.7	HITS and PageRank indexes across number of authors	92
6.8	Number of Papers of top 10 authors found by HITS and PageRank for 25 Topics	93
6.9	Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 25 Topics	93
6.10	Number of Papers of top 10 authors found by PageRank compared to first top 10 authors with more papers for 25 Topics	94
6.11	HITS and PageRank indexes across number of authors	95
6.12	Number of Papers of top 10 authors found by HITS and PageRank for 50 Topics	95
6.13	Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 50 Topics	96
6.14	Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 50 Topics	96
6.15	HITS and PageRank indexes across number of authors	97
6.16	Number of Papers of top 10 authors found by HITS and PageRank for Keywords Based network	98
6.17	Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for Keywords Based network	98
6.18	Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for Keywords Based network	99
6.19	HITS and PageRank indexes across number of authors	100
6.20	Number of Papers of top 10 authors found by HITS and PageRank for 15 Topics	100
6.21	Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 15 Topics	101
6.22	Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 15 Topics	101
6.23	HITS and PageRank indexes across number of authors	102
6.24	Number of Papers of top 10 authors found by HITS and PageRank for 25 Topics	103
6.25	Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 25 Topics	103
6.26	Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 25 Topics	104

6.27	HITS and PageRank indexes across number of authors	105
6.28	Number of Papers of top 10 authors found by HITS and PageRank for 50 Topics	105
6.29	Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 50 Topics	106
6.30	Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 50 Topics	106

Chapter 1

Introduction

This work aims to study and analysis different Virtual Communities of Practice, specifically scientist networks of collaboration. This chapter purpose is to present the problem that is being address with this thesis, its objectives and its context, followed by a brief explanation of the expected results, the methodology and how this thesis is going to be structured is presented.

1.1 The community and key-member detection problem

The different social networks that have been emerging over the past years are just the online expression of what has always been a common and natural sense of humans to gather around a specific subject, with the feeling of belonging to a *community*.

This new social structures has become complex and bigger networks [39], taking multiple forms. The new services offered on the Internet allow people anywhere in the world to communicate and exchange ideas, interest, information, problems and many others.

A *community* can be understood as a network that represent complex relationships, hard to modelling and changing over time. According to [19], *communities* are groups of vertices which probably share common properties and/or play similar roles within a graph. According to the characteristics and the way of sharing knowledge of a community, they can be classified as *Communities of Interest*, and as *Communities of Practice*.

In [41] a *Community of Practice* (CoP) is defined as groups of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an ongoing basis. Otherwise, a *Community of Interest* (CoI)

can be understood as a group of people that share common ideas about a given passion, but may not know about each other outside the area and its members are quite heterogeneous [18]. The community study has helped to understand how information spreads, how humans and other entities organize themselves and as they create knowledge, how new relationships are created. If a community was created under Web technologies and works through the Web, it is called a *Virtual Community* (VC) [4], and according to its participation, it can be a *Virtual Community of Interest* (VCoI) or a *Virtual Community of Practice* (VCoP).

The underlying structure of a community is one of the main properties that allows to know how their members are connected. As stated in [19], a natural network [7] is not a random graph, i.e., the distribution of edges among its vertices is not homogeneous, revealing a high level of order and organization. This feature is called the community structure.

In [19], the author enunciate the importance of actual community study and their concrete applications, which vary from identify customers in a network of purchases for recommendation systems of online retailers, to cluster large graphs to store data efficiently. The analysis of the structure of a community can be addressed with different approaches such as the classification of vertices according to their position in the network - to detect which of them are more connected and which ones are union nodes between sub communities (sharers of knowledge) - , or study the hierarchical organization of the network (community composed by smaller communities and so on). These analysis has long been carried out with Data mining techniques, using text mining and other approaches [33, 2], and most recently has being combined with Social Network Analysis (SNA) [37], field that take advantage of graph representation to model the interactions between any type of node such as humans, organizations, devices, etc., to search for a better understanding and vision of the community [32, 16, 22, 25, 34, 35].

The organization of the community contents and the links structure let, not only a better usage for their members, but also assures the stability and growth of the community over time. The analysis of Virtual Communities (VC) has become a challenge in the field of Social Network Analysis (SNA), as it enables the discovery of key members, enhances the community administration and reveal useful information of user's behaviour such as sub areas of knowledge. For these tasks, it is relevant the analysis of the data generated in the community and the relationships that their members develop along time. Nevertheless, SNA

provides with techniques for analysing the social structure of a community and Data mining by their own, provides methods for extracting, processing and discover patterns in a data set, applying any of these methods gives a partial vision of any virtual social structure, even a combination of these tools can left out hidden relationships, thereby it may not provide a full insight of the virtual social structure.

Therefore, the main contribution of this thesis is to take into account the semantic of the content created by the members of two Virtual Communities of Practice as well as the structural properties of the networks, to study if adding that type of information, new underlying structures of the communities can be found for a better characterization, in a scientific VCoP of the Ubiquitous Computing area, a sub discipline of Computer Science, and a scientific VCoP of the Web Semantic area. For the representation of those VCoPs, it was considered that a scientific community is composed by the researchers and the set of all the scientific articles written and presented by them in the *International Conferences on Pervasive and Ubiquitous Computing* and *The International Semantic Web and European Semantic Web Conference series*, respectively.

This thesis propose two methods to consider the semantic (contents of the papers) of the articles in order to compare them with the structure of the community given by the relations of authorship of the articles. The first method, Keywords based, uses the underlying concepts of an article represented by their keywords; the second one is a Topic Model that uses all the content of an article. Both methods will be tested with some of the available data of both VCoPs mentioned. To evaluate the quality of this approach SNA metrics such as, HITS and PageRank algorithms, density, degree of the network among others will be used.

1.2 Objectives

1.2.1 General Objectives

The main objective of this work is to study two Virtual communities of practice of a scientific field in terms of the social structure formed by its members when adding the content that they generate in order to identify key members, sub communities and the main topics that each of one research.

1.2.2 Specific Objectives

- To prove whether adding the semantic meaning of the publications of two VCoPs using the methodology SNA-KDD, provide better insights in the knowledge discovery of information in scientific virtual communities of practice.
- To detect communities of two VCoP based on their members and the relations formed by them by different forms: by authorship, by the keywords of their articles and by the topics of research of their works.
- To discover key members (experts) of the VCoPs based on their relationships and the content of their works.
- To extract the topics of research of both fields that the papers of their authors treat about, and to know how they evolved over time.

1.3 Expected Results

- A representation of both VCoP through graphs, considering different forms of relation between their members.
- Communities detected by graph representations and a characterization of them with SNA metrics.
- A comparison of key members findings across the different relationships built by the methodology of this thesis.
- Topic inferring given by the content of the articles that the members of the VCoPs produced, and a graph built by the relation of this topics and the authors.

1.4 Methodology

The methodology of this thesis used a hybrid process called SNA-KDD, proposed by [33], and is based in the Knowledge Discovery in Databases (KDD) of the field of Data Mining, and Social Network Analysis (SNA).

To address the study of the VCoPs, it is necessary to obtain a representation of the inner structure of both networks, in order to filter them later with the addition of the semantics

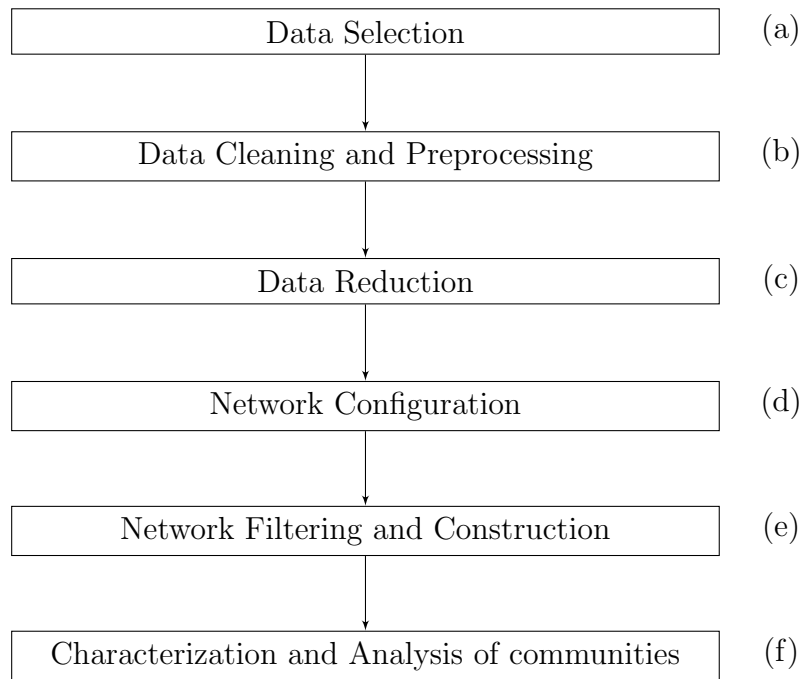


Figure 1.1: Methodology according to SNA-KDD

of the contents forming a reduced network, this is called the *original network*. Firstly, it is necessary to extract and select the data (a) of both VCoP, taking into account the researches names, the entire articles (PDFs) and the year and conference of publication. These datasets need to be cleaned from lost data and unified under a unique format. Also, for the pre-process and data reduction it is necessary to apply text mining techniques to filter words such as numbers and articles (Stop Words) and to transform words conjugated into their root word (Stemming) (b).

The next step (c) is where the techniques of adding semantics are applied to the datasets. One technique requires the keywords of all articles to represent each paper, while the other technique applies LDA model to the set of all papers to represent each paper with a set of topics.

The part of Network Configuration (d) corresponds to the way the community is going to be represented as a graph, the type of nodes and the interactions between them according to the members participation.

In the network construction stage (e), three types of graphs are going to be made with the nodes defined previously and the interactions determined by: (1) the authorship of each paper, (2) the keywords of each papers and (3) the topics of each paper; this for both dataset.

Finally, with the graphs constructed, SNA metrics are going to be calculated to analyse the social structure of each community, a detection of key members and a characterization of and evolution of the topics of research found with the addition of semantics (f).

1.5 Thesis Structure

In chapter 2 a revision of the related work for this thesis is presented. A brief review on SNA techniques and metrics commonly used in literature, the problem of Community detection and how to discover and measures communities and what and how a key member can be found in a community with techniques likes HITS and PageRank algorithms. Also, is presented the theory behind the two methods proposed to add semantics in this thesis, TF-IDF from text mining and LDA topic model.

In chapter 3 is shown the methodology used in this thesis, explaining how the hybrid approach of KDD-SNA is used. As well, a detail explanation of how the networks of both VCoPs are designed and constructed adding the semantics of the content that the authors (members) produced with the two methods proposed for that.

In chapter 4, an description of the communities in study is presented, specially an explanation of the domain of research and how the datasets were formed for the realization of this thesis.

In chapter 5, the results of building the 3 types of networks proposed by the methodology are shown for both VCoP in study. As well, the study and analysis of topics and key member detection are discussed.

In chapter 6 the conclusions of this work can be found along with the future work that the analysis of VCoP offers.

Chapter 2

Related Work

2.1 Communities and Social Networks

In the last years, the use of computers has changed dramatically. With the appearance of Internet in the 90s and its massiveness in the past decade, has led to an explosive creation and growing of new services in the Internet. This has allowed people to communicate and share knowledge like never before, breaking down geographical barriers.

The Internet has allowed new collective activity for people with the emergence of new social institutions with specific characteristics [31]. With this recent development of diverse technology, new forms of connecting people have surfaced around these social structures.

These social structures can take the form of *social networks*, *virtual communities*, *virtual communities of practice*, *virtual communities of interest*, etc.

This collective activity of social structures has enabled the appearance of new social technologies in the Internet, such as forums, blogs, social networks, live chats, messaging services, picture sharing networks, as well as, platforms for the organizations of multiple activities, such as conferences, ONGs, companies and new entrepreneurs.

2.1.1 Online Social Networks

Nowadays, the way of people to work, share information, entertainment, etc., has changed dramatically. This has been possible by the formation of social networks.

According to [31] a social network is a set of social institutions (people, groups, organizations, etc.) associated with any social character, such as friendship, co-working or information exchange. These relationships form a network structure.

A group of people who interact online, implicitly build a social network, though not be

defined explicitly [10].

2.1.2 Virtual Communities

According to many researchers [31], a community can be define as a group of people sharing some common interests, experiences and / or needs, which are linked between each other through social relations, where important resources can be found for interested members, creating a sense of identity. The members of a community usually develop strong interpersonal feelings of belonging and mutual need.

Virtual communities are the type of communities created, maintained and accessed through Internet. These can be seen as a social network, given that the network of computers connecting their members could be represented as a set of links with social meaning.

There are different types of communities studied depending on the nature of the interactions between its members:

Communities of Interest [24]: are those communities in which members share the same interest in any topic (and therefore they all have a common background). Examples of this type of community: music bands fan club, groups of people interested in the planets of the solar system, groups of people interested on environment, among many other.

Communities of Practice [40]: are those in which groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly. The Gephi software community, the Open Source community and communities inside companies, among others, are examples of communities of practice.

Communities of Purpose [12]: are those where it members share the same objective. The buyers of a virtual library sharing the goal of finding and buying a book is an example of a community of purpose. The functional purpose that connects the members to the community is disbanded once the goal is reached.

Virtual Communities of Practice (VCoP)

According to [40], *communities of practice are formed by people who engage in a process of collective learning in a shared domain of human endeavor: a tribe learning to survive, a band of artists seeking new forms of expression, a group of engineers working on similar problems, a clique of pupils defining their identity in the school, a network of surgeons exploring novel*

techniques, a gathering of first-time managers helping each other cope. And states that community of practice is characterized by three essential components:

- *The Domain*: comprehends the specific area of interest that members of a community share and are identified for. Members of a community are committed to the domain and have competences that distinguish them from people outside the community.
- *The Community*: the interest that the members of a community shows for the domain, move them to collaborate, engage and join in different activities, discussions, asking for help and share information. In this process, relationships are form enabling them to feel part of a group, the community.
- *The Practice*: which distinguish a community of practice from others communities are that members are not only interested in the domain, but they are also practitioners. Since they become part of the community they develop resources, experiences, tools in order to enhance their ability to perform the task related to the domain.

The communities that will be analysed in this thesis, are both virtual communities of practice, since they fulfil the three mentioned conditions of a VCoP:

- *Domain*: each community has a specific domain. In particular, for the first dataset, the domain is Ubiquitous and Pervasive computing; for the second one, the domain is the Semantic Web.
- *Community*: for both communities, its members are researchers that work together in academic research within the context of specific conferences.
- *Practice*: members of the community actively engage in working on different lines of research within the same domain, publishing papers, presenting their work in conferences, collaborating with other researchers, improving methodologies, and using related work from others members.

2.2 Social Network Analysis

The social structures mentioned in the previous section can be studied and modeled as networks with techniques from the area of Social Network Analysis, since individuals forms social relations in this types of structures.

The field of Social Network Analysis has experienced a huge growth from the past decade. Nevertheless this field has been used since the mid-1930s as a research topic in social and behavioural sciences. The early researchers started to develop social theory with formal mathematics, statistical and computing and later to adopt techniques of graph theory, clustering and grouping [13].

This multidisciplinary field is based on the importance of representing the relationships among interacting units as a network, following the conviction that the units are interdependent entities (not autonomous) and can be of any nature: humans, animals, web pages, institutions, proteins, etc. These units form complex connections forming a network, this ties also can be of any nature (behavioural interactions, economic ties, affective evaluation, flow of resources, customers purchases, ideas, etc.), this perspective of network is the base of the analysis in SNA [37].

2.2.1 Graph Theory to represent Networks

To analyse a network based on the concepts of SNA graph theory is not only a useful technique, but also necessary to represent the concept of relationship between units and their structural properties. Graph can be extended almost to any discipline, consisting on nodes and arcs. Nodes (or vertex) are objects that are connected by links called arcs or edges. According to the direction of the arcs, a graph can be directed or undirected. In the first type of graph, the links have one order, i.e., it is not the same as an arc from node i to node j than vice versa $(i, j) \neq (j, i)$. In the second type, the links are defined by the connected nodes and the order is not relevant, $(i, j) = (j, i)$. Mathematically, a node v is defined as an item of a set $V = \{1 \dots n\}$. The arcs e , for a directed graph, are defined as ordered pairs of nodes belonging to the set of arcs $E \subset V \times V$. The graph G is completely defined by the form $G = (V, E)$.

A graph $G = (V, E)$ can be represented by his *Adjacency matrix*, A . Let $V = \{1, \dots, i, j, \dots, n\}$ be ordered, A is the $n \times n$ matrix, where for a pair of nodes i and j , $A_{ij} = 1$ if they are connected by an arc, $A_{ij} = 0$, otherwise.

Some notation used in graph theory is helpful for SNA [15]:

1. Subgraph

If a graph $G1 = (V1, E1)$ and another graph $G2 = (V2, E2)$, $G1$ is subgraph of $G2$,

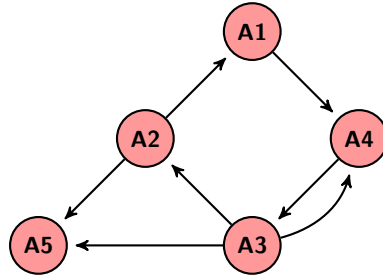


Figure 2.1: Directed Graph

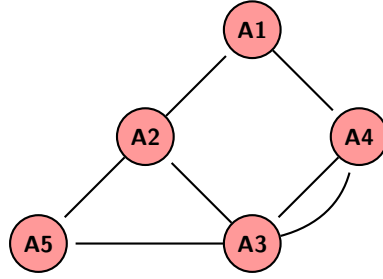
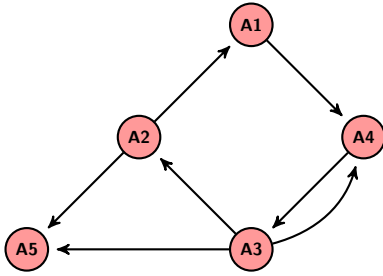


Figure 2.2: Undirected Graph



$$A = \begin{matrix} & \begin{matrix} A1 & A2 & A3 & A4 & A5 \end{matrix} \\ \begin{matrix} A1 \\ A2 \\ A3 \\ A4 \\ A5 \end{matrix} & \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

Adjacency Matrix representation

Figure 2.3: Exmample of Adjacency Matrix representation of a graph

denoted as $G1 \subseteq G2$ then:

- $V1 \subseteq V2$
- $E1 = E2 \cap (V1 \times V1)$

2. Adjacency Nodes

A node $i \in E$ is adjacent to a node $j \in E$ if an arc (i, j) exists between them.

3. Neighbourhood

Given a graph $G = (V, E)$, let $i \in V$ be a node, the neighbourhood of i is the set:

$$N_G(i) = \{j \in G \mid (i, j) \in G\}$$

4. Order

The order of a graph $G = (V, E)$ is the number of vertices, v_G :

$$v_G = |V_G|$$

5. Size

The size of a graph $G = (V, E)$ is the number of edges $|E_G|$.

6. Walks

For a graph G if $e_k = i_k i_{k+1} \in G$ for $k \in [1, n]$, the sequence $W = e_1 e_2 \dots e_n$ is a walk of length n from i_1 to i_{n+1} (i_k is adjacent to $i_{k+1} \forall i \in [1, n - 1]$).

Graph theory is useful for SNA, not only for studying simple graphs, but also for the study of structures with multiple interactions, permitting multiple arcs to connect with an edge pair (many-to-many instead of one-to-one), graph theory called this structures *hypergraphs*. A real case of this type of graphs is the one produced by the comments on the social network, Facebook. A hypergraph can capture the complexity of certain networks, specially for those cases richness in social interactions, but in practice the most common representation to model any type of networks are the simple graphs, because they are generally simpler and easier to construct and analyze [7].

2.2.2 Metrics in SNA

Supporting the analysis of networks, graph theory is the base for constructing measures within a graph to determine node and edges properties.

Degree

Given a graph $G = (V, E)$, the degree of a node is the number of its *neighbourhood*, i.e., the number of edges incident to a vertex:

$$d_G(i) = |N_G(i)| = \sum_j A_{ij}$$

If $d_G(i) = 0$, then i is an **isolated** node in G , instead if $d_G(i) = 1$, i is a **leaf** of the graph. The **minimum degree** and the **maximum degree** of G are defined as:

$$\delta(G) = \min\{d_G(i) \mid i \in G\}$$

and

$$\Delta(G) = \max\{d_G(i) \mid i \in G\}$$

For a simple graph it applies that:

$$0 \leq d_G(i) \leq n - 1$$

for a n number of nodes.

For a directed graph, the *Average Degree* is $\overline{d_G} = \frac{d_G}{n}$, and for an undirected graph $\overline{d_G} = \frac{2d_G}{n}$.

Though this is a basic metric of SNA, this is one of the most used metrics for network analysis.

Density

The idea behind the concept of the density of a graph is to know its size according to the arcs presents in comparison to all the possible arcs between all the nodes in the graph.

If the number of nodes in a graph $G = (V, E)$ is n , the maximum possible arcs between all the nodes would be:

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

The density of a graph is the proportion of the arcs actually present in the graph to the maximum possible $\binom{n}{2}$, defined as follows [37]:

$$\delta = \frac{|E|}{n(n-1)/2} = \frac{2|E|}{n(n-1)}$$

This metrics can take values in the range $[0, 1]$, where a value of 0 means there are absolutely no edges present, and 1 if all the possible arcs are present, if that is the case, the graph is so-called a *Complete graph*. This measurement captures the degree of connectivity of nodes in the graph. Irrelevant connections makes more difficult to cluster and detect sub communities within the graph. Therefore, it would be desirable to delete irrelevant arcs which would decrease the density of the graph.

2.3 Community detection

In a social network, is commonly the appearance of densely connected groups of vertices, with only sparser connections between the different groups. The ability to detect such groups - called communities- could be of relevance in the study of a network, revealing new ties and nodes groups [28].

The study of community structure in networks has long been studied for multiple disciplines, and it has been closely related to graph partitioning ideas from graph theory and computer science and to hierarchical clustering from sociology [29]. Many algorithms have been proposed in the literature to the optimization problem that involves partitioning the graph because precise formulations are known computationally intractable. This algorithms goes from: *divisive algorithms* that are able to detect inter-community links and remove them later from the network, *agglomerative algorithms* that merge similar nodes/communities in a recursively way and *optimization methods* that are based on the maximisation of an objective function [9].

2.3.1 Modularity

In order to detect the quality of the community structure, i.e., the quality of the partitions obtained by the community detection algorithm, a metric called **Modularity** is often used and is calculated as defined by Newman [26]:

$$Q = \frac{1}{2m} \sum_c \sum_{i,j \in c} (A_{ij} - \frac{k_i k_j}{2m})$$

where A_{ij} is the adjacency weighted matrix between node i and j , $k_i = \sum_j A_{ij}$ is the sum of the weights of all vertex coming in or out of the node i , c_i is the community to which arc i is assigned to, and $m = \frac{1}{2} \sum_{ij} A_{ij}$.

This metric is a scalar value in the range $[-1, 1]$ and measures the density of vertex inside communities as compared to vertex intra communities. This means that modularity can be understood as the number of edges occurrence within groups minus the expected number in an identical network with edges placed at random [28]. The values of modularity can be positive or negative, yet according to Newman, positives values of modularity indicate that the network has a possible presence of community, therefore the author suggest to look for

divisions with high modularity, as it indicates a network with many possibilities of having well defined and clear partitions forming communities.

In this thesis the algorithm used to find communities in both VCoPs is the one proposed by Blondel et al.[9], which is a heuristic method based on modularity optimization, implemented by the software Gephi [5], as well as the modularity.

2.4 Key members in Virtual Communities

Key members plays a fundamental role in the construction and development of any community, as they not only produces many of the contributions, but also encourage other members to participate, share knowledge, bring up new members and can moderate the different types of activities inside the community.

As this thesis aims to add the content of the works produced by the community, discovery techniques will be applied over the arcs that finds the different methods proposed in the methodology over the VCoPs in study.

2.4.1 Discovery Techniques

In the literature of SNA there are techniques to discover what members of a community are the most influential based in the idea of measure each member interaction with a visualization of the graphical representation of the community. Because key members are are at the heart of the community, it is common to apply core algorithms, like HITS and Pagerank to rank members according to its importance given by its connections to others to obtain key members based in their respective topics of interest.

HITS

In the area of Web search problem, Kleinberg proposed an algorithm to rank pages relevant to a search topic using the structure of links of each one of them, this algorithm called HITS (Hyperlink-induced topic search). In [21] Kleinberg describes his algorithm as a tool for extracting information effectively in an web environment, based on the discovery of *authoritative* information sources on any search topic.

The problem that tries to address this algorithm, is that for a search made in the web by a human, a text based ranking system of the web pages referring to that topic results on a

huge number of pages (that may not even be as relevant to the topic) causing a big problem of sorting them, not even ensuring useful information for the person.

Kleinberg raises a different ranking system for web pages as the result of a query in the web based on a classification of them, as *authoritative* or *hubs* pages. A page is an *authority* if it contains relevant and valuable information for an specific topic, i.e, if the creator of a page p include a link to a page q in p , he has “conferred” authority on q (excluding links for navigational purposes such as “Return to main menu”). While a page can be a *hub* if it advertise an *authoritative* page, i.e., a page that contain useful links towards the *authoritative* pages, helping to the search engine to point in the right direction.

HITS classifies which pages are goods *authoritative* pages and which are good *hubs*. This is done by assigning an *authoritative* weight and a *hub* weight to every page, depending on how many times a page is pointed by and how many times a page points to.

With this classification, Kleinberg construct focused direct subgraphs of the web, $G = (V, E)$ where the nodes correspond to the pages, and an edge $(p, q) \in E$ is a link from page p to q . This graphs takes in consideration the weights for the arcs as stated above. The algorithm starts with a “root” subgraph containing pages with a high occurrences of the search words, then another subgraph is constructed with all edges that comes out and in of the “root” subgraph, this is called “seed” subgraph and probably have a lot of *authoritative* pages to the topic. HITS dynamically update the weights of the “seed” subgraph based on that a good hub increases the *authority* weights of the pages that it points, and a good *authority* increases the weights of the pages that point to it.

By the way this algorithm takes graph theory to model web pages and its connections, it has been useful for SNA in discovery of key members. As stated in [21], HITS can be used in social networks, as with the notion of *authority* one can measure the use of link structure in standing, impact and influence of members in a community. In this thesis, an implementation of HITS algorithm, by the software Gephi [5], will be used to rank authors based on his *authority* in the community network, using the weights of the arcs, depending on how many times an author worked with another author.

PageRank

PageRank is the algorithm proposed by Sergey Brin and Lawrence Page [11] and it is the actual algorithm that the web search engine, Google, utilizes for its query search. This

algorithm was created near 1998, and follows the same basic ideas of the algorithm HITS of using the link structure of the web graph to make a system of ranking of every web page related to a certain search topic. The difference is that HITS use the “seed” subgraph for every query that changes with a different search, hence the ranking of those collection of nodes (web pages) also change.

Unlike HITS, PageRank use the web link structure as an entire directed graph, an overcome the problem of connectivity of the network since, by the heterogeneous nature of the web, a lot of web pages are merely descriptive and does not contains links to another web pages, making the graph unconnected.

The algorithm model a web surf considering that a user can navigate through a page following its links, but suddenly change to another page that wasn’t pointed by the previous page with a small, but positive probability. To recover this behaviour, Brin and Page add a constant p to the *PageRank Matrix*, M , as it follows:

$$M = A(1 - p) + Bp$$

where p is the mentioned probability constant named *damping factor*. A is the adjacency weighted matrix of the graph, and $B = \frac{1}{n} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}$ for a n number of web pages. Therefore, each page have a $\frac{1}{n}$ probabilities of being chosen.

Then, the algorithm follows the task to compute the PageRank vector, that is the unique eigenvector corresponding to the eigenvalue 1 of the *PageRank Matrix* until it converges, resulting in a ranking of web pages.

As one can imagine, this ranking system have been used in SNA to sort members communities by the way the algorithm treats the web network like a graph with interactions, similarly to a community network. In this thesis, an implementation of this algorithm by the software Gephi [5] will be used to discover new key members according to its ranking in the graph formed by the community in each method.

2.5 Keywords Based Text Mining

One of the main objectives of this thesis includes the addition of semantic of the content of publications, specifically, one technique includes the extraction of keywords of every paper in each VCoP. Unfortunately, there are papers within each dataset that doesn't show keywords. Thus, it is necessary to represent this sub-set of papers with no keywords, with some sort of words collection that can simulate keywords, i.e., words that represent the content of the paper.

To accomplish that all papers have keywords, it is proposed to use a technique of text mining, called *TF-IDF* over the title and abstract of the articles with no keywords, in which *TF* stands for *term frequency* and *IDF* for *inverse document frequency*, this is defined by [36]:

$$TF - IDF = m_{ij} = \frac{n_{ij}}{\sum_k^{|\nu|} n_{kj}} \times \log\left(\frac{|\mathcal{C}|}{n_i}\right) \quad (2.1)$$

where ν is the vector of words given by the entire set of words that forms the vocabulary of all the papers. A word $w \in \{1, \dots, |\nu|\}$ is the basic unit of the sequence of words S formed by all the text in the title plus the abstract such that $\mathbf{w} = (w_1, \dots, w_S)$. Therefore, we have a corpus compose by a collection of P "short" papers as $\mathcal{C} = (\mathbf{w}_1, \dots, \mathbf{w}_{|P|})$.

With this, the resulting would be the most representative words of a paper, each with a score given by *TF-IDF*, to later choose those ones that overcome certain threshold δ . With this procedure, an entire corpus of keywords of all papers in the datasets is obtained, and every paper can be represented as a vector from the vector of vocabulary, ν .

To make possible the network configuration, the set of keywords of each paper is compared with the set of keywords of every other paper, with a measure of similarity (such as the *Cosine Similarity*), and then determined if a paper is similar to another in terms of its contents, thereby an edge can be establish between the creators of those publications.

2.6 Topic model: Latent Dirichlet Allocation

The probabilistic model, called Latent Dirichlet Allocation (LDA) is a topic model based on a Bayesian model proposed by [8]. This model allow to reduce the dimensionality of the content in a document by modelling the words in it by multiple topics or concepts inferred

from the same text. LDA is an unsupervised learning method that is based on the idea that a document can be considered as a mixture of topics, which for the model, are latent variables *hidden* in the text, and are modeled as a probability distribution over the set of words, giving as a result topics which can be latter interpreted and named depending on the probabilities that a topic presents to generate a specific word given the documents. This model assumes that each document, from a set of documents, exhibits topics with different proportions. For example, the topic *Economy* can generate words like *growth*, *inflation*, *capital*, *stocks*, where a document in particular can have a 40% of their words that correspond to *Economy*, a 20% to *Sociology*, another 20% to *Education* and a 10% to *Politics*.

Each document is transformed from the set of words S to a set of topics τ . A document is a composition formed by a collection of words ($\mathbf{w} = w^1, \dots, w^S$), the model then, tries to determine the probability distribution $p(\theta, z, \mathbf{w}|\alpha, \beta)$ from τ to generate \mathbf{w} . Then, the corpus of documents D can be represented as a convex distribution of topics that best suits the words of the text composed by $w \in \{1, \dots, |\nu|\}$, represented by the vocabulary ν .

$$p(\theta, z, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{s=1}^S p(z_s|\theta)p(w^s|z_s, \beta) \quad (2.2)$$

The model is described as follows: first, θ is the distribution of topics for each document, i.e., for the s 'th word in document, the topic for that word realization z_s is drawn using the distribution θ . On the other hand, $\vec{\phi}_z = p(w^s|z_s, \beta)$ is the distribution of words given the topic z , i.e., if the topic of the s word to be generated is z ($z_s = z$), then the word is draw using the probabilities $\vec{\phi}_z$. Finally, α and β are prior parameters of the Dirichlet distributions of the model, and captures the prior distribution of words and topics.

Integrating over the random variable θ and summing over topics z the equation (2.2), the marginal distribution of a document can be deduced:

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{s=1}^S \sum_{z_s \in \tau} p(z_s|\theta)p(w^s|z_s, \beta) \right) d\theta \quad (2.3)$$

To address the final goal of LDA of obtain the probability of a corpus of documents, i.e., the likelihood of the model, is necessary to take the product of the marginal probabilities of a document (2.3), for all documents:

$$p(D|\alpha, \beta) = \prod_{d=1}^M \int p(\theta_d|\alpha) \left(\prod_{s=1}^{S_d} \sum_{z_{ds} \in \tau} p(z_{ds}|\theta_d) p(w^{ds}|z_{ds}, \beta) \right) d\theta_d \quad (2.4)$$

In LDA, the distribution of words given a topic are modeled as a multinomial variable, as each topic within a document. To estimate the model, it is assumed that the parameters of each of these distributions, θ and $\vec{\phi}_z$, follow a Dirichlet distribution of parameters α and β , respectively. The parameters of the model can be drawn from the posterior distribution using Bayesian methods, in particular Gibbs sampling methods, besides optimization approaches.

In this thesis, a variation of LDA model developed by Hoffman et al. in [20] will be used, implemented in the library *Gensim* developed by Radim Rehurek [30]. This implementation is an online variational Bayes (VB) algorithm for LDA based on stochastic optimization for large corpus of documents, which approximates the posterior as well as traditional LDA batch VB algorithm.

Chapter 3

Methodology

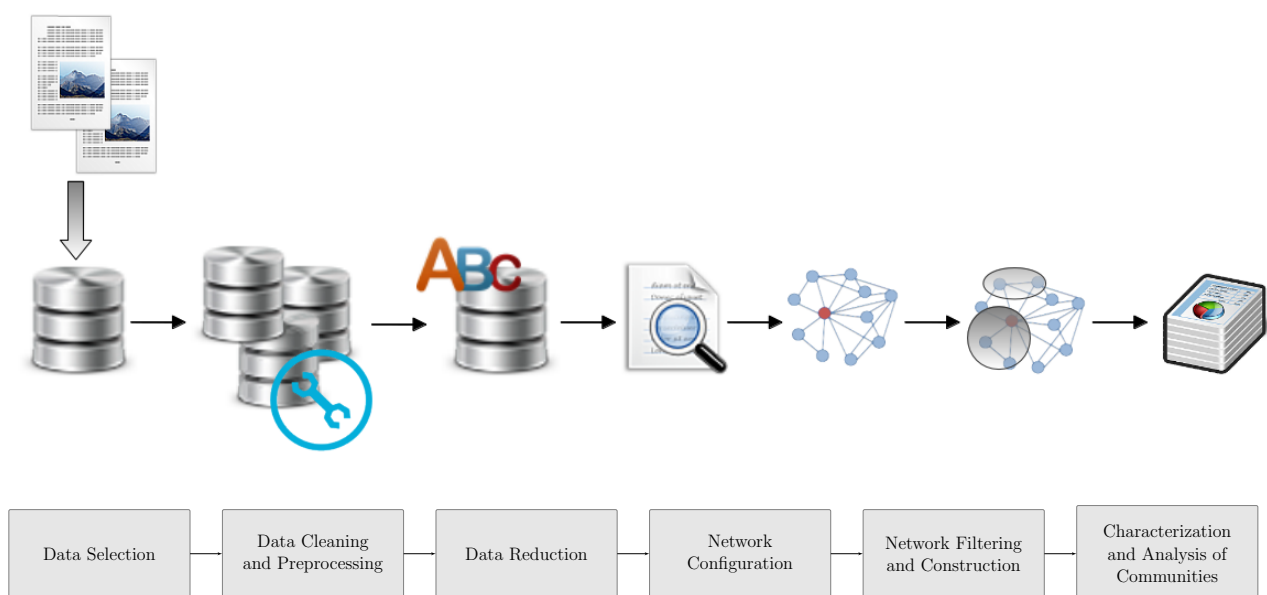


Figure 3.1: SNA-KDD Methodology

As mentioned in Chapter 1, the methodology proposed is an approach developed by Ríos & Aguilera in [33], called *SNA-KDD methodology*. This consist in a process that take advantage of the Knowledge Discovery in Database (KDD) process of Data mining field, to mix it with Social Network Analysis techniques (SNA). The first steps of the methodology are the traditional way from extracting data to process it and then analyze it with SNA techniques.

3.1 Data Selection

In this first stage, the necessary data is gathered for both VCoPs. As mentioned before, the content of papers and name of authors will structure the datasets for this works.

For the International Conferences on Pervasive and Ubiquitous Computing series, papers of the proceedings of every conference consider in the series will be obtained in PDF format. Likewise, for the Semantic Web Conference series the papers will be extracted from every conference proceeding. In addition, the Semantic Web Conference Corpus, <http://data.semanticweb.org/> provides RDF files that contains the metadata of each conference that are going to be query to extract authors names and year of the paper publication. All this data is transformed to plain text and store in a database.

3.2 Data Cleaning and Preprocessing

In order to consider the semantic meaning of a text, simple measures of cleaning and preprocessing needs to be apply to the dataset. Data cleaning consider to remove all of the sections of a paper that wouldn't be used for the analysis, such as the References part, figures and tables.

Later, in order to improve the quality of the text analysis two techniques of text mining will be used: Stop Words and Stemming.

3.2.1 Stop Words

In a document there are many different words proper of the human language such as adjectives, articles, pronouns, substantives and others. A Stop Word is a functional word that allows users to form its written speech and comes in the form of articles and conjunctions, such as *is*, *you*, *we*, *in*, *a*, also numbers and expressions like mathematical characters. For text analysis these type of words must be filtered out, mainly because they don't contribute to the meaning of the context in the speech.

In this thesis, a common list of Stop Words is used to remove them from the vocabulary of the entire corpus of both datasets.

3.2.2 Stemming

Stemming is a text analysis technique that helps to transform words to the stem of the word, i.e., the "root" of the word. An example of this technique would be the words *eating* and *eats* taken to *eat*. This facilitate the work for the analysis since any computer program would consider them as three different words when in fact are just derived words from one word.

This allows to reduce the type of words making easy to map words and to determine the similarity of a text with another.

To remove the suffix of all the words, this thesis considers the Porter Stemmer suffix stripping algorithm for English implemented by NLTK initiative of Natural Language Tool Kit for the programming language Python, <http://www.nltk.org/>.

3.3 Data Reduction

3.3.1 Keywords Based Network

This method incorporates the content of an article to the analysis through the keywords that every paper naturally has. Since a keyword is a word that tries truthfully to represent in the best way the content of a document, a set of keywords should, completely represent it. Since in the scientific articles is commonly used these type of representation, this thesis propose to use the collection of keywords of every article published in the VCoPs as a measure to determine whether a paper is similar to another (with Cosine Similarity), allowing to determined then, if an edge must exist between paper's authors. This method, is a new way of designing the scientist networks in study, with semantic content.

For the cases in which a paper don't have keywords, a text mining technique has to be used, *TF-IDF*:

$$TF - IDF = m_{ij} = \frac{n_{ij}}{\sum_k^{|\nu|} n_{kj}} \times \log\left(\frac{|\mathcal{C}|}{n_i}\right) \quad (3.1)$$

This technique is applied to the title and abstract section of each paper without keywords, obtaining a list of words with a score from $[0, 1]$. All words with a score over 0.15 will be considered as a good representation, letting each paper to has at least 3 or 4 words.

Later, when every paper has a small set of words representing it (3 to 8 words), another reduction of dimensionality is needed, given that, the vocabulary formed by the keywords for the entire corpus is still too sparse to form real similarities between one paper to another. To resolve this, another method will be used, apply a reduction of the vocabulary based in synonyms.

An open web thesaurus of words will be used to find synonyms of every keyword in the vocabulary with the idea to replace those words that has the same synonyms with one of the

words. This will be done using The Big Huge Thesaurus API <http://words.bighugelabs.com/api.php> and then making a search and match of words, leaving a small collection of words for each paper.

3.3.2 Topic Based Network

Another way to consider the semantic meaning of the text of every paper is with a reduction of the content into topics that can represent each article, i.e., instead of determine whether a paper is similar to another based on all of its words, the similarity is going to be calculated based in the topics that represents each publication.

A topic is a collection of most probable words that constitute one specific subject. A document, such as a scientific article, can be represented by a mixture of topics. A probabilistic model is the most suitable way to accomplish these, LDA is a Bayesian model that consider topics as a latent variable occult in the document and modeled them as a probability distribution over the set of words. These can be interpreted at posterior as the probability of a topic to generate a word given the set of documents.

This model transform each document from the set of words S that is compound, to a set of topics τ . A document is a composition formed by a collection of words ($\mathbf{w} = w^1, \dots, w^S$), the model tries to compute the probability distribution $p(\theta, z, \mathbf{w}|\alpha, \beta)$ from τ to generate \mathbf{w} .

$$p(\theta, z, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{s=1}^S p(z_s|\theta)p(w^s|z_s, \beta) \quad (3.2)$$

Then, the corpus of entire documents to analyse D can be represented as a convex distribution of topics that best suits the words of the text composed by $w \in \{1, \dots, |\nu|\}$, represented by the vocabulary ν .

With this model, each paper in both dataset will be represented by a couple of topics, allowing to compare them to each other calculating a similarity measure to, later, build an edge with the author of a paper that, in similarity, is close to another paper of another author.

3.4 Network Configuration

To detect the sub communities among researchers of an specific topic, the procedure involves modeling the papers and authors data into a network which contains the information of both

interactions (co-authorship) and semantics (contents of their papers).

Interactions, such as co-authorship, provide information about relationships among authors. If two authors work together frequently, any reasonable sub-community detection technique should classify them in the same sub-community with high probability. However, interactions could also occur rarely between two authors, and lead to create links between those authors when they have only worked together for singular and specific projects. A better community detection technique should incorporate also information about their research content and identify if those authors have academic interests in common. Given that, this methodology also incorporates semantic information into the network. This information is provided by keywords of each author's publications, or from the full text information contained on them.

In order to build this network two steps are required. First, text data from publications (e.g. keywords, abstract, full text) has to be extracted and transformed into a more suitable and useful data, in the same format. In this work, two procedures for transforming text into numerical data are considered. The first, Keyword Based Network Configuration, uses a *TF-IDF* transformation of keywords of publications, which are previously reduced to a smaller set of words using synonyms relationships. The second, Topic Based Network Configuration, uses LDA technique to capture the main topics on the full text of the papers. Both methods transform each publication text data into a numerical vector. After each publication text data is transformed into numerical representations, similarity measure between authors are computed.

Second, the similarity measure, among co-authorship information, will be used to construct the network of authors as nodes and their relationships as edges. The relationships account for both co-authorship and similarity of research interests.

Finally, some SNA techniques are performed over the generated network. In addition, both networks constructed using Keyword Based and Topic Based Network Construction techniques will be compared to a network generated using only co-authorship information, i.e., without incorporating semantic information.

This section describes the methodology used for each one of the two steps previously mentioned.

Similarity

The purpose of this step is to characterize the strength of the relationship between two authors, i and j , by setting the weight of the edge between them a_{ij} using how similar are the papers they have published.

First, let \vec{g}_p be the the numerical representation of paper p . The numerical representation corresponds to *TF-IDF* for Keyword Based and to the Document-Topic distribution Φ_d for Topic Based Network Configuration. In both cases, characterize numerically the text information used. Once the text information is reduced to a numerical representation, similarities measures between authors can be computed using similarity between papers numerical representations. As the text is written at paper-level, the similarity between text data have to be computed from similarity between papers. Let S_{pq} be the similarity index between papers p and q , defined as follows:

$$S_{pq} = \frac{\vec{g}_p \cdot \vec{g}_q}{\|\vec{g}_p\| \|\vec{g}_q\|} = \frac{\sum_{l=1}^n g_{lp} g_{lq}}{\sqrt{\left(\sum_{l=1}^n g_{lp}^2\right) \left(\sum_{l=1}^n g_{lq}^2\right)}} \quad (3.3)$$

where $x \cdot y$ is the dot product and $\|\cdot\|$ the Euclidean norm. Note that, given that all $g_{ip} \geq 0$, then that means that $S_{pq} = 1$ if the papers have the same text, and $S_{pq} = 0$ if there is absolutely no similarity between those papers.

However, the previous equation describes similarities among papers, not between authors. In order to compute authors similarities using the papers similarities, only those interactions higher than a certain threshold θ are going to be considered. Let \mathcal{P}_i and \mathcal{P}_j the sets of papers for authors i and j , respectively. Then, given a threshold parameter θ , the weight of arc $i - j$, a_{ij} is:

$$a_{ij} = \sum_{p \in \mathcal{P}_i} \sum_{q \in \mathcal{P}_j} S_{pq} \cdot \mathbf{1}\{S_{pq} \geq \theta\} \quad (3.4)$$

where $\mathbf{1}\{S_{pq} \geq \theta\}$ is 1 if $S_{pq} \geq \theta$, and 0 otherwise. In other words, a_{ij} is the sum of the similarities of all papers of i with all papers of j that are relevant (greater or equal than θ).

3.4.1 Network Construction

The purpose of this step is to create a graph $G = (V, E)$ where V is the set of authors, and E is the set of relationships between authors. The following statement is used to define the set of edges E . Let i and j be two authors in V and δ be a threshold parameter, then the arc $(i, j) \in E$ if and only if two properties hold:

1. Authors i and j are coauthors of at least one paper, i.e., $\exists p$ such that $p \in \mathcal{P}_i$ and $p \in \mathcal{P}_j$.
2. The similarity between authors i and j is at least above a threshold δ , i.e., $a_{ij} \geq \delta$.

In addition, the weight of an edge (i, j) is $w_{ij} = \mathbf{1}\{a_{ij} \geq \delta\} \cdot a_{ij}$

Given this condition, Algorithm 1 describes the steps to create the network.

Algorithm 1 Network Construction

```
1: Set  $E \leftarrow \emptyset$ 
2: Set  $w_{ij} = 0 \forall i, j \in V$ 
3: for each  $i \leftarrow 1, \dots, |V|$  do
4:   for each  $j \leftarrow i + 1, \dots, |V|$  do
5:     for each  $p \in \mathcal{P}_i$  do
6:       if  $p \in \mathcal{P}_j$  then
7:         if  $a_{ij} \geq \delta$  then
8:           Set the weight  $w_{ij} = a_{ij}$ 
9:           Add the arc  $(i, j) \rightarrow E$ 
10:        go to next  $j$  in step 4
11:       else
12:         go to next  $j$  in step 4
13: return  $E$  and  $w$ 
```

3.5 Characterization and Analysis of communities

Finally, some metrics of SNA will be applied to the original networks formed without semantic information, and compared to those applied to the networks formed with the content of the papers: the keywords based network and the topic based network. Density, average degree, modularity and others will be used.

Additionally, this thesis focus its analysis in the discovery of key members, using HITS and PageRank algorithm to compare whether the authors founds by this methods are of importance.

To make a more profuse analysis, the characterization will consider the topics found with the LDA model, to make an analysis of the main subjects of research that the two VCoPs has had in time.

Chapter 4

Community and Key members discovery on real scientific VCoPs

In chapter 1 was defined what a community is, specifically a virtual community and whether they can be of practice or interest. In this thesis, the methodology is applied in two virtual communities of practice, in particular, scientific virtual communities corresponding to sub disciplines of Computer Science, Ubiquitous Computing and Semantic Web.

A scientific network of collaboration can be any community formed by people who make research in a specific area of knowledge, and commonly, publish it in the different platforms across the world, such as journals or magazines, in conferences, congress and other. According to Ductor et al., *scientific collaboration involves the exchange of opinions and ideas and facilitates the generation of new ideas. Access to new and original ideas may in turn help researchers be more productive. It follows that, other things being equal, individuals who are better connected and more “central” in their professional network may be more productive in the future.* [17].

The study of scientific networks is not new in literature, existing several approaches of study: bibliometrics (networks formed by the citations between papers) and scientometrics (science whose study covers scientific production levels in the varied fields of research.).

Nonetheless, these disciplines are quite distinct, according to Newman, from coauthorship networks; the nodes in a citation network are typically papers, not authors, and the links between them are citations, not coauthorship [27]. Furthermore, none of this disciplines take into account the whole semantics of the articles published by scientists to form new networks, which this thesis do take into consideration.

4.1 Ubiquitous Computing

Ubiquitous Computing is defined as the field of research that tries to integrate deeply computer and human systems. This emerging area has been defined by his ability to change certain paradigms of the computational models of this century. Researchers of this area establish that systems should be involved as an inseparable part of our daily experiences being, simultaneously imperceptible. Computers and devices should be immersed with our movements and to the interactions we have with the environment [23].

The first ideas and settlements in this area of research were structured by M. Weiser [38], who had a first glimpse of what Ubiquitous Computing was in 1991, describing it as a computational environment where different systems are intertwined in daily life and fade into the background.

In [23] the authors mention that Ubiquitous Computing is a concept derived from two previous terms in this field, *Mobile Computing* and *Pervasive Computing*. The first, promoted the physical ability of computing services to move, while the other encourages to make computers as invisible as possible. Therefore, it is an interdisciplinary field that covers many topics. Both authors, stated in 2002, that this is an emerging area, which is positioned at an early stage of development, and therefore have many challenges ahead. Currently, researchers in this area not only research on various topics, but they also are modeling and imagining future technologies that will join us in the future.

Given the current emergence of this area as a consolidated discipline within the Computer Science field, different studies have been presented in recent years. Some authors, including Weiser [38] and Lyytinen et al. [23] have created initiatives to study the area. One of the most profuse work that exists in the field is the one made by Zhao and Wang [42], who propose an analysis of *Scientometrics* in the Ubiquitous Computing field through articles of the area. Their study is based on bibliographic citations contained in an article, which are used to find the main subject of study, its authors, and other relationships. The methodology uses a program called *CiteSpace* developed by Chen [14], which achieves a data visualization based on the literature review of all items in the dataset.

4.2 International Conferences on Pervasive and Ubiquitous Computing

The chosen dataset to study the Virtual community of practice of Ubiquitous Computing was the one formed by the conferences of Ubicomp and Pervasive Computing, constituting the International Conferences on Pervasive and Ubiquitous Computing series, being held since 2001 to the present.

This thesis covers all the papers presented at this conferences, forming a dataset of 740 papers written by 1,920 authors in 13 years of activity, over 23 meetings. A list of the conferences is shown below:

- Ubicomp 2001: Ubiquitous Computing, Third International Conference Atlanta, Georgia, USA.
- Pervasive Computing, 1st International Conference, Pervasive 2002, Zürich, Switzerland.
- UbiComp 2002: Ubiquitous Computing, 4th International Conference, Göteborg, Sweden.
- UbiComp 2003: Ubiquitous Computing, 5th International Conference, Seattle, WA, USA.
- Pervasive Computing, 2nd International Conference, Pervasive 2004, Vienna, Austria.
- UbiComp 2004: Ubiquitous Computing: 6th International Conference, Nottingham, UK.
- Pervasive Computing, 3rd International Conference, Pervasive 2005, Munich, Germany.
- UbiComp 2005: Ubiquitous Computing, 7th International Conference, UbiComp 2005, Tokyo, Japan.
- Pervasive Computing, 4th International Conference, Pervasive 2006, Dublin, Ireland.
- UbiComp 2006: Ubiquitous Computing, 8th International Conference, UbiComp 2006, Orange County, CA, USA.
- Pervasive Computing, 5th International Conference, Pervasive 2007, Toronto, Canada.
- UbiComp 2007: Ubiquitous Computing, 9th International Conference, UbiComp 2007, Innsbruck, Austria.
- Pervasive Computing, 6th International Conference, Pervasive 2008, Sydney, Australia.

- UbiComp 2008: Ubiquitous Computing, 10th International Conference, UbiComp 2008, Seoul, Korea.
- Pervasive Computing, 7th International Conference, Pervasive 2009, Nara, Japan.
- UbiComp 2009: Ubiquitous Computing, 11th International Conference, UbiComp 2009, Orlando, Florida, USA.
- Pervasive Computing, 8th International Conference, Pervasive 2010, Helsinki, Finland.
- UbiComp 2010: Ubiquitous Computing, 12th International Conference, UbiComp 2010, Copenhagen, Denmark.
- Pervasive Computing, 9th International Conference, Pervasive 2011, San Francisco, CA, USA.
- UbiComp 2011: Ubiquitous Computing, 13th International Conference, UbiComp 2011, Beijing, China.
- Pervasive Computing, 10th International Conference, Pervasive 2012, Newcastle, UK.
- The 2012 ACM Conference on Ubiquitous Computing, Ubicomp '12, Pittsburgh, PA, USA.
- The 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '13, Zurich, Switzerland.

4.3 The International Semantic Web and European Semantic Web Conference series

The Semantic Web was born with the World Wide Web Consortium (3WC) as an initiative to provide a standard for the way the knowledge was being shared over the Internet: *The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.* According to Antoniou et al., the term Semantic Web comprises techniques that promise to dramatically improve the current web and its use [3].

Tim Berners-Lee is one of the main contributors for the development of the area, who, in 2001, coined the term *Web of data*, to insinuate that the web should be a tool that enables users to share, find and combine data in a useful yet simply way. *The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined*

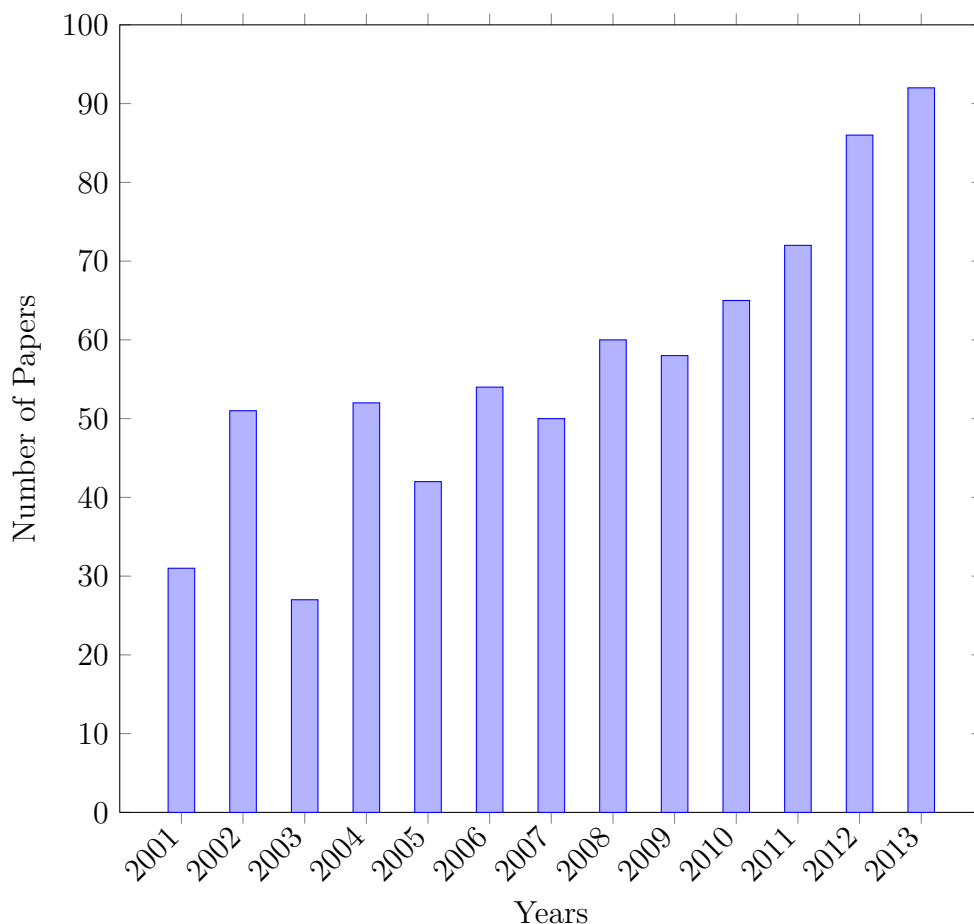


Figure 4.1: Number of Papers per Year in Conferences on Pervasive and Ubiquitous Computing

meaning, better enabling computers and people to work in cooperation [6].

This field has developed in several knowledge domains, such as frameworks to describe knowledge (RDF), ontologies (family of knowledge representation areas), query language (SPARQL), web rule languages frameworks (RIF), structure of contents (XML schemas), to name a few.

The second virtual community of practice chosen in this work, was the one formed by this area of research, mainly because the “practice” the research of the semantic web, it is very active and has a large number of members. This dataset is formed by 4,105 papers of 8,120 authors from 2001 to 2013. This Conference Series contemplate 37 conferences and the web that includes the metadata of them is called the Semantic Web Conference Corpus, <http://data.semanticweb.org/>, being a helpful tool and a central point for researchers: *In previous conferences, data has been hosted at the conference site. This introduces potential*

problems of curation and sustainability. At data.semanticweb.org we intend to provide a permanent, central, home for this conference metadata [1]. The named conferences are:

- 10th International Conference on Dublin Core and Metadata Applications (DC-2010)
- Digital Humanities 2010 (DH2010)
- 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW2012)
- 3rd European Semantic Web Conference (ESWC2006)
- 4th European Semantic Web Conference (ESWC2007)
- 5th European Semantic Web Conference (ESWC2008)
- 6th Annual European Semantic Web Conference (ESWC2009)
- 7th Extended Semantic Web Conference (ESWC2010)
- 8th Extended Semantic Web Conference (ESWC2011)
- 9th Extended Semantic Web Conference (ESWC2012)
- 10th ESWC 2013 (ESWC2013)
- 3rd Future Internet Symposium (FIS2010)
- 5th International Conference on Semantic Systems (I-Semantics 2009)
- The First International Semantic Web Conference (ISWC2002)
- The Second International Semantic Web Conference (ISWC2003)
- The Third International Semantic Web Conference (ISWC2004)
- The Forth International Semantic Web Conference (ISWC2005)
- 5th International Semantic Web Conference (ISWC2006)
- 6th International and 2nd Asian Semantic Web Conference (ISWC2007+ASWC2007)
- 7th International Semantic Web Conference (ISWC2008)
- 8th International Semantic Web Conference (ISWC2009)
- 9th International Semantic Web Conference (ISWC2010)
- The 11th International Semantic Web Conference (ISWC2012)
- The 12th International Semantic Web Conference (ISWC2013)
- The Sixth International Language Resources and Evaluation Conference (LREC2008)
- The Second International Conference on Web Reasoning and Rule Systems (RR2008)
- The Third International Conference on Web Reasoning and Rule Systems (RR2009)

- The Fourth International Conference on Web Reasoning and Rule Systems (RR2010)
- The Fifth International Conference on Web Reasoning and Rule Systems (RR2011)
- The 5th International Symposium on Rules: Research Based and Industry Focused (Barcelona) (RuleML2011-Europe)
- The First Semantic Web Working Symposium (SWWS2001)
- 16th International World Wide Web Conference (WWW2007)
- 17th International World Wide Web Conference (WWW2008)
- 18th International World Wide Web Conference (WWW2009)
- 19th International World Wide Web Conference (WWW2010)
- 20th International World Wide Web Conference (WWW2011)
- 21st International World Wide Web Conference (WWW2012)

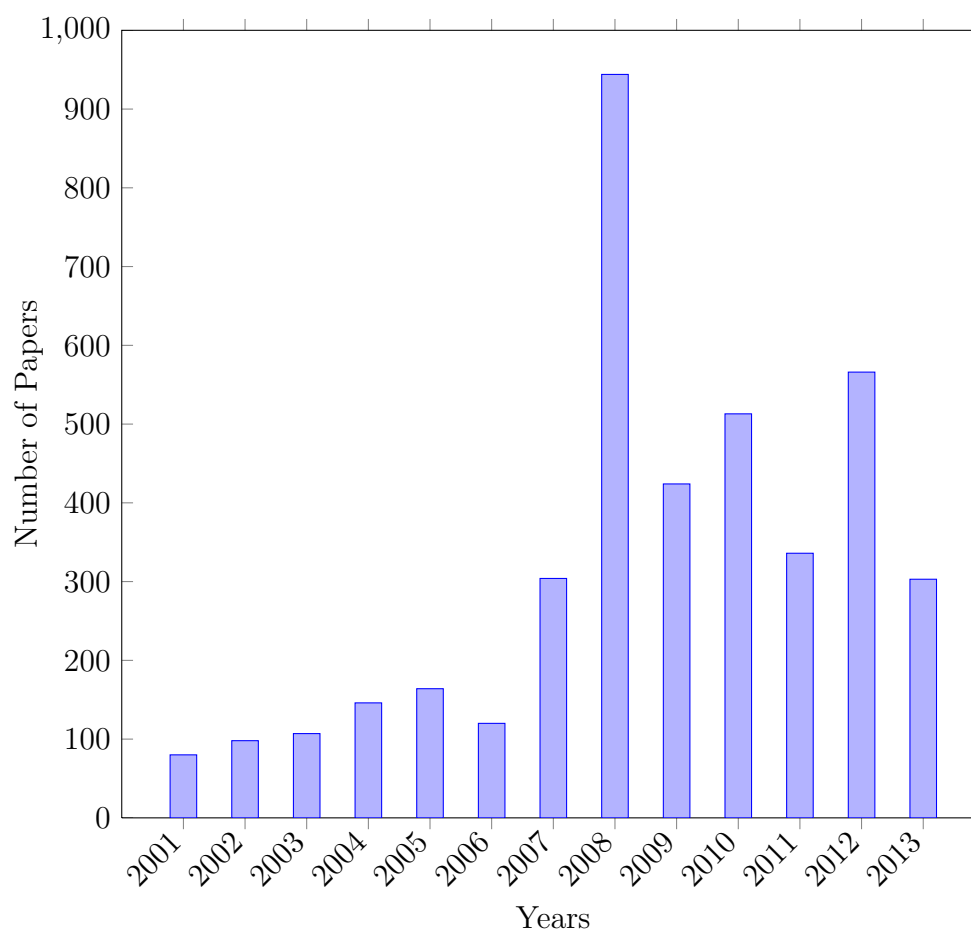


Figure 4.2: Number of Papers per Year in Semantic Web Conference series

Chapter 5

Results and Discussion

This chapter aims to present the main results of the application of the methodology proposed to the datasets described in the previous chapter. The networks built with the methodology were set as undirected graphs because the nature of the edges is coauthorship of papers, which has no “direction”. Given all the authors for a paper, there is no clear relationship rule between two authors that worked together, because there isn’t an order that indicates a direction from one to another. For example, between the fourth and fifth author of a paper it is not obvious to whom should an edge be directed to. Even, a simple rule as an edge goes from the first author to the others authors would not represent a hierarchy between authors because there is no unique rule for deciding the order of authorship of the article.

This chapter is organized as follows: first, the topic based model applied to both datasets is shown, in order to understand that topics were used as an input to build a topic based network in each dataset. This part also has an analysis of the topics found by the model. Second, all the networks built are presented for each dataset: the original network without semantic information, the keywords based network and the topic based network. Finally, the key members discovery analysis is presented, showing which key members were found by the algorithms HITS and PageRank and its performance.

5.1 Topic Analysis

In this section, a description of the topics founded with the LDA model is made. For reasons of interpretability, the results below are the ones obtained for the model with 15 topics, the results for the model with 25 and 50 topics can be found in the Appendix Section 6.1.

As stated in Section 2.6, the probabilistic model LDA generates as output the inferred

topics from the documents, in this case, papers, that represent a reduce semantic representation of the document, that is, the topics describe the content of the papers using a lower dimensional vector than words appearance.

5.1.1 Topic Analysis Ubiquitous and Pervasive Computing Conferences

The model describes both the topic composition and the words distribution by topic. The words most likely to be generated by topic are presented in Table 5.1, as well as their corresponding probability in Table 5.2.

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 1	locat	particip	studi	technolog	privaci	inform	work	user	base	peopl
Topic 2	activ	sensor	signal	time	measur	data	gp	model	figur	locat
Topic 3	user	mobil	comput	design	applic	time	phone	sensor	context	inform
Topic 4	user	devic	game	inform	design	locat	base	time	comput	displai
Topic 5	data	user	applic	sensor	design	devic	time	model	phone	comput
Topic 6	home	user	base	sensor	studi	work	time	devic	gestur	comput
Topic 7	user	technolog	data	tag	inform	comput	design	time	locat	studi
Topic 8	data	locat	user	time	inform	base	sensor	activ	model	comput
Topic 9	data	user	particip	studi	servic	comput	base	provid	home	time
Topic 10	activ	user	displai	data	time	work	comput	object	sensor	particip
Topic 11	data	time	user	studi	place	mobil	particip	work	locat	inform
Topic 12	data	user	time	context	sensor	applic	activ	base	posit	devic
Topic 13	user	locat	time	data	comput	devic	inform	activ	particip	applic
Topic 14	sensor	user	data	devic	base	locat	comput	activ	time	imag
Topic 15	user	inform	data	mobil	devic	time	applic	work	provid	studi

Table 5.1: Display of 10 words for each Topic

With the first two or fourth words of every topic, and knowing the field, an understandable interpretation of each topic can be made, such as designating a name to each topic.

Topic	Prob. 1	Prob. 2	Prob. 3	Prob. 4	Prob. 5	Prob. 6	Prob. 7	Prob. 8	Prob. 9	Prob. 10
Topic 1	0.015	0.008	0.007	0.006	0.005	0.005	0.005	0.005	0.005	0.004
Topic 2	0.008	0.006	0.005	0.005	0.005	0.005	0.005	0.004	0.004	0.004
Topic 3	0.009	0.007	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.005
Topic 4	0.023	0.007	0.006	0.005	0.005	0.005	0.004	0.004	0.004	0.004
Topic 5	0.01	0.01	0.008	0.007	0.006	0.005	0.005	0.005	0.005	0.004
Topic 6	0.009	0.006	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004
Topic 7	0.007	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.004
Topic 8	0.009	0.008	0.007	0.006	0.006	0.006	0.006	0.005	0.005	0.005
Topic 9	0.009	0.008	0.007	0.006	0.005	0.005	0.005	0.005	0.005	0.004
Topic 10	0.008	0.008	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.004
Topic 11	0.01	0.007	0.006	0.006	0.005	0.005	0.005	0.005	0.004	0.004
Topic 12	0.009	0.007	0.006	0.005	0.005	0.005	0.005	0.005	0.005	0.004
Topic 13	0.009	0.008	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.004
Topic 14	0.012	0.009	0.008	0.007	0.005	0.005	0.005	0.005	0.004	0.004
Topic 15	0.012	0.007	0.006	0.006	0.005	0.005	0.005	0.005	0.004	0.004

Table 5.2: Probabilities of words in 15 Topics

From the Tables 5.1 and 5.2, it is possible to manually group topics into concepts categories. Given that, papers belong to an already specific area of computer science, is not rare that many of the topics found share a high similarity of words (they can share the same research idea with slightly changes). For example, from Tables 5.1 and 5.2, it is possible to infer that the Topic 1, Topic 8 and Topic 13 are topics that deal with the user location, with location data such as GPS, or studies were participation involves location data. These topics can be grouped under the overall concept **Location Data**, which is useful for social network analysis.

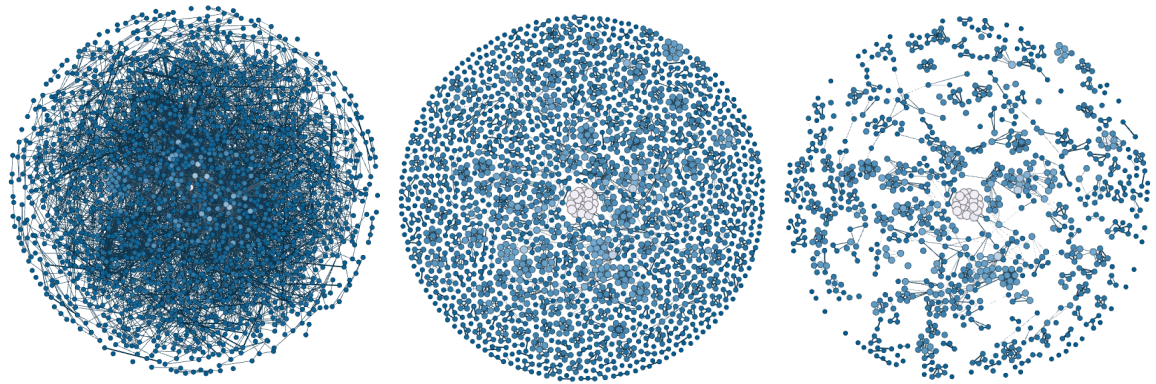
Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 1	locat	particip	studi	technolog	privaci	inform	work	user	base	peopl
Topic 8	data	locat	user	time	inform	base	sensor	activ	model	comput
Topic 13	user	locat	time	data	comput	devic	inform	activ	particip	applic

Table 5.3: Topics grouped under the concept **Location Data**

In Figure 5.1 a comparison of the networks filtered and without filter can be observed. In Figure 5.1a the basic network can be seen without no type of filtering, i.e, without any semantic information, two nodes that are connected is because those two authors wrote a paper together. In Figure 5.1b the Topic based network is presented with the filter of 15 topics extracted from the entire dataset of 740 documents. Then in Figure 5.1c, the same graph of Figure 5.1b is shown, but with the filter of the concept **Location Data**, where only the main papers that presented this particular concept as a main subject of research where filtered (over 0.5 probability of belonging to Topic 1 or Topic 8 or Topic 13). As expected, the density of such a filtered graph decreased, which suggest that using these type of visualization techniques in graphs can help to have a better insight of what the community is and for analysts to observe in a much clear way the network.

Using a deeper analysis into the topics, it is possible to observe that the papers with highest probabilities within the grouped topics under the concept **Location Data**, in fact, talk about research made with location data, as is observed in the list below:

- Cellular data meet vehicular traffic theory: location area updates and cell transitions for travel time estimation (2012).
- Safeguarding Location Privacy in Wireless Ad-Hoc Networks (2007).



(a) Original Network without topic filtering (b) All Topics based Network, 15 topics (c) Topic Network filtered Topic "Location Data"

Figure 5.1: Social Network Visualization (a) Network with no filter, (b) Topic based network and (c) Topic based network filter with Topic concept "Location Data"

- COPDTrainer: a smartphone-based motion rehabilitation training system with real-time acoustic feedback (2013).
- Living for the Global City: Mobile Kits, Urban Interfaces, and Ubicomp (2005).
- Shake Well Before Use: Authentication Based on Accelerometer Data (2001).
- Locality and privacy in people-nearby applications (2013).
- Is Context-Aware Computing Taking Control away from the User? Three Levels of Interactivity Examined (2003).

For the filtered network of Topic based with an specific topic, another benefit is that ranking algorithms run faster obtaining results in shorter times. In Table 5.4 it is possible to observe that using the PageRank ranking, one can filter by a specific topic its main researchers, in this case, the top ten authors that write more influentially about the concept "Location Data".

Complete Network		Topic Based Network (All Topics)		Topic Based ("Location Data")	
Member	PageRank	Member	PageRank	Member	PageRank
Gregory D. Abowd	0.0045	Norman M. Sadeh	0.0013	Conor Haggerty	0.0023
Gaetano Borriello	0.0037	Hong Lu	0.0012	Chris Greenhalgh	0.0020
Sunny Consolvo	0.0032	Christof Roduner	0.0011	Xianghua Ding	0.0020
James Scott	0.0029	Roy Want	0.0011	John Bowers	0.0020
Anind K. Dey	0.0028	Andrew T. Campbell	0.0011	Janet van der Linden	0.0019
Anthony LaMarca	0.0027	George Colouris	0.0011	Cuong Pham	0.0019
Steve Benford	0.0025	Lorrie Faith Cranor	0.0010	William R. Hazlewood	0.0019
Shahram Izadi	0.0023	Allison Woodruff	0.0010	Alex Butler	0.0019
Gerhard Tröster	0.0023	John Bowers	0.0010	Brian D. Ziebart	0.0019
Timothy Sohn	0.0022	Ryan Libby	0.0010	Deborah Estrin	0.0019

Table 5.4: Top 10 members by PageRank score over the complete network and Topic based network for all topics and for concept "Location Data"

In order to better recognize and interpret topics, names were given to each topic considering the most important words given the probabilities in each topic. In table 5.5 a possible list of names can be observed for each topic:

Topic	Topic Name
Topic 1	Studies of Location
Topic 2	Studies with Sensors
Topic 3	User's Mobiles
Topic 4	Games for user's devices
Topic 5	Applications with user's data
Topic 6	Home Technologies for users
Topic 7	Technologies for users
Topic 8	Location Data
Topic 9	Data from users
Topic 10	Display of user activities
Topic 11	Studies with time data from users
Topic 12	Context aware data
Topic 13	Location and Time data from users
Topic 14	Users Sensors measurements
Topic 15	Mobile users data

Table 5.5: Manually given names for each Topic

Topics over Years

In this section, a small analysis was done over the topics found by the LDA model in the dataset across years. The idea is to visualize how topics have been changing over time. With this goal in mind, using the information of paper's year and the topic they belong to, the evolution of topics across years was plotted, including the presence of each topic by year. In Figure 5.2 the data can be observe:

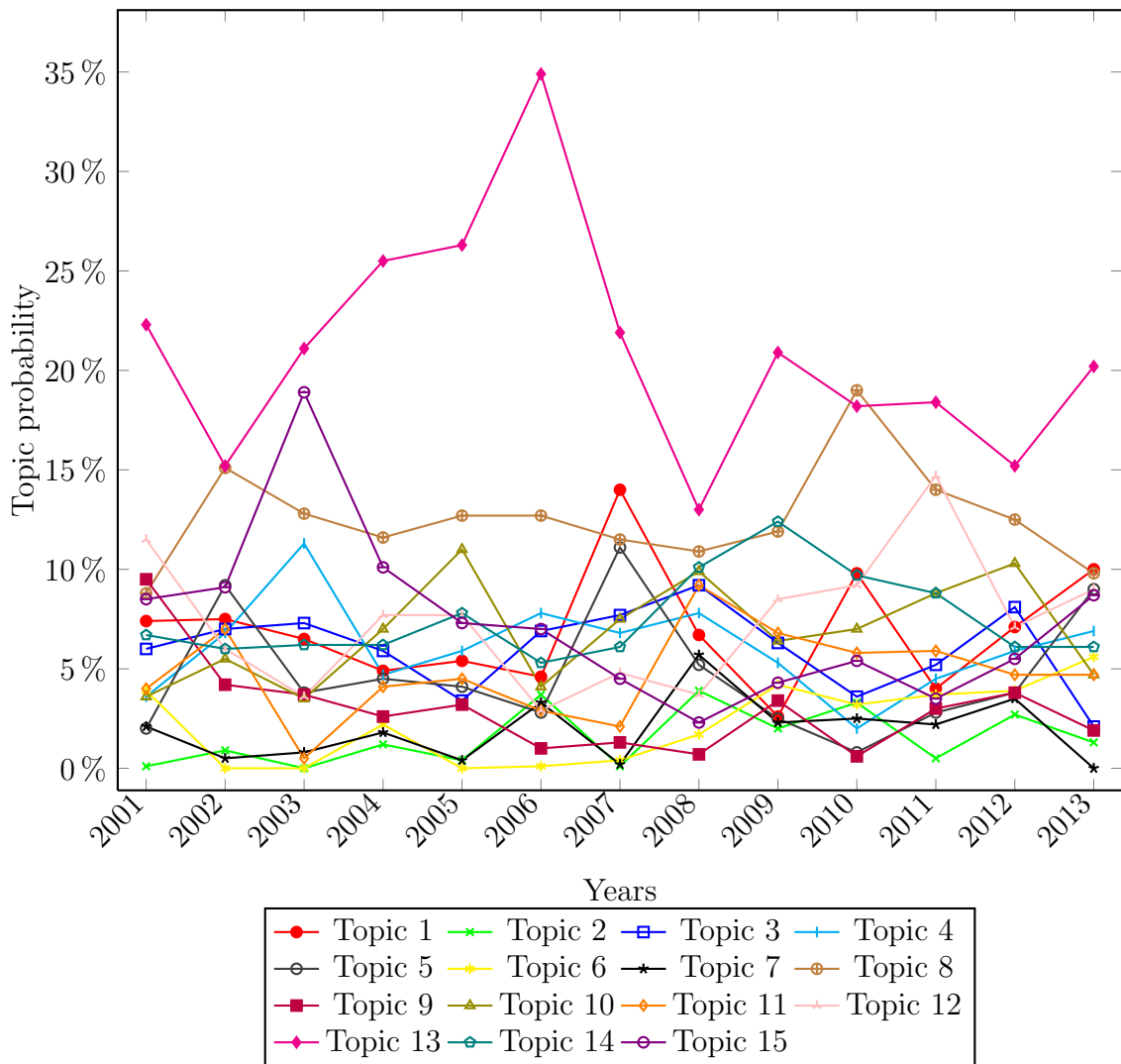


Figure 5.2: Topic Probability over the years in Ubiquitous and Pervasive Computing Conferences

It can be seen that Topic 13 has been, in the overall, the topic with more paper presence, which can be interpreted as “Location Data”, concept that was grouped along with topic 1 and 8, and these make sense, since this trend is the one that has been dominating the development of the technology and the research in the area, taking its peak at the year 2006, year in which later, curiously, the first Iphone touch was released and technologies as the gps and maps were made massive and accessible in most of the phones. This can be called a *hot trend*.

Moreover, in 2011, a smaller peak, can be observed. This corresponds to Topic 12, the one that can be interpreted as data from user with context aware consideration such as time

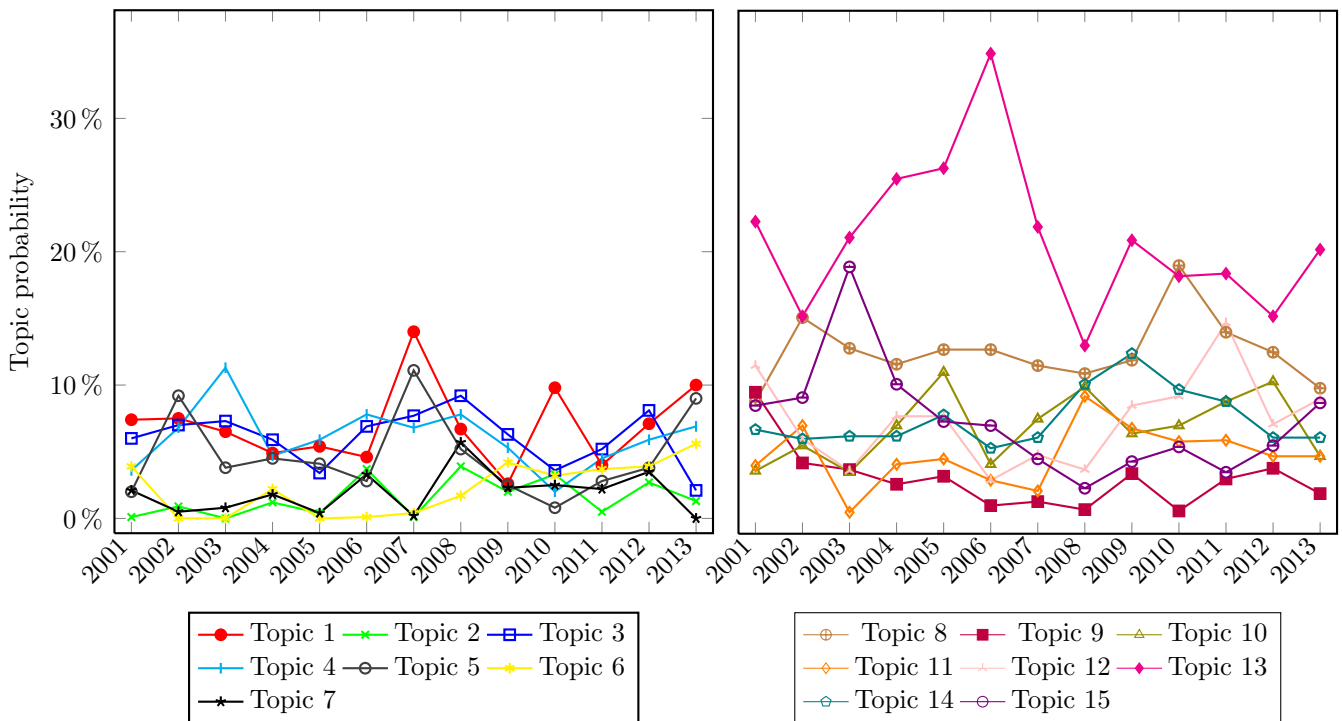


Figure 5.3: Topic Probability over the years in Ubiquitous and Pervasive Computing Conferences

and location, “Context-aware data”. This topic has also been a trend lately, with the rapid development of mobile applications that uses data from the context of the user. Therefore, researchers has also been developing technologies following this trend.

5.1.2 Topic Analysis Semantic Web Conference Series

The main words of the 15 topics of the LDA model are shown as follows, for the dataset of Semantic Web. As expected, most of the topics have similar words. This can be explain as the field is already specific in a particular subject of research, so the members of this community make they research in similar lines of investigation. Despite this, the model achieved a good separation of topics, separating the different papers in different topics. In Table 5.6 the first 10 words of each topic can be observed.

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 1	topic	model	document	data	review	spammer	text	featur	new	player
Topic 2	user	model	algorithm	social	network	set	time	queri	result	search
Topic 3	set	queri	algorithm	cluster	method	learn	score	rank	result	url
Topic 4	user	model	network	social	lyon	inform	number	differ	data	task
Topic 5	question	languag	word	answer	model	web	inform	set	phrase	document
Topic 6	rdf	properti	tripl	queri	semant	owl	web	rule	graph	sparql
Topic 7	entiti	path	extract	relat	node	page	worker	web	code	
Topic 8	servic	qo	nash	composit	set	task	comput	node	problem	web
Topic 9	set	ontolog	graph	rule	relat	algorithm	base	node	map	semant
Topic 10	queri	search	data	time	simul	match	fresh	result	result	search
Topic 11	data	queri	web	servic	applic	model	user	process	http	provid
Topic 12	model	word	wikipedia	topic	commun	link	network	featur	set	data
Topic 13	model	event	crimin	page	ontolog	web	site	set	annot	base
Topic 14	node	domain	label	ontolog	model	set	gener	base	concept	valu
Topic 15	data	queri	web	entiti	semant	document	link	search	inform	ontolog

Table 5.6: Display of 10 words for each Topic

The probabilities of each word in every topic, i.e., the probability distribution can be seen in Table 5.7.

Topic	Prob. 1	Prob. 2	Prob. 3	Prob. 4	Prob. 5	Prob. 6	Prob. 7	Prob. 8	Prob. 9	Prob. 10
Topic 1	0.016	0.009	0.007	0.006	0.005	0.005	0.005	0.005	0.005	0.005
Topic 2	0.014	0.009	0.008	0.008	0.007	0.007	0.007	0.006	0.005	0.005
Topic 3	0.012	0.011	0.008	0.007	0.007	0.006	0.006	0.006	0.006	0.006
Topic 4	0.015	0.007	0.007	0.007	0.006	0.006	0.005	0.005	0.005	0.005
Topic 5	0.012	0.01	0.009	0.009	0.008	0.006	0.006	0.005	0.005	0.005
Topic 6	0.019	0.013	0.01	0.01	0.01	0.008	0.008	0.008	0.007	0.006
Topic 7	0.013	0.012	0.009	0.007	0.007	0.007	0.007	0.006	0.006	0.005
Topic 8	0.035	0.02	0.007	0.007	0.006	0.006	0.005	0.005	0.005	0.005
Topic 9	0.013	0.013	0.012	0.008	0.007	0.007	0.006	0.006	0.005	0.005
Topic 10	0.051	0.008	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.004
Topic 11	0.014	0.012	0.009	0.006	0.006	0.006	0.005	0.005	0.005	0.005
Topic 12	0.012	0.012	0.007	0.006	0.006	0.005	0.005	0.005	0.004	0.004
Topic 13	0.014	0.013	0.009	0.007	0.007	0.006	0.006	0.005	0.005	0.005
Topic 14	0.007	0.007	0.006	0.006	0.005	0.005	0.005	0.005	0.004	0.004
Topic 15	0.015	0.012	0.012	0.011	0.01	0.009	0.009	0.008	0.007	0.007

Table 5.7: Probabilities of words in Topics

Following the same procedure of grouping topics made with the other dataset, for this dataset, the Topic 1 and Topic 12 could be grouped under the concept of **Topics models**, since they talk about modeling words in documents by topics, something similar as one of the methods of this thesis, such as LDA and others topic models like LSA, PLSI, etc. This is a common and well known area of study inside Semantic Web field.

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 1	topic	model	document	data	review	spammer	text	featur	- (minus sign)	new
Topic 12	model	word	- (minus sign)	wikipedia	topic	commun	link	network	featur	set

Table 5.8: Topics grouped under the concept “Topics models”

Using the same analysis as the first dataset, a visualization over the concept “Topics

Models” can be made. Figure 5.4 shows an insight of what a filtered graph is when applying one concept filter.

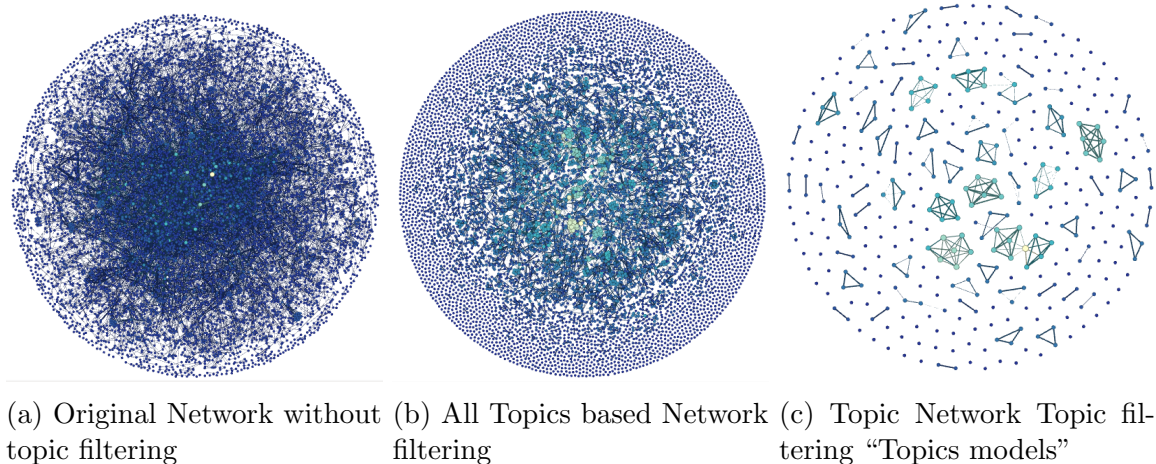


Figure 5.4: Social Network Visualization (a) Network with no filter, (b) Topic based network and (c) Topic based network filter with Topic concept “Topics models”

In Figure 5.4c a notorious reduction of density is noted, making it easier to visualize collaboration networks within this specific concept of research, “Topics Models”. Likewise the other dataset, the Table 5.9 shows the scores of PageRank of the concept “Topics Models”. It is clear that the scores are higher for the filtered network with one topic than the complete network with no semantic information and the network filtered by all topics (in this case 15). This relationships hold because there are lesser nodes, therefore experts in that field appear easily.

Complete Network		Topic Based Network (All Topics)		Topic Based (“Topics Models”)	
Member	PageRank	Member	PageRank	Member	PageRank
Zheng Chen	0.0020	Fabio Casati	0.0010	Neil Fraistat	0.0069
Lei Zhang	0.0010	Jesus Contreras	0.0004	Lynne Siemens	0.0062
Jiajun Bu	0.0010	Giovanni Tummarello	0.0004	Neel Sundaresan	0.0062
Yong Yu	0.0010	Carole A. Goble	0.0004	Susan Brown	0.0057
Wolfgang Nejdl	0.0010	Xiangyang Xue	0.0004	Ilyas Potamitis	0.0055
Steffen Staab	0.0010	Peter Wittenburg	0.0004	Todor Ganchev	0.0055
Axel Polleres	0.0010	Alessandro Mazzei	0.0004	Claire Warwick	0.0052
Soeren Auer	0.0010	Guan Luo	0.0004	Steven Krauwer	0.0049
Peter Haase	0.0010	Hong-luan Liao	0.0004	Jon Sanchez	0.0046
Ian Horrocks	0.0022	Ryan Libby	0.0010	Eva Navas	0.0046

Table 5.9: Top 10 members by PageRank score over the complete network and Topic based network for all topics and for topic “Topics Models”

A deeper understanding of the concept can be done by looking at the papers with highest

probability of showing Topic 1 and Topic 12. The following papers are a sample of that subset:

- Web Video Topic Discovery and Tracking via Bipartite Graph Reinforcement Model.
- Vanishing Point(s) and Communion.
- Modeling Online Reviews with Multi-grain Topic Models.
- Corpus Co-Occurrence Dictionary and Wikipedia Entries as Resources for Semantic Relatedness Information.
- Professor or Screaming Beast? Detecting Anomalous Words in Chinese.
- Building the Valency Lexicon of Arabic Verbs.

Likewise, with this dataset names were given to the topics for a better understanding and recognition:

Topic	Topic Name
Topic 1	Topics models
Topic 2	Social network models
Topic 3	Query set
Topic 4	User model networks
Topic 5	Data communication
Topic 6	RDF field
Topic 7	Semantic paths between entities
Topic 8	Quality of Service
Topic 9	Ontology set
Topic 10	Search Query
Topic 11	Web services
Topic 12	Word model
Topic 13	Event modeling
Topic 14	Ontologies domain
Topic 15	Web search

Table 5.10: Manually given names for each Topic

Topics over Years

In this section, as well, the same small analysis of the topics found by the LDA model in the dataset across years was perform. The goal is to visualize how topics have been evolving over time and relating with the research of the VCoP. In Figure 5.5 the evolution of topics over time can be observed:

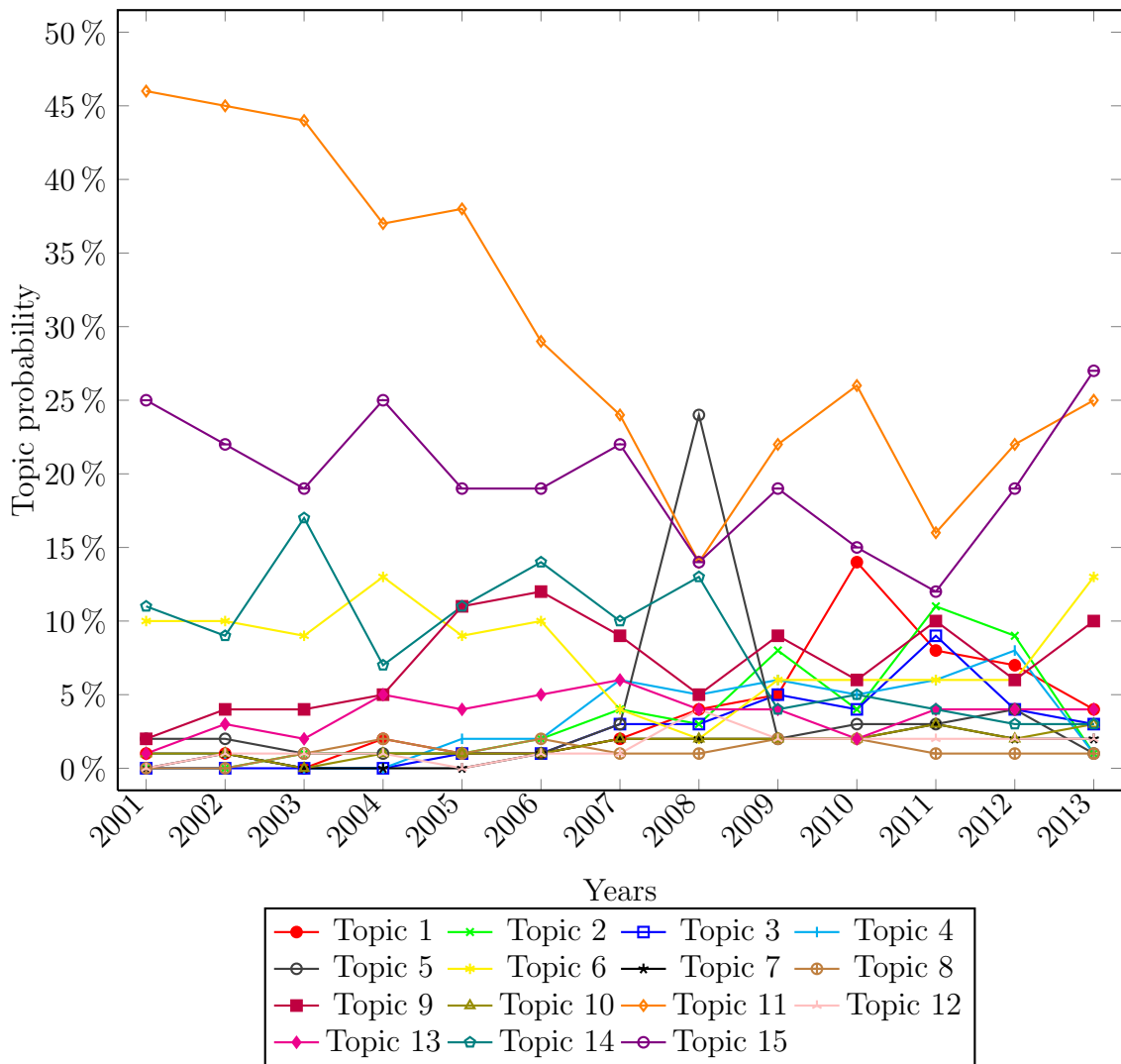


Figure 5.5: Topic Probabilities over the years in Semantic Web Conference series

From the Figure 5.5, it is possible to observe that the first year of this dataset, 2001, the topic that was more prominent among papers, is Topic 12. This topic can be interpreted using the papers with more probability of showing Topic 12:

- Corpus Co-Occurrence Dictionary and Wikipedia Entries as Resources for Semantic Relatedness Information.
- Professor or Screaming Beast? Detecting Anomalous Words in Chinese.
- Building the Valency Lexicon of Arabic Verbs.
- The Extended Architecture of Hantology for Japan Kanji.
- Improving NER in Arabic Using a Morphological Tagger.

It seems that a trend in Semantic Web were models that analyse different languages and

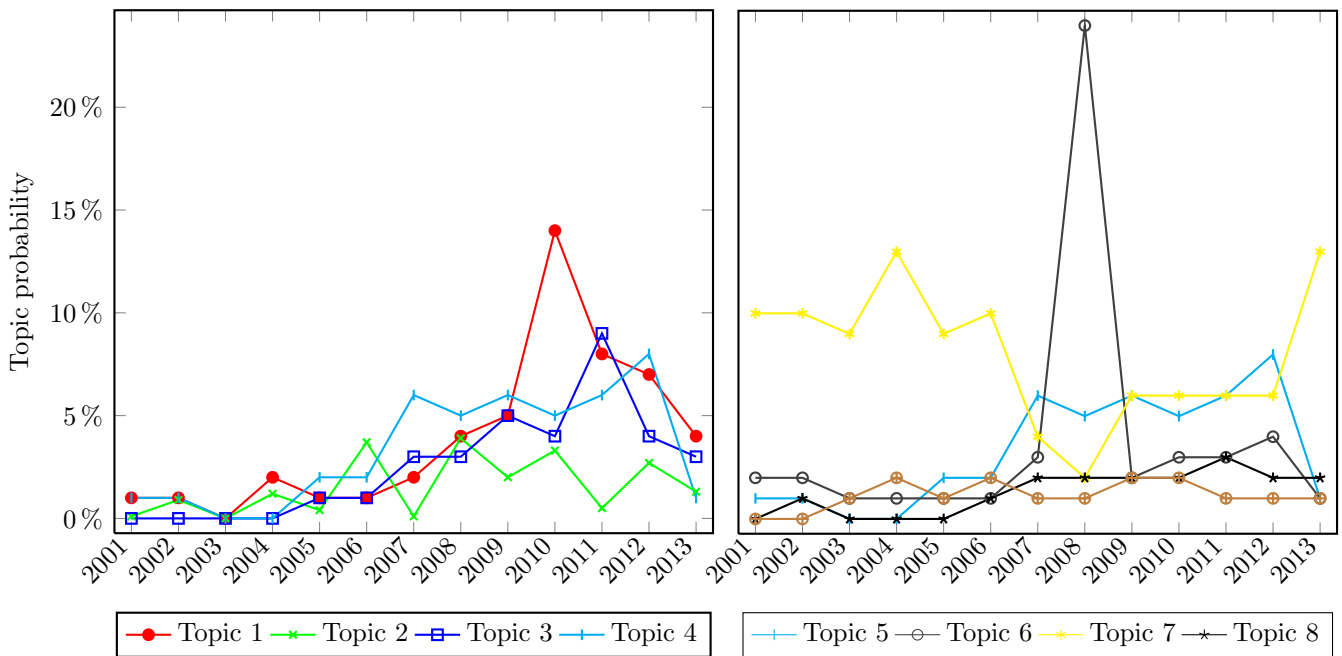


Figure 5.6: Topic Probability over the years in Semantic Web Conference series

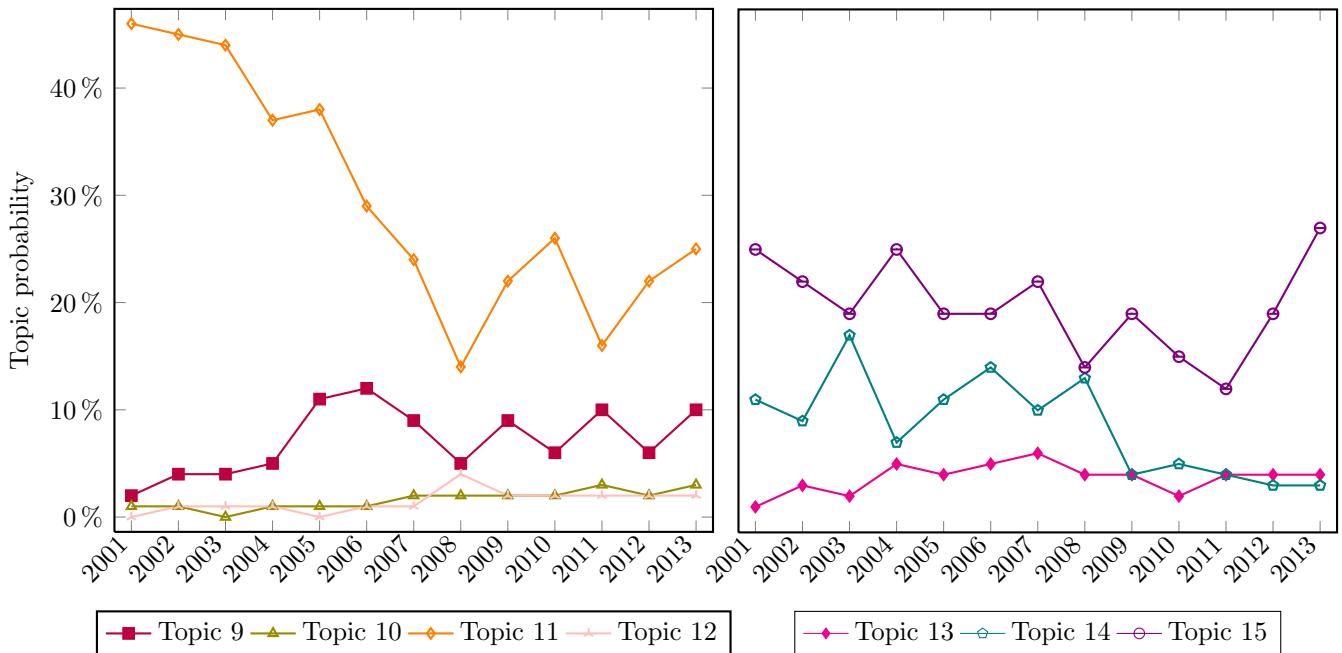


Figure 5.7: Topic Probability over the years in Semantic Web Conference series

the specific structure of them using documents of words. Additionally, in the conferences that took place the last year, the Topic 5 was the most present in the papers presented for that called. This topic can be interpreted as query languages as a service in which a question is done and an answer is expected, maybe some sort of web service translate models, since it has to do with words.

5.2 Ubiquitous and Pervasive Computing Conferences Networks

In this section, the results of the Ubiquitous and Pervasive Computing Conferences are presented, for the Original Network as for the Keywords based network and Topic based network.

5.2.1 Original Network - Ubiquitous and Pervasive Computing Conferences

This network corresponds to the network formed only by the coauthorship of the papers, i.e., an edge exists between authors only if they wrote an article together. This network has 1,920 nodes connected by 5,452 edges.

Nodes = 1,920
Edges = 5,452

Density = 0.003
Modularity = 0.903
Average Degree = 5.679
Average Weighted Degree = 6.485
Network Diameter = 13

Table 5.11: Original Network Statistics - Ubiquitous and Pervasive Computing Conferences

From the Table 5.11, it is possible to observe some structural statistics of the network. As one can see, an author work in average, with almost 5.7 other persons. In addition, an author collaborate in average, with 6.5 other people.

Moreover, the density of this network shows that a 0.3% of the possible connections of the nodes in graphs are covered, i.e., almost one third of the graph is connected.

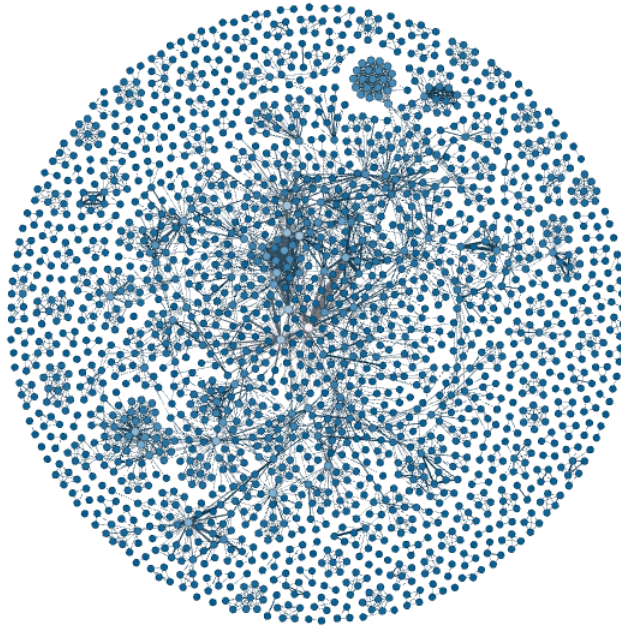


Figure 5.8: Original Network with average degree display

The results from community detection algorithm are shown in Table 5.12. For this network the number of communities found is 212 with a modularity of 0.903, that is, this network has a structure of community and can be easily partitioned into smaller groups.

Modularity = 0.903		Communities = 212	
Community Number		Number of Members	
1st		121 (6.3%)	
2nd		108 (5.62%)	
3rd		91 (4.74%)	

Table 5.12: Communities and modularity for Original Network

In order to compare if the modularity of this network is relatively high, it is necessary to compare it with a graph of the same size (equal number of nodes and edges) but randomly generated, this is, all of its edges are randomly assigned.

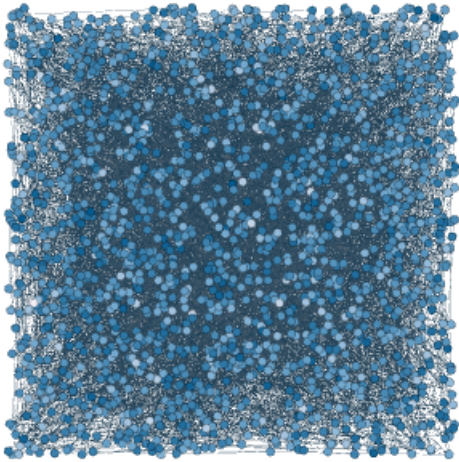
Nodes = 1,920
Edges = 5,452
Modularity = 0.386


Table 5.13: Random graph of same size as original network

Table 5.13 shows the modularity of the random generated network. It is clearly that the modularity of the original network, 0.903 it's much greater than the randomly generated network, 0.386. This means that the VCoP itself, without any semantic information, has a structure of a community, given the existing relations between its members.

5.2.2 Keywords Based - Ubiquitous and Pervasive Computing Conferences

This section presents the properties of the network built by relations based on the keywords of every paper from all authors in the dataset.

Nodes = 1,920
Edges = 1,866
Density = 0.001
Modularity = 0.672
Average Degree = 1.944
Average Weighted Degree = 1.571
Network Diameter = 12

Table 5.14: Keywords Based Network Statistics - Ubiquitous and Pervasive Computing Conferences

As seen in Table 5.14, the number of edges decreased from 5,452 in the original network to 1,866 with the Keywords based network. This decrease is because not all authors that wrote together necessarily share the same investigation topics, i.e., let's say the author *A* wrote with the author *B* in an article of *mining data from telecommunication*, but the author *A* is a scientist from the field of *Data mining* and the author *B* is an expert in *Telecommunications*, therefore, in the original network they are connected, but in Keywords based network the ties of their papers are probably removed since their investigation is not similar in terms of content.

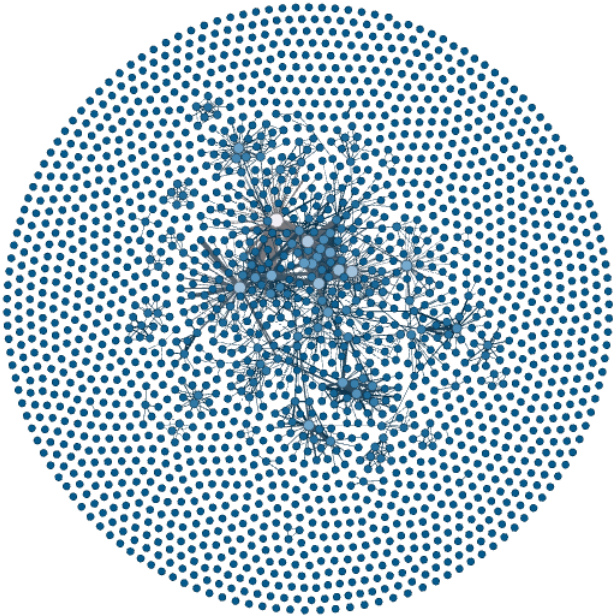


Figure 5.9: Keyword Based Network with average degree display

The reduction of density in the network becomes obvious in Figure 5.9 in comparison to Figure 5.8, as irrelevant edges are removed and therefore, more nodes are disconnected.

Modularity = 0.672 Communities = 1,095	
Community Number	Number of Members
1st	126 (6.56%)
2nd	119 (6.2%)
3rd	89 (4.64%)

Table 5.15: Communities and modularity for Keywords Based Network

Table 5.15 shows that a high modularity is also achieve yet even when the relationships are

based on content instead of connections of people who wrote together. However, modularity is lower than original network. Even though the big communities are roughly the same size as in the original network, there is an increase in the number of communities due to a big set of not connected individuals. This fact can explain the decrease in modularity. This can also be seen in Figure 5.9 compare to Figure 5.8.

Nodes = 1,920 Edges = 1,866
Modularity = 0.812

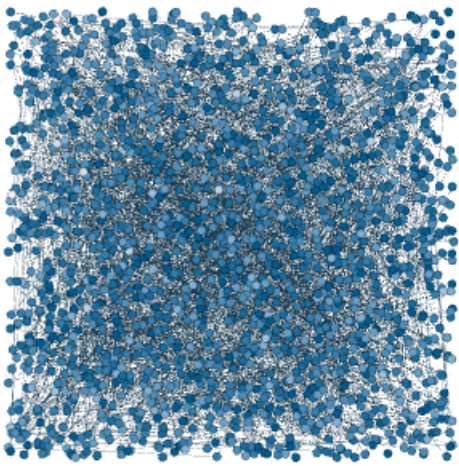


Table 5.16: Random graph of same size as keyword based network

As in original network, the modularity of the keyword based network is compared to the randomly generated network of the same size. In Table 5.16 it is possible to observe that, in this case, the modularity of the keyword based network, 0.672, is below than the value from the random network, 0.812, showing that probably the network form by keywords doesn't allow a good clustering of the network.

In Figure 5.10 a small word cloud can be seen. This was made with the purpose of showing the type of Keywords that the papers contained. They are designed such that the keyword with most occurrences is bigger and vice versa.

Nodes = 1,920
Edges = 2,913
Density = 0.002
Modularity = 0.991
Average Degree = 3.034
Average Weighted Degree = 1.316
Network Diameter = 5

Table 5.17: 15 Topic Based Network Statistics - Ubiquitous and Pervasive Computing Conferences

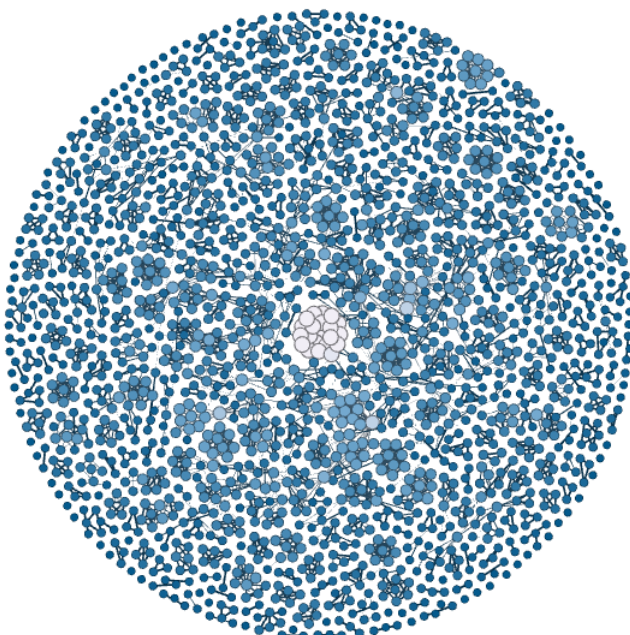


Figure 5.11: 15 Topic Based Network with average degree display

Modularity = 0.991	Communities = 652
Community Number	Number of Members
1st	29 (1.15%)
2nd	18 (0.94%)
3rd	15 (0.78%)

Table 5.18: Communities and modularity for 15 Topic Based Network

Comparing with the original network, the 15 Topic Based network has more number of communities with fewer members. In contrast with Keywords Based network, all communities

in Topic Based network even the largest ones are smaller.

Nodes = 1,920
Edges = 2,913
Modularity = 0.617

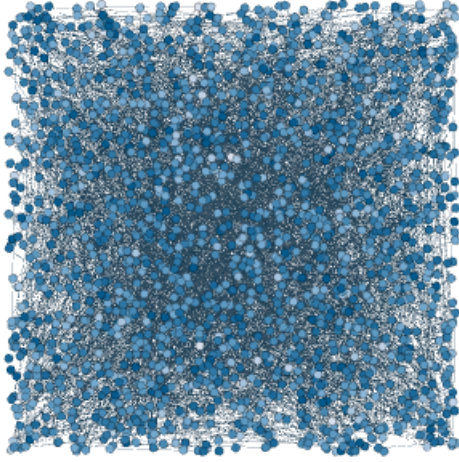


Table 5.19: Random graph of same size as topic based network

Likewise the other networks, a randomly graph was generated of the same size of the topic based network. The results from Table 5.19, shows that the modularity of the topic based network, 0.991, is much greater than its random peer of 0.617. This is evidence that using topics from LDA model is a good method to extract the community structure of a VCoP.

Tables 5.20, 5.21, 5.22 and 5.23 shows the same properties that the network built with 15 topics. The different metrics are very similar for 15, 25 and 50 topics.

Nodes = 1,920
Edges = 2,788
Density = 0.002
Modularity = 0.991
Average Degree = 2.904
Average Weighted Degree = 1.273
Network Diameter = 5

Table 5.20: 25 Topic Based Network Statistics - Ubiquitous and Pervasive Computing Conferences

Modularity = 0.991	Communities = 677
Community Number	Number of Members
1st	29 (0.94%)
2nd	29 (0.94%)
3rd	14 (0.73%)

Table 5.21: Communities and modularity for 25 Topic Based Network

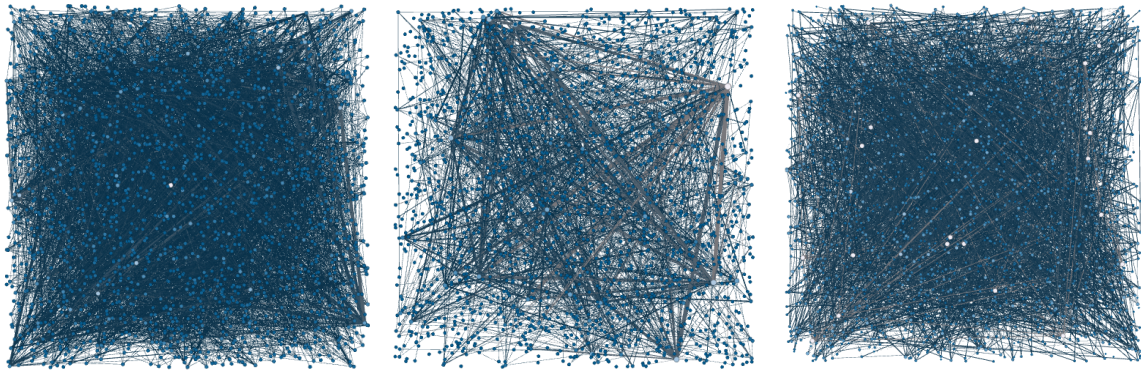
Nodes = 1,920
Edges = 2,790
Density = 0.002
Modularity = 0.991
Average Degree = 2.906
Average Weighted Degree = 1.271
Network Diameter = 6

Table 5.22: 50 Topic Based Network Statistics - Ubiquitous and Pervasive Computing Conferences

Modularity = 0.991	Communities = 688
Community Number	Number of Members
1st	23 (1.2%)
2nd	18 (0.94%)
3rd	17 (0.89%)

Table 5.23: Communities and modularity for 50 Topic Based Network

In the Figure 5.12 a visualization of the Original networks and networks built by the two methods proposed can be observed. The purpose of this figure is to show clearly that with the two methods a density reduction can be achieved, eliminating weak ties between authors that not necessarily match the research of each member. Moreover, it can be seen that the Original graph is very cluttered, meanwhile the others are very clearly and allow to find some nodes of importance.



(a) Original Network (b) Keywords Based Network (c) 15 Topic Based Network

Figure 5.12: Ubiquitous and Pervasive Computing Conferences

Summarizing, the main results of the three networks can be compared in Table 5.24:

Nodes = 1,920 Edges = 5,452	Nodes = 1,920 Edges = 1,866	Nodes = 1,920 Edges = 2,913
Density = 0.003	Density = 0.001	Density = 0.002
Modularity = 0.903	Modularity = 0.672	Modularity = 0.991
Average Degree = 5.679	Average Degree = 1.944	Average Degree = 3.034
Avg. Weighted Degree = 6.485	Avg. Weighted Degree = 1.571	Avg. Weighted Degree = 1.316
Network Diameter = 13	Network Diameter = 12	Network Diameter = 5

Table 5.24: Comparison of structural properties for Original network, Keywords Based network and Topic Based network

5.3 Semantic Web Conference Series Networks

In this section, the results of the Semantic Web Conference Series are presented, for the Original Network as for the Keywords based network and Topic based network.

5.3.1 Original Network - Semantic Web Conference Series

Following the same framework as previous dataset, this network corresponds to the network formed only by the coauthorship of the papers, i.e., an edge exists between authors only if they wrote an article together. This network has 8,120 nodes connected by 20,332 edges. Table 5.25 shows other properties of interest for this network. As one can see, an author work in average, with almost 5 other people. In addition, an author in average, have 5.7 collaborations.

Moreover, the density of this network shows that a 0.1% of the possible connections of the nodes in graphs are covered, i.e., almost one third of the graph is connected.

Nodes = 8,120
Edges = 20,332

Density = 0.001
Modularity = 0.904
Average Degree = 5.008
Average Weighted Degree = 5.787
Network Diameter = 22

Table 5.25: Original Network Statistics - Semantic Web Conference Series

In Figure 5.13, a visualization of the network can be seen where nodes are grouped by their average degree.

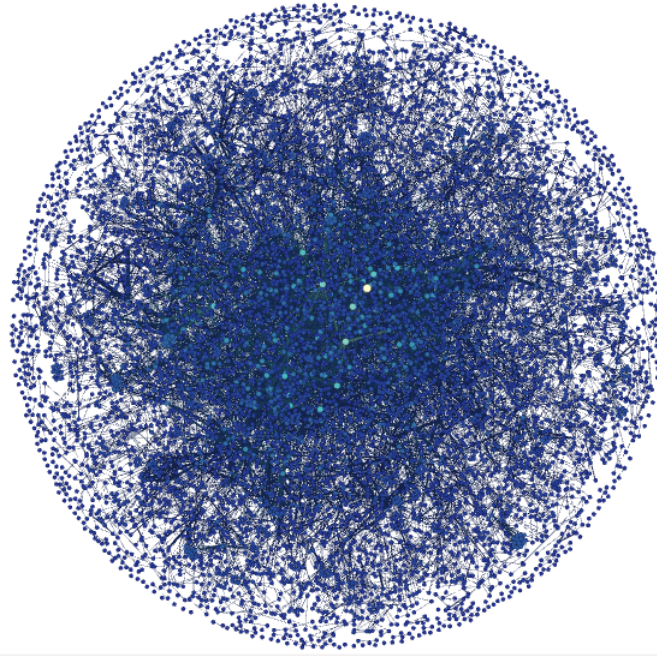


Figure 5.13: Original Network with average degree display

Modularity = 0.904 Communities = 1,132	
Community Number	Number of Members
1st	520 (6.4%)
2nd	274 (3.37%)
3rd	263 (3.24%)

Table 5.26: Communities and modularity for Original Based Network

As seen in Table 5.26, this network presents a high modularity, partitioning it in 1,132 communities. This can be stated as compared to the modularity of a random generated network of the same size that the original network (8,120 nodes and 20,332 edges). Comparing the modularity of the original network, 0.904, with the one shown in Table 5.27, 0.425, the original network modularity is greater than the random network. This fact shows that the existing relations that this VCoP has between its members shape a community structure and its capability of being partitioned.

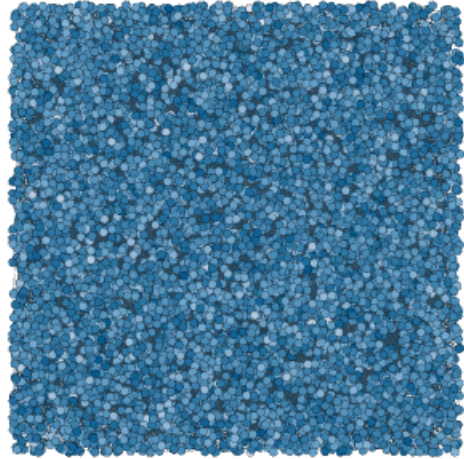
Nodes = 8,120
Edges = 20,332
Modularity = 0.425


Table 5.27: Random graph of same size as original network

5.3.2 Keywords Based - Semantic Web Conference Series

For the Keywords Based method in this dataset, the network has lesser edges than the original, 13,897, and also a decrease in density, as seen in Table 5.28:

Nodes = 8,120
Edges = 13,897
Density = 0.000
Modularity = 0.715
Average Degree = 3.423
Average Weighted Degree = 9.339
Network Diameter = 21

Table 5.28: Keywords Based Network Statistics - Semantic Web Conference Series

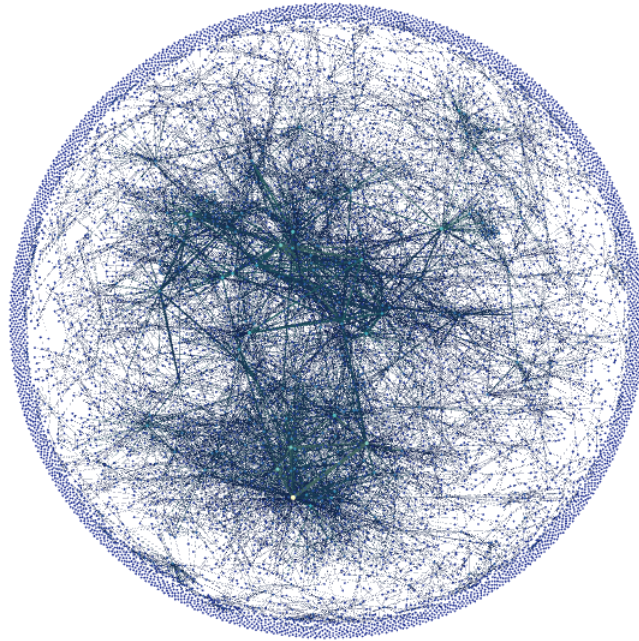


Figure 5.14: Keywords Based Network with average degree display

From Table 5.29, it can be seen that, similar to the case of the other dataset, the modularity index experimented a decrease, but the main communities are bigger in number of members. This fact may be due to the elimination of irrelevant edges that produces that a number of nodes remain isolated and the rest of the nodes stay very well connected.

Modularity = 0.715		Communities = 2,580	
Community Number		Number of Members	
1st		980 (12.07%)	
2nd		545 (6.71%)	
3rd		461 (5.68%)	

Table 5.29: Communities and modularity for Keywords Based Network

As with the original network, the modularity of the keyword based network has to be compared with the modularity of a randomly generated network of the same size of nodes and edges.

Nodes = 8,120
Edges = 13,897
Modularity = 0.579

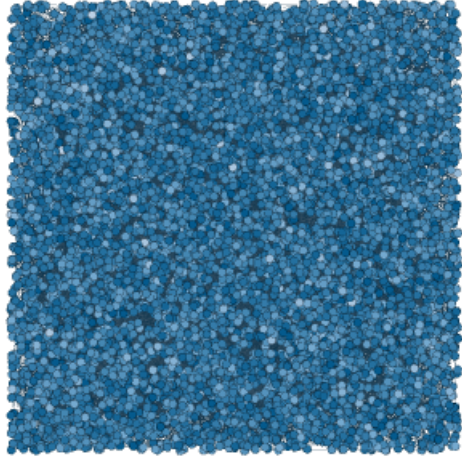


Table 5.30: Random graph of same size as keyword based network

From Table 5.30, is observed that the modularity of the keyword based is greater than the random generated graph, 0.715 and 0.579, respectively. This shows that the network built on keywords for this dataset has a community structure and can be clustered.

In Figure 5.15 a small word cloud can be seen, likewise dataset one. This was made with the purpose of showing the type of keywords that the papers included. They are designed such that the keyword with most occurrences is bigger and the small ones lesser occurrences. This helps to visualize the type of semantic content that members of the community produce.

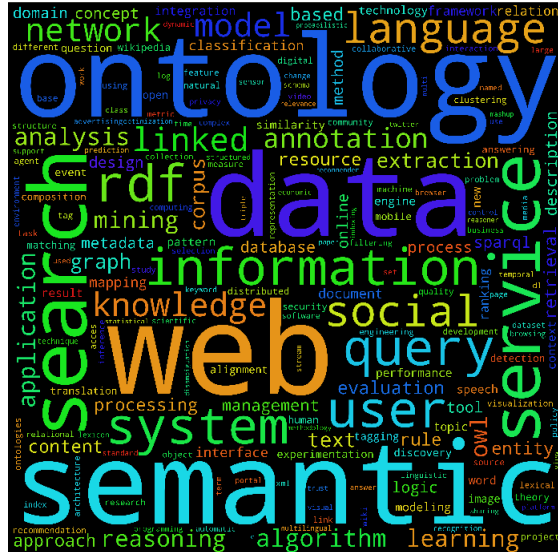


Figure 5.15: Keywords Cloud based on word occurrences

5.3.3 Topic Based - Semantic Web Conference Series

In this section, the results of the Topic based networks are shown for this dataset. Three LDA model were generated, one with 15 topics, another of 25 topics and a final one of 50 topics. The iterations in which the LDA model converge where decided by the perplex of each number of iterations. The parameters to form the edges of this networks were fixed at $\theta = 0.1$ and $\delta = 0.1$.

Nodes = 8,120
Edges = 8,502

Density = 0.000
Modularity = 0.995
Average Degree = 2.094
Average Weighted Degree = 0.764
Network Diameter = 29

Table 5.31: 15 Topic Based Network Statistics - Semantic Web Conference Series

From Table 5.31, it can be seen that with LDA model, the number of edges decreases from 20,332 in the original network to 8,502 for the model with 15 topics, along with a reduction in density and other properties.

Nevertheless, this network shows one of the highest modularity of all networks, 0.995. This means that this network, apparently, has a better structure of clusters.

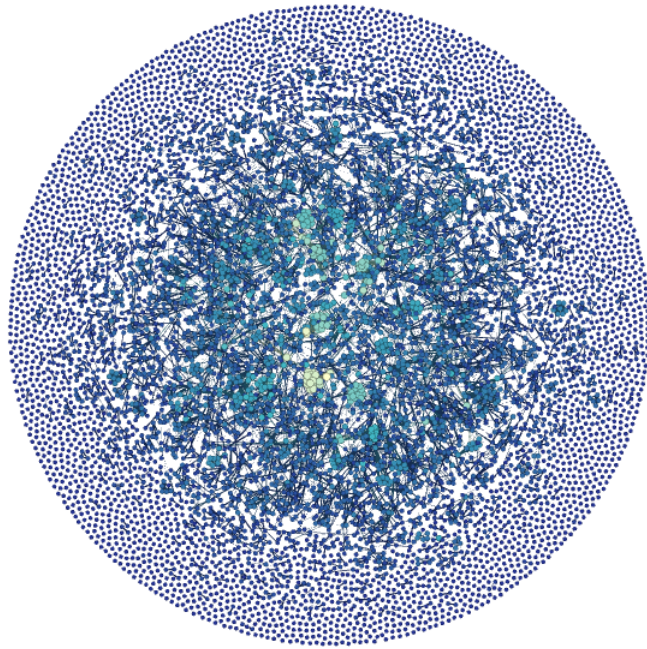


Figure 5.16: 15 Topic Based Network with average degree display

Modularity = 0.995		Communities = 3,987	
Community Number		Number of Members	
1st		57 (0.7%)	
2nd		51 (0.63%)	
3rd		49 (0.6%)	

Table 5.32: Communities and modularity for 15 Topic Based Network

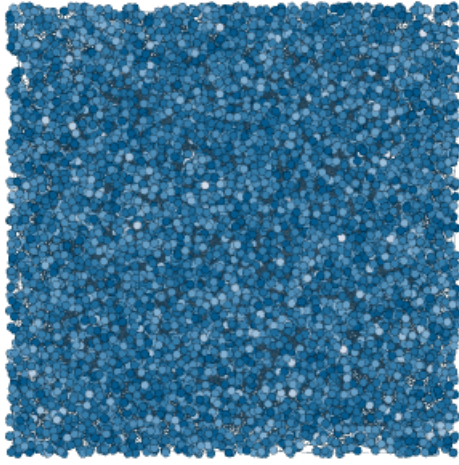
Nodes = 8,120
Edges = 8,502
Modularity = 0.810


Table 5.33: Random graph of same size as topic based network

As in original network and keyword based network, the modularity of this network is also compared with a randomly generated network. For this case, the modularity of the topic based network, 0.995, is greater than the one presented by the random network, 0.810. This results shows that the network built with topics obtained by LDA model has a structure of community, evidencing that the semantic information based relations formed a community structure.

Tables 5.34, 5.35, 5.36 and 5.37 shows the same properties of the networks, but for the model with 25 and 50 topics.

Nodes = 8,120
Edges = 8,245
Density = 0.000
Modularity = 0.996
Average Degree = 2.031
Average Weighted Degree = 0.743
Network Diameter = 25

Table 5.34: 25 Topic Based Network Statistics - Semantic Web Conference Series

Modularity = 0.996 Communities = 4,086	
Community Number	Number of Members
1st	57 (0.7%)
2nd	49 (0.6%)
3rd	48 (0.59%)

Table 5.35: Communities and modularity for 25 Topic Based Network

Nodes = 8,120
Edges = 8,075
Density = 0.000
Modularity = 0.996
Average Degree = 1.989
Average Weighted Degree = 0.730
Network Diameter = 18

Table 5.36: 50 Topic Based Network Statistics - Semantic Web Conference Series

Modularity = 0.996 Communities = 4,175	
Community Number	Number of Members
1st	56 (0.69%)
2nd	53 (0.65%)
3rd	38 (0.47%)

Table 5.37: Communities and modularity for 50 Topic Based Network

In the Figure 5.17 a three visualization of the original network and networks built by the two methods, Keywords and Topic based, proposed can be observed. As well as in the previous dataset, in this figure is shown clearly that with the two methods a density reduction can be accomplish, and weak edges between authors are gone. Additionally, it can be seen that the Original graph is very cluttered, much greater than the first dataset, meanwhile the others are very much clearer and allow to see nodes and edges in programs such as Gephi very easily.

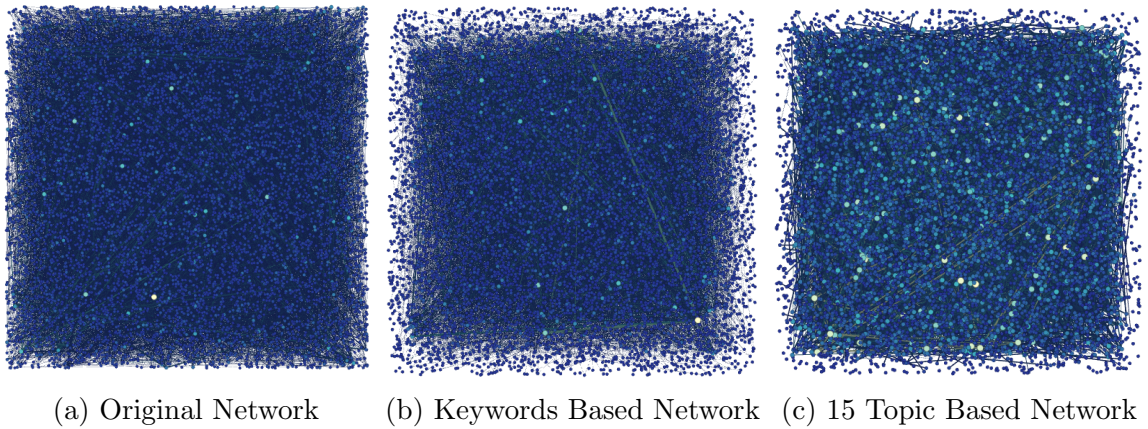


Figure 5.17: Semantic Web Conference Series Networks

Summarizing, the main results of the three networks can be compared in Table 5.38:

Nodes = 8,120 Edges = 20,332	Nodes = 8,120 Edges = 13,897	Nodes = 8,120 Edges = 8,502
Density = 0.001	Density = 0.000	Density = 0.000
Modularity = 0.904	Modularity = 0.715	Modularity = 0.995
Average Degree = 5.008	Average Degree = 3.423	Average Degree = 2.094
Avg. Weighted Degree = 5.787	Avg. Weighted Degree = 9.339	Avg. Weighted Degree = 0.764
Network Diameter = 22	Network Diameter = 21	Network Diameter = 29

Table 5.38: Comparison of structural properties for Original network, Keywords Based network and Topic Based network

5.4 Key Members discovery

In order to evaluate the key members discovery technique, the key members of each network (Traditional, Keyword Based and Topic Based Networks) should be compared with an external set of possible authors that are key members. The ideal procedure to recover this set would be to ask the administrators of the conferences to generate such set so that this will serve as a benchmark. However, this procedure is not available for any of the two data sets. Another way to obtain this benchmark set would be to ask experts (researchers) of the area for help to generate it. Nonetheless, this procedure may be biased by the sub area of such consulted expert.

Given this situation, a benchmark set of authors is generated using information of Google Scholar as shown in the Figures 5.18, 5.19 and 5.20. The procedure is the following. For each data set, the main concepts of the area are identified. For example, for the Pervasive and Ubiquitous computing data set, the concepts *Pervasive Computing* and *Ubiquitous Computing*, among others, are identified. Then a query on Google Scholar is performed by searching for the most cited authors on each one of those identified concepts (over 2000 citations). Next, the resulting list of searched authors in the query is crossed with the list of authors that belong to the data set, being that an author that does not have a paper on the data set will not be identified as key member by any of the networks. Finally, the remaining set of authors is compared with the key members set found by HITS and PageRank algorithm for each network.

Even though this procedure allows to generate a benchmark data set, it is accepted that this technique has some weak points. First, there are important authors on the data set that do not have a profile on Google Scholar, i.e., authors that belong to the authorship list of papers, but do not have a profile. In consequence, those authors will not appear on a search by the main concepts identified, and therefore, their citations are not counted.

Second, even though citations are a indicator of academic influence, they are correlated with both the number of years in activity and the number of authors of each sub area of research. Hereby, using the number of citations may not be the best criterion for identifying a benchmark set of key members. And third, the most influential researchers of the field could not match the most influential researchers among those who have participated in the conference. This could happen because some influential authors doesn't send their work to

the conferences in study but to specific journals.

Although, this procedure presents the weak points described before, it has both methodological as well as conceptual advantages. From the methodological perspective, it provides an unbiased source of data to generate the set. Google Scholar provides information about fields of interest and number of citations for authors. From the conceptual perspective, the most cited authors should be correlated with the authors with more papers, and this last variable correlated with the connections that an author has within a particular network. Given this two points, in this framework the benchmark set extracted from Google Scholar is used to evaluate the key member discovery methods.

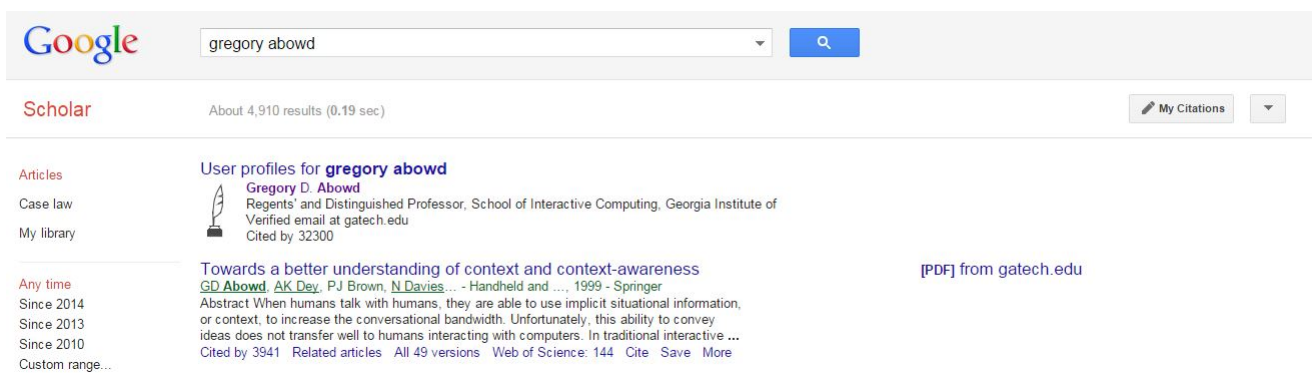


Figure 5.18: Search of an author in Google Scholar

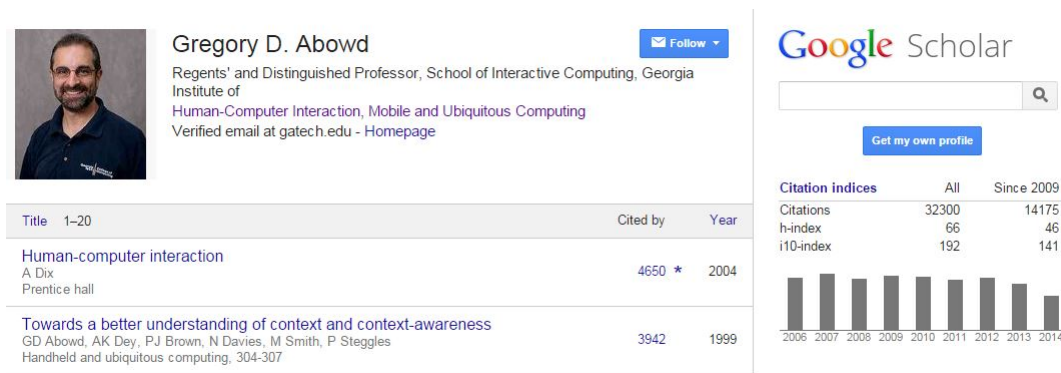


Figure 5.19: Profile of an author in Google Scholar

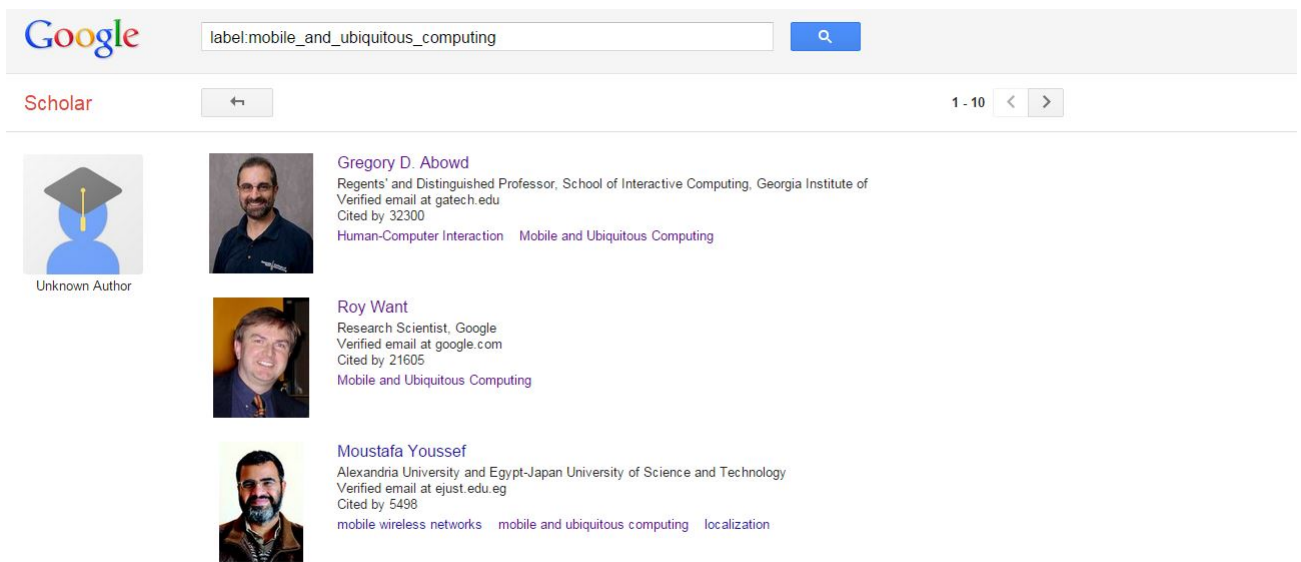


Figure 5.20: Search for investigation topics in Google Scholar

5.4.1 Key Members discovery in Ubiquitous and Pervasive Computing Conferences

The next tables shown the ranking obtained by the algorithm HITS and PageRank compared with the number of papers of every author, because that metric can give a glimpse of the importance of each author, since it is their level of production, and can be expected that authors with more papers written have more experience and more connections with other members.

Ranking	Number of Papers	HITS	PageRank
1	Gregory D. Abowd	Gregory D. Abowd	Gregory D. Abowd
2	Anind K. Dey	Gaetano Borriello	Anind K. Dey
3	Shwetak N.Patel	Sunny Consolvo	Gerhard Tröster
4	Gerhard Tröster	James Scott	James Scott
5	Sunny Consolvo	Anind K. Dey	Gaetano Borriello
6	Anthony LaMarca	Anthony LaMarca	Sunny Consolvo
7	James Scott	Steve Benford	Yvonne Rogers
8	Gaetano Borriello	Shahram Izadi	Anthony LaMarca
9	Khai N. Truong	Gerhard Tröster	Albrecht Schmidt
10	Stephen S. Intille	Timothy Sohn	Timothy Sohn

Table 5.39: Original Network - Ubiquitous and Pervasive Computing Conferences

As expected, in Table 5.39 it can be seen that the authors found by the algorithms are

mostly the same as the ranking by number of papers, that is because in the Original network no semantic information was added.

Ranking	Number of Papers	HITS	PageRank
1	Gregory D. Abowd	Gregory D. Abowd	Gregory D. Abowd
2	Anind K. Dey	James Scott	Anind K. Dey
3	Shwetak N.Patel	Gaetano Borriello	James Scott
4	Gerhard Tröster	Anind K. Dey	Albrecht Schmidt
5	Sunny Consolvo	Sunny Consolvo	Gerhard Tröster
6	Anthony LaMarca	Anthony LaMarca	Gaetano Borriello
7	James Scott	Albrecht Schmidt	Sunny Consolvo
8	Gaetano Borriello	Shwetak N.Patel	Anthony LaMarca
9	Khai N. Truong	Gerhard Tröster	Steve Benford
10	Stephen S. Intille	Steve Benford	Shwetak N.Patel

Table 5.40: Keyword Based Network - Ubiquitous and Pervasive Computing Conferences

Ranking	Number of Papers	HITS	PageRank
1	Gregory D. Abowd	Zhuoqing Morley Mao	Norman M. Sadeh
2	Anind K. Dey	Anthony D. Joseph	Hong Lu
3	Shwetak N.Patel	Keith Sklower	Christof Roduner
4	Gerhard Tröster	Kevin Lai	Roy Want
5	Sunny Consolvo	Lakshminarayanan Subramanian	Andrew T. Campbell
6	Anthony LaMarca	Yan Chen	George Coulouris
7	James Scott	Jimmy S. Shih	Lorrie Faith Cranor
8	Gaetano Borriello	Ion Stoica	Allison Woodruff
9	Khai N. Truong	Matthew Caesar	John Bowers
10	Stephen S. Intille	Weidong Cui	Ryan Libby

Table 5.41: Topic Based Network - 15 Topics - Ubiquitous and Pervasive Computing Conferences

Observing the rankings with both methods, new authors appears, nonetheless the biggest changes in rankings are shown by the Topic Based network as seen in Table 5.41, this can be explained because LDA model is a probabilistic model that infer topics, meanwhile keywords have been placed by a person, having an expert criterion, so the links are more alike with the number of papers that an author have.

In order to know the performance of the different ranking algorithms used, HITS and Pagerank, in the different methods, traditional performance measures of Data Mining (typi-

cally used in document retrieval) were computed: Recall ¹, Precision ² and F-measure ³. This type of metrics allows to measure how well an algorithm is capable of discover the relevance of different instances present in a dataset, in this case, the different experts.

	Recall	Precision	F
HITS	15%	17%	0.1587
HITS + Keywords	18%	20%	0.1905
HITS + LDA	21%	23%	0.2222
PageRank	15%	17%	0.1587
PageRank + Keywords	15%	17%	0.1587
PageRank + LDA	15%	17%	0.1587

Table 5.42: Key Members discovery recovery measures

From the Table 5.42, the better results are obtained by the HITS algorithm when filtering a network with topic based model. Though, the results are in overall not too high due to the limitations of the benchmark used.

In the following series of graphs for the dataset of Ubiquitous and Pervasive Computing Conferences, the ranking of authors obtained by the algorithms HITS and PageRank are shown in comparison with the number of papers of each author and also in comparison with the top 10 authors with more papers.

In Figures 5.21, 5.22 and 5.23, the comparison is made for the Keyword Based network authors:

¹Number of relevant cases retrieved (the probability that a relevant document would be retrieved in a random selection).

²Fraction of the correct cases returned (the probability that a retrieved document, randomly chosen, is relevant).

³Measure of a test accuracy that takes into consideration its precision and recall.

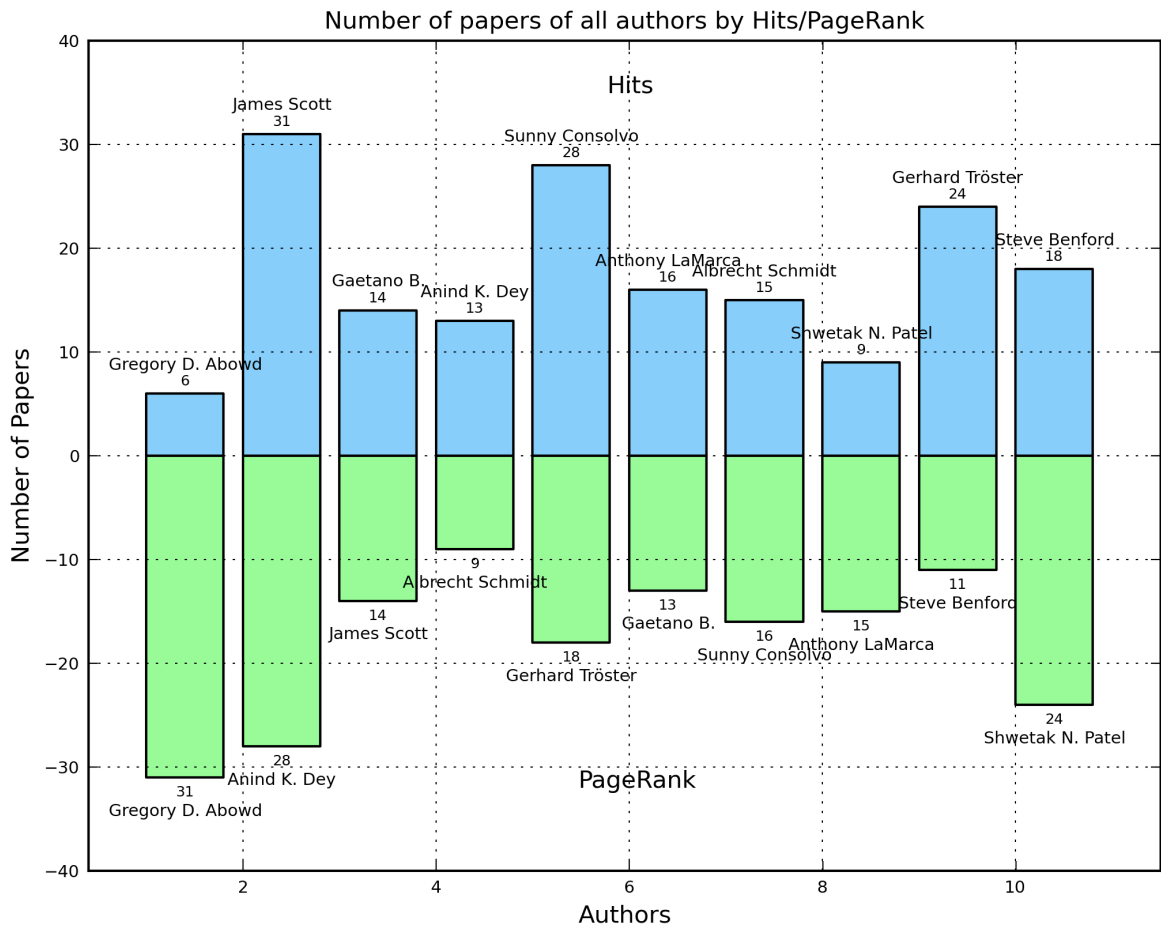


Figure 5.21: Number of Papers of top 10 authors found by HITS and PageRank for Keyword Based network

From the graph 5.21 it is observable that both algorithms rank almost the same authors but in a different order.

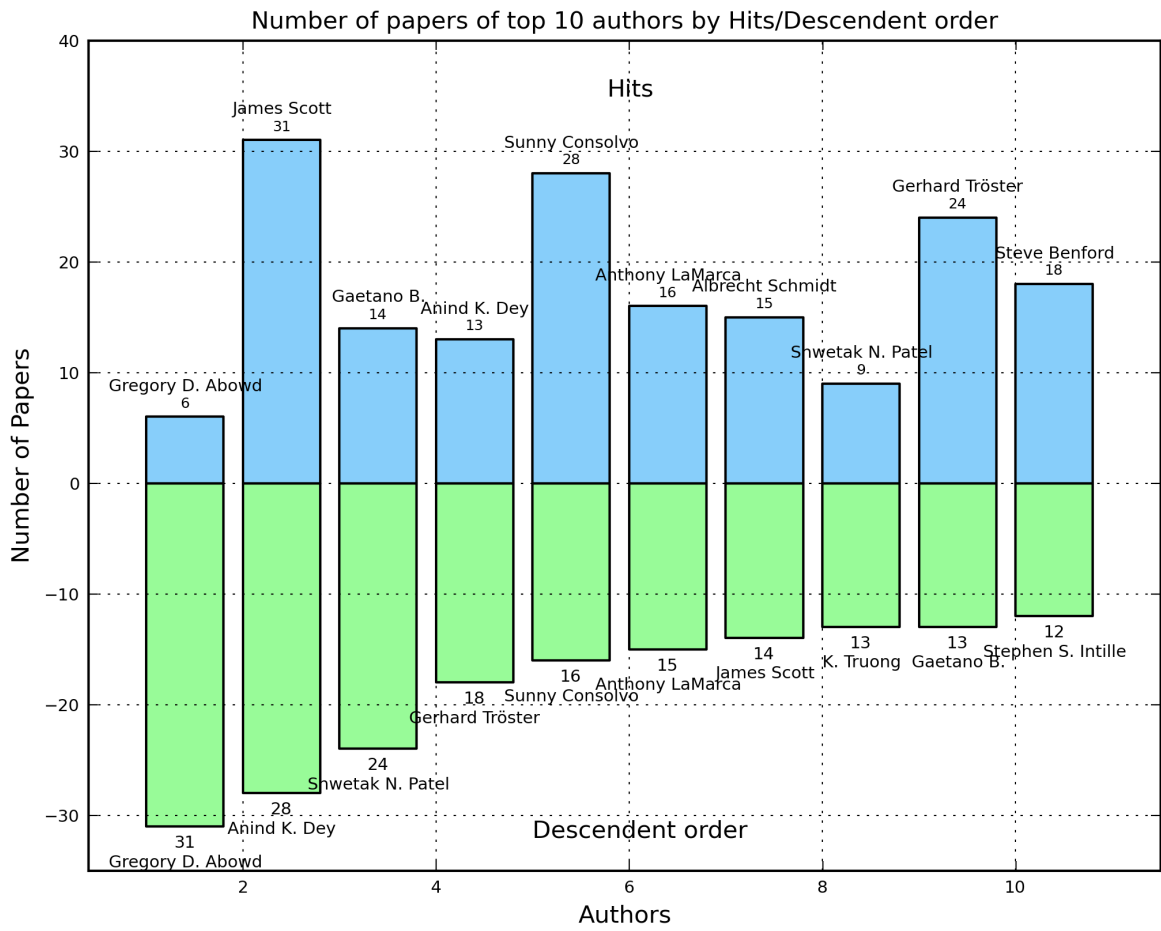


Figure 5.22: Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for Keyword Based network

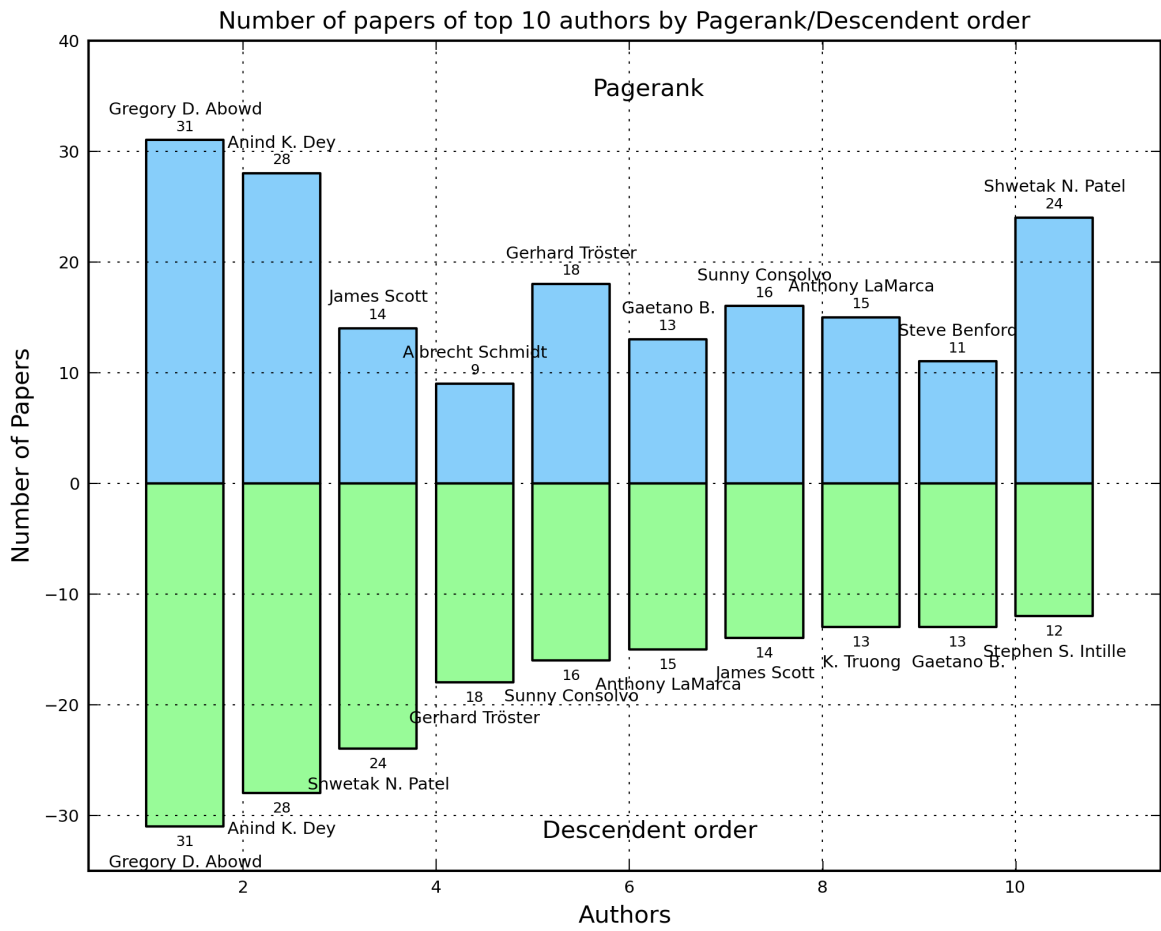


Figure 5.23: Number of Papers of top 10 authors found by PageRank compared to first top 10 authors with more papers for Keyword Based network

For the Topic based method, in the Figure 5.24, it is possible to compare the authors found by the algorithm HITS with their number of papers to the authors found with PageRank and their papers.

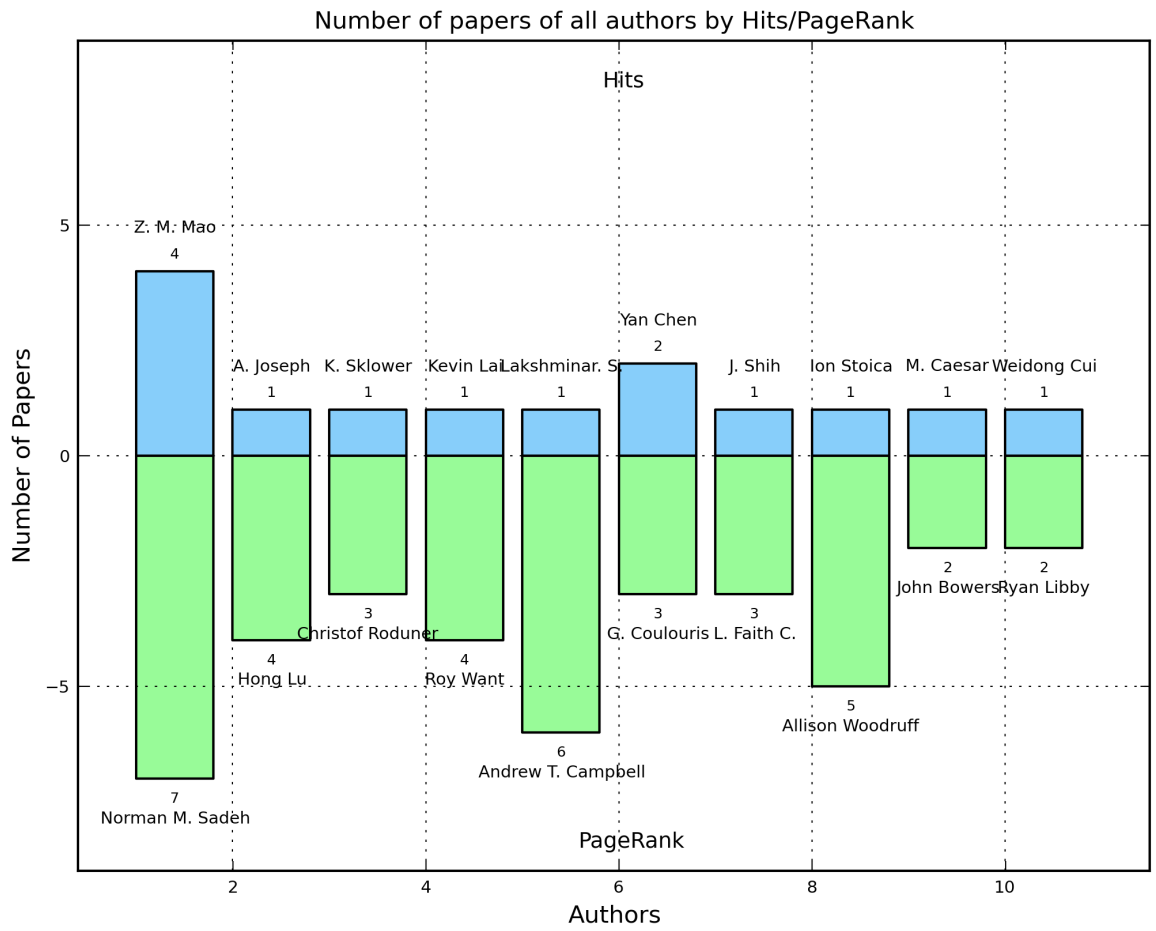


Figure 5.24: Number of Papers of top 10 authors found by HITS and PageRank for 15 Topics

In Figure 5.25 and 5.26, the same comparison is made but taking into account the top 10 authors with more papers in the entire dataset compared to HITS and PageRank, respectively.

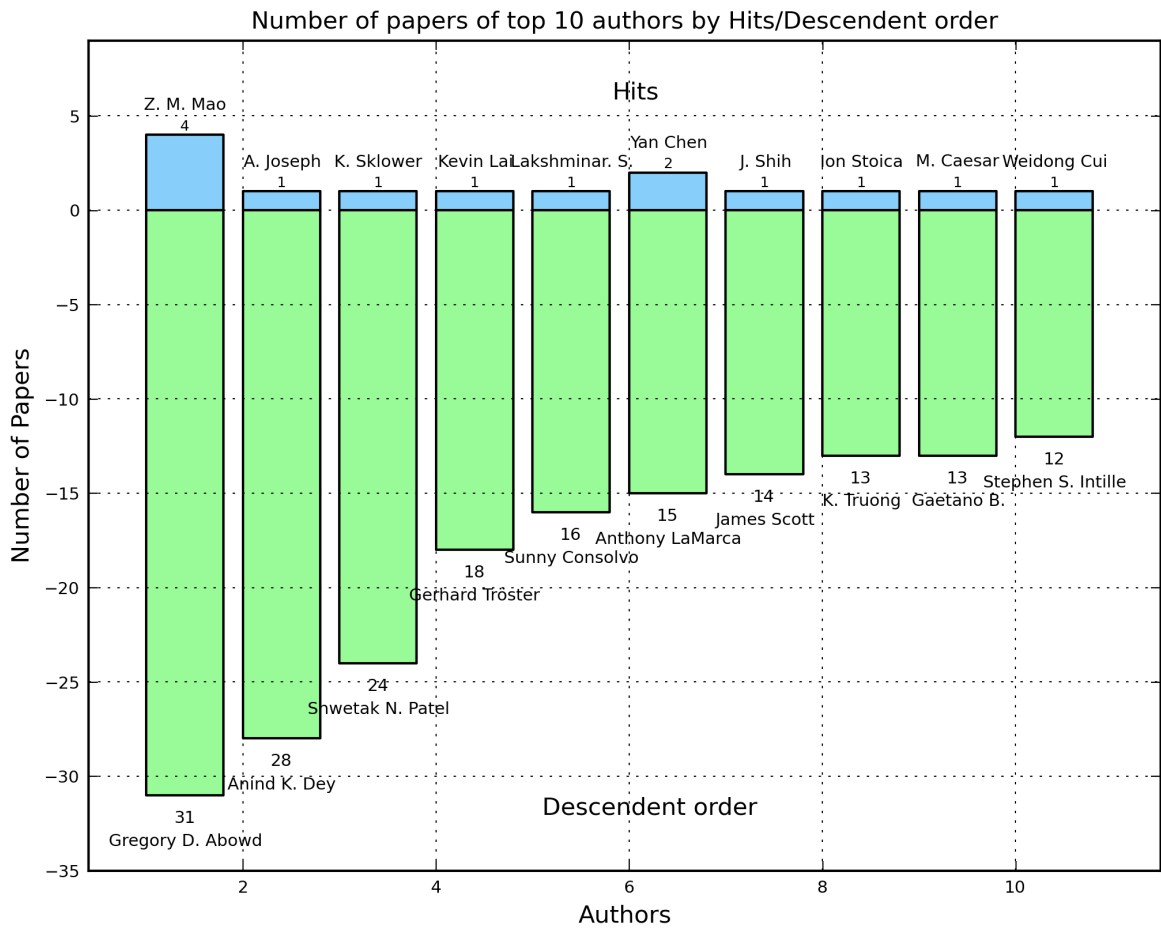


Figure 5.25: Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 15 Topics

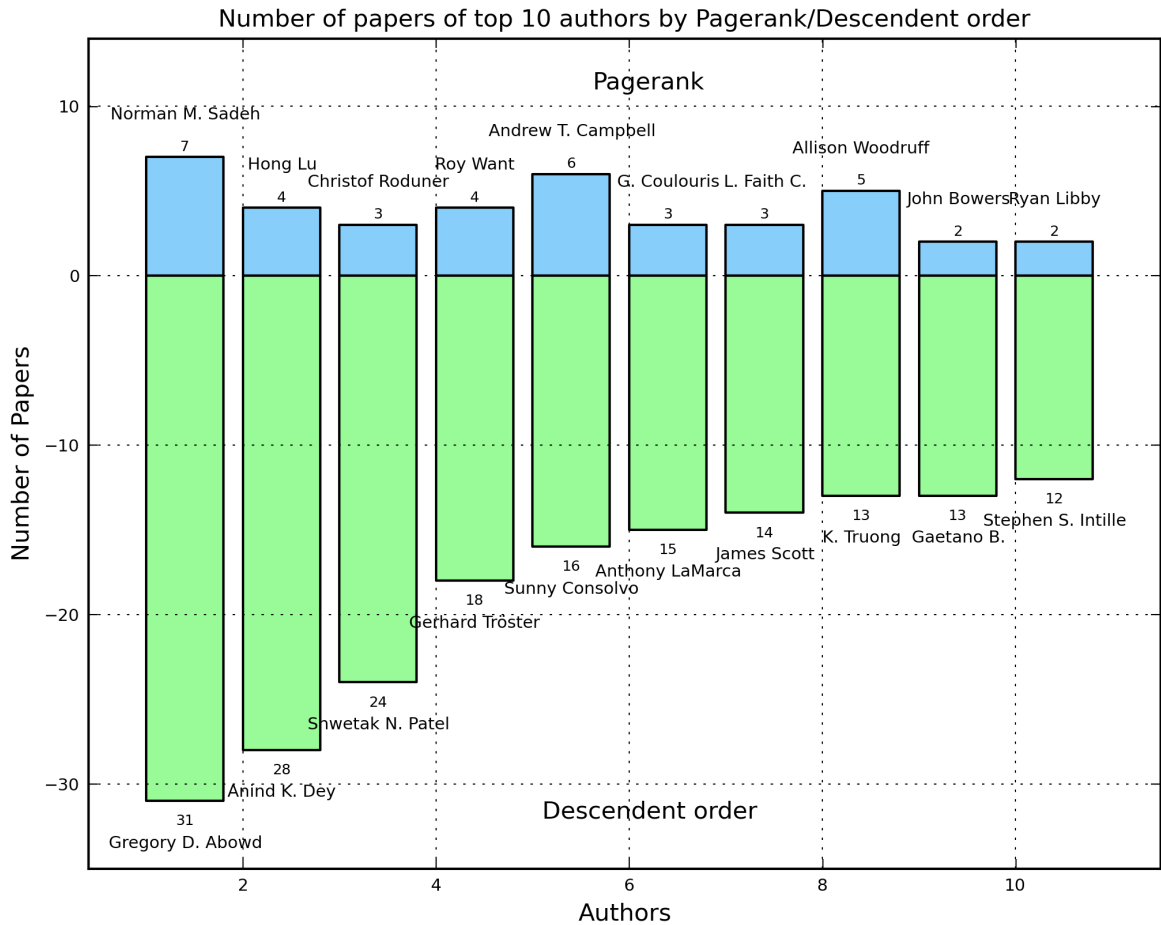


Figure 5.26: Number of Papers of top 10 authors found by PageRank compared to first top 10 authors with more papers for 15 Topics

An important result is the one discovered by PageRank algorithm in Topic Based Network, this algorithm ranked in 4th place the author *Roy Want*. The appearance of this author is relevant because this person appears 2nd in the Google Scholar benchmark when search was made with the tag *Ubiquitous Computing*, but didn't appear neither in the Original network nor in the top 10 list of authors with more papers. This means that adding semantic information is useful in terms on finding other key members that the Original network didn't show using HITS and PageRank.

5.4.2 Key Members discovery in Semantic Web Conference Series

As with the previous results for the first dataset, rankings were also built for this dataset. The tables below shown the same type of comparison between authors with more papers and the authors found by the algorithms in Keyword based and Topic based networks.

Ranking	Number of Papers	HITS	PageRank
1	Bijan Parsia	Zheng Chen	Zheng Chen
2	Stefan Decker	Yong Yu	Lei Zhang
3	Soeren Auer	Lei Zhang	Jiajun Bu
4	Ian Horrocks	Jiajun Bu	Yong Yu
5	Zheng chen	Peter Haase	Wolfgang Nejdl
6	Pascal Hitzler	Soeren Auer	Steffen Staab
7	Abraham Bernstein	Luciano Serafini	Axel Polleres
8	Yong Yu	Ian Horrocks	Soeren Auer
9	Asuncion Gomez Perez	Yue Pan	Peter Haase
10	Peter Haase	Wolfgang Nejdl	Ian Horrocks

Table 5.43: Original Network - Semantic Web Conference Series

Ranking	Number of Papers	HITS	PageRank
1	Bijan Parsia	Zheng Chen	Zheng Chen
2	Stefan Decker	Yong Yu	Wolfgang Nejdl
3	Soeren Auer	Lei Zhang	Jiajun Bu
4	Ian Horrocks	Jiajun Bu	Yong Yu
5	Zheng chen	Peter Haase	Lei Zhang
6	Pascal Hitzler	Soeren Auer	Steffen Staab
7	Abraham Bernstein	Ian Horrocks	Peter Haase
8	Yong Yu	Luciano Serafini	Axel Polleres
9	Asuncion Gomez Perez	Yue Pan	Soeren Auer
10	Peter Haase	Wolfgang Nejdl	Ian Horrocks

Table 5.44: Keyword Based Network - Semantic Web Conference Series

Ranking	Number of Papers	HITS	PageRank
1	Bijan Parsia	Sara Tonelli	Fabio Casati
2	Stefan Decker	Francesco Saverio Nucci	Jesus Contreras
3	Soeren Auer	Vincenzo Croce	Giovanni Tummarello
4	Ian Horrocks	Anastasia Moutzidou	Carole A Goble
5	Zheng chen	Gerard Casamayor	Hiroyuki Kitagawa
6	Pascal Hitzler	Maria Myllynen	Xiangyang Xue
7	Abraham Bernstein	Leo Wanner	Peter Wittenburg
8	Yong Yu	Horacio Saggion	Alessandro Mazzei
9	Asuncion Gomez Perez	Virpi Tarvainen	Guan Luo
10	Peter Haase	Tarja Koskentalo	Hong-luan Liao

Table 5.45: Topic Based Network - 15 Topics - Semantic Web Conference Series

As observed in Tables 5.43, 6.3 and 6.4 the same type of results were obtained, coincident with the first dataset, Topic based networks discover a lot of new authors, possible experts.

	Recall	Precision	F
HITS	14%	37%	0.2075
HITS + Keywords	14%	37%	0.2075
HITS + LDA	5%	13%	0.0755
PageRank	16%	40%	0.2264
PageRank + Keywords	14%	37%	0.2075
PageRank + LDA	4%	10%	0.0566

Table 5.46: Key Members discovery recovery measures

From Table 5.46 the results are slightly different from the first dataset, showing that in this case, the algorithms in the Keywords based network deliver better results than the others networks, of precision and recall, when retrieving an author randomly than the other networks. Nonetheless, neither of the algorithms is too good when searching for authors, it may also due to the imprecise benchmark.

In the Appendix Section 6.1, the graphs comparing the authors found by HITS and PageRank can be found for this dataset, for both Keyword Based and Topic Based networks.

Chapter 6

Conclusions and Future Work

In this thesis a community and key members methodology was applied to a novel dataset of Virtual communities of Practice, formed by two scientist networks of collaboration describe by the conferences in which their members publish their work and present it to all members. This methodology was an hybrid approach called SNA-KDD that mixes techniques of Social Networks Analysis with Data mining techniques, specially text mining methods, to discover not only structural properties but also that allows to aggregate the meaning of the content that members of the communities produce. This methodology was based in two methods to model and take into account semantic information into networks, to later analyse them. The first one, considered that Keywords was a good representation of an academic paper, and de the second one considered that a topic probabilistic model, called LDA can infer the latent topics of the papers in study.

The methods proposed use the content as a factor to decide whether a paper was similar to another, hence an author can be connected to another, i.e., an edge is built between them, allowing to design and construct new networks.

From the analysis and study of the networks built by both methods in the two communities it is possible to conclude that:

1. It is possible to model scientific collaboration networks using the SNA-KDD methodology, and study them as Virtual communities of practice. According to the results, it is possible to see that the metric modularity for all the networks built were highly, almost near 1, leaving evidence that these type of networks behave like communities, hence can be studied with SNA techniques and present structures that can be partitioned easily.

2. The type of relationships that were considered, given the nature of the network, can be modeled as networks, although they don't have a highly participation over the Internet, the relations are stronger than a post or an answer in a forum, because work with somebody else is a truly example of sharing knowledge. This can be demonstrated with the results of the metric average degree, and average weighted degree, because work with 4 or 5 members of a community in the form of scientific research is a greater tie than having a lot of replies from another user in an Internet interaction.
3. Key members detection was not as good as expected since the benchmark of comparison was not at precisely as one can desire. Hence, the algorithms used, HITS and PageRank, were not able to found as much as key members one wishes, they do find possible key members that are not visible from the traditional relationship. Therefore, adding semantics to study the VCoPs does provide with new key members, but one should have a good benchmark to probe it.
4. The results obtained by the topic probabilistic model LDA are very precise when obtaining topics, because even though a paper does not belong into a category of participation structured on a web, it talks about a much specific subject, thing that the model rescue in the topics. This proves that the methodology based on semantics helps to provide useful insights, extracting the main subjects of research out of a big corpus.
5. In terms of graph properties, with both methods in the VCoPs in study a clear reduction of density graph was accomplished, showing that adding semantic content to model the relationships can improve structural results besides eliminate relations that were not relevant, helping to clean the graph in order to seek for key members and detect sub communities.
6. In all the networks built by the methodology proposed (except for keyword based network of Ubiquitous and Pervasive Computing Conferences) the modularity was much greater than the modularity of the randomly generated network built of the same size, showing clearly that both VCoP have communities structure and that relations built on semantic information give as a result networks with community structure.

6.1 Future Work

Given the novelty of this approach with a completely different type of dataset than the traditional VCoP like a forum or a blog, the new lines of research that this work may present are:

1. Model the scientific networks with more relationship data, such as References. This type of data also generates a network of relations between an author who cite another researcher (take into account for his investigation), and some of them probably authors from the same community and others from related fields, thereby it would be interesting to see how these communities intertwine.
2. For the nature of these communities, the algorithms of ranking occupied for web pages are not the best suitable algorithms to search for key members. It would represent an interesting line of research to seek for better and more suitable algorithms taking into consideration the nature of the relations of the scientific collaboration networks.

Bibliography

- [1] Semantic web conference corpus. <http://data.semanticweb.org/>, 2014. Accessed: 25-11-2014.
- [2] Héctor Alvarez, Sebastián A Ríos, Felipe Aguilera, Eduardo Merlo, and Luis Guerrero. Enhancing social network analysis with a concept-based text mining approach to discover key members on a virtual community of practice. In *Knowledge-Based and Intelligent Information and Engineering Systems*, pages 591–600. Springer, 2010.
- [3] Grigoris Antoniou and Frank Van Harmelen. *A semantic web primer*. MIT press, 2004.
- [4] A.L. Barabási. *Linked: The New Science of Networks*. Perseus Pub., 2002.
- [5] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. Gephi: an open source software for exploring and manipulating networks. *ICWSM*, 8:361–362, 2009.
- [6] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.
- [7] Nicolás Ignacio Bersano-Méndez, Satu Elisa Schaeffer, and Javier Bustos-Jiménez. Metrics and models for social networks. In *Computational Social Networks*, pages 115–142. Springer, 2012.
- [8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
- [9] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.
- [10] John Breslin and Stefan Decker. The future of social networks on the internet: The need for semantics. *Internet Computing, IEEE*, 11(6):86–90, 2007.
- [11] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

- [12] Linda Carotenuto, William Etienne, Michael Fontaine, Jessica Friedman, Michael Muller, Helene Newberg, Matthew Simpson, Jason Slusher, Kenneth Stevenson, et al. Communityspace: toward flexible support for voluntary knowledge communities. In *Changing Places workshop, London*, 1999.
- [13] Peter J Carrington, John Scott, and Stanley Wasserman. *Models and methods in social network analysis*, volume 28. Cambridge university press, 2005.
- [14] Chaomei Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, 2006.
- [15] Diane J Cook and Lawrence B Holder. *Mining graph data*. John Wiley & Sons, 2006.
- [16] Lautaro Cuadra. Metodología de búsqueda de sub-comunidades mediante análisis de redes sociales y minería de datos. Master’s thesis, Universidad de Chile, 2011.
- [17] Lorenzo Ductor, Marcel Fafchamps, Sanjeev Goyal, and Marco J van der Leij. Social networks and research output. *Review of Economics and Statistics*, (0), 2011.
- [18] Gerhard Fischer. Communities of interest: Learning through the interaction of multiple knowledge systems. In *Proceedings of the 24th IRIS Conference*, pages 1–14. Department of Information Science, Bergen, 2001.
- [19] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.
- [20] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.
- [21] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.
- [22] Gaston L’Huillier, Sebastián A Ríos, Hector Alvarez, and Felipe Aguilera. Topic-based social network analysis for virtual communities of interests in the dark web. In *ACM SIGKDD Workshop on Intelligence and Security Informatics*, page 9. ACM, 2010.
- [23] Kalle Lyytinen and Youngjin Yoo. Ubiquitous computing. *Communicationsof the ACM*, 45(12):63, 2002.

- [24] Jay Marathe. Creating community online. *Durlacher Research Ltd*, 1999.
- [25] Ricardo Muñoz and Sebastián A Ríos. Overlapping community detection in vcop using topic models. In *KES*, pages 736–745, 2012.
- [26] Mark EJ Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.
- [27] Mark EJ Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5200–5205, 2004.
- [28] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
- [29] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.
- [30] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. <http://is.muni.cz/publication/884893/en>.
- [31] Sebastián A Ríos and Felipe Aguilera. Web intelligence on the social web. In *Advanced Techniques in Web Intelligence-I*, pages 225–249. Springer, 2010.
- [32] Sebastián A Ríos, Felipe Aguilera, Francisco Bustos, Tope Omitola, and Nigel Shadbolt. Leveraging social network analysis with topic models and the semantic web (extended). *Web Intelligence and Agent Systems*, 11(4):303–314, 2013.
- [33] Sebastián A Ríos, Felipe Aguilera, and Luis A Guerrero. Virtual communities of practice’s purpose evolution analysis using a concept-based mining approach. In *Knowledge-Based and Intelligent Information and Engineering Systems*, pages 480–489. Springer, 2009.
- [34] Sebastián A Ríos and Ricardo Muñoz. Dark web portal overlapping community detection based on topic models. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*, page 2. ACM, 2012.

- [35] Sebastián A Ríos and Ricardo Muñoz. Content patterns in topic-based overlapping communities. *The Scientific World Journal*, 2014, 2014.
- [36] Gerard Salton and Michael J McGill. Introduction to modern information retrieval. 1983.
- [37] Stanley Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.
- [38] Mark Weiser. Some computer science issues in ubiquitous computing. *Communications of the ACM*, 36(7):75–84, 1993.
- [39] Barry Wellman and Milena Gulia. Virtual communities as communities. *Communities in cyberspace*, pages 167–194.
- [40] Etienne Wenger. *Communities of practice: Learning, meaning, and identity*. Cambridge university press, 1998.
- [41] Etienne Wenger, Richard Arnold McDermott, and William Snyder. *Cultivating communities of practice: A guide to managing knowledge*. Harvard Business Press, 2002.
- [42] Rongying Zhao and Ju Wang. Visualizing the research on pervasive and ubiquitous computing. *Scientometrics*, 86(3):593–612, 2011.

Appendix

A Original Network HITS - PageRank plots Ubiquitous and Pervasive Computing Conferences

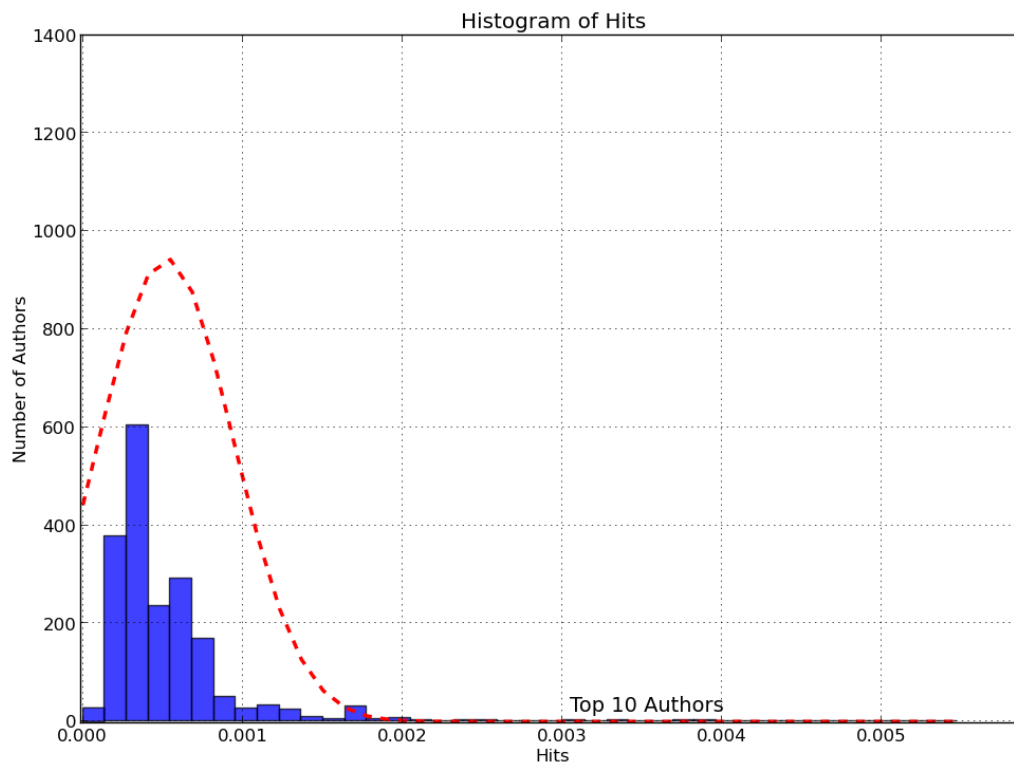


Figure 6.1: Histogram of HITS for Original Network

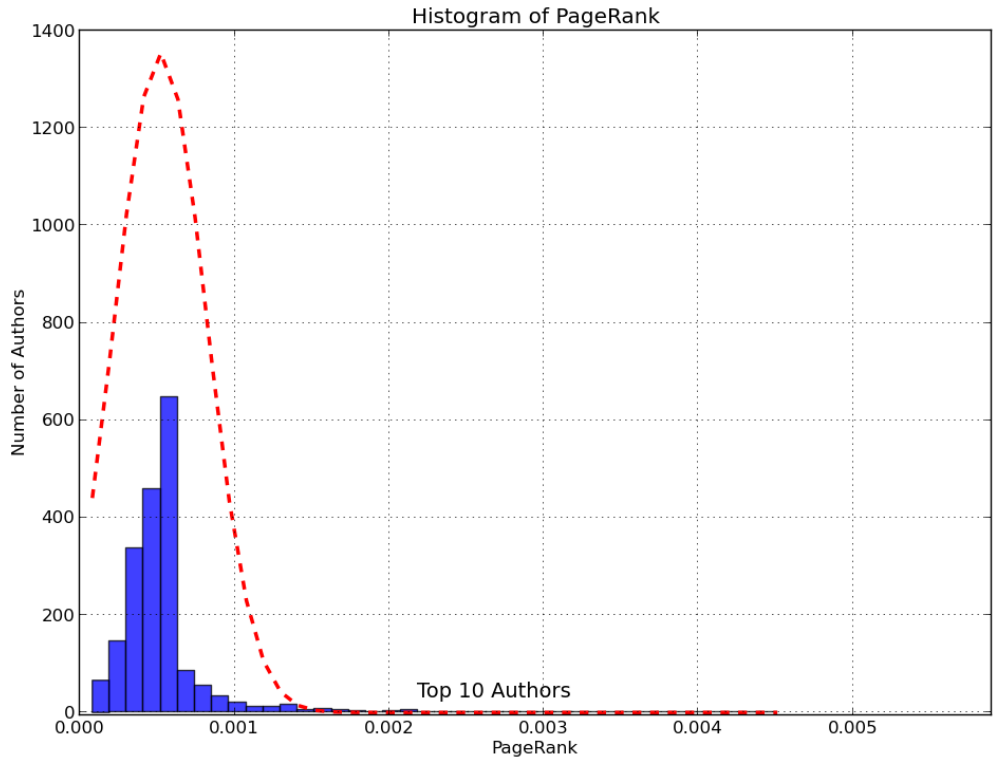


Figure 6.2: Histogram of PageRank for Original Network

B Keywords Based Network HITS - PageRank plots Ubiquitous and Pervasive Computing Conferences

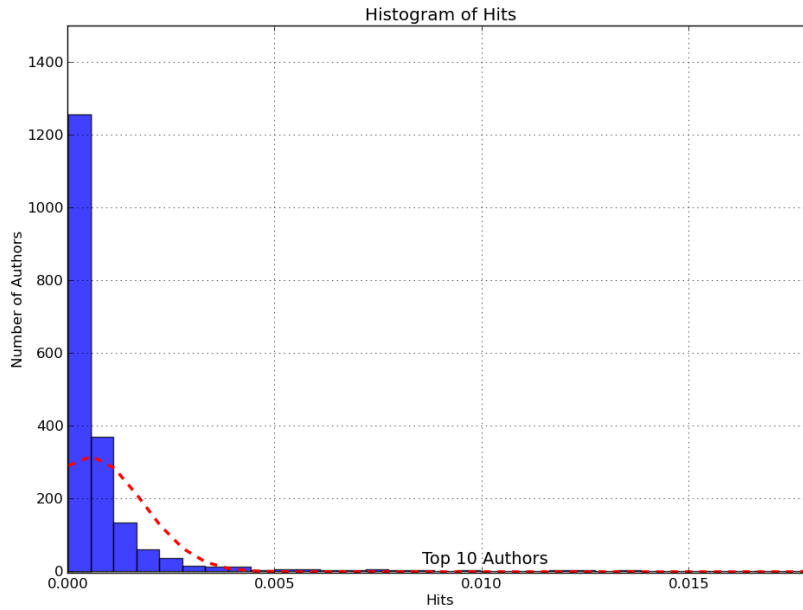


Figure 6.3: Histogram of HITS for Keyword Based Network

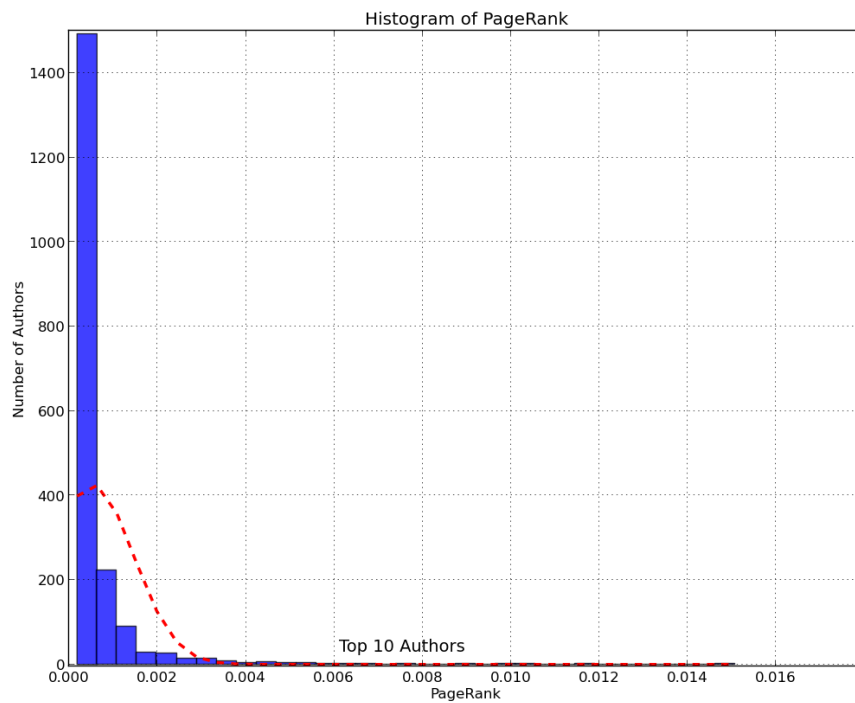


Figure 6.4: Histogram of PageRank for Keyword Based Network

C Topic Based, 15 topics Network HITS - PageRank plots Ubiquitous and Pervasive Computing Conferences

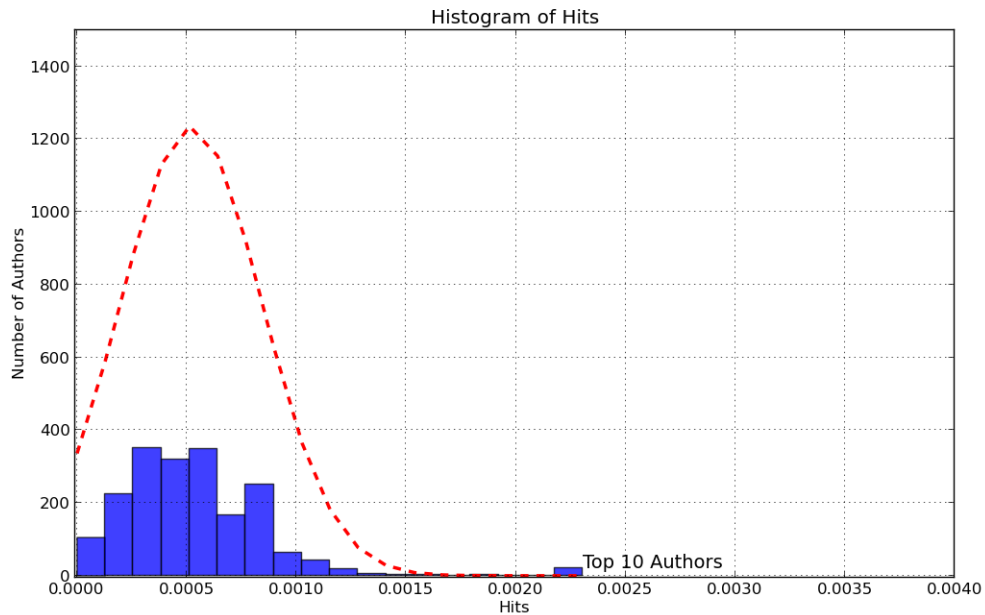


Figure 6.5: Histogram of HITS for Topic Based Network

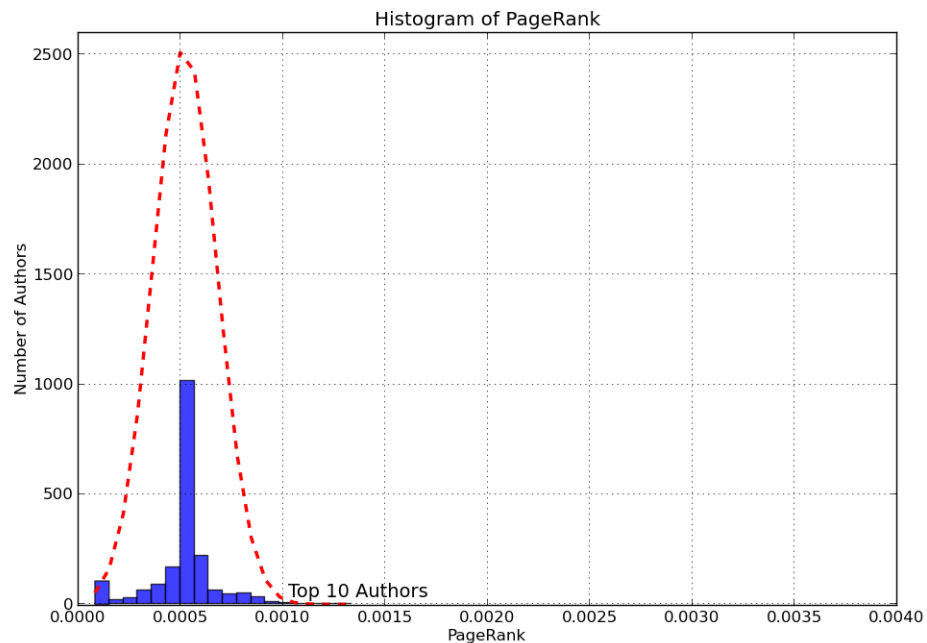


Figure 6.6: Histogram of PageRank for Topic Based Network

D Key Members discovery in Ubiquitous and Pervasive Computing Conferences for 25 Topics

Ranking	Number of Papers	HITS	PageRank
1	Gregory D. Abowd	Zhuoqing Morley Mao	Xianghua Ding
2	Anind K. Dey	Per Johansson	Alexander Varshavsky
3	Shwetak N.Patel	George Porter	Paul Dourish
4	Gerhard Tröster	Alexander Varshavsky	Eric Paulos
5	Sunny Consolvo	Yifei Jiang	Andrew T. Campbell
6	Anthony LaMarca	Yan Chen	Qin Lv
7	James Scott	Li Shang	Holger Junker
8	Gaetano Borriello	Lei Tan	Roy Want
9	Khai N. Truong	Takashi Suzuki	Emmanuel Munguia Tapia
10	Stephen S. Intille	Lakshminarayanan Subramanian	John Bowers

Table 6.1: Topic Based Network - 25 Topics - Ubiquitous and Pervasive Computing Conferences

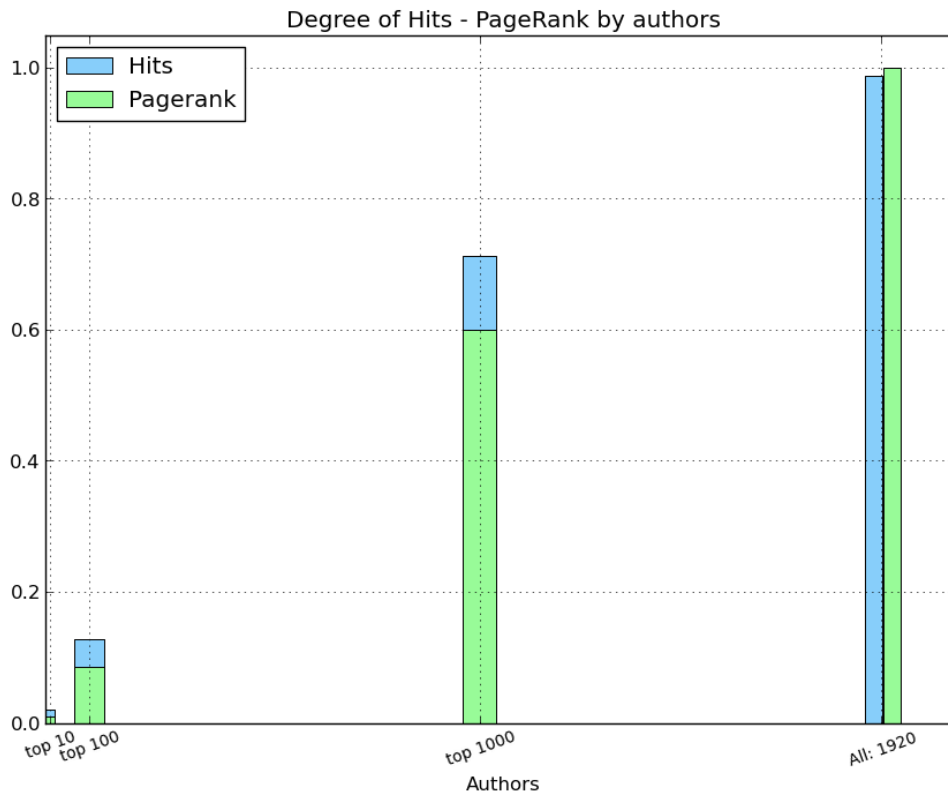


Figure 6.7: HITS and PageRank indexes across number of authors

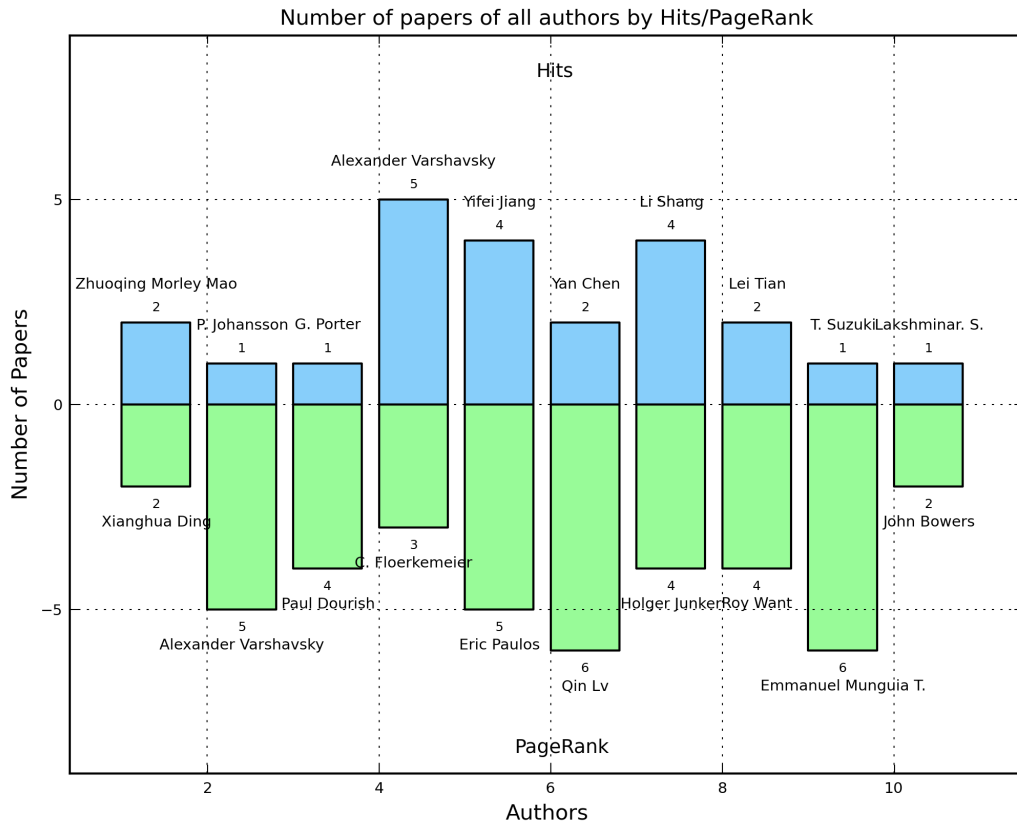


Figure 6.8: Number of Papers of top 10 authors found by HITS and PageRank for 25 Topics

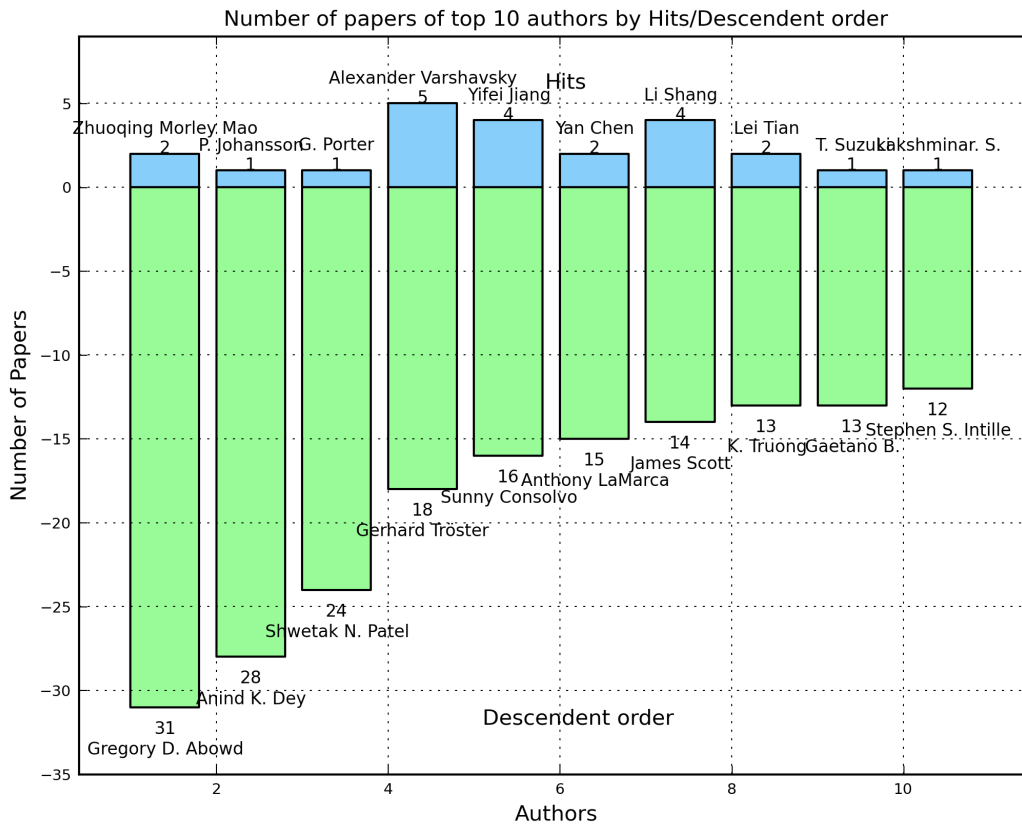


Figure 6.9: Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 25 Topics

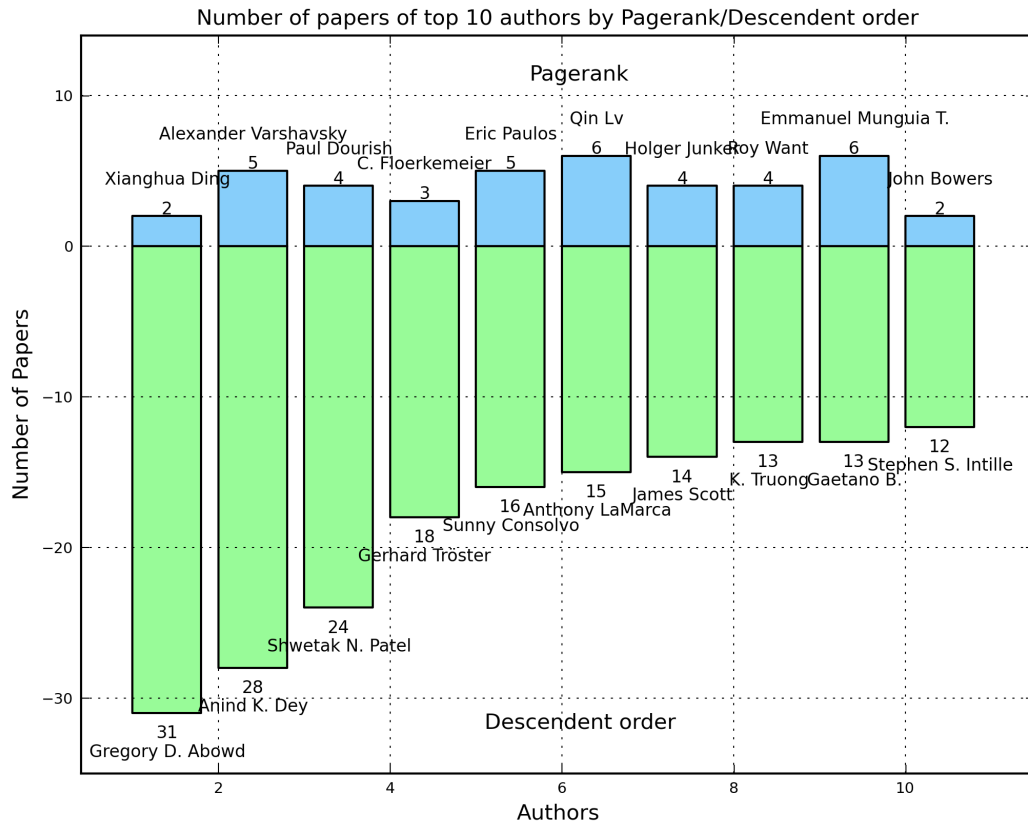


Figure 6.10: Number of Papers of top 10 authors found by PageRank compared to first top 10 authors with more papers for 25 Topics

E Key Members discovery in Ubiquitous and Pervasive Computing Conferences for 50 Topics

Ranking	Number of Papers	HITS	PageRank
1	Gregory D. Abowd	Zhuoqing Morley Mao	Martin Flintham
2	Anind K. Dey	Per Johansson	Roy Want
3	Shwetak N. Patel	Jimmy S. Shih	George Coulouris
4	Gerhard Tröster	Keith Sklower	John Bowers
5	Sunny Consolvo	Kevin Lai	Paul Dourish
6	Anthony LaMarca	Lakshminarayanan Subramanian	Eric Paulos
7	James Scott	Matthew Caesar	Stacey Kuznetsov
8	Gaetano Borriello	Mukund Seshadri	Zachary Pousman
9	Khai N. Truong	Randy H. Katz	Nick Tandavanitj
10	Stephen S. Intille	George Porter	Chanyou Hwang

Table 6.2: Topic Based Network - 50 Topics - Ubiquitous and Pervasive Computing Conferences

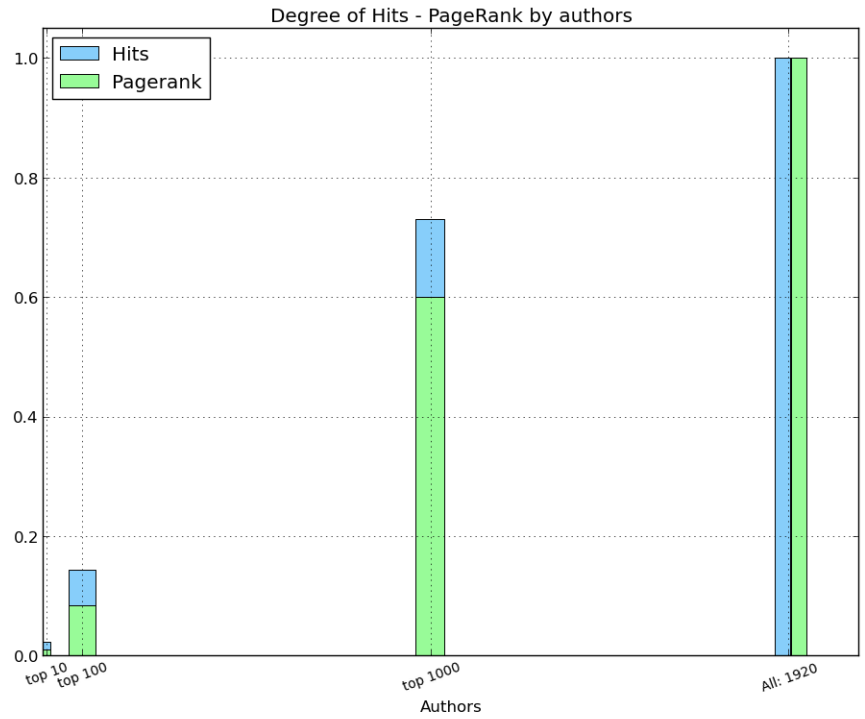


Figure 6.11: HITS and PageRank indexes across number of authors

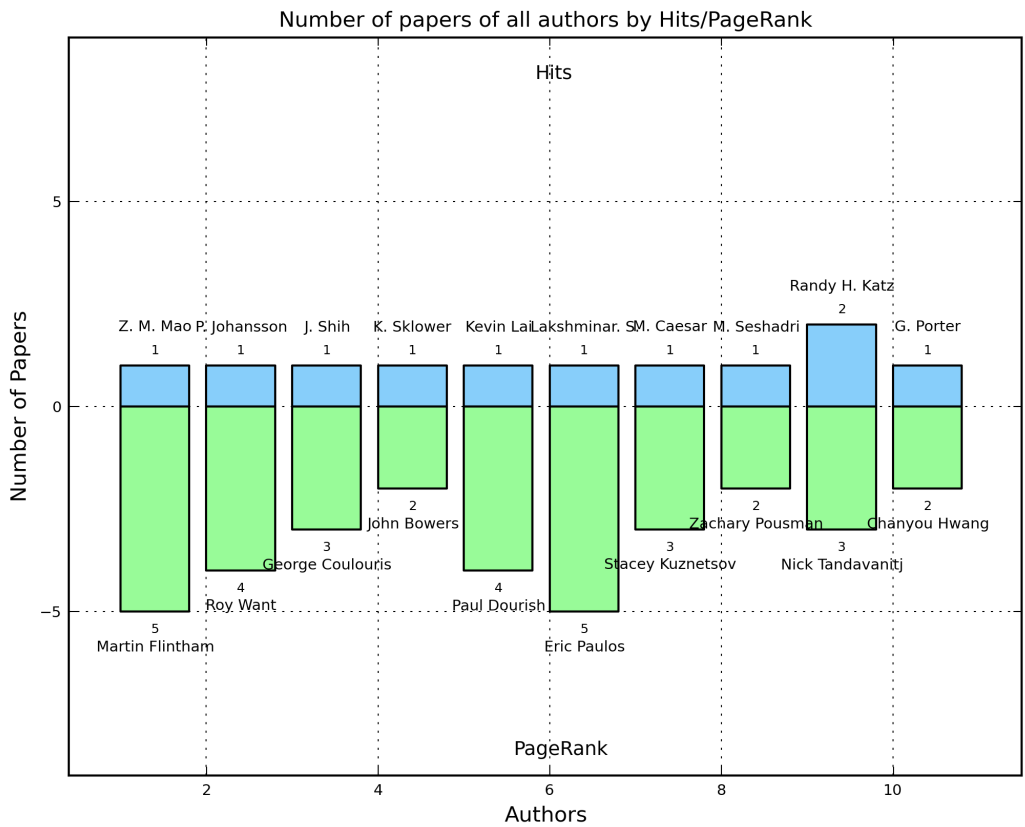


Figure 6.12: Number of Papers of top 10 authors found by HITS and PageRank for 50 Topics

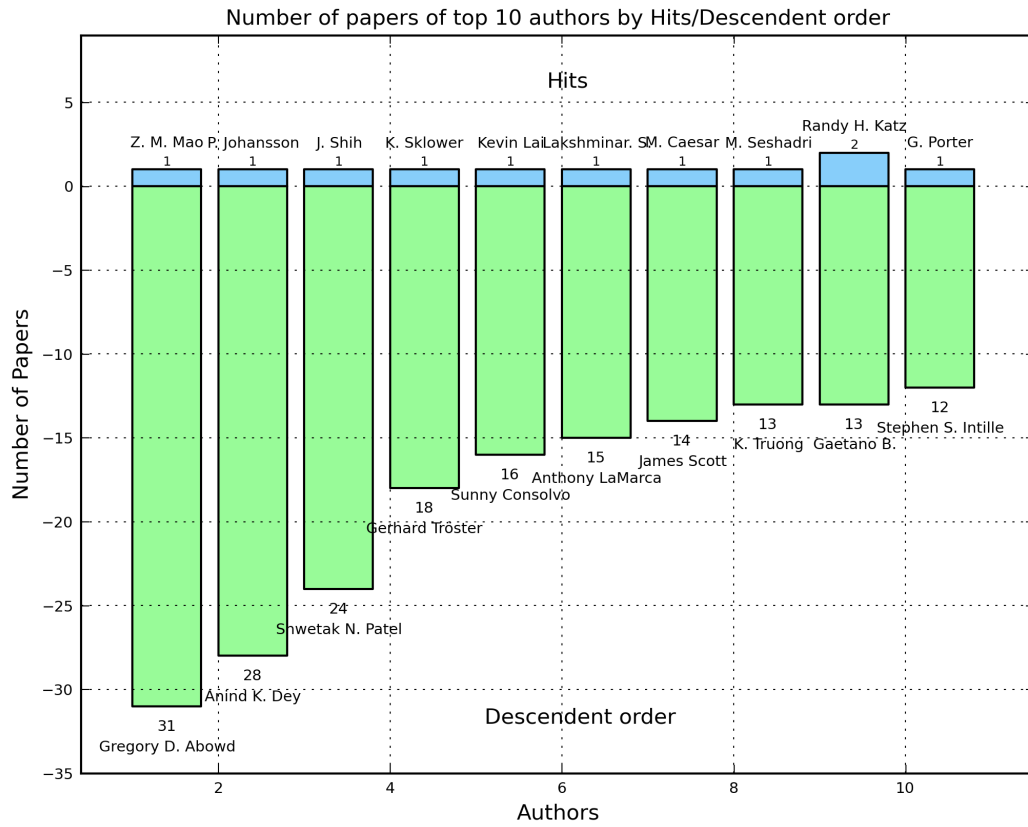


Figure 6.13: Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 50 Topics

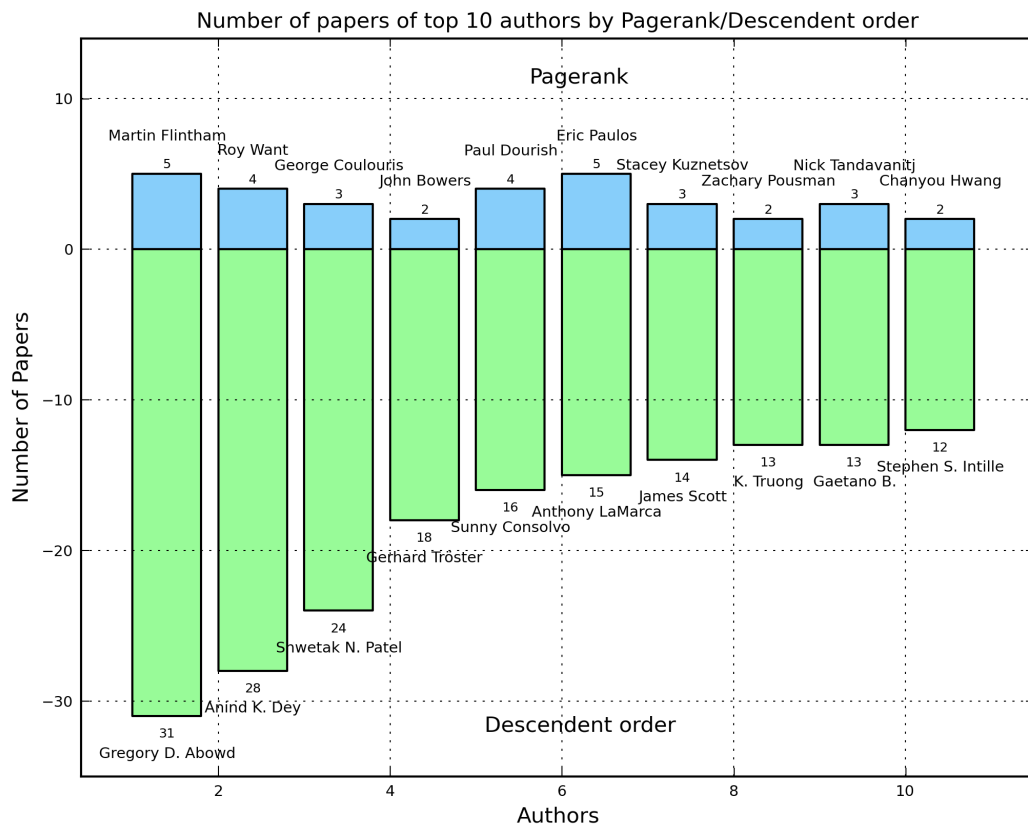


Figure 6.14: Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 50 Topics

F Key Members discovery in Semantic Web Conference Series for Keywords Based network

Ranking	Number of Papers	HITS	PageRank
1	Bijan Parsia	Zheng Chen	Zheng Chen
2	Stefan Decker	Yong Yu	Wolfgang Nejdl
3	Soeren Auer	Lei Zhang	Jiajun Bu
4	Ian Horrocks	Jiajun Bu	Yong Yu
5	Zheng chen	Peter Haase	Lei Zhang
6	Pascal Hitzler	Soeren Auer	Steffen Staab
7	Abraham Bernstein	Ian Horrocks	Peter Haase
8	Yong Yu	Luciano Serafini	Axel Polleres
9	Asuncion Gomez Perez	Yue Pan	Soeren Auer
10	Peter Haase	Wolfgang Nejdl	Ian Horrocks

Table 6.3: Keyword Based Network - Semantic Web Conference Series

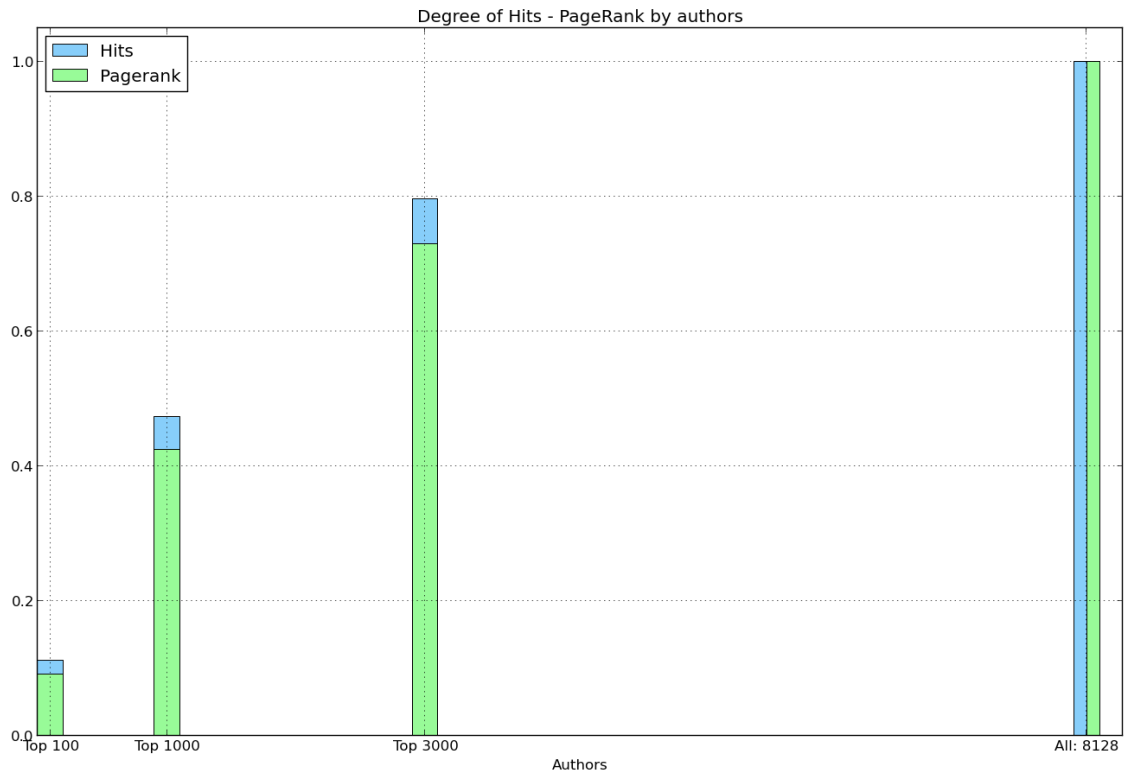


Figure 6.15: HITS and PageRank indexes across number of authors

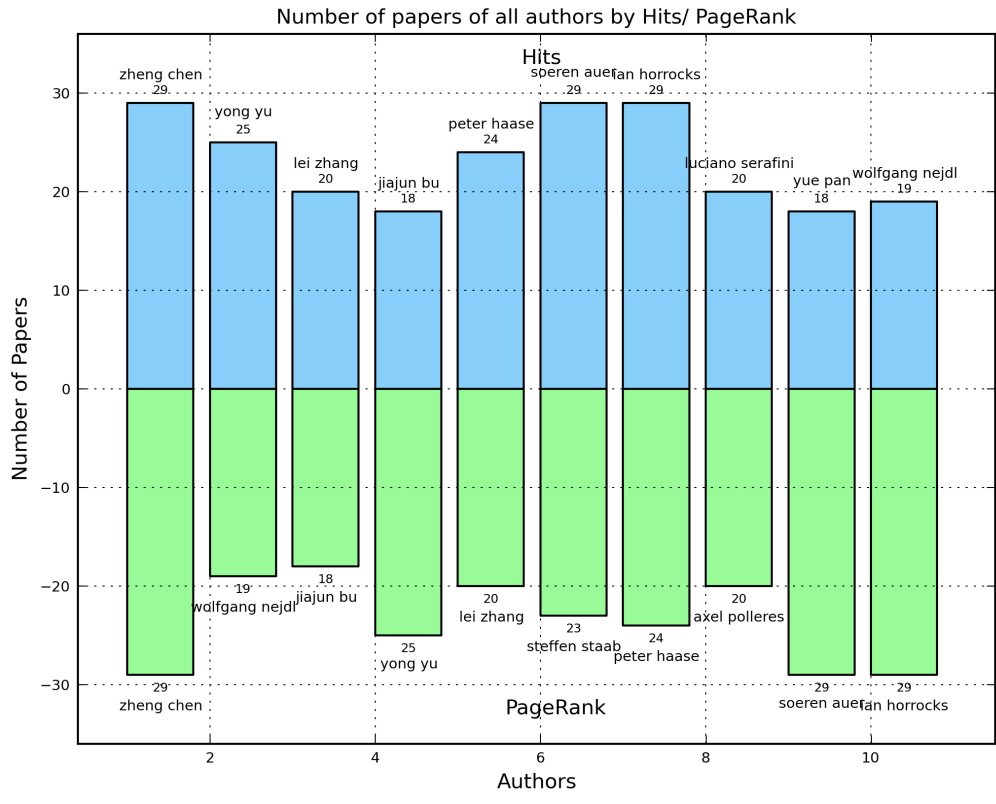


Figure 6.16: Number of Papers of top 10 authors found by HITS and PageRank for Keywords Based network

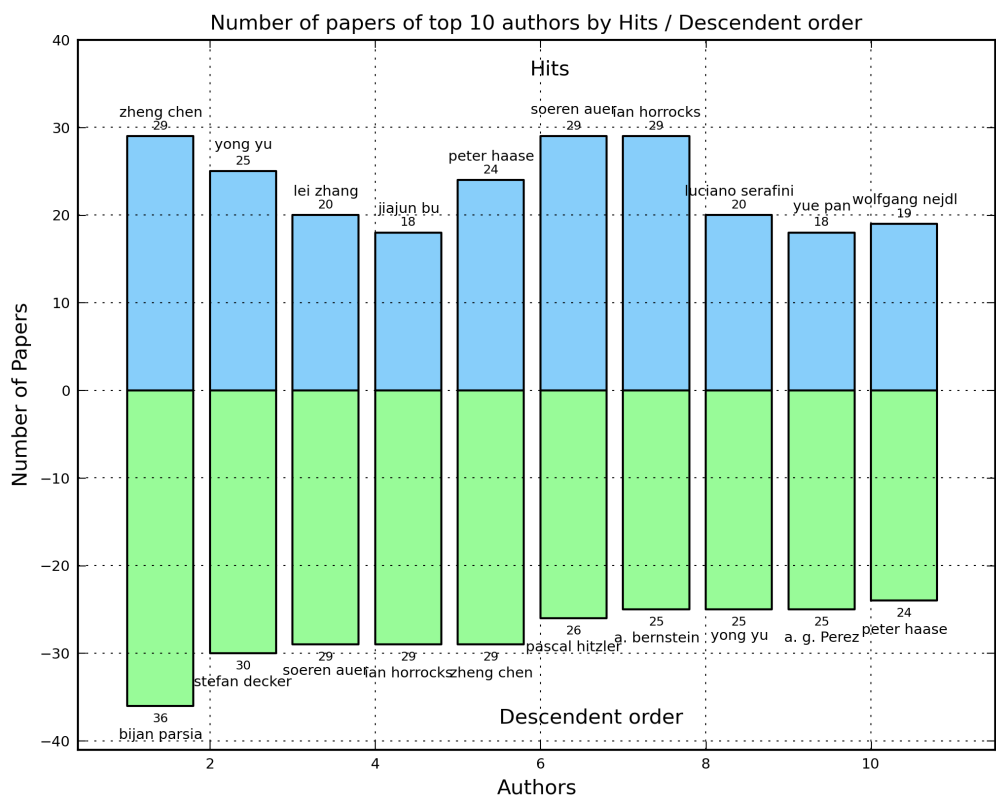


Figure 6.17: Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for Keywords Based network

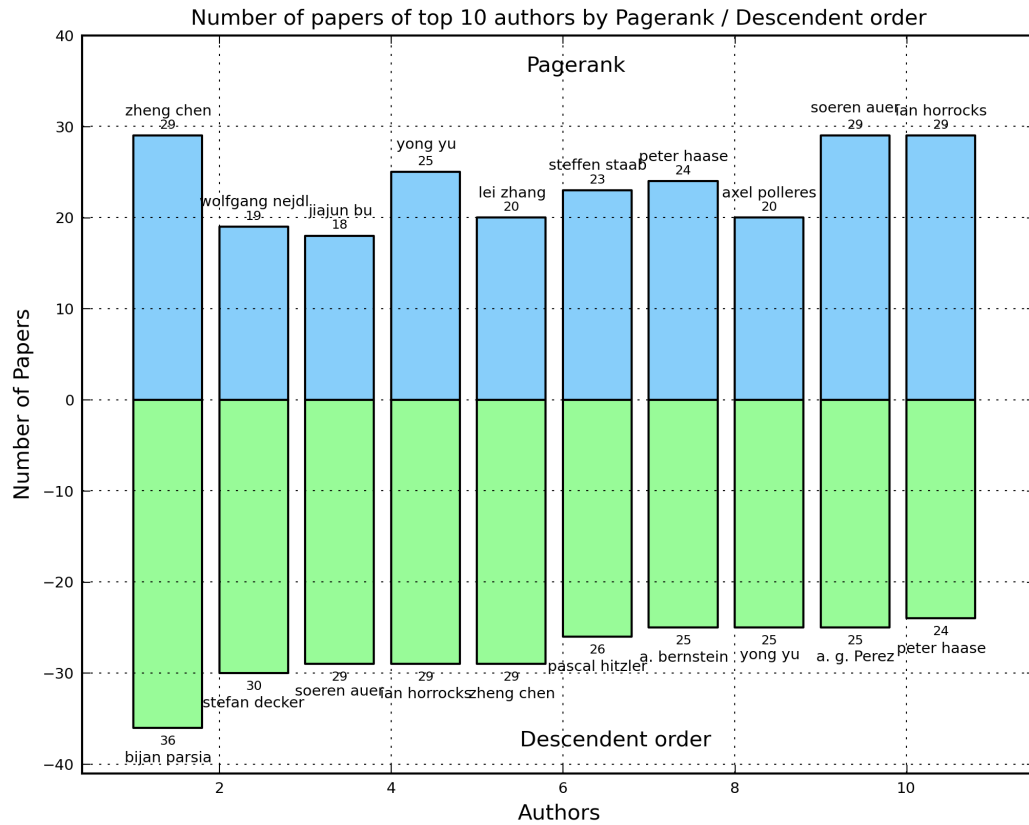


Figure 6.18: Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for Keywords Based network

G Key Members discovery in Semantic Web Conference Series for 15 Topics

Ranking	Number of Papers	HITS	PageRank
1	Bijan Parsia	Sara Tonelli	Fabio Casati
2	Stefan Decker	Francesco Saverio Nucci	Jesus Contreras
3	Soeren Auer	Vincenzo Croce	Giovanni Tummarello
4	Ian Horrocks	Anastasia Moutzidou	Carole A Goble
5	Zheng chen	Gerard Casamayor	Hiroyuki Kitagawa
6	Pascal Hitzler	Maria Myllynen	Xiangyang Xue
7	Abraham Bernstein	Leo Wanner	Peter Wittenburg
8	Yong Yu	Horacio Saggion	Alessandro Mazzei
9	Asuncion Gomez Perez	Virpi Tarvainen	Guan Luo
10	Peter Haase	Tarja Koskentalo	Hong-luan Liao

Table 6.4: Topic Based Network - 15 Topics - Semantic Web Conference Series

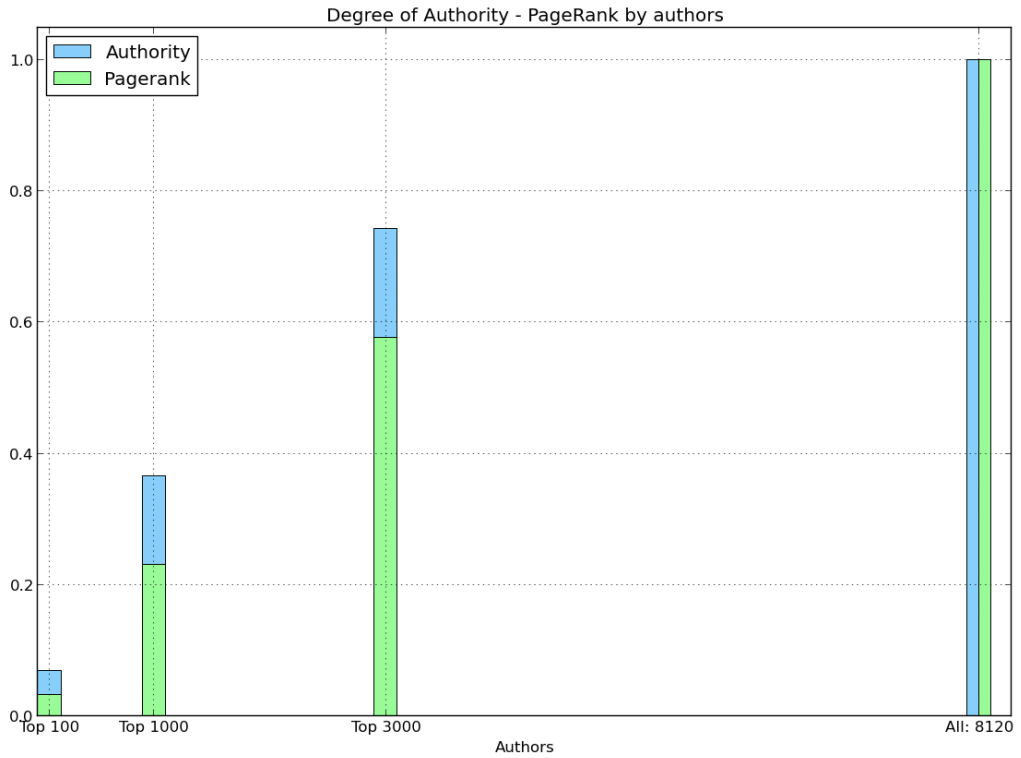


Figure 6.19: HITS and PageRank indexes across number of authors

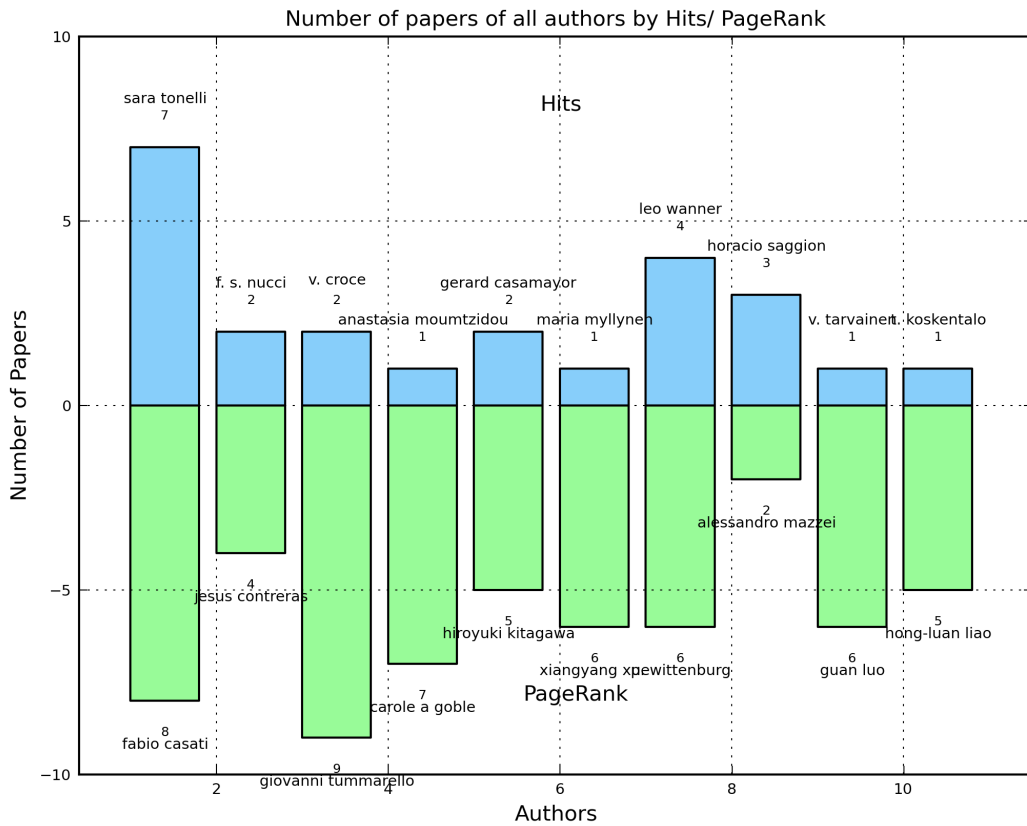


Figure 6.20: Number of Papers of top 10 authors found by HITS and PageRank for 15 Topics

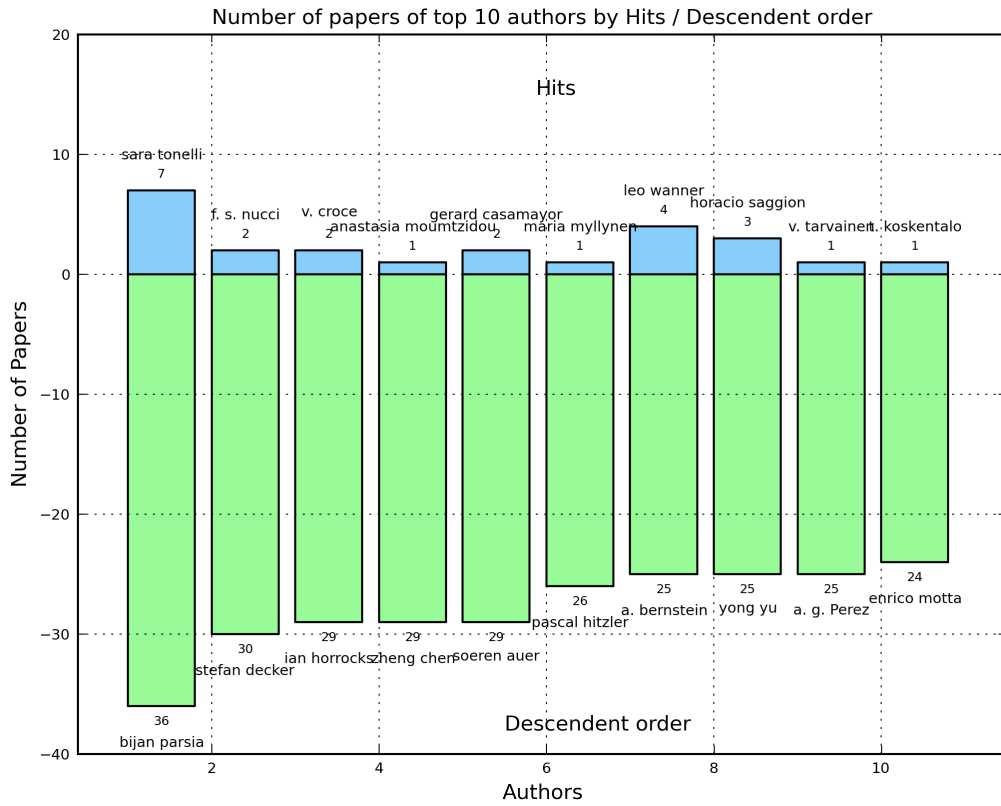


Figure 6.21: Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 15 Topics

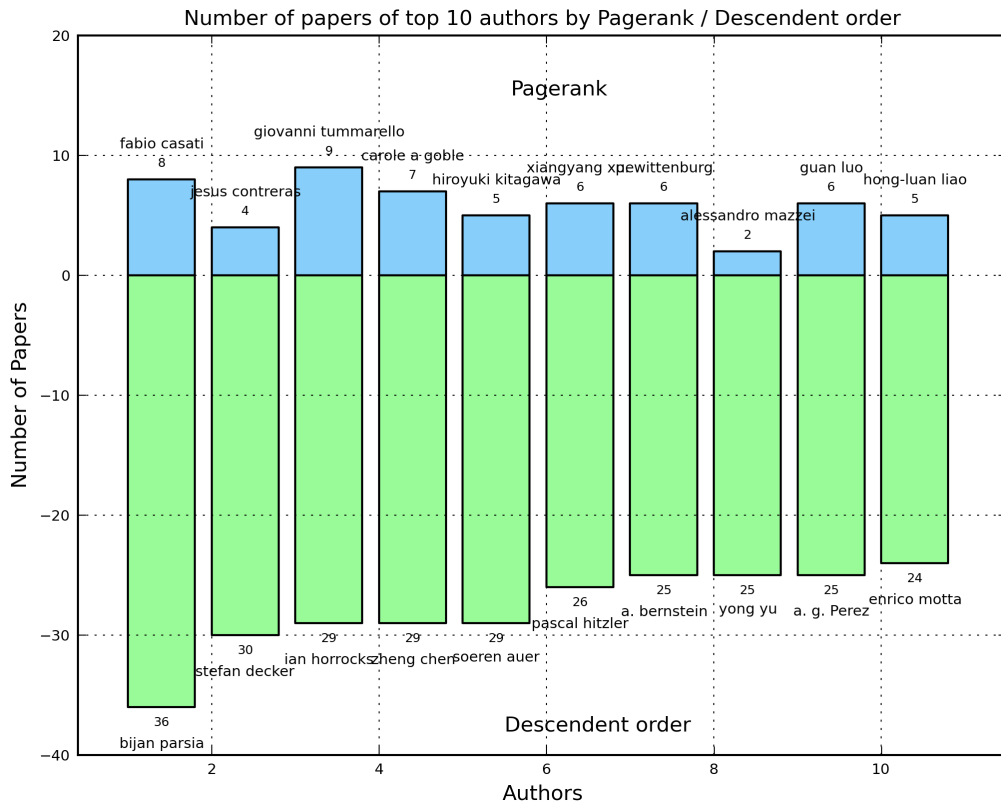


Figure 6.22: Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 15 Topics

H Key Members discovery in Semantic Web Conference Series for 25 Topics

Ranking	Number of Papers	HITS	PageRank
1	Bijan Parsia	Vincenzo Croce	Guan Luo
2	Stefan Decker	Francesco Saverio Nucci	Hiroyuki Kitagawa
3	Soeren Auer	Ari Karppinen	Aimin Pan
4	Ian Horrocks	Juergen Mossgraber	Yorick Wilks
5	Zheng chen	Gerard Casamayor	Raymond Fergerson
6	Pascal Hitzler	Virpi Tarvainen	Kuansan Wang
7	Abraham Bernstein	Nadjet Bouayad Agha	Hang Li
8	Yong Yu	Ulrich Buegel	Masahiro Hamasaki
9	Asuncion Gomez Perez	Harald Bosch	Guo Tong Xie
10	Peter Haase	Norihide Kitaoka	T.H. Tse

Table 6.5: Topic Based Network - 25 Topics - Semantic Web Conference Series

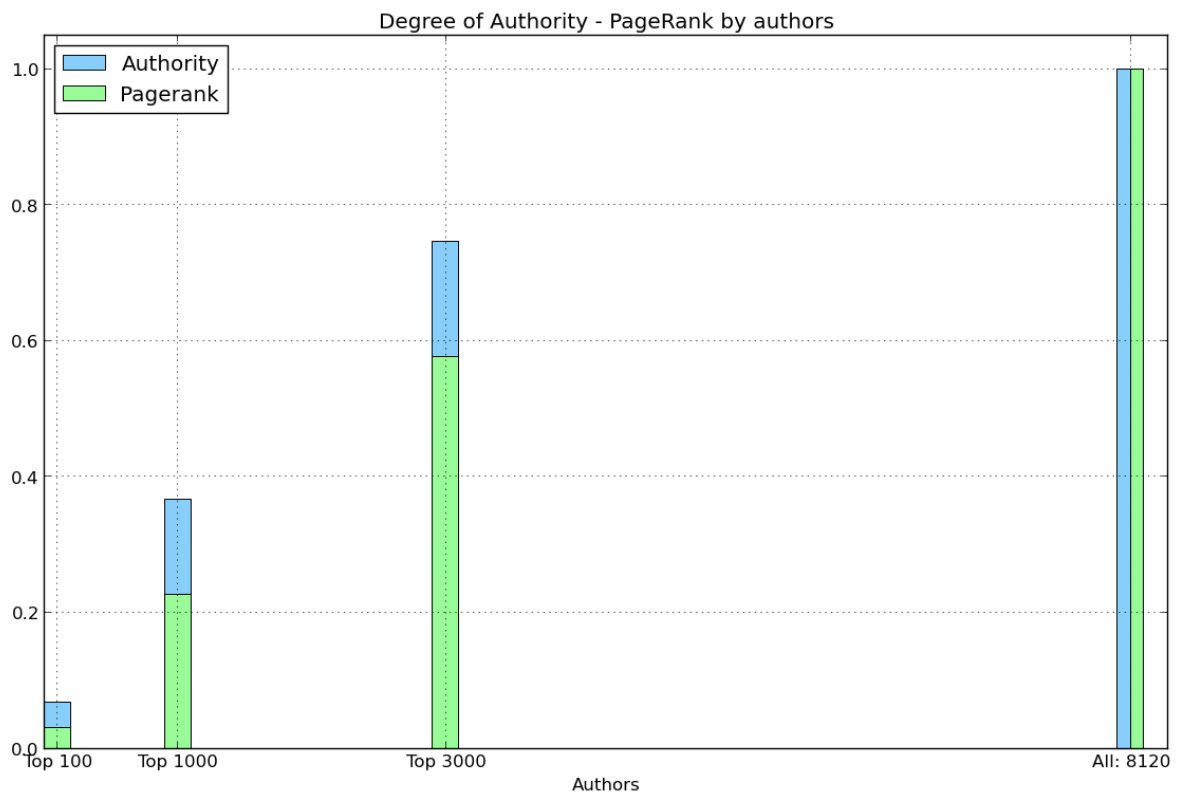


Figure 6.23: HITS and PageRank indexes across number of authors

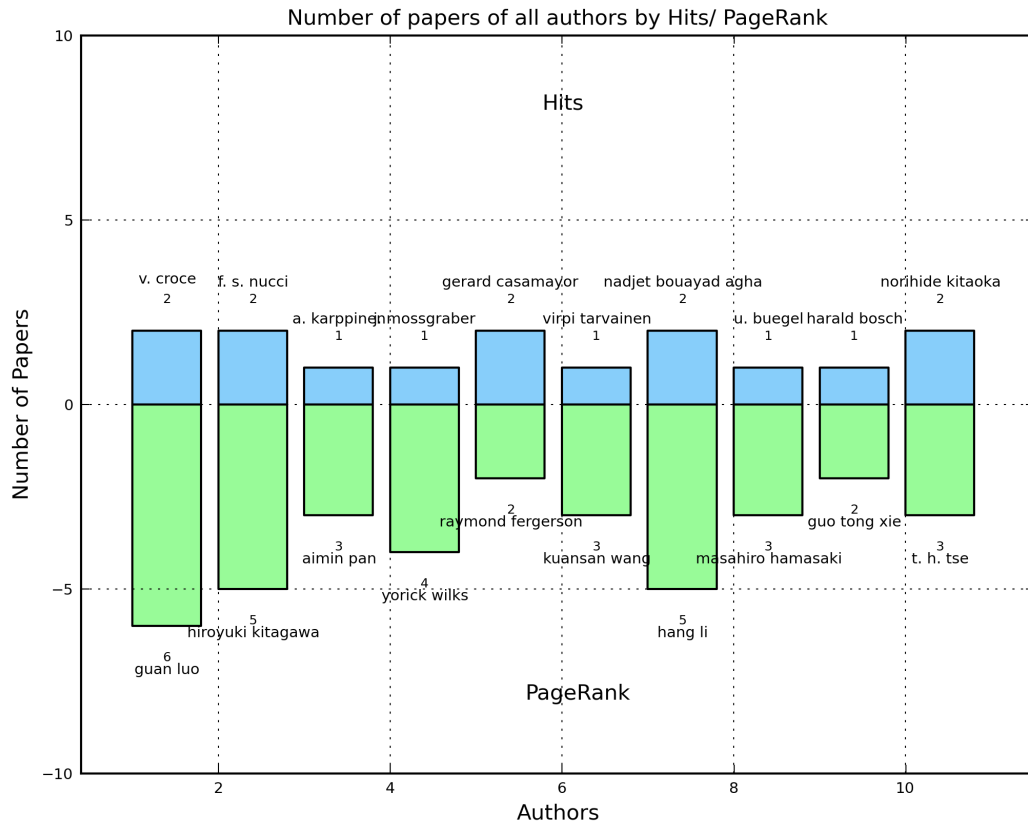


Figure 6.24: Number of Papers of top 10 authors found by HITS and PageRank for 25 Topics

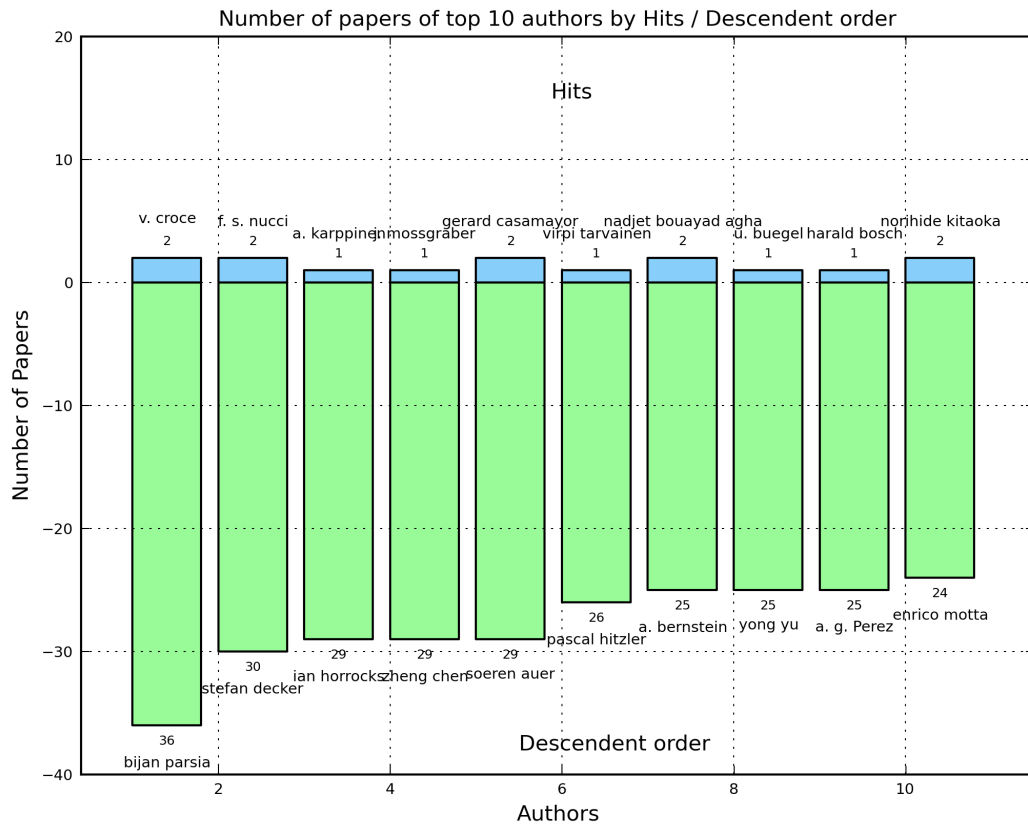


Figure 6.25: Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 25 Topics

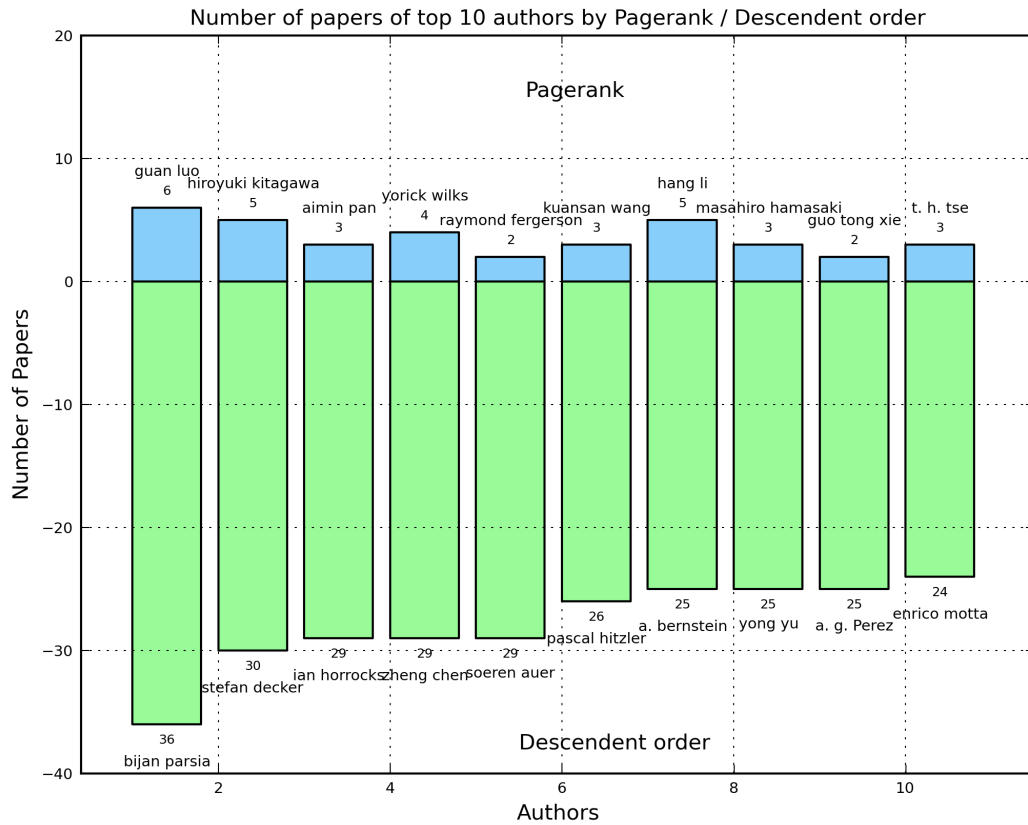


Figure 6.26: Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 25 Topics

I Key Members discovery in Semantic Web Conference Series for 50 Topics

Ranking	Number of Papers	HITS	PageRank
1	Bijan Parsia	Vincenzo Croce	Monica Monachini
2	Stefan Decker	Francesco Saverio Nucci	Chu Ren Huang
3	Soeren Auer	Nadjet Bouayad Agha	Guan Luo
4	Ian Horrocks	Desiree Hilbring	Raymond Fergerson
5	Zheng chen	Gerard Casamayor	Kuansan Wang
6	Pascal Hitzler	Anastasia Mountzidou	Sebastian Schaffert
7	Abraham Bernstein	Ioannis Kompatsiaris	Thierry Declerck
8	Yong Yu	Juergen Mossgraber	Guo Tong Xie
9	Asuncion Gomez Perez	Harald Bosch	Massimo Poesio
10	Peter Haase	Chiyomi Miyajima	Christian Borgs

Table 6.6: Topic Based Network - 50 Topics - Semantic Web Conference Series

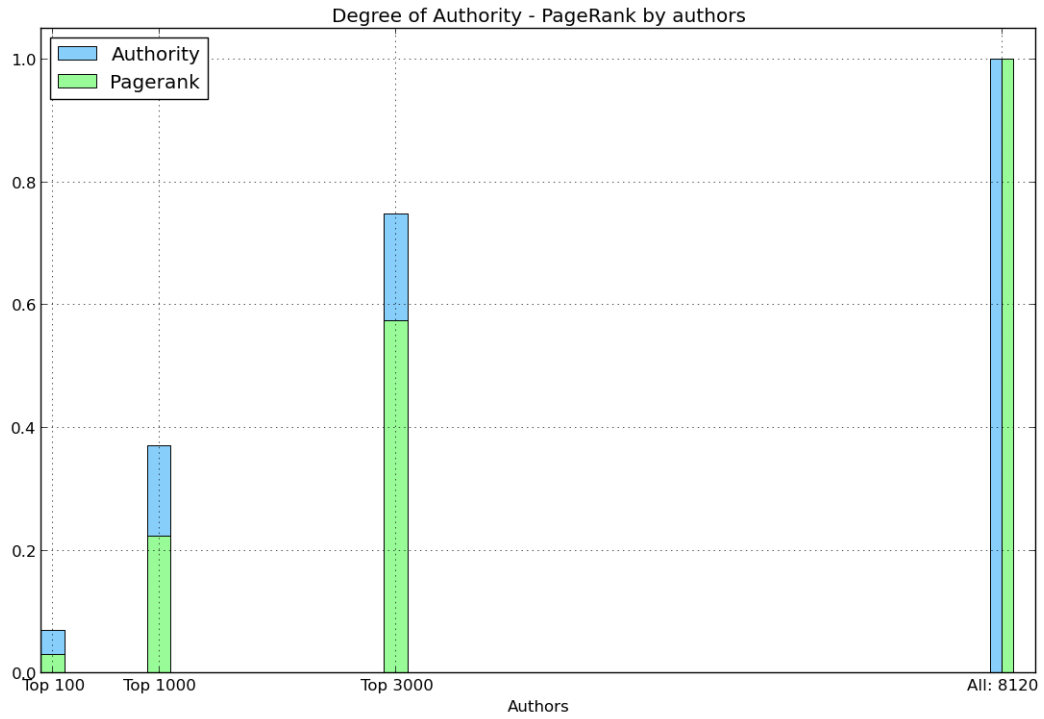


Figure 6.27: HITS and PageRank indexes across number of authors

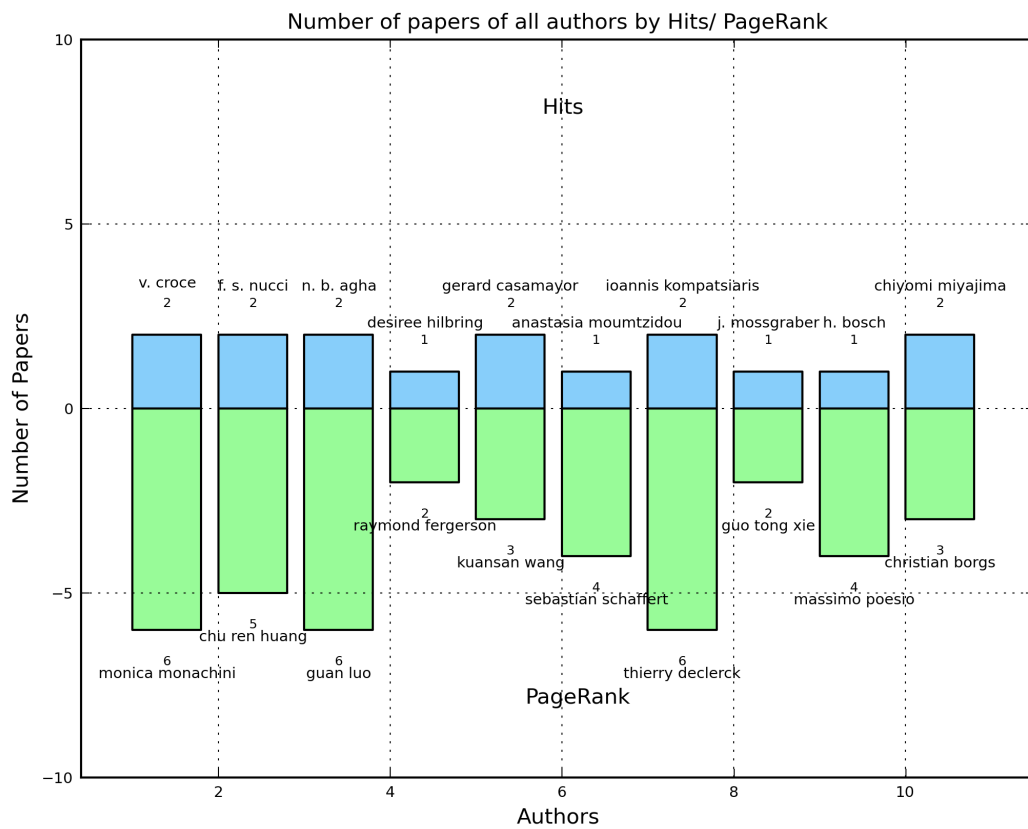


Figure 6.28: Number of Papers of top 10 authors found by HITS and PageRank for 50 Topics

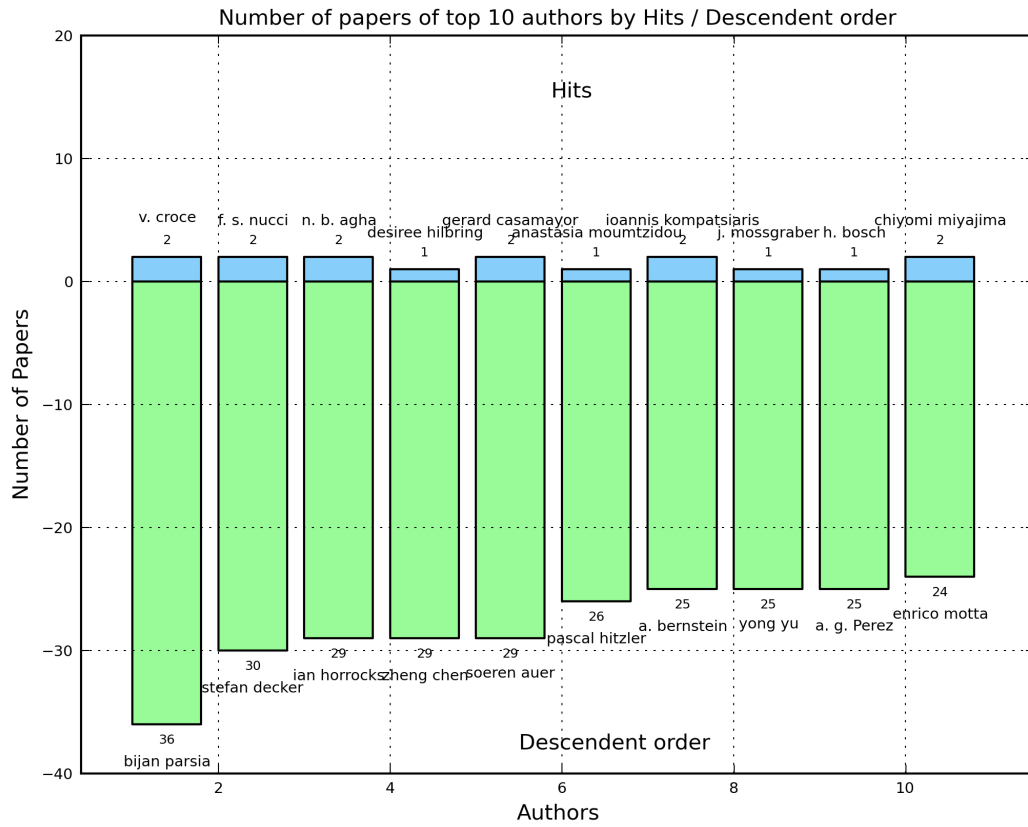


Figure 6.29: Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 50 Topics

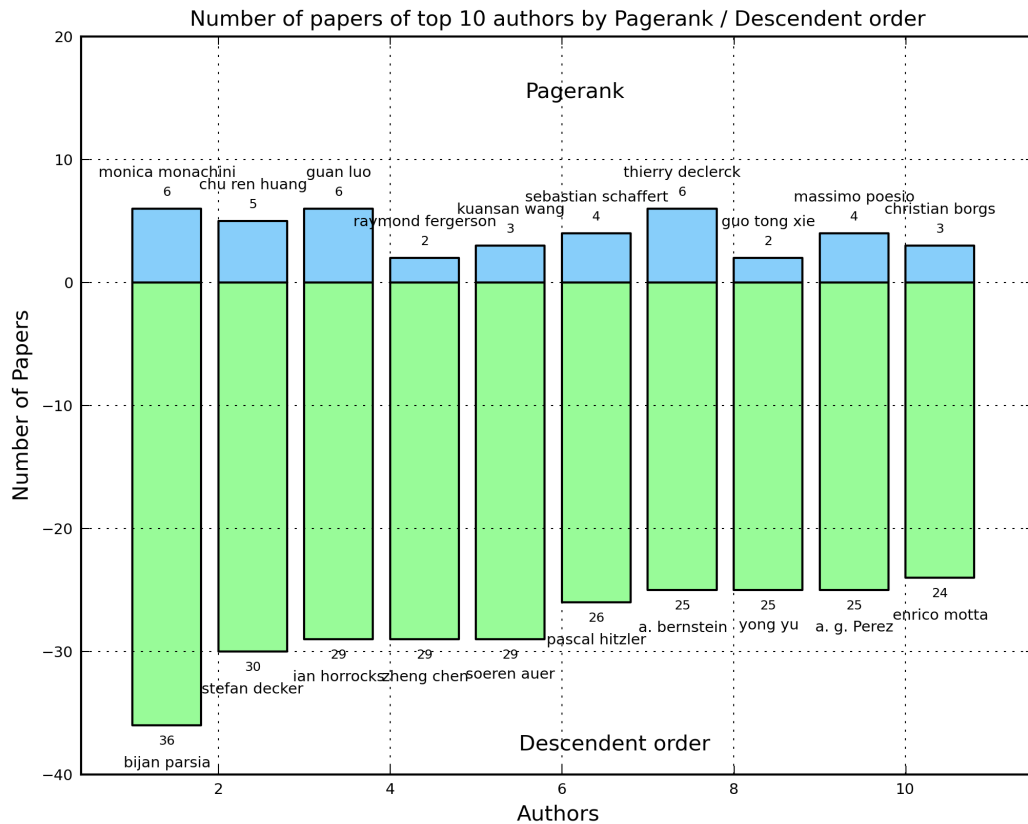


Figure 6.30: Number of Papers of top 10 authors found by HITS compared to first top 10 authors with more papers for 50 Topics

J Topic Analysis for 25 topics in Ubiquitous and Pervasive Computing Conferences

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 1	home	user	particip	us	design	comput	inform	technolog	time	locat
Topic 2	user	place	devic	time	us	game	design	locat	provid	applic
Topic 3	data	us	comput	devic	tag	user	inform	space	particip	applic
Topic 4	user	interact	us	object	time	base	comput	devic	data	figur
Topic 5	data	user	particip	home	inform	time	technolog	displai	design	us
Topic 6	locat	devic	user	us	inform	comput	sensor	kei	mobil	applic
Topic 7	us	home	activ	app	predict	data	user	locat	sensor	share
Topic 8	user	context	comput	us	applic	sensor	data	devic	inform	base
Topic 9	locat	data	time	particip	mobil	us	place	devic	user	sensor
Topic 10	user	particip	locat	applic	context	us	time	interact	comput	data
Topic 11	data	displai	user	devic	inform	phone	applic	us	comput	interact
Topic 12	locat	us	time	base	user	commun	inform	mobil	model	social
Topic 13	sensor	activ	data	user	devic	time	locat	us	comput	studi
Topic 14	user	activ	comput	locat	applic	sensor	context	time	us	data
Topic 15	user	mobil	us	locat	inform	devic	time	activ	phone	model
Topic 16	trajectori	taxi	detect	energi	gp	anomal	trace	cell	household	drive
Topic 17	pose	walk	speed	method	devic	estim	train	featur	data	vector
Topic 18	sensor	bodi	time	measur	rotat	swim	power	figur	robot	swimmer
Topic 19	user	data	time	activ	locat	us	power	comput	devic	model
Topic 20	locat	user	data	inform	comput	privaci	us	base	work	particip
Topic 21	user	data	time	posit	us	base	locat	inform	estim	result
Topic 22	devic	match	context	schema	user	wkh	group	method	place	comput
Topic 23	activ	object	data	us	model	time	comput	particip	work	locat
Topic 24	user	data	inform	time	activ	base	sensor	us	model	comput
Topic 25	user	servic	data	time	activ	perform	figur	applic	base	differ

Table 6.7: Display of 10 words for each 25 Topics

Topic	Prob. 1	Prob. 2	Prob. 3	Prob. 4	Prob. 5	Prob. 6	Prob. 7	Prob. 8	Prob. 9	Prob. 10
Topic 1	0.009	0.007	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005
Topic 2	0.012	0.008	0.007	0.006	0.006	0.006	0.006	0.005	0.005	0.005
Topic 3	0.008	0.007	0.007	0.006	0.006	0.006	0.006	0.005	0.005	0.005
Topic 4	0.011	0.007	0.006	0.006	0.005	0.005	0.004	0.004	0.004	0.004
Topic 5	0.007	0.007	0.007	0.006	0.006	0.006	0.006	0.005	0.005	0.005
Topic 6	0.014	0.013	0.011	0.006	0.006	0.006	0.005	0.005	0.005	0.005
Topic 7	0.009	0.008	0.007	0.007	0.005	0.005	0.005	0.005	0.005	0.005
Topic 8	0.011	0.009	0.007	0.007	0.006	0.005	0.005	0.005	0.005	0.004
Topic 9	0.011	0.009	0.007	0.007	0.006	0.006	0.006	0.005	0.005	0.004
Topic 10	0.015	0.009	0.008	0.007	0.007	0.007	0.006	0.005	0.005	0.005
Topic 11	0.008	0.008	0.008	0.007	0.006	0.006	0.006	0.005	0.005	0.005
Topic 12	0.013	0.008	0.007	0.006	0.006	0.006	0.005	0.005	0.005	0.004
Topic 13	0.013	0.012	0.01	0.008	0.006	0.005	0.005	0.005	0.004	0.004
Topic 14	0.011	0.008	0.007	0.006	0.005	0.005	0.005	0.005	0.005	0.005
Topic 15	0.009	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.004
Topic 16	0.012	0.01	0.008	0.006	0.006	0.005	0.004	0.004	0.004	0.003
Topic 17	0.007	0.006	0.006	0.004	0.004	0.004	0.003	0.003	0.003	0.003
Topic 18	0.013	0.009	0.006	0.005	0.005	0.005	0.005	0.005	0.004	0.004
Topic 19	0.008	0.008	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005
Topic 20	0.017	0.014	0.007	0.006	0.006	0.005	0.005	0.005	0.005	0.005
Topic 21	0.01	0.01	0.008	0.006	0.005	0.005	0.005	0.004	0.004	0.004
Topic 22	0.008	0.007	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.004
Topic 23	0.01	0.009	0.007	0.006	0.006	0.006	0.005	0.005	0.004	0.004
Topic 24	0.008	0.007	0.007	0.006	0.005	0.005	0.005	0.004	0.004	0.004
Topic 25	0.008	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.004	0.004

Table 6.8: Probabilities of words in 25 Topics

K Topic Analysis for 50 topics in Ubiquitous and Pervasive Computing Conferences

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 1	user	comput	context	inform	base	time	us	locat	work	devic
Topic 2	locat	data	mobil	us	sensor	user	game	player	posit	activ
Topic 3	locat	devic	user	data	comput	model	us	time	mobil	interact
Topic 4	data	learn	featur	model	hous	sensor	server	activ	us	replica
Topic 5	time	user	model	data	app	differ	rate	predict	home	base
Topic 6	user	data	us	inform	particip	comput	technolog	time	studi	privaci
Topic 7	data	activ	applic	user	base	time	sensor	us	work	mobil
Topic 8	home	sensor	activ	us	work	user	time	event	data	studi
Topic 9	sensor	devic	data	activ	comput	us	user	model	time	locat
Topic 10	user	place	servic	sensor	devic	data	provid	activ	studi	us
Topic 11	displai	comput	time	inform	user	behavior	point	interact	experi	design
Topic 12	locat	user	time	base	us	model	devic	applic	comput	set
Topic 13	data	sensor	time	place	user	us	locat	algorithm	base	work
Topic 14	data	user	particip	network	model	design	displai	applic	us	comput
Topic 15	devic	tag	messag	number	user	comput	host	time	applic	inform
Topic 16	segwai	tactil	navig	drive	user	instruct	ar	turn	interfac	rout
Topic 17	activ	locat	us	time	data	user	sensor	base	work	particip
Topic 18	memori	record	answer	subject	audio	question	retriev	search	problem	data
Topic 19	applic	data	sensor	user	room	comput	devic	us	inform	design
Topic 20	displai	inform	user	activ	design	peopl	particip	differ	data	time
Topic 21	tag	interact	sensor	user	time	comput	us	detect	bodi	base
Topic 22	user	data	time	comput	phone	locat	displai	differ	sensor	servic
Topic 23	user	us	particip	phone	data	comput	applic	studi	base	time
Topic 24	time	mobil	data	user	home	us	behavior	particip	activ	studi
Topic 25	user	comput	locat	inform	interact	devic	applic	mobil	base	time
Topic 26	node	mobil	inform	movement	ey	figur	locat	comput	estim	landmark
Topic 27	activ	air	qualiti	school	goal	data	household	tm	particip	indoor
Topic 28	user	inform	data	imag	us	travel	model	context	mobil	encount
Topic 29	wkh	school	teacher	student	dqg	child	activ	wr	children	studi
Topic 30	user	locat	applic	us	data	particip	time	comput	technolog	mobil
Topic 31	game	user	model	player	us	time	sensor	experi	pervas	opinion
Topic 32	user	data	locat	time	comput	context	base	commun	inform	us
Topic 33	match	schema	context	model	user	predict	time	attribut	particip	figur
Topic 34	user	activ	data	time	us	base	devic	applic	comput	object
Topic 35	inform	user	children	devic	locat	mobil	point	interact	social	base
Topic 36	particip	inform	us	activ	studi	user	work	data	comput	design
Topic 37	activ	inform	time	comput	zone	chang	estim	locat	work	figur
Topic 38	data	particip	locat	devic	share	work	studi	inform	home	us
Topic 39	user	activ	data	sensor	featur	locat	time	us	model	devic
Topic 40	locat	user	context	inform	devic	comput	design	applic	provid	us
Topic 41	locat	place	share	user	particip	time	data	social	visitor	us
Topic 42	user	devic	locat	time	data	inform	us	mobil	design	place
Topic 43	frequenc	home	signal	laptop	us	power	tag	devic	gener	accuraci
Topic 44	data	user	devic	locat	base	measur	us	time	signal	comput
Topic 45	activ	sensor	data	recognit	class	failur	train	learn	model	accuraci
Topic 46	sensor	home	data	inform	design	locat	particip	node	technolog	comput
Topic 47	resourc	resolv	queri	descript	strand	map	network	fact	twine	gps
Topic 48	user	?	applic	rate	devic	us	mobil	figur	bluetooth	comput
Topic 49	devic	user	us	applic	studi	particip	mobil	phone	data	design
Topic 50	user	block	context	event	base	sensor	chang	inform	data	us

Table 6.9: Display of 10 words for each 50 Topics

Topic	Prob. 1	Prob. 2	Prob. 3	Prob. 4	Prob. 5	Prob. 6	Prob. 7	Prob. 8	Prob. 9	Prob. 10
Topic 1	0.009	0.008	0.007	0.007	0.006	0.006	0.005	0.005	0.005	0.004
Topic 2	0.01	0.009	0.007	0.006	0.006	0.006	0.006	0.005	0.005	0.004
Topic 3	0.009	0.009	0.008	0.008	0.007	0.006	0.006	0.006	0.005	0.004
Topic 4	0.011	0.011	0.011	0.01	0.008	0.007	0.006	0.006	0.005	0.005
Topic 5	0.007	0.007	0.007	0.007	0.006	0.005	0.005	0.005	0.004	0.004
Topic 6	0.01	0.008	0.008	0.007	0.007	0.007	0.006	0.005	0.005	0.005
Topic 7	0.01	0.009	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005
Topic 8	0.014	0.012	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.004
Topic 9	0.023	0.01	0.009	0.006	0.006	0.006	0.006	0.006	0.005	0.005
Topic 10	0.013	0.009	0.007	0.007	0.006	0.005	0.005	0.005	0.005	0.004
Topic 11	0.007	0.006	0.006	0.005	0.005	0.004	0.004	0.004	0.004	0.004
Topic 12	0.016	0.008	0.007	0.006	0.006	0.006	0.006	0.005	0.005	0.005
Topic 13	0.014	0.011	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.005
Topic 14	0.009	0.007	0.006	0.005	0.005	0.005	0.004	0.004	0.004	0.004
Topic 15	0.008	0.008	0.007	0.006	0.006	0.005	0.005	0.005	0.005	0.005
Topic 16	0.012	0.009	0.008	0.006	0.005	0.005	0.005	0.005	0.005	0.005
Topic 17	0.008	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.005	0.004
Topic 18	0.01	0.008	0.006	0.006	0.005	0.004	0.004	0.003	0.003	0.002
Topic 19	0.012	0.009	0.008	0.008	0.006	0.006	0.005	0.005	0.005	0.004
Topic 20	0.012	0.007	0.007	0.006	0.006	0.005	0.005	0.005	0.005	0.004
Topic 21	0.013	0.008	0.006	0.005	0.005	0.005	0.005	0.005	0.005	0.004
Topic 22	0.013	0.007	0.007	0.006	0.006	0.006	0.006	0.005	0.005	0.005
Topic 23	0.01	0.008	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.005
Topic 24	0.012	0.008	0.007	0.007	0.006	0.006	0.005	0.005	0.005	0.004
Topic 25	0.012	0.008	0.008	0.007	0.006	0.006	0.006	0.005	0.005	0.004
Topic 26	0.015	0.014	0.01	0.01	0.008	0.006	0.006	0.006	0.006	0.006
Topic 27	0.012	0.01	0.009	0.009	0.008	0.006	0.006	0.006	0.005	0.005
Topic 28	0.01	0.008	0.007	0.006	0.005	0.005	0.005	0.005	0.005	0.005
Topic 29	0.01	0.01	0.008	0.007	0.007	0.006	0.006	0.006	0.005	0.005
Topic 30	0.01	0.01	0.006	0.006	0.006	0.006	0.006	0.005	0.004	0.004
Topic 31	0.021	0.01	0.008	0.007	0.006	0.005	0.005	0.004	0.004	0.004
Topic 32	0.012	0.009	0.007	0.006	0.006	0.005	0.005	0.005	0.005	0.004
Topic 33	0.011	0.011	0.007	0.007	0.006	0.006	0.005	0.005	0.005	0.004
Topic 34	0.011	0.01	0.009	0.007	0.006	0.005	0.005	0.005	0.005	0.005
Topic 35	0.012	0.01	0.008	0.007	0.006	0.006	0.006	0.005	0.005	0.005
Topic 36	0.009	0.007	0.007	0.006	0.006	0.006	0.005	0.005	0.005	0.005
Topic 37	0.008	0.008	0.007	0.005	0.005	0.005	0.005	0.004	0.004	0.004
Topic 38	0.012	0.011	0.01	0.008	0.007	0.006	0.006	0.005	0.005	0.005
Topic 39	0.01	0.009	0.008	0.006	0.006	0.006	0.005	0.005	0.005	0.005
Topic 40	0.009	0.008	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.005
Topic 41	0.021	0.014	0.01	0.008	0.007	0.007	0.005	0.005	0.005	0.005
Topic 42	0.014	0.007	0.007	0.007	0.006	0.006	0.006	0.005	0.005	0.005
Topic 43	0.012	0.01	0.008	0.008	0.006	0.005	0.005	0.005	0.005	0.004
Topic 44	0.015	0.008	0.006	0.005	0.005	0.005	0.005	0.005	0.005	0.004
Topic 45	0.017	0.014	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.006
Topic 46	0.008	0.008	0.007	0.006	0.005	0.005	0.005	0.004	0.004	0.004
Topic 47	0.023	0.019	0.015	0.013	0.01	0.008	0.008	0.008	0.007	0.006
Topic 48	0.014	0.008	0.007	0.006	0.005	0.005	0.004	0.004	0.004	0.004
Topic 49	0.008	0.007	0.006	0.005	0.005	0.005	0.005	0.004	0.004	0.004
Topic 50	0.014	0.013	0.008	0.008	0.006	0.005	0.005	0.005	0.005	0.004

Table 6.10: Probabilities of words in 50 Topics

L Topic Analysis for 25 topics in Semantic Web Conference Series

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 1	entiti	annot	page	document	web	mention	search	label	ndcg	form
Topic 2	lda	sid	text	model	fe	jacquard	user	mapreduc	comment	post
Topic 3	servic	web	ontolog	semant	cloud	tag	base	process	rule	user
Topic 4	cheater	review	qualiti	opinion	segment	summari	product	crowdsourc	data	evalu
Topic 5	queri	search	languag	dwell	word	trail	url	data	result	null
Topic 6	word	data	languag	set	text	differ	new	result	inform	queri
Topic 7	rule	queri	rdf	semant	reason	set	relat	properti	map	data
Topic 8	data	queri	valu	set	stream	base	user	result	tag	type
Topic 9	network	social	node	user	data	number	distribut	account	web	time
Topic 10	model	twitter	number	user	algorithm	similar	set	page	featur	result
Topic 11	model	equilibrium	price	player	imag	advertis	network	agent	mechan	revenu
Topic 12	queri	algorithm	set	search	optim	result	rank	problem	document	comput
Topic 13	data	servic	web	inform	user	semant	cloud	model	locat	applic
Topic 14	user	web	page	applic	time	model	content	data	inform	servic
Topic 15	spammer	spam	farm	follow	user	link	ontolog	farmer	social	axiom
Topic 16	model	topic	data	document	entiti	featur	queri	learn	base	inform
Topic 17	user	event	social	inform	content	locat	network	campaign	time	recommend
Topic 18	crimin	site	spam	url	web	account	data	model	link	follow
Topic 19	web	agent	servic	data	set	httpi	base	model	result	inform
Topic 20	graph	node	agent	let	set	vm	rule	regim	algorithm	comput
Topic 21	question	answer	type	label	queri	set	attribut	languag	model	form
Topic 22	semant	rdf	domain	web	ontolog	opal	model	data	knowledg	relat
Topic 23	ontolog	entiti	properti	relationship	type	semant	annot	rdf	relat	class
Topic 24	servic	path	evenc	event	data	graph	web	process	semant	match
Topic 25	ontolog	concept	semant	set	base	similar	game	map	result	align

Table 6.11: Display of 10 words for each 25 Topics

Topic	Prob. 1	Prob. 2	Prob. 3	Prob. 4	Prob. 5	Prob. 6	Prob. 7	Prob. 8	Prob. 9	Prob. 10
Topic 1	0.02	0.013	0.009	0.007	0.007	0.007	0.006	0.006	0.006	0.005
Topic 2	0.017	0.01	0.009	0.009	0.008	0.006	0.006	0.005	0.005	0.005
Topic 3	0.01	0.007	0.007	0.007	0.007	0.007	0.006	0.006	0.005	0.005
Topic 4	0.026	0.021	0.011	0.011	0.01	0.008	0.007	0.007	0.007	0.007
Topic 5	0.048	0.016	0.008	0.007	0.007	0.007	0.006	0.006	0.005	0.005
Topic 6	0.009	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.004	0.004
Topic 7	0.016	0.014	0.013	0.009	0.008	0.008	0.008	0.007	0.007	0.007
Topic 8	0.019	0.011	0.007	0.006	0.006	0.006	0.006	0.006	0.006	0.005
Topic 9	0.022	0.015	0.01	0.008	0.008	0.006	0.006	0.005	0.005	0.005
Topic 10	0.012	0.01	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005
Topic 11	0.017	0.014	0.013	0.011	0.01	0.01	0.008	0.008	0.007	0.007
Topic 12	0.021	0.015	0.013	0.007	0.006	0.006	0.006	0.006	0.006	0.006
Topic 13	0.02	0.011	0.009	0.006	0.006	0.006	0.005	0.005	0.005	0.005
Topic 14	0.023	0.012	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.005
Topic 15	0.039	0.013	0.011	0.01	0.009	0.009	0.008	0.007	0.006	0.006
Topic 16	0.015	0.011	0.008	0.007	0.007	0.006	0.006	0.006	0.006	0.006
Topic 17	0.021	0.012	0.01	0.009	0.008	0.007	0.006	0.006	0.006	0.005
Topic 18	0.043	0.015	0.012	0.01	0.008	0.007	0.007	0.006	0.006	0.005
Topic 19	0.009	0.008	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.004
Topic 20	0.015	0.013	0.009	0.008	0.007	0.006	0.006	0.006	0.006	0.005
Topic 21	0.033	0.025	0.01	0.009	0.008	0.007	0.007	0.006	0.005	0.005
Topic 22	0.016	0.013	0.012	0.012	0.01	0.009	0.007	0.007	0.006	0.005
Topic 23	0.024	0.019	0.009	0.009	0.008	0.008	0.007	0.007	0.007	0.006
Topic 24	0.03	0.017	0.016	0.015	0.014	0.009	0.008	0.008	0.007	0.006
Topic 25	0.027	0.008	0.008	0.007	0.006	0.006	0.006	0.005	0.005	0.005

Table 6.12: Probabilities of words in 25 Topics

M Topic Analysis for 50 topics in Semantic Web Conference Series

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 1	network	user	servic	web	social	friend	composit	model	prefer	task
Topic 2	cluster	data	dataset	set	measur	method	spectral	evalu	stream	comput
Topic 3	inform	rdf	sparql	dataset	data	web	semant	link	user	tripl
Topic 4	rule	reason	owl	data	logic	fact	gpu	base	semant	set
Topic 5	question	answer	path	queri	mr	star	nca	qa	evalu	parliament
Topic 6	servic	qo	web	workflow	provid	semant	comput	model	requir	approach
Topic 7	graph	equilibrium	price	rdf	queri	relat	set	algorithm	match	properti
Topic 8	jacquard	email	mtl	sbm	process	action	activ	character	workflow	semant
Topic 9	repair	set	atom	reason	logic	ontolog	preserv	axiom	dl	follow
Topic 10	user	tag	document	similar	selector	web	steiner	page	term	valu
Topic 11	path	queri	properti	element	express	complex	sparql	defin	semant	languag
Topic 12	model	user	rank	algorithm	document	click	featur	set	relev	inform
Topic 13	spammer	spam	follow	link	farm	ount	web	account	model	semant
Topic 14	featur	learn	model	train	lyon	topic	set	classifi	summari	text
Topic 15	div	seller	scroll	tree	trace	tempor	page	llt	match	monitor
Topic 16	ontolog	queri	rdf	semant	data	set	relat	base	context	web
Topic 17	relat	inform	semant	tag	capitalist	web	social	cuv	model	knowledg
Topic 18	phrase	word	languag	model	lexicon	translat	depend	bilingu	wo	concept
Topic 19	entiti	semant	mention	languag	ontolog	domain	knowledg	base	set	type
Topic 20	link	osn	hierarchi	data	network	inform	uh	set	hierarch	differ
Topic 21	wiki	articl	classif	wr	categori	dst	rdi	nmp	hierarchi	jrfl
Topic 22	rout	network	valu	ema	lq	vd	peer	queri	number	sk
Topic 23	ontolog	rdf	web	semant	data	map	base	rule	set	servic
Topic 24	event	causal	tempor	data	metadata	media	relat	extract	burst	entiti
Topic 25	farmer	farm	butterfli	crowdturfing	region	collusionrank	link	follow	model	social
Topic 26	set	ri	let	defin	properti	rule	node	semant	satisfi	model
Topic 27	code	applic	css	inventori	depend	javascript	event	ui	function	client
Topic 28	node	algorithm	network	edg	data	set	cost	number	site	graph
Topic 29	queri	data	search	model	user	result	time	inform	document	set
Topic 30	opal	model	field	data	domain	requir	web	segment	type	annot
Topic 31	question	answer	game	cheater	user	worker	task	model	network	reward
Topic 32	httpi	http	content	web	us	inform	languag	word	user	anarchi
Topic 33	xist	revis	opn	knowledg	conver	rule	msc	owl	supervi	data
Topic 34	word	text	languag	annot	data	entiti	document	model	us	corpu
Topic 35	agent	contribut	set	qualiti	network	user	probabl	model	trust	pb
Topic 36	entiti	wikipedia	name	categori	disambigu	taxonomi	mention	link	mentionrank	capit
Topic 37	predict	model	yt	xt	sdh	smooth	user	learn	argument	agent
Topic 38	ontolog	chang	knowledg	new	semant	base	user	web	set	inform
Topic 39	att	speaker	event	semant	web	speech	earthquak	coment	annot	provid
Topic 40	social	page	featur	network	web	set	inform	form	extract	domain
Topic 41	action	polic	model	class	context	safe	owl	properti	ontolog	sfw
Topic 42	user	time	number	set	model	evalu	result	gener	queri	click
Topic 43	web	http	applic	server	messag	data	content	secur	servic	page
Topic 44	ontolog	data	user	semant	web	base	map	properti	queri	result
Topic 45	action	spell	word	model	correct	cloudspel	el	state	misspel	es
Topic 46	topic	crimin	model	url	photo	locat	distribut	player	lda	april
Topic 47	imag	user	twitter	web	set	tweet	base	model	page	entiti
Topic 48	geo	vac	nsga	yit	form	din	model	field	set	cq
Topic 49	ontolog	chain	rdf	class	owl	reason	map	queri	air	resourc
Topic 50	queri	pattern	tripl	bit	document	data	hash	result	index	set

Table 6.13: Display of 10 words for each 50 Topics

Topic	Prob. 1	Prob. 2	Prob. 3	Prob. 4	Prob. 5	Prob. 6	Prob. 7	Prob. 8	Prob. 9	Prob. 10
Topic 1	0.026	0.022	0.018	0.016	0.013	0.01	0.008	0.006	0.006	0.005
Topic 2	0.037	0.035	0.013	0.01	0.01	0.009	0.006	0.006	0.006	0.006
Topic 3	0.013	0.012	0.011	0.011	0.011	0.01	0.009	0.007	0.007	0.006
Topic 4	0.032	0.016	0.011	0.008	0.007	0.007	0.007	0.006	0.006	0.006
Topic 5	0.03	0.021	0.017	0.012	0.009	0.009	0.008	0.007	0.006	0.006
Topic 6	0.083	0.023	0.018	0.013	0.01	0.008	0.007	0.007	0.007	0.007
Topic 7	0.02	0.011	0.011	0.01	0.008	0.008	0.008	0.007	0.007	0.006
Topic 8	0.02	0.01	0.009	0.004	0.004	0.004	0.004	0.004	0.004	0.004
Topic 9	0.013	0.011	0.009	0.009	0.007	0.007	0.006	0.006	0.006	0.006
Topic 10	0.014	0.014	0.011	0.008	0.008	0.008	0.007	0.006	0.006	0.006
Topic 11	0.088	0.015	0.012	0.01	0.009	0.008	0.008	0.008	0.008	0.007
Topic 12	0.018	0.013	0.012	0.009	0.009	0.008	0.007	0.007	0.006	0.006
Topic 13	0.043	0.027	0.015	0.014	0.011	0.008	0.008	0.008	0.007	0.005
Topic 14	0.017	0.014	0.012	0.011	0.01	0.01	0.008	0.007	0.007	0.007
Topic 15	0.019	0.015	0.013	0.012	0.012	0.011	0.01	0.009	0.008	0.007
Topic 16	0.019	0.014	0.01	0.008	0.007	0.007	0.006	0.006	0.006	0.006
Topic 17	0.012	0.011	0.01	0.01	0.009	0.008	0.008	0.007	0.007	0.007
Topic 18	0.029	0.022	0.02	0.016	0.015	0.012	0.012	0.01	0.007	0.007
Topic 19	0.015	0.008	0.007	0.007	0.006	0.006	0.006	0.006	0.005	0.005
Topic 20	0.013	0.012	0.01	0.008	0.007	0.007	0.007	0.006	0.006	0.005
Topic 21	0.018	0.01	0.01	0.01	0.01	0.009	0.007	0.007	0.006	0.006
Topic 22	0.012	0.011	0.011	0.011	0.009	0.008	0.006	0.006	0.006	0.006
Topic 23	0.011	0.01	0.01	0.007	0.006	0.006	0.005	0.005	0.005	0.004
Topic 24	0.079	0.017	0.014	0.011	0.01	0.009	0.009	0.009	0.007	0.007
Topic 25	0.031	0.011	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006
Topic 26	0.015	0.011	0.01	0.009	0.009	0.008	0.008	0.007	0.007	0.007
Topic 27	0.033	0.021	0.017	0.014	0.013	0.012	0.01	0.01	0.01	0.009
Topic 28	0.023	0.016	0.01	0.01	0.008	0.008	0.008	0.007	0.007	0.007
Topic 29	0.055	0.017	0.017	0.01	0.01	0.009	0.006	0.006	0.005	0.005
Topic 30	0.013	0.01	0.01	0.009	0.008	0.006	0.006	0.005	0.005	0.005
Topic 31	0.038	0.031	0.017	0.017	0.015	0.014	0.013	0.01	0.01	0.008
Topic 32	0.05	0.009	0.009	0.007	0.006	0.006	0.005	0.005	0.005	0.005
Topic 33	0.019	0.017	0.009	0.008	0.008	0.007	0.006	0.005	0.005	0.005
Topic 34	0.014	0.007	0.006	0.006	0.006	0.006	0.006	0.006	0.005	0.005
Topic 35	0.027	0.011	0.011	0.01	0.007	0.007	0.007	0.006	0.006	0.005
Topic 36	0.101	0.05	0.029	0.011	0.011	0.008	0.008	0.007	0.007	0.006
Topic 37	0.032	0.009	0.009	0.009	0.008	0.008	0.008	0.007	0.007	0.007
Topic 38	0.012	0.008	0.007	0.007	0.006	0.006	0.006	0.006	0.005	0.005
Topic 39	0.011	0.008	0.007	0.007	0.006	0.006	0.006	0.006	0.005	0.005
Topic 40	0.017	0.01	0.008	0.008	0.008	0.007	0.007	0.006	0.006	0.005
Topic 41	0.033	0.018	0.014	0.009	0.008	0.007	0.007	0.007	0.007	0.007
Topic 42	0.017	0.012	0.008	0.006	0.005	0.005	0.005	0.005	0.005	0.004
Topic 43	0.013	0.009	0.009	0.008	0.008	0.008	0.008	0.007	0.007	0.006
Topic 44	0.022	0.012	0.011	0.009	0.009	0.007	0.007	0.006	0.006	0.006
Topic 45	0.023	0.015	0.015	0.014	0.01	0.007	0.006	0.006	0.006	0.005
Topic 46	0.037	0.02	0.02	0.013	0.013	0.012	0.009	0.008	0.008	0.008
Topic 47	0.01	0.01	0.01	0.009	0.006	0.006	0.005	0.005	0.005	0.005
Topic 48	0.045	0.023	0.008	0.008	0.007	0.007	0.006	0.006	0.005	0.005
Topic 49	0.023	0.023	0.018	0.012	0.01	0.009	0.007	0.006	0.006	0.005
Topic 50	0.032	0.011	0.009	0.009	0.009	0.008	0.008	0.007	0.007	0.007

Table 6.14: Probabilities of words in 50 Topics