UNIVERSITY OF SOUTHAMPTON

Enhancing the Automation of Forming Groups for Education with Semantics

by

Asma Ounnas

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Engineering and Applied Science
Department of Electronics and Computer Science

November 2010

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND APPLIED SCIENCE
DEPARTMENT OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Asma Ounnas

Many approaches to learning and teaching rely upon students working in groups. Formation of optimal groups can be a time consuming and complex task, particularly when the list of participants is unknown in advance. This research investigates the implementation of semantics to enhance computer-supported group formation in education using two approaches: The first approach uses semantics to express the criteria specified by the person forming the groups. The group formation in this approach is modelled as a constraint satisfaction problem where the criteria is a set of constraints that we aim to minimise their violation while processing the groups. The second approach uses Semantic Web domain ontologies in describing the participants to enrich the data used in calculating the similarity between the participants when the group formation is processed using a heuristic approach such as clustering algorithms.

We run a number of experiments that include real datasets from higher education classes, simulated datasets, Web-based datasets, and user studies, to evaluate the research. The results proved that in both approaches, implementing semantics improved the generated groups, in that, using semantics to model group formation's constraints generates an optimised grouping in terms of constraint satisfaction that exceeds the performance of existing applications, particularly in terms of the number of constraints it can handle; and that using semantics to model the participants' data enhances their satisfaction with the groups they are allocated to.

Contents

List of Figures

List of Tables

List of Algorithms

Declaration of Authorship

I, Asma Ounnas declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

    - Ounnas, A., Davis, H. and Millard, D. (2009) A Framework for Semantic Group Formation in Education. Journal of International Forum of Educational Technology and Society, 12. pp. 43-55. ISSN 1436-4522.
    - Ounnas, A., Davis, H. and Millard, D. (2008) A Framework for Semantic Group Formation. In: The 8th IEEE International Conference on Advanced Learning Technologies (ICALT 2008), July 1st- July 5th, 2008, Santander, Cantabria, Spain. pp. 34-38.

- Ounnas, A., Davis, H. and Millard, D. (2008) Semantic Web-based Group Formation for E-learning. In: PhD Symposium in the 5th European Semantic Web Conference 2008.
- Ounnas, A., Millard, D. and Davis, H. (2007) A Metrics Framework for Evaluating Group Formation. In: ACM Group'07, 4-7th November 2007, Sanibel Island, Florida, USA. pp. 221-224 .
- Liccardi, I., Ounnas, A., Pau, R., Massey, E., Kinnumen, P., Lewthwaite, S., Midy, M. A. and Sakar, C. (2007) The role of social networks in students learning experiences. ACM SIGCSE Bulletin (December Issue) . pp. 224-237.
- Ounnas, A., Davis, H. C. and Millard, D. E. (2007) Semantic (Group Formation). In: Knowledge Web PhD Symposium 2007 (KWEPSY2007), in the 4th European Semantic Web Conference (ESWC 2007), 3-7th, June 2007, Innsbruck, Austria.
- Ounnas, A., Davis, H. C. and Millard, D. E. (2007) Semantic Modeling for Group Formation. In: Workshop on Personalisation in E-Learning Environments at Individual and Group Level (PING) at the 11th International Conference on User Modeling UM2007, Corfu, Greece.
- Ounnas, A., Davis, H. C. and Millard, D. E. (2007) Towards Semantic Group Formation. In: The 7th IEEE International Conference on Advanced Learning Technologies (ICALT 2007), July 18-20, 2007, Niigata, Japan. pp. 825-827.
- Ounnas, A., Liccardi, I., Davis, H. C., Millard, D. E. and White, S. A. (2006) Towards a Semantic Modeling of Learners for Social Networks. In: International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL) at the AH2006 Conference, 20-23 June, Dublin, Ireland. pp. 102-108.

Acknowledgements

I would like to thank all those who help me throughout my research. I am deeply indebted to my supervisors Prof. Hugh Davis and Dr. David Millard from the School of Electronics and computer Science, University of Southampton, whose help, stimulating suggestions, and encouragement helped me in all the time of research and writing of this thesis.

Also, a big thank you to Prof. James Hendler, and Prof. Hal Abelson for their helpful comments on the contribution of this research to the field during my participation in the Web Science Research Initiative exchange program; and to Prof. Dame Wendy Hall who facilitated my participation in the program.

Thanks also to Dr Mark Weal, Dr Mike Poppleton, Dr Bob Walter, Dr Gary Wills, and Dr Su White, for allowing me to use data from their classes and events to run my experiments; and to my colleague Areeb Al Owisheq for proof reading this thesis.

Thanks also to all my friends who helped me through the good and the bad times over the years I dedicated to this research, particularly Tarek Nechma, Kheiredine Derouiche, and Dr Ilaria Liccardi in whom I found a caring friend and a motivating collaborator.

I am also thankful to my family for the emotional support and motivation to complete this work. I am particularly thankful to my beloved sister Dr Houda Ounnas, my aunt Zahia Braouyk, and above all my mother Nassima Braouyk, a loving teacher who is always there for me.

Abbreviations

AI	**A**rtificial **I**ntelligence
ASP	**A**nswer **S**et **P**rogramming
CoP	**C**ommunities **o**f **P**ractice
CSCW	**C**omputer **S**upported **C**ollaborative **W**ork
CSGF	**C**omputer **S**upported **G**roup **F**ormation
CSP	**C**onstraint **S**atisfaction **P**roblem
GF	**G**roup **F**ormation
CSP	**C**onstraint **S**atisfaction **P**roblem
ECS	school of **E**lectronics and **C**omputer **S**cience
FOAF	**F**riend **o**f **a** **F**riend
MRV	**M**inimum **R**emaining **V**alues
NoP	**N**etworks **o**f **P**ractice
OGF	**O**pportunistic **G**roup **F**ormation
OWL	**W**eb **O**ntology **L**anguage
RDF	**R**esource **D**escription **F**ramework
RIF	**R**ule **I**nterchange **F**ormat
SKOS	**S**imple **K**nowledge **O**rganization **S**ystem
SN	**S**ocial **N**etwork
SPARQL	**S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage
SVM	**S**upport **V**ector **M**achine
SW	**S**emantic **W**eb
WTC	**W**illingness **t**o **C**ommunicate
WWW	**W**orld **W**ide **W**eb
W3C	**W**orld **W**ide **W**eb **C**onsortium
ZPD	**Z**one of **P**roximal **D**evelopment

To . . . My Mother

Chapter 1

Introduction

1.1 Motivation and challenges

For decades, group formation has been a subject of study in many domains including psychology, sociology, philosophy, and education (Owens et al., 1998), particularly after the popularity of employing collaborative learning methodologies in the classroom increased. In education, teachers form or influence the formation of groups of students for different types of collaborative activities. For the grouping to be efficient, teachers need to take into account any criteria (constraints) that can influence the performance of the group as a whole and that of the individuals within the group, including the data to be collected from the user. For example, a teacher aiming to group students for software engineering projects may be concerned with the even distribution of the students across the groups in terms of their previous experience in the subject of study to ensure all groups are balanced and hence will have an equal opportunity in performing well in the task.

In addition to that, the teacher has to ensure that all the students are allocated to the groups that would relatively maximise their benefits from participating in the collaborative work. For example if a female student qualifies best to be a leader in group work, then she should not be allocated in a group with other leaders as this might create a

negative conflict. At the same time, she should not be grouped with a group of otherwise all-male participants as she might be left out by the other members within the group. In another scenario, a teacher might motivate a collaborative activity by identifying the students who would achieve a successful collaboration if they worked together, and then recommend potential collaborators to each other, motivating the emergence of communities of learners. The teacher can identify the potential collaborators based on their interests or learning styles depending on the targeted collaborative activity. In fact, recommendation of potential collaborators is not limited to school kids or higher education students. In research, academics look for collaborators all the time, as might be observed in publications. For this reason, forming groups can also be finding researchers with similar interests to recommend them to each other for research projects.

In this context, group formation can be viewed as a personalisation of the individuals allocation to groups, and a personalisation of the groups themselves by creating a culture for the group through its participants. As individuals tend to be different, the problem of satisfying all the learners needs is complex; and negotiating the allocations to reach consensus is a challenging task. As the number of constraints grows larger, or the data about the participants more complex, reaching agreements for the formation becomes more complicated, especially in highly heterogeneous, multidimensional, distributed groups.

So far there have been few efforts in automating the process of group formation in learning under few constraints. However, most of the existing systems are developed for a specific type of group formation that is usually associated with a collaborative e-learning system that groups the learners based on their state (progress) within the designed learning activity. Moreover, the evaluation of the group formation efficiency within most system is not well reported, and the systems are very different in principles and aim to be compared.

With the growing use of e-learning technologies such as forums, online classes, and resources share sites, the data available about the participants can be large and messy. In fact, since the explosion of social networking, profiling users and gathering data about them became a bit too common, to the point that most users volunteer their data wether in an explicit way or through their recorded behaviour in the games applications they use. Although, this is not that common in education yet, most likely due to more

restrictions on the ethics of data gathering, the amount of data available in the Web is usually not small. These lead us to the though that if the gathering of this data was more meaningful than just gathering strings and keywords, more powerful results can be harvested. Unfortunately, there are not many e-learning based group formation mechanisms that make use of the semantics of the data harvested from the Web. In fact, there aren't many applications that use Web based profiling for group formation -in education- in the first place.

1.2 Research overview

It is not a hidden fact that this century is the century of data and data analysis. There is just too much data, and leaders in many fields, from economics to sociology to technology to advertising, are realising this fact. The proof is in their investment in information systems. Thanks to the larger provider of information, the Web that is, there is no limit to what can be done. Surely, we can do a lot with the data available to us from the Web, but we can do even more if it is represented in a meaningful way, that both we, and the machine can understand and use.

Automation wise, semantics, or the study of meaning, is not new to Computer Science, but it is relatively new to the study of the Web, and although might seem to have started slow, when the Semantic Web was first introduced, it is now starting to show an impact in everyday realities. The potential of the Semantic Web has allowed the semantic formation of social networks to be successful. Given that a social network is nothing but a large group of people, we trust that the problem of group formation, in this context Web-based group formation, can as well benefit from employing Semantic Web technologies in describing the participants.

Similarly, wether Web-based or not, the semantics of the group formation's constraints can be represented to model the goals of the collaboration activity in question. This will give more meaning to the way a teacher would want the negotiation of the groups to be performed.

In this thesis, therefore, we hypothesis that semantics can be used in various ways when automating the process of forming groups for education. We also hypothesis that the usage of Semantic Web technologies can improve the quality of the formed groups.

1.3 Research hypothesis

The hypotheses of this thesis can be stated as follows:

- Forming groups for collaboration in education can be modelled in different ways based on the aim of the collaboration activity. Two main elements of forming groups are: the participants' data and the criteria or constraints that need to be addressed to achieve the desired collaboration activity. Modelling the semantics of both these elements can improve the results of the group formation, where the quality of the groups is measured in two different ways: as the satisfaction of the participants with their groups, or as the satisfaction of the constraints that the teacher set up for the collaboration to be successful. The latter is based on the assumption that when a teacher uses a Constraint Satisfaction Problems approach to form groups, the students' satisfaction is not the main needed measure for the quality of the groups, as measuring the constraints' satisfaction becomes an indictor of this approach's success.

- The problem of forming groups for education can be modelled as a constraint satisfaction problem to optimise the allocations of participants to groups. The semantics of the teachers' chosen criteria can be modelled as a set of strong and weak constraints depending on their importance in achieving the collaboration goals. This approach aim at modelling a number of variables such as the students' demographics, experiences, preferences, and relationships. From studying existing literature on group formation applications for education, we noticed that most toll try to target a large number of variables when forming the groups. This is because some teachers try to form groups based on the available or collected student data, whereas some teachers prefer allocating students to groups at random. The way in which the students are allocated would depend on the objective of the learning activity. In this thesis, we try to model many variables for these teachers who prefer this type of tools, and to whom this approach will add value to the the way they form groups. We note that the constraint satisfaction approach provides us with optimal solutions but would not scale when the dealing with larger datasets such as the ones available from virtual universities.

- The problem of forming groups for education with a larger dataset can be solved

using heuristic approach, where the solution might not be optimal, but can be enhanced if the participants' description is enriched with Semantic Web ontologies.

To investigate the hypothesis, we implement two different ways of group formation, the first approach models the problem of forming groups as a constraint satisfaction problem, where a list of constraints is identified and the aim is to minimise the violation of the constraints. The constraint satisfaction problem is expressed in logic programming, and the violation of the constraints is reported to the person forming the groups. We compare the performance of this approach to existing tools for allocating students to groups to facilitate cooperative learning.

The second approach is a heuristic one, we use clustering algorithms to measure the similarity between the participants in order to allocate them to the right cluster. The similarity is related to the criteria proposed by the teacher. We use Semantic Web ontologies to enrich our data about the participants as we aim to improve the similarity measure between the semantically described participants through inference. We measure the users' satisfaction to compare the results of the clustering alone to the results of clustering with the addition of the ontology. We then report on the performance of the two approaches and conclude with a discussions on the usability of each approach in education, and the future directions of improving the performance of both.

We emphasise that this research does not include the following:

- Proving that any particular set of constraints leads to better results in terms of the performance of the groups. This, however, can be measured in terms of user's satisfaction.
- Claiming that any particular algorithm leads to best grouping.

To evaluate the hypothesis, we studied a number of topics that define the scope of this research. In particular we focused on the following topics, where the topic of this thesis falls in between:

- Collaboration in education and the problem of forming groups. Understanding the depth of this problem goes back to the theories of social learning and types of groups.

- Different types of algorithms can be used to form different types of groups for different collaboration objectives.
- Semantics and knowledge representation, particularly Semantic Web applications and ontologies

1.4 Contributions

This thesis investigates the use of semantics to model different elements in forming groups, particularly the constraints of the groups formation and the participant's data. With the methodology and findings of this thesis, the researcher believes that the key contributions of this work can be summarised as follows:

- Major contribution: the major contribution of this research is the integration of semantics into the process of two different approaches to forming groups for education. Given that the problem of forming groups is a well known problem in many research domains, therefore there has been many algorithms implemented to automate the formation of groups whether for education or otherwise. In this thesis we study the way in which semantics can be integrated into two of these algorithms (constant satisfaction and clustering) in order to generate better grouping. In these two approaches we define a good group as a group that satisfied the teacher' s constraints in the constraint satisfaction approach, and a group that improved the participants' satisfaction in the clustering approach. This contribution is presented in Chapter 7 for the constraint satisfaction approach, and chapters 8 and 9 for the clustering approach. We list the some minor contributions of this thesis below:
- A framework for evaluating different aspects of group formation such as the user satisfaction with the groups they have been allocated to, or the number of constraints that have been violated cross the cohort of participants. This contribution is presented in chapter 6, section 6.2
- A tool that allocates a number of students to groups based on a set of constraints chosen by the teacher. The tool solves the allocation to groups as a constraints satisfaction problem. This is presented in Chapter 7, section 7.3

- A model of describing learners. The researcher has created an ontology based on studying the literature on learner profile standards in educations such as PAPI and IMS LIP, this ontology is presented in chapter 7, section 7.3.2
- A model for describing Computer Science related topics used to describe CS researchers' interests. The researcher created a domain ontology that maps instances of an existing folcsonomy for describing ECS research interests to the existing ACM classification of CS topics, also modelled as an ontology for this research. This minor contribution is presented in chapter 9, section 9.2

1.5 Outline of research chapters

This report investigates the relevant literature to the problem of forming groups for e-learning, state a hypothesis for the research of the PhD, and provides a proposed solution to analyse the hypothesis with an outline of the future work needed to draw a conclusion on the soundness of the hypothesis. The organisation of this thesis is as follows:

Chapter 2 Introduces the reader to social learning theories of development referencing key researchers in this field. The chapter also introduces the principles of collaborative learning, cooperative learning, interaction, and computer supported collaborative learning. The chapter concludes with an insight into the current and future directions of applying social software technologies to enhance the learning experience.

Chapter 3 provides an introduction to group formation and the different approaches to allocate students to groups. The chapter lists the types of groups used in education from simple teams to complex networks.

Chapter 4 discusses existing computer-supported applications developed to facilitate forming groups in general and for education in particular. In this chapter, we focus on the different attributes of these applications and the technologies they implement as we discuss their performance and limitations.

Chapter 5 sheds the light on the Semantic Web, its concepts and potential in enhancing

e-learning applications in general; and as a potential technology for enhancing the automation of group formation in particular.

Chapter 6 discusses modelling group formation as a constraint satisfaction problem to obtain the optimal allocation of students to groups. The chapter also propose a metrics framework for evaluating group formation based on constraint satisfaction.

Chapter 7 introduces a framework that implements group formation as a constraint satisfaction problem in education using logic programming (Datalog) with strong and weak constraints to describe the constraints of forming groups, and Semantic Web ontologies to describe the participants. The performance of this approach is evaluated at the end of the chapter.

Chapter 8 introduces modelling group formation with a heuristic approach. In this chapter, we implement clustering algorithms to form groups of participants in education. We analyse the results of the clustering in comparison to our constraint satisfaction approach.

Chapter 9 builds on the results of chapter 8 by implementing the same clustering approach to the same datasets to form groups, but this time we add an implementation of a Semantic Web domain ontologies to describe the participants. We compare the results of the clustering algorithm before adding the data semantics through the ontology and after adding the semantics. We run a user study to confirm our results.

Chapter 10 concludes the thesis by reviewing the key points and linking them to the achieved findings. The chapter also discusses the shortcomings of the thesis tool and suggests various enhancements. The chapter ends with some future research directions.

Chapter 2

Social Learning

"To learn is to work collaboratively to establish and maintain a community of knowledgeable peers" (Bruffee, 1984),

The social dimension of learning has always been of a great importance to both teachers and learners. For a long time, however, learning has been studied in cognitive psychology as an individual process. Psychology articles related to teaching and learning had generally examined issues of cognition from an individualistic perspective (Voss et al., 1995). However, due to the emergence of few theories over the last century, it became evident that learning is a social process. Research on social learning started with theories from cognitive science that promoted a psychology that focused on "meaning making", therefore entailing some form of constructivism (Bruner, 1990). It was not until the 1980s and 1990s, that research witnessed the "sociocultural revolution" with its focus on learning and on the acquisition of intellectual skills through social interaction.

Similarly, for many years, theories of collaborative learning tended to focus on how individuals function in a group. This reflected a position which was dominant both in cognitive psychology and in artificial intelligence in the 1970s and early 1980s, where cognition was seen as a product of individual information processors. More recently, however, the group itself has become the unit of analysis and the focus has shifted to

more emergent, socially constructed, properties of the interaction (Dillenbourg et al., 1996).

Over the last decade of the century and up to this date, the development of the World Wide Web has been changing many areas of human activity, and among them learning. Traditionally learning has been perceived as time and place dependent (Warschauer, 1997), instructor-driven (Bruffee, 1984), supported by linear learning design (Dillenbourg et al., 1996) (Alani et al., 2003). Consequently, learning has been aimed at mass participation rather than being personalized, which resulted in the learning processes not being suitable for every potential learner. However, as e-learning emerged, the perception of the learning process changed dramatically to be: time and place independent, learner-driven, supported by adaptive learning design.

2.1 Socially oriented theories of learning

2.1.1 Piaget and socio-congnitive theory

The socio-cognitive conflict theory of Piaget (1985) suggests that social interaction leads to higher levels of reasoning and learning, in that social interaction creates cognitive conflict. This is reflected in the fact that the contradiction between the learner's existing understanding and what the learner experiences creates disequilibrium, which in turn leads the learner to question his or her beliefs and experiment with new ideas. In Piaget's words: "disequilibrium forces the subject to go beyond his current state and strike out in new directions" (Piaget et al., 1985).

Piaget further suggests that the social exchanges between peers are more likely to lead to cognitive development that exchange between, for example, a student and a teacher. This observation was premised on the belief that among age peers, there is mutual control over the interaction, whereas if the learner's partner's cognitive level was too much in advance of the learner's, the process of "striking out in new directions" is less likely to take place. This hypothesis, that peer interaction in particular provides greater opportunities for learning, was argued to be valid when the learning involves transformation of perspective, whereas, gaining new skills or strategy might be best obtained by working with more skillful and experienced partners (Damon, 1984).

2.1.2 Vygotsky and socio-cultutal theory

Prior to Piaget's socio-cognitive theory, one of the evolutionary theories that illustrates the role of social processes as a mechanism for learning, is the socio-cultural theory of Vygotsky (Vygotsky and Cole, 1978). The theory claims that the social interaction plays a fundamental role in the development of cognition. In his book, Lev Vygotsky, a Russian psychologist, suggested that "the social dimensions of consciousness is primary in time and in fact. The individual dimension of consciousness is derivative and secondary". This notion explains that the mental functionality of the individual is not simply derived from social interaction, but rather, the specific structures and processes revealed by individuals can be traced to their interactions with others.

Vygotsky recognized that ideas have social origins, that they are constructed through communication with others, and that an individual's cognitive system is a result of communication in social groups and cannot be separated from social life (Vygotsky, 1997), (Vygotsky and Cole, 1978).

Vygotsky also introduced the concept of the Zone of Proximal Development (ZPD), that is, the realm of potential learning that each learner could reach within a given developmental span under optimal circumstances and with the best possible support from the teacher and others in the environment. Vygotsky stressed that collaborative learning, either among students or between students and a teacher, is essential for assisting each student in advancing through his or her own ZPD, that is, the gap between what the learner could accomplish alone and what he or she could accomplish in cooperation with others who are more skilled or experienced (Vygotsky, 1997). In this model, the teacher acts as a facilitator, who assists students, not as a model but rather as a guide, while students collaborate to create connections between new ideas, and prior knowledge.

In comparison to Piaget's theory, while the socio-cognitive approach focused on individual development in the context of social interaction, the socio-cultural approach focuses on the causal relationship between social interaction and individual cognitive change. The basic unit of analysis in this context is social activity, from which individual mental functioning develops.

2.1.3 Social constructivism

Descending from Vygotsky's social development theory, postmodern constructivist perspectives reject the notion that the locus of knowledge is in the individual, but rather that learning and understanding are regarded as inherently social, and that cultural activities and tools are regarded as essential to conceptual development (Palincsar, 1998).

In the 1990s, research in learning and teaching found a new focus point that investigates a social constructivist perspective. This perspective focus on the interdependence of social and individual processes in the co-construction of knowledge, and the multiple ways in which knowledge could be structured among individuals working together, as groups could attain more success that individuals working alone. In this context, thought, learning and knowledge are not just influenced by social factors, but are social phenomena. From this perspective, and according to Barbara Rogoff (1998), a children psychologist, cognition is a collaborative process.

Palincsar, in (1998), reviewed the influence of social and cultural factors on cognition, drawing from Piagetian and Vygotsikan accounts. In this empirical research, the author illustrated the application of institutional analysis to investigate schooling as a cultural process, and the application of interpersonal analysis to examine how interactions promote cognition and learning, particularly, the acquisition of expertise and assessment practice, such as explaining one's thinking to another leads to deeper cognitive processing.

2.1.4 Situated Learning

Other social constructivist concepts include context and situated cognition. Here, the context (i.e., setting and activity) in which knowledge is developed cannot be separated from learning (Rogoff and Lave, 1984) (Lave and Wenger, 1991). Thus, learning is fully situated or located within a given context. Learning occurs while people participate in the socio-cultural activities of their learning community, transforming and constructing their understanding and responsibilities as they participate.

Lave and Wenger (1991), argue that learning is a function of the activity, context, and culture in which it occurs, where social interaction is a critical component of situated learning. Called the "situated learning theory", the authors describe learning as a

process of being part of a community of practice, where the learners become involved in a "community of practice" that represents certain beliefs and behaviors, and as the newcomer learner moves from the border of this community to its center, they become more active and engaged within the culture and thus take the role of expert or old-timer.

In their work, Lave and Wenger (1991) also stress that learning needs to be understood in relation to the development of human identity. In learning to be, in this perspective becoming a member of a community of practice, an individual is developing a social identity. In turn, the identity under development shapes what the person comes to know, how he or she assimilates knowledge and information (Brown and Duguid, 2002).

The concept of shared cognition is deeply intertwined with the situated cognition theory. In this research, the environment is an integral part of cognitive activity, and not merely a set of circumstances in which context-independent cognitive processes are performed (Dillenbourg et al., 1996). Thus, the focus is placed largely on the social context, not only the temporary group of collaborators, but the social communities in which these collaborators participate.

In a community of learners, ideas are best shaped through reflective inquiry with other people (teachers, peers, and so on), who help the learner negotiate his or her own ZPD. In a strong learning community, these people provide scaffolding, consisting of multiple forms of assistance that can be removed bit by bit as the learner becomes more proficient in the skill they aimed to learn.

Moreover, this theory claims that situated learning is usually unintentional rather than deliberate, so it is more effective for the learner to belong to a self-selecting communities rather than being assigned to some group. This is certainly the case in the classroom where the learner knows their colleagues and might have preference of whom to interact with. However, in a distance learning world that employs distance e-learning and virtual universities concepts (Skolnik, 2000) (Rada, 2001) (Cavanaugh, 2004), identifying fellow learners for interaction might not be so simple, and recommendation might be necessary to help the learner achieve their objectives.

In (1998), Wegerif provided evidence of the importance of considering the social side of learning when designing a course. The author showed that individual success or failure

on the course depended on the learner feeling as insider or outsider in the learning process. The paper gave an insight into designing courses that support community building and collaborative learning.

2.2 Collaboration in Education

As seen in the previous section, collaboration has long been considered an effective approach to learning. Since at least the 1960s, before the advent of networked personal computers, there was considerable investigation of cooperative learning by education researchers such as Stahl et al. (2006).

According to Bruffee (1984), although the term "collaborative" learning was coined, and the basic idea first developed in the 1950s and 1960s by a group of British secondary school teachers and researchers studying post-graduate education. Collaborative learning began to interest American college teachers and other communities widely only in the 1980s.

In this paper, the author reported that a common phenomenon was that new college students experienced difficulty adapting to the college classroom and performed poorly although they had excellent secondary preparation on paper. The teachers took action to overcome this difficulties by mandated programs and providing tutoring and counselling programs staffed by graduate students and other professionals. However, although many solutions were tried, they failed as the undergraduates refused to use the assistance.

Taking hints about the social organisation of learning, some college faculty members guessed that students were refusing help because the kind of help provided seemed merely an extension of the work, the expectations, and above all the social structure of traditional classroom learning, which left them unprepared in the first place.

To provide an alternative to the traditional classroom teaching, some colleges turned to peer tutoring, where teachers could reach students by organising them to teach each other. This new form of indirect teaching, where the teacher sets the problem and organises students to work it out collaboratively, was referred to as "collaborative learning". The students' work tended to improve when they received the assistance from peers, and learned from the students they helped, and from the activity of helping itself.

Albert Bandura (1969), a psychologist specialising in social cognitive theory and self-efficacy, and famous for his social learning theory, analysed similar behaviours in children learning, where the child must rely on siblings and peers as models for specific models of behaviour that parents and teachers do not ordinary provide.

2.2.1 Collaborative learning

Collaborative Learning is broadly known as a situation in which two or more people learn or attempt to learn something together. In Dillenbourg (1999), the author claims that a precise definition of what collaborative learning means is highly negotiable. In practice some people may consider a collaboration to be three to four participants performing an activity together for twenty minutes, while others may see it more of a forty professional people working on a problem solving task for a year. Therefore, Dillenbourg suggests some dimensions in which the nature of the collaboration can be more precise:

- Group size and time span: defines the number of people involved in the collaboration and the duration of the collaboration, where the optimal group size depends on the situation such as the group, the task, and the context (Nunamaker et al., 1991).
- Learning: defines the collaboration aim and the learning gain from performing it, e.g. following a course, performing a problem solving activity, or learning from lifelong work practice.
- Collaboration: defines the different forms of interaction the participants of the collaboration may use such as face-to-face computer-mediated, synchronous or otherwise.

The forms of collaboration differ in purpose, length, the complexity of tasks, and the degree of formality (Dillenbourg, 1999). Examples of the most widely used forms of collaboration are: group discussions, where students share views on issues; peer coaching; group projects, where students cooperate to solve a specific problem; study groups, where struggling students can seek help from better students in their group; social groups, where students have a common interest they practice or discuss together.

As defined by Bruffee (1993), collaborative learning is a reacculturative process that helps students become members of the knowledge communities whose common property is different from the common property of knowledge communities they already belong to. In this paper, the author argues that collaborative learning provides a particular kind of social context for conversation, a particular kind of community -a community of status equals (i.e. peers)- where students learn "the skill and partnership of re-externalised conversation" (Bruffee, 1984).

From this point of view, and the literature mentioned in the previous section, social constructivism philosophy is the foundation for collaborative learning, and hence, collaborative learning focuses on social relationships in a community of learners. Communities in this context, usually self-selected, are facilitated by the teacher. Therefore, the role of the teacher in collaborative learning involves engaging students in conversation among themselves and motivating the emergence of learners' communities (Bruffee, 1984).

Literature in educational psychology and cognition showed the importance of collaborative learning. Several studies investigating higher education learning found that students who follow in-class collaborative learning procedures and actively interact with each other are more satisfied with their learning experience and evaluate their courses more favourably than students who are exposed to the traditional lecture method (Bligh, 1972). This is due to the fact that collaborative activities enhance learning by allowing individuals to exercise, verify, solidify, and improve their mental models through discussions and information sharing during the problem-solving process, while working on the assigned academic task (Alavi, 1994).

According to (Cuseo, 2002), research in many disciplines has shown that learning within groups improves the students learning experience by enabling peers to teach each other and learn from each other in various ways. There are several forms of collaborative work that allow the students to learn from each other in different ways. The variety of forms of collaboration has proven to increase the students attention, interest, and motivation, hence, ensuring they are actively involved in the learning process both mentally and socially.

2.2.2 Cooperative Learning

Cooperative learning is defined as a group learning activity organised so that learning is dependent on the socially structured exchange of information between learners in groups, and in which each learner is held accountable for his or her own learning and is motivated to increase the learning of others (Oxford, 1997). Cooperative learning therefore promotes positive interdependence, where students within the learning groups should feel responsible for one another's learning (Alavi, 1994).

According to Alavi (1994), cooperative learning research evolved from the work of educational or social psychologists such as Johnson and Johnson (1986) and Slavin (1989). It involves social (interpersonal) processes by which a small group of students work together (i.e., cooperate and work as a team) to complete an academic problem-solving task designed to promote learning. Researchers in this field, found that compared to individual and/or competitive instructional methods, collaborative instructional methods involving cooperative procedures are more effective in promoting student learning and achievement (Slavin, 1993), including promoting critical thinking (Gokhale, 1995).

Johnson et al. (1998) assert that "what we know about effective instruction indicates that cooperative learning should be used when we want students to learn more, like school better, like each other better, like themselves better, and learn more effective social skills". Numerous studies indicate that compared to competitive or individualistic learning experiences, cooperative learning is more effective in promoting intrinsic motivation and task achievement, generating higher-order thinking skills, improving attitudes toward the subject, developing academic peer norms, heightening self-esteem (often through assigning roles), increasing time on task, creating caring and altruistic relationships, and lowering anxiety and prejudice (Slavin, 1993) (Alavi, 1994) (Gokhale, 1995) (Oxford, 1997) (Johnson et al., 1998) (Dillenbourg, 1999). Cooperative learning, however, is not a replacement for individual learning, as these two approaches can target different objectives and different learning activity.

2.2.3 Cooperative versus Collaborative learning

In research papers, the terms collaborative learning and cooperative learning are often used interchangeably (Dillenbourg, 1999). However, in the literature that distinguish

the two terms, cooperative learning, as compared with collaborative learning, is considered more structured, highly organised, more prescriptive to teachers about classroom techniques, more directive to students about how to work together in groups, has specific aims, and is more targeted to the secondary school population than to postsecondary or adult education. On the other hand, collaborative learning, when compared with cooperative learning, seems less technique-oriented, less prescriptive, and more concerned with acculturation into the learning community (Oxford, 1997).

Therefore, practically, cooperative learning refers to a particular set of classroom techniques that foster learner interdependence as a route to cognitive and social development. Whereas, collaborative learning has a "social constructivist" philosophical base, which views learning as construction of knowledge within a social context, and which therefore encourages acculturation of individuals into a learning community.

For cooperative learning, when designing the learning activity, the teacher should decide the level of structure of the activity, often referred to as scripting, where a collaboration script is a set of instructions that explains how the group members should interact, how they should cooperate and how they should solve the problem. When teachers engage students in cooperative learning, they usually provide them with global instructions such a "do this task in a group of 4". These instructions usually come with implicit expectations with respect to the way students should work together. In this context, scripts bridge the gap between collaborative learning and traditional instructional design (Dillenbourg, 2002). In his research, Dillenbourg warns about teachers over scripting cooperative learning, as this might reduce the effectiveness of collaboration as a social construction of knowledge, by failing to trigger the cognitive, social, and emotional mechanisms that are expected to occur during collaboration.

2.2.4 Enhancing the effectiveness of collaboration in education

To increase the effectiveness or efficiency of interacting groups, it is relevant to increase the group process gains and/or reduce the group process losses. Process gains refer to the synergistic aspects of the group interaction that improve group performance relative to the individual member performance. Process losses, on the other hand, refer to certain aspects of the group interactions that impair group performance relative to the efforts of individual members working alone (Oxford, 1997). The following summarises

some examples of group process gains and process losses as adapted from the work of Nunamaker et al. (1991):

Group Process Gains: These include:

- A group as a whole generates more information and alternatives compared to the average group member;
- Groups are more effective and objective in evaluation and error detection tasks;
- Working in a group may motivate the individual member to perform better;
- Interactions among the group members leads to synergies.

Group Process Losses: These include:

- Member participation in the group process is fragmented (i.e., group members should take turns in speaking);
- One or a few individual members may dominate group discussions and monopolize the group's time;
- Fear of negative evaluation (evaluation apprehension) causes members to withdraw and avoid participating in the group discussions;
- Higher volumes of information generated during the group process creates a condition of information overload for individual members

When facilitating a collaborative activity, the teacher should take these gains and losses into account. In empirical studies, these effects are usually evaluated through questionnaires that evaluate the participant's reactions, such as perceived skill development, self-reported learning, interest in learning the subject, evaluation of classroom experience, and evaluation of group learning experience, also referred to as group case analysis (Alavi, 1994).

2.2.5 Interaction

Another relevant concept of collaboration is interaction. Interaction refers to the situation in which people act upon each other, and the personal communication, which

is facilitated by an understanding of key elements such as willingness to communicate, style differences, and group dynamics.

According to Oxford (1997), in educational settings, interaction involves teachers, learners, and others acting upon each other and consciously or unconsciously interpreting (i.e., giving meaning to) those actions. Therefore, interaction involves meaning, however, it might or might not involve learning new concepts. Elements related to interactions are:

Willingness to communicate is related to a feeling of comfort, high self-esteem, extroversion, low anxiety, and perceived competence, whereas unwillingness to communicate (i.e., communication apprehension) is associated with the opposites: discomfort, low self-esteem, and introversion.

Several studies found a greater amount of student participation according to three measures:

1. percentage of student talk versus teacher talk,
2. directional focus of student talk (toward other students or toward the teacher),
3. equality of student participation (Warschauer, 1996)

Learning styles can be defined as the general approaches students use to learn a new subject or tackle a new problem. Identifying learning styles is usually processed through surveys or a long interaction with the students.

Teachers and learners need to understand style conflicts if they have taken the time to identify and discuss their own preferred styles. Understanding the style preferences of individual learners helps the teacher design lessons that provide a range of collaborative activities suitable for all the people in the class, neither slighting nor favouring a particular set of individuals (Oxford, 1997). However, employing learning styles in teaching and learning is a debatable topic, in that people change their learning styles in relation to the context and time of the learning activity. It might be also useful for the learners to be pushed to change their learning styles and be introduced to new learning styles.

Group dynamics is related to the life of the group, which involves four stages: group formation, conflict, cohesion, and problem solving (Frank and Brownell, 1989).

The fist stage, involved creating the bond to the group by understanding its culture. It's a known fact that the group, which is richer in resources than any single individual, affects members' attitudes, such as confidence and satisfaction, and these attitudes influence interaction.

Groups provide guidelines for behaviour within the group, that might be very different from behaviour outside the group/culture of the community of participants. These guidelines offer standards for self-evaluation, and help learners maintain energy. Harris (1992), an organisational behaviour specialist, describes three types of group cultures:

- The authoritarian/bureaucratic culture: uses rules, laws, rewards, and punishments to control members, with the desired end being compliance to authority.
- The compromise/supportive culture: uses interpersonal or group commitment, discussion, and agreement, with the desired goal of consensus.
- The performance/innovative culture: emphasises internally controlled, highly individualistic ideas, with the goal being self-actualisation and individual achievement.

According to (Dillenbourg et al., 1996), two types of interaction have been universally referred to in collaborative learning research: *negotiation*, often referred to within the Vygotskian "cooperation" approach as an indicator of joint involvement in task solutions, and *argumentation*, as a possible means for resolving socio-cognitive conflict. Both types of interaction would relate to different cultures of collaboration.

Depending on the teacher and the group members, the classroom can contain any of these types of group cultures (Oxford, 1997). It is important for the teacher to motivate the collaboration with awareness to the type of group culture that would occur within the group, as the participation of the students will depend on this culture.

2.3 Computer Supported Collaborative Learning (CSCL)

According to Stahl et al. (2006), Computer Supported Collaborative Learning (CSCL) arose in the 1990s in reaction to software that forced students to learn as isolated individuals. The exciting potential of the Internet to connect people in innovative ways provided a stimulus for CSCL research.

So far, many technologies have been introduced to support different areas or collaborative learning such as discussion forums, co-authoring, and collaborative brain storming. In terms of empirical research, the initial goal was to establish whether and under what circumstances collaborative learning was more effective than learning alone (Dillenbourg et al., 1996). Consequently, early research on CSCL involved studies of the potential and benefits of employing CSCL and Computer Mediated Communication (CMC) in supporting collaborative and cooperative learning, and user interaction, mainly through empirical studies (Warschauer, 1997).

Research on CMC involved designing and developing CSCL systems to support interaction such as interface design (Baker and Lund, 1997), electronic meetings systems (Nunamaker et al., 1991), and collaboration scripts (Dillenbourg, 2002), that supports sociocultural values that form the foundation of collaborative learning and interaction (Warschauer, 1997).

The main outcome of empirical studies is that the structured interface is able to promote an interaction that enables learners to effectively collaborate in problem-solving, using flexibly structured CMC (Baker and Lund, 1997). This approach proved CMC to be more effective than traditional strategies[1] to learning in terms of learning outcomes and student affective reactions (Alavi, 1994). Research in this area also showed that CMC plays a significant role in promoting collaborative and cooperative learning mainly through few features that distinguish it from other communication media. The features listed below are adapted from the work of Warschauer (1997):

Many-to-many communication: occurs when any member of the group may initiate interaction with any or all of the others. Studies conducted on the social dynamics of CMC have found that CMC results in communication that is more equal in

[1] In this context, we refer by traditional strategies to the instructivist methodology of learning

participation than face-to-face discussion, with those who are traditionally shut out of discussions benefiting most from the increased participation.

Researchers found that in discussions held electronically, women made the first proposal of a solution to a problem as often as men, whereas in face-to-face discussions men made the first proposal five times more often (McGuire et al., 1987). It was also found that proposals by higher status users, such as graduate students compared to undergraduates, were invariably favored during in-person discussion groups, whereas proposals by lower status and higher status users were selected equally as often in electronic discussion groups.

Researchers justified the occurrence of greater equality in CMC with the following factors (Sproull and Kiesler, 1991) (Warschauer, 1997):

- CMC reduces social context clues related to race, gender, handicap, accent, and status.
- CMC reduces nonverbal cues, such as frowning and hesitating, which can intimidate people, especially those with less power and authority.
- CMC allows individuals to contribute at their own time and pace.

Time and place independence: Time and place independent communication extends the potential of online collaboration in several ways, and supports distance learning.

Long distance exchanges: Researchers argued that CMC makes long distance exchanges faster, easier, less expensive, and more natural, with interaction between learners occurring on a frequent rather than occasional basis. In addition to that, by adding many-to-many communication, an entire group of students can have regular access to interacting with any or all of another group of students, and students from many different schools can interact together as well. This feature is nowadays displayed in the way students interact in social networking sites.

Hypermedia links: CMC allows multimedia documents to be published and distributed via links among computers around the world. This characteristic is related to the World Wide Web and has implications for collaborative learning. Hypermedia can provide access to up-to-date, authentic information, which can then be incorporated into classroom collaborative activities.

CMC allows personalisation of content as it allows students to collaborate using information related to their own personal interests gathered from a variety of sites all over the world. The most effective collaborative activities involve not just finding and using information, but rather actively making use of technologies to construct new knowledge together. The use of the Web allows access to rich material that integrates textual, audio, and audiovisual material together. In the recent years, the usage of repositories, blogs, wikis, and other forms of social web sites with user generated content for education have increased in higher education[2], allowing a smooth collaboration between participants. This is discussed in more details in section 2.4.

Through these features, CMC provides an impressive array of new ways to link learners. When viewed in the context of socio-cultural learning theory, which emphasises the educational value of creating cross-cultural communities of practice and critical inquiry, these features make online learning a potentially useful tool for collaborative language learning.

2.3.1 Challenges for CSCL

Although research in CSCL and CMC is well established, there is a number of challenges to be considered. These challenges are mostly inherited from social problems existing in distance learning (Cavanaugh, 2004). Some of these challenges are:

1. Socialisation: includes providing sufficient attention given to the learner and lack of face to face interaction.

2. Groups management: for collaborative learning is difficult as learners may not know each other, and describing them effectively for the teacher to allocate them into groups is difficult. As collaborative learning is associated with communities of learners, ensuring the right people join the right community is essential to the benefit of both the individual participants and the community life.

[2]See the following link for a list of UK universities that use social web for education. The list was compiled by the author of this thesis as a part of a JISC funded research project: http://wiki.semtech.ecs.soton.ac.uk/index.php/Survey_of_Semantic_Technology_use_across_UK_universities

3. Learner suitability: as seen before, students have different demographics, interests, preferences, learning experience, learning styles and ability which may be mismatched to the collaborative activity. This effect can be seen as a personalisation of the learning at the individual level, or the group level, as communities of learners emerge based on sharing interests and goals.

Personalisation is a very important issue that is applied in many domains other than learning. Since the evolution of e-commerce websites begun, providing personalised services and products to the users has become crucial to the satisfaction of the customers. Nowadays, many websites have some level of customising services to the customer liking. Amazon for example uses recommender systems to provide similar products to what the user has bought or looked at, or to what similar users have purchased. The same perspective applies in e-learning and adaptive hypermedia domain, a more quality service that caters most for the learner needs and preferences.

Adaptive Web-based educational systems emerged as an alternative to the traditional "one-size-fits-all" approach. These systems build a model of the goals, preferences, and knowledge of each individual student and use this model throughout the interaction with the student in order to adapt to the needs of that student (Brusilovsky and Nijhavan, 2002). Using this concept in designing learning with awareness to generating sound pedagogy, increases the potential effectiveness of the learning experience (O'Keefee et al., 2006) (Dagger et al., 2004). Providing automated personalisation of services is therefore crucial to learning, which includes the social nature of learning. In relation to this research, we recognise two types of personalisation:

Personalisation at the individual level: This type is the most common one in e-learning which is getting the learner what they want. The aim of collaboration in learning is to provide a way for students to learn from each other while ensuring that every student will have a learning benefit from participating in the collaboration whether it is learning from a more competent peer, learning by teaching a less competent peer, or just cooperating with a peer to solve a problem or discuss a topic. When forming groups, a teacher has to consider the possible individual benefit of the student from participating in the group and aim at personalising their allocation to the group for each individual in the class.

Personalisation at the group level: This notion is concerned with the idea that if learners are similar enough to create self-selecting networks for themselves, then we can provide members of the same community with similar services. Considering the group as an entity against other groups can enable the relation between the members of the group, i.e., the group goal such as the common interest or practice, to be the base of the personalisation. In particular, social networks and communities of practice can be very useful in this context, as they are based on the links between people who share a common interest. Research on recommendation (and recommender systems), in particular, can make a good use of social networks by recommending material to the members of the same network.

2.4 Present and future of social learning

Recent years have seen the emergence of Web 2.0, in which users are not only passive recipients of the featured content, but actively engaged in constructing it. The Internet technologies of the subsequent generation have been profoundly social, as listservs, Usenet groups, discussion software, groupware, and Web-based communities have linked people around the world. However during the past few years, social software such as blogs, wikis, trackback, podcasting, video-blogs, and social networking tools such as MySpace and Facebook (Alexander, 2006) create new dimension for online social presence.

In his book *"Here Comes everybody"*, Shirky (2008) argues that if people can share their work in an environment where they can also converse with one another, they will begin talking about the things they have shared. Here, Shirky (2008) quotes author activist *Cory Doctorow* stating: "Conversation is king. Content is just something to talk about". In education, collaboration, being of a social nature, involves this exact concept. Bruffee (1984) argued that what students do when working collaboratively, for example on their writing, is not write or edit or, least of all, read proof. What they do is converse. They talk about the subject and about the assignment. The author also argues that collaborative learning as a classroom practice models more than how knowledge is established and maintained, but how knowledge is generated, how it changes and grows as well, which is modelled in this new way of user generated content.

In his paper on e-learning 2.0, Downes (2005) argues that Web 2.0 is not a technological revolution, it is a social revolution. That Web 2.0 is an attitude not a technology, as it's about enabling and encouraging participation through open applications and services.

It wasn't long before educators began to notice something different happening when they began to use tools such as wikis and blogs in the classroom, which led them to think how can they implement social software in their teaching (Liccardi et al., 2007). Alexander (2006) argues that educators noticed that, instead of discussing pre-assigned topics with their classmates, students found themselves discussing a wide range of topics with peers worldwide. Students' blog posts are often about something from their own range of interests, rather than on a course topic or assigned project.

In this way, the e-learning application becomes, not an institutional or corporate application, but a personal learning centre, where content is reused and remixed according to the student's own needs and interests promoting personalisation due to the openness nature of education in providing accessible material to all. Alexander (2006) argues that unlike those in the corporate world, those in higher education tend to share their ideas and their outcomes openly and proudly. By engaging in exactly this sort of sharing, educators can capitalise on social networking technologies in ways that will benefit both teachers and students.

Goggins et al. (2007) studied how these social tools influence group behaviour and community formation in online systems. They found that tools have a very important role in the development of interaction as they support social awareness and collective action among users. For example, results from empirical experiments showed that using wikis improved the work product of students, in virtual groups, in comparison to exchange of files.

Blogs, given bloggers propensity for linking, not to mention some services ability to search links, blogs and other platforms readily lead a searcher to further sources. Students can search the blogosphere for resources such as political commentary, as well as analyse how a story, topic, idea, or discussion changes over time and explore different views on it as blogs are more about posts, rather than just pages.

Dippold (2009) observed that due to the opportunities for self-reflection and interac-

tive learning offered by blogs, they have become one of the emerging tools in language pedagogy and higher education. The author researched to what extent blogs can facilitate peer feedback and what issues need to be addressed for them to be a valuable tool in this process. Through experimentation with students' blogs, questionnaires, and focus groups, she argued that blogs are potentially valuable tools for peer feedback, but entail the need to address specific issues regarding the choice of CMC tool for feedback tasks, training in the use of interactive online tools and the roles of teachers and students.

Social Networks, Yang and Tang (2003) investigated the effects of social networks on students' performance in online education to analyse which social relations are linked with the students' academic performance. Through empirical studies, they found that advising networks and friendship networks variables are positively related to the student performance both face-to-face and online, whereas adversarial networks are negatively related. They also found that while advising and adversarial networks variables are good determinants of students' performance, friendship networks are not.

Lampe et al. (2006) observed that as large numbers of college students have become avid Facebook[3] users in a short period of time, students are largely employing Facebook to learn more about people they meet offline, and are less likely to use the site to initiate new connections. Lampe et al. found that Facebook users anticipate their profiles being searched and viewed by peers, not faculty, administration within the campus community, or outsiders. Similar results were reported by Berg et al. (2007). The strongest expectations are that peers who have some sort of offline connection - either by virtue of prior friendship, common classes, or having met at a social event - constitute the audience for one's profile, although students also primarily use Facebook to find offline connection.

Social bookmarking, used by a single person or groups, is used in education for collaborative information discovery. Alexander (2006) suggests that researchers (students, faculty, staff) can quickly set up a social bookmarking page for their personal and professional inquiries, store links, and find people with similar interests, which has been happening in some universities such as University of Pennsylvania[4] and Harvard[5].

[3]Facebook.com is a social network site originally aimed at college students. Facebooks primary distinction is that participation is structured by offline social networks, initially membership in a university community.

[4]Pennsylvania Tag Library: http://tags.library.upenn.edu/

[5]The H2O project: http://h2obeta.law.harvard.edu/home.do

Clusters of interests reveals patterns of research preference.

In the EDUCAUSE conference, Butler and Chatfield (2009) suggested that social media is everywhere and that educators need to learn to embrace it. That educators are no longer gatekeepers of knowledge as the role of educators is changing. Sarah Robbins[6], featured speaker in the conference, claimed that college students and lifelong learners are no longer limited to learning by the approved technologies and methodologies for a specific department, and they no longer have to fit into specific institutions. Instructors podcasts - such as material in iTunesU[7] and YouTube - and written materials are often offered online to the public, not restricted to their current students. Educators therefore need to adapt to this change to keep their students engaged with their teaching. In a questionnaire run by Berg et al. (2007), a faculty member commented "I had e-mailed a student about coming in for a meeting. I waited three days with no response. I tried contacting the same student through Facebook and received a response in fifteen minutes."

Berg et al. (2007) also suggests that as students connect more frequently through social networking technologies, their expectations grow regarding their connections with and between campus professionals. Therefore, educators are obligated to think in terms of student satisfaction and long-term success. Considering how students use social technologies can help educators build a strong network of information, and think differently about how to offer core services and communicate with students and with each other.

Berg et al. (2007) investigated what aspects of social software might translate into new ways for creating better and more effective student and academic services. At the University of Wisconsin-Madison, they run questionnaires with students and faculty to employ the suggestions on using social systems to improve the delivery of services such as enrolment, campus communications, e-learning, advising, and involvement activities, in a way that match the expectation of this generation of students. Suggestions included: providing photos and profiles within the course management system so that they could get to know each other before meeting face-to-face for a group project; providing mass postings when a campus deadline is approaching or to offer opt-in class chats that may or may not involve the faculty member; allowing anonymous e-chats with faculty and

[6]The podcast for Sarah's presentation is available at:
http://www.educause.edu/blog/gbayne/E08PodcastSocialMediaandEducat/167993

[7]http://www.apple.com/education/mobile-learning/

advisors and posting pop-up alerts about campus safety.

The idea of learning as the steady supply of facts of information, that learning is mere information absorption, is no longer positively embraced (Brown and Duguid, 2002). The shift in implementation of social learning using Web 2.0 changed the production and consumption of content and the way students interact on a large scale. Learning is much more demand driven and personalised. People learn in response to their need (Brown and Duguid, 2002), and with the available tools this is made more clear to educators who need to fit their teaching to the expectation of the new generation, especially now that Web 2.0 meta-services, like social software before them, are heading for the mobile, wireless world. As suggested by Alexander (2006), in the future it will be more widely recognised that the learning comes not from the design of learning content but in how it is used.

2.5 Summary

Based on the social theories of development introduced in this chapter, we note that while the socio-cognitive theory focus on changes in perspectives or restructuring of concepts, the socio-cultural approach emphasises acquiring understanding and skills. From this perspective, researchers concluded that the former theory promotes collaboration between peers of equivalent intellectual ability, whilst tutoring or guidance may be necessary in fostering the latter. Therefore, how groups should be composed with respect to skills and abilities may depend upon what learning outcomes one is interested in and the tasks involved.

Similarly, studies that distinguish collaborative and cooperative learning show that each approach involves a different philosophy of working together. While cooperative learning is more structured and relies on the teacher specifying the features of the collaboration, collaborative learning is more free of instruction, as participation is self-selected into communities of peers. Designing groups for cooperative and collaborative learning is therefore different, but it is, in either case, constrained with a motive that justify allocating the students to groups for a specific task, or joining a specific community. Whether the constraint is related to expertise, skills, abilities, learning styles, demographics, interests or preferences, or even a combination of these characteristics, identifying it to

achieve the goal of the collaboration is essential to the process.

Although group composition might be different in cooperative and collaborative learning, the student's willingness to participate is key factor in the success of the collaboration and interaction with peers. Allocating the students to the right group or recommending them to join the right community that will maximise their benefit will have a positive effect on the students willingness to communicate. Thus, to maximise the gain for the individual and the group, and minimise the loss, forming groups should be done with awareness to this issues.

At this point, we recognise that the problem of forming groups is beyond ignorance. In summary, grouping plays an important role in education, and to reveal its significance and strength, we dedicate chapter 3 to present this topic.

Chapter 3

Group Formation

Group psychology, whether in education or other domains, has been researched for many years. Groups are defined as "a set of two or more persons who are linked through interaction" (Biddle, 1979). Research on group formation theory and practice is very wide due to the different usage of groups. In this chapter, we discuss different aspects of grouping varying from social definitions to complex structure and topology, but we keep a particular focus on grouping for education.

3.1 General definition of group formation

Formation of groups can have different definitions in different domains. The definition however, usually depends on the purpose of the groups, the type of groups, the way it affects the individuals participating in the formation, and a set of assumptions on the way the groups should behave. For example, Lloyd et al. (1999) define a formation, within collaborative virtual environments, as a mechanism that supports the explicit definition and use of groups of participant's through some expression of mutual focus of interest, and introduce group effects. In this context, the authors emphasise that a formation affects a participants' experience through two principle relationships:

- Individuals contribution to the formation: how do individuals contribute to the

overall effect of being part of the group.

- Formations contribution to each individual: how does the formation affect the individual members.

A formation effect may be introduced by some arbitrary combination of formation attributes and individual attributes using a suitable combination function that describes how these attributes contribute to the formation. For example, a formation attribute could be a role in the team (such as a leader), and the function can be that the distribution of the individuals to groups should involve the allocation of a leader in each group. Thereby, we can consider the components of the formation to be: the members of the group formation, the attributes of the formation, the value of the formation (the formation influence on the individual), and the combination function of the attributes (formation criteria or constraints).

3.2 Importance of group formation in learning

The general definition of group formation in education is "putting learners into groups for educational purposes". As discussed in the previous chapter, organising collaborative learning effectively requires doing more than throwing students together with their peers with little or no guidance or preparation. To do that is merely to perpetuate, perhaps even aggravate, the many possible negative efforts of peer group influence: conformity, anti-intellectualism, intimidation, and leveling-down of quality, which lead to group losses. This requires teachers to create and maintain a demanding academic environment that makes collaboration-social engagement in intellectual pursuits -a genuine part of students' educational development (Bruffee, 1984).

For the collaboration to be successful, the different forms of collaboration require different types of groups; and for the groups to be successful, the approach used in group formation has to be considered carefully. The significance of the formation relies on its impact on the group performance and the individual gain of being a part of the group.

In the learning domain, teachers often have to deal with group formation manually which can sometimes turn into a very complex task which has led researchers to

investigate several techniques for automating this process through the use of computer-supported group formation, as will be discussed in the next chapter. However, in most existing research, the applications developed only model a limited range of group types.

3.3 Taxonomy of groups

There are several factors that contribute to the variation of groups. Groups can vary along dimensions such as the size of the group, the duration of the group work, the aim of the group, which is usually presented in the type of the task to be carried by the group and the degree of its formality, and the cohesion of the group (Winter, 2004). The major types of groups are:

3.3.1 Teams

Teams are a planned group of people that collaborate together on a well-defined task or set of tasks. Teams can be as small as a pair of students discussing some aspects of the course or helping each other; or large task-oriented teams working together to solve a complex problem (Cuseo, 2002). Hackman distinguishes teams from other types of groups in that they must meet the following three criteria (Hackman, 1990), (Hackman et al., 2000):

1. They are intact social systems, complete with boundaries, interdependence among members, and differentiated member roles.
2. They have one or more tasks to perform.
3. They operate within an organisational context.

Whether in education or organisations and human resources research, the structure of teams usually depends on the aim of the task. For example, the team may consist of students sharing a common interest; alternatively its members may have been deliberately chosen to cover the full range of academic ability. The teacher may select different strategies depending on the desired learning outcome (Cuseo, 2002). With respect to knowledge, the types of teams are often termed (Hoppe, 1995):

- Complementary teams: where the student with good knowledge in a particular topic can help the student with low knowledge on the same topic to improve the latters competency.
- Competitive teams: where students with similar levels of knowledge in a particular topic stimulate each other under the pressure of competitiveness within the team to improve each others learning competency.
- Problem solving teams: where the team is given a problem that cannot be solved by one individual member of the group, and requires the knowledge of all the members.

In the distance learning domain, teams take a more virtual shape. A virtual team is a group of individuals working towards a common goal who do not interact face to face, or may or may not be geographically close to each other (Edwards et al., 1996). Virtual teams work on interdependently across space, time, cultures, and organisational boundaries on temporary, non-occuring projects with shared purpose while using technology (Michailidou and Economides, 2007). Virtual teams are not very popular comparing to virtual groups such as communities and networks. The next chapters discuss the formation of virtual groups in more details. In some literature, the word "group" is often used to describe a "team" as opposed to any other type of groups (Goggins et al., 2007).

3.3.2 Communities

Communities, also known as communities of knowledge, are an informal group of people that develop a shared way of working together to accomplish some activity (Andriessen et al., 2001). The goal of the community is generally diverse even if the community has been formed to deal with a specific topic. The membership of a community is usually self-selected and self organised. Literature on the emergence of communities, particularly through Web 2.0 technologies has shown the power of group action -through blogs, tags, and other online groups technology- (Shirky, 2008). Theory and practice of this domain is currently studied in multidisciplinary research as part of web science, particularly studying the macro-micro effect of creating and contributing to communities.

In learning, Bruffee (1984) defines a community of knowledgeable peers as a group of people who accept, and whose work is guided by, the same paradigms and the same

code of values and assumptions.

According to McDermott (1999), communities differ from teams in that the former are driven by the value they provide to individual members while teams are driven by the value they deliver in the results they produce. Moreover, the set of interdependent tasks that leads to the defined objective forms the heart of a team, while in the community the heart is formed by the shared knowledge. In addition to that, teams progress by moving through a work plan, while communities develop by discovering new areas to share current knowledge and develop new knowledge.

There are various definitions for various types of communities. Similar to the variations of groups in general, the types of communities differ regarding the following aspects: purpose of the community, boundary of membership, formalisation of set-up; formalisation of co-ordination in terms of members roles, size of community, composition of the community in terms of expertise, and frequency and type of interaction whether it is face to face and/or via computer supported tools (Wenger and Snyder, 2000), (Andriessen et al., 2001).

The most well known type for knowledge sharing communities is **Communities of Practice (CoP)**. According to Wenger (1998), CoP are groups of people informally bound together by a common interest in some subject or a shared expertise (practice) and collaborate to share ideas or find solutions. CoPs tend to have an organic, spontaneous, and informal nature that makes them resistant to supervision and interference from management.

As mentioned in the previous chapter, Wenger argues that learning a practice involves becoming a member of a "community of practice". In his research, he shows the importance of the group (CoP) to both what people learn and how, which involves the concept of practice. Here, people are able to affect one another and the group as a whole. In their book, "The Social Life of Information", the authors Brown and Duguid (2002) argue that ideas and knowledge may be distributed across the group, and not held individually, Therefore, this type of groups allows for highly productive and creative work to develop collaboratively. Here, the members can recruit one another or allow themselves to be found by interested searchers, which is now made easy with the availability of social tools (Shirky, 2008).

Observing Communities of Practice in organisations, Brown and Duguid (2002) described some properties of the Community of Practice. They noticed that in getting the job done, members ignored roles, they had overlapping knowledge, a common work identity, and relatively blurred boundaries.

Communities of Practice differ from Teams in that the latter are usually created by managers -in our case instructors- to complete specific projects. Instructors tend to select team members on the basis of their ability to contribute to the team goals, and the team disbands once the project is completed. CoPs, however, are informal, they organise themselves by setting their own agendas and establish their own leadership. The membership in CoPs is self-selected, where people know when and if they should join. They know if they can contribute to the community and if they will gain from joining it. In addition to that, existing members of the community, when they invite someone to join, they operate on a gut sense of the prospective member's appropriateness for the group (Wenger, 1998) (Goggins et al., 2007).

According to literature, there are other types of communities besides CoPs such as Communities of Interest (CoIs) and Communities of Commitment (CoCs) (Collison, 2000). The variation of these types of communities resides in the level of formality and contract value, i.e. the degree to which a community has to deliver concrete results. Unlike CoCs, CoPs have low formality and contract value. The classification of communities can be also based on their virtuality. A virtual community is one that has some form of computer system facilitating the communication between the members as a central element to its theory (Preece, 2000).

The terminology in communities research can be confusing due to the fact that different names can be found for concepts that are similar to CoPs, such as virtual communities, knowledge communities, and occupational communities (Andriessen et al., 2001). In learning, the teacher often guides the students to form communities based on their interests or preferences to encourage discussion on different topics within the community. Consequently, in this research we will focus on the characteristics along which communities can diverge in the domain of learning and sharing knowledge. Since CoPs are the closest concept to sharing an interest and a social practice with a low formality on the deliverables within the learning domain, in comparison to a community within an organisation, we use the term to refer to communities of learners in this context.

3.3.3 Networks

Similar to communities, there are many types of networks. Here we consider two common sorts of networks:

Intensional Networks: Intensional networks are an informal collection of collaborators who are selected to accomplish a specific task (Winter, 2004). This type of grouping differ from teams in that it is less formal, has a shorter temporal duration, and low group cohesion. The members are not required to be familiar with each other as long as they can cooperate to deliver the task. The definition of intensional networks is similar to that of what is referred to as **Networks of Practice (NoPs)** or "occupational groups".

Brown and Duguid (2002) identify networks of practice as networks that link people to each others whom they may never get to know but who work on similar practices. This type of networks share the concept of practice with Communities of Practice in that people have practice and knowledge in common. Brown and Duguid distinguish NoPs from CoPs in that the former consists of members from various organisations and are much larger but with less interaction between the members than CoPs that are internally focused, tight-knit groups who work together on the same or similar tasks. Hence, the members know each other, which results in high reciprocity, whereas the participants in NoPs work on a similar domain, but may never meet, do not take action, and produce little (creative) knowledge. CoPs can be subsections of larger networks of practice, where a community's culture and common identity distinguish its members from other communities within a network of others. NoPs can also be subsections of larger networks, often referred to as "network society". Information in the larger network does not travel uniformly throughout the network, it travels according to the topography, that is formed of NoPs and CoPs.

Social Networks: Social networks are a social structure of nodes that represents individuals (or organisations) and the relationships between them within a certain domain. Social networks are very widely studied in sociology, mathematics, and computer science.

As adapted from the work of Yang and Tang (2003), in education, there are three types of networks a student can be involved in:

- **Friendship Networks:** are based on friendship relations, which can emerge between two people only if and when their paths cross, such that they have to meet before they can mate. People are more likely to meet if they share, for instance, the same living, school, or work environment, or if their social networks overlap. The authors claim that increased visibility and exposure increase the likelihood of becoming friends. Therefore, a student who is central in a friendship network has more opportunities to access resources that may be important to successful academic performance. Thus, this centrality measure was found to be positively related to the students' performance.

- **Advice Networks:** consist of relations through which individuals share resources such as information, assistance, and guidance that are related to the completion of their work. The advice network is more instrumental-oriented than the friendship network, which is more social-oriented. Advice networks were also found to be positively related to the student academic performance.

- **Adversarial Networks:** adversarial relations refer to relations that may involve negative exchanges. This kind of relations causes emotional distress, anger, or indifference. Adversarial relations have been demonstrated empirically to be detrimental to student performance and satisfaction, and therefore, are negatively related to academic performance.

Cho et al. (2007) investigated the relationships between communication styles , social networks, and learning performance in a CSCL community. Their results showed that both individual and structural factors developed collaborative learning social networks. In particular, learners who possessed high willingness to communicate (WTC as discussed in section 2.2.5) or occupied initially peripheral network positions were more likely to explore new network linkages. They also found that the resultant social network properties significantly influences learner's performance to the extent that central actors in the emergent collaborative social network tended to get higher final grades. Based on their findings, they suggested that communication and social networks should be central elements in a distributed learning environment.

In science, social networks are usually presented by a network, which is usually represented as a graph. The participants in the network are usually referred to as *vertices, nodes, individuals, agents, or players*. The links between the nodes are usually called

edges, ties, and sometimes *arcs* when the network is directed. Directed networks have their edges defined by a source node where the edge starts and a sink node, where the edge goes to. They are useful is modelling non-mutual relationships such as trust, where A trusting B, does not necessarily mean that B trusts A. The study of networks involves identifying and measuring the properties and the topology of the network based on the study of the network graph. Examples of these properties are shortest path, the diameter, components, and clustering (Jackson, 2008). Definitions of these elements are included in future chapters.

In his popular work on the science of networks, Barabási (1999; 2002) introduced scale-free networks and a description of their topology and evolution. Social networks, just like the Internet, the WWW, biology, and economy networks, follow this model, where the distribution of the number of vertices to the number of links follows the power law. This distribution shows a number of members (vertices) having a large amount of links, and the rest of the vertices having a relatively smaller links. Barabaśi called these highly linked vertices "*hubs*". Due to this interesting distribution of power in the network, social networks that follow a scale-free topology are studied with different properties as opposed to random networks.

In addition to the distribution of links, another interesting property of networks is the weight of the links. Weighted social networks are usually built based on the intensity level of the relationship, such as similarity, degree of knowing, collaborating, or trust between the members (nodes of the network) (Jackson, 2008). This type of networks is very useful in modelling real life relationships and inferring the intensity of these relationships, such as the similarity between 2 people who both share interests with another person, but do not connect to each other directly. This is can particularly useful in recommendation systems.

The study of weighted networks hasn't been a popular topic in general. In his paper, Newman (2004a) shows that weighted networks can be easily studied by mapping them to an unweighted network with multi-edges. From a social aspect though, the way in which these nodes are connected has been an interesting topic of research, where different types of ties between them have been identified. In (Alani et al., 2003) and (Goecks and Mynatt, 2004), the authors discuss the notion of describing connections in terms of *strong and weak ties* or *formal and informal relationships* respectively. In

this context, a strong tie is a one established directly between two people in the same network, whereas a weak tie is a relationship between two people connected through another person (by two levels of separation). An example of studying social networks, are Newman's analysis of the scientific co-authorship networks (Newman, 2004b), which investigates many properties such as number of authors per paper, number of papers per author, shortest paths, centrality and betweenness, weighted networks, and other properties that reveal a significant amount of information about the network and the participants.

Virtual networks, also termed Web-based networks, are the most popular type of virtual groups. With various technologies developed to facilitate modelling networks on the Web, the growth of the number of websites that connects networks of people is rapidly growing to the point that raised an abbreviation mocking their prevalence ASN (Yet Another Social Network) (Alexander, 2006).

Examples of online social networks are: Facebook[1], mySpace[2], and LinkedIn[3]. The word of mouth concept behind social networking sites and communities within these networks has been taken into great consideration by marketing services. Social software technologies such as twitter are now used as an official way to advertisement. This is due to the fact that social conscience economists acknowledge that recommendation is stronger within a network of friends or colleagues who trust each other, as these networks motivate persuasion. This social component is now a part of many companies' business model, and is now expanding to education.

Ning[4] for example is an online platform for people to create their own social networks, launched in October 2005. Ning competes with large social sites like MySpace and Facebook by appealing to people who want to create their own social networks around specific interests with their own visual design, choice of features and member data. The unique feature of Ning is that anyone can create their own social network for a particular topic or need, catering to specific membership bases. For education, Ning has lunched a network names "Classroom 2.0"[5] to encourage discussions between those interested in Web 2.0 and collaborative technologies applications for education.

[1] http://www.facebook.com/
[2] http://www.myspace.com/
[3] http://www.linkedin.com/
[4] http://www.ning.com/
[5] http://www.classroom20.com/

In education and sociology, social networks differ from communities of practice in that the latter are defined by Barab et al. (2002), as "persistent sustained social networks of individuals". In other words, a more connected and maintained form of a social network. This is mainly due to the fact that people join communities because they share something, such as a topic of interest, with everyone in the community, making it a more specified network where members tend to stay in the network for as long as they are still connected to that specific element (topic), that made the community come together. There are cases where a community grows large in size due to its evolution, and ends up being split into smaller communities. This topic however, is out of the scope of this thesis. Table 3.1 shows the difference between the group types according to Wenger and Snyder (2000).

	CoPs	Project Teams	Informal Networks
Powers	Develop capability, build and exchange knowledge	Accomplish a specific task	Disseminate information
Boundary	Knowledge domain	Assigned charter	Scope of relationships
Motivation	commitment, identification with group's expertise	Project goals and requirements	Mutual needs
Time Scale	As long as there is interest in maintaining the group	Temporary	As long as people have a reason to connect

TABLE 3.1: Comparison of various grouping (Wenger and Snyder, 2000)

Back to the topology of these groupings, according to Girvan and Newman (2002), well known physics scientists in the domain of networks structure and topology, communities are a common part of the social network, in which the topology shows dense clusters connected by few edges. The clusters are defined as communities within the network. Just as small world, power law, and clustering coefficient, community structure is a property of studying social networks. Newman's and Girvan's work in this area will be described in more details in the coming chapters.

3.4 Group formation process

In learning, when the need for a collaborative activity is defined, the type of the group that best suits the aim of the collaboration is determined (by the instructor or the

learner). A specific group formation approach (as described in section 3.5) is then chosen to carry out the formation process, which takes place by going through three stages regardless of what formation approach is selected (Wessner and Pfister, 2001):

1. Initiating the formation process: Here the initiator starts the formation of the chosen group type. The initiator can be the instructor, the learner, or a system representing the instructor or the learner.

2. Identifying group members: This is where the formation initiator chooses who should join which group. This is usually done based on the learners profiles and the requirements for joining the groups.

3. Negotiating the formation: in this stage, the initiator has to ensure the formation satisfies members of the group(s) in addition to the criteria (constraints) of the initiator, and hence the collaboration.

For all types of groups, in stage (1) and (2) of group formation, the initiator has to consider two problems:

1. Modeling: In stage (2), the requirements needed to identify the members of each group serve as parameters for the formation. In this context, the initiator needs to identify what parameters need to be modeled for profiling the learners and processing the formation.

2. Satisfying the criteria: In stage (3), forming the groups with intention to maximise the benefits for each student within the group is not an easy task. When the formation aims to construct balanced groups in terms of the formation parameters, this approach may conflict with the best interests of individual students. These factors create the complexity of the group formation in terms of violating the criteria set for the group composition.

In the next section, we discuss this complexity in relation to the different group formation approaches.

3.5 Group formation approaches

There are three different approaches to constructing groups. In this section, we discuss the use of these approaches to generating the different types of groups mentioned in section 3 of this chapter. We also discuss how these approaches execute stages (1), (2), and (3) of the group formation process in relation to the factors of the formation complexity discussed in the previous section.

3.5.1 Randomly selected groups

Where the formation is initiated by the instructor who assigns students to groups at random, this approach is usually used for forming temporary informal groups (mainly teams). Without any need for negotiations, and as there are no attributes or constraints for the selection of group members; this is the simplest way to form groups.

3.5.2 Self-selecting groups

In this type, the formation is initiated by learners who are allowed to choose which group they want to belong to, and can negotiate with whom they want to work with. Here, the assignment of participants involves identifying potential peers who fullfill the requirements to join the learners group. This formation is extensively used in identifying communities and networks where the members get together for a common interest. This approach can be used in teams as well where students choose their partners based on interests, preferences, similarities, friendship, and trust; or form with the attention to the need of recruiting members with technical skills, knowledge, expertise and ability that can fulfill the task (Owens et al., 1998). Hence, unless the group is formed with attention to educational diversity, the groups generated from this formation tend to be homogeneous. The effectiveness of this formation lies on the efficiency of the negotiations with the identified peers to join the learners group.

Adamic et al. (2003), researchers at HP Labs and Google, studied the users of an online student centre at Stanford called “club Nexus” and found that two students were likely to be friends if their interests overlapped, and that the likelihood rose if the shared interests were more specific (Two people who like fencing are likelier to be friends that

two people who like football). Shirky (2008) argues that the net effect is that "it's easier to like people who are odd in the same ways you are odd, but it's harder to find them". This fact can be used in recommendation-based systems such as the ones used in dating and match making websites.

3.5.3 Instructor-selected groups

This approach is often referred to as "criterion-based selection" (Oxford, 1997). Group formation in this type is initiated by the instructors. Although teachers can create or direct/motivate the creation of students communities and social networks by considering the social ties and shared interests among students, this approach is most popular in task-oriented teams and intensional networks, due to their cooperative nature as opposed to communities that enjoy a collaborative nature.

Table 3.2 shows the support of the different group formation approaches for building different types of groups. The shaded cells highlight the best technique to form each type. From the table, we observe that both communities of practice and social networks are better formed using self-selecting approach due to their self-organised nature. Hence, for the formation to be effective for collaborative learning, the instructor has to provide a degree of self-organisation within these groups. In contrast, teams and intensional networks perform better when they are controlled by the instructor.

Groups \ Approach	Random	Self-selected	Instructor-selected
Teams	partial	partial	full
Communities of Practice	none	full	none
Intensional Networks	none	partial	full
Social Networks	none	full	none

TABLE 3.2: Group formation techniques' support for building the different types of groups

In instructor-selected grouping, the formation process depends on the instructors aim from the assigned collaborative task. The assignment of students to groups will involve either:

(a) The simultaneous distribution of all the students in the class over n groups, where $n = \frac{\textit{The Number of Students in the Class}}{\textit{The Optimal Group Size}}$; Or

(b) Choosing few students from the entire class to form one group, such as a class committee

The last case (b) happens when the collaboration is only needed for a number of students in the class such as using sample students from the whole class population or selecting top students for a specific challenge. In this case, forming the group relies on identifying the potential participants by querying on the list of all students without need for negotiation. For the instructor, though it might not seem so obvious, these two types of forming groups are different, and this difference is reflected in the types of algorithms that were developed accordingly to approach the different problems. These algorithms, not specifically developed for learning, will be discussed in the following chapters.

The first case (a) is concerned with distributing students evenly to construct balanced groups in terms of the formation criteria, while considering the students maximum benefits from participating in the groups, in order to ensure active involvement of all students simultaneously, as well as fairly even performances from all the groups. In this context, the formation becomes more challenging than just negotiating with peers or choosing a group of students from the whole class. Hence, this approach is regarded as the most complex formation. As the number of constraints grow, reaching agreements for the formation becomes even more difficult, especially in heterogeneous grouping.

For example, a teacher wanting to group students for software engineering projects may base the formation on their learning styles and preferences in working under a specific supervisor, while making sure that: they are distributed evenly across the groups in terms of their grades and ethnic origin; each group has a balanced combination of team roles such as a leader and an implementer; and that no female can be grouped with a group of otherwise all-male participants. The group formation is therefore not an easy task to do manually, especially if the number of students involved in the collaboration, or the number of constraints, is large. In the next chapter we review existing research on automating group formation to facilitate its process for both instructor and learner. Table 3.3 summarises the stages of processing group formation for the different approaches of grouping students.

Regardless of the approach, we can also distinguish the ways in which group formation criteria is applied to forming groups as:

<table>
<tr><th></th><th>(1) Initiator</th><th>(2) Identifying members</th><th>(3) Negotiating the formation</th></tr>
<tr><td>Random</td><td>instructor</td><td>random</td><td>none</td></tr>
<tr><td>Self-selecting</td><td>learner</td><td>identify potential peers</td><td>negotiate with the identified peers to join the learners group</td></tr>
<tr><td rowspan="4">Instructor-selected</td><td rowspan="4">instructor</td><td colspan="2">From one group:</td></tr>
<tr><td>identify potential learners</td><td>query on potential learners - no negotiation needed</td></tr>
<tr><td colspan="2">Group all students:</td></tr>
<tr><td>distribute students over groups</td><td>ensure fairness of formation + maximising every students benefit</td></tr>
</table>

TABLE 3.3: Formation process in different group formation approaches

- Homogeneous: the members of the group are similar in relation to the criteria.
- Heterogeneous: the members of the group are different in relation to the criteria.
- Rule based: the criteria is to apply a specific rule such as do not allocate exactly one female to an all-male group.

According to cooperative-learning research, structured forms of teacher-assigned heterogeneous grouping -in the instructor-based approach- can enhance relations among classmates, promote learner-to-learner tutoring, increase tolerance, decrease prejudice, and promote cross-cultural understanding (Salvin and Oickle, 1981; Kagan, 1985), although such grouping involves increased thought, effort, and energy on the part of the teacher. Heterogeneous grouping can be done on the basis of proficiency, background, ethnicity, gender, or other factors. Michailidou and Economides (2007) studied the impact of diversity in learner-learner interactions in collaborative virtual teams through a social cultural perspective. Social differences include gender, race, class, or age. Cultural differences refer to matters like how an individual's cognition, values, beliefs and study behaviours are influenced by culture. The authors found that social and cultural differences influence an individual's performance in a learning environment.

Olsen and Kagan (1992) argue that random grouping or interest-based grouping -can be self-selecting- can provide a perception of fairness, although it can also create possible incompatibilities and "loser teams". Homogeneous grouping, for example according to proficiency or other factors, can ease classroom management but can create group labelling problems and inhibit learner-to-learner tutoring opportunities (Oxford, 1997).

From a personalisation perspective (Harrigan et al., 2009), it is often arguable that there is a potential conflict between a learner's preferred learning style and an optimal learning strategy that the teacher might see best fit the learning situation. In this situations, it is difficult to please the learner and do what is best for them from a pedagogical standpoint.

3.6 Discussion and Summary

In learning, most systems use the term group to refer to a team. In this research we want to include the need and study of communities and social networks of learners as well. The significance of providing networks and communities of learners relies on their impact on the students learning experience, and the different type of group work they can motivate, in terms of collaborative versus cooperative learning. Moreover, we are not just concerned with groups (communities and networks) on the web, the formation of the group takes place on the web, but the group dynamics may take place offline as well; and since this research does not involve developing computer supported applications for students interaction within the group, the formation of the groups here does not depend on the degree of virtuality. Therefore, we do not assume that the members' communication will be through a virtual community, but we realise that the communication can be face to face as well. Consequently, we aim at providing a solution to allocating students to groups whether they are distributed geographically or not.

As discussed in the previous chapter, computer-mediated collaborative learning provides an opportunity for enhancing the educational process through the application of information technology. It is important, however, to point out that computer-mediated collaborative learning represents a departure from the traditional instructional method. Therefore, adequate time and effort by the instructor should be allocated for successful implementation of this new teaching method. It is believed that the positive educational outcomes of computer-mediated collaborative learning make for high returns on investment of instructor time and effort. As will be discussed in the coming chapters, this is the case for allowing the instructor to form groups of students using automated or semi-automated group formation methods.

Chapter 4

Computer-Supported Group Formation

Computer-Supported Group Formation (CSGF) is the use of computer programs and algorithms to facilitate and simulate (often visualising) the allocations of people to groups. This involves developing applications that usually take a dataset of people and make decisions on which group they should belong to based on the patterns and characteristics of the dataset, and the algorithm used to process the dataset into groups. Given the variety of objectives for creating groups, many existing algorithms can be used to support a CSGF application. The choice of which algorithm should be implemented is highly related to the data and the type of groups to be produced.

4.1 Group formation algorithms

Due to the different applications of collaboration and hence group formation, there are several algorithms that were developed for different scenarios and domains of group formation. Examples of these domains are: distributed processors communication and computer networks, economics, multiplayer games, and dating services. The algorithms range from simple greedy algorithms that distribute participants across the stable groups

based on some attribute, to mathematical models of multiagent systems (overlapping or non-overlapping formation) where group membership is dynamic.

4.1.1 Optimisation algorithms

Optimisation is an approach to solving problems (usually problems that have many solutions) by selecting a solution that minimises the output of the cost function. The cost function (also referred to as the *utility function*) is any function that takes a guess at a solution and return a value that is higher for worse solutions and lower for better solutions. Optimisation algorithms use this function to set solutions and to search possible solutions for the best one, also referred to as the optimal solution. The cost function might have many variables to consider, and it's not always clear which is the best variable to change in order to improve the result (Segaran, 2007). The challenge of solving problems using this approach is representing the problem and deciding on the cost function. Once this has been done, the algorithm can be reused for other problems and datasets that can have similar structure to that problem.

Greedy algorithms: aim to find the optimal solution to a problem, but are short-sighted in their approach in that they take decisions on the basis of information at hand without taking into consideration the effect these decisions may have in the future. Greedy algorithms are easy to invent, easy to implement and most of the time quite efficient. Many problems cannot be solved correctly by greedy approach (Korte and Lovsz, 1981).

Hill climbing: Similar to all optimisation algorithms, Hill climbing can be used to solve problems that have many solutions, some of which are better than others. It starts with a random (potentially poor) solution, and iteratively makes small changes to the solution, each time improving it a little. When the algorithm cannot see any improvement anymore, it terminates. Ideally, at that point the current solution is close to optimal, but it is not guaranteed that hill climbing will ever come close to the optimal solution (Koza et al., 2003).

Simulated anealing: inspired by alloy cooling in physics, this algorithm starts with a random guess at a solution, and tries to improve it by determining the cost for similar solutions that are a small distance away and in a random direction for the solution in

question. If the cost is lower, this becomes the new solution; if it's higher, it becomes the new solution with a certain probability depending on the *temperature* (the variable in question). This variable starts at high and decreases slowly, so that the algorithm is more likely to accept worse solutions at the beginning, in order to prevent getting stuck in a local minimum (Haupt and Haupt, 2004).

Genetic algorithms: inspired by evolutionary theory, this family of algorithms starts with several random solutions called the *population*. The strongest members of the population (those with lowest cost) are chosen and modified either through slight changes (mutation) or through trait combination, called crossover or breeding. This creates a new population, known as the next generation, and over successive generations, the solutions improve. The process stops when a certain threshold is reached, when the population has not improved over several generations, or when a maximum number of generations has been reached. The algorithm returns the best solution that has been found in any generation (Goldberg and Holland, 1988).

4.1.2 Constraint Satisfaction Problems (CSP)

Also related to the optimisation family, many problems in computer science and particularly Artificial Intelligence can be represented as a constraint satisfaction problem. Examples of these problems are machine vision, scheduling, graph problems, floor planning design, planning for genetic experiments, and satisfiability problems. Solving these problems differ from using constraint propagation to backtracking to search for possible solutions (Kumar, 1992).

A Constraint Satisfaction Problem (referred to as CSP) is defined by a set of variables, $X_1, X_2, \ldots, X_n$, and a set of constraints $C_1, C_2, \ldots, C_m$. Each variable has a non-empty domain D_i of possible values (Nadel, 1989), (Frost and Frost, 1997), (Russell and Norvig, 2002).

Each constraint C_i involves some subset of the variables and specifies the allowable combinations of values for that subset. A state of the problem is defined by an assignment of values to some or all of the variables, $X_i = v_i, X_j = v_j, \ldots$. An assignment that does not violate any constraints is called a consistent or legal assignment. A complete assignment is one in which every variable is mentioned, and a solution to a CSP is a

complete assignment that satisfies all the constraints. Some CSPs also require a solution that maximises an objective function. A CSP can be presented as a graph where variables are nodes and constraints are links called *arcs* between the variables.

Depending on the domain size, If the maximum domain size of any variable in a CSP is d, then the number of possible complete assignments is $O(d^n)$, that is, exponential in the number of variables. Finite-domain CSPs include Boolean CSPs, whose variables can be either true or false. Boolean CSPs include as special cases some NP-complete problems, such as 3SAT, which will be described in the next section.

4.1.2.1 Constraint satisfaction algorithms and solvers

Backtracking: commonly used to solving CSPs, the backtracking search is used for a depth-first search that choses values for one variable at a time and backtracks when a variable has no legal values left to assign. In other words, the variables are instantiated sequentially, and as soon as all the variables relevant to a constraint are instantiated, the validity of the constraint is checked. If a partial instantiation violates any of the constraints, backtracking is performed to the most recently instantiated variable that still has alternatives available. Due to its run-time complexity (for most non-trivial problems is still exponential), various domain-specific heuristic functions were developed to enhance the performance of the algorithm (Russell and Norvig, 2002), (Kumar, 1992).

Variable and value ordering: Given that backtracking simply selects the next unassigned variable in the order given by the list of variables, this approach improves the performance of solving the CSP by ordering the variables to be assigned, in other words, finding which variable should be assigned next, and in what order should its values be tried. Ordering variable and values usually uses a heuristic called *minimum remaining values (MRV)*, where the most constrained variable (the one with the fewest legal values that would satisfy the constraint) will be assigned first (Bacchus and Grove, 1999), (Frost and Dechter, 1995), (Sadeh and Fox, 1996)

Forward checking: In order to reduce the search space, whenever a variable X is assigned, this process looks at each unassigned variable Y that is connected to X, and deletes from Y's domain the values that are inconsistent with the value chosen for X. Using forward checking eliminates branches of the search tree reducing the search

significantly, particularly when used with the MRV heuristic (Bacchus and Grove, 1999).

Constraint propagation: A more powerful technique than forward checking, this approach propagates the implications of a constraint on one variable onto other variables. One method of constraint propagation is *arc*[1] *consistency* (Nadel, 1989). An arc from X to Y is consistent if for every value of X, there is a value of Y that is consistent with that value for X. If a value x in X doesn't have any consistent value in Y, then the arc between X and Y can be made consistent by deleting x. By checking all arcs for consistency, and therefore deleting values causing inconsistency, the search space gets reduced (Russell and Norvig, 2002) (Kumar, 1992) Correia and Barahona (2004).

Backjumping: Since backtracking goes back to the last instantiated variable, a more intelligent approach to backtracking is to go all the way back to one of the set of variables that caused the failure. This set is called *the conflict set* of a variable X and is composed of the previously assigned variables that are connected to X by constraints. The backjumping method backtracks to the most recent variable in the conflict set (Dechter and Frost, 2002).

Solvers: CSP solvers are usually associated with the concept of satisfiability. Satisfiability is the problem of determining if the variables of a given boolean formula can be assigned in such a way as to make the formula evaluate to *TRUE*. Equally important is to determine whether no such assignments exist, which would imply that the function expressed by the formula is identically *FALSE* for all possible variable assignments. In this case, the function is said to be *unsatisfiable*; otherwise it is satisfiable. To emphasise the binary nature of this problem, it is frequently referred to as *Boolean or propositional satisfiability*. The shorthand *SAT* is also commonly used to denote it, with the implicit understanding that the function and its variables are all binary-valued. Many CSP problems, such as graph colouring problems, planning problems, and scheduling problems, can be easily encoded into SAT. SAT problems are proven to be NP-complete problems where NP-completeness only refers to the run-time of the worst case instances, but many practical instances can be solved much more quickly. SAT solvers are usually implemented using backtracking search such as the most used DPLL algorithm. However, modern solvers have been implemented using backjumping and look-ahead algorithms (Zhang and Malik, 2002).

[1] an arc refers to a directed arc in the constraint graph

Another type of solving CSPs is *Answer set programming (ASP)*, a form of declarative programming oriented towards difficult (primarily NP-hard) search problems. ASP is based on the stable model (answer set) semantics of logic programming. In ASP, search problems are reduced to computing stable models[2], and answer set solvers (programs for generating stable models) are used to perform search. The computational process employed in the design of many answer set solvers is an enhancement of the DPLL algorithm and, in principle, it always terminates (unlike Prolog query evaluation, which may lead to an infinite loop). ASP includes all applications of answer sets to knowledge representation and the use of Prolog-style query evaluation for solving problems arising in these applications (Lifschitz, 2005). Due to its knowledge representation quality, Answer set programming is currently combined with description logics for implementing Semantic Web applications (Eiter et al., 2008).

A relatively related research is the work of Preece et al. (2006), who introduced a Semantic Web ontology to model CSPs and particularly soft constraints in OWL DL. The work creates a Constraints Interchange Format (CIF) that builds on the existing Semantic Web Rule Language (SWRL). Unfortunately, although very promising, there was not enough evaluation on the ontology for us to use this work.

4.1.3 Clustering algorithms

Clustering is the assignment of objects into groups (called clusters) so that objects from the same cluster are more similar to each other than objects from different clusters. Similarity is usually assessed according to a distance measure or a similarity measure. Selecting a distance (dissimilarity) or a similarity measure is an important step in any clustering, as the distance determines how the similarity of two elements is calculated. This measure influences the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. Common distance and similarity functions are: *the Euclidean distance, the Manhattan distance, Pearson similarity, the Cosine Coefficient similarity and the Tanimoto coefficient similarity* (Segaran, 2007) (Tan et al., 2006).

[2]The stable model, or answer set, is used to define a declarative semantics for logic programs with negation as failure (deriving *not p* from failure to derive *p*). This is one of several standard approaches to the meaning of negation in logic programming, along with program completion and the well-founded semantics (instead of only assigning propositions true or false, it also allows for a value representing ignorance). The stable model semantics is the basis of answer set programming.

Clustering is highly used in detecting and discovering groups and particularly communities that are often a part of a large network. Lately, clustering techniques have been widely used in clustering users interests and tags in social networking sites such as flikr and del.ico.us to find relations between topics and users (Yeung et al., 2009). Clustering in this context reveals patterns that can be used to support recommendation of new links such as topics investigated by similar users, or interests and merchandise that are similar to what the user likes. As detailed in the next section, in learning, clustering is usually used to discover homogeneous groups of students. There is a family of clustering algorithms that are used for different objectives. Three major types of clustering are:

Hierarchical clustering: is a method of cluster analysis which seeks to build a hierarchy of clusters, usually presented in a tree or dendrogram, in which closely related pairs of vertices have lowest common ancestors that are lower in the tree than those of more distantly related pairs. Strategies for hierarchical clustering generally fall into two types:

- Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as the algorithm moves up the hierarchy.
- Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Partitional clustering: as opposed to hierarchal clustering, this approach aims at identifying a number of groups within the network or the dataset where this number is known in advance of running the algorithm, therefore partitioning the network to exactly the given number of groups (Tan et al., 2006). The most used algorithms are K-means and fuzzy C-means.

K-means creates k groups by placing k centroids at random in the space, then iterating to allocate the participants to the groups with the closest centroid. The algorithm recalculates the average distances in each group to alter the centroids and usually stops after a given number of iterations or when there are no different results found. In this approach, sometimes referred to as *hard clustering*, the data is divided into distinct clusters, where each data element belongs to exactly one cluster. These indicate the strength of the association between that data element and a particular cluster (Hamerly and Elkan, 2004).

In fuzzy clustering however, each point has a degree of belonging to clusters, rather than belonging completely to just one cluster. In other words, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

Spectral clustering: these algorithms aim at clustering data using eigenvectors and eigenvalues, such that for N data points, the algorithm defines an NxN matrix M where the value of M_{ij} is the distance from point i to point j. The algorithm then takes the second eigenvector of this matrix and partitions the data points according to whether the corresponding entry in the second eigenvector is positive or negative, creating a partition of the entries into two sets. The algorithm can be used recursively to subdivide these partitions into more partitions (Bach and Jordan, 2004).

Network clustering: In (2004), Newman and Girvan introduced a new algorithm for network clustering. The discovery of community structure in this context is closer to the divisive approach. The author's algorithm iteratively removes edges from the network breaking it into communities. The edges removed are identified by using one of the set of edge *betweenness measure*[3] such as the shortest path betweenness. This algorithm proved more efficient than traditional divisive clustering because the betweenness of edges is recalculated every time the edge with the highest betweenness is removed. The algorithm functions well with around 10000 vertices and with a complexity of $O(n^3)$ on sparse networks. In later years, Newman proved that this algorithm performs well with weighted networks as well by mapping the weighted network to a mutli-edge network (Newman, 2004a). The highlight of this approach (Hierarchal clustering and community discovery) is finding groups without a restriction of knowledge of the number of groups within the network or dataset.

Mark Newman introduced few measures and algorithms for discovering communities within networks of large sizes. Later, in 2006, Newman revised the concept of modularity, the measure of the strength of a community structure, and introduced a new algorithm that takes the strength of optimising the modularity from similar approach used in graph partitioning that uses eigenvectors. Just as hierarchal clustering, this algorithm was directed to discovering communities without knowing the number of the resulting

[3]Edge betweenness measure is a measure that calculates the centrality of an edge being between vertices

groups in advance. The recorded complexity of this spectral algorithm was in $O(n^2 \log n)$, and takes around 27000 vertices, processing the algorithm in around 20 to 30 minutes on a standard personal computer (Newman, 2006).

4.1.4 Classication algorithms

Classification algorithms are usually used to determine or predict wether a point in the dataset belongs to a class or not, therefore classifying it to one of two classes. The classification can either be in made in a linear space or a multi-dimentional space. A popular algorithm for classification are the *Support Vector Machines (SVMs)* (Gunn, 1998) (Bennett and Campbell, 2003). SVMs are the most efficient approach to creating classifiers to solve matching (classifying) problems such as determining if two people are a good match in a dating website based on their interests, age, location, and so on. Like other classifying algorithms SVMs process the problem by drawing a line between the two classes. However, SVMs excel in performance by finding the line that is as far away as possible from the two classes, by identifying the points at the margin of each class and using only these points, called *supporting vectors*, to draw the classifying line. Deciding where a point falls in space, as in one of the categories, is determined by which side of the line it is. The main idea behind using classification algorithms in people's datasets is predicting the likelihood that two people will make a good match or not; the fact that it is designed for two classes makes it an unfit application for forming more than two groups. However, this set of algorithms is highly useful in recommending partners, friends, or collaborators, and many websites with large datasets apply them.

4.1.5 Example applications

Most of the existing research on formation algorithms focuses on group formation in artificial intelligence, which includes the concepts of team formation and coalition formation (Cohen et al., 1997), practically in mathematical models of multi-agent systems (research on self interested agents).

Ambroszkiewicz et al. (1998) introduced algorithms for dynamic non overlapping team formation by self-interested mobile agents in distributed environment where the agents are viewed as reactive agents augmented with knowledge, goal, and cooperation

mechanism. Their research focuses on the communication and consistent decision making inside the team. The algorithms involve the expansion of the team from a single agent, and then shrinking the team once it reaches inconsistency. Also on self interested agents, de Cote et al. (2006) investigated the cooperation of these agents within the economy dynamics domain, where each agent needs to maximise its own welfare while cooperating for a common goal. The proposed research aims at using game strategy where agents (players) can learn interactively from their opponents while accelerating each other in the learning to form a market equilibrium that maximises the system utility.

As for networks and communities, most algorithms for forming and searching them are based on graph theory. In (Yu and Singh, 2003), the authors studied strategies using agents that search dynamic social networks, modelled as referral graphs, where the agents act autonomously based on local knowledge.

In learning, clustering algorithms are usually applied to allocate students to heterogeneous groups by clustering the students into homogeneous groups, and then allocate individuals from the homogeneous cluster in different groups (heterogeneous clusters) (Myller et al., 2002). Another family of algorithms that has been applied in learning is greedy algorithms where students who stand out with respect to a constraint are allocated first to the groups, and then students with less profile -with respect to the constraint- are added to the group using a loop until all students have been allocated (Redmond, 2001). Since the stable marriage algorithm is reliable in one to one matching, it is usually applied in learning to match students with their preferred topic for a project, or preferred possible supervisor.

4.2 Computer-Supported Group Formation for Education

In this section, we discuss a number of applications that were developed to automate the process of forming groups of students. We discuss the approach they follow, the criteria they use to allocate the students, and their method for evaluating their results.

4.2.1 Existing applications

In 1995, Hoppe introduced an intelligent tutoring system that allows the learners to initiate a group formation when they have a problem (a learner-helper group). Based on the learners models, the system displays a list of all potential peer learners that can help; the learner then selects a helper from the list, and the latter can accept or reject the invitation to help the learner. Parameters here are based on learning experience and competency criteria in the subject of the collaboration. In Mühlenbrock's work (2005; 2006), context information such as the learners geographical location from PCs, Phones, and PDAs were added to the model. Unfortunately, no evaluation of the application was provided by the authors.

A team from Osaka University in Japan (Ikeda et al., 1997) and (Inaba et al., 2000) introduced Opportunistic Group Formation (OGF), where an intelligent system detects the appropriate situation to start a collaborative learning session and sets up a learning goal for the learner. The system takes into account the modelling of learning goals for each learner. Based on individual goals as well as the whole group, the system negotiates with the agents of all the learners in order to come to an agreement and to form a learning group so that each member of the group can obtain some educational benefit. Unfortunately, there is no literature on the architecture of the developed systems or their evaluation.

In similar research (Soh et al., 2006), (Zhang et al., 2005), the authors introduce a multi-agent intelligent system called I-MINDS where the instructor, each student and each group is represented by an intelligent agent. The student agent profiles the student and finds compatible students to form the students body group. The agents communicate, and form coalitions dynamically. For the group formation, in an auction, each student agent bids to join its favourite group based on their previous performance in group activities. Therefore, the formation is constrained by the learners previous performance. The collaboration, formation, and learner profiles updates are all processed in real-time. The students profile in this research is built dynamically based on how active the student is during the real-time collaboration. At the beginning of this research (Soh, 2004), the author intended to provide a group formation based on positive interdependence and hence joint intentions where the students depend on each other for goal satisfaction, rewards, resources, division of labor, roles, and so on. However,

when I-MINDS was introduced (Soh et al., 2006), the authors did not mention how this theory was put into practice, and only mentioned that their application considers the students previous performance as a constraint to the formation. Soh et al. evaluated their I-MINDS system by measuring its effectiveness in terms of its ease of usability by the instructor and the students. The group formation itself was evaluated against the performance of the teams, measured based on the teams outcomes and their responses to a series of questionnaires that evaluates team-based efficacy, peer rating, and individual evaluation.

Also supporting Opportunistic Group Formation systems, in (Wessner and Pfister, 2001), the course author defines at which points in a distributed web based course a collaborative activity should occur. The system then uses knowledge about the collaboration context in real-time such as whether the student has performed this collaboration before, how often, and how fast in order to form appropriate groups. The formation here follows a self-selected approach. Although the authors did not present any results of evaluating their system, they mentioned that the comprehensibility of the group formation algorithms and the satisfaction of learning groups to be a key factor of the overall approach acceptance.

On recommending expert collaborators, Vivacqua and Lieberman (2000) introduces Expert Finder, an agent that automatically generates user models by classifying both novice and expert knowledge of the participants. The agent autonomously analyses documents created in the course of routine work to rank the experts for recommendation to the learner who initiated the expertise search. For evaluation, the authors compare the results (the formed groups) generated from the system to manual generated results of the same participants sample. This technique is frequently used in recommender systems (McDonald and Ackerman, 2000b).

Redmond (2001) introduces a computer program to aid the assignment of students projects groups. This technique is used to form instructor-based group formation for all (part time and distance) learners in the class simultaneously. The students are grouped based on the time slot they prefer to do the group work in, and then allocate the projects to the groups based on the members preferences in the group. The group formation is processed using a greedy algorithm where the program starts with the tightest constraint the student with the fewest time slots rated highly and tries to find a compatible group

for them. This process repeats until all students have been assigned to a group. The formed groups are manually checked for even distribution of grades, and the students who are left unassigned are manually allocated to groups. To measure the efficiency of the program, the author introduced an evaluation formula that calculated the rating of group assignments by subtracting an unassigned penalty representing the program failure in assigned some students from the sum of all formed group overall rating.

Christodoulopoulos and Papanikolaou (2007) presented a web-based group formation tool that supports the instructor in automatically creating homogeneous and heterogeneous groups based on up to three criteria and the learner in negotiating the grouping. The tool employs a clustering algorithm (Fuzzy C-means) for homogeneous grouping while heterogeneous groups are generated using Random Selection algorithm. For each student, the clustering algorithm gives the probability of the student belonging to each group. This helps the instructor to manually adjust the formation since the generated clusters may not be of same size. The probabilities enable swapping students who are unsatisfied with their allocation. The preliminary evaluation of the tool was satisfactory although it was tested on 18 groups with one criterion (constraint on Learning Styles).

Tobar and de Freitas (2007), introduced a rule-based tool that aims at reducing the time teachers spend creating groups for learning. The tool takes into account the students characteristics that are required by the rule (hard constraint that can not be violated). The characteristics available in this tool are taken form the IMS learner specification (Wilson and Jones, 2002). The results returned from the tool can be manually modified by the instructor. Unfortunately, there was no evaluation of the performance of this tool.

Lugano et al. (2004), studied data from self-rated questionnaire together with statistics of the learners real activity in a collaborative learning environment called EDUCOSM. The authors considered the students motivation (learning goal orientation) and social skills (social group roles). The results of analysing self-perception in actual behaviour showed a low correlation of pre-test results with learning outcomes. The authors explained this observation by the high initial expectations being lowered later by factors such as high workload or technical problems with the system.

de Faria et al. (2006), introduced an approach of forming groups for collaborative learning of computer programming. The groups were formed based on the students

programming style generated by a tool implemented to automatically assess the style of the programs submitted by the students. Analysing the students programs assists in finding characteristics that evidence significant differences such as program quality, which would be relevant enough to motivate the students to discuss them.

Graf and Bekele (2006), proposed a mathematical model for building heterogeneous groups based on the students personality traits (group work attitude, interest for the subject, achievement motivation, self confidence, and shyness), their level of performance in the subject, and fluency in the language of instruction, where each of these attributes is ranked on a one to three scale. The authors use the Ant Colony Optimisation algorithm to allocate each student to the most appropriate group that would maximise the diversity of that group while keeping the deviation between the groups minimum. The authors show that their approach is scalable (around 500 students) despite the problem being NP-hard.

Cavanaugh (2004), described Team-Maker, a web-based system that aims at reducing instructors time in allocating students to groups. The system takes some students characteristics such as gender, skills, and students schedules, and the instructors criteria for the creating of homogeneous or heterogeneous groups, and applies a Hill Climbing algorithm to get the optimal solution. The authors show that the system outperforms manual group formation, but do not mention the complexity of the system or how well it performs as the number of constraints (instructors criteria) grows.

Wang et al. (2007) introduced a computer-supported heterogeneous grouping system called DIANA. The system uses a genetic algorithm to form fair groups in terms of heterogeneous grouping such that all groups have the same level of diversity. The system uses the students characteristics (thinking styles) collected from questionnaires. It takes up to 7 variables and allocates 3 to 7 members per group. The authors stated that questionnaires based evaluation of the research on a class of 66 students showed that DIANA performs better than random allocations to groups. Although, the authors did not discuss the complexity or scalability associated with the application of the algorithm.

Table 4.1 shows a summary comparison of the discussed Computer Supported Group Formation applications. We refer to the applications that are not provided with a name with their first author name.

Tools	Formation Features					
	Approach		Principles		Algorithm	Modeled characteristics
	Self-selecting	Instructor Based	Opportunistic	Simultaneous for all students in the class		
Hoppe	yes		yes		Rules & inference	Knowledge in a specific domain
Inaba	yes		yes		Multi-agent System	Learning goal
Soh	yes		yes		Multi-agent System	Performance in previous teamwork
Wessner		yes	yes		Multi-agent System	Knowledge on students state within the designed learning
Vivacqua	yes		yes		Profile Matching	Expertise in a specific domain e.g., Java Programming skills
Redmond		yes		yes	Greedy algorithm	Preferred time slots and Preferred projects
DIANA		yes		yes	Genetic algorithm	Psychological variables (thinking styles) but can take up to 7 variables
Team-Maker		yes		yes	Hill Climbing	Any variable
Graf		yes		yes	Ant Colony Optimisation	Performance and Personality traits
Tobar		yes		yes	Rule based	IMS LIP
Christodoulopoulos		yes		yes	Fuzzy C-Means	Knowledge and learning styles

Table 4.1: Existing CSGF applications in e-learning

4.2.2 Limitations of current applications

From the literature, and in terms of satisfying the criteria set for the group formation, we observe the following (Ounnas et al., 2007b):

Modeling: With regards to the data modelled and used to form the groups, we noticed the following:

- Most systems only model a fixed set of attributes, which do not allow for the formation of different types of groups, and hence the implementation of different collaborative activities (only supports some types of teams). This prevents having multi-dimentional groups, in other words, grouping based on many variables.
- None of the existing efforts discuss the performance of the relative application in handling the group formation when the data about the user is incomplete, for example, if a new student with no record in the university joins the collaboration activity.

Satisfying the criteria: With regard to the restrictions (constraints) that are used to obtain a desirable type of groups, we noticed the following:

- Many systems use Opportunistic Group Formation (Wessner and Pfister, 2001), (Hoppe, 1995), (Soh et al., 2006), (Inaba et al., 2000), which does ensure satisfaction of the participants in the group through negotiation, but does not discuss the efficiency of the negotiation if all students in the class are grouped simultaneously. In addition to that, OGF is usually more beneficial in short-term groups, in that every time the learner establishes the relationship with the group members, they get new collaborators. Furthermore, these systems are based on self-selecting group formation (Hoppe, 1995), (Inaba et al., 2000), (Vivacqua and Lieberman, 2000), (Soh et al., 2006), which is not always the most efficient approach in forming teams for learning, as it does not ensure balanced grouping.
- As observed in using existing group formation tools, another common problem in forming groups is the orphans problem; these are the students who remain unassigned to any group at the end of the formation. In existing tools, such as (Redmond, 2001) and (Tobar and de Freitas, 2007), this problem remains unsolved.

Instead, most tools return the names of the orphans for the instructor to allocate them manually to some group, or rearrange the formation by swapping the orphans with other members, the fact that decreases the efficiency of the automated formation.

- Based on the reported results, most applications can only take a small fixed number of constraints. So far, DIANA seems to handle the highest number of constraints, which is currently limited to 7, and only for homogeneous grouping. We hypothesis that this fact is related to the limitation and complexity of the algorithms implemented in the tools and therefore raises issues of scalability of the systems.

Evaluation: In addition to the lack of providing results on the performance of the applications in some of the literature, a limitation of most group formation applications is the exclusive reliance on the groups performance measures indicators such as members responses to questionnaires or post-tests to draw inference about the group formation system performance. From a learning viewpoint at least, group formation efficiency is clearly a multi-dimensional concept, which implies that multiple efficiency indicators besides perceived performance need to be employed. While different formation constraints might result in different formulas for calculating efficiency, these constraints can be related to group formation efficiency in a more abstract way. If so, consideration of defining this relation together with other group formation related measures is required.

4.3 Other Computer-Supported Group Formation systems

There are many applications outside the learning domain that employ different theories and algorithms of group formation. Many of these applications employ joint attention theories that are usually applied to multiagent team formation or coalition formation. The formation of groups in multiagent systems usually follows a self selecting approach, where the agents decide which group they want to join or quit. Examples of these systems is LOTTO that Legras and Tessier (2003) introduced as an extension to their overhearing-based group formation framework OTTO (Organising Teams Through Overhearing). This framework allows a team to dynamically re-organise itself in accordance to the evolution of the communication possibilities between agents.

Regarding group formation, both social networks and communities of practice can have an overlapping structure; in addition to that, they can both have a self-organised dynamic formation (i.e. constant allowance of formation refinements). In this context, using team formation, coalition, and network formation algorithms (Jackson, 2003) are a very good solution to modeling the dynamics. Researchers from the University of Maryland, introduced the notion of enabling team formation from existing social networks of agents (Gaston et al., 2004) (Gaston and desJardins, 2005), such that if a failure of an agent member takes place within the team, a neighbouring agent in the network can take its place in the team to prevent the task from failing.

In this research however, we are interested in static team formation such as software engineering project teams, where the team members have to collaborate without negotiating joining other teams once the allocation has been announced. In real life teaching, within a university course, this is more likely (frequent) to happen than dynamic team formation. The reason behind this is that most existing dynamic group formation systems are connected with a real-time collaboration environment where the students are asked to carry the collaborative tasks using those environments facilities (including the formation system). These systems are usually aware of the individuals (agents) communications and decision making inside a group. In this research, we aim at providing a less complex solution to group formation that can use minimum data about the distributed users without constraining the users on the way they will communicate within the groups after the formation has taken place.

In this research, we also emphasise that the instructor facilitates the collaboration by allocating students into groups, but does not assign roles or distribute the subtasks among the members within the groups once the formation is done. Role assignment is hence self-organised by the members themselves in support of constructive learning theories. Therefore, the instructor facilitates the smooth run of the dynamics within the group but does not control the dynamics once the collaboration has taken place.

4.4 Generating communities and social networks

The use of social networking research can be very beneficial when it comes to analysing the networks and identifying the appropriate people that each individual in the net-

work share similarities with. The study of communities of practice, which refers to the process of social learning that occurs when a self selecting group of people who have a common interest in some subject collaborate to share ideas or find solutions, is becoming increasingly important (Schwen and Hara, 2003).

As many people need to share and collaborate through the Internet, algorithms of grouping them into communities to facilitate the communication become crucial. In (Hamasaki and Takeda, 2003) the authors present a method to connect two people in a network through another person who knows them both. Their method, which they call The Neighbouring Matchmaker, suggests that each individual in the network would analyse the relationship between the other nodes that are connected to them, and recommends a new tie between these two nodes depending on how suitable they are for each other. Other methods of generating communities of practice rely on Expertise Recommenders (McDonald and Ackerman, 2000a). These methods however, neither ensure a system that is capable of building a social network that keeps track of the people with whom a person has interacted and the information about these interactions, nor do they address security and privacy control over shared information, the problem which raises the issue of trust in the network.

As for social networks, most applications for forming and searching them are based on graph theory and searching algorithms (Jackson, 2003). On searching for expertise in social networks, Zhang and Ackerman (2005) studied the social characteristics of various searching algorithms of finding expertise, in order to understand the tradeoffs involved in the design of social network based searching engines.

4.5 Discussion and summary

In this chapter, we discussed different algorithms and the objectives behind using them in a problem such as forming groups. Although the algorithms have different complexities, it's the type of data, the representation of the problem, and type of groups one want to produce that matters in choosing an algorithm. For example, while optimisation algorithms might find the best solution, their complexity is higher than using a clustering algorithm that can process larger sets of data, but might not handle restrictions on the size of groups in the most efficient way. In the second part of the chapter, we discussed

different applications of Computer Supported Group Formation within and outside the learning domain. We noticed the lack of support to the formation of groups based on controlled (instructor based) approach. We also noticed that most applications can only take few constraints and model few characteristics of the student population. In addition to that, evaluating how good are the formed groups have not been covered by many of the implementations, leaving a gap in determining wether an application is good or not, and in what sense.

This chapter concludes our literature review on forming groups. The coming chapters will focus our implementation of automating the group formation process and evaluating it, but most significantly, on using semantics to improve the results of automating group formation. Therefore, before getting to the main contribution of this work, we review some literature on the Semantic Web, a technology we use in our work discussed on the last chapters.

Chapter 5

The Potential of the Semantic Web in Group Formation

As mentioned in previous chapters, the data used to describe the set of participants to be grouped is a very crucial element to forming groups. Describing the participants with as much data as possible to achieve a good grouping is highly beneficial to the quality of the groups, and therefore the data representation is very significant in this topic. Over the past decade, new technologies associated with knowledge representation, linking data, and inferring new concepts from existing ones offered a new approach to modelling datasets for a large variety of problems that span different fields from biology to personalisation of browsers, but particularly used to enhance Web applications. These technologies form a new layer to the Web called the Semantic Web, that uses logic, mainly description logic (Nardi and Brachman, 2003), to introduce meaning (semantics) to the Web at the machine level.

5.1 What is the Semantic Web?

The Semantic Web (SW) is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation

(Berners-Lee et al., 2001). Aiming at improving the current state of the World Wide Web and allowing it to reach its full potential, the key idea behind the Semantic Web is the use of machine-processable Web information. According to the main vision behind the Semantic Web (Berners-Lee et al., 2001), the key technologies to enable this include:

- Explicit metadata: the Semantic Web relies on machine-processable metadata in addition to the traditional text-based manipulation.
- Ontologies: in the Semantic Web, ontologies are a set of knowledge terms, including the vocabulary, the semantic interconnections (relationships between concepts) and some axioms that define the domain.
- Logic and inferencing: In addition to the structured collections of information in the ontologies, inference rules are used to conduct automated reasoning.
- Agents: are computer programs that work autonomously on behalf of a person. They receive tasks to accomplish, make certain choices and give answers. The Semantic Web promotes the idea that agents, that were not expressly designed to work together, can transfer data among themselves when the data comes with semantics (Hendler, 2001).

The goal of the Semantic Web is to give meaning to resources in order to make them understandable for software agents and other programs and improve interoperability. Resources are therefore associated with metadata. To realise this goal, the World Wide Web Consortium (W3C)[1] proposes a stack of standards, often called the "Semantic Web layer cake". The layers in the semantic cake also represent the proceeding of the Semantic Web development. The most important building blocks of the Semantic Web are ontologies that are used to explicitly represent the meaning of terms in vocabularies and the relationships between these terms (Ding et al., 2005a). In general, ontologies provide a shared and common understanding of a domain that can be communicated between people and heterogeneous and distributed applications systems (Klein et al., 2003).

Ontologies can be expressed in the Web Ontology Language (OWL)(W3C, 2004a) that is built on top of the Resource Description Framework (RDF)(W3C, 2004b). The

[1] http://www.w3.org/

Semantic Web layer cake constantly evolves as related research develops and the defined layers get more technical support. Figure 5.1 shows the latest layer cake at the time of writing.

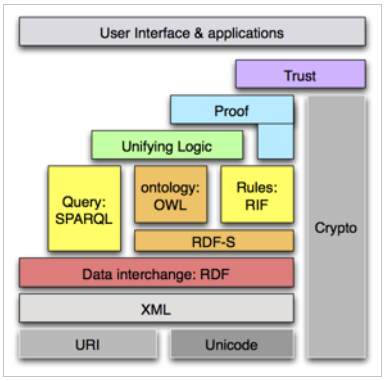


FIGURE 5.1: The Semantic Web Stack. (Berners-Lee et al., 2006)

Since the Semantic Web research is rapidly gaining interest from both researchers and developers, the Semantic Web layer cake takes a new shape every time new standards are developed. In (Shadbolt et al., 2006), the Semantic Web development was revisited to monitor the development of the components of the layer cake, and discuss other topics within the development of the Semantic Web such as rules and inference, and triple stores.

In (Berners-Lee et al., 2006), as part of the "Engineering the Web" research within the Web Science project, the authors discussed the revision of the Semantic Web stack by emphasising the idea of having of Rule Interchange Format (RIF) (W3C, 2010) -which is still under development at the time of writing- alongside OWL as another extension of RDF Schema that is used to model Web ontologies. Next to these layers sits the query language SPARQL (W3C, 2008), another recommendation from W3C. In addition to these modifications, a new layer that represents the need for effective user interfaces and applications sits on top of the stack (Berners-Lee et al., 2006).

As the research advances, more standards becomes available. For example, another recommendation of the W3C is Simple Knowledge Organisation System (SKOS) (W3C, 2009b), a data model aimed for sharing and linking knowledge organisation systems via the Web. SKOS can be used on its own, or in combination with formal knowledge representation languages such as the OWL.

5.2 Why the Semantic Web?

The potential of the Semantic Web as represented is everything that the Web was initiated to do when it was invented. The following quote from a talk given by Tim Berners-Lee at the W3C in 1997, describes the original dream behind the creation of the Web, and thus demonstrates the characteristics of which the Semantic Web involves in order to improve the current Web and lead it to its full potential:

The Original Dream ...

> *"The web was designed to be a universal space of information ... this universality is essential to the web: it loses its power if there are certain types of things you cant link ...* ***And what was the purpose of all this universality? The first goal was to enable people to work together better. Although the use of the web across all scales is essential to the development of the web, the original driving force of the web was collaboration at home and at work. The idea was that by building a hypertext Web, a group of whatever size would force itself to use a common vocabulary, to overcome its misunderstandings,*** *and at any time to have a running model-in the Web- of its plans and reasons ... The Web was designed as an instrument to prevent misunderstandings."*
>
> *Based on the talk at the W3C, London, December 3, 1997 (Fensel et al., 2005)*

From this vision, we realise that the illustration of the power of the Semantic Web starts at the point where people can share knowledge and collaborate free from size and place boundaries, which is what is exactly needed for building service oriented organisations such as e-learning systems where different services from different sources are employed in one framework. This kind of architecture needs to be supported by some level of interoperability. The fact that draw e-learning researchers to the Semantic Web.

In more detail, the Semantic Web enjoys a number of characteristics that distinguish the capabilities of its applications from other technologies:

- Given that the construction of ontologies means classifying terms in hierarchies of classes and specifying the relationships between all these classes, developers produce metadata to describe their resources, which can help finding and implementing these resources by other developers.

- Interoperability and increased connectivity is possible through a commonality of expression. This allows systems to interoperate more easily; even if developers would not fully agree on some terminology or classifications in some ontologies, the systems could still distinguish the agreed upon metadata from the non-agreed upon ones. In addition to that, the systems would realise the difference between the different interpretations of concepts if they are expressed in different ways. Thus, this feature would still be beneficial even if it does not seem practical to some researchers to assume the achievement of commonality of ontologies and concepts (Marshall and Shipman., 2003).

- Vocabularies can be combined and used together: e.g. a description of a book using Dublin Core metadata can be augmented with specifics about the book author using the Friend-of-a-Friend (FOAF) vocabulary. In (Hendler, 2001), James Hendler envisions that instead of a few large, complex, consistent ontologies that great number of users share, a great number of small ontological components consisting largely of pointers to each other. In this vision, Web users will develop these components in much the same way that web content is created.

- Vocabularies can be easily extended: e.g. extensions of the FOAF vocabulary.

- One of the most important features of the Semantic Web is allowing reasoning and inference capabilities where inference rules allow new relationships to be derived from previously declared ones.

- Intelligent search with more granularity and relevance: e.g. a search can be personalised to an individual by making use of their identity and relationship information (Sah and Hall, 2007).

Despite the great potential of implementing Semantic Web technologies, there are few limitations that stand in the way of this vision. For example, providing the vocabulary through ontology engineering provides a rich way of describing data and analysing relationships between concepts. However, engineering ontologies can be a painful task.

It takes valuable time, effort, and experts' input to engineer, evaluate, and map an ontology.

5.3 E-learning in the semantic age

The potential of the Semantic Web in education was obvious to many e-learning researchers. The implementation of the Semantic Web in this field started growing in the beginning of the new millennium, and by 2007, many applications to demonstrate this potential were put in place, and now workshops and conferences are dedicated to this field.

5.3.1 Ontologies for e-learning

As mentioned in the first section of this chapter, ontologies are a powerful mechanism for achieving a common understanding of terms in e-learning, especially in systems with more than one author. In an e-learning environment, Grigoris and van Harmelen (2004) differentiated between three sorts of knowledge, and therefore three types of ontologies: content ontologies, context ontologies, and structure ontologies:

The content ontology, also known as domain ontology, describes the fundamental concepts of the domain where the learning takes place and the relationships between them to allow the inference of new relations. One type of ontology is used in modelling learning domains within specific subjects, and are used in inference, for example, in personalising content delivered to students in adaptive learning systems (O'Keefee et al., 2006), (Stojanovic et al., 2001) .

The second type is the context ontology (also called pedagogy ontology), which addresses the pedagogical issues of learning and describes its processes. For example, learning material can be classified as lectures, tutorial, analysis, discussion, example, exercise and so on. Unfortunately, there aren't any context ontologies known to the e-learning community yet.

The structure ontology is used to define the logical structure of the learning material. Hierarchical and navigational relations such as *hasPart, isPartOf, isBasedOn* and so on are used to define the knowledge of this type of ontology. Ontologies are usually

written in OWL, which comes in three different flavours[2], where each differs in its level of expressiveness and reasoning OWL Lite, OWL DL, OWL Full as described by W3C. Research in OWL is constantly evolving. Recently, researchers introduces OWL 2, an more expressive version of OWL 1 (W3C, 2009a).

5.3.2 The contribution of Semantic Web technology

The key idea of the Semantic Web, namely, shared meaning expressed as machine-processable metadata, establishes a promising approach for satisfying the e-learning requirements discussed in previous chapters. For instance, Semantic Web technologies can support:

- Describing, certifying, annotating, extending and reusing learning material (learning objects). In (Gasevic et al., 2004) the authors proposed an approach to use content structure and domain ontologies to enhance learning objects. In another research by the Edutella P2P network (Brase and Painter, 2004), the developers used learning objects classifies using domain ontologies and annotated with Dublin Core and LOM metadata using RDFS.
- Adaptive learning: enabling teachers to improve learning adaptation and flexibility for single and group users:
  - Personalisation of content: Personalised courses can be designed through semantic querying, and learning materials can be retrieved in the context of actual problems, as decided by the learner.
  - Personalisation of learning design: knowledge can be accessed in any order the learner wishes, according to her interests and needs. Of course, appropriate semantic annotation will still set constraints in cases where prerequisites are necessary. But overall nonlinear access will be supported (Millard et al., 2006) (Markellou et al., 2005).
- Searching learning material: semantic querying and the conceptual navigation of learning materials, locating resources, re-using and sharing resources, and annotation, and so on can be easily achieved when the resources are enriched with

[2]http://www.w3.org/TR/owl-features/

semantics. In (Woukeu et al., 2003), the authors introduced an ontological hypertext framework called Ontoportal to be used in building educational web portals based on domain ontologies, where the educational resources within the portal are semantically interconnected.

- Interoperability and integration: The Semantic Web can provide a uniform platform for the processes of e-learning systems of virtual organisations, and learning activities can be integrated in these processes. This solution may be particularly valuable for virtual universities that involve different educational systems. Interoperability in e-learning systems mostly relies on semantic conceptualisation and ontologies, common standardised communication syntax, and large-scale service-based integration of educational content and functionality provision and usage (Aroyo and Dicheva, 2004).

- Collaborative learning: with the support of Semantic Web technologies to building social networks, communities of practice and semantic wikis, collaborative learning and co-authoring are enabled in an easy and efficient way (Millard et al., 2006).

- Constructivist learning: Based on the support for personalised learning and semantic querying, the implementation of Semantic Web technologies also supports learner-driven e-learning, such that the learner can access information anytime, in any order and up to their preferences and level of understanding.

Millard et al. (2006) describe the pedagogical view of semantics and how semantic enrichment can improve learning and learning management. In addition to some of the points mentioned above, the authors also describe the use of semantics in other areas such as adaptive assessment, feedback agents, recommender agents, and analytical tools for learning and the production of learning material, student management, timetable management, record keeping and quality assurance for learning management.

Tiropanis et al. (2009) surveyed a number of Semantic Web technologies that are currently used or could be used for learning and teaching[3]. With over 35 applications that have potential to change the learning experience, the technologies have been categorised to collaboration, search, repositories, infrastructure and so on.

[3]http://semtech-survey.ecs.soton.ac.uk/technology

5.4 Semantic Web and group formation

Unlike teams, the formation of Communities of Practice and social networks based on the Semantic Web is developing rapidly. However, none of this research has been applied to support the social dimension of learning.

Generating Communities of Practice: Lawrence and m.c. schraefel (2006) argued that existing research on CoPs and Web-Based Social Networks (WBSN) on the Semantic Web does not employ the concept of a virtual community as it was defined by Preece (2000). The authors introduced a new group structure called Internet Based Community Network (IBCN) that fulfills the definition of a community by bridging the gap between the existing concepts of WBSN and a virtual community.

Identifying Social networks: In the Semantic Web, the study of social networks relies generally on describing people, their attributes and the relationships between them using ontologies, and in particular on ranking these relationships to determine the strength or the type of the relationship between the users in the graph. As mentioned before, in (Alani et al., 2003), the algorithm is based on calculating the weight of the paths between two users. A popular research on web-based social networks by Golbeck investigates the construction of social networks for different aims such as movie recommendation, and e-mail spam filters (Golbeck and Hendler, 2004).

The most popular method of identifying social networks and communities of practice in the Semantic Web, as described by Finin et al. (2005), remains the use of people's descriptions with the Friend of a Friend FOAF vocabulary, created by Brickley and Miller (2005). This is reflected in the way that although social network applications such as Friendster[4], Okrut[5], LinkedIn[6], Facebook[7] , and SongBuddy[8] are not based on FOAF, some of them have already begun reading and exporting FOAF files (Alani et al., 2005). The relationships between people in FOAF are described using the *"knows"* attribute; and although FOAF is the second most popular ontology with more than 1.5 millions of FOAF documents generated (Ding et al., 2005a), descriptions using *"knows"* are not sufficient to generate a reliable network, as it does not give any information

[4]http://www.friendster.com/
[5]http://www.orkut.com/
[6]https://www.linkedin.com/
[7]http://facebook.com/
[8]http://www.songbuddy.com/

the type and strength of relationships between the file holder and the people he or she knows. For this reason, a number of ontologies add more specification to FOAF such as the Relationship ontology[9].

Other than being a very popular ontology for describing people and managing communities, there are many key benefits of using FOAF to generate social networks:

- FOAF makes it possible to locate people with similar interests, which is essential to building communities and groups in general. There are a number of ontologies that were developed to extend the FOAF vocabulary in order to model the interests, hobbies, and preferences of a person, such as the Skill ontology[10], The FOAFCorp ontology[11], which extends FOAF to describe in more detail the structure and interconnections of corporate entities (such that a person can present which company or committee he or she works in), and the Description of a Career ontology (DOAC)[12], which describes the professional capabilities of a person as in a (Europass) CV such that employers will be able to find employees that fit their requirements. Information included in a FOAF+DOAC file describes the education, working experience, publications, spoken languages and other skills of a person. Breslin et al. (2005) introduced an ontology called SIOC[13] (Semantically-Interlinked Online Communities) that combines terms from existing vocabularies including FOAF with new vocabulary to describe the relationships between concepts in the realm of online community sites so that online communities can be semantically interlinked.

- FOAF is supported by a number of tools such as FOAF-a-Matic[14], an interface which enables the user to easily create FOAF files; FOAFBot[15], an IRC bot that provides access to a knowledge base created by spidering FOAF files; FOAFNaut[16], a SVG-based FOAF user interface that enables the user to enter a person's email address and explore who they know; FOAFMe![17], which offers a simple editor for the users FOAF file, and provides a way for browsing connected FOAF files;

[9] http://vocab.org/relationship/
[10] http://www.licef.teluq.uquebec.ca/ontology/
[11] http://rdfweb.org/foafcorp/intro.html
[12] http://ramonantonio.net/doac/
[13] http://sioc-project.org/ontology
[14] http://www.ldodds.com/foaf/foaf-a-matic.html
[15] http://usefulinc.com/foaf/foafbot
[16] http://www.foafnaut.org/
[17] http://foafme.opendfki.de/

FOAF Explorer[18], developed to browse the virtual neighbourhoods of friends in much the same way the "regular" web is browsed, presenting the information and assertions in a human-readable format; and FOAFMap, an online service providing geo-location with FOAF and Google Maps mashup, as a mix of both Semantic Web and Web 2.0 technologies (Passant, 2006).

- Another practical significance of FOAF is that by encoding the email address of the person described in the file, FOAF expresses identity by allowing unique user IDs across applications and services without compromising privacy (Ding et al., 2005b).

- FOAF has already been used in a number of social networks projects, whether it is for social/entertainment, business, dating, blogging, photos, religion, media sharing, marketing, research, and so on (Richter et al., 2006). Examples of such networks are: eCademy[19], FilmTrust[20], LiveJournal[21], Tribe[22], and Videntity[23].

- Trust in FOAF: Another benefit in using FOAF is that in addition to the Web of Trust ontology that supports authorship and copyrights, trust has long been an issue when considering sharing information or being a member of a community. In (Kruk, 2001), the author discusses authentication and access control problems in social networks generated from the knows attribute in FOAF, where a percentage friendship evaluation based on reification on the $< foaf : knows >$ statements was presented. The evaluation in his paper is based on finding the closest distance between two people and the highest level of friendship.

  Golbeck et al. (2003), introduced a trust module as an extension to the FOAF ontology that enables the user to rank the level of his or her trust in other people they know from 1 to 10, where the trust is given in general or in specific subjects. The authors investigate the calculations of trust between the users based on the reputation ranking and infer the trust relationships between individuals based on the algorithm introduced in the paper. This method was also used for augmenting email filtering by prioritising mail from trustable colleagues. Using the degree of trust derived from the extended FOAF files, users can prioritise incoming email

[18] http://xml.mfd-consult.dk/foaf/explorer/
[19] http://www.ecademy.com
[20] http://trust.mindswap.org/FilmTrust
[21] http://www.livejournal.com
[22] http://www.tribe.net
[23] http://videntity.org

and thus filter out those with low trust values (Golbeck and Hendler, 2004). Golbeck and Mannes (2006) introduced an algorithm to infer trust relationships in networks with continuous rating systems using provenance information and trust annotations for content filtering and websites personalisation. Applications of the algorithm were applied to the FilmTrust project, which uses trust to compute personalised recommended movie ratings and to order reviews. Golbeck and Hendler (2006), also introduced a binary rating based system was also introduced.

Also using FOAF, Mika (2005) presented a system called FLINK that employs semantic technology for reasoning with personal information extracted from a number of electronic information sources including web pages, emails, publication archives and FOAF profiles. FLINK aims at extracting, aggregating, and visualising online social networks, and have been applied to construct a social network for the Semantic Web community.

5.5 Summary

In this chapter, we have discussed the potential of the Semantic Web in general and in e-learning in particular. So far, Semantic Web technologies have been applied in different domains of e-learning, but mainly in describing recourses, conceptualisation of knowledge, and annotation and navigation of learning material. We also researched the use of Semantic Web technologies in generating communities of practice and social networks and the different issues of the current research such as trust and relationships inferences. The Semantic Web shows great potential in forming groups, in that it is backed up with knowledge representation through ontologies and inference rules that can enrich the dataset to be used to form groups. In the coming chapters, we describe the two models for forming groups and the use of semantics to improve it.

Chapter 6

Constraint-based Group Formation

In chapter 4, we have discussed many algorithms that can be used to represent a group formation problem, we have also discussed the limitations of existing applications in relation to how many goals and how many restrictions or constraints they can handle to produce the desired grouping, and how they evaluate the quality of the produced groups. From analysing these applications, and looking at the different elements one should consider in solving a group formation problem, it came to our attention that forming groups can be represented and solved as a constraint satisfaction problem aiming at finding the optimal group formation. This chapter is dedicated to describing the potential of representing and evaluating group formation in education as a constraint satisfaction problem.

6.1 Group Formation as a CSP

With respect to the group formation process, in instructor-selected groups discussed in section 3 of chapter 2, we propose the following definition of constraint-based group formation to aid analysing the hypothesis.

6.1.1 What do we mean by Group Formation?

In this research, we refer by constrained group formation to the allocation of individuals to groups that satisfies a set of constraints, where a constraint is any attribute or condition of the formation. Constrained web based group formation is therefore an allocation of distributed users to groups using web technologies. For any group formation:

- Based on the desired type of groups, the formation can be either overlapping where an individual can be a member of two groups at one time, or non-overlapping where an individual can be only a member of one group at any given time.
- The formation can be either stable where the membership of an individual does not change or evolve over time, or dynamic where the individual can join and leave the groups dynamically.
- The formation can either map the entire set of individuals to groups, or a subset of the individuals to groups, where the remaining of the individuals are referred to as non-members.
- With respect to the dynamic nature of the individual within the group, whether the formation is stable or dynamic; the individuals position in the group can change, usually from being a new member to the centre of the group by taking some leadership roles.

In this research, we define a team as a non-overlapping group of students, where each student can be a member of only one team at a time. When modelled in a graph, a team is a set of vertices where the relation between the members is the edges connecting them, and whose collective skill set fulfills the skill requirements for a given task.

For networks and communities, we only consider networks with a fixed number of nodes (i.e. the number of students at any given time is a constant), thereby, graphs where new vertices may be added during the time of the formation are out of the scope of this research.

6.1.2 Group Formation model for education

In this research, we are interested in instructor-based like formation. If the instructor models the collaboration goals for the individuals and the groups as a set of requirements (constraints), then, the success of group formation in this context is defined by the satisfaction of the constraints that define these goals. To facilitate the evaluation of group formation, we propose an analytical metrics framework that defines what we mean by formation success. To achieve this, we first make the following assumptions (Ounnas et al., 2007c):

1. The group formation is non-overlapping: each participant in the class should belong to exactly one group,
2. All groups should have the same optimal number of participants (i.e. all groups have a similar size),
3. All formed groups are stable: we do not consider dynamic groups in this study.

In this context, the instructor may have to form balanced groups of students in terms of expected performance, such that no group will have all the top students, while another have weak ones. In other terms, all groups will have an equal chance to perform well and achieve the goals of the collaborative activity, although this may conflict with the best interest of individual students. Therefore, to form the groups, the instructor has to think about modeling the collaboration goals in a way that satisfies both the task of the collaboration that the students have to achieve as a group, in addition to the needs of the students. The following definitions explain the framework for the constraints satisfaction that is discussed in the next chapters.

Definitions:

Constraints: we define a constraint as any parameter, variable, or condition that affects the process of the group formation (i.e. in Computer Supported Group Formation, the variables that influence the systems decision of allocating participants to appropriate groups). We define the finite set of all possible constraints as $C = \{c_1, c_2, c_3 \dots c_Q\}$.

Collaboration Task: we define task t as the task of the collaboration activity that the instructor intends for the students groups to perform. In education, the instructor usually selects a set of collaboration goals $\{\alpha_1, \alpha_2, \dots \alpha_K\}$ that assist in achieving the

task (i.e. helps the collaborative activity to achieve maximum learning gain for the groups and individuals participants). For instance, using our previous example, if the task is a software engineering group project, then example goals can be that all groups are to be balanced in terms of students experience in the field; no female student can be allocated alone in an all-male group; and groups should be multicultural in terms of students nationalities

Collaboration Goal: we define a collaboration goal α to achieve task t as a set of constraints $\alpha = \{(c_1, v_1), (c_2, v_2), \dots (c_L, v_L)\}$ that the instructor chooses to model the requirements for achieving the goal, where each constraint c_j of α is associated with a value $v_j \in \mathbb{R}$ that represents the importance of the constraint c_j in achieving the goal α. We define $A \subseteq (C, \mathbb{R})$ as the finite set of all possible goals. In the example above, the constraints for modeling the goals can be respectively: for each group {average percentage of members experience average $\simeq$ percentage of members experience in the next group}; {number of females $\neq 1$};{number of international students ≥ 1, number of international from the same country $<$ number of participants in the group}. The last goal is presented with two constraints. The constraints can overlap between the goals with different values v for each goal: $(c_1, v_1) \in \alpha_1 \wedge (c_1, v_1\prime) \in \alpha_2$.

Participants: we define the finite set of all individual participants (all students in the class) $P = \{p_1, p_2, p_3 \dots p_M\}$, where $M = |P| > 1$ is the size of the class.

Groups: we define a group g as a set of participants that have at least 2 elements in it (i.e. $|g| > 1$), where each participant $p_i \in g$ is a member of the group. We define the set of all possible groups $G = \{g_1, g_2 \dots g_O\}$ such that $\forall p_i \in P : G = \mathcal{P}(P) - (\{p_i\} \cup \emptyset)$.

Cohort: we define a cohort as the set of pairwise disjoint groups $\{g_1, g_2, g_3 \dots g_N\}$ that include all participants in the class. We define $G_X \subseteq G$ the set of all cohorts that have all their groups of size X (for each element g_i in the set: $|g_i| = X \pm d$ where d is the acceptable deviation for the group size). Therefore, each cohort in G_X has cardinality $N \simeq M \div X$ and $X > 1$.

Formation: We define a relation R from $\mathcal{P}(A)$ to G_X that maps a set of goals to a set of N disjoint groups. This relation can be any algorithm applied to the set of goals. Therefore, for each set of goals, there is more than one possible set of grouping (allocating students to groups) and therefore more than one possible cohort. This is

because although if participants p_i, p_j have similar characteristics in relation to the constraints modeling the goals, then the cohort with p_i in group g_k is not the same as the cohort with p_j in group g_k. We refer to each single grouping of R as a *formation*. We say that a formation is defined by the set of goals that determines the cohort: $form(\alpha_1, \alpha_2, \ldots \alpha_K) = \{g_1, g_2, g_3 \ldots g_N\}$. Figure 6.1 shows a simplified diagram of the relationships between the collaboration goals and the group formation (the values of constraints v are not shown).

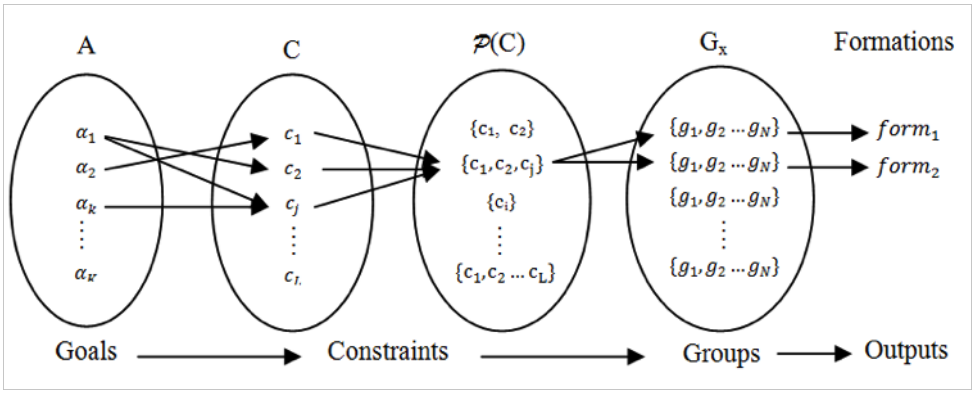


FIGURE 6.1: Example representation of group formation

6.2 Metrics framework for evaluating group formation

We aim at allowing the teachers a degree of freedom to choose the goals and therefore the constraints they want to base the formation on. The success of group formation in this context is defined by the satisfaction of the constraints that define these goals. As mentioned before, each set of goals can generate a set of formations, where each formation is associated with a different constraints satisfaction value. We propose the following framework that aims at maximising the constraints satisfaction by calculating the quality of the generated formations. Again we assume the following:

- Each participant in the class should belong to exactly one group (i.e. non-overlapping group formation)
- All groups should have the same optimal number of participants (i.e. all groups have a similar size), and all formed groups are stable.

A naïve approach: A naïve way to evaluate efficiency is through calculating the ratio or percentage of meeting the constraints:

Well-formed Group: we define a well-formed group, as one that satisfies all the constraints in a formation.

Malformed Group: we define a malformed group, as one that violates any constraint in a formation.

$$\textit{Formation Effectiveness} = \frac{\#\ \textit{well formed groups}}{\#\ \textit{all formed groups}} * 100$$

$$\textit{Formation Failure} = \frac{\#\ \textit{malformed groups}}{\#\ \textit{all formed groups}} * 100$$

However, the reason we cannot use this approach to evaluate formation efficiency is that in education, if we define the problem as above, then, situations where formation efficiency is 70% will look highly positive, although 30% of the groups will be 100% malformed, hence their members will receive no gain from the collaboration at all. A better way to evaluate the formation is through the following non naïve approach:

6.2.1 Formation metrics

1. Constraint Satisfaction Quality

   We refer by constraint satisfaction quality to how well the constraints of a goal α_k were satisfied in the formation of the groups (allocation of students). We use this metric to evaluate the *formation quality* later on.

   - Group Constraint Satisfaction Quality: we use this metric to refer to how well a group g_i is formed in relation to how well the students allocation (to that group) satisfied a constraint c_j. For each group g_i in the formed cohort, and for each c_j in the set of constraint of $\alpha_k (c_j \in \alpha_k)$ we define a function $f_{cg}(g_i, c_j)$ that determines whether g_i satisfies the constraint c_j such that:

   $$f_{cg}(g_i, c_j) = \begin{cases} v_j & \text{if} \quad c_j \text{ is satisfied} \\ 0 & \text{if} \quad c_j \text{ is not satisfied} \end{cases} \tag{6.1}$$

   Where $v_j \in \mathbb{R}$ is a value that represents the importance of the constraint c_j in achieving the goal α_k

   - Cohort Constraint Satisfaction Quality: we use this metric to refer to how well were all the groups formed in terms of satisfying the constraint c_j of

goal α_k. We define the function f_{cG} that calculates the degree to which the formed groups are balanced (i.e. clustered together) in terms of c_j. The Cohort Constraint Satisfaction is therefore defined by:

$$f_{cG} = \sigma \quad \text{and} \quad \overline{f_{cg}} = \frac{1}{N}\sum_{m=1}^{N}(f_{cg}(g_i, c_j)) \tag{6.2}$$

Where the groups standard deviation $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_cg(g_i, c_j) - \overline{f_cg})^2}$

6.2.2 Goal satisfaction metrics

1. Goal satisfaction Quality

   We use this metric to refer to how well the groups were formed in terms of satisfying a goal α_k within the collaboration task t.

   - Group Goal Satisfaction Quality: we use this metric to refer to how well a group g_i is formed, in terms of how well the students allocation (to that group) satisfied the goal α_k. We define Group Goal Satisfaction Quality for goal α_k as a function $f_g(g_i, \alpha_k)$ that calculates the quality of a group g_i in terms of α_k and therefore all constraints of goal $\alpha_k = \{c_1, c_2 \dots c_L\}$ such that:

   $$f_g(g_i, \alpha_k) = f_g(c_1, c_2 \dots c_L) = \frac{1}{L}\sum_{j=1}^{L} f_{cg}(g_i, c_j) \tag{6.3}$$

   - Cohort Goal Satisfaction Quality: we refer by Cohort Goal Satisfaction to how well were all the groups formed in terms of satisfying the collaboration goal α_k and hence the constraints that model it. We define the function f_G that calculates the degree to which the formed groups are balances in relation the goal α_k. The Cohort Goal Satisfaction is therefore defined by:

   $$f_G(\alpha_k) = \sigma_\alpha \quad \text{and} \quad \overline{f_g} = \frac{1}{N}\sum_{i=1}^{N}(f_g(g_i, \alpha_k)) \tag{6.4}$$

   Where $\sigma_\alpha = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_g(g_i, c_j) - \overline{f_g})^2}$, and $\overline{f_g}$ is the mean of the groups goal satisfaction quality $f_g(g_i, \alpha_k)$.

2. Formation Quality

   We refer by formation quality to how well were the groups formed in terms of satisfying all the goals for the collaboration task t.

- Group Formation Quality: This metric evaluates how well was a group formed in terms of all the goals. Similar to the previous calculation of group quality, for each group g_i

$$f_{fg}(g_i, t) = f_{fg}(g_i, \alpha_1, \alpha_2 \dots \alpha_K) = \frac{1}{L} \sum_{j=1}^{L} f_g(g_i, \alpha_j) \tag{6.5}$$

- Cohort Formation Quality: This metric evaluates how well the cohort was formed in terms of all the goals and therefore the task. We define the function $f_{fG}(t) = f_{fg}(g_i, \alpha_1, \alpha_2 \dots \alpha_K)$ that calculates the degree to which the formed groups are balanced in relation the collaboration task t. The Formation Quality is therefore defined by:

$$f_{fG}(t) = \sigma_f \quad \text{and} \quad \overline{f_{fg}} = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} (f_{fg}(g_i, t)) \tag{6.6}$$

  Where the standard deviation $\sigma_f = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (f_{fg}(g_i, t) - \overline{f_{fg}})^2}$ and $\overline{f_{fg}}$ is the mean of the groups formation quality $f_{fg}(g_i, t)$.

To analyse how useful (effective) are the constraints for a given goal, and the goals for a given task, we need to evaluate the formation quality of all possible formations over many runs using the same set of constraints for the goals. For each goal, if the resulted formation quality is constantly high, then if the goal satisfaction quality is high, and the constraint satisfaction quality for that goal is low, we consider that constraint to have a low significance in modeling that goal. Similarly, if the quality of the goal satisfaction is low, but the quality of constraint satisfaction for that goal is high, then the constraint has a low significance in modeling the goal. However, if the formation quality is low, then the constraint significance will be undefined despite the state of the goal satisfaction and the constraint satisfaction. For a large number of evaluated formations, we can evaluate the behaviour (consistency) and therefore the reliability of the constraints and goals, and consequently, the effectiveness of the formation using these constraints in the collaboration.

6.2.3 Optimal formation

We define the optimal formation $form_{opt}(t)$ of the relation R as the optimal cohort that can result from the set of goals, such that the formation quality metrics f_{fG} and $\overline{f_{fG}}$

are maximised. We refer by $f_{optG}(t)$ and $\overline{f_{optG}}$ to objective functions of the quality of $form_{opt}(t)$.

Algorithm 1: Calculating group formation quality

Given task t, $form_{opt}(t)$, $f_{optG}(t)$;
begin
  foreach *formation* $form_n$ **do**
    foreach *group* g_i *in the cohort* **do**
      foreach *goal* $\alpha_k \in t$ **do**
        foreach *constraint* $c_i \in \alpha_i$ **do**
          calculate $f_{cj}(g_i, c_j)$;
        end
        calculate $\underline{f_g}(g_i, \alpha_k)$;
        calculate $\overline{f_{cg}}$;
      end
      calculate $\underline{f_{fg}}(g_i, t) = f_{fg}(g_i, \alpha_1, \alpha_2, \ldots \alpha_K)$;
      calculate $\overline{f_g}$;
    end
    calculate $\underline{f_{fG}}(t) = f_{fG}(\alpha_1, \alpha_2, \ldots \alpha_K)$;
    calculate $\overline{f_{fg}}$;
    if $f_{fG}(t) < f_{optG}(t) \wedge \overline{f_{fg}} > \overline{f_{optG}}$ **then**
      $form_{opt} \longleftarrow form_n$;
    end
  end
  return $form_{opt}$;
end

So far, we assumed that to achieve the collaboration task, a CSGF system would apply the optimal formation to the given set of participants. However, unless the system is appointed to the optimal formation (or uses an optimiser), it will select a formation at random. A possible way to know which formation is optimal is for the system to search for the optimal formation by calculating the quality of each possible formation generated by the set of given goals as shown in Algorithm 1. These calculations however, mean that the system has to generate all the possible formations in order to return the optimal one. Given the number of formed cohorts, the number of groups in each cohort, the number of goals, the number of constraints for each goal, the worst-case complexity of searching for the optimal solution is high. However, the implementation of an appropriate algorithm, together with a reliable description of the students can reduce this complexity by generating few possible formations.

6.2.4 Group productivity quality metric

In addition to the metrics associated with the constraint satisfaction in forming the groups, another metric, usually used to evaluate the quality of work within groups in the Group Productivity Metric. We refer by quality $Q(t)$ to how well the group achieved the collaborative task t specified by the teacher. This is a measure of the quality of the groups outcome (sometimes referred to as output or reward) against an absolute scale defined by the teacher or an examiner of the groups output. In learning, this is usually given in the form of grades or credit to the group. If both the collaboration goal and quality measure are defined by the teacher, then this is a consistent measure. In this thesis, we do not use this measure for evaluating the group formation approaches and results, but we are listing it as a possible way for evaluating groups.

6.2.5 Perceived formation satisfaction metrics

We use this metric to refer to how well the formation was perceived in terms of participants satisfaction with the allocations to groups:

- Individual Perceived Formation Satisfaction: we use this metric to refer to how pleased is the individual with being allocated a member of the group. Individual satisfaction is usually evaluated using self-assessment questionnaires. Since the questionnaires are usually composed of statements on the Likert scale or the 6 points scale, the satisfactions can be given a weight s_i for each individual p_i where s_i can be the weight mean of the questions results.

- Group Formation Satisfaction: we use this metric to refer to the individual satisfactions of all the members of the group in order to evaluate how much is the group satisfied. This metric can be also used to monitor the interactions values of the collaboration such as assistance and conflicts. We define the Group Formation Satisfaction for each group g_i as:

$$f_{sg}(g_i) = \frac{1}{L}\sum_{j=1}^{L} s_j \tag{6.7}$$

  where L is the number of participants in the group.

- Cohort Perceived Formation Satisfaction: The cohort perceived satisfaction f_sG can be defined in terms of the average and the standard deviation of all the formed groups such that for N groups:

$$f_{sG}(t) = \sigma_f \quad \text{and} \quad \overline{f_{sg}} = \frac{1}{N}\sum_{i=1}^{N}(f_{sg}(g_i)) \tag{6.8}$$

where $\sigma_f = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_{sg}(g_i) - \overline{f_{sg}})^2}$ is the standard deviation for the groups.

6.3 Summary

In this chapter, we introduced the concept of Constraint Satisfaction Problems and a model of writing group formation in education as a constraint satisfaction problem where the student participants are the variables, and the domain is the groups they can be allocated to. The constraint are the restrictions on the groupings, which also model the goals the instructor aims to achieve from setting the group formation. Given that a constraint in a CSP graph is a restriction between two variables, in our case two students, finding the right solver, will generate these restrictions from the goals (written as constraints) we feed in the problem.

At the end of the chapter, we introduced a metrics framework to evaluating group formation quality in terms of constraint satisfaction, together with other metrics that can be used on the side to evaluate the quality of the groups in terms of the participant's satisfaction and productivity. The person concerned with forming the groups can use the metrics that are relevant to their grouping and the objectives of their collaborative activity. They can either use some of the constraints or all together to evaluate the outcome of the group formation. In the next chapter, we analyse an application that we have developed that takes the concept of presenting student group formation as a CSP, and we evaluate its performance using out metrics framework.

Chapter 7

Semantic Constraint-based Group Formation

As described in the previous chapter, the problem of forming groups of students for collaborative learning can be modelled as a constraint satisfaction problem. In this chapter, we use that model to implement a framework that takes the data from the students, and pass it to the system where the instructor specifies a set of constraints on the group formation, together with the desired number of participants per group. We use Datalog to express the semantics of the constraints. This application is also supported by Semantic Web technologies to describe the students, but the focus in this chapter is on the use of semantics to express the constraints.

Given that our focus is on education, we analyse the different variables that can be used to describe the students and the constraints such as demographics, experience, and preferences. We carried an observational study to investigate how different variables affect the students' satisfaction with their groups allocations, and to show that the more data and constraints available in the group formation, the better outcome in terms of student satisfaction.

We evaluated this framework, and recorded the results from analysing its performance on datasets of students modelled specifically for different types of groups to achieve

different types of collaborative activities.

7.1 Modelling the participants

As mentioned in previous chapters, in learning, the type of grouping used to facilitate the collaboration is determined based on the objectives of the collaborative activity the teacher is introducing to the students. After specifying the type of groups to be used, the teacher is faced with the task of allocating students to groups and deciding who should work with whom in each group. The formation of the groups can aim to create: diverse groups, where the students population is evenly distributed over the groups in terms of grades, gender, nationality, and so on; complementary groups, where the group members complement each other to perform a specific task (e.g. non like-minded students); or similarity-based groups (like minded students) where the students share a common feature or interest.

However, regardless of whether the teachers aim to form homogeneous or heterogeneous groups, there are a number of variables they have to consider in the formation of the groups. These variables are used to model the students and support the constraints; and combinations of these variables are used to model the collaboration goals. This enables the teachers to initiate different formations with different combinations of the modelled constraints. Harrigan et al. (2009) assessed the dimensions that can be the subject of adaptation based on what teachers and learners think is significant in the adaptation of learning. They found the most wanted features were: learner knowledge, learning goals, interests, learning styles, and user roles. These dimensions are relevant in designing a learning activity in general, and can be therefore applied to a collaborative learning activity.

In (Winter, 2004), these attributes are categorised in four dimensions that model a team: knowledge and functional skills such as level of expertise, teamwork behaviour such as social roles, task type, and context and situation such as geographical location. In this paper, we refine this model to involve all types of groups. In this context, we categorise the variables into three types:

- Task-related: these parameters model the students in relation to the course they are taking or the task of the group work, examples of the variables in this category

are: experience, education level, knowledge, skills, abilities (cognitive and physical), grades, interests, preferences of topics and experts the student want to work with, and so on (Jackson et al., 1995).

- Relation-oriented: these parameters are independent from the topic of the collaboration as they involve more personal information on the student such as gender, age, culture (race, ethnicity, national origin, and so on), social status, personality and behavioural style, social ties, trust between members, and so on (Jackson et al., 1995).

- Context-related: these parameters hold information on the context features of the students and their environment such as geographical location, availability schedules, the communication tools used for the collaboration, and so on (Mühlenbrock, 2005). These variables are usually useful for part time and distance learners.

For each of the group types introduced in the previous section, there are some specific parameters that need to be modelled for the formation of that type. Table 7.1 illustrates the mapping between the range of these attributes and the different types of groups. In particular, except for teams that can be formed for different reasons (complementary, similarity, and so on), and thus can use any range of constraints; the table shows the variables that are crucial to the formation of each type of grouping. For example, communities of practice are usually formed based on the topic or practice that the members are interested in, but also in the type of relationship and trust between the members.

For an efficient modelling of students, we need to model a large range of attributes that can be considered for different formations such as: expertise, grades, skills, preferences, gender, ethnicity, age, team roles, interests which includes academic and social interests, social ties, and trust.

As mentioned before, roles are a crucial part of team definition; there are many efforts to define team roles (in terms of personality and role theory) in the psychological interdependence within teams. Some of these efforts are: Belbin roles (Belbin, 2004), Myers Briggs, and Keirsey roles (Higgs, 1996). The identification of these roles is usually processed using self-evaluation inventories. In this research, we choose to model the students possible roles in a group using Belbin roles (a table explaining these roles can be

	Variables		
Group Type	Task-related	Relation-related	Context-related
Teams	interests, topic preferences, experience, expert preferences, skills, abilities	Expertise, demographics, relationships, trust	Geographical location, availability schedules, communication tools
Communities of Practice	interests, topic preferences, experience, expert preferences	Expertise, relationships, trust	None
Intentional Networks	Skills, abilities, experience	None	Geographical location, availability schedules, communication tools
Social Networks	Interests, topic preference	Social ties, trust	None

TABLE 7.1: The different variables needed to be modelled for the formation of different groups

found in appendix C). The reason for this is that in addition to its cost, the Myers Briggs inventory has to be supervised by a professional. Although the Belbin self inventory has been criticised for its consistency regarding the reliability of the team roles discovered by filling the inventory due to the self-perception factor, research that employed the inventory to study teams of students tasking software engineering group projects showed that considering the Belbin roles can impact positively on the performance of the teams (Stevens, 1998) (Winter, 2004), and can provide a prediction of the teams performance based on the composition of the roles within the groups (Johansen, 2003). As described in later sections, we use the concept of ontologies for modelling.

Another variable that is usually identified through surveys is the student's learning style. Here, we use the learning styles as described by Honey and Mumford (1992), which includes theorist, reflector, pragmatist, and activist.

7.2 Observational study

To match the growing need of forming groups with higher flexibility, we started analysing what constraints do teachers consider when forming groups. We studied the possible students features that can be relevant to forming different types of groups by investi-

gating the available literature on collaborative learning theories (Ounnas et al., 2007b), and asking teachers what constraints they employ for different educational goals.

As a case study on group formation, we conducted an observational study that aims at observing the following:

- what do teachers consider in practice when forming groups for education
- how well the groups have worked in relation to the variables and constraints considered bytthe teacher
- how well would the groups perform if the teacher had taken further variables into account.

The study was run with 67 undergraduate students taking a software engineering group projects course (SEG) in the School of Electronics and Computer Science at the University of Southampton. The students were manually grouped by the course organisers into 11 groups of 5 to 6 students, based on the following constraints:

- All groups have to be balanced in terms of the students previous grades to ensure that all groups have an equal opportunity in performing well in the project.
- To avoid minorities, a female cannot be allocated to an all-male group to prevent her from being cast away by the members.
- International students from the same country cant be all members of the same group.

The module organisers used a script to allocate the students based on their marks, then manually swapped some of them to redistribute females and international students. To analyse the dynamics of the groups and how other criteria affect them, we distributed two questionnaires to the class:

Questionnaire (1)[1] at the beginning of the course, we asked the students to fill in a form to get information about their previous experience in software engineering, teamwork, their gender, nationality (to detect minorities), and Belbin team roles to check which role can each student play within their group. Belbin roles are typically

[1]Questionnaire (1) is available in Appendix A

used in industry and training activities to discover the best roles a participant can play in a group (Belbin, 2004). There are 8 Belbin roles, and according to these roles; a balanced team is composed of:

- One leader: Coordinator (CO) or Shaper (SH), and not both in the same group to avoid conflicts,
- A Plant (PL): to stimulate ideas and insure creativity,
- A Monitor/Evaluator (ME) to maintain honesty,
- One or more Implementers (IM) to executed actions, Team Worker (TW) to ensure cooperation in the group, Resource Investigator (RI) to explore opportunities and secure resources, or a Completer/Finisher (CF) to ensure all tasks are completed on time.

Each person usually plays more than one Belbin role within the team. However, a member usually scores high in only one or two roles. In our study, we collected both the first and second roles for each student. Detailed results are included in appendix C.

Due to some students dropping out of the course and others not filling in the questionnaires, we collected data from 9 groups out of the whole 11. Table 7.2 illustrates the results collected from questionnaire (1) showing Belbin roles in each group. The numbers in the cells demonstrate how many members in the group have that role as their strongest role. The distribution of Belbin roles is taken from the first and second strongest roles[2].

Questionnaire (2)[3]: at the end of the course, we distributed a 17 questions based evaluation form where the student is asked to rank the key elements that measure their group performance, dynamics, and the individual satisfaction with the group work on a 1 to 6 scale. In particular, we analysed creativity, motivation, leadership, group cohesion, satisfaction with contribution of members and the group output.

Given that in some groups, only one or two students returned the questionnaire, we were only able to use the data from groups 1, 3, 4, 5, 7, and 8. Table 7.2 shows

[2]Belbin roles are usually calculated as primary and secondary (back-up) roles, the results shown in the table bellow are mainly the primary roles, with the Resource Investigator (RI) role taken from secondary roles as there were no primary RIs in the class.

[3]Questionnaire (2) is available in Appendix B

Groups \ Roles	IM	CO	SH	PL	RI	ME	TW	CF
1	1			1	1	2	1	
2	3						1	
3	1		2		1	1	1	
4	1	1	1			1	1	1
5	1		1	1			3	1
6	4		1		1			
7	3	2	1					
8	1	1	1				2	
9	3			2	1			
Total	18	4	7	4	3	4	9	2

TABLE 7.2: Results of observational study (distribution of Belbin Roles)

these groups in shaded colour. The results showed that the majority of the groups were satisfied with the group output (the software), and no members (minorities) were isolated which can be related to the fact that the teams were formed to be balanced in terms of grades and gender. However, constant conflicts were reported in the groups that had no leader or more than one strong leader (groups 1, 7 and 8). The groups that did not have a plant member such as groups 3 reported a lack of innovation, while groups with a Plant responded well (group 1 and 5).

From the study, we observed the relation and effect of possible group formation constraints on the students perceived satisfaction. However, despite the benefits of having a number of constraints in achieving the educational goal of the collaboration, negotiating the students allocation to groups manually gets more complex and time consuming as the number of constraints grows, even if the teacher had the required data about the student.

The case study provided us with an initial understanding of the domain characteristics and relevant problems in forming groups, which support our findings from analysing existing literature in the area. Our analysis from both the literature and the case study yielded various ideas for possible computer support in both modelling the constraints and evaluating the formation of groups.

7.3 Framework structure

To optimise allocating students to groups, we propose a framework to assist the teacher in forming groups based on their chosen set of constraints. The framework handles the group formation process based on the following concepts:

- Modeling the students features: we model a large range of features that can be considered for different group formations using the concept of Semantic Web ontologies, which can form a reliable dynamic learner profile (Ounnas et al., 2007a). In this context, semantic modelling provides meaningful descriptions of the students and the relationships between them.
- Negotiating the group formation: we express the students allocation problem as a Constraint Satisfaction Problem (CSP). The negotiation process of allocating each student to their most appropriate group can then be handled by a constraint satisfaction solver.

We emphasise that, in this research, we are not concerned with proving that any particular set of constraints leads to better results in terms of the performance of the groups; neither do we claim that any particular algorithm leads to best grouping. Figure 7.1 shows an overview structure of the framework, which is based on the following components:

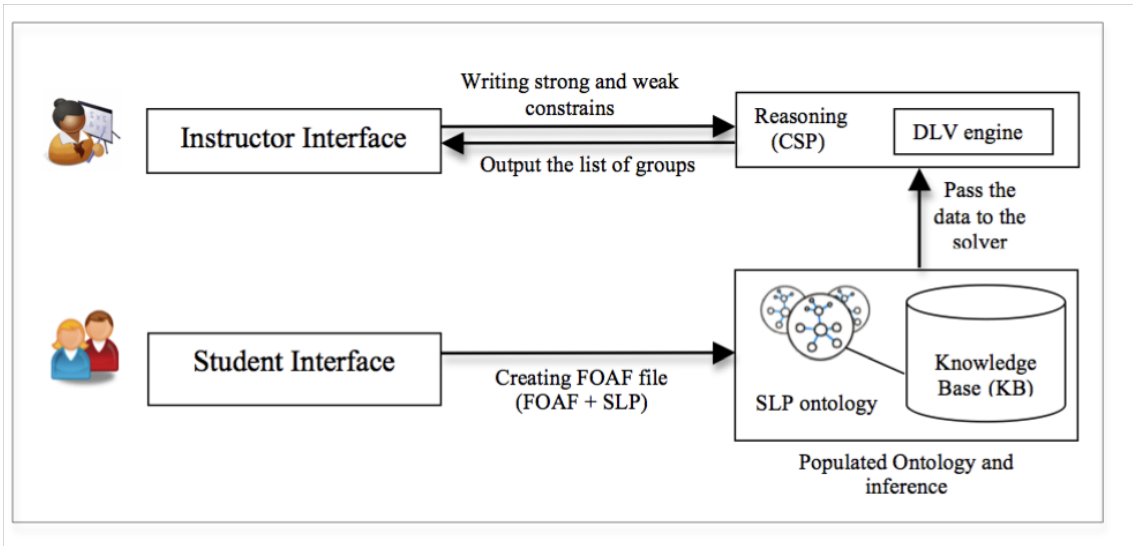


FIGURE 7.1: Semantic group formation framework

7.3.1 The student interface

The student can enter their data through a web-based form composed of four parts: the students personal data, their interests and preferences, information about their course such as the modules they are taking, and a list of their friends/colleagues taking similar courses if possible. The students can update their data at any time. To avoid the form style in the interface, and to make it more desirable to students to update regularly, we aim to make it a Web 2.0 like interface such as the ones used by social networking sites. So far, the characteristics we found can be easily collected from the students are:

- gender: being male or female
- nationality: being a string representing the person's nationality, and can infer other properties such as *european*, or *non home student*
- age: being a numerical value, a category such as 18-20, or binary such as mature student and non-mature students
- previous marks to demonstrate competency: these are presented as numerical values, or alphabet grade,
- interests: keywords representing academic or general social interests and hobbies. The same applies to grouping based on expertise on a topic or having a preference of a topic or a person,
- preferences: a preference of working on a specific topic,
- friends: names of people related to this person through friendship or previous collaboration,
- team roles: as represented by Belbin roles,
- learning styles: as represented by Honey and Mumford model,

Unlike demographics and interests, the last two items have to be collected through surveys. Students, and people in general, do not like completing surveys are are only willing to fill them if it is compulsory to do so. However, to evaluate how good forming groups in a computer based system with multidimensional variables and a high number of constraint, we had to keep these variables (roles and learning styles) for evaluation, even if it's on simulated data as will be illustrated later on in this chapter.

7.3.2 The ontology

We created an ontology called Semantic Learner Profile (referred to in this paper as SLP[4]). The ontology extends the FOAF ontology[5]. The learners characteristics that the ontology describes were chosen based on a comparison of existing learner profiles such as PAPI, IMS LIP and eduPerson (Ounnas et al., 2006). Therefore, the ontology describes a large range of students personal, social, and academic data such as learning styles, preferred modules, topics, and collaborators (Ounnas et al., 2007a). The semantic representation of these data, to which the instructor constraints can be mapped to, allows inferences to generate more data. This feature of using semantics enables the framework to handle incomplete data in a more effective way (this is explained in more details in section 4). Since the FOAF ontology is very popular (Ding et al., 2005b), employing it would allow using data from any other ontology that can be mapped to FOAF.

Once the student submits the profile data through the student interface, an RDF file is created (FOAF + SLP). The file is processed using Jena, a Semantic Web inference engine (Carroll et al., 2004), and instances of the ontology are then stored in a database. Listing 7.2 shows an example of a students FOAF file extended with the SLP ontology. In this figure, the file holds information about the students name, gender, Belbin role, preferred module, topics of interest, and friends (classmate).

```
<rdf:RDF
xmlns:rdf=``http://www.w3.org/1999/02/22-rdf-syntax-ns#''
xmlns:foaf=``http://xmlns.com/foaf/0.1/''
xmlns:rdf=``http://users.ecs.soton.ac.uk/ao05r/slp.owl''>
   <foaf:Person rdf:nodeID=``asma''>
        <foaf:name>Asma Ounnas</foaf:name>
        <foaf:gender>Female</foaf:gender>
        <slp:belbin>Implementer</slp:belbin>
        <slp:interest>e-Learning</slp:interest>
        <slp:interest>Semantic Web</slp:interest>
        <slp:preferredModule>CS1004</slp:preferredModule>
  </foaf:Person>
</rdf:RDF>
```

LISTING 7.1: Example DLV code for a small size problem

[4]available at http://users.ecs.soton.ac.uk/ao05r/slp.owl

[5]as introduced before, Friend Of A Friend (available at http://xmlns.com/foaf/spec/), is an existing ontology that describes people for building communities and social groupings

7.3.3 The instructor interface

Through this web-based interface, the instructor can select which constraints they care about for the formation they are initiating such as the students gender, their team role, their learning style and so on. For each of these variables, the instructor can constraint the group formation to be heterogeneous, homogeneous or follow some rule in that aspect. So far, it is possible to add constraints on the following characteristics:

- gender: homogeneous grouping, heterogeneous grouping, or restriction that the number of participants of a specific gender should be larger than one in each group to prevent isolation (mixed groups).
- nationality: homogeneous grouping, heterogeneous grouping, or restriction for isolation prevention as with gender. All these can be applied on the explicit nationalities or after inference, for example, inferring who are the overseas students, and then applying restriction for isolation prevention.
- age: homogeneous grouping, heterogeneous grouping, or restriction for isolation prevention mainly if the dataset has mature students.
- previous marks: homogeneous grouping, heterogeneous grouping, or enforcing heterogeneously on a specific part of the population, such as a restriction for distributing previously failed students through groups.
- team roles: homogeneous grouping, heterogeneous grouping, or restriction for a specific combination of belbin roles, such as distribution of leaders (being a shaper or co-ordinator), implementer, and plants; or restriction that no shaper and coordinator can be in the same group.
- interests: homogeneous grouping, heterogeneous grouping
- learning styles: homogeneous grouping, heterogeneous grouping

Other characteristics such as preferences and trust values between participants can be easily implemented. The interface is illustrated in figure 7.2.

For each constraint, the teacher can select a number of characteristic and the conditions on them, for example the groups are to be heterogeneous in gender and mark

distribution, homogenous in learning style, and so on. They are also provided with an option that enables them to set a priority value for each constraint. Ranking the importance of the constraints to the group formation enables the application to manage compromises based on these priorities as explained in the next section.

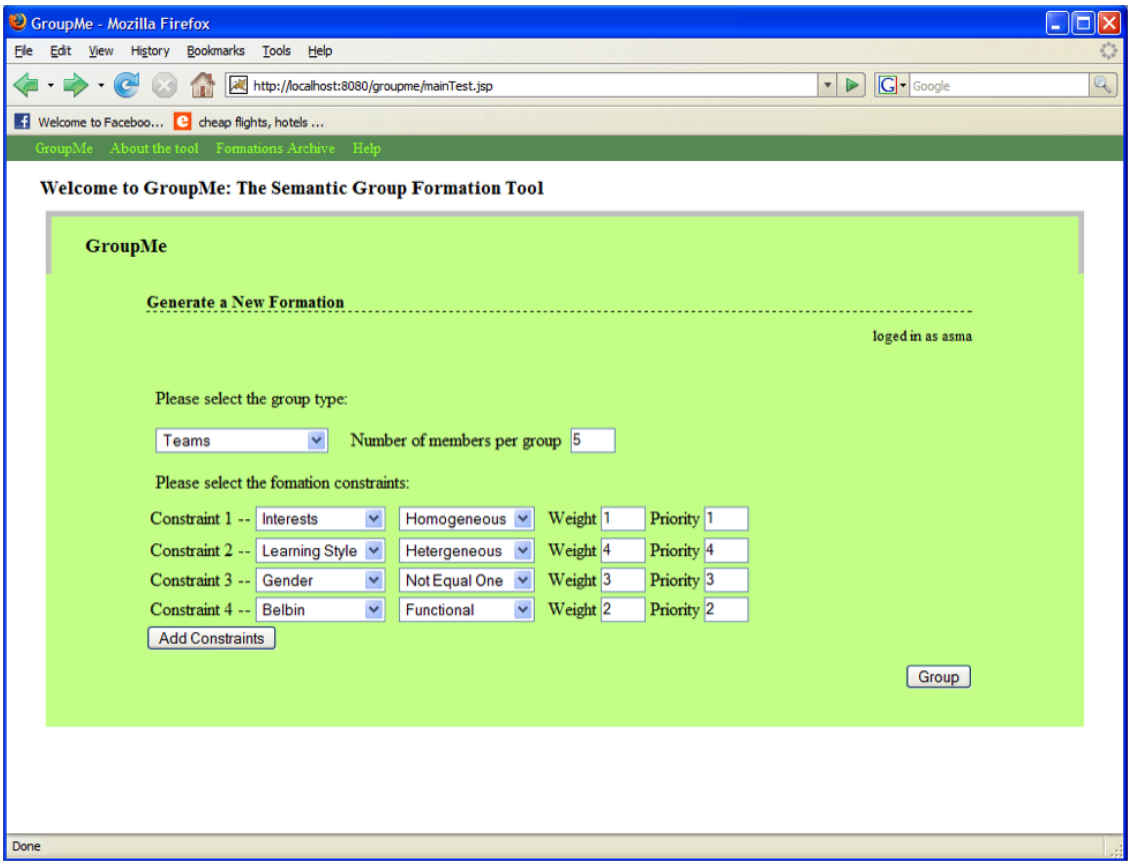


FIGURE 7.2: The instructor interface

7.3.4 The group generator

As the core component of the framework, the group generator is responsible for negotiating the allocation of students into groups. The generator is based on a DLV solver, an implementation of Disjunctive Logic Programming, an advanced formalism for knowledge representation and reasoning[6]. DLVs native language is Disjunctive Datalog extended with constraints, true negation and queries (Leone et al., 2006), where Datalog is a query and rule language for deductive databases that is syntactically a subset of Prolog.

[6]Disjunctive logic programs are logic programs where disjunction is allowed in the heads of the rules and negation may occur in the bodies of the rules.

DLV performs a simple forward checking algorithm (Kumar, 1992) on the data provided by the learners and the instructor in order to allocate students to groups. The students data is automatically transformed from the SQL database to an Extensional Database (EDB) in the form of predicates that the solver can read as an input. Figure 7.3 shows an example of this knowledge base where predicates of the form student(name,role,gender). show the students family name, Belbin role, and gender.

Through the instructor interface, we feed the list of constraints specified by the teacher. The constraints are written into a DLV program, modeled as a constraint satisfaction problem as illustrated in figure 7.3. The domain for the constraint satisfaction problems are the groups, and the variables are the student participants, such that every student can have a value from the domain being their allocated group number.

Here, we use two types of constraints: strong constraints and weak constraints (Buccafurri et al., 1997). The former are used to specify the conditions that have to be satisfied by the system in all cases. An example of these constraints would be that each student can be a member of only one group. The weak constraints are used to specify the conditions that are preferably satisfied, but can be violated if the system would not be able to find a solution otherwise. These constraints are given a priority level according to their importance in the group formation through the instructor interface. For example, in figure 7.3, the instructor considers having only one leader (shaper or coordinator) in each group to be more important than having an implementer in each group by assigning these constraints priority levels 3 and 1 respectively.

Although we allow the instructor to choose any constraint to be strong or weak through the interface, we encourage the usage of weak constraints in all constraints side from the one restricting the number members in each group, and therefore restricting the number of groups to be generated. This is because finding a solution is easier when a constraint is relaxed than when dealing with a hard constraint. By allowing the instructor to use strong constraints however, we allow them to learn some features of the dataset they fed into the framework. For example, if allocating at least one leader in each group is a strong constraint, and the framework fails to return a solution than that means that the dataset contains a small number of leaders, in other words, less leaders than the number of groups. Recommending a specific choice of constraints is out of the scope of this thesis.

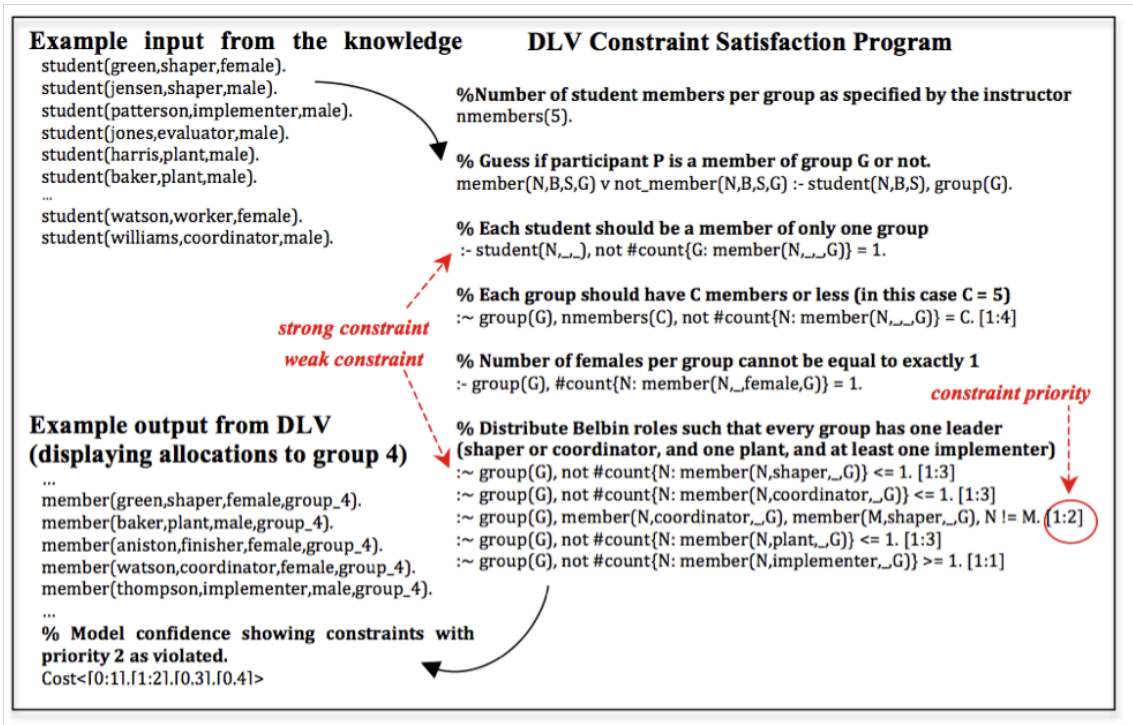


FIGURE 7.3: Example DLV program

Depending on the data provided and the constraints, DLV outputs more than one solution to the problem (i.e. more than one grouping of the students). Each solution is referred to as a model. The optimal model is hence the best grouping of students in relation to the given constraints and input data. The best model is calculated as an objective function that minimises the number of violated constraints.

Unlike other computer supported instructor-based group formation tools (Redmond, 2001) (Tobar and de Freitas, 2007), our approach does not leave student orphans. Based on the negotiation of the constraints satisfaction through optimisation, all students are allocated to some group, even if some constraints are violated. The best model is computed and the confidence of the computation (formation) is returned to the instructor: For instance, if the instructor wants only one leader per group to avoid conflicts, and gave the constraint priority 2, but the number of leaders is larger than the number of students; then some of the groups will have more than one leader. Here, a constraint of that priority is violated. Hence, the confidence of how good is the group formation is decreased. Together with the model, the confidence is computed in terms of violated constraints, and then returned as an output solution.

DLV outputs the model as a list of predicates. Figure 7.3 illustrates an example

output predicates of the form *"member(name,role,gender,group)"* showing the students family name, Belbin role, gender, and the group they are allocated to. This output data is then stored in an SQL database and then returned to the instructor through the instructor interface as a list of groups. To ensure a good practice interface design, the interface output can be dynamically manipulated by the instructor in case modifications or swapping students around the groups is preferred.

7.4 Evaluation

As mentioned in section 4.2.2 of chapter 4, most existing Computer Supported Group Formation applications are only evaluated against few metrics that do not always reflect their efficiency in forming appropriate groupings, but rather assume that a positive group output such as groups' marks can be interpreted as a success of the followed group formation approach. In this context, groups are usually evaluated based on their performance (Higgs et al., 2003) and effectiveness (Bateman et al., 2002) where effectiveness is measured in terms of the group synergy, performance objectives, skills, use of resources, and innovation. These metrics involve the groups quantified output on an absolute scale, the satisfaction of the members, and the dynamics within the group, in terms of communication and conflict.

In this study however, we are more concerned with the evaluation of the group formation rather than how well the groups perform. Therefore, efficiency here is defined in terms of the formation quality introduced in the constraints satisfaction framework in chapter 6. As discussed before, the choice of constraints has a significant impact on the predicted performance of the group. However, in this research, we consider the choice of constraints to be the responsibility of the formation initiator (in this case, the instructor). Furthermore, we do not measure the stability of the groups formed, as wether they will do well or fall apart.

7.4.1 Real data

For an initial evaluation, we used the framework to allocate students to groups within two courses in the University of Southampton including the software engineering course on the following 2 years. However, since the instructors of these courses had only a

maximum of three constraints (on previous marks, gender, and international students)[7], the framework returned a best model in both cases with no violations in both courses. We also used this framework to group students from a foundation course with 27 students, whose instructor wanted to allocate them for a heterogeneous grouping of gender and nationality, with a restriction not to isolate females and overseas students.

7.4.2 Simulated data

To monitor the performance of the proposed framework, more evaluation scenarios were set. The group formation framework was used with a range of constraints, different in content and number, and with different class sizes ranging from small class of 30 students to 100 students.

The simulated data was based on the population statistics collected from our observational study and confirmed from the UK Higher Education Statistics Agency (HESA[8]) on Computer Science demographics statistics for UK universities for 2006/2007. Tables 7.3, 7.4, 7.5 and 7.3 show the data distribution for gender, nationality[9], grades, and team roles respectively. These distributions are used to create a dataset generator for the simulation.

Gender	Distribution
Male	80%
Female	20%

Table 7.3: Gender data distribution

Nationality	Distribution
British	71%
EU	10%
Overseas	19%

Table 7.4: Nationality data distribution

From the evaluation, we observed that our system can take up to 11 constraints on 11 dimensions (participant characteristics) on a typical dataset such as the one in the SEG before the solver starts taking a long time to process the groups. Recalling the table from section 4.2.1, we add our application to the table to compare it to existing group formation applications. Table 7.7 shows this comparison.

[7]It is more common to have a small set of constraints due to the difficulty of data collection
[8]http://www.hesa.ac.uk/
[9]The nationalities were broken down into more details for both the EU and overseas categories.

Grade	Distribution
First (A)	17%
2:1(B)	36%
2:2 (C)	27%
3 (D)	14%
Fail (F)	6%

TABLE 7.5: Grades data distribution

Team role	Distribution
Implementer	30%
Team Worker	19%
Plant	13%
Shaper	12%
Monitor Evaluator	10%
Coordinator	7%
Completer Finisher	6%
Resource Investigator	3%

TABLE 7.6: Team roles data distribution

Although the results are still more efficient than any of the existing tools, the maximum number of constraints can be enhanced with the addition of some heuristics to the solvers algorithm. With the heuristics implemented, the solver can be prevented from running out of time during the computation.

The group formation quality here is measured against the satisfaction of the constraints chosen by the teacher. This includes calculating the average of how many constraints have been satisfied for each group. The formation quality is then calculated in terms of the standard deviation of the constraints satisfaction of each group, and therefore for the cohort in general. Formulas for calculating the formation quality are detailed in our evaluation framework, as described in in section 6.2. In this context, we don't measure the quality of the constraints themselves if they will lead to a good formation or not, neither do we take the students satisfaction with the allocations. Since teachers are given the freedom to choose the constraints and their importance, we trust they will choose the constraints that best fit their students needs and the collaborative task they are trying to achieve through the group formation. It is a part of our future work however, to evaluate the quality of different constraints in relation to different sets of data. Through many runs of experiments, the quality of the constraints can be measured as described in the evaluation framework in section 6.2.

Example Experiment: This experiment was designed to evaluate forming groups

Tools	Formation Features					
	Approach		Principles		Algorithm	Modeled characteristics
	Self-selecting	Instructor Based	Opportunistic	Simultaneous for all students in the class		
Hoppe	yes		yes		Rules & inference	Knowledge in a specific domain (1 dimension)
Inaba	yes		yes		Multi-agent System	Learning goal (1 dimension)
Soh	yes		yes		Multi-agent System	Performance in previous teamwork (1 dimension)
Wessner		yes	yes		Multi-agent System	Knowledge on students state within the designed learning (1 dimension)
Vivacqua	yes		yes		Profile Matching	Expertise in a specific domain e.g., Java Programming skills (1 dimension)
Redmond		yes		yes	Greedy algorithm	Preferred time slots and Preferred projects (2 dimensions)
DIANA		yes		yes	Genetic algorithm	Psychological variables, e.g. thinking styles (7 dimensions)
Team-Maker		yes		yes	Hill Climbing	Any variable (3 dimensions)
Graf		yes		yes	Ant Colony Optimisation	Performance and Personality traits, 3 dimensions
Tobar		yes		yes	Rule based	IMS LIP (4 dimensions)
Christodoulopoulos		yes		yes	Fuzzy C-Means	Knowledge and learning styles (2 dimensions)
Our approach			yes	yes	Constraint satisfaction (DLV)	demographics, learning styles, grades, interests, preferences, Belbin roles, friends (11 dimensions)

TABLE 7.7: Existing CSGF applications in e-learning

with constraint satisfaction based framework where the data is similar to the one used in real classes such as the one in the SEG course. The experiment was run with 66 students to be formed into 11 groups of 6, a typical size of class.

The data: the sample datasets used contain the following information about the students: first name, surname, gender, nationality, grade, and belbin role, and a key identifier (the student email address).

The constraints: Based on the collaboration task, we provide a set of goals, where each goal has a set of weighted and prioritized constraints as follows:

- Goal 1: The groups should be heterogeneous in gender, this includes:
  - Constraint C1: Number of females should not be equal to one to avoid having a female in an all-male group
  - Constraint C2: Number of males should not be equal to one
- Goal 2: The groups should be multicultural, this includes:
  - Constraint C3: Number of international students should not be equal to one to avoid these students being marginalised by other members
  - Constraint C4: Number of home students should not be equal to one
- Goal 3: The groups should be balanced in terms of previous grades, so they can have an equal chance in performing well in the collaboration task. This includes:
  - Constraint C5: The number of each grade is less than or equal to the number of members in the group to ensure heterogeneous groups
  - Constraint C6: If a member has a Third, none of the other members should have a Fail, to ensure weak students are distributed evenly through the groups
- Goal 4: Every group should have one leader to direct the group and at least one implementer to write the software. For this goal, we use Belbin roles to identify implementers and leaders (shapers or coordinators). This includes:
  - Constraint C7: Number of shapers should be less than or equal to one
  - Constraint C8: Number of coordinators should be less than or equal to one
  - Constraint C9: Allocate at least one implementer in each group

– Constraint C10: If you allocate a shaper, then do not allocate a coordinator in the group to avoid conflicts

We used the interface to form the groups of students with 10 constraints (all constraints from the list), 9 constraints (all except C9), 7 constraints (all except C1, C2, and C5), 3 constraints and 2 constraints. The program run and terminated successfully and produced groups with no constraints violation with the given dataset. Therefore, to challenge the framework, we evaluated its performance with an incomplete dataset.

Incomplete data: In real situations, data can be lost as a result of unaccuracy or due to self-perception based data collection. Data collected from the Web, in particular, can be incomplete. We tested the systems performance with incomplete data, by deleting data at random with an equal distribution on each characteristic considered for that group formation. Results showed that the systems still performs well (for example from no violations to 2 violations) when the data is down to 50% incomplete for a formation with 3 constraints, a moderate number of constraints for creating learning groups. However, as the number of constraints grows, the performance decreases accordingly when the data is incomplete. We observed that with low data completeness (50% and 30%), when using a higher number of constraints, such as 9 and 10 constraints, the complexity of the computation increases to the point where the solver would not return a solution within the given time limit. Because the solver keeps computing possible models (formations) in order to return the optimal one, a higher number of constraints in relation to the number of variables (students) and the nature of the populated incomplete dataset might cause the computation to extend for days.

Table 7.8 shows the results of forming groups with incomplete data with a different number of constraints. For each scenario (with a different number of constraints), we calculated the number of constraints violations (NCV), and the cohort formation quality (FQ) as the mean of the group's satisfaction as explained in our cohort group formation evaluation metrics framework in section 6.2. The cells with no results show the case where the solver runs out of time during the computation (Ounnas et al., 2009).

The nature of the dataset: When the data is 100% complete, the violation of the constraints depends on the dataset. We ran similar experiments to the one above with 100 students, and 150 students, with a similar population distribution. Similar

Data Completeness	With 2 Constraints		With 3 Constraints		With 7 Constraints		With 9 Constraints		With 10 Constraints	
	NCV	FQ	NCV	FQ	NCV	FQ	NCV	FQ	NCV	FQ
100%	0	100%	0	100%	0	100%	0	100%	0	100%
90%	2	91%	2	94%	3	96%	3	97%	5	95%
70%	2	91%	2	94%	3	96%	4	96%	6	94%
60%	2	91%	2	94%	3	96%	5	95%	7	93%
50%	2	91%	2	94%	10	87%	12	87%	-	-
30%	3	86%	6	82%	11	85%	-	-	-	-

TABLE 7.8: Results of forming groups with complete and incomplete data

results were obtained: in that with different numbers of constraints, the violation is still relatively small, and processing the computation takes only few minutes. We noticed from several runs, that the solver usually takes up to 9 to 10 minutes and returns a solution on standard computer. If the solver takes longer than that, it is most likely getting stuck looking for a solution.

However, when the population changes, the solver might get stuck. This is because, it's the dataset that determines the number of violated constraints, meaning, two different dataset, but similar in size, with the same set of constraints, will give different results, such that one might terminate with a solution, and the other might run for a long time, which might cause the solver to get stuck. To analyse this fact in more depth, we evaluated the framework with different datasets that include different attributes such as people's interests.

Data Size: As described in chapter 6, in a CSP, a constraint is a restriction between two variables, for this reason, although, in our representation of the problem, a constraint in DLV is only one single line of code, the solver will write it internally as a set of constraints, each linking the variables mentioned in that constraint. The same approach occurs with rules, as the solver maps the data to the given rules. This fact can be used to estimate the size of the problem in terms of how may constraints and rules does a particular dataset give. For instance, the simplified DLV code in listing 7.2 allocates 6 students to 2 groups based on similarity of interest:

```
group(group1). group(group2).

student(s1).
student(s2).
student(s3).
```

```
student(s4).
student(s5).
student(s6).

has_interest(s1,math).
has_interest(s2,physics).
has_interest(s3,literature).
has_interest(s4,math).
has_interest(s5,math).
has_interest(s5,physics).
has_interest(s6,literature).

same_interest(X,Y) :- has_interest(X,I), has_interest(Y,I). %(rule 1)
diff_interest(X,Y) :- has_interest(X,I1), has_interest(Y,I2), I1 != I2. %(rule 2)

members(3).
member(X,G) v not_member(X,G) :- student(X), group(G).
:- student(X), not #count{G: member(X,G)} = 1. %(strong constraint)
:~ group(G), nmembers(C), not #count{X: member(X,G)} = C [1:2]  %(weak constraint)
:~ member(X,G), member(Y,G), X != Y, diff_interest(X,Y). [1:1]  %(weak constraint)
```

LISTING 7.2: Example DLV code for a small size problem

After processing the data, the problems appears to have: 36 rules, 6 strong constraints, and 54 weak constraints. This means that for rule 1 and rule 2, it generated: (number of students*2) rules, as each rule relates two variables to each other. For rule 3, which relates the student to the group, this generated (number of students*number of groups) rules. It generated 6 strong constraints for each person counted that has to belong to only one group, and 54 weak constraints composed of counting the number of couples of students with different interest, which adds to 32 constraints plus 12 for counting the number of members per group (number of groups*number of students).

Data Partition: We also noticed that the number of groups to be produced increases the problem size. The larger the number of groups to be generated, the larger the size of the problem. We tested this property, by running an experiment similar to the interests one, but with real data taken from the interests description of people in the Learning Societies Lab at the School of Electronics and Computer Science. As opposed to reality, in this experiment, each participant had only one interest. The data had 44 people to be grouped in a homogeneous way. We noticed that the numbers of rules and weak constraints are a linear function of the number of groups. For example processing the data in this case to generate one group gives: 132 rules and 26 weak constraints. To

generate 2 groups: 264 rules, and 54 weak constraints, and so on as shown in figure 7.4.

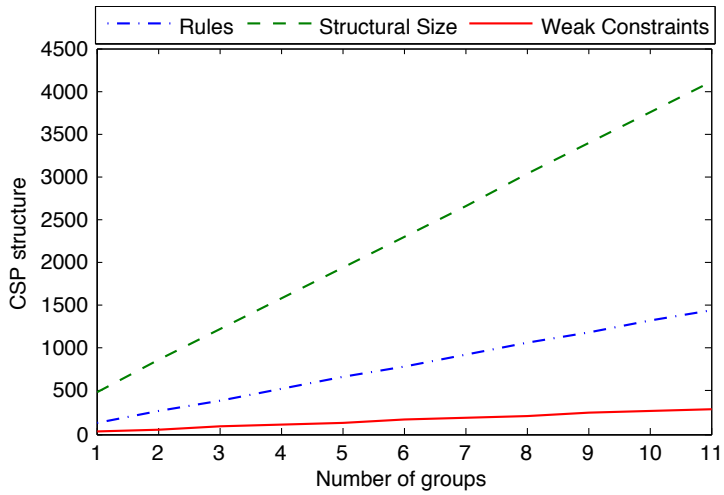


FIGURE 7.4: Number of constraints and rules generated in relation to the number of wanted groups

7.5 Summary

Modelling Group Formation as a CSP and implementing the model using inference and strong and weak constraints is a nice way to present what an instructor would want in specifying properties of forming a group. The knowledge representation used in this approach is straight forward and can be used to model any characteristics of a learner. However, although, our framework proved efficient in forming groups with many constraints that exceeded existing tools for forming groups for education, a major drawback is the size of the problem as the size of the dataset grows, and the nature of the dataset varies. These factors where causing a large number of violations starts to slow the processing of finding a solution.

Based on this, we describe another way of forming groups with large datasets based on an heuristic approach, that can handle returning a solution despite of the data size and nature. In this method, we aim to improve the results of forming groups using semantics when describing the participants. Here, the semantics of the data are imbedded in Semantic Web ontologies. This approach uses clustering as base for forming the groups, and domain ontologies to form inferences to enhance measuring distances between people.

Chapter 8

Clustering Based Group Formation

In the previous chapter, we researched the use of an optimisation approach to form groups based on a set of constraints. This approach proved efficient when the dataset is small in size; but due to its complexity, it failed to deliver results of the same standards as the dataset grew larger. In this chapter, we use a heuristic approach to form groups of participants. Clustering algorithms are known to handle a large dataset with a low complexity, at least in comparison with optimisation algorithms. In this chapter, we use clustering to form groups within a research community. We aim to test if the performance of clustering improves when we add semantics to the data using a Semantic Web domain ontology. We implement the semantics and discuss these results in the next chapter.

8.1 Methodology

In chapter 4.1.3, we covered the literature on clustering algorithms and their use in forming groups (clusters). Clustering algorithms are widely used in discovering groups based on similarity or distance between the entities to be grouped. They can process large datasets. Although the results are not guaranteed to be optimal, many heuristics

can be set to adjust the allocations to the clusters.

For this study, we use the K-means algorithm to perform the clustering of people into K groups based on one variable, in this case, the participants' interests. We use K-means as a simple algorithm that would partition our data into exactly K non-overlapping. As mentioned in chapter 4.1.3, this algorithm is used when the number of groups is known is advance, which is the case in most grouping aimed to fulfil a collaboration (or cooperative learning) activity in education. There are many algorithms that are aimed for community discovery, but here we are interested in non-overlapping small sized groups.

Given that our CSP approach did not deliver good results when each participant is associated with more than one value for the variable to be considered, for example if the variable is the user's interests, more than one interest keyword can be used to describe the participant. In this chapter, we use the same examples (participants' interests) as our variable for the group formation.

We use a dataset of people's interests that is related to education, but not a traditional classroom of students as the one used in the previous chapter. We use data of a research community formed of academic staff and postgraduate students, and we aim to put them in groups based on the similarity of their research. This simulates a situation where one would want to perform research brain storming activity, or to simply recommend potential collaborator to researchers.

In order to analyse the dataset and the clusters generated by K-means, we follow the following procedure:

1. We use a real dataset taken from the list of scientists in the school of Electronics and Computer science (ECS). This dataset contains information about each academic and postgraduate students in the school including their interests, papers, projects, and seminars they have given. The most useful items of information for the evaluation are the interests and collaboration information for these scientists.

2. We view the data as a network where participants are connected if they share interests. This will allow us to visualise the connection between people, particularly if the weights of the edges represent the similarity between the participants. The numerical description of the network will also allow us to view the difference

between the dataset before adding any semantics, and after adding inferences as a result of implementing a domain ontology of interests. The network is represented as an adjacency matrix (Newman, 2008). This is explained in more details in the following sections.

3. We apply the clustering algorithm to the network and obtain a set of clusters (groups), in this case K-means.

4. We analyse the resulting groups (clusters) for the datasets before implementing the ontology and after implementing the ontology. We evaluate the quality of the groups in both cases against a user study, where we calculate the participant's perceived satisfaction with the list of people they are allocated with in the groups. This calculation is based on a questionnaire we gave to the participants where they would indicate the list of people they think share their interests.

8.2 The dataset

The dataset for this study is taken from the list of the school of Electronics and Computer Science (ECS) members and their interests. Each member in ECS has an HTML profile page containing a list of their interests as keywords (i.e. not mapped to any ontology) as illustrated in figure 8.1. These interests are manually chosen by the person to represent their research and general preferences. The list of all people in ECS contains academics, research staff, postgraduate and undergraduate students[1]. For the evaluation, we initially only selected academics, research staff, and postgraduate students, giving a total of 842 participants. Unfortunately, only 236, a little above one third of them had explicitly stated their interests at the time of writing. The URL of the interest itself points to the list of all members that share that interest[2]. This list shows 1476 interests. Each interest in the list is accompanied by a number that shows how many people share it, for example, the interest *"social networking (7)"* shows that 7 people in ECS have stated that they are interested in *"social networking"* spelled in this exact way. Some of these interests, however, correspond to only one person. Each interest keyword is linked to the list of people associated with it, and any user can add it to their interest list publicly or internally.

[1]The list of ECS people can be found at: http://www.ecs.soton.ac.uk/people/
[2]The list of all interests can be found in http://www.ecs.soton.ac.uk/interests/

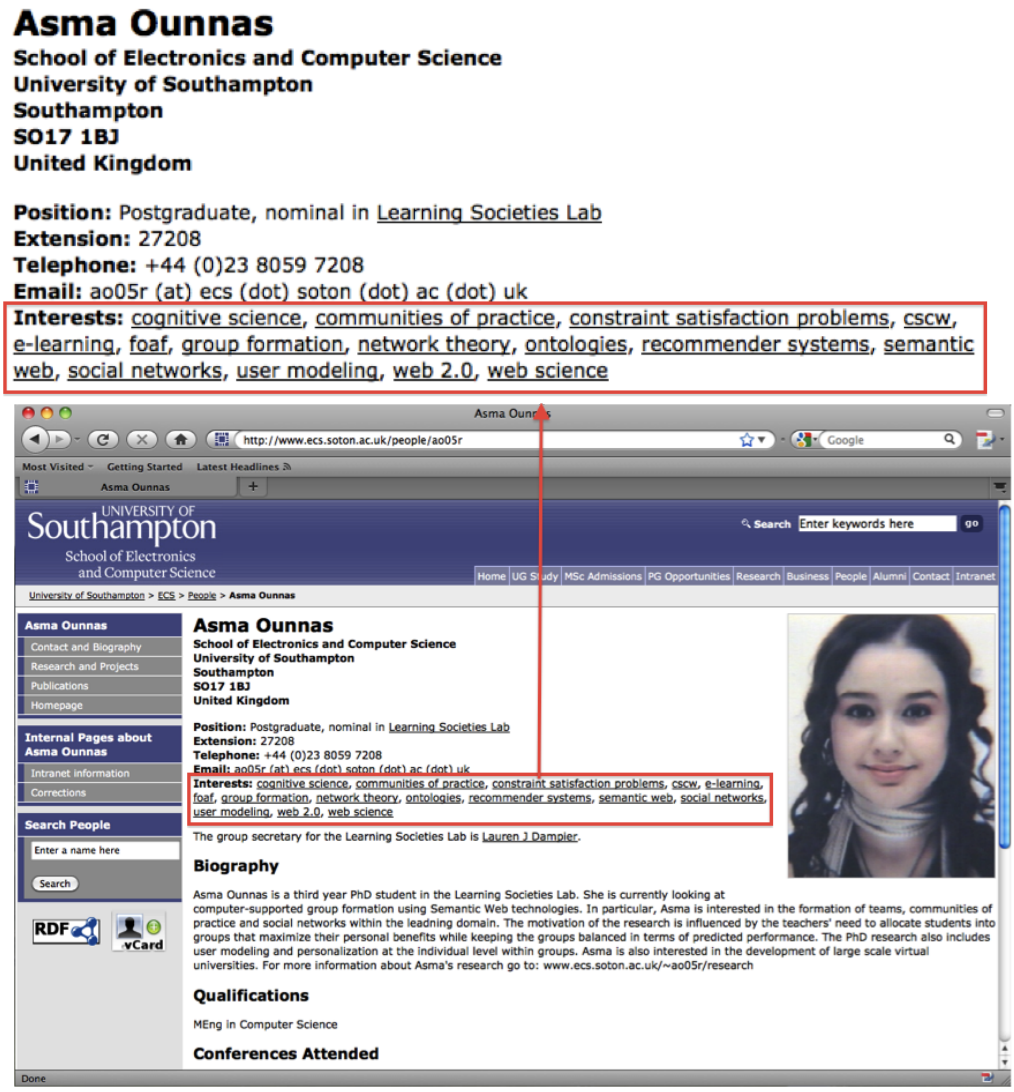


FIGURE 8.1: An example ECS page

To simplify the experiments, we use subsets of the ECS dataset. The reason behind this choice is the distribution of the data across the people and the communities. The ECS dataset is composed of smaller communities each representing a research group, where the people in a research group share more with each other than with the people in groups as they collaborate more with each other. When we first analysed the ECS interests dataset, we noticed that side from the members of 2 communities, over 80% of the people do not include any interests keywords to describe themselves in their ECS pages profile; making the data from these communities insignificant to what this study tries to achieve. Therefore, we only used the data from the 2 communities where people describe themselves. To illustrate this fact, we include figure 8.2 that shows the ECS interests network. From this network, we can observe that the dense cluster is the 2

communities (datasets) we are using in our experiments, and the shallow end of the network, where most nodes are connected to a maximum 3 nodes, is the remaining communities. These nodes have fewer links because they have very few interests, and there are only few nodes as we only scrapped the list of people in ECS who have at least one interests. The nodes' sizes and colours in the figure are adapted to their centrality in the network.

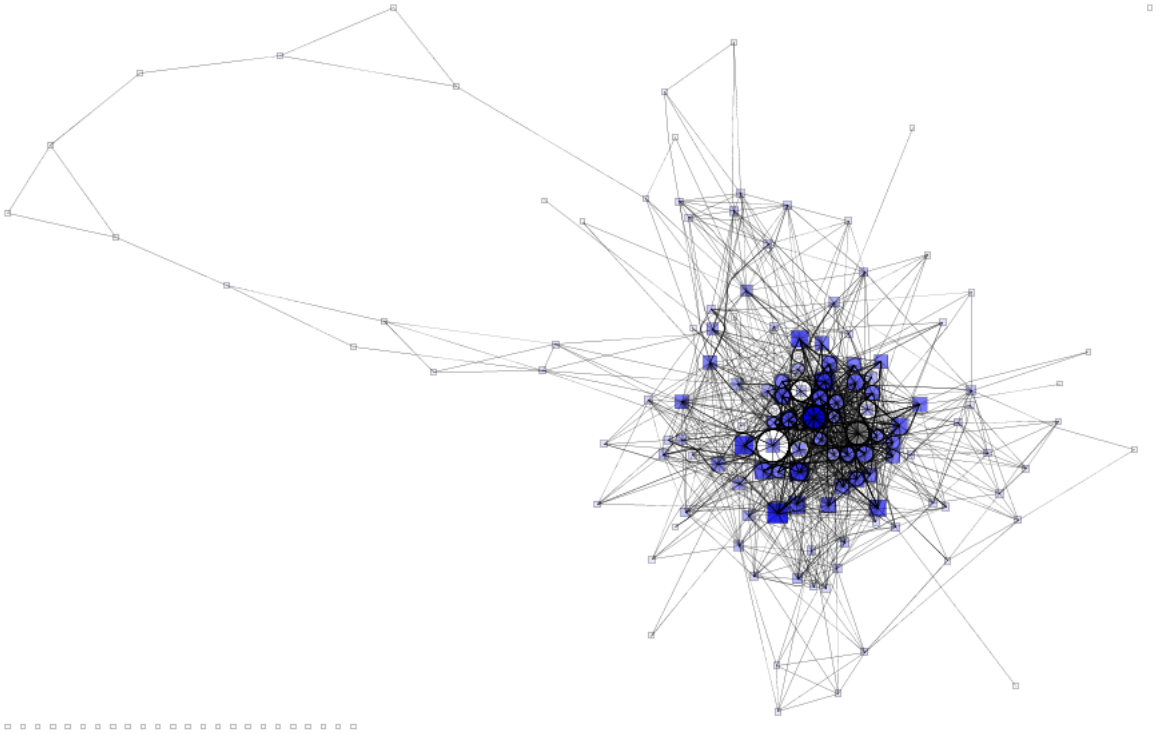


FIGURE 8.2: The ECS network

The two datasets we used in our study are the following:

- **The LSL dataset:** containing data of people in the Learning Societies Lab research group. The group had 35 members[3] who are interested in studying computer applications in education. Given that we access the data from the Web, our data scraper only mined the data for the people who have their interests on public display, reducing the dataset to 28 people. In total, the participants had 161 interests at the time of processing the data. This dataset has been used by (Yang et al., 2009) to research the creation of social academic networks.

- **The WebFest dataset:** containing data of people from ECS who participated in

[3]Counting only academics and postgraduate students, at the time of writing

a WebFest, a research related event that took place in the School in May 2009 to study various topics relating to the applications of the World Wide Web, Semantic Web, and Web 2.0, and their impact on daily human activities. Most participants in this research were from two main communities in the ECS: the Intelligence Agents and Multimedia (IAM) research group (the main group running the event), and some members of the LSL research group introduced in the previous dataset. Similar to the LSL dataset, our scraper only mined the data available on public display on the ECS pages. The dataset has 61 members. In total, the participants had 325 interests at the time of processing the data.

The participants' profiles in the ECS pages are linked to their corresponding, automatically generated, RDF files. In these files, concepts and properties are defined by the ECS ontology and the FOAF ontology. The RDF in listing 8.1 presents a fragment of the profile presented in figure 8.1:

```
<rdf:RDF>
...
<ecs:Person rdf:about="http://id.ecs.soton.ac.uk/person/9520">
...
<ecs:hasGivenName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
Asma</ecs:hasGivenName>
<foaf:givenname rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
Asma</foaf:givenname>
<ecs:hasInterest rdf:resource="http://id.ecs.soton.ac.uk/interest/group_formation"/>
<ecs:hasInterest rdf:resource="http://id.ecs.soton.ac.uk/interest/cognitive_science"/>
<ecs:hasInterest rdf:resource="http://id.ecs.soton.ac.uk/interest/semantic_web"/>
...
<ecs:hasInterest rdf:resource="http://id.ecs.soton.ac.uk/interest/ontologies"/>
<ecs:hasInterest rdf:resource="http://id.ecs.soton.ac.uk/interest/foaf"/>
<ecs:hasInterest rdf:resource="http://id.ecs.soton.ac.uk/interest/cscw"/>
<foaf:homepage rdf:resource="http://www.ecs.soton.ac.uk/~ao05r/"/>
<foaf:mbox rdf:resource="mailto:ao05r (at) ecs (dot) soton (dot) ac (dot) uk"/>
...
</ecs:Person>
</rdf:RDF>
```

LISTING 8.1: Participant profile in RDF

We use the ECS profiles' data to form groups of people based on the similarity of their interests.

8.3 Creating the network

Networks or graphs have in recent years emerged as an invaluable tool for mathematically representing, describing, and quantifying complex systems in many branches of science from the World Wide Web and the internet to social and biochemical systems as mentioned in section 3.3.3 of chapter 3. Networks often exhibit hierarchal organisation, in which vertices divide into groups that further subdivide into groups of groups, and so forth over multiple scales, such as communities on social networks.

Clauset, Moore, and Newman (2008; 2007) presented general techniques for inferring hierarchal structure from networks data and show that the existence of hierarchy can simultaneously explain and quantitatively reproduce many commonly observed topological properties of networks, such as degree distributions, high clustering coefficients and short path lengths. They further show that knowledge of hierarchical structure (refer to section 4.1.3) can be used to predict missing connections in partly known networks with high accuracy, and for more general network structures than competing techniques.

Network properties: According to the literature, there is a number of properties that can be observed or calculated to provide an interpretation of the network's data. These properties reveal more information about the network topology and will allow us to compare different networks:

- **Shortest path:** A fundamental concept in graph theory is the geodesic, the shortest path of vertices and edges that links two given vertices. Calculating the shortest path between any vertices can give an estimate to the average distance of reaching any node in the network.
- **Network diameter:** Given all the shortest paths between all pairs of nodes, the diameter is the longest path (number of edges) of these paths.
- **Clustering coefficient:** This measure assesses the degree to which nodes tend to cluster together in a graph. The coefficient ranges from 0 to 1, 1 being the maximum clustering.
- **Number of components:** This is the number of connected cluster of nodes, if each node is connected to at least one node in the network, the entire graph will be one component.

Network visualisation: using visualisation tools allows the representation of the information in a human readable format, providing visualisation of important organisational features of the network, which can be a useful tool for practitioners in generating new hypotheses about the organisation of networks. We used three visualisation tools: "yEd"[4], "SocNet"[5], and "Pajek"[6] to observe and analyse the networks.

Building the network: We build the interests network by representing each person as a node, and creating a link between two nodes if they share the same interest. Given that some people might share more that one interest, we produce a weighted network, where the weight of an edge represents how many interests are shared between the nodes. The weights are therefore a measure of similarity between the participants of the network. Given that this network is an undirected graph, it is represented by a symmetric adjacency matrix.

The adjacency matrix of a finite directed graph G on n vertices is the $n \times n$ matrix where the non-diagonal entry a_{ij} is the number of edges from vertex i to vertex j, and the diagonal entry a_{ii} is the number of edges from vertex i to itself. In our case, since every person share the interests with themselves, the distance from a node to itself is 0, and is not presented by a loop edge in the graph. An example adjacency matrix is shown bellow as table 8.1 illustrates.

	A	B	C	D
A	0	1	1	1
B	1	0	1	0
C	1	1	0	1
D	1	0	1	0

TABLE 8.1: An example adjacency matrix

To build the adjacency matrix, we first create a people/interest matrix that records the interests that each person has. The matrix size is $P \times I$ where P is the size of the people's list and I is the size of the interests' list. The diagonal entries a_{ij} of the matrix represent the weight of person i having interest i. In this case, since the ECS pages do not include weights for each interest, but rather a simple list of interests for each person, the matrix entries a_{ij} will be 1 if that person has that interest in their page, and 0 if they don't have it in their page. Example of this matrix is shown bellow as table 8.2. In

[4]http://www.yworks.com/en/products_yed_about.html
[5]http://socnetv.sourceforge.net/
[6]http://vlado.fmf.uni-lj.si/pub/networks/pajek/

this study, all interests are considered, even if it is not share, in other words, only one person is interested in it. Due to the nature of this interests (keywords), the interest *"social network"* and the interest *"social networks"* are not considered the same. This is visible from the list of all interests in ECS[7].

	semantic web	web 2.0	badminton	php	pervasive computing	e-learning
Alice	0	1	0	1	0	1
Bob	0	0	1	0	1	0
Carol	1	0	0	1	0	1
Dave	1	1	1	0	0	1
Eve	0	0	0	0	0	1
Ivan	0	0	0	0	0	1

TABLE 8.2: Example of the people/interests matrix

8.3.1 Similarity measures

To create the adjacency matrix from the people/interest matrix, we calculate the similarity of each two participants. We use the cosine coefficient similarity measure to normalise the distance between any given participants, which is also the weight of edges between the codes in the network. Given two participants with a list of n interests for each as their vectors, the cosine similarity of the two vectors is a mathematical measure of how similar two vectors are on a scale of 0 to 1, 1 being that the vectors are either identical, or that their values differ by a constant factor. The cosine similarity $cos(\theta)$ for vectors A and B is calculated as follows:

$$cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=0}^{n} a_i b_i}{\sqrt{\sum {a_i}^2}\sqrt{\sum {b_i}^2}} \tag{8.1}$$

The maximum value of this measure represents the maximum similarity between the two participants, in this case, cos equals 1 is the maximum value for similarity, whereas 0 is the minimum value representing the fact that the participants do not share any interests. To confirm the results of using the similarity measure on out datasets, we also calculated the similarity using the tanimoto coefficient[8] on the binary data (for the

[7] http://www.ecs.soton.ac.uk/interests/

[8] Tanimoto similarity T is calculated as the number of overlapping items N_c between sets A and B and divides it by the sum of all items minus the number of shared items: $T = \frac{N_c}{(N_a + N_b - N_c)}$

similarity calculations where all values are 1 if that participant has that interest, and 0 if they don't), and a simple similarity measure that only takes into account the sum of the shared interests. Similar results were obtained to those generated using the cosine similarity measure.

Using this similarity measure, the distance between each participants is calculated and stored as the entry a_{ij} between participants i and j respectively. The adjacency matrix created from the people/interests matrix shown in table 8.2 is illustrated in table 8.3. Given that the adjacency matrix is symmetric, we only create half of it.

	Alice	Bob	Carol	Dave	Eve	Ivan
Alice	1	0	1/2	2/5	1/3	1/3
Bob	0	1	0	1/5	0	0
Carol	1/2	0	1	2/5	1/3	1/3
Dave	2/5	1/5	2/5	1	1/4	1/4
Eve	1/3	0	1/3	1/4	1	1
Ivan	1/3	0	1/3	1/4	1	1

TABLE 8.3: Participants' adjacency matrix example

8.4 K-means clustering

Similar to the assumptions taken in modelling group formation as a constraint satisfaction problem, our main interest remain to form non-overlapping groups of the participants. Given that the number of groups is known in advance, we use a simple K-means algorithm to partition the data to K groups as explained in algorithm 2.

Although it can be proved that the procedure will always terminate, the K-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres. The K-means algorithm can be run multiple times to reduce this effect. In our case, we run the algorithm for few thousands iterations.

Algorithm 2: K-means Algorithm

Given K groups;
begin
  Create K initial centroids c_i to represent the initial groups and place each centroid into the space represented by the objects (participants) that are being clustered;
  foreach *participant* p_j *to be grouped* **do**
    foreach *centroid* c_i **do**
      calculate the similarity $S_j i$ of the participant to this centroid;
      if *it* $S_j i$ *is higher than the similarity to the previous centroid* **then**
        assign p_j to cluster i;
      end
    end
  end
  When all participants have been assigned, recalculate the positions of the K centroids (average) Repeat the outer loop until the centroids no longer move (this may take few thousands iterations)
  return K *clusters*;
end

8.5 The results

The results generated from creating the networks and running K-means on the data are illustrated in the following subsections for both the LSL and the WebFest datasets. In the next chapter, we will discuss these results in comparison to the results obtained from the datasets enriched with a Semantic Web ontology that represents the participants' interests.

8.5.1 The LSL network

We generated the network for the LSL dataset and plotted it using Pajek. The network is illustrated in figure 8.3, and holds the following properties as shown in table 8.4:

Number of nodes	28
Number of edges	234
Number of components:	6
Clustering coefficient	0.591999
Shortest Path	1.63636
Network diameter	4

TABLE 8.4: Properties of the LSL network before applying the semantics

The number of components is high due to the fact that 5 nodes are not linked to any other nodes. In other words, based on their explicit interests, the participants

represented by these nodes do not share any interests with the rest of the network, which in itself is a component of 23 participants.

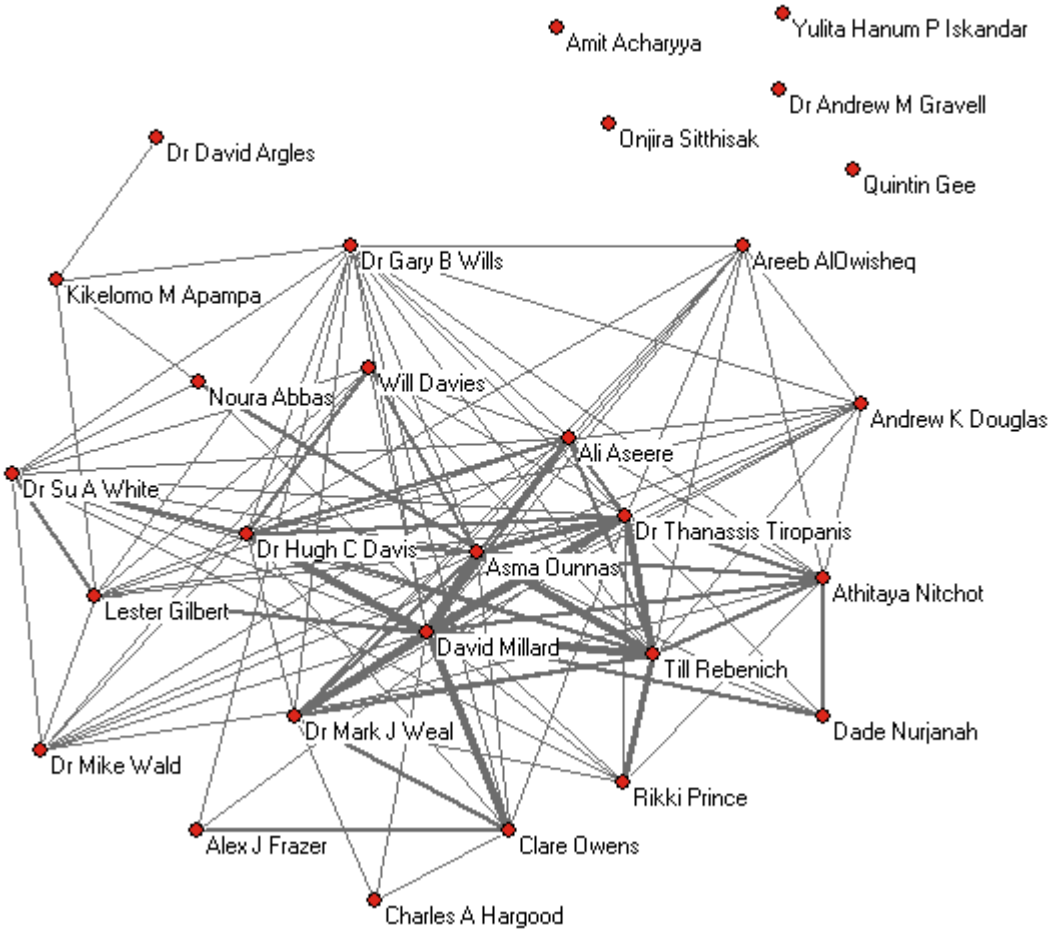


FIGURE 8.3: The LSL network

We run the K-means algorithm with the LSL data with K=5. Table 9.4.1.1 shows the members of each cluster. Comments on these memberships are discussed in the next chapter as we evaluate the participants' satisfaction with the members of the groups they have been allocated to.

Cluster	Participants
1	Amit, Yulita, Onjira, Quintin, Andy
2	Noura, Areeb, Kikelomo, Will, David, Hugh, Lester, Mike
3	Ali, Till, Mark,
4	Andrew,Athitaya, Dade, Asma, Clare, Rikki, Thanassis, Dave
5	Alex, Su, Gary

TABLE 8.5: Clusters created based on the LSL dataset

8.5.2 The WebFest network

Similar to the LSL dataset, we generated the network for the WebFest dataset and plotted it using Pajek. The network is illustrated in figure 8.4, and holds the following properties as illustrated in table 8.6:

Number of nodes	61
Number of edges	1620
Number of components:	9
Clustering coefficient	0.752953
Shortest Path	1.63
Network diameter	4

TABLE 8.6: Properties of the WebFest network before applying the semantics

Although the network shows one big dense cluster, the number of components is 9 due to the fact that 8 nodes are not linked to any other nodes. In other words, based on their explicit interests, the participants represented by these nodes do not share any interests with the rest of the network. The WebFest network is illustrated in figure 8.4.

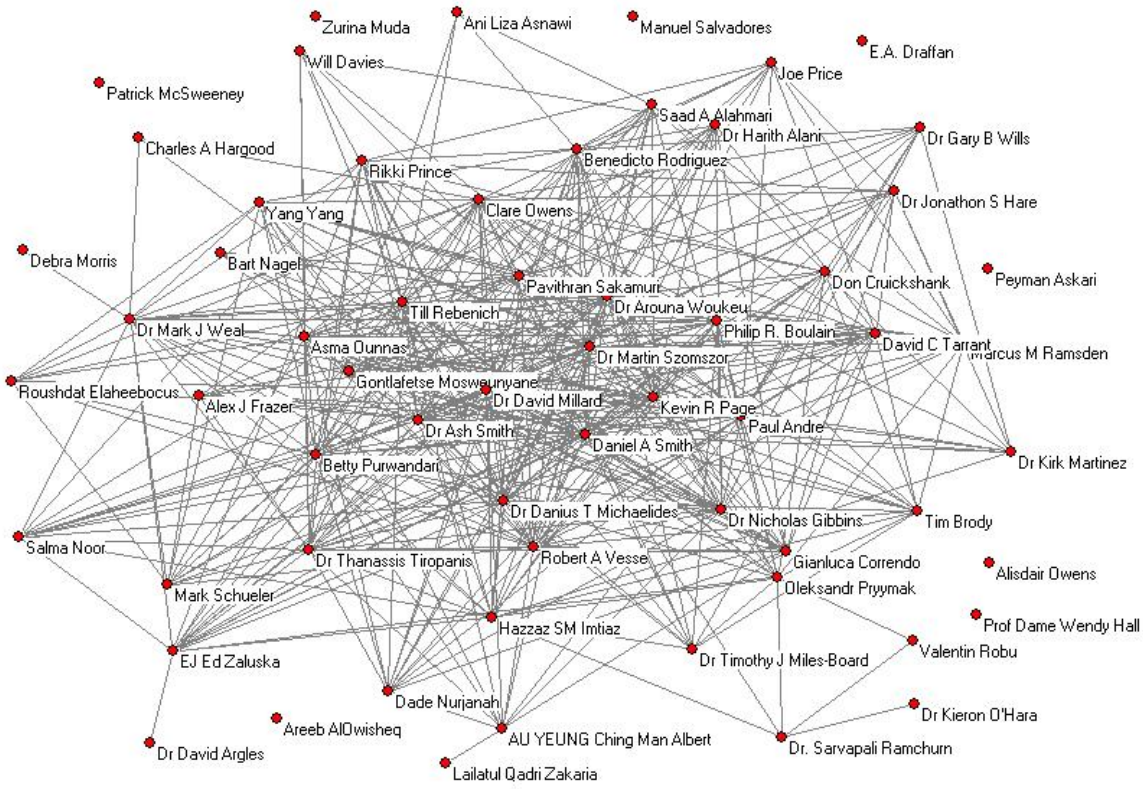


FIGURE 8.4: WebFest Network

We run K-means on the WebFest dataset with k=8. Table 9.11 shows the members

of each cluster. Similar to the LSL dataset, we discuss these memberships in the next chapter as we evaluate the participants' satisfaction with the members of the groups they have been allocated to.

Cluster	Participants
1	Dr David Argles, EJ Ed Zaluska, Dr Thanassis Tiropanis, Peyman Askari, Prof Dame Wendy Hall, Alisdair Owen, Manuel Salvadores, Mark Schueler, Zurina Muda, Saad A Alahmari
2	E.A. Draffan, Dr Danius T Michaelides, Dr Ash Smith, Roushdat Elaheebocus, Philip R. Boulain, Robert A Vesse
3	Marcus M Ramsden, Dr Jonathon S Hare, Tim Brody, Dr Timothy J Miles-Board, Daniel A Smith, Don Cruickshank, Pavithran Sakamuri, Bart Nagel, Dr Gary B Wills
4	Rikki Prince, Ani Liza Asnawi, Till Rebenich, Areeb AlOwisheq, Debra Morris, Dr Mark J Weal
5	Asma Ounnas, Paul Andre, Will Davies, Yang Yang, Dr Kieron O'Hara, Dr Kirk Martinez, Lailatul Qadri Zakaria, AU YEUNG Ching Man Albert
6	Dr Arouna Woukeu, Joe Price, Benedicto Rodriguez, Gontlafetse Moswe-unyane, Salma Noor, Betty Purwandari, Oleksandr Pryymak, Hazzaz SM Imtiaz, Dr. Sarvapali Ramchurn, Valentin Robu
7	Dr Martin Szomszor, Dr Nicholas Gibbins, Kevin R Page, David C Tarrant, Dr Harith Alani, Gianluca Correndo
8	Patrick McSweeney, Clare Owens, Dr David Millard, Charles A Hargood, Alex J Frazer, Dade Nurjanah

TABLE 8.7: Clusters created based on the WebFest dataset without the ontology

8.6 Summary

Clustering, as an approach for forming groups, can process large datasets in comparison to optimisation as constraint satisfaction problems. However, when the data is incomplete, the resulting clusters do not project real life, and often groups the participants at random when their data does not project their similarity. In the next chapter, we add Semantic Web ontologies to enhance the inference of similarities between participants in order to obtain better results from clustering algorithms.

Chapter 9

Semantic Clustering Based Group Formation

In this chapter, we use the clustering approach introduced in the previous chapter to form groups of people. However, using the same datasets, we add a Semantic Web ontology to improve the distance measure between the participants, and therefore the results of the group formation. The ontology represents the participants interests, mapping them to the ACM classification of Computer Science subjects.

9.1 Methodology

We use the K-means algorithm discussed in the previous chapter to perform the clustering of people into K groups based on a variable, that in this case is their interests. We then add an ontology of topic interests to the participants' description and we run the clustering algorithm again to see how the results improve. We compare the results against the network of the participants linked by edges that represent the property of sharing interests. The edges are weighted, and the weights are the similarity of interests.

To evaluate the value of adding semantics to the description of participants in order to improve the allocations of these participants to groups, we use the same datasets (LSL

and WebFest) taken from the list of scientists in the school of Electronics and Computer science (ECS). This datasets contain information about the interests of each academic and postgraduate student in the school. To evaluate the studies we:

1. add semantics to the specified characteristic (e.g people's interests), by linking the available data to the interests topic ontology. We add 3 types of semantics. We use the concepts of the ECS ontology. We map these concepts to the ACM Classification to represent them in a hierarchy. We then add the related, alternative, and hidden labels to identify the relationships between the concepts using SKOS.

2. as in the previous chapter, based on the new data, we build a network of scientists where edges represent their relationship for a specific characteristic, e.g. share interests or co-authorship. The network is represented as an adjacency matrix.

3. apply a clustering algorithm to the network and obtain a set of groups (K-means in this case).

4. analyse the groups (clusters) that form in the network in comparison to the ones generated by K-means before implementing the ontology. These are the results obtained from the previous chapter.

5. we run a user study to collect the participants' satisfaction with the groupings. In this study, we ask each participant to choose a number of participants that they think share their interests. Based on their responses, we compare their answers against the groups generated by the clustering algorithm before and after adding the ontology.

Given that we explained building the network, and the clustering algorithm in the previous chapter, details on these steps will not included in this section.

9.2 Building the ontology

We build a Semantic Web domain ontology to describe the interests of the participants in the LSL and the WebFest datasets. The ontology contains the participants interests as keywords (ECS interests) mapped to a hierarchy of interests to model the relationships between these keywords. Given that most ECS interests are Computer Science

keywords, we had to provide an efficient Hierarchy of Computer Science subjects to map the ECS interests. For this reason, we employ the ACM classification of Computer Science subjects, a well known hierarchy for describing computer Science publications and conferences. In this section, we describe our motivation behind using the ACM classification as a base for our ontology to represent the participants' interests.

9.2.1 The ACM classification

The ACM classification system was first published in 1964, and has gone through six revisions since to reflect the change in Computer Science research interest. Revised versions appeared in 1982, 1983, 1987, 1991, and the current version in 1998. For 20 years, it served as the primary and most generally used system for the classification and indexing of the published literature of computing.

The ACM Computing Classification System (CCS)[1] is hierarchically structured in four levels: three outer levels, coded by capital letters and numbers, and an uncoded fourth level of subject descriptors. Thus, for example, one branch of the hierarchy contains:

I. Computing Methodologies, which contains:

I.2 Artificial Intelligence, which contains:

I.2.4 Knowledge representation formalisms and methods, which contains:

Temporal logic (as a subject)

The classification is used to describe the topics of research papers to be published by ACM Press, which allows proper indexing and retrieval information in the ACM portal. The classification is a hierarchy of computer science topics organised in categories. The highest categories, eleven in total, are associated with a letter of the alphabet from A to K. The latest top categories in the classification are illustrated in table 9.1. Each of these categories has subcategories which themselves are divided into subcategories.

Due to the number of evolving topics, the ACM CCS does not include all the terms and topics in Computer Science, but rather, the user of the system, most likely an author

[1] http://www.acm.org/about/class/1998

Alphabet	Category
A	General Literature
B	Hardware
C	Computer Systems Organisation
D	Software
E	Data
F	Theory of Computation
G	Mathematics of Computing
H	Information Systems
I	Computing Methodologies
J	Computer Applications
K	Computing Milieux

TABLE 9.1: The ACM classification categories

of a paper, needs to use the terms or categories that are closest to the topic of their paper. The ACM CCS contains some items such as names of programming languages that are not an explicit part of the classification. These uncoded items are referred to as *Implicit Subject Descriptors*[2].

Implicit Subject Descriptors (also called "Proper Noun Subject Descriptors") are proprietary names of products, systems, languages, and prominent people in the computing field, along with the category code under which they are classified. For example, "C++" is under "D.3.2 Language Classifications". Listing is alphabetical by name. The sorting of people's names is by first name, not surname. There was only one name in the ECS interests' list, which is not in the ACM descriptors list. This name is "Ted Nelson".

9.2.2 Editing the ACM classification

In 1998, the ACM proposed processes for making annual changes, and recommends a future total revision (Coulter et al., 1998). The CCS remains a four-level, hierarchical taxonomy with 11 unchanged top-level nodes. At each of the three lower levels, index terms were added, retired, or revised, with increasing frequency through levels 2, 3, and 4.

Items at levels 2 and 3 are sometimes cross-referenced to indicate close relationships. As intended in the original design of the CCS, lower-level nodes (and their associated terms) allow the tree to expand and, occasionally, contract most easily to accommodate computing's rapidly changing nature. The heuristic is that a word that appears

[2]http://portal.acm.org/lookup/ccsnoun.cfm

frequently but is not an index term might be considered as a new concept for the CCS.

Retirement of terms from the CCS is facilitated by a count of frequency of usage of CCS terms over the past three years. If any term has been used less than five times in each of the past three years to index documents, it will be automatically deemed appropriate to be retired unless the Maintenance Committee sees some reason not to retire the term. Consequently, some of the newer major subfields of computing, such as Computational Science and Human-Computer Interaction, are not clearly represented in the CCS. Moreover, some major categories of the CCS (e.g., category E: "Data") have become increasingly irrelevant in the modern literature, and ought to be redesigned or combined with other categories (e.g., perhaps as Data and Databases) to reflect a more contemporary and enriched major subject category.

Despite the fact that many would consider the ACM classification to be old or slowly evolving in comparison to the field of Computer Science, many researchers used the classification to represent computing related interests. Stefanov (2003), for example, developed an ontology covering the Computing Education domain based on the ACM CCS, although not much details were reported on its applications. Mirkin et al. (2008) proposed a method to map clusters of ontology classes of lower level onto a subset of high level classes in a way that the latter can be considered as a generalised description of the former. The authors mapped a list of research topics that represents the research of Computer Science Research Organisation to the ACM classification (as an ontology). Using their method, Mirkin et al. (2008) can describe the research interest of the organisation, in other words, which top ACM category does the organisation most fit in to.

9.2.3 Adding concepts to the classification

Implicit Subject Descriptors do not appear as part of the formal scheme because they are too numerous to include without making the formal scheme too cumbersome. The ACM classification maintenance team claims that the list is dynamic and sees frequent updates as new names are introduced. However, the list does not include all the computer science terms one might express as an interest or as a descriptor for an ACM publication. For example despite it's popularity, PHP is not in the list of descriptors. The list also classifies some terms under more than one category, for example "Prolog" is classified as:

F.4.1, H.2.3, D.3.2, I.2.3, and I.2.5. In order to map all the ECS interests to the ontology, we need to first allocate each interest to the correct category of the classification. Given that the classification does not include all the terms in the ECS interest list, processing these allocations involves following a procedure to ensure each term is added to the right category of the classification. Figure 9.1 illustrates an example simple addition of terms to the ACM classification. The procedure for allocating ECS interests terms to the classification works as follows:

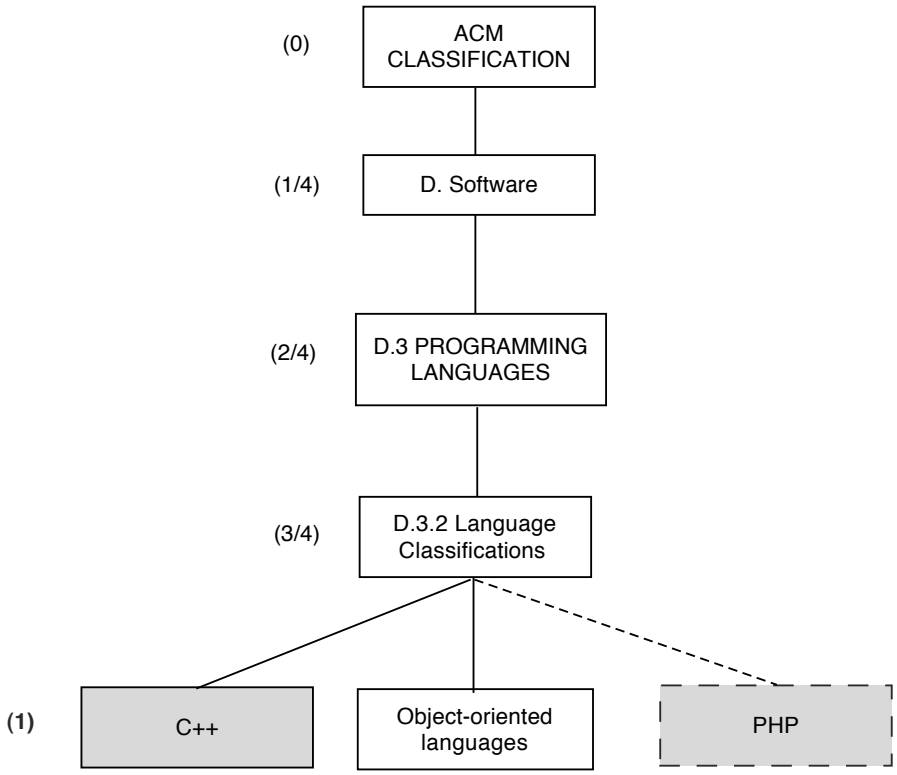


FIGURE 9.1: Adding terms to the ACM Classification

- **Representing the hierarchy**[3]**:** we converted the ACM classification to an OWL ontology, where all categories and subjects are of type *acm:Classification* (an OWL class). The relationships between the categories and subjects in the hierarchy are expressed through *SKOS:broader* and *SKOS:narrower* relations. These two relations are inverse of each other and express the fact that a concept, in our case an ACM category, is broader or more general than another concept, and vice versa. Although each category has only one broader category, the terms from the Implicit Subject Descriptors such as "Prolog" are added to have more than one broader category as specified by the ACM.

[3]The ACM classification hierarchy is available at http://users.ecs.soton.ac.uk/ao05r/acm/acm.owl

- **Related concepts:** given that some categories in the ACM classification are related to each other, we expressed this relationships using the *SKOS:related* relation, which shows that two concepts are related, but not equivalent to each other.

- **Alternative Labels:** we created relationships between the ECS interests based on their labels (syntax) using the *SKOS:altLabel* relation. This included identifying the following terms, and expressing them as alternative labels:

  - Plurals: such as *"social network"* and *"social networks"*
  - Format between verb and noun: such as *"social network"* and *"social networking"*
  - Spelling: such as *"e-learning"* and *"elearning"*, and *"personalisation"* and *"personalization"'*
  - Abbreviations: such as *"cscw"* for *"computer supported collaborative work"* and emph"soa" for *"service oriented architecture"*
  - Composed terms: such as *"ubiquitous and pervasive computing"* and *"accessibility and usability"*

- **Hidden Labels:** we identified misspelled interests such as *"mobile computng"*, and we used the *SKOS: hiddenLabel* relationship to represent their relationships as recommended by the SKOS specification. This shows that an interest might have one or more hidden labels which represent a misspelling of it.

- **Mapping the ECS interests to the ACM classification:** For each interest in the ECS interests list:

  - If interest already appears in the ontology, then associate it with the relevant concept. For example, the ECS interest "information systems" matches the ontology class *H. Information systems*, which would mean that the interest would have the same children (narrower categories) in the hierarchy as the ACM classification *H. Information systems*.
  - Otherwise, if the interest appears in slightly different format from one of the concepts of the ontology, for example as plural or in a verb format, then allocate in the same position as the corresponding singular noun concept of the ontology.

– Otherwise, if the interest is not in the ontology, then check if it is in the Implicit Subject Descriptors' list. If it is in this list, then allocate it in the ontology according to the Implicit Subject Descriptors' classification.

– Otherwise, add the interest as a narrower concept to the appropriate class in the ACM classification. To decide on where to allocate the concepts, we used Wikipedia, and the ACM portal with keywords from publications of one of the people who have that interest in ECS.

9.2.4 Using Wikipedia and the ACM portal to allocate concepts

Wikipedia[4] is a multilingual, Web-based encyclopaedia, that is written collaboratively by volunteers and is available for free. The majority of Wikipedia pages have been manually assigned to one or multiple categories. For example, figure 9.2 shows the Wikipedia categories (super classes) for the page "Ubiquitous computing".

FIGURE 9.2: Example of wikipedia categories

Wikipedia have been used by few researchers as a reliable resource to aid generating taxonomies, ontologies, and as a semantic network which serves as a basic for computing the semantic relatedness of words. This is because it is: domain independent, up-to-date, multilingual, has a disambiguation page, has infoboxes, has lists and categories (Wu and Weld, 2007).

Researchers also relied on wikipedia as it provides a wide coverage online encyclopedia developed by a large number of users. In their research, Ponzetto and Strube (2007) used methods based on connectivity in the network and lexico-syntactic patterns to label the relations between wikipedia categories. As a result they were able to derive a large scale taxonomy.

In a similar research, Suchanek et al. (2008, 2007) developed YAGO, a large ontology derived from Wikipedia's info-boxes and category pages and WordNet that proved to have high coverage and precision. YAGO is available online, however, unfortunately we

[4]http://www.wikipedia.org/

can not use it for this research as it does not create relationships, particularly subclass relationships between topics in a domain dependent way as needed in our case; but rather defines the nature of a word such as Paris *is-a* Capital, and *is-a* noun.

In addition to that, Wu and Weld (2007) developed a prototype system called KYLIN, that autonomously extracts structured data from wikipedia and regularises its internal link structure. Szomszor et al. (2008) also used wikipedia categories to semantically model the interests of users, where the interests are tags from the user's Flickr and Del.icio.us accounts to generate richer user profiles. Auer and Lehmann (2007) presented a method for revealing this structured content by extracting information from template instances such as the ones in infoboxes and categories. From our observation, the wikipedia categories confirmed the classification of most of the concept that are already classifies in the ACM portal.

The ACM Portal[5] contains all the ACM publications and some non-ACM publications with a classification to the ACM CCS based on the keywords filled by the author when submitting the publication. According to the ACM portal, some papers (topics) can be mapped to more than one concept, these are referred to as "Primary Classification" and "Additional Classification". Therefore, we allowed that ECS interests can be mapped to more than one ACM classification concept[6]. For example: "Adaptive educational hypermedia" is classifies under "Hypermedia" and "computers use for education".

Percentages of mapping the ECS interests to the ACM classification are shown in table 9.2. The percentages in the table represents how many of the ECS interests were allocated with assistance from the corresponding resource.

Resource	Percentage
ACM CCS	34%
ACM Portal	22%
Wikipedia	19%
ACM Portal + Keywords	18%
ACM Implicit Subjects	7%

TABLE 9.2: Percentage of ECS interests concept classification within the ACM CCS

Some of the interests were allocated such that they have more than one ACM broader category. For example, the interest *"semantic web"* has *"Knowledge Representation*

[5]Accessed from http://portal.acm.org/portal.cfm

[6]The classification for ECS interests in relation to the ACM classification is available at http://users.ecs.soton.ac.uk/ao05r/ecs2acm.owl

Formalisms and Methods" and *"World Wide Web"*. And some of the ECS interests are the broader category for other ECS interests. For example, the interest *"semantic web"* has 17 narrower ECS interests such as *"semantic wiki"*, *"semantic web services"*, and so on.

9.2.5 Evaluating the ontology mapping

We run an expert review to evaluate the accuracy of mapping the ECS interests to the ACM classification. The expert review took the form of a short questionnaire that asks the reviewer to answer the following questions in relation to the ontology.

- Does the hierarchy/ontology as presented seem to be modelled correctly?
- Do you think the ECS interests are well mapped to the ACM classification?
- Do you find any unintended redundancy, and if so where is it?
- Is there anything you believe should be added or moved within the hierarchy?
- Any additional comments?

4 participants filled in the expert review, all were from the school of Electronics and Computer Science (ECS). The suggestions collected from the experts regarding amendments needed were taken into account. Most of the suggestions resulted in adding new relationships between the ECS interests. For example, one of the experts suggested that the terms: *"computer supported learning"* and *"technology enhanced learning"* are alternative labels to the term *"e-learning"*. The final version of the ontology has 360 concepts. 247 concepts are in the ECS ontology and 110 are ACM concepts.

There are some professional ways to evaluating ontology mapping such as measuring data representation, data precision conflicts, data unit conflicts, naming conflicts, and aggregation conflicts (Kaza and Chen, 2008). Ontology mapping and matching researchers introduced a number of tools to evaluate mappings and surveys that evaluate these tools (Kalfoglou and Schorlemmer, 2003), (Conroy et al., 2009). Kaza and Chen (2008) introduced an evaluation of ontology mapping techniques, while Noy and Musen (2002) and Shvaiko and Euzenat (2005) evaluated ontology-mapping tools.

In this research, we do not use a tool to evaluate our mapping, and details of these tools are outside the scope of this thesis. Here, we consider the expert review results to be satisfactory to enable us to use the ontology to prove our hypothesis.

9.3 Inferring the interests

After mapping the ECS interests to the ontology, we recalculate the similarity between the participants and therefore the weights of the edges connecting them by inferring the new interests. Given that the inference of weights is based on the relationships of the ontology, we revisit the adjacency matrix and weight distribution as follows:

The Hierarchy: For each interest the participant have, we add the broader interests to their interest set with different weights depending on the level (depth) of the category, such that the higher the category (further on top of the interest in question), the lower the weight it receives.

Related concepts: We also add the related interests so that a related interest holds a fragment of the weight of the current interest based on how many interests are related to the current interest. For example, if the weight of the current interest is 1.0, then the weight of its only one related interest would be augmented by 0.5, but if the weight of the interest is 1.0, and it has 2 related interests, then the weight of each of its interest would be augmented by 0.33.

Alternative and hidden labels: Interests that are hidden labels or alternative labels to the current interest hold the same weight as the latter. For example, if Alice has interest *social networking* with weight 0.5, then her inferred weight for *social networks* would also be 0.5, which means that if Bob has interest *social networks* with weight 1.0, then Alice and Bob now have an interest in common, which although obvious, would not be detected without the semantics as seen in chapter 8 where the clustering is performed based on the set of interests as literals.

9.4 The results

In order to evaluate our approach of clustering participants with the aid of the Semantic Web based domain ontology, we compare the results of clustering with semantics to the results obtained from clustering without semantics as described in the previous chapter. We use the exact same datasets introduced in previous chapter. We create networks based on the adjacency matrices holding the similarity measures between the participants (now described using the ontology), and we run K-means algorithm to generate the groups. We compare the generated groups in terms of the participants' satisfaction obtained from a user study. The results are compared with the ones obtained in the previous chapter in order to observe the impact of using ontologies to infer the participants' interests on clustering-based group formation.

9.4.1 The LSL dataset

This is the same dataset introduced in the previous chapter under the same name. The dataset has 28 participants from the LSL research lab, each associated with a number of keywords they chose to describe their interests (mostly academic interests).

9.4.1.1 The network and the groups

The network generated from processing the LSL dataset with the ontology is illustrated in figure 9.3. The network's properties are shown in table 9.3. The table also shows the difference between this network and the one generated from the same dataset before the ontology was used to infer the interests. We observe that the number of components has decreased, this is due to the fact that the new network has 3 less lonely nodes. These nodes are now connected to other nodes in the main components, due to the fact that the inferring their interests resulted in discovering some connections (common interests) with other nodes. For the same reason, the density of the network increased significantly by almost doubling the number of edges (connections between nodes).

We run K-means on the newly generated dataset with $K = 5$ (an average size of a group). Table 9.5 shows the membership of the participants within the clusters. In relation to the results obtained from running the K-means on the data before applying

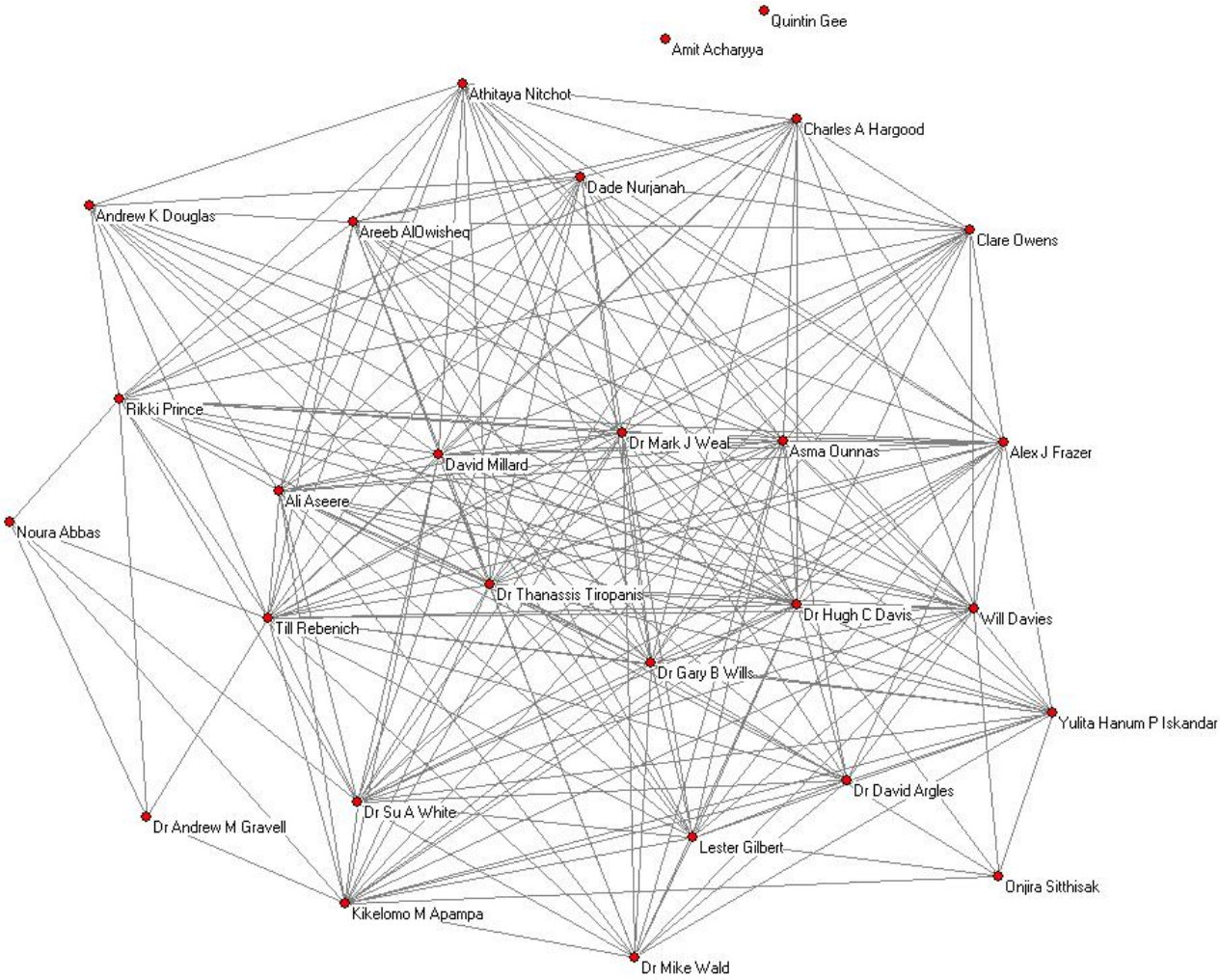


FIGURE 9.3: LSL network after using the ontology

Network Property	Value in New Network	Value in Non-semantic Network
Number of nodes	28	28
Number of edges:	518	234
Number of components	3	6
Clustering coefficient	0.858816	0.591999
Shortest Path	1.20308	1.63636
Network diameter	2	4

TABLE 9.3: LSL network properties after inferring the interests in comparison to their values before using the ontology

the ontology, we noticed 17 changes as 17 participants moved to a new group whereas 12 participants stayed in the same group.

Cluster	Participants
1	Amit, Yulita, Onjira, Quintin, Andy
2	Noura, Areeb, Kikelomo, Will, David, Hugh, Lester, Mike
3	Ali, Till, Mark,
4	Andrew,Athitaya, Dade, Asma, Clare, Rikki, Thanassis, Dave
5	Alex, Su, Gary

TABLE 9.4: Clusters created based on the LSL dataset before the ontology

Cluster	Participants
1	Amit, Quintin
2	Will, David, Lester, Gary, Mike, Kikelomo, Yulita, Onjira
3	Andrew, Ali, Rikki, Till, Dade, Athitaya, Thanassis, Su, Hugh
4	Alex, Areeb, Charlie, Asma, Clare, Dave, Mark
5	Noura, Andy

TABLE 9.5: Clusters created based on the LSL dataset with the ontology

We observe that the semantics-based clustering shows more defined themes of interests. To study this effect, we calculated the top interests within each group for both clustering results of the dataset: without and with the implementation of the interests' semantics using the ontology. We calculated the weights of each interest within each group in a similar way to calculating tag clouds, where the tag (interest) that is most shared appears with a a bigger weight.

The interests' weights are calculated such that for each interest i, the weight of i is equal to the sum of the weights of each occurrence of i in the group. For example, if a group has participants A, B, and C, where A has interest i with weight 1, B has interest i with weight 0.5, and C does not have that interest (weight=0), then the weight of i within the group is 2.5. If an interest is not shared, in other words, only one participant has that interest, the weight will not be calculated, and the interest is discarded. Table 9.6 and table 9.7 show the interests weights for for the groups created with the clustering alone, and the clustering with the aid of the ontology respectively. In both tables, group 1 does not show any interests. This is because the participants within the group do not have any interests in common, due to the fact that group 1 is still the group of lonely nodes, even after implementing the ontology.

From table 9.7, we observe that some themes are emerging, such that: group 2 is

	Top interests in order of their weight
Group 1	-
Group 2	e-learning (2), security (2), project management (2), computer-assisted assessment (2)
Group 3	semantic web (4), e-learning (2), pervasive computing (2),
Group 4	semantic web (5), web 2.0 (5), e-learning (3), pervasive computing (3), narrative (3), web science (3)
Group 5	e-learning (3), e-assessment (2), hci (2), virtual research environment (2), learning and teaching (2)

TABLE 9.6: Interests weights for the LSL dataset groups before implementing their semantics

	Top interests in order of their weight
Group 1	-
Group 2	computer use in education (6.5), computer-assisted assessment (6.5), computers and education (4.99), computing milieux (4.2)
Group 3	e-learning (7.33), semantic web (7.33), information systems (6.81), world wide web (6.39), web 2.0 (5.91), hypertext/hypermedia (4.24)
Group 4	information systems (30.08), world wide web (9.07), information interfaces and presentation (7.36), semantic web (6.83), web science (5.5), information storage and retrieval (5.82), hypertext/hypermedia (5.41), web 2.0 (4.58), narrative (4.5), user interfaces (4)
Group 5	software engineering (3), software (2.5), agile methods (2)

TABLE 9.7: Interests weights for the LSL dataset groups after implementing their semantics

generally interested in e-learning and assessment; group 3 is generally interested in e-learning and Web studies: World Wide Web, Semantic Web, Web 2.0, hypertext and hypermedia; group 4 is generally interested in information systems, user interfaces, and web science; and group 5 seem to be generally interested in software engineering. Given the fact that the inference generates more interests, we only show interests with weight higher that 4 in table 9.7. We made an exception for group 5 as it only has 2 participants, which would have similar weights to the other top interests in other groups if the weights were averaged. However, our objective from the calculation is to observe the top (the ranking) of the interests rather than just the weight.

9.4.1.2 Participants' satisfaction

In order to compare the results obtained from the clustering, we asked each participant to pick up to 5 people from the list of participants that they think share their interests. The participants' answers would represent the people they would choose to be in their group if the grouping was based on similarity of interests. This user study questionnaire is available in appendix D. Based on the participants' responses to the questionnaire, we calculated the participants' satisfaction with both sets of clusters C1 and C2, such that:

- C1 is the set of clusters (groups) we created without the aid of the ontology in the previous chapter,
- C2 is the set of clusters (groups) that we created with the aid of the ontology in this chapter.

The individual satisfaction of each participant is calculated as a ratio of the number of people in their chosen list who have been allocated to their clustering generated group to the total number of people they have chosen (maximum 5, given that k=5, and the dataset has 28 participants). We then calculated the group satisfaction for each group based on the average and standard deviation of the individuals' satisfactions within the group. Then, we calculated the cohort's satisfaction based on the average of all groups' satisfactions. We finally compared these satisfactions of the clustering with the ontology to the groups' satisfaction of the clustering without the ontology. The exact formulas for calculating the averages and the standard deviations for the group satisfaction and the cohort satisfaction are described in section 6.2.5.

Out of the 28 participants in the dataset, 23 replied, 16 of them had their individual satisfaction increased in comparison to their satisfaction with the non-semantic clustering, 5 did not have their satisfaction changed, and only 2 had their satisfaction decreased. Figure 9.4 shows the individual satisfaction percentage of change in comparison to clustering without semantics. Table 9.8 shows the average and the standard deviation of the individual satisfaction with the clustering results for both C1 (clustering without semantics), and C2 (clustering with semantic). As shown in the table, the average satisfaction of the participants with the groups generated by the semantic-based clustering

is much higher that their satisfactions with the groups generated by the non-semantic clustering from the previous chapter. In fact, the average satisfaction has more than doubled thanks to the implementation of inferences using the ontology.

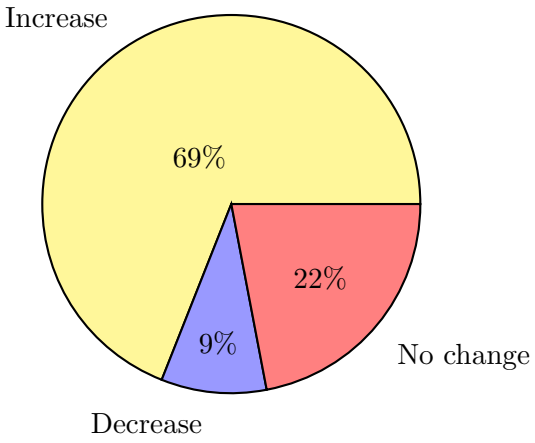


FIGURE 9.4: LSL dataset individual satisfaction change in comparison to clustering without semantics

	C1 grouping	C2 grouping
Average Satisfaction	0.243	0.592
Standard Deviation	0.227	0.301

TABLE 9.8: Individual satisfaction with the group allocations for the LSL dataset: groups formed with clustering "C1" & groups formed with clustering + the ontology "C2"

Table 9.9 shows the results of the groups' satisfaction and the cohort satisfaction. From the results in the table, we can see that the cohort is much more satisfied with the semantic-based group allocations, proving that using semantics to enhance the results of the clustering generated better results than just clustering the data as it is without inference. Although the standard deviation for the groups' satisfaction is lower before implementing the semantics, the average group satisfaction in this case was very low, meaning that most participants for each group agreed that their satisfaction is low. To prove that this is not just the case with the LSL dataset, we perform the same procedure to study the impact of using semantics to enhance the clustering of the WebFest dataset.

9.4.2 The WebFest dataset

This is the same dataset introduced in the previous chapter under the same name. The dataset has 61 participants from the school of Electronics and Computer Science

		C1 grouping	C2 grouping
Group 1	Average	0	-
	Std Dev	-	-
Group 2	Average	0.2	0.64
	Std Dev	0.18	0.20
Group 3	Average	0.05	0.34
	Std Dev	0.1	0.25
Group 4	Average	0.42	0.68
	Std Dev	0.04	0
Group 5	Average	0.36	1
	Std Dev	0.14	0.19
Cohort	Average	0.20	0.66
	Std Dev	0.18	0.26
	Avg Dev	0.14	0.19

TABLE 9.9: Group satisfaction with the group allocations for the LSL dataset: Groups formed with clustering "C1" & Groups formed with clustering + the ontology "C2"

at the University of Southampton. Particularly, the people in this datasets have all participated in a research event called "WebFest" that relates to the study, application, and development of the World Wide Web. Similar to the LSL dataset, the WebFest dataset is a list of people each associated with a number of keywords they chose to describe their interests (mostly academic interests). The evaluation methodology used on this dataset is the same methodology we used on the LSL dataset.

9.4.2.1 The network and the groups

The network generated from processing the WebFest dataset with the ontology is illustrated in figure 9.5 (Observe that the network is so dense, that the shortest path and diameter are minimal, and the clustering coefficient is very high). It is a known fact that all the participants in this dataset have some interest in the World Wide Web as a topic of research, therefore the nodes representing them in the network are highly connected, and with inferring more interests, it is only natural that the number of edges would double. The network's properties are shown in table 9.10. The table also shows the difference between this network, and the one generated from the same dataset before the ontology was used to infer the interests. Similar to the results obtained from the LSL dataset, the number of components has decreased, in this case, 7 lonely nodes are now connected to the main component of the network, and only one remains out with no

interests to share with the remaining nodes. We checked the interests of this participant, and they none of them were related to research, which would explain the fact that the inferences applied by the ontology on their interests, did not create any new links.

Network Property	Value in New Network	Value in Non-semantic Network
Number of nodes	61	61
Number of edges:	3216	1620
Number of components	2	9
Clustering coefficient	0.946541	0.858816
Shortest Path	0	1.63
Network diameter	0	4

TABLE 9.10: WebFest network properties after inferring the interests in comparison to their values before using the ontology

We run the same K-means algorithm on the new semantic-based WebFest dataset with $K = 8$. The resulting clusters are shown in table 9.12, which illustrates the membership of the participants within the clusters. In relation to the results obtained from running the K-means on the data before applying the ontology, we noticed 52 changes, as 52 participants moved to a new group whereas only 9 participants stayed in the same group.

Cluster	Participants
1	Dr David Argles, EJ Ed Zaluska, Dr Thanassis Tiropanis, Peyman Askari, Prof Dame Wendy Hall, Alisdair Owen, Manuel Salvadores, Mark Schueler, Zurina Muda, Saad A Alahmari
2	E.A. Draffan, Dr Danius T Michaelides, Dr Ash Smith, Roushdat Elaheebocus, Philip R. Boulain, Robert A Vesse
3	Marcus M Ramsden, Dr Jonathon S Hare, Tim Brody, Dr Timothy J Miles-Board, Daniel A Smith, Don Cruickshank, Pavithran Sakamuri, Bart Nagel, Dr Gary B Wills
4	Rikki Prince, Ani Liza Asnawi, Till Rebenich, Areeb AlOwisheq, Debra Morris, Dr Mark J Weal
5	Asma Ounnas, Paul Andre, Will Davies, Yang Yang, Dr Kieron O'Hara, Dr Kirk Martinez, Lailatul Qadri Zakaria, AU YEUNG Ching Man Albert
6	Dr Arouna Woukeu, Joe Price, Benedicto Rodriguez, Gontlafetse Moswe-unyane, Salma Noor, Betty Purwandari, Oleksandr Pryymak, Hazzaz SM Imtiaz, Dr. Sarvapali Ramchurn, Valentin Robu
7	Dr Martin Szomszor, Dr Nicholas Gibbins, Kevin R Page, David C Tarrant, Dr Harith Alani, Gianluca Correndo
8	Patrick McSweeney, Clare Owens, Dr David Millard, Charles A Hargood, Alex J Frazer, Dade Nurjanah

TABLE 9.11: Clusters created based on the WebFest dataset before the ontology

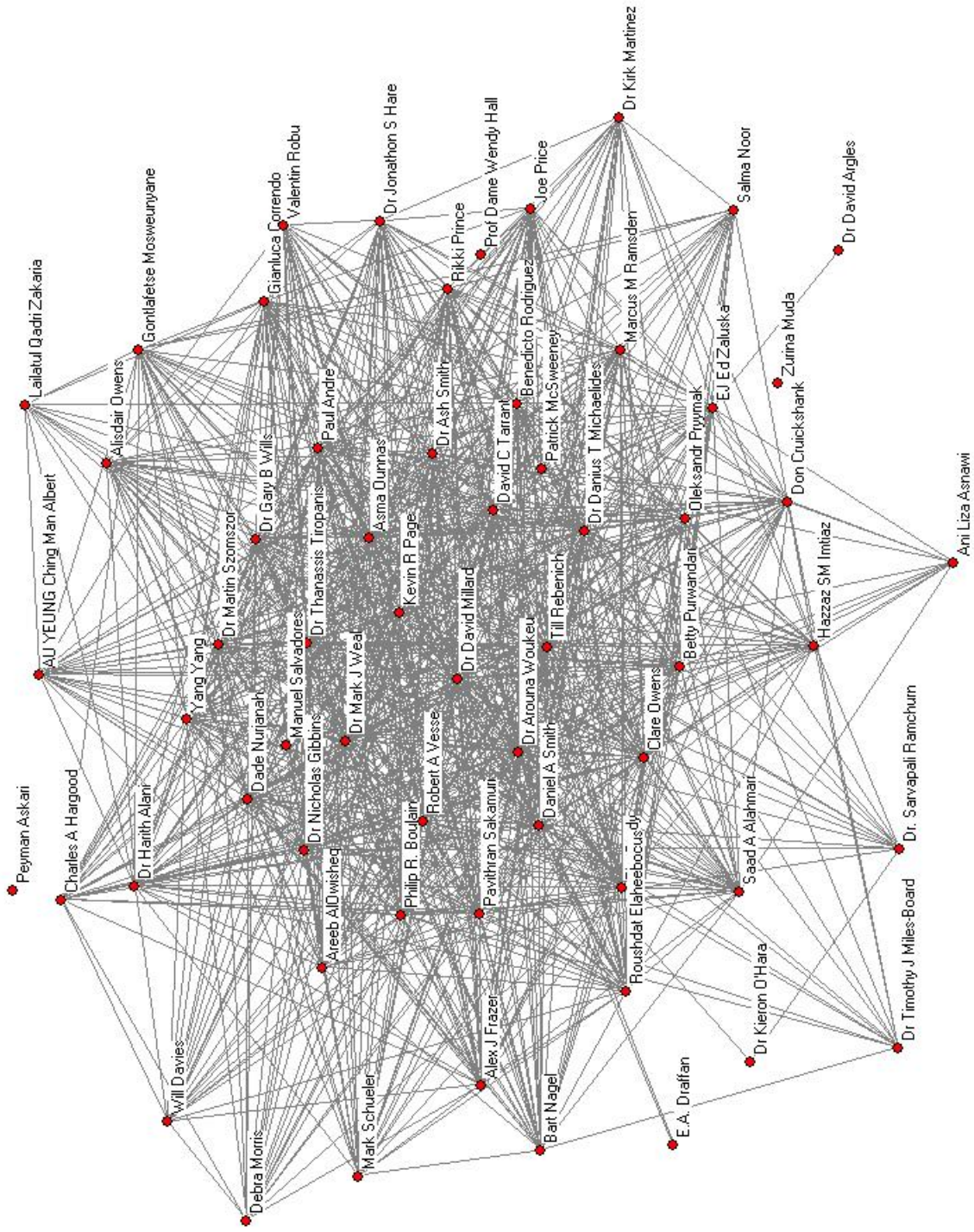


FIGURE 9.5: WebFest network after using the ontology

Cluster	Participants
1	Dr David Argles, EJ Ed Zaluska, E.A. Draffan, Dr Gary B Wills, Will Davies, Dr Arouna Woukeu, Debra Morris
2	Joe Price, Rikki Prince, Till Rebenich, Ani Liza Asnawi, Bart Nagel
3	Philip R. Boulain, Dr Nicholas Gibbins, Dr Danius T Michaelides, Dr Timothy J Miles-Board, Dr Martin Szomszor, Yang Yang, Gianluca Correndo, Dr Harith Alani, Benedicto Rodriguez, Manuel Salvadores, Kevin R Page
4	Alisdair Owens, Ash Smith, Dr Jonathon S Hare, Gontlafetse Mosweunyane, Salma Noor, Lailatul Qadri Zakaria, Don Cruickshank, Roushdat Elaheebocus
5	Zurina Muda, Oleksandr Pryymak, Saad A Alahmari, Robert A Vesse, Pavithran Sakamuri, Areeb AlOwisheq
6	Peyman Askari, Valentin Robu, Dr. Sarvapali Ramchurn
7	Asma Ounnas, Dr David Millard, Mark Schueler, Dr Thanassis Tiropanis, Betty Purwandari, Dr Mark J Weal, Dade Nurjanah, Prof Dame Wendy Hall, Hazzaz SM Imtiaz, Dr Kieron O'Hara, AU YEUNG Ching Man Albert
8	Alex J Frazer, Charles A Hargood, Patrick McSweeney, Paul Andre, Marcus M Ramsden, Tim Brody, David C Tarrant, Clare Owens, Dr Kirk Martinez, Daniel A Smith

TABLE 9.12: Clusters created based on the WebFest dataset with the ontology

Similar to the LSL dataset, we analysed the interests and their weights for each group to see if the groups have any themes. Given that the participants of this dataset are all specifically interested in the the World Wide Web and the Semantic Web, these two interests appear on the top of the interests list of each group with hight weights. When we applied the ontology inferences, the concepts that are broader than Semantic Web and WWW in the classification, such as Information systems, also had very high weights for each group. Therefore, it is more meaningful to ignore these interests in this study, as they form the theme of the entire dataset rather than just being a theme for a group or two.

Without the Semantic Web and WWW interests, the results of analysing the interests' themes for each group are not as clear as the ones we obtained from the LSL dataset. As tables 9.13 and 9.14 show, not all groups have clear themes. But the results after implementing the semantics of the interests are still better than the clustering results before the ontology. Table 9.13 shows that before the ontology, the only groups that seem to have a theme are: group 7, where the interests seem to be generally related to the ontologies and Semantic Web knowledge representation standards; and group 3 that seem to be interested in some programming languages (although when clustering, the algorithm is not aware of the fact that these interests are programming languages).

Table 9.14 shows that after the inferences, more themes emerged for some groups. For example, group 6 has participants generally interested in artificial intelligence. Group 3 is interested in knowledge representation related to the Semantic Web, Group 7 is interested in Web 2.0, and Group 8 is interested user interfaces. These results confirm that using the ontology allowed a better representation of the participants' interests, which with the clustering allowed the the participants to share more interests within the groups.

	Top interests in order of their weight
Group 1	semantic web (4), distributed Systems (2), social networking (2), security (2), pervasive computing (2)
Group 2	semantic web (5), rdf (3), hypertext (3)
Group 3	World Wide Web (4), linux (4), ajax (4), hci (3), perl (3), php (3), semantic web (3)
Group 4	semantic web (3), pervasive computing (3), agile methods (2), mobile computing (2), e-learning (2)
Group 5	semantic web (7), web science (2), ontology (2), e-learning (2), recommender systems, web 2.0 (2)
Group 6	semantic web (7), pervasive computing (4), artificial intelligence (3), multi-agent systems (3), linux (3)
Group 7	semantic web (6), ontologies (6), rdf (4), owl (3), web 2.0 (3)
Group 8	narrative (3), hci (2), semantic wiki (2), web science (2), hypertext (2)

TABLE 9.13: Interests weights for the WebFest dataset groups before implementing their semantics

	Top interests in order of their weight
Group 1	distributed systems (3.5), e-learning (3.5), computer uses in education (4) , user interfaces (3)
Group 2	software (3.16), hypertext/hypermedia (2.08), network architecture and design (2.33)
Group 3	knowledge representation formalisms and methods (15.04), artificial intelligence (11.17), ontology (6.82), owl (4)
Group 4	ubiquitous computing (3)
Group 5	software engineering (2.66), web 2.0 (2.75)
Group 6	distributed artificial intelligence (4), agent-based computing (3.5), artificial intelligence (2.84)
Group 7	web 2.0 (8.74), web science (5.5), social network (4.83)
Group 8	user interfaces (8), multimedia (6), hci (5), programming languages (3.58), personal computing (3.03), eprints (3)

TABLE 9.14: Interests weights for the WebFest dataset groups after implementing their semantics

9.4.2.2 Participants' satisfaction

Similar to our methodology of comparing the results obtained from the clustering with and without the semantics with the LSL dataset, we asked each participant to pick up to 8 people from the list of participants that they think share their interests. As with the LSL dataset, the maximum number of people to choose from the list is related to the number of members a group might have after the clustering. Many of the participants chose less than 8 people from the list (an average of 6 people for each participant).

The participants' responses represent the people they would choose to be in their group if the grouping was based on similarity of interests. The questionnaire used in this user study is available in appendix E. We calculated the participants' satisfaction with the two sets of clusters C1 and C2, where C1 is the set of clusters we created without the aid of the ontology in the previous chapter, and C2 is the set of clusters that we created with the aid of the ontology in this chapter.

The individual satisfaction of each participant is calculated as a ratio of the number of people in their chosen list who are in their clustering generated group to the number of people they have chosen (maximum 8, given that k=8, and the dataset has 61 participants). We then calculated the groups' satisfaction based on the average and standard deviation of the individuals' satisfactions within the groups. We finally compared the groups' satisfaction of the clustering with the ontology to the groups' satisfaction of the clustering without the ontology.

Out of the 61 participants in the dataset, 40 replied, 31 of them had their individual satisfaction increased in comparison to their satisfaction with the non-semantic clustering, 5 did not have their satisfaction changed, and 4 had their satisfaction decreased. Figure 9.6 shows the individual satisfaction percentage of change in comparison to clustering without semantics. Table 9.15 shows the average and the standard deviation of the individual satisfaction with the clustering results for both C1 (clustering without semantics), and C2 (clustering with semantic). As shown in the table, similar to the LSL dataset, the average satisfaction of the participants with the groups generated by the semantic-based clustering is much higher than their satisfactions with the groups generated by the non-semantic clustering from the previous chapter. In fact, the average satisfaction has more than doubled thanks to the implementation of inferences using the ontology.

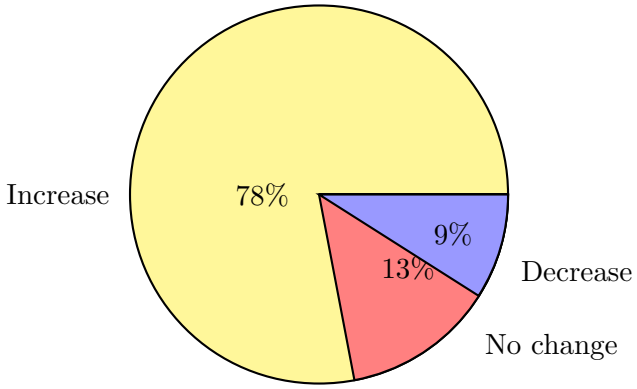


FIGURE 9.6: WebFest dataset individual satisfaction change in comparison to clustering without semantics

	C1 grouping	C2 grouping
Average Satisfaction	0.158	0.403
Standard Deviation	0.174	0.258

TABLE 9.15: Individual satisfaction with the group allocations for the WebFest dataset: Groups formed with clustering "C1" & Groups formed with clustering + the ontology "C2"

Table 9.16 shows the results of the groups' satisfaction and the cohort satisfaction. From the results in the table, we can see that similar to our findings with the LSL dataset, the cohort is much more satisfied with the semantic-based group allocations, proving that using semantics to enhance the results of the clustering generated better results than just clustering the data as it is without inference. In fact, our results show that the average satisfaction for the cohort when implementing semantics is 3 times the satisfaction without the semantics. We noticed that the standard deviation for the groups' satisfaction and the cohort satisfaction was slightly lower before implementing the semantics, however, the average group satisfaction in this case was very low, showing that most participants for each group agree that their satisfaction is low. Therefore, the fact that the standard deviation is higher when using semantics is not a negative factor.

9.5 Discussion

In this chapter, we studied the impact of using Semantic Web domain ontologies on the results of a simple clustering algorithm such as K-means to form groups of participants. We showed that adding semantics to the dataset resulted in creating better groups, from the participants' satisfaction perspective to say the least. This is because the ontology

		C1 grouping	C2 grouping
Group 1	Average	0.156	0.542
	Std Dev	0.187	0.193
Group 2	Average	0	0.226
	Std Dev	0	0.099
Group 3	Average	0.148	0.636
	Std Dev	0.122	0.207
Group 4	Average	0.056	0.233
	Std Dev	0.089	0.251
Group 5	Average	0.132	0.2
	Std Dev	0.114	0.14
Group 6	Average	0	0.65
	Std Dev	0	0.15
Group 7	Average	0.278	0.568
	Std Dev	0.123	0.156
Group 8	Average	0.405	0.486
	Std Dev	0.193	0.186
Cohort	Average	0.147	0.443
	Std Dev	0.139	0.191
	Avg Dev	0.104	0.173

TABLE 9.16: Group satisfaction with the group allocations for the WebFest dataset: Groups formed with clustering "C1" & Groups formed with clustering + the ontology "C2"

inferred more links between the participants that would not have been discovered otherwise. Through this mechanism, it put the interests' keywords into context. For example some of the participants such as *"Areeb"* had few interests but she was linked to so many participants with weight=1, so in terms of k-means clustering, she would have been equally well allocated to any group with any of her adjacent participants.

However, after the ontology's inference, the weights changed, and the semantics of the interests' keywords played a role such that two keywords with the same broader-class increased the weight of that broader-class creating a context of these keywords, and enabling *"Areeb"* to be allocated with participants more similar context, in other words, participants with similar weights and similar broader-classes. This factor provided a better recommendation of group allocations to each participant in the dataset. Figure 9.7 shows the difference to the cohort satisfaction for both the datasets we used: the LSL, and the WebFest dataset.

Given the datasets used in this study, we showed that clustering can handle larger datasets in less time than the constraint satisfaction approach we discussed in chapter

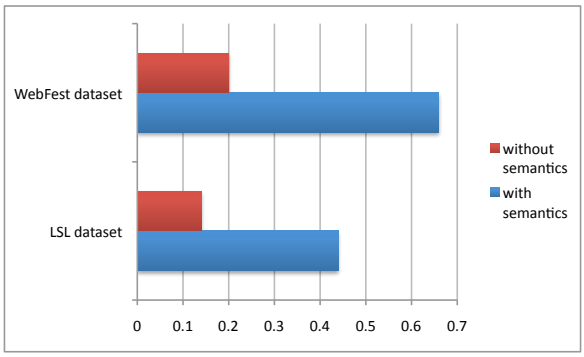


FIGURE 9.7: Increase of cohort satisfaction with the implementation of the interests' semantics

7. With clustering, we only had one constraint: grouping participants based on the similarity of their interests. When we added the ontology to perform inferences on the set of interests' keywords. Although the ontology contains few hundreds concepts that adds to the size of the computation (the data about each person increased in size), the clustering algorithm runs almost as fast as without adding the semantics. The algorithm still processed the data very quickly, for only 2 minutes with the LSL data, and 3 and a half minutes with the WebFest data, which is by far more efficient than the optimisation approach discussed in chapter 7 in terms of run time.

Although clustering algorithms do not guarantee optimal solutions, the group formation we obtained from the clustering when adding semantics were satisfying. If we had run the same datasets with the constraint satisfaction solver discussed in chapter 7, the solver would have run out of time with the size of the datasets before it would have found the optimal solution. We run the two experiments with a small number of participants from overlapping groups to prove the impact of adding semantics to the clustering approach. The obtained results proved for each experiment that the change in the participants' satisfaction is clear. Although the participants of the experiments overlapped, the results from only one of the experiments would have been enough to show the impact of the ontology on the clustering results.

9.6 Summary

Ontological knowledge structures play an important role in the quality of the formed groups, in that it adds meaning to the relationships between participants through the

use of inference of data that would not be otherwise be used by the clustering algorithm. In our studies we used one ontology to add semantics to the list of ECS interests, so that groups can be formed or potential collaborators could be recommended.

The clustering approach we followed in this chapter and the previous one is based partitioning the data into non-overlapping groups. A simple change to the algorithm or an implementation of another heuristic approach such as community discovery algorithms can create overlapping groups or communities that could also benefit from adding semantics to the dataset. Overall, regardless of the algorithm, implementing ontologies or other knowledge representation approaches that enrich the datasets used for allocating participants to groups is significant to discovering new connections between the participants, which is visible in the networks we generated with the ontology. These new connections allow better findings of computer supported group formation approaches.

Chapter 10

Conclusions and Future Work

This thesis provided a proof of concept that modelling the semantics of forming groups through the knowledge representation of the data about the participants and the constraints of the group formation enhances the quality of the groups obtained. In this chapter, we reflect on the achieved objectives and findings. We then conclude with possible follow-up work and further research directions.

10.1 Research justification

In this thesis, we used 2 approaches in forming groups for education:

The first approach was to model group formation as a Constraint Satisfaction Problem that aims at finding the optimal grouping of students. In this approach, we aimed at modelling the semantics of the teachers collaboration goals. We modelled these constraints as strong and weak Datalog constraints with different weights depending on their importance in the collaboration, an approach close to reality. We implemented this model in a tool that is capable to process multi-dimensional group formation, in other words, a group formation with many constraints on many variables such as the students' demographic and performance data. This application was evaluated against the constraints satisfaction and violation, where we used an evaluation framework to assess

the quality of the groups. The results showed that with example real data, on normal size classes of students in higher education, this approach performs well and returns an optimal solution is a short time (a maximum 3 minutes with 12 constraints). This performance is better than many existing group formation applications in educations. This approach processes the data and returns the optimal solution. However, although performs very well on normal size classes, this approach falls short on performance when the dataset is larger. Looking for optimal grouping consumes time, sometimes too much time to the point that the standard computer can not handle, and the process halts.

The second approach was to model group formation in a more heuristic way using a simple clustering algorithm. From this approach, we aimed at forming groups from larger datasets, datasets that CSP approach could not handle. We run the clustering on the dataset that the CSP failed to process, and the results showed that clustering as expected returned a solution in a matter of seconds, but the solution might not have been optimal. We wanted to show that if we enrich the participants' data with semantics, we can improve the results of clustering. We implemented a Semantic Web domain ontology, which we evaluated with an expert review, and run it with the same datasets scraped from the Web. The results of a user study showed that the users' responses to whom they are more similar to are more satisfied when implementing the semantics. In fact, this satisfaction was more than doubled, while the processing time didn't. We did not run the clustering approach with classes of students as it is expected to return similar results to existing group formation applications. We were only interested in a heuristic approach when we know the CSP model would not deliver a solution. Here, it is about the size of the dataset.

From both approaches, we proved that adding semantics deliver considerably better results than without the semantics, which is the aim of this thesis.

10.2 Research findings

In this section, we want to summarise the research findings, and add few points of discussions regarding the results of this work. In this thesis, we have demonstrated the following:

- Multi-dimensional group formation can be modelled as a CSP with strong and

weak weighted constraints in Datalog, and performs well or average size classes of students. An optimal solution can be obtained from this approach the average size dataset.

- Results from group formation obtained from clustering algorithms can be enhanced with the use of Semantic Web domain ontology when describing the participants.
- Semantic Web ontologies as knowledge representation mechanism we be built up on existing ontologies, which encourages the re-use of data.
- Group formation can be evaluated is different ways depending on the objectives of the collaborative activity wether through calculation constraints satisfaction for the criteria chosen by the teacher, or the learners' satisfaction and outcome (performance).
- Semantic group formation automation approaches for education can be helpful to both forming groups of students for cooperative or collaborative learning, and to finding new links between the participants, such as researchers in higher education, to recommend new potential collaborators.

In relation to social learning, a teacher can chose between a CSP based group formation mechanism, and a clustering based mechanism. Grouping for cooperative learning with it's more instructive nature might be better modelled as a CSP. This is because teacher designing a cooperative learning activity tend to have many criteria on how the groups should interact and achieve. A teacher aiming for a collaborative learning activity, with its more self-selected nature, might relax the constraints, and choose a mechanism that can model similarity or distance between the participants well. A clustering approach with added semantics on the participants' data can be more appropriate for collaborative learning.

Clustering algorithms are widely used in community discovery as discussed in the literature review. Therefore, a clustering algorithm with added semantics can be implemented in community discovery research for better recommendation of community membership. In this thesis, we used the interests attribute to show that domain ontologies can enhance group formation, but we can generalise that on any attribute that describes the participants' data, inference of new links through ontologies will provide better results.

As mentioned earlier, the use of semantics can add a lot to education, wether in describing the people or in describing the criteria of the groups. Semantics and Knowledge representation provide a better explanation of real life just as seen in both modelling the constraints and the interests of the participants in this thesis. Due to this fact, semantics provide better solutions, as shown in the user study of our last chapter, and sometimes optimal solutions as shown in the CSP approach to group formation. The disadvantage of using semantics is the added complexity to the problem as seen in the CSP approach. Obviously when adding semantic to the clustering approach, the algorithm took longer than without the semantics as there is more data to process. Only, the processing time was not much visible, but it can be if, for example, the dataset had 1000 participants.

Another limitation is the domain dependency. Domain ontologies are usually not a favourite to write by engineers, and are usually better automatically generated, for example, from folksonomies. Of course, an automated ontology would not match the quality of an ontology written by a knowledge engineer and agreed upon with a number of experts in that domain. The manual writing of domain ontologies for a specific application defies the aim of the Semantic Web, and automation in general. However, the more domain ontologies there are, the less we have to write, as one can only pick the one they want to use for a specific application. This is why in this research, we used the ACM classification, and the ECS ontology, both existing vocabularies, and just created an ontology to map them. The re-use of ontologies is regarded as a good practice that demonstrates the original vision of the Semantic Web.

10.3 Future work

In relation to the topics covered in this thesis, there is a number of points that can be further researched:

More algorithms: In this thesis, we used both CSP and clustering to form groups of students, but the aim of the research was to study the effects of adding semantics to improve group formation. In the future, other optimisation and heuristic algorithms including other implementations of CSP and clustering algorithms can be implemented to further study the impact of implementing semantics. Due to the nature of most education related group

formation, we focused on overlapping groups, a type of grouping also common in organisations. For future research, we can widen research to include overlapping grouping, a more common grouping to research and communities of practice within social networks. Many algorithms can be tested with semantics, and implemented into the tool we developed for CSP based group formation. Another future piece of work, would be to enhance the tool to return a confidence in the groups generated, and improve the user interface based on a user evaluation.

More Web-based data: In this research, we used the researchers' interests as an attribute to form groups, and mainly discover new connections that can be used in recommending collaborators. For future research, we can study another common attribute in describing researchers: their publications, and the keywords harvested from their publications. Unlike the networks generated from the interests' datasets used in this thesis, co-authorship networks follow a power law distribution, where the hubs are researchers who authored the most, and leaf nodes or new joiners might benefit more from recommendations. Whereas in the interests' networks, the most connected people are the ones that either have very common research interests, or described themselves with an above average number of keywords, or both. Studying co-authorship networks might generate different results in terms of semantics, although might generate less interesting new links.

In addition to this, we can add more inference rules to test other attributes in education, such as inferring learning styles and personality types from data in the Web. This will involve more data mining, and a probability framework that can set a confidence in the rules. For example, we can set the following rule: if participant A is good is a captain of a football team, and A has has founded a group for fellow colleagues in a social networking site, then A is a leader; and set a weight that we are 70% confident that rule applies. Mining data and inferring behaviour is a growing topic now that social networking sites provide a lot of this type of data. For example, Singla and Richardson (2008) found that chatting in social networks and personal behaviour are correlated, in that people who chat more with each other are more similar in more than one level, such as interests, age, and location.

Furthermore, more domain ontologies to describe students and researchers in education can be added to provide more inference on such data available from the Web. This can include recommendation for collaboration rather than just non-overlapping group formation, such as the inference rules used nowadays in dating sites. This can make use of the large amount of data provided by students in websites, forums and social networking sites for the benefit of education.

Incomplete data: As future extension to this work, we can research, for different algorithms, how does an algorithm for group formation performs with incomplete data. To do this, we can study the quality of the groups generated against a number of factors such as students' satisfaction and goal satisfaction, We study the quality of the groups as we add more and more semantics. For example an ontology that is only 3 levels is less complete with the same ontology with 4 levels in the hierarchy. We can run each algorithm with many datasets different in size and context and different inference rules, to discover at what point does the completeness of data stop making a big difference in the algorithm's performance. This can then help researchers predict how good an algorithm will perform based on the datasets they have, and therefore, have a confidence in the algorithm's results.

10.4 Summary

In this thesis, we have research the topics of forming groups in education from different angels, including literature on the theories of social learning that provided a foundation of collaborative and cooperative learning, the motivation behind forming groups. We researched the different types of groups and the different ways of forming groups including the different algorithms that have been used up to the date to automate the process of allocating people to groups. But the aim of out thesis was to show that the process and results of forming groups can be enhanced if we implement the semantics of the participants' data and the criteria set to form the groups. We proved using an evaluation framework and user studies that this is the case, that semantics of the grouping constraints and participants' data does improve the generated groups. The results and the research covered in this thesis

can be used as a foundation base to more research to come on both computer supported collaboration or education, by the researchers, or any researcher in this fields.

Appendix A

Observational Study: Questionnaire 1

All information that you provide on this questionnaire will be kept strictly confidential and will have absolutely no effect on your grade. This questionnaire will be handled anonymously as all the information you provide will be mainly used as a part of a PhD research on the improvement of students group formation. For further information about the project that this questionnaire relates to, please contact me Asma Ounnas at the Learning Societies Lab, Building 32, level 3, room 3069, Highfield, or via e-mail at ao05r@ecs.soton.ac.uk

Group # (Please specify your group number)

1. Demographic data:
   (a) Please delete as appropriate:
      - Gender: Female Male
      - Age: younger or aged 22 older than 22
   (b) Please specify your first spoken language:
   (c) Experience: Circle the rate that best describes your knowledge before you started the course from 1 to 5 (1 being no previous experience and 5 being very experienced) in the following
      - Technical/creative teamwork: 1 2 3 4 5
      - Software engineering: 1 2 3 4 5

2. The Belbin Self-Perception Inventory The following is from Belbins original work on team roles as appeared in his book(Belbin, 2004). For each of the following sections, distribute ten (10) points among the 8 sentences that you think best describe your behaviour. These points may be distributed among several sentences: in extreme cases they might be spread among all 8 sentences or ten points may be given to a single sentence. Enter the points in the spaces in front of each sentence. For example, for section 1, you might give five points to statement 2, two points to each statement 4 & 5, and one point to statement (Suggestion: Read all of the sentences, crossing out the ones that are not true or hardly true, then distribute points among those sentences left.)

   I. What I believe I can contribute to the team:

   1. I think I can quickly see and take advantage of new opportunities.
   2. I can work well with a very wide range of people.
   3. Producing ideas is one of my natural assets.
   4. My ability rests in being able to draw people out whenever I detect they have something of value to contribute to group objectives.
   5. My capacity to follow through has much to do with my personal effectiveness.
   6. I am ready to face temporary unpopularity if it leads to worthwhile results in the end.
   7. I can usually sense what is realistic and likely to work.
   8. I can offer a reasoned case for alternate courses of action without introducing bias or prejudice.

   II. If I have a possible shortcoming in teamwork, it could be that:

   1. I am not at ease unless meetings are well structured and controlled and generally well conducted.
   2. I am inclined to be too generous towards others who have a valid viewpoint that has not been given proper airing.
   3. I have a tendency to talk too much once the group gets on to new ideas.
   4. My objective outlook makes it difficult for me to join in readily and enthusiastically with colleagues.

5. I am sometimes seen as forceful and authoritarian if there is a need to get something done.
6. I find it difficult to lead from the front, perhaps because I am over responsive to group atmosphere.
7. I am apt to get caught up in ideas that occur to me and so lose track of what is happening.
8. My colleagues tend to see me as worrying unnecessarily over detail and the possibility that things may go wrong.

III. When involved in a project with other people:

1. I have an aptitude for influencing people without pressurising them.
2. My general vigilance prevents careless mistakes and omissions being made.
3. I am ready to press for action to make sure that the meeting does not waste time or lose site of the main objective.
4. I can be counted on to contribute something original.
5. I am always ready to back a good suggestion in the common interest.
6. I am keen to look for the latest in new ideas and developments.
7. I believe my capacity for judgment can help to bring about the right decisions.
8. I can be relied upon to see that all essential work is organised.

IV. My characteristic approach to group work is that:

1. I have a quite interest in getting to know colleagues better.
2. I am not reluctant to challenge the views of others or to hold a minority view myself.
3. I can usually find a line of argument to refute unsound propositions.
4. I think I have a talent for making things work once a plan has to be put into operation.
5. I have a tendency to avoid the obvious and to come out with the unexpected.
6. I bring a touch of perfectionism to any job I undertake.
7. I am ready to make use of contacts outside the group itself.
8. While I am interested in all views I have not hesitation in making up my mind once a decision has to be made.

V. I gain satisfaction in a job because:

1. I enjoy analysing situations and weighing up all of the possible choices.
2. I am interested in finding practical solutions to problems.
3. I like to feel I am fostering good working relationships.
4. I can have a strong influence on decisions.
5. I can meet people who may have something new to offer.
6. I can get people to agree on a necessary course of action.
7. I feel in my element where I can give a task my full attention.
8. I like to find a field that stretches my imagination.

VI. If I am suddenly given a difficult task with limited time and unfamiliar people:

1. I would feel like retiring to a corner to devise a way out of the impasse before developing a line.
2. I would be ready to work with the person who showed the most positive approach.
3. I would find some way of reducing the size of the task by establishing what different individuals might best contribute.
4. My natural sense of urgency would help to ensure that we did not fall behind schedule.
5. I believe I would keep cool and maintain my capacity to think straight.
6. I would retain a steadiness of purpose in spite of the pressures.
7. I would be prepared to take a positive lead if I felt the group was making no progress.
8. I would open up discussions with a view to stimulating new thoughts and getting something moving.

VII. With reference to the problems to which I am subject to working in groups:

1. I am apt to show my impatience with those who are obstructing progress.
2. Others may criticise me for being too analytical and insufficiently intuitive.
3. My desire to ensure that work is properly done can hold up proceedings.

4. I tend to get bored rather easily and rely on one or two stimulating members to spark me off.
5. I find it difficult to get started unless the goals are clear.
6. I am sometimes poor at explaining and clarifying complex points that occur to me.
7. I am conscious of demanding from others the things I cannot do myself.
8. I hesitate to get my points across when I run up against real opposition.

Appendix B

Observational Study: Questionnaire 2

Group # (Please specify your group number)

All information that you provide on this questionnaire will be kept strictly confidential and will have absolutely no effect on your grade. This questionnaire will be removed from your group report and handled anonymously. The information you provide will be mainly used as a part of a PhD research on the improvement of students group formation. For further information about the project that this questionnaire relates to, please contact me Asma Ounnas at the Learning Societies Lab, Building 32, level 3, room 3069, Highfield, or via e-mail at ao05r@ecs.soton.ac.uk

Please tick the number from 1 to 6 that best addresses the question. For questions 11 and 12 give comments:

1. Regardless of your teams plan for distributing tasks, rate the overall participation of the team members (Ideally everyone should contribute equally to the team. Do you feel every member has roughly contributed 1/6 of the effort?)

   not equally 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ equally

2. To what extend do you think leadership has emerged within your team?

   low 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ high

3. Regardless of whether you had a single leader:

a - How effective was the strategy your team used to allocate tasks to members?

poor 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ great

b - Was the quality of decision making in technical aspects effective?

not at all 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ very

4. How well do you think your abilities were used in the project?

not 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ very

5. How well were other team-members abilities used?

not 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ very

6. How strongly do you feel that you were included in the team?

weak 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ strong

7. How good was the mutual cooperation and helpfulness within your team? (if there was any)

poor 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ great

8. Was your teams approach or solutions conventional or innovative?

conventional 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ very innovative

9. How motivated was your team?

not at all 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ very

10. How willing would you be to work with this team again?

not at all 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ very

11. What is the worst aspect (or feature) of the team that caused the most problems (if any)?

12. What is the best aspect (or feature) of the team?

13. Rate how extensive was your experience (if you had any) in team working in technical or creative projects before taking this course

none 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ extensive

14. Rate how good you think you were in team work before taking the course?

    weak 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ great

15. Rate how good you think you are in team work after taking the course?

    weak 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ great

16. Rate your overall satisfaction with the teams outputs

    poor 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ great

17. Rate the overall quality of the teamwork regardless of its outputs

    poor 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ great

Appendix C

Observational Study: The Results

C.1 Study 1: Demographic Analysis

Experiment instrument: Questionnaire 1 in A

Study objectives: The aim of the questionnaire was to observe the demographics of students population taking a Software Engineering course[1] and their distribution across the groups. The demographics will be used in a simulation of the class to evaluate the research hypothesis and implementation. The questionnaires also aim to investigate the constraints used by the group formation initiators in the composition of the groups.

Methodology: To realise the objectives of the study, we investigated the parameters that can be considered in the formation of the Software engineering project groups. We have given the questionnaires to 9 groups out of 11 groups. Each group has 6 to 7 members that have already been formed based on the following two rules:

- No female can be alone in an all-male group.
- Students have to be evenly distributed across the groups based on their grades.

The students were assigned to groups based on their grades then manual swapping

[1]The course runs in the school of Electronics and Computer Science in the University of Southampton. Hence the results of the study are a typical computer science population in the UK higher education system.

for females found alone in a group was performed. The groups were formed by the course organisers in the first term for another course, that is 3 months before the beginning of the software engineering projects, which gave the students some time to know each other, the range of skills they have, and the strengths and weaknesses of each member.

The questionnaire, found in Appendix A, asks the students to provide information on the following: gender, age, first spoken language (to monitor home and international students), previous experience in software engineering previous experience in team work, and team roles by filling a Belbin self-perception inventory. The questionnaire was given to the students in the second term.

The questionnaires were anonymous and only the group number was recorded to keep track on the group dynamic in future study. The questionnaires were given to each group in their group meeting and were supervised by the research investigator. Unfortunately, few members of some groups dropped from the course, and some did not fill in the questionnaire, which reduced the number of members in some groups to 4 or 5 members.

The participating groups had the following demographics:

Group	Male	Female
1	5	0
3	5	0
4	5	0
5	7	0
6	4	2
7	4	2
8	3	2
9	5	0
11	2	2
Total	40	8

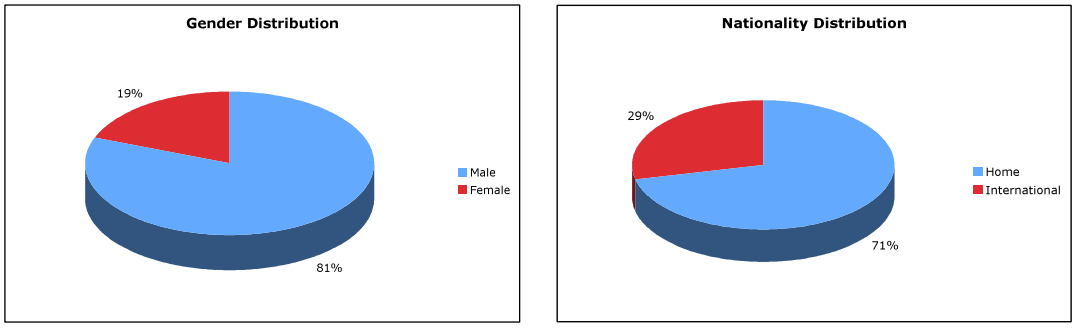
TABLE C.1: Observational Study: Gender distribution

In terms of these Belbin roles, to achieve the best team balance, there should be:

- One Coordinator or Shaper (not both) for leader
- A Plant to stimulate ideas
- A Monitor/evaluator to maintain honesty and clarity

Group	Home	International
1	4	1
3	5	0
4	2	3
5	6	1
6	2	4
7	5	1
8	4	1
9	4	1
11	3	1
Total	35	13

TABLE C.2: Observational Study: Nationality distribution



(a) Gender Distribution (b) Nationality Distribution

FIGURE C.1: Participants' demographics distribution

– One or more Implementer, Team worker, Resource investigator or Completer/Finisher to make things happen

Results: The outcome of the questionnaires responses were as follows:

– Gender distribution: as shown in figure C.1(a).

– Nationality distribution: as shown in figure C.1(b).

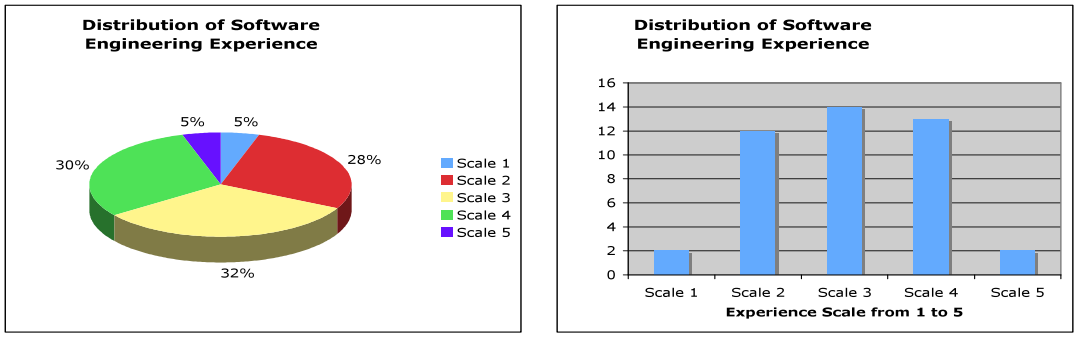
– Age: as the study showed that all the students who taking this course were 22 years old or younger, i.e. no mature students taking the course, it was not relevant to illustrate the distribution of age.

– Software engineering previous experience distribution: as shown in figure C.2, the distribution was on 1-5 Likert scale, which shows that the majority of the students are close to the mean 3.

– Teamwork previous experience distribution: as shown in figure C.3, the distribution was on 1-5 Likert scale, which shows that the majority of the students

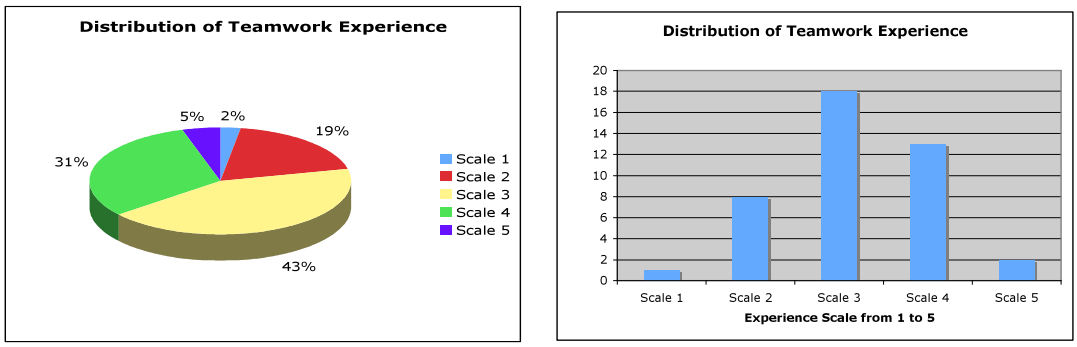
Role	Symbol	Typical features	Positive qualities	Allowable weakness
coordinator	CO	Calm, self-confident, controlled	A capacity for treating and welcoming all potential contributors on their merits and without prejudices. Strong sense of objectiveness	No more than ordinary in teams of intellect or creative ability
shaper	SH	Highly strung	Drive and readiness to challenge inertia, ineffectiveness, complacency or self-deception	Proneness to provocation, irritation and impatience
plant	PL	Individualistic, serious-minded, unorthodox	Genius, imagination, intellect, knowledge	Up in the clouds, inclined to disregard practical details or protocol
Monitor evaluator	ME	Sober, unemotional, prudent	Judgment, discretion, hard-headedness	Lacks inspiration or the ability to motivate others
Implementer	IM	Conservative, dutiful, predictable	Organising ability, practical common sense, hard-working, self-discipline	Lack of flexibility, unresponsiveness to unproven ideas
Completer finisher	CF	Painstaking, orderly, conscientious, anxious	A capacity for follow-through, perfectionism	A tendency to worry about small things, a reluctance to "let go"
Resource investigator	RI	Extroverted, enthusiastic, curious, communicative	A capacity for contacting people and exploring anything new. An ability to respond to challenge	Liable to lose interest once the initial fascination has passed
Team worker	TW	Socially oriented, mild, sensitive	Ability to respond to people and situations, and to promote team spirit	Indecisiveness at moments of crisis

TABLE C.3: Observational Study: Brief description of Belbin Roles(Belbin, 2004)



(a) Percentage (b) Number of responses

FIGURE C.2: Software Engineering Experience Distribution

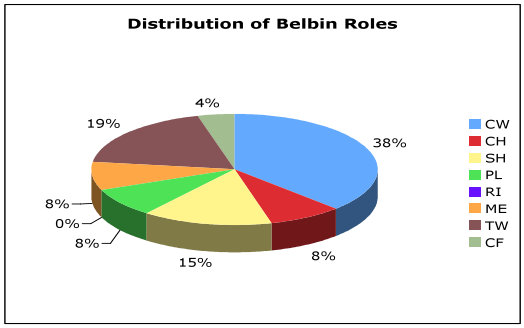


(a) Percentage (b) Number of responses

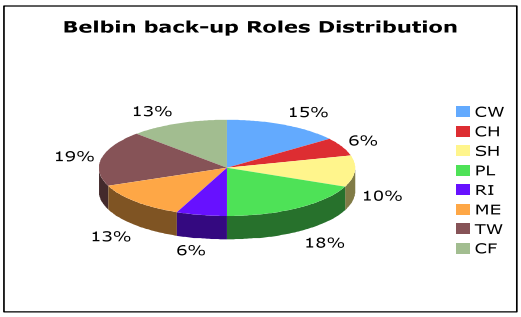
FIGURE C.3: Teamwork Experience Distribution

are close to the mean 3, but few of them ranked their experience as 1 or 2 on the scale.

– Belbin team roles distribution: as shown in figure C.4(a). Table C.4 shows the strong or primary roles available in each group. The numbers in the cells represent the Belbin role score calculated from the self-perception inventory. Only scores that are greater than or equal to 10 are considered. The highest scores (that exceed 20) are highlighted in bold in the table. As expected, most of the SEG students (population) are implementers (CW). Surprisingly, the study showed that there were no strong resource investigators (RI) in the population, this is probably due to the fact that the course organisers provide the students with a detailed specification of the project, so the students did not perceive themselves as recourses finders. The number of plants PL, which present creative members, is relatively low for a computer science population. Table C.5 shows the back up roles (second strongest roles) that a student can shift to during the course of the project. The back-up roles show a larger



(a) Primary roles



(b) Back-up roles

FIGURE C.4: Percentage distribution of Belbin roles

Group \ Role	CW	CH	SH	PL	RI	ME	TW	CF
1	12			13		**22**, 15	12	
4	16	16	12				13	13
3	15		15, 16			14	15	
5	12		15	**20**			**26**, 13, 17	14
6	31,**26**, 14, 26		14			16		
7	15, 17, **20**	18, 15	14					
8	13	13	15				16, 17	
9	19, 18, 19			12 ,18				
11	14, 14, 19						17	
Total	18	4	7	4	0	4	9	2

TABLE C.4: Observational Study: Belbin team role distribution for each group

Group \ Role	CW	CH	SH	PL	RI	ME	TW	CF
1	17, 13			12	11		12	
4	10	12	12	12, 12				
3					13	10, 11	13	13
5	13	14	11			12, 10	12	
6				11, 11			11	11, 17
7	12			14		13, 13	15	13, 13
8	12	14		12			12	14
9				14, 14	10		16, 13	
11	16		12, 11				14	
Total	7	3	5	9	3	6	9	6

TABLE C.5: Observational Study: Belbin back up team role distribution for each group

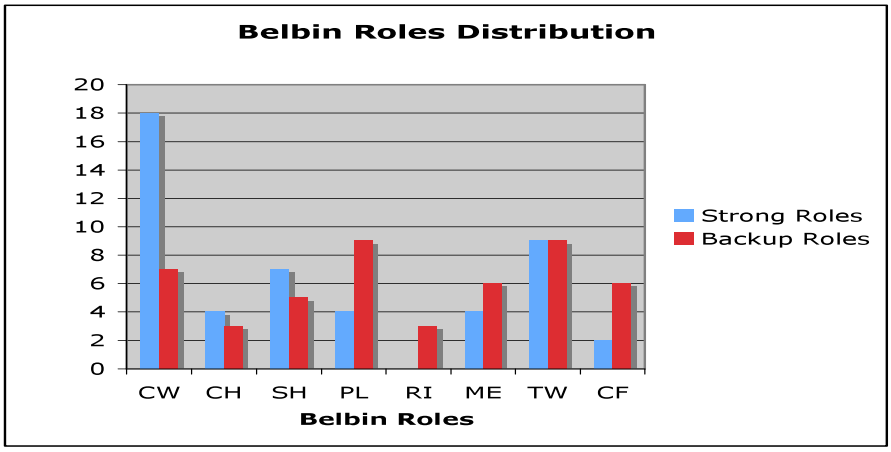


FIGURE C.5: Distribution of Belbin roles

number of plants and team workers in relation to strong roles. The secondary roles also show that there are 3 students who are resource investigators. Figure C.4(b) shows the distribution of back up roles, and figure C.5 compares them with primary roles.

C.2 Study 2: Evaluation

Experiment instrument: Questionnaire 2 in Appendix B

Study objectives: The questionnaire is more of an evaluation of the group formation carried by the course organisers this year in terms of the perceived teamwork success by the students. It has 17 questions, where the student is asked to rank the key elements that measure both the group performance, and the individual satisfaction of the group work.

Methodology: Questionnaire 2 has 17 questions. Each question is a statement that the student can rank from 1 to 6 (6 point scale) depending on how much they agree with the question. Questions 11 and 12 ask the student to give some comments on the bad and good aspects of the group respectively. The questions were designed to monitor the following aspects of group work:

- Q1: perceived contribution of individuals and fairness of task distribution among members.
- Q2, Q3: Members management, decision making and leadership.
- Q4, Q5: Use of team members skills.

- Q6, Q7: Team cohesion.
- Q8: Creativity.
- Q9: Motivation.
- Q13, Q14: Previous experience in team projects.
- Q15: Experience gained from the course in team projects.
- Q10, Q11, Q12: Evaluation of the team.
- Q16, Q17: Satisfaction with the team.

The questionnaires were given to the same data sample (all groups of SEG students). The questionnaires were given (online) by the course organiser for the students to fill in and attach in their final project reports of the course. Unfortunately, only 6 groups returned the questionnaires.

Results: We analyse the results of the questionnaire for each group given the answers of the students for each group. We relate the results to the ones collected from the previous one (the Belbin roles). The responses from the groups are as follows:

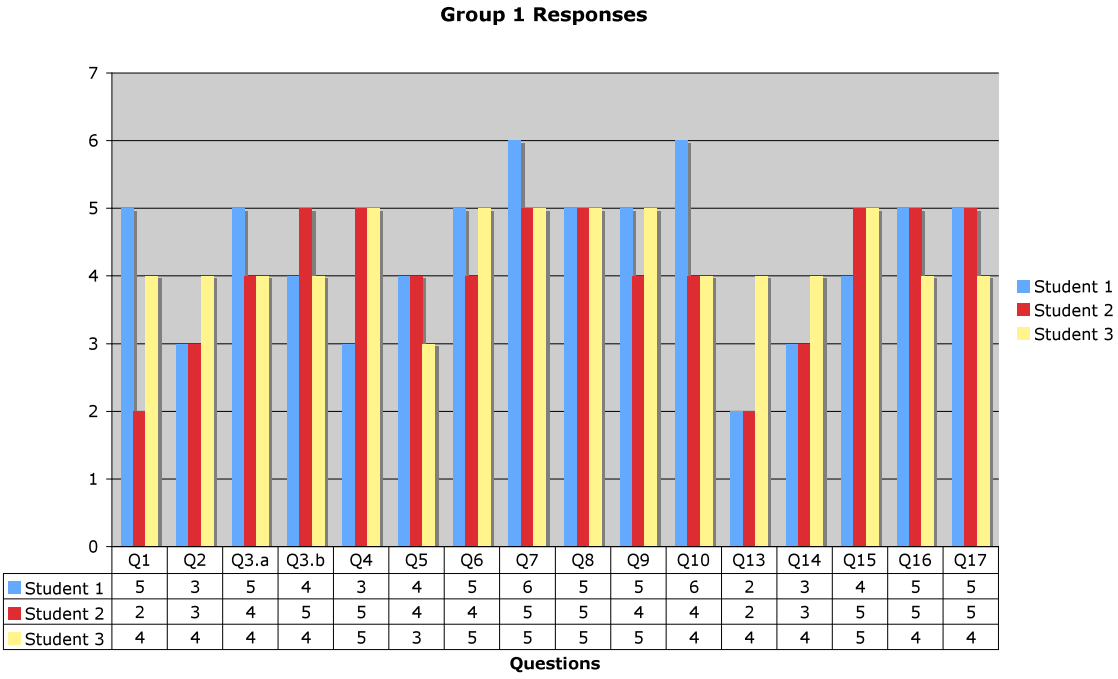
Group1: responses depicted in figure C.6

- Q11: bad distribution of workload, unexpected absence.
- Q12: very good mutual cooperation, good mixed range of skills, members got on well, constructive discussion, no unnecessary argument.

The group members seem to have different opinions on the equal contribution to the project as the standard deviation is quite high, which they also reported on question 11 as the worst aspect of the group. They also reported a low response to agreements on decision making and good use of members skills. The reason for that can be because the group didnt have any clear leader (also shown in the Belbin team roles) to coordinate the tasks. The group reported the highest perceived creativity in relation to other groups, which can be a consequence of having a strong plant in the group, unlike most other groups.

Group3: responses depicted in figure C.7

- Q11: lack of decisiveness on difficult issues, some members needed to be pushed in the right direction rather than being proactive, time was sometimes



	Q1	Q2	Q3.a	Q3.b	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q13	Q14	Q15	Q16	Q17
Student 1	5	3	5	4	3	4	5	6	5	5	6	2	3	4	5	5
Student 2	2	3	4	5	5	4	4	5	5	4	4	2	3	5	5	5
Student 3	4	4	4	4	5	3	5	5	5	5	4	4	4	5	4	4

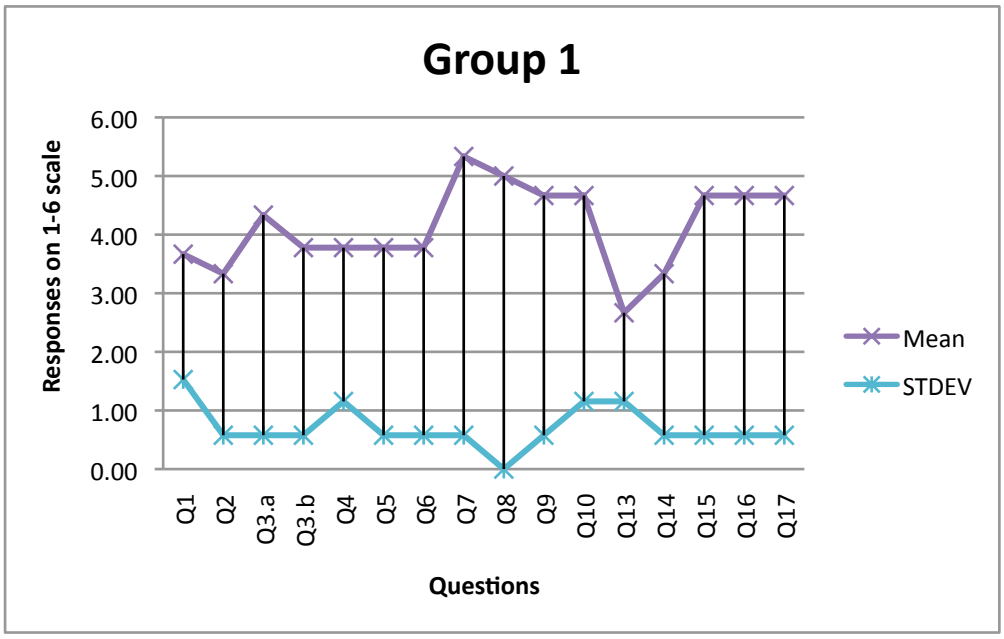
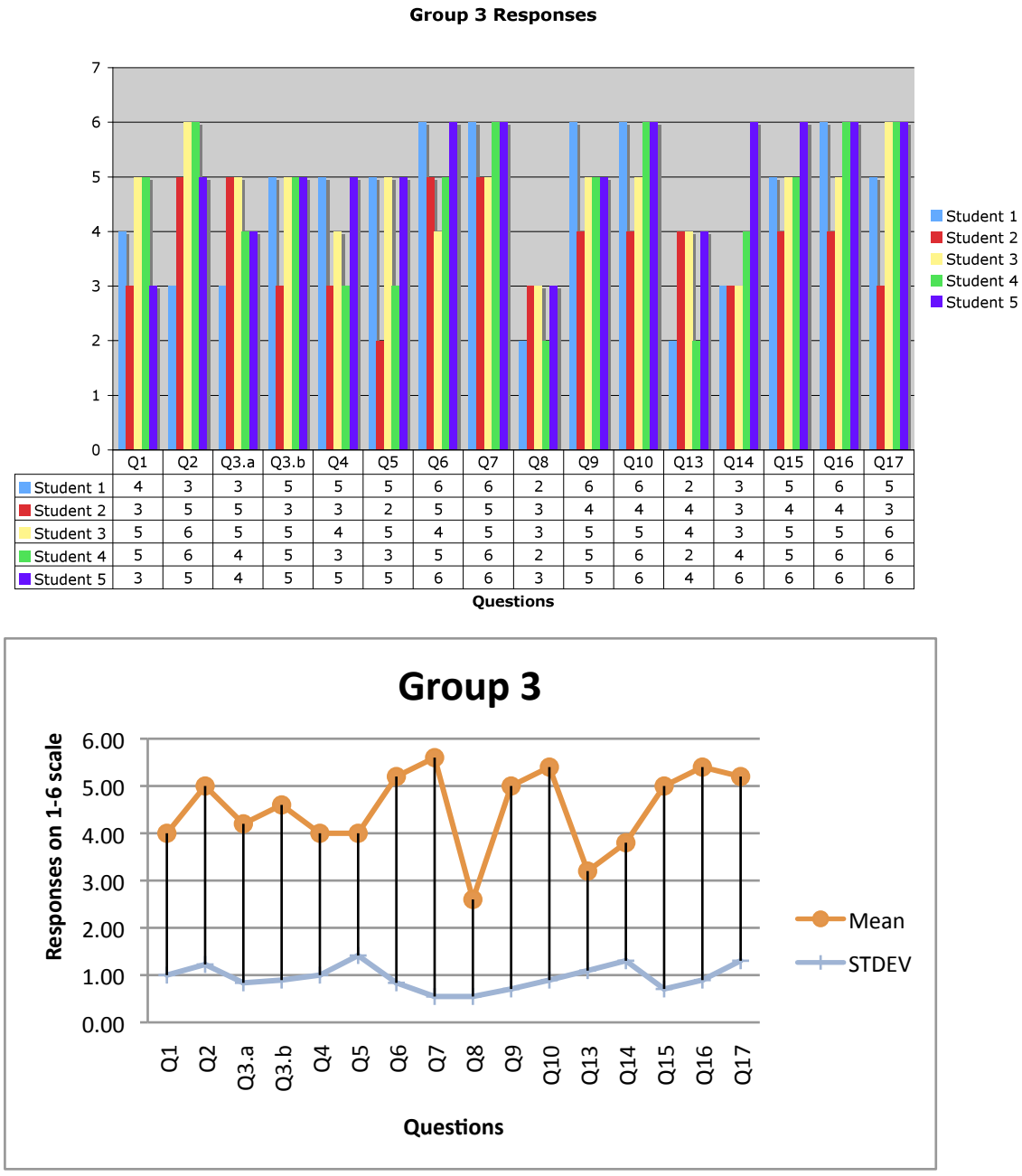


FIGURE C.6: Group 1 responses to questionnaire 2



	Q1	Q2	Q3.a	Q3.b	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q13	Q14	Q15	Q16	Q17
Student 1	4	3	3	5	5	5	6	6	2	6	6	2	3	5	6	5
Student 2	3	5	5	3	3	2	5	5	3	4	4	4	3	4	4	3
Student 3	5	6	5	5	4	5	4	5	3	5	5	4	3	5	5	6
Student 4	5	6	4	5	3	3	5	6	2	5	6	2	4	5	6	6
Student 5	3	5	4	5	5	5	6	6	3	5	6	4	6	6	6	6

FIGURE C.7: Group 3 responses to questionnaire 2

spent on trivial tasks, the group members were easy going which made it hard to get solid decisions, one of the members was not always around doing work when others were.

– Q12: relaxed atmosphere, good dynamic between all team members, constructive criticism to each others work, the members agreement on most decisions, and no disagreements on any major decision.

The group members seem to have different opinions on the way the teams abilities

were used. This is probably due to the fact that the equal contribution to the project -as the standard deviation- is quite high, which they also reported on questions 11 as the worst aspect of the group. The reason for that can be because the group didnt have any clear leader (also shown in the Belbin team roles) to coordinate the tasks. From Q8, the members perceived the creativity of their solution to be low. According to Belbin roles, creativity is associated with the plant team role which is missing from the group even in back up roles.

Group4: responses depicted in figure C.8

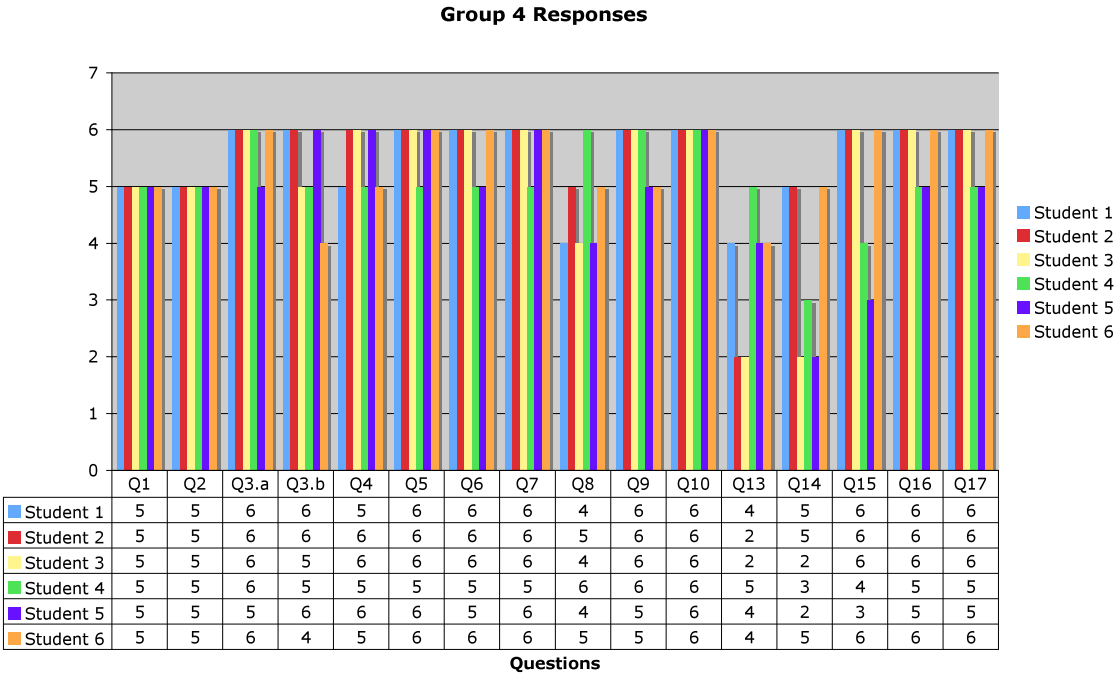
- Q11: none.
- Q12: the group members got on well and were not shy raising issues or asking questions to each other, helpfulness and trust, each individual had different skills which were used effectively, team members had the ability (motivation) to deliver without having to be asked to do so.

Group 4 members seem to agree on most responses to the questionnaire. However, they seem to have very different experiences in teamwork which may have contributed to the maturity of the members in delivering their tasks without having to be pushed as well as to the range of skills they have. From the first study, we notice that group 4 has a very good combination of team roles. Apart from monitor/evaluator, all the team roles are present as a primary or secondary role.

Group5: responses depicted in figure C.9

- Q11: conflict in ideas, some tasks was poorly allocated to members, time management problems because of some members not recording their contribution and efforts in allocated tasks.
- Q12: good communication, cooperation, teamwork, every member took the project seriously and contributed considerable efforts to the team, team members got on well and made decisions together, excellent diversity of skills.

Group 5 seem to have high responses to the emerging of leadership within the group, from questionnaire 1 we can see that the group has different strong leaders -both shapers and chairman- and a very strong plant, which might be the reasons for the conflicts reported in Q11. The group also has many team workers which might be the reason behind the high perceived cooperation and good communication of



	Q1	Q2	Q3.a	Q3.b	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q13	Q14	Q15	Q16	Q17
Student 1	5	5	6	6	5	6	6	6	4	6	6	4	5	6	6	6
Student 2	5	5	6	6	6	6	6	6	5	6	6	2	5	6	6	6
Student 3	5	5	6	5	6	6	6	6	4	6	6	2	2	6	6	6
Student 4	5	5	6	5	5	5	5	5	6	6	6	5	3	4	5	5
Student 5	5	5	5	6	6	6	5	6	4	5	6	4	2	3	5	5
Student 6	5	5	6	4	5	6	6	6	5	5	6	4	5	6	6	6

Questions

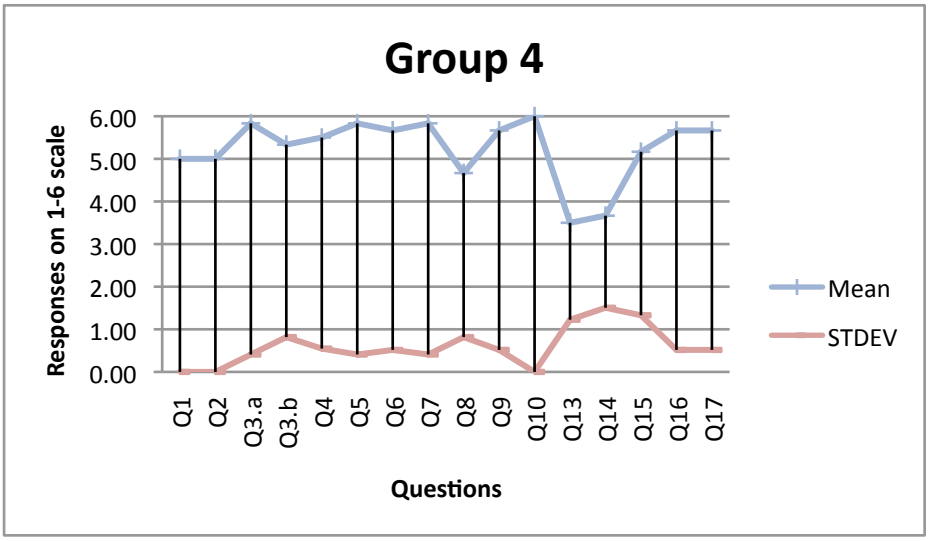
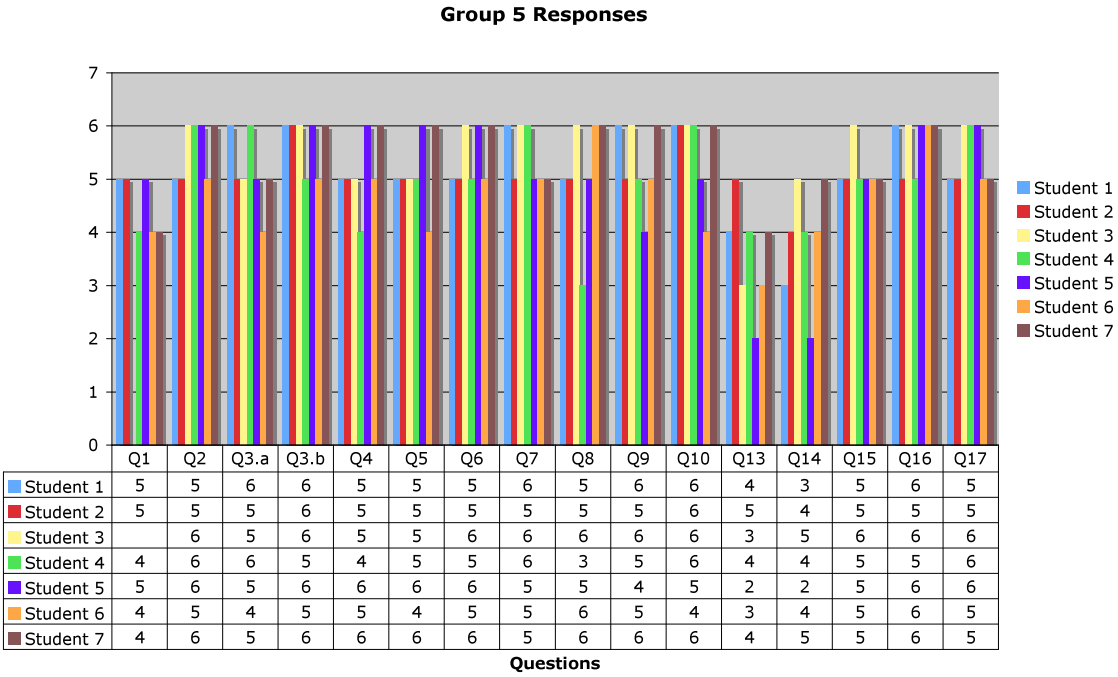


FIGURE C.8: Group 4 responses to questionnaire 2

the group. Because of the strong plant, the group reported a very high creativity comparing to other groups.

Group8: responses depicted in figure C.10

- Q11: one member not contributing to the project (this member was eliminated from the course), some luck of communication,
- Q12: members got on well, no major arguments, all members worked hard on their tasks.



	Q1	Q2	Q3.a	Q3.b	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q13	Q14	Q15	Q16	Q17
Student 1	5	5	6	6	5	5	5	6	5	6	6	4	3	5	6	5
Student 2	5	5	5	6	5	5	5	5	5	5	6	5	4	5	5	5
Student 3		6	5	6	5	5	6	6	6	6	6	3	5	6	6	6
Student 4	4	6	6	5	4	5	5	6	3	5	6	4	4	5	5	6
Student 5	5	6	5	6	6	6	6	5	5	4	5	2	2	5	6	6
Student 6	4	5	4	5	5	4	5	5	6	5	4	3	4	5	6	5
Student 7	4	6	5	6	6	6	6	5	6	6	6	4	5	5	6	5

Questions

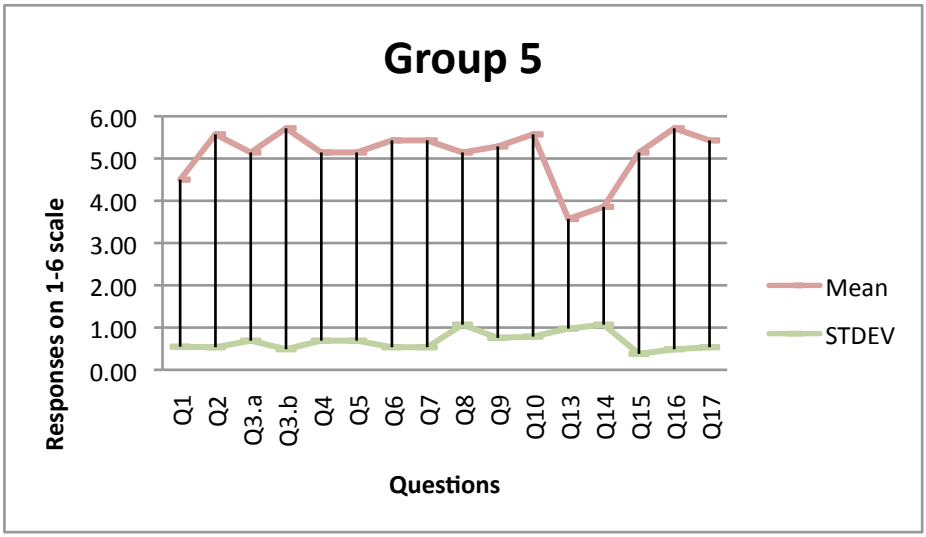
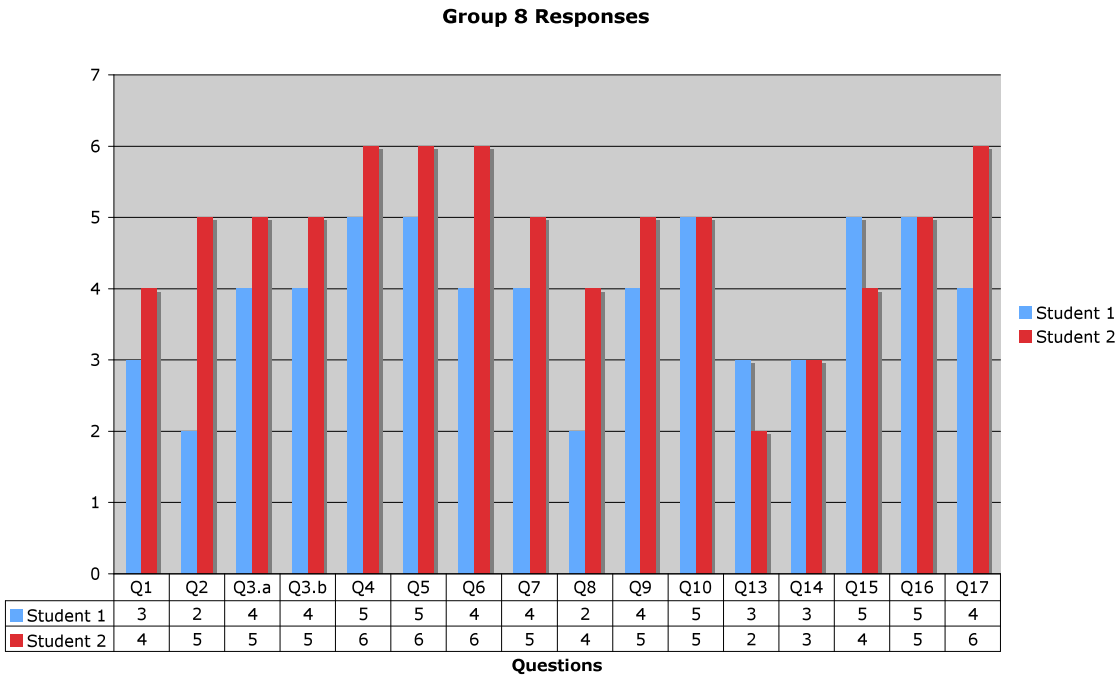


FIGURE C.9: Group 5 responses to questionnaire 2

Unfortunately, only 2 out of 5 responses were received from group 8, so the observations on the group cannot be highly reliable. In addition to that, the group had one member not contributing to the group at all, which can be the reason for the relatively perceived non equality of contribution to the project. This may explain the slight loss of motivation and the high use of the members skills to cover the expected contribution of the missing member. The group had relative leadership from the Belbin roles; however, responses to Q2 seem to show a difference in opinion of the students perception of emerging leadership. Moreover, although one of



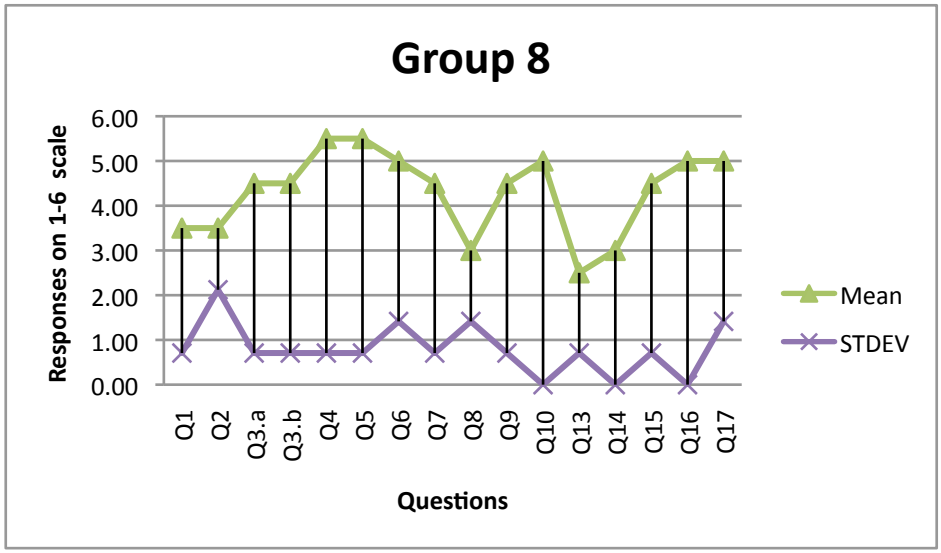
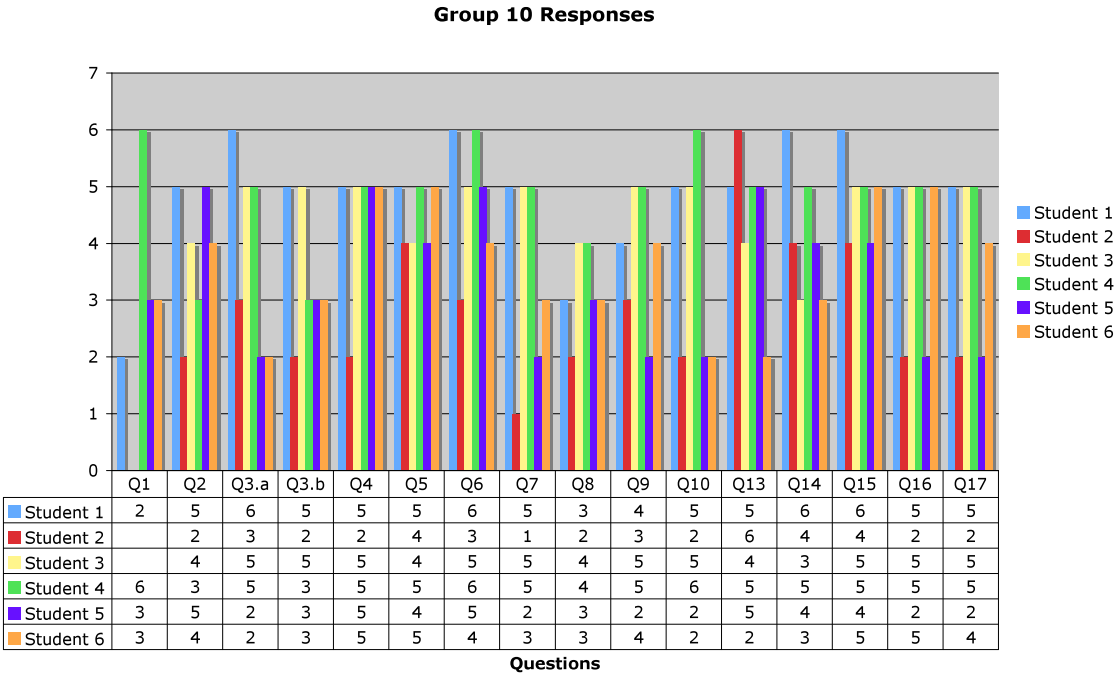


FIGURE C.10: Group 8 responses to questionnaire 2

the members was a plant as a backup role, the creativity of the solution is not as high as expected.

Group10: responses depicted in figure C.11

- Q11: differences in experience might have caused arguments in the team, conflict between 2 members throughout the project, lack of abilities in certain areas, and luck of motivation
- Q12: the members got along well, devotion of certain members



	Q1	Q2	Q3.a	Q3.b	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q13	Q14	Q15	Q16	Q17
Student 1	2	5	6	5	5	5	6	5	3	4	5	5	6	6	5	5
Student 2		2	3	2	2	4	3	1	2	3	2	6	4	4	2	2
Student 3		4	5	5	5	4	5	5	4	5	5	4	3	5	5	5
Student 4	6	3	5	3	5	5	6	5	4	5	6	5	5	5	5	5
Student 5	3	5	2	3	5	4	5	2	3	2	2	5	4	4	2	2
Student 6	3	4	2	3	5	5	4	3	3	4	2	2	3	5	5	4

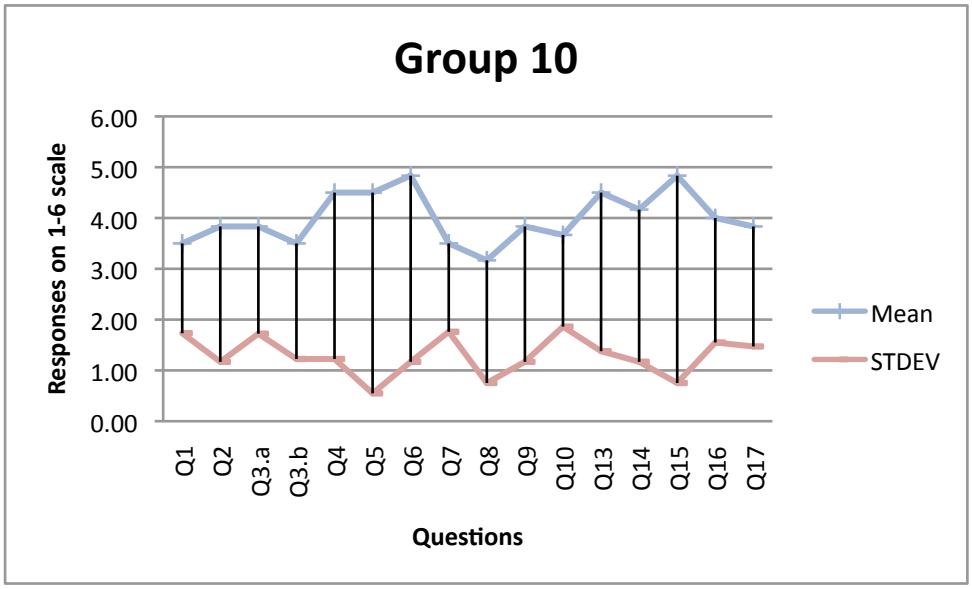


FIGURE C.11: Group 10 responses to questionnaire 2

The group members seem to have very different perceptions in many aspects of the teamwork such as satisfaction with the team, team cohesion and cooperation, leadership, and contribution. The members claim a conflict between some members that remained for the whole length of the project which can be the reason for the different perceptions. Unfortunately, the group did not participate in the first questionnaire, so correlation to the first study cannot be inferred. A prediction of the reason might be based on the idea that conflicts usually result from having more than one strong leader.

In general: The results from question 13, 14, and 15 on previous experience are very similar for all groups which also match the results of the first study. However, the satisfaction of the individuals with their groups is very interesting. The perceived perception for groups 4 and 5 seems consistent with all group members; however, for groups 8 and 10, the satisfaction is quite low, although the perceived team cohesion is high. This means that students perceive a satisfying team as a team with a good output rather than considering the dynamics of the group. Although groups 1, 5, and 8 had only one international student in them, isolation was not detected. This is because group 1 had conflicts which may have been the reason for their response to Q6. Group 8 didnt return all the questionnaires, and group 5 responded well to the question with a low standard deviation.

Appendix D

User Study: The LSL Dataset

Please pick up to 5 people from the following list, whom you think share the most interests with you:

Noura Abbas

Amit Acharyya

Areeb AlOwisheq

Kikelomo M Apampa

Ali Aseere

Will Davies

Andrew K Douglas

Alex J Frazer

Charles A Hargood

Athitaya Nitchot

Dade Nurjanah

Asma Ounnas

Clare Owens

Yulita Hanum P Iskandar

Rikki Prince

Till Rebenich

Onjira Sitthisak

David Argles

Hugh Davis

Quintin Gee

Lester Gilbert

Andy Gravell

Thanassis Tiropanis

Mike Wald

Mark J Weal

Su White

Gary Wills

Appendix E

User Study: The WebFest Dataset

Please pick up to 8 people from the following list, that you think share the most research interests with you:

Areeb AlOwisheq

Dr David Argles

Will Davies

E.A. Draffan

Alex J Frazer

Charles A Hargood

Patrick McSweeney

Dade Nurjanah

Bart Nagel

Asma Ounnas

Clare Owens

Rikki Prince

Marcus M Ramsden

Till Rebenich

Dr Thanassis Tiropanis

Dr Mark J Weal

Dr Gary B Wills

Dr Arouna Woukeu

Saad A Alahmari

Dr Harith Alani

Paul Andre

Peyman Askari

Albert (AU YEUNG Ching Man)

Philip R. Boulain

Tim Brody

Gianluca Correndo

Don Cruickshank

Dr Nicholas Gibbins

Prof Wendy Hall

Dr Jonathon S Hare

Hazzaz SM Imtiaz

Dr Kirk Martinez

Dr Danius Michaelides

Dr Timothy J Miles-Board

Gontlafetse Mosweunyane

Zurina Muda

Salma Noor

Alisdair Owens

Kevin R Page

Oleksandr Pryymak

Betty Purwandari

Dr. Sarvapali Ramchurn

Valentin Robu

Benedicto Rodriguez

Manuel Salvadores

Mark Schueler

Dr Ash Smith

Daniel A Smith

Dr Martin Szomszor

David C Tarrant

Robert A Vesse

EJ "Ed" Zaluska

Roushdat Elaheebocus

Pavithran Sakamuri

Yang Yang

Debra Morris

Dr Kieron O'Hara

Dr David Millard

Bibliography

A. L. A. Barabási and R. Albert. Emergence of scaling in random networks. *Science (New York, N.Y.)*, 286(5439):509–512, October 1999. ISSN 1095-9203.

L. A. Adamic, O. Buyukkokten, and E. Adar. A social network caught in the web. *First Monday*, 8, 2003.

H. Alani, S. Dasmahapatra, K. O'hara, and N. Shadholt. Ontocopi -using ontology based network analysis to identify communities of practice. *IEEE Intelligent Systems*, 18:18–25, 2003.

H. Alani, Y. Kalfoglou, K. O'Hara, and N. Shadbolt. Towards a killer app for the semantic web. In *4th International Semantic Web Conference (ISWC 2005)*, 2005.

M. Alavi. Computer-mediated collaborative learning: an empirical evaluation. *MIS Q.*, 18(2):159–174, 1994. ISSN 0276-7783.

B. Alexander. Web 2.0: A new wave of innovation for teaching and learning? *In Educause Review*, 41(2):32–44, 2006.

S. Ambroszkiewicz, O. Matyja, and W. Penczek. Team formation by self-interested mobile agents. In *Selected Papers from the 4th Australian Workshop on Distributed Artificial Intelligence, Multi-Agent Systems*, pages 1–15, London, UK, 1998. Springer-Verlag. ISBN 3-540-65477-1.

J.H.E. Andriessen, M. Soekijad, M. Huis in 't Veld, and J. Poot. Dynamics of knowledge sharing communities. Technical report, Telematica Instituut, Enschede, 2001.

L. Aroyo and D. Dicheva. The new challenges for e-learning: The educational semantic web. *Educational Technology & Society*, 7(4):59–69, 2004.

S. Auer and J. Lehmann. What have innsbruck and leipzig in common? extracting semantics from wiki content. *The Semantic Web: Research and Applications*, pages 503–517, 2007.

Fahiem Bacchus and Adam Grove. Looking forward in constraint satisfaction algorithms, 1999.

Francis R. Bach and Michael I. Jordan. Learning spectral clustering. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

M. Baker and K. Lund. Promoting reflective interactions in a computer-supported collaborative learning environment. *Journal of Computer Assisted Learning*, 13: 175–193, 1997.

A. Bandura. Social-learning theory of identificatory processes. In David A Goslin, editor, *Handbook of Socialization Theory and Research*, pages 213–261. Rand McNally & Company, 1969.

S. A. Barab, M. Barnett, and K. Squire. Developing an empirical account of community of practice: characterizing the essential tensions. *The Journal of the Learning Science*, 11(4):489–542, 2002.

A. Barabási. *Linked: The New Science of Networks*. Perseus Publishing, 2002.

B. Bateman, F. C. Wilson, and D. Bingham. Team effectiveness - development of an audit quastionnaire. *Journal of Management Development*, 21(3):215–226, 2002.

R. M. Belbin. *Management Teams: Why They Succeed Or Fail*. Elsevier Butterworth-Heinemann, second edition, 2004.

Kristin P. Bennett and Colin Campbell. Support vector machines: Hype or hallelujah? *SIGKDD Explorations*, 2, 2003.

J. Berg, L. Berquam, and K. Christoph. Social networking technologies: A "poke" for campus services. *EDUCAUSE Review*, 42(2):32–44, 2007.

T. Berners-Lee, W. Hall, J. Hendler, K. O'Hara, N. Shadbolt, and D.J. Weitzner. A framework for web science. *Found. Trends Web Sci.*, 1(1):1–130, 2006. ISSN 1555-077X.

T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web, 2001.

B. J. Biddle. *Role theory : expectations, identities, and behaviors*. Academic Press, New York, 1979. ISBN 012095950X.

D. Bligh. *What's the Use of Lectures?* Penguin Books, 1972.

J. Brase and M. Painter. Inferring metadata for a semantic web peer-to-peer environment. *Educational Technology & Society*, 7(2):61–67, 2004.

J.G. Breslin, A. Harth, U. Bojars, and S. Decker. *Towards Semantically-Interlinked Online Communities*. Springer-Verlag, 2005.

D. Brickley and L. Miller. *FOAF Vocabulary Specification -Namespace Document 27 July 2005 - ('Pages about Things' Edition)*. http://xmlns.com/foaf/0.1/, 2005. Accessed 12 May, 2009.

J.S. Brown and P. Duguid. *The Social Life of Information*. Harvard Business School Press, Boston, MA, USA, 2002. ISBN 1578517087.

K. A. Bruffee. Collaborative learning and the "conversation of mankind". *College English*, 46(7):635–652, 1984.

K. A. Bruffee. *Collaborative Learning: Higher Education, Interdependence, and the Authority of Knowledge*. Johns Hopkins University Press, 2715 North Charles Street, Baltimore, 1993.

J. S. Bruner. *Acts of meaning*. Harvard University Press, Cambridge, MA, 1990.

P. Brusilovsky and H. Nijhavan. A framework for adaptive e-learning based on distributed re-usable learning activities. In *Proceedings of World Conference on E-Learning, E-Learn 2002*, pages 154–161, 2002.

F. Buccafurri, N. Leone, and P. Rullo. Strong and weak constraints in disjunctive datalog. In *LPNMR '97: Proceedings of the 4th International Conference on Logic Programming and Nonmonotonic Reasoning*, pages 2–17, London, UK, 1997. Springer-Verlag. ISBN 3-540-63255-7.

D. Butler and C. Chatfield. 2008 educause annual conference. *Serials Review*, 35 (2):106–107, 2009.

J.J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A.Seaborne, and K. Wilkinson. Jena: implementing the semantic web recommendations. In *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 74–83, New York, NY, USA, 2004. ACM. ISBN 1-58113-912-8.

C. Cavanaugh. *Development and Management of Virtual Schools: Issues and Trends*. Information Science Publishing (an imprint on Idea Group Inc), 2004.

H. Cho, G. Gay, B. Davidson, and A. R. Ingraffea. Social networks, communication styles, and learning performance in a cscl community. *Computers & Education*, 49(2):309–329, 2007.

C. E. Christodoulopoulos and K. A. Papanikolaou. A group formation tool in an e-learning context. In *Proceedings of the 19th IEEE ICTAI'2007*, pages 117–123, 2007.

A. Clauset, C. Moore, and M. Newman. Structural inference of hierarchies in networks. *Statistical Network Analysis: Models, Issues, and New Directions*, pages 1–13, 2007.

A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008. ISSN 0028-0836.

P. Cohen, H. Levesque, and I. Smith. On Team Formation. In *Contemporary Action Theory. Synthese*, volume 2: Social Action, pages 87–114. Kluwer Academic Publishers, 1997.

C. Collison. Connecting the new organisation. how bp amoco encourages post-merger collaboration. *Knowledge Management Review*, 7, 2000.

Colm Conroy, Co)supervisors Declan O'sullivan, and David Lewis. Towards ontology mapping for ordinary people, 2009.

Marco Correia and Pedro Barahona. Machine learned heuristics to improve constraint satisfaction. *Advances in Artificial Intelligence SBIA 2004, Lecture Notes in Computer Science*, 3171/2004:103–113, 2004.

N. Coulter, J. French, E. Glinert, T. Horton, N. Mead, R. Rada, A. Ralston, C. Rodkin, B. Rous, A.Tucker, P. Wegner, E. Weiss, and C. Wierzbicki. Computing classification system 1998: Current status and future maintenance, report of the ccs update committee. *Computing Reviews*, 39(1):1–62, 1998.

J. B. Cuseo. Igniting student involvement, peer interaction, and teamwork: A taxonomy of specific cooperative learning structures and collaborative learning strategies, 2002.

D. Dagger, V. Wade, and O. Conlan. A framework for developing adaptive personalized elearning. In Griff Richards, editor, *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2004*, pages 2579–2587, Washington, DC, USA, 2004. AACE.

W. Damon. Peer education: The untapped potential. *Journal of Applied Developmental Psychology*, 5(4), Oct-Dec:331–343, 1984.

E.M. de Cote, A. Lazaric, and M. Restelli. Learning to cooperate in multi-agent social dilemmas. In *AAMAS '06: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 783–785, New York, NY, USA, 2006. ACM. ISBN 1-59593-303-4.

E. S. J. de Faria, J. M. A. Coello, and K. Yamanaka. Forming groups for collaborative learning in introductory computer programming courses based on students programming styles: An empirical study. In *In Proceedings of the ASEE/IEEE Frontiers in Education Conference*, 2006.

Rina Dechter and Daniel Frost. Backjump-based backtracking for constraint satisfaction problems. *Artif. Intell.*, 136(2):147–188, 2002. ISSN 0004-3702.

P. Dillenbourg. What do you mean by collaborative learning? In *P. Dillenbourg (Ed) Collaborative-learning: Cognitive and Computational Approaches*, pages 1–19. Elsevier, Oxford, 1999.

P. Dillenbourg. Over-scripting cscl: The risks of blending collaborative learning with instructional design. *Three worlds of CSCL. Can we support CSCL*, pages 61–91, 2002.

P. Dillenbourg, M. Baker, A. Blaye, and C. O'Malley. The evolution of research on collaborative learning. In *E. Spada & P. Reiman (Eds) Learning in Humans and Machine: Towards an interdisciplinary learning science*, pages 189–211. Elsevier, Oxford, 1996.

L. Ding, P. Kolari, Z. Ding, S. Avancha, T. Finin, and A. Joshi. Using ontologies in the semantic web: A survey. Technical report, University of Maryland, 2005a.

L. Ding, L. Zhou, T. Finin, and A. Joshi. How the semantic web is being used: An analysis of foaf documents. In *HICSS '05: Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4*, page 113.3, Washington, DC, USA, 2005b. IEEE Computer Society. ISBN 0-7695-2268-8-4.

D. Dippold. Peer feedback through blogs: Student and teacher perceptions in an advanced german class. *ReCALL*, 21(1):18–36, 2009. ISSN 0958-3440.

S. Downes. E-learning 2.0. *eLearn*, 2005(10):1, 2005.

J.S. Edwards, K.C. Burgess Yakemovic, D.P. Cowan, T.J. Gaiser, J. Gancz, E. Levin, J. Vezina, and E. Wynn. E-team: forming a viable group on internet. In *SIGCPR'96: Proceedings of the 1996 ACM SIGCPR/SIGMIS conference on Computer personnel research*, pages 161–172, New York, NY, USA, 1996. ACM. ISBN 0-89791-782-0.

T. Eiter, G. Ianni, T. Lukasiewicz, R. Schindlauer, and H. Tompits. Combining answer set programming with description logics for the semantic web. *Artificial Intelligence*, 172(12-13):1495–1539, 2008. ISSN 0004-3702.

D. Fensel, J. Hendler, H. Lieberman, W. Wahlster, and T. Berners-Lee. Spinning the semantic web: Bringing the world wide web to its full potential, 2005.

T. Finin, L. Ding, and L. Zhou. Social networking on the semantic web. *The Learning Organization*, 12:418–435, 2005.

A. D. Frank and J. L. Brownell. *Organizational Communication and Behavior: Communicating to Improve Performance*. Wadsworth Publishing, 1989.

Daniel Frost and Rina Dechter. Look-ahead value ordering for constraint satisfaction problems, 1995.

Daniel Hunter Frost and Hunter Frost. *Algorithms and Heuristics for Constraint Satisfaction Problems*. PhD thesis, University of California, Irvine, 1997.

D. Gasevic, J. Jovanovic, and V. Devedzic. Enhancing learning object content on the semantic web. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT'04)*, pages 714–716. IEEE Computer Society, Washington, DC, USA, 2004.

M. E. Gaston and M. desJardins. Agent-organized networks for dynamic team formation. In *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 230–237, New York, NY, USA, 2005. ACM. ISBN 1-59593-093-0.

M. E. Gaston, J. Simmons, and M. desJardins. Adapting network structures for efficient team formation. In *AAMAS-04 Workshop on Learning and Evolution in Agent-based Systems*, 2004.

M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the Unites States of America (PNAS)*, 99(12):7821–7826, June 2002. ISSN 0027-8424.

J. Goecks and E. D. Mynatt. Leveraging social networks for information sharing. In *ACM Computer Supported Collaborative Work*, volume 6(3), pages 328–331, 2004.

S. P. Goggins, J. Laffey, and I. Tsai. Cooperation and groupness: community formation in small online collaborative groups. In *GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work*, pages 207–216, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-845-9.

A. Gokhale. Collaborative learning enhances critical thinking. *Journal of Technology Education*, 7, 1995.

J. Golbeck and J. Hendler. Reputation network analysis for email filtering. In *the First Conference on Email and Anti-Spam*, Address, 2004.

J. Golbeck and A. Mannes. Using trust and provenance for content filtering on the semantic web. In *Models of Trust for the Web (MTW) Workshop, In the 15th International World Wide Web Conference (WWW2006)*, Address, 2006.

J. Golbeck, B. Parsia, and J. Hendler. Trust networks on the semantic web. In *Cooperative Intelligent Agents 2003*, Address, 2003.

Jennifer Golbeck and James A. Hendler. Inferring binary trust relationships in web-based social networks. *ACM Trans. Internet Techn.*, 6(4):497–529, 2006.

David E. Goldberg and John H. Holland. Genetic algorithms and machine learning. *Mach. Learn.*, 3(2-3):95–99, 1988. ISSN 0885-6125.

S. Graf and R. Bekele. Forming heterogeneous groups for intelligent collaborative learning systems with ant colony optimization. *In LNCS on Intelligent tutoring Systems*, pages pp. 217–226, 2006.

A. Grigoris and F. van Harmelen. *A Semantic Web Primer*. MIT Press, Cambridge, MA, USA, 2004. ISBN 0262012103.

Steve R. Gunn. Support vector machines for classification and regression. Technical report, Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, May 1998.

J.R. Hackman, editor. *Groups that work (and those that don't)*. Jossey-Bass, San Francisco, 1990.

J.R. Hackman, R. Wageman, T.M. Ruddy, and C.L. Ray. Team effectiveness in theory and in practice. In Edwin A. Locke Cary L. Coope and, editor, *Industrial and Organizational Psychology*, pages 110–129. Blackwell, USA, 2000.

M. Hamasaki and H. Takeda. Find better friends? - re-configuration of personal networks by the neighborhood matchmaker method -. In *In Proceedings of the SWFAT 03, Nara, pages. 73-76.*, Address, 2003.

Greg Hamerly and Charles Elkan. Learning the k in k-means. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

M. Harrigan, M. Kravcik, C. Steiner, and V. Wade. What do academic users really want from an adaptive learning system? In *UMAP*, pages 454–460, 2009.

T. E. Harris. *Applied organizational communication: Perspectives, principles, and pragmatics*. Routledge, 1992.

Randy L. Haupt and Sue Ellen Haupt. *Practical Genetic Algorithms*. Wiley-Interscience, 2004. ISBN 0471455652.

J. Hendler. Agents and the semantic web. *IEEE Intelligent Systems*, 16(2):30–37, 2001. ISSN 1541-1672.

M.J. Higgs. A comparison of myers briggs type indicator profiles and belbin team roles. Technical report, Henley Business School, University of Reading (Henley Working Paper Series, HWP 9640), 1996.

M.J. Higgs, A. Plewnia, and G. Ploch. Influence of team composition on team performance and dependence on task complexity. Technical report, Henley Business School, University of Reading (Henley Working Paper Series, HWP 0314), 2003.

P. Honey and A. Mumford. *The manual of learning styles*. Peter Honey Publications, 1992.

H.U. Hoppe. Use of multiple student modeling to parametrize group learning. In J. e. Greer, editor, *In Proceedings of the 7th World Conference on Artificial Intelligence in Education (AI-ED'95)*, pages 234–241, 1995.

M. Ikeda, S. Go, and R. Mizoguchi. Opportunistic group formation. In B. D. Boulay and R. Mizoguchi, editors, *Proceedings of AI-ED 97 Artificial Intelligence in Education: Knowledge and Media in Learning Systems*, pages 166–174, 1997.

A. Inaba, T. Supnithi, M. Ikeda, R. Mizoguchi, and J. i. Toyoda. How can we form effective collaborative learning groups? In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems.*, pages 282–291, 2000.

M. O. Jackson. A survey of models of network formation: Stability and efficiency. Game Theory and Information 0303011, EconWPA, Mar 2003.

M. O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008. ISBN 978-0-691-13440-6.

S. E. Jackson, K. E. May, and K. Whitney. Understanding the dynamics of diversity in decision-making teams. In R. A. Guzzo and E. Salas, editors, *Team decision-making effectiveness in organizations*, pages 204–261, 1995.

T. K. Johansen. *Predicting a Team's Behaviour by Using Belbin's Team Role Self Perception Inventory.* PhD thesis, University of Stirling, 2003.

D. W. Johnson, R. T. Johnson, and E. J. Holubec. *Circles of Learning: Cooperation in the Classroom.* Interaction Book Company Edina, MN, 5 edition, 1998. ISBN 0939603128.

R. T. Johnson and D. W. Johnson. Action research: Cooperative learning in the science classroom. *Science and Children*, 24:31–32, 1986.

S. Kagan. Co-op co-op: A flexible cooperative learning technique. In R Slavin, S Sharan, S Kagan, Hertz-Lazarowitz, C Webb, and R Schmuck, editors, *Learning to cooperate cooperating to learn*, pages 437–452. NY: Plenum Press, 1985.

Yannis Kalfoglou and Marco Schorlemmer. Ontology mapping: the state of the art. *The Knowledge Engineering Review*, 18(1):1–31, January 2003.

Siddharth Kaza and Hsinchun Chen. Evaluating ontology mapping techniques: An experiment in public safety information sharing. *Decis. Support Syst.*, 45(4): 714–728, 2008. ISSN 0167-9236.

M. Klein, J. Broekstra, D. Fensel, F. van Harmelen, and I. Horrocks. Ontologies and schema languages on the web. In H. Lieberman D. Fensel, J. Hendler and w. Wahlster, editors, *Booktitle*, pages 96–100. MIT Press, Address, 2003.

B. Korte and L. Lovsz. Mathematical structures underlying greedy algorithms. In Ferenc Gcseg, editor, *Fundamentals of Computation Theory*, volume 117 of *Lecture Notes in Computer Science*, pages 205–209. Springer Berlin / Heidelberg, 1981. 10.1007/3-540-10854-$8_2$2.

John R. Koza, Martin A. Keane, Matthew J. Streeter, William Mydlowec, Jessen Yu, and Guido Lanza. *Genetic Programming IV : Routine Human-Competitive Machine Intelligence (Genetic Programming).* Springer, 2003. ISBN 1402074468.

S. R. Kruk. Foaf-realm - control your friends' access to resources, 2001.

V. Kumar. Algorithms for constraint-satisfaction problems: a survey, 1992. ISSN 0738-4602.

C. Lampe, N. Ellison, and C. Steinfield. A face(book) in the crowd: social searching vs. social browsing. In *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 167–170, New York, NY, USA, 2006. ACM. ISBN 1-59593-249-6.

J. Lave and E. Wenger. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, 1991. ISBN 978-0-521-42374-8.

K.F. Lawrence and m.c. schraefel. Bringing communities to the semantic web and the semantic web to communities. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 153–162, New York, NY, USA, 2006. ACM. ISBN 1-59593-323-9.

F. Legras and C. Tessier. Lotto: group formation by overhearing in large teams. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 425–432, New York, NY, USA, 2003. ACM. ISBN 1-58113-683-8.

N. Leone, G. Pfeifer, W. Faber, T. Eiter, G. Gottlob, S. Perri, and F. Scarcello. The dlv system for knowledge representation and reasoning. *ACM Trans. Comput. Logic*, 7(3): 499–562, 2006. ISSN 1529-3785.

I. Liccardi, A. Ounnas, R. Pau, E. Massey, P. Kinnumen, S. Lewthwaite, M. Midy, and C. Sakar. The role of social networks in students' learning experiences. *ACM SIGCSE Bulletin (December Issue)*, pages 224–237, December 2007.

V. Lifschitz. What is answer set programming? Technical report, Department of Computer Science, University of Texas, 2005.

D. Lloyd, S. Benford, and C. Greenhalgh. Formations: explicit group support in collaborative virtual environments. In *VRST '99: Proceedings of the ACM symposium on Virtual reality software and technology*, pages 162–163, New York, NY, USA, 1999. ACM. ISBN 1-58113-141-0.

G. Lugano, P. Nokelainen, M. Miettinen, J. Kurhila, and H. Tirri. On the relationship between learners' orientations and activity in cscl. In *ICALT '04: Proceedings of the IEEE International Conference on Advanced Learning Technologies*, pages 759–761, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7695-2181-9.

P. Markellou, I. Mousourouli, S. Spiros, and A. Tsakalidis. Using semantic web mining technologies for personalized e-learning experiences. In *Web-Based Education: Proceedings of the Fourth IASTED International Conference (WBE-2005)*, 2005.

C. C. Marshall and F. M. Shipman. Which semantic web? In *The fourteenth ACM Conference on Hypertext and Hypermedia*. ACM Press, New York, NY, 57-66, 2003.

R. McDermott. Nurturing Three Dimensional Communities of Practice: How to the most out of human networks. *Knowledge Management Review*, 1999.

D. W. McDonald and M. S. Ackerman. Expertise recommender: A flexible recommendation system and architecture. In *ACM CSCW 00*, 2000a.

D.W. McDonald and M.S. Ackerman. Expertise recommender: a flexible recommendation system and architecture. In *CSCW '00: Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 231–240, New York, NY, USA, 2000b. ACM. ISBN 1-58113-222-0.

T. W. McGuire, S. Kiesler, and J. Siegel. Group and computer-mediated discussion effects in risk decision making. *Journal of Personality and Social Psychology*, 52(5):917–930, 1987.

A. Michailidou and A. A. Economides. Gender and diversity in collaborative virtual teams. *Computer Supported Collaborative Learning: Best Practices and Principles for Instructors*, pages 199–224, 2007.

P. Mika. Flink: Semantic web technology for the extraction and analysis of social networks. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3):211–223, October 2005.

D. Millard, F. Tao, K. Doody, A. Woukeu, and H. Davis. The knowledge life cycle for e-learning. *International Journal of Continuing Engineering Education and Lifelong Learning: Special Issue on Application of Semantic Web Technologies in E-learning*, 16: 110–121, 2006.

B. Mirkin, S. Nascimento, and L. M. Pereira. Representing a computer science research organization on the acm computing classification system. In Peter W. Eklund and Ollivier Haemmerl, editors, *ICCS Supplement*, volume 354 of *CEUR Workshop Proceedings*, pages 57–65. CEUR-WS.org, 2008.

M. Mühlenbrock. Formation of learning groups by using learner profiles and context information. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education AIED-2005*, 2005.

M. Mühlenbrock. Learning group formation based on learner profile and context. *International Journal on E-Learning*, 5:19–24, 2006.

N. Myller, J. Suhonen, and E. Sutinen. Using data mining for improving web-based course design. In *ICCE '02: Proceedings of the International Conference on Computers in Education*, page 959, Washington, DC, USA, 2002. IEEE Computer Society. ISBN 0-7695-1509-6.

B. A. Nadel. Constraints satisfaction algorithms. *Computaional Intelligence*, 5(4):188–224, 1989. ISSN 0824-7935.

D. Nardi and R. J. Brachman. An introduction to description logics. In *Description Logic Handbook*, pages 1–40, 2003.

M. E. J. Newman. Analysis of weighted networks, July 2004a.

M. E. J. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. *Lecture notes in physics*, 650:337–370, 2004b. ISSN 0075-8450.

M. E. J. Newman. Modularity and community structure in networks. *PNAS*, 103(23): 8577–8582, June 2006.

M. E. J. Newman. Mathematics of networks. In L. E. Blume and S. N. Durlauf, editors, *The New Palgrave Encyclopedia of Economics (2nd Edition)*. Palgrave Macmillan, 2008.

M.E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69(2 Pt 2), February 2004. ISSN 1539-3755.

Natalya F. Noy and Mark A. Musen. Evaluating ontology-mapping tools: Requirements and experience. In *In Proceedings of OntoWeb-SIG3 Workshop at the 13th International Conference on Knowledge Engineering and Knowledge Management*, pages 1–14, 2002.

J. F. Nunamaker, A. R. Dennis, J. S. Valacich, D. Vogel, and J. F. George. Electronic meeting systems. *Communications of the ACM*, 34(7):40–61, 1991. ISSN 0001-0782.

I. O'Keefee, A. Brady, O. Conlan, and V. Wade. Just-in-time generation of pedagogically sound, context sensitive personalized learning experiences. *International Journal on E-Learning*, 5(1):113–127, 2006.

R. E. W. B. Olsen and S. Kagan. About cooperative learning. In Carolyn Kessler, editor, *Cooperative language learning:A teacher's resource book*. NJ: Prentice Hall, 1992.

A. Ounnas, H. C. Davis, and D. E. Millard. Semantic modeling for group formation. In *Workshop on Personalisation in E-Learning Environments at Individual and Group Level (PING) at the 11th International Conference on User Modeling UM2007*, June 2007a.

A. Ounnas, H. C. Davis, and D. E. Millard. Towards semantic group formation. In *The 7th IEEE International Conference on Advanced Learning Technologies (ICALT 2007)*, pages 825–827, 2007b.

A. Ounnas, H. C. Davis, and D. E. Millard. A framework for semantic group formation in education. In *Journal of International Forum of Educational Technology and Society*. IEEE IFETS, 2009. ISBN 1436-4522.

A. Ounnas, I. Liccardi, H. C. Davis, D. E. Millard, and S. White. Towards a semantic modeling of learners for social networks. In *Proceedings of International Workshop on Applications of Semantic Web Technologies for E-Learning (SW-EL) at the AH2006 Conference, pp.102-108*. Lecture Notes in Learning and Teaching, 2006.

A. Ounnas, D. E. Millard, and H. C. Davis. A metrics framework for evaluating group formation. In *ACM Group'07*, pages 221–224. ACM New York, NY, USA, 2007c.

D. A. Owens, E. A. Mannix, and M. A. Neale. Strategic formation of groups: Issues in task-performance and team member selection. In D. H. Gruenfeld, editor, *Research on managing groups and teams: Composition*, volume 1, pages 149–165. JAI Press, 1998.

R. L. Oxford. Cooperative learning, collaborative learning, and interaction: Three communicative strands in the language classroom. *Modern Language Journal*, 81:443–456, 1997.

A. S. Palincsar. Social constructivist perspectives on teaching and learning. *Annual Review of Psychology*, 49, 1998.

A. Passant. Foafmap: Web2.0 meets the semantic web. In Sören Auer, Chris Bizer, and Libby Miller, editors, *CEUR Workshop Proceedings ISSN 1613-0073*, volume 183, June 2006.

J. Piaget, T. Brown, and K. J. Thampy. *The equilibration of cognitive structures: The central problem of intellectual development*. University of Chicago Press Chicago, 1985.

S. P. Ponzetto and M. Strube. Deriving a large scale taxonomy from wikipedia. In *The 22st National Conference on Artificial Intelligence*, pages 1440–1445, 2007.

A. Preece, J. Z. Pan, S. Chalmers, P.Gray, and C. Mckenzie. Handling soft constraints in the semantic web architecture. In *Proceedings of WWW 2006, Edinburgh, UK*, 2006.

J. Preece. *Online Communities: Designing Usability and Supporting Sociability*. John Wiley & Sons, September 2000. ISBN 0471805998.

R. Rada. *Understanding Virtual Universities*. Intellect Books, 2001.

M. A. Redmond. A computer program to aid assignment of student project groups. In *SIGCSE'01: Proceedings of the thirty-second SIGCSE technical symposium on Computer Science Education*, pages 134–138, New York, NY, USA, 2001. ACM. ISBN 1-58113-329-4.

Susanne Richter, Robert Tolksdorf, and Rainer Eckstein. Foaf and other ways of describing a person. Technical report, Freie Universität Berlin, Institute for Computer Science, 2006.

B. Rogoff. Cognition as a collaborative process. In D. Kuhn & R.S. Siegler, editor, *Handbook of Child Psychology: Cognition, perception and language (5th ed.)*, volume 2, pages 679–744. NY: Wiley, 1998.

B. Rogoff and J. Lave. *Everyday Cognition: Its Development in Social Context*. Cambridge Mass.: Harvard University Press, 1984.

S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition)*, chapter Constraint Satisfaction Problems, pages 137–160. Prentice Hall, December 2002. ISBN 0137903952.

Norman M. Sadeh and Mark S. Fox. Variable and value ordering heuristics for the job shop scheduling constraint satisfaction problem. *Artificial Intelligence*, 86:1–41, 1996.

M. Sah and W. Hall. Building and managing personalized semantic portals. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1227–1228, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7.

R. E. Salvin and E. Oickle. Effects of cooperative learning teams on student achievement and race relations: Treatment by race interactions. *Sociology of Education*, 54(3):174–180, 1981.

T. M. Schwen and N. Hara. Community of practice: A metaphor for online design? *The Information Society*, page 257270, 2003.

T. Segaran. *Programming collective intelligence*. O'Reilly, 2007. ISBN 9780596529321.

N. Shadbolt, T. Berners-Lee, and W. Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101, 2006. ISSN 1541-1672.

C. Shirky. *Here Comes Everybody: How Change Happens When People Come Together*. Pinguin Books, 2008.

Pavel Shvaiko and Jérôme Euzenat. A survey of schema-based matching approaches. In *Journal on Data Semantics IV*, chapter 5, pages 146–171. Lecture Notes in Computer Science, Springer, 2005.

P. Singla and M. Richardson. Yes, there is a correlation - from social networks to personal behavior on the web. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 655–664, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2.

M. Skolnik. The virtual university and the professoriate. In S. Inayatullah and J. Gidley, editors, *The University in Transformation: Global Perspectives on the Futures of the University*, pages 55–67. Praeger Press, 2000.

R. E. Slavin. Research on cooperative learning: An international perspective. *Scandinavian Journal of Educational Research*, 33:231– 243, 1989.

R. E. Slavin. *Educational Psychology: Theory And Practice*. Allyn & Bacon, 1993. ISBN 0205151612.

L. Soh. On cooperative learning teams for multiagent team formation. In *Technical Report WS-04-06 of the AAAI'2004 Workshop on Forming and Maintaining Coalitions and Teams in Adaptive Multiagent Systems*, pages pp. 37–44, 2004.

L.-K. Soh, N. Khandaker, X. Liu, and H. Jiang. A computer-supported cooperative learning system with multiagent intelligence. In *AAMAS'06, pp 1556-1563*, 2006.

L. Sproull and S. Kiesler. *Connections: New ways of working in the networked organization.* Cambridge, MA: MIT Press, 1991.

G. Stahl, T. Koschmann, and D. Suthers. CSCL: An Historical Perspective. In R. K. Sawyer, editor, *Cambridge Handbook of the Learning Sciences*. Cambridge University Press, 2006.

K. Stefanov. Computing ontology creation. In *CompSysTech '03: Proceedings of the 4th international conference conference on Computer systems and technologies*, pages 656–660, New York, NY, USA, 2003. ACM. ISBN 954-9641-33-3.

K. T. Stevens. *The Effects of Roles and Personality Characteristics on Software Development Team Effectiveness*. PhD thesis, Faculty of Virginia Polytechnic Institute and State University, Address, 1998.

L. Stojanovic, S. Staab, and R. Studer. elearning based on the semantic web. In *In WebNet2001 - World Conference on the WWW and Internet*, pages 23–27, 2001.

F. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In Carey L. Williamson, Mary Ellen Zurko, and Prashant J. Patel-Schneider, Peter F. Shenoy, editors, *16th International World Wide Web Conference (WWW 2007)*, pages 697–706, Banff, Canada, 2007. ACM. ISBN 978-1-59593-654-7.

F. Suchanek, G. Kasneci, and G. Weikum. YAGO - a large ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics*, 6(3):203–217, September 2008.

M. Szomszor, H. Alani, I. Cantador, K. O'Hara, and N. Shadbolt. Semantic modelling of user interests based on cross-folksonomy analysis. In *7th International Semantic Web Conference (ISWC)*, October 2008.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Cluster analysis: Basic concepts and algorithms. In *Introduction to Data Mining*. Addison-Wesley, 2006.

T. Tiropanis, H.C. Davis, D.E. Millard, and M. Weal. Semantic technologies for learning and teaching in the web 2.0 era - a survey. In *WebSci'09: Society On-Line*, 2009.

C. M. Tobar and R. L. de Freitas. A support tool for student group definition. In *In Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference*, 2007.

A. Vivacqua and H. Lieberman. Agents to assist in finding help. In *ACM CHI00*, pages 65–72, 2000.

J. F. Voss, J. Wiley, and M. Carretero. Acquiring intellectual skills. *Annual Review of Psychology*, 46, 1995.

L. S. Vygotsky. *Educational Psychology*. St. Lucie Press, Florida, 1997.

L. S. Vygotsky and Michael Cole. *Mind in society: the development of higher psychological processes / L. S. Vygotsky; edited by Michael Cole ... [et al.]*. Harvard University Press, Cambridge, 1978. ISBN 0674576284 0674576292.

W3C. Owl web ontology language. Technical report, World Wide Web Consortium, February 2004a.

W3C. Resource description framework (rdf). Technical report, World Wide Web Consortium, RDF Working Group, February 2004b.

W3C. Sparql query language for rdf. Technical report, World Wide Web Consortium, January 2008.

W3C. Owl 2 web ontology language. Technical report, World Wide Web Consortium, October 2009a.

W3C. Skos simple knowledge organization system. Technical report, World Wide Web Consortium, August 2009b.

W3C. Rif working group. Technical report, World Wide Web Consortium, June 2010.

D. Y. Wang, S. S. J. Lin, and C. Sun. Diana: A computer-supported heterogeneous grouping system for teachers to conduct successful small learning groups. *Computers in human Behaviours*, pages 1997–2010, 2007.

M. Warschauer. Motivational aspects of using computers for writing and communication. In M. Warschauer and R. Kern, editors, *Network-based language learning: Concepts and practice*. Cambridge University Press, New York, 1996.

M. Warschauer. Computer-mediated collaborative learning: Theory and practice. *The Modern Language Journal*, 81(4):470–481, Winter 1997.

R. Wegerif. The social dimension of asynchronous learning networks. *Journal of ALN*, 2: 34–49, 1998.

E.C. Wenger. *Communities of Practice: Learning, Meaning and Identity*. Cambridge University Press, 1998.

E.C. Wenger and W.M Snyder. Communities of practice: The organizational frontier. *Harvard Business Review*, 78(1):139–145, 2000.

M Wessner and H. Pfister. Group formation in computer-supported collaborative learning. In *GROUP '01: Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*, pages 24–31, New York, NY, USA, 2001. ACM. ISBN 1-58113-294-8.

S. Wilson and P. R. Jones. what is...IMS- IMS Learner Information Packaging? Technical report, JISC, 2002.

M. Winter. Developing a group model for student software engineering teams. Master's thesis, University of Saskatchewan, 2004.

A. Woukeu, G. Wills, G. Conole, L. Carr, L. Kampa, and W. Hall. Ontological hypermedia in education: A framework for building web-based educational portals. In *World Conference on Educational Multimedia, Hypermedia & Telecommunications (ED-MEDIA'03)*, pages 349–357, 2003.

F. Wu and D. S. Weld. Autonomously semantifying wikipedia. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9.

H. Yang and J. Tang. Effects of social network on students' performance: A web-based forum study in taiwan. *Journal of Asynchronous Learning Networks*, 7(3):93–107, September 2003.

Y. Yang, C. M. A. Yeung, M. J. Weal, and H. C. Davis. The researcher social network: A social network based on metadata of scientific publications. In *Proceedings of WebSci'09: Society On-Line*, March 2009.

C. M. A. Yeung, N. Gibbins, and N. Shadbolt. Contextualising tags in collaborative tagging systems. In *20th ACM Conference on Hypertext and Hypermedia*, June 2009.

B. Yu and M.P. Singh. Searching social networks. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 65–72, New York, NY, USA, 2003. ACM. ISBN 1-58113-683-8.

J. Zhang and M. S. Ackerman. Searching for expertise in social networks: a simulation of potential strategies. In *GROUP '05: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 71–80, New York, NY, USA, 2005. ACM. ISBN 1-59593-223-2.

L. Zhang and S. Malik. The quest for efficient boolean satisfiability solvers. In *CAV '02: Proceedings of the 14th International Conference on Computer Aided Verification*, pages 17–36, London, UK, 2002. Springer-Verlag. ISBN 3-540-43997-8.

X. Zhang, L. Soh, X. Liu, and H. Jiang. Intelligent collaborating agents to support teaching and learning. In *IEEE International Conference on Electro Information Technology 2005*, 2005.