University of California

Los Angeles

Structure and evolution of scientific collaboration networks in a modern research collaboratory

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Information Studies

by

Alberto Pepe

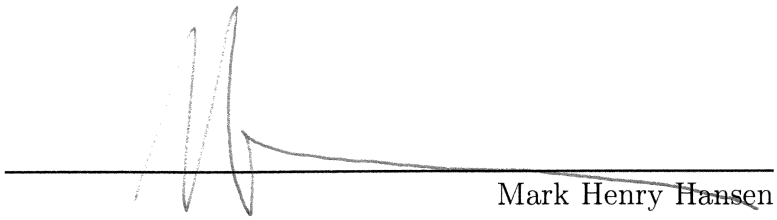
2010

© Copyright by

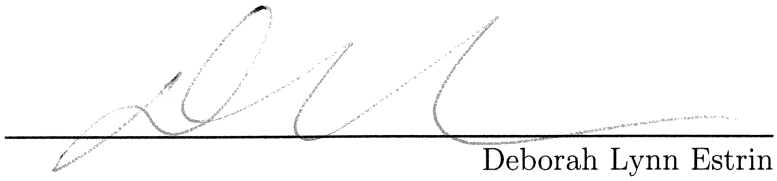
Alberto Pepe

2010

The dissertation of Alberto Pepe is approved.

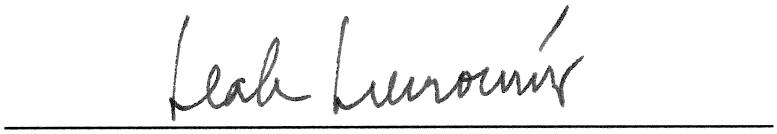


Mark Henry Hansen

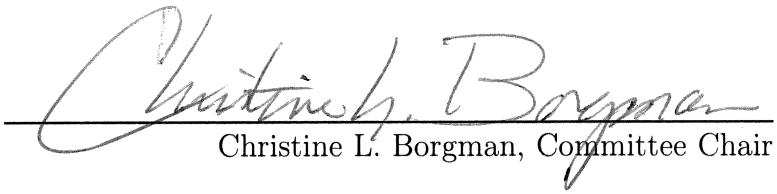


Deborah Lynn Estrin

Ramesh Srinivasan



Leah A. Lievrouw



Christine L. Borgman, Committee Chair

University of California, Los Angeles

2010

A Manduria

Table of Contents

1 Introduction 1

1.1 Making science 1

1.2 From laboratory science to collaboratory science 4

1.3 The scientific collaboratory as a complex system 8

1.4 A network approach 14

1.5 Organization of this dissertation 16

2 Studying scientific collaboration 17

2.1 CENS: Center for Embedded Networked Sensing 17

2.2 A scenario of scientific collaboration at CENS 19

2.3 Problem statement 23

2.4 Review of related literature 30

2.4.1 Literature on coauthorship and scholarly collaboration . . 30

2.4.2 Literature on online communication 34

2.4.3 Literature on acquaintanceship and social relationships . . 38

2.5 Contribution of this dissertation 42

3 Research methods, data and instruments 45

3.1 Foundations of network analysis 46

3.1.1 Basic properties of networks 47

3.1.2 Community structure 51

3.1.3 Homophily and assortative mixing 55

3.2 Data and instruments 58
3.2.1 The CENS bibliographic record 58
3.2.2 The CENS mailing list archive 64
3.2.3 The social network survey instrument 68
3.3 Summary 73

4 Results: Network topology and socio-academic configuration . 75
4.1 Coauthorship network 75
4.2 Communication network 80
4.3 Acquaintanceship network 84
4.4 Socio-academic configuration 98
4.5 Summary 102

5 Results: Structural analysis 104
5.1 Detection of community structure 105
5.2 Comparative analysis of community structure 111
5.2.1 Tests for statistical independence 111
5.2.2 Community structure across networks 115
5.2.3 Community structure and socio-academic configuration . . 122
5.3 Summary 133

6 Results: Evolutionary analysis 135
6.1 Slicing the networks by their temporal component 135
6.2 Evolution of network topology 142

6.3 The dynamics of preferential attachment 148
6.4 Evolution of the socio-academic configuration 155
6.5 Network evolution and socio-academic configuration 163
6.5.1 Assortative mixing in the coauthorship network 165
6.5.2 Assortative mixing in the communication network 176
6.5.3 Assortative mixing in the acquaintanceship network 179
6.6 Sequential relationship: acquaintanceship and coauthorship 184
6.7 Physical proximity at the CENS headquarters 190
6.8 Summary 195

7 Discussion 198
7.1 Summary of the results 198
7.2 A fluid, non-cliquish small-world 207
7.3 The role of interpersonal networks 212
7.4 Reflections on the notion of complexity 216
7.5 Lessons learned amid two modes of studying science 221

8 Conclusion 228
8.1 Limitations and assets of this study 229
8.2 Future work 233

A Appendix 236
A.1 Survey invitation letter 236
A.2 Text of the informed consent form 237

A.3 Online survey instrument 239
A.4 Comparative network data 241
A.5 Description of software code and tools 242
A.6 Seating chart of 3551 Boelter Hall 246

References **247**

List of Figures

2.1 A CENS nestbox 20

2.2 Images obtained from the imaging sensor in the nestbox. 21

3.1 A small example network. 47

3.2 Three examples of maximal cliques. 50

3.3 A network partitioned into three structural communities. 53

3.4 Comic: Interdisciplinary Madness 59

3.5 Temporal distribution of survey responses. 72

4.1 Weighted coauthorship network. 78

4.2 Weighted communication network. 83

4.3 Four possible classes of acquaintanceship ties. 86

4.4 Weighted acquaintanceship network (directed). 87

4.5 Academic profile of survey respondents and entire population. . . 91

4.6 Coauthorship degree distribution of survey respondents and entire population 92

4.7 Weighted acquaintanceship network (undirected). 97

5.1 Community membership as node metadata. 106

5.2 Coauthorship network: detected community structure. 107

5.3 Communication network: detected community structure. 109

5.4 Acquaintanceship network: detected community structure. 110

5.5 Two fictitious networks and detected community structure. 112

5.6 Anecdotal example: overlap of coauthorship and acquaintanceship communities. 119

5.7 Socio-academic information as node metadata. 122

5.8 Anecdotal example: overlap of coauthorship community and socio-academic profile. 132

6.1 Evolution of the coauthorship network. 139

6.2 Evolution of the communication network. 140

6.3 Evolution of the acquaintanceship network. 141

6.4 Degree distributions of the coauthorship network 149

6.5 Degree distributions of the communication network. 152

6.6 Degree distributions of the acquaintanceship network. 154

6.7 Evolution of discrete assortativity mixing in the coauthorship network 166

6.8 Evolution of discrete assortativity mixing in the communication network 177

6.9 Evolution of discrete assortativity mixing in the acquaintanceship network 180

6.10 Anecdotal example: evolution of coauthorship and acquaintanceship. 185

6.11 The coauthorship and acquaintanceship networks of Boelter Hall 3551. 194

A.1 Screenshot of the first page of the survey instrument. 239

A.2 Screenshot of the second page of the survey instrument. 240

A.3 Seating chart of the CENS headquarters (3551 Boelter Hall, UCLA) as of March 2010. 246

List of Tables

3.1 Basic statistics for the collected bibliographic data. 63

3.2 Basic statistics for the collected mailing list logs. 67

3.3 Basic statistics for the collected social survey data. 73

4.1 Topological properties of the coauthorship network. 79

4.2 Topological properties of the communication network. 84

4.3 Socio-academic profile of survey population and respondents . . . 90

4.4 A summary of the data collected in the social network survey. . . 93

4.5 A summary of the data collected in the second part of the social network survey. 95

4.6 Topological properties of the acquaintanceship network. 98

4.7 Socio-academic profile of the coauthorship, communication and acquaintanceship networks of collaboration 100

5.1 Contingency table displaying the community membership distribution of two fictitious networks. 113

5.2 Contingency table: communication vs. coauthorship networks. . . 116

5.3 Contingency table: communication vs. acquaintanceship. 117

5.4 Contingency table: coauthorship vs. acquaintanceship. 117

5.5 Independence tests between collaboration networks. 121

5.6 Collapsed categories (academic departments and positions). 125

5.7 Contingency tables: coauthorship vs. socio-academic community membership. 127

5.8 Contingency tables: communication vs. socio-academic community membership. 128

5.9 Contingency tables: acquaintanceship vs. socio-academic community membership. 129

5.10 Independence tests: collaboration networks and socio-academic profile. 130

6.1 Fundamental network statistics of the collaboration networks, 1999-2010 143

6.2 Components of the coauthorship network (by year). 144

6.3 Evolution of the CENS socio-academic configuration, 2003-2007 . 158

6.4 CENS Faculty dynamics, 2003-2007 161

6.5 Discrete assortativity coefficients of the collaboration networks. . . 164

6.6 Academic affiliation pairs in the coauthorship network. 168

6.7 Academic department pairs in the coauthorship network. 171

6.8 Academic position pairs in the coauthorship network. 174

6.9 Country of origin pairs in coauthorship network. 175

6.10 Academic department pairs in the communication network. 178

6.11 Academic affiliation pairs in the acquaintanceship network. 181

6.12 Academic department pairs in the acquaintanceship network. . . . 183

6.13 Affiliation pairs in the coauthorship and acquaintanceship networks: summary and statistical comparison 187

6.14 Department pairs in the coauthorship and acquaintanceship networks: summary and statistical comparison 189

6.15 Physical proximity pairs in the coauthorship and acquaintanceship networks of Boelter Hall 3551. 192

6.16 Assortativity by workplace location (seating row). 193

7.1 Summary of the main results found for the coauthorship, communication, and acquaintanceship networks. 199

A.1 Basic statistics for a number of published bibliographic, communication, and social networks. 241

Acknowledgments

This dissertation is about social networks and my own social network deserves considerable credit for its completion. First and foremost, I would like to thank my chair and advisor, Christine Borgman, for providing tremendous support on many levels. I am indebted to her for her enduring advice, mentoring, and patience. In difficult moments of my doctoral research, she also offered emotional support and encouragement. I will especially miss her salon-style dinner parties that gave me the opportunity to meet an extraordinary circle of scholars, friends, and collaborators. I feel truly honored to have been one of her PhD students.

I would also like to thank the other members of my dissertation committee: Leah Lievrouw and Ramesh Srinivasan for offering friendly guidance and extensive insight into research methods and sociological theory, Mark Hansen for nurturing my intellectual curiosity, and Deborah Estrin for welcoming me at CENS and being always available for a chat.

My research has also benefited from the support of numerous other faculty at UCLA, especially Jonathan Furner, Greg Leazer, Jean-François Blanchette, Sharon Traweek, Dana Cuff, and Victoria Vesna. Throughout my doctoral research, I have also had the chance to collaborate with faculty from other institutions. I am particularly grateful to members of the NSF-funded Monitoring, Modeling, Memory project. I thank Geoffrey Bowker, Paul Edwards, Tom Finholt, and Susan Leigh Star for many stimulating conversations throughout the years.

Working as a Graduate Student Researcher in the Statistics and Data Practices team of CENS, I have had the fortune to be surrounded by an outstanding circle of colleagues. I thank David Fearon, Andrew Lau, Matthew Mayernik,

Katie Shilton, Jillian Wallis, and Laura Wynholds for patiently reading and reviewing practically every single paper I have worked on in the past four years.

Financial and material support for this research has come from Microsoft Research. My sincere and heartfelt gratitude goes to Tony Hey, Corporate Vice President of External Research. Without his generous support, this dissertation could not have been written. I am also thankful to Catharine Van Ingen and Cathy Marshall of Microsoft Research for their advice throughout my entire doctoral career.

There are three other scholars who deserve a very special mention: Herbert Van De Sompel, Johan Bollen, and Marko Rodriguez. I am indebted to them for introducing me to the nuts and bolts of scholarly publishing, information architecture, online data mining, and network analysis. Very importantly, they taught me how to make scientific research *fun*. Without their support and sustaining cheer this dissertation would have been very different—maybe better? ;-)

Many friends have inspired me throughout these past four years. It is difficult to express how grateful I am to them. I wish I had the space to describe in detail how each one of them made my doctoral career memorable. I limit myself to listing here the ones who were particularly close to my research and influenced me with their ideas: Anil Bawa-Cavia, Yassine Bennani, Antoine Blin, Matteo Cantiello, Amandine Chaillous, Talar Chahinian, Leila Chirayath Janah, Giulio Dimitri, Daniel Costa, Melissa Fernandez, Boris Mangano, Mauro and Marco of Genera.tv, Jennafer Leanne McCabe, Fabio Oliva, Veronica Olivotto, Francesco Schiff, Basil Singer, Jeremi Sudol, Karen Van Godtsenhoven, Viktor Venson, Mario Vollera, Cara Walsh, Shane Willcox, and Spencer Wolff.

Heartfelt thanks to Simo Bennani, Gauvain Haulot, Andrew Price, Calli Ryals, Jeremie Senior, and Chris Starr for bringing vitality and comfort in the

solitary months of dissertation writing. They made this pretty town of Los Angeles look even prettier. I will miss you and Los Angeles very much.

Dulcis in fundo, un messaggio ai miei genitori, ai miei fratelli e alla nonna Gentile: grazie di cuore. Da buon emigrante, ho dedicato questa tesi a Manduria. Ma la mia è una dedica figurativa: il mio pensiero di camminante alla casa, alla terra, e alla famiglia. Questa tesi è dedicata a voi, che avete confidato in me fin dall'inizio. *Vi voglio bene.*

Vita

1979	Born, Manduria, Italy.
2002	B.Sc. Astrophysics, University College London, United Kingdom.
2003	M.Sc. Computer Science, University College London, United Kingdom.
2004	Research fellow, Scientific Computing and Data Visualization Group, CINECA — InterUniversity Consortium, University of Bologna, Italy.
2004-2006	Marie Curie fellow, Information Technology Department, CERN — European Organization for Nuclear Research, Geneva, Switzerland
2006-present	Graduate Student Researcher, Statistics and Data Practices Team, CENS — Center for Embedded Networked Sensing, Los Angeles, CA, USA
Summer 2007	Visiting fellow, Digital library Research and Prototyping Group, LANL — Los Alamos National Laboratory, Los Alamos, NM, USA
2009	Teaching assistant, Art, Science and Technology, Department of Design — Media Arts, University of California, Los Angeles, USA

Publications

Alberto Pepe and Marko A. Rodriguez. Collaboration in sensor network research: an in-depth longitudinal analysis of assortative mixing patterns. *Scientometrics* (in press), 2010.

Marko A. Rodriguez, **Alberto Pepe**, and Joshua Shinavier. The dilated triple. In B. Chbeir and A. Hassanien (eds.), *Emergent Web Intelligence: Advanced Semantic Technologies*. Springer (in press), 2010

Jillian C. Wallis, Matthew S. Mayernik, Christine L. Borgman, and **Alberto Pepe**. Digital libraries for scientific data discovery and reuse: from vision to practical reality. *Proceedings of the ACM IEEE Joint Conference on Digital Libraries* (in press), 2010.

Alberto Pepe and Corinna di Gennaro. Political protest Italian-style: The blogosphere and mainstream media in the promotion and coverage of Beppe Grillo's Vday. *First Monday*, vol. 14, no. 12, 2009. URL: `http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2740/2406`

Alberto Pepe, Matthew S. Mayernik, Christine L. Borgman, and Herbert Van De Sompel. From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. *Journal of the American Society for Information Science & Technology* vol. 61, no. 3, pp. 567 - 582, doi:10.1002/asi.21263, Wiley, 2010.

Alberto Pepe, Sasank Reddy, Lilly Nguyen, and Mark Hansen. Twitflick: visualizing the rhythm and narrative of micro-blogging activity. *Proceedings of Digital Art and Culture (DAC) conference*, 2009. URL: `http://escholarship.org/uc/item/6rw4n69h`

Marko A. Rodriguez and **Alberto Pepe**. Faith in the Algorithm, Part I: Beyond the Turing Test. *Proceedings of the AISB Symposium on Computing and Philosophy, The Society for the Study of Artificial Intelligence and Simulation of Behaviour*, 2009. URL: `http://arxiv.org/abs/0903.0200`

Alberto Pepe Gentile. Reinventing airspace: Spectatorship, fluidity, intimacy at PEK T3. *ACE: Architecture, City & Environment*, vol. 4, no. 10, pp. 9-19, 2009.

Alberto Pepe. Socio-epistemic analysis of scientific knowledge production in little science research. *tripleC (Cognition, Communication, Co-operation)*, vol. 6, no. 2, pp. 134-145, 2009.

Marko A. Rodriguez, Vadas Gintautas, and **Alberto Pepe**. A Grateful Dead analysis: The relationship between concert and listening behavior. *First Monday*, vol. 14, no. 1, 2009 URL: `http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2273/2064`

Ramesh Srinivasan, **Alberto Pepe**, and Marko A. Rodriguez. A clustering-based semi-automated technique to build cultural ontologies. *Journal of the American Society for Information Science & Technology*, vol. 60, no. 3, pp. 608 620, 2009.

Alberto Pepe and Johan Bollen. Between conjecture and memento: shaping a collective emotional perception of the future. *Proceedings of the AAAI Spring Symposium on Emotion, Personality, and Social Behavior*, 2008. URL: `http://arxiv.org/abs/0801.3864`

Marko A. Rodriguez† and **Alberto Pepe**†. On the relationship between the structural and socioacademic communities of a coauthorship network. *Journal of Informetrics*, vol. 2, no. 3, pp. 195-201, doi:10.1016/j.joi.2008.04.002, Elsevier, 2008.

Christine L. Borgman, Jillian C. Wallis, Matthew S. Mayernik, and **Alberto Pepe**. Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. *Proceedings of the ACM IEEE Joint Conference on Digital Libraries*, pp. 269-277, doi:10.1145/1255175.1255228, ACM, 2007.

Alberto Pepe, Christine L. Borgman, Jillian C. Wallis, and Matthew S. Mayernik. Knitting a fabric of sensor data resources. *Proceedings of the ACM IEEE International Conference on Information Processing in Sensor Networks*, ACM, 2007.

Alberto Pepe and Joanne Yeomans. Protocols for scholarly communication. *Astronomical Society of Pacific Conference Series*, vol. 377, 2007.

Jillian C. Wallis, Christine L. Borgman, Matthew S. Mayernik, and **Alberto Pepe**. Moving archival practices upstream: An exploration of the life cycle of

† Co-first authors

ecological sensing data in collaborative field research. *International Journal of Digital Curation*, vol. 3, no. 1, 2007. URL: `www.ijdc.net/index.php/ijdc/article/view/67`

Jillian C. Wallis, Christine L. Borgman, Matthew S. Mayernik, **Alberto Pepe**, Nithya Ramanathan, and Mark Hansen. Know thy sensor: Trust, data quality, and data integrity in scientific digital libraries. *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries*, vol. 4675/2007, pp. 380-391, doi:10.1007/978-3-540-74851-9_32, Springer, 2007.

Alberto Pepe. The digital library as a social and cooperative environment: experience perspective of the CERN document server. *La biblioteca su misura*. Milano: Editrice Bibliografica, 2006.

Alberto Pepe, Jean-Yves Le Meur, and Tibor Simko. Dissemination of scientific results in high energy physics: the CERN document server vision. *Proceedings of Computing in High Energy and Nuclear Physics (CHEP) Conference*, 2006.

Abstract of the Dissertation

Structure and evolution of scientific collaboration networks in a modern research collaboratory

by

Alberto Pepe

Doctor of Philosophy in Information Studies

University of California, Los Angeles, 2010

Professor Christine L. Borgman, Chair

This dissertation is a study of scientific collaboration at the Center for Embedded Networked Sensing (CENS), a modern, multi-disciplinary, distributed laboratory involved in sensor network research. By use of survey research and network analysis, this dissertation examines the collaborative ecology of CENS in terms of three networks of interaction: coauthorship of scholarly publications, communication activity on mailing lists, and interpersonal acquaintanceship. This study exposes the topology, structure, and evolution of these networks in relation with the disciplinary and institutional arrangements of CENS. Findings indicate that CENS collaboration networks have fluid, non-cliquish, small-world topologies, and are free of prestige-based mechanisms. Further analysis reveals that structural communities in the coauthorship and acquaintanceship networks overlap considerably. They also exhibit little disciplinary and institutional diversity locally, although CENS becomes more inter-disciplinary over time. Overall, results of the structural and evolutionary analyses point to the importance of interpersonal relationships for accomplishing scientific work in distributed environments.

CHAPTER 1

Introduction

1.1 Making science

In a 1918 *Nature* article on the making of science, British-American physicist Charles E. Kenneth Mees identified three stages of scientific knowledge production:

> "The increase of scientific knowledge can be divided into three steps: first, the production of new knowledge by means of laboratory research; secondly, the publication of this knowledge in the form of papers and abstracts of papers; thirdly, the digestion of the new knowledge and its absorption into the general mass of information by critical comparison with other experiments on the same or similar subjects." [1, p. 355].

How is scientific knowledge produced nowadays? Nearly a century later, Mees' distillation seems remarkably accurate and fairly up to date. The general mechanisms by which science is conducted have remained fairly stable over time. What has changed considerably in the past one hundred years is the level of academic and popular interest in these mechanisms. Domains as diverse as philosophy, history of science, logic, sociology, psychology and cognitive science have increasingly become interested in the study of scientific knowledge and its production

mechanisms. Different disciplines have brought forward different perspectives, theories, and methods. Studies of science grounded in sociology, for example, analyze the processes of scientific activity in the social, cultural and political context in which science takes place [2, for a review]. Cognitive scientists focus on the study of the scientific mind and the mental processes underlying scientific reasoning [3, for a review]. Information science contributes to this body of research by studying scientific practices and the artifacts generated in the process of scientific knowledge production. Information scientists employ digital collections of scholarly papers and bibliographic repositories in order to model and analyze networks of scientific collaboration. Although theoretically grounded in their own domains, these approaches to the study of science do not exist in isolation; they have often borrowed principles and concepts from each other.

An example of this disciplinary cross-fertilization is the sub-domain of psychology that deals with scientific creativity and problem solving. Traditional studies of scientific reasoning were concerned mostly with the study of the scientific mind as an individual entity and as the sole source of new ideas, insights and discoveries. The process of knowledge generation was studied in terms of cognitive capability, such as scientists' ability to generate and test hypotheses [4], and the ability to use analogical reasoning to construct novel links between known and unknown scientific facts [5]. Recently, these approaches have been implemented by new research trends, oftentimes grounded in sociology and anthropology, that locate the process of knowledge production in the social and cultural context in which it takes place. In a review of the history of scientific creativity, Dean K. Simonton [6] refers to an *internal zeitgeist*, defined by the subjective individual capability of scientific thinking, and an *external zeitgeist*, defined by the context in which scientific thinking takes place shaped by broader social, temporal, cultural and political dimensions. He notes that

> "Galileo became a great scientist only because he had the fortune of being born in Italy during the time when it became the center of scientific creativity. Similarly, Newton's creative genius could appear only because he lived in Great Britain when the center had shifted there from Italy. If Galileo and Newton had switched birth years without changing national origins, then neither would have secured a place in the annals of science." [6, p. 134].

This example, although specifically grounded in the field of psychology, illustrates a broader academic trend — that a theoretical framework for the study of scientific knowledge production has been reformulated in terms of the collective dimension in which science takes place. Studies of science and its knowledge production mechanisms focused on the *individual* have been progressively been implemented with studies that account for the *collective*. For these latter investigations, science making is not only an individual endeavor; it is also a collective, distributed process, involving the close interaction among a number of social, cultural, technological, economical, and political dimensions. This "collective" component of scientific knowledge production — the ensemble of collaborative processes and practices that take place in science making — is the focus of this dissertation.

By conceptualizing modern scientific research as a collective, distributed, collaborative activity, I study its advancement in a multi-disciplinary, multi-institutional research enterprise focused on the development and application of wireless sensing systems. My investigation relies on quantitative analyses of coauthorship, communication and acquaintanceship patterns among scientists with relation to the social and academic context in which those scientists are embedded.

1.2 From laboratory science to collaboratory science

In *Laboratory Life*, one of the earliest and most notable ethnographic studies of science, sociologists Bruno Latour and Steve Woolgar explored the construction of scientific facts in a biological research laboratory [7]. The aim of that investigation was to capture the latent minutiae of scientific activity by *in situ* monitoring of the scientists, working in their most natural working environment: the scientific laboratory. This "anthropology of science" took place in the late seventies when the scientific laboratory, in the traditional sense, was certainly the most obvious environment in which to study the production and the dissemination of scientific knowledge. Much collaboration in the traditional laboratory was driven by physical proximity and physical exchange of paper literature: books, journals, preprints, and articles.

With the advent of digital collaborative platforms in the past two decades, scientific research has changed considerably. Some modern scientific and engineering centers, for example, operate in a very distributed fashion by heavily relying on electronic communication and on distributed computing resources. The laboratory, as a collaborative research environment, has extended beyond the traditional laboratory walls: *the intellectual space in which scientific collaboration takes place no longer corresponds to a single physical working environment.* In turn, the laboratory, in the traditional sense, ceases to pose a reliable framework for the study of modern scientific activity.

The term "collaboratory" was coined in the early nineties from a blend of the words "collaboration" and "laboratory", to mark the importance of computer-supported collaboration work in science. The earliest definition of collaboratory described it as:

> "a center without walls in which the nation's researchers can perform their research without regard to geographical location — interacting with colleagues, accessing instrumentation, sharing data and computational resources, and accessing information in digital libraries" [8, p. 854].

In this definition, the requirement of physical and geographical proximity of the researchers ceases to exist in favor of a novel organization of scientific activity. The use of computer-based communication technologies relaxes the constraints of distance and time imposed by traditional paper-based laboratory work. Nowadays, both large-scale and smaller collaboratories represent a substantial portion of the ecology in which scientific knowledge production takes place. Modern collaboratory research extends beyond national, institutional, and disciplinary boundaries to make up dedicated networks of people, information, artifacts, technologies, and ideas dispersed around worldwide locations and institutions. More recent definitions of "collaboratory" stress the needs to solve problems simultaneously and remotely, to access and distribute datasets, and to provide flexible, informal interaction among colleagues [9].

Contemporary research on collaboratories tends to frame them in the context of scholarly and scientific cyberinfrastructure initiatives [10]. The National Science Foundation's cyberinfrastructure program promises to build a human-centered, comprehensive infrastructure "needed to capitalize on dramatic advances in information technology" [11, p. 4] via a wide range of initiatives aimed at integrating "high performance computing; data, data analysis and visualization; collaboratories, observatories and virtual organizations; and, education and workforce development" [11, p. 6]. This definition of cyberinfrastructure places collaboratories among other types of working environments, such

as large-scale observatories and virtual organizations. These working environments may vary in size, arrangement, resources, and aims. It is beyond the scope of this dissertation to provide a taxonomy of these emerging organizational arrangements. Yet, it is important to stress that their heterogeneity reflects the need to accommodate the range of national, institutional and disciplinary requirements for collaboration, computer-supported communication platforms, and ever increasing computational power. The result is an ecology of working environments, from large-scale observatories, such as the Virtual Observatory (`http://www.ivoa.net`), providing and managing astronomical data from archives and observatories worldwide, to virtual organizations, such as the Geosciences Network (GEON, `http://www.geongrid.org/`), grouping a dozen projects and institutions to enhance and integrate geoscience research.

Cyberinfrastructure initiatives in the form of collaboratories and virtual organizations are relatively recent and novel types of scientific organization. Being founded upon the notions of multi-disciplinary research, and modern multi-sited collaboration, these initiatives have attracted investments from major funding bodies, both at the national and international level. The function of collaboratories and cyberinfrastructure centers have received particular attention from sociologists of science and scholars interested in computer-supported collaborative work. Much of this research is especially concerned with the social, human, and organizational arrangements of these working environments. The pioneering work of Star and Ruhleder [12] has paved the way in this direction. Using the notion of "infrastructural inversion," initially introduced by Bowker [13], they propose to study infrastructural complexity with a focus on *relations*, rather than *things*. Recent research in this field, has built around this and similar notions.

Overall, scholars are finding that the social organization of cyberinfrastruc-

ture initiatives is in disarray. For example, Lee, Dourish and Mark [14] perform an ethnographic study of a large-scale cyberinfrastructure effort to uncover its human infrastructure— the people, organizations, networks, and social practices that support the technical enterprise. They find a hybrid human organization that mixes the old and new: traditional organizational forms and work practices embedded in new social and technological contexts. In other work, Cummings and Kiesler [15] analyze the multi-disciplinary and multi-university components of research collaborations finding that scientific coordination and reporting patterns are hindered not by the presence of multiple disciplines, but by the physical distance between collaborating institutions. This last finding points to a tensional disconnect between the underlying mission of many cyberinfrastructure initiatives—to foster interdisciplinary and multi-sited research—and their factual organization. This dissertation explores this tension in the context of a specific research collaboratory.

The collaboratory under study is the Center for Embedded Networked Sensing (CENS), a multi-disciplinary[1], multi-institutional research enterprise involved in wireless sensing research. CENS is presented in much detail in the next chapter. At this point, it is sufficient to note that the modus operandi of CENS fits very well within the general notion of collaboratory research presented above. More specifically, one can regard CENS as a collaboratory embedded within a larger and emerging cyberinfrastructure initiative. In order to investigate such a multi-

[1]A note about terminology. Specialized literature often conflates the terms *multi-disciplinary* and *inter-disciplinary*. For the purpose of this dissertation, *multi-disciplinary* refers to work "combining or involving several separate academic disciplines", i.e., work that involves the presence of different disciplines. By contrast, *inter-disciplinary* refers to work "of or pertaining to two or more disciplines or branches of learning; contributing to or benefiting from two or more disciplines", i.e., work that involves both the presence of and the interaction among different disciplines. The term *transdisciplinary*, not used in this dissertation, is often used to refer to work that is developed within one discipline but is then borrowed by others. The definitions above are taken from The Oxford English Dictionary. 2nd ed. 1989. OED Online. Oxford University Press.

faceted research environment I transcend the traditional study of the laboratory — the laboratory as a physical space — and analyze the distributed, highly collaborative nature of scientific research, via a network analysis. In this context, I formulate a definition of collaboratory that I employ throughout the rest of this dissertation, as follows:

> *A collaboratory is an array of inter-related physical and virtual environments in which researchers — possibly from different locations, affiliations and disciplines — use computer-supported technologies to produce scientific knowledge interacting formally and informally, solving problems, exchanging data, computing resources, and ideas.*

1.3 The scientific collaboratory as a complex system

The notion of collaboratory introduced in the previous section emphasizes two key factors related to the process of knowledge production in modern scientific research. First, scientific knowledge production takes place in a heterogeneous array of inter-related, distributed physical and virtual environments. Second, it is a collective, highly-collaborative endeavor that heavily relies on technological infrastructure in the form of computer-supported information and communication technologies (ICTs). These elements reveal that scientific research does not necessarily *happen* in a well-delimited physical space, suitable for in-depth, location-based investigations. Rather, scientific research permeates an array of physical and virtual collaborative environments. In these environments, interactions take place among a number of heterogeneous components. Even such a high level description of the prototypical scientific collaboratory reveals this underlying notion: that collaboratories can be regarded, essentially, as *systems*.

Collaboratories, as systems, are environments consisting of heterogeneous components — people (scientists, engineers, students, staff), artifacts (articles, data), technologies (communication technologies, technical equipment, analytical tools), ideas, and power structures.

Scientific activity in a collaboratory is marked by myriad *small-scale* interactions among these components. For example, scientists interact with other scientists, exchanging ideas, producing data and tools, writing new articles, and so on. This layer of small-scale interactions constitutes the fabric of social, cultural, technological, economical, and political activities that drive and define scientific knowledge production in a collaboratory. When analyzed at a broader level, a typical collaboratory also exhibits *large-scale* phenomena; for example, its mission and research agenda might change over time. The combination of a number of small-scale interactions might account for such phenomenon. For example, a large number of scientists might have been exposed to a new trend of literature; or a prestigious faculty member might have joined the collaboratory, influencing the research aims of a circle of collaborating scientists.

The scenario presented here suggests that the collaboratory described thus far is a type of *complex system*. A complex system is a system made up of a large number of components that interact in such a way that their collective behavior is not a simple combination of their individual behaviors [16]. As such, a complex system is one whose behavior is neither regular, nor random. Describing complexity in the realm of social systems, Niklas Luhmann places complex social systems at the intersection between systems in which every element can be related to every other element (regular) and those in which this is not the case (random) [17]. Other descriptions of social complexity frame it in terms of *agency* (acting at the microscopic level) and *structure* (emerging at the macroscopic level). Agency

is the level of interactions between individual agents. The regular, recurrent interplay of interactions at the microscopic level produces stable relations that become structure: rules, values, ethics and morals which both constrain and enable agency. Research on social complexity is aimed at understanding the *dialectic* between agency and structure [18]. As explained in the next chapter, this dissertation explores this dialectic by investigating the interdependence between small-scale interactions of scholarly, social, and communication activity and the overarching disciplinary and institutional arrangement of the collaboratory.

The study of complex systems is an academic discipline of its own. Complexity science is a necessarily interdisciplinary endeavor, focused on understanding the consequences of combining many small- and large-scale system-environment dynamics [19, 20, 21, for a review]. Complex systems manifest themselves in many biological, social, technological and organizational settings, but resist a single, universal definition. They are best described by exploring their defining characteristics. Since the subject of study of this dissertation is a collaboratory conceptualized as a complex system, and because this dissertation employs a number of analytical methods for the study of complex networks, it is opportune to delineate some key notions that are central to the organization and function of many complex systems. I discuss the theoretical underpinnings of emergence and boundary flexibility, and how they apply to studies of collaboratory research.

Emergence. Emergence is a core defining characteristic of complex systems. The first definition of emergence dates back to the nineteenth century with English philosopher G.H. Lewes, but many trace its roots to Aristotle [22, for a historical review]. In layman terms, the emergence paradigm can be explained as: *the whole is different from the sum of its parts*. In more technical terms, emergence happens when novel and coherent structures, patterns, and properties

arise during the process of self-organization in complex systems. In particular, complex systems exhibit large-scale patterns that cannot necessarily be deduced from the intrinsic properties of the individual components and their small-scale interactions [23, for a review].

The emergent nature of collaboratories is evident from the multitude of small-scale interacting components that constitute them and their repercussion on their large-scale configuration [24]. Collaboratories are inherently social systems and their function is solidly built around relations: between people, artifacts, technologies, etc. In a collaboratory, an example of small-scale interaction is the intensification of scholarly collaborations between researchers from different disciplines, while an example of large-scale phenomenon could be a shift of its research agenda towards new disciplines. Reflecting on the emergent component of cyberinfrastructure initiatives, Finholt notes that the development of collaboratories into their present conceptual and functional arrangement has not been guided by a master plan, but rather by emerging, adaptive mechanisms: "systems have emerged through a combination of prodding by visionaries, appropriation of technology designed for other purposes, and the marketing of low-cost high-performance personal computers" [9, p. 78]. The result is an infrastructure that functions at multiple scales of action: technological, socio-organizational, and institutional [25]. It is via continuous rearrangements, transfers, and interactions across and among these scales that a scientific collaboratory exhibits its emergent nature.

Boundary flexibility. Computing the range, population, configuration and boundaries of a system is never a trivial endeavor. As philosopher Cilliers reminds us:

> "In order to be recognisable as such, a system must be bounded in

> some way. However, as soon as one tries to be specific about the boundaries of a system, a number of difficulties become apparent" [26, p. 139].

Specifying what belongs inside a system and what belongs outside can be an arduous task. This difficulty has to do with the inherently open and dynamic nature of complex systems. Biological complex systems provide a convenient example to describe this boundary problem. A cell, for example, is delimited by a membrane which discriminates between what's inside (*life*) and what's outside (the living environment). Although the membrane is a salient and defining feature of the cell as a complex system, it is not a rigid boundary: it is active, "opening and closing continually, keeping certain substances out and letting others in" [27, p. 7]. Cilliers emphasizes that the boundary of a complex system resists the traditional notion of physical, rigid, system-independent *agent of closure*. He notes two peculiar aspects of boundaries in complex systems. The first is related to the very nature of the boundary:

> "We often fall into the trap of thinking of a boundary as something that separates one thing from another. We should rather think of a boundary as something that constitutes that which is bounded" [26, p. 140].

The second aspect relates to the spatiality of the boundary:

> "The propensity we have towards visual metaphors inclines us to think in spatial terms. A system is therefore often visualised as something contiguous in space. [...] Social systems are obviously not limited in the same way. Parts of the system may exist in totally different spatial locations" [26, p. 141].

This conceptualization of "boundary" applies particularly well to the notion of collaboratory. Unlike other types of organizational arrangements that enforce, or at least attempt to enforce, the existence of clear boundaries (e.g., institutions, departments, governmental centers), collaboratories are systems with flexible delimiters. In collaboratory research, there is a continuous, adaptive flow of interactions between heterogeneous actors — people, artifacts, ideas, etc. It is the "inside" activities—the processes—that determine the existence of a collaboratory as a distinct entity, regardless of its latent structure. Moreover, collaboratory systems consist of an array of physical and virtual environments, thus their boundaries are perforce untraceable in a solely physical space. Reflecting on the heterogeneity of domains, artifacts, spaces, and people that characterize modern scientific work, Star and Griesemer posited the idea of *boundary objects*—abstract or concrete objects that are "both plastic enough to adapt to local needs and the constraints of the several parties employing them, yet robust enough to maintain a common identity across sites" [28, p. 393]. Boundary objects may "inhabit several communities of practice and satisfy the informational requirements of each of them" [29]. The notion of boundary object is useful in this context as it poses a useful means to study the intersection among different domains, communities, and understandings. In more recent research, Charlotte Lee proposes to employ boundary objects to delineate, move across and push the boundaries of the communities of practices that adopt them [30]. As discussed more in detail in the methodology chapter of this dissertation (§ 3.2.1), the boundary problem—the impossibility to unequivocally specify the boundaries of a scientific research environment—has repercussions on the results and outcomes of investigations of this kind.

With these notions in mind, I regard collaboratories as dynamic, open, and emergent ecologies in which modern scientific research takes place in a highly

collaborative and distributed manner in an array of physical and virtual environments.

1.4 A network approach

Complex ecologies, including collaboratories and other scholarly and scientific collaboration endeavors, are frequently studied by the use of network analysis. Network analysis, also known as *graph theory*[2], is a specific branch of discrete mathematics that deals with the description and analysis of networks. But the conceptual notions underlying *a way of thinking in network terms* are widely applicable beyond mathematics and have been repeatedly used as a frame of reference in a number of different contexts: in the social and cognitive sciences, and in the humanities, for example. This dissertation employs a network approach to study scientific collaboration.

Why networks? Networks are the most convenient structure to represent and analyze interactions among the components of a system. Networks provide an abstract representation of a system, and allow researchers to study its function and organization. Network analysis is especially advantageous when dealing with complex ecologies, such as real-world social, biological and technological networks. The networks that represent these systems are inherently complex, i.e., they exhibit non-trivial topologies with characteristics that are neither random nor regular.

From a mathematical point of view, a network can be simply described as a set of *nodes* (also called vertices) with connections between them, called *edges*. This simple mathematical scheme can be used to represent many systems in the form

[2]Networks are oftentimes referred to as *graphs*, especially in mathematical literature. This dissertation uses these terms interchangeably.

of networks. In the real world, there are innumerable examples of networks: the Internet, social circles of friends (both online and offline), scholarly collaborations, food webs, postal delivery routes, aviation routes, metabolic networks, cellular networks, genetic networks, and neural networks, to name a few. Networks are ubiquitous. As Fritjof Capra reminds us, "wherever we see life, we see networks" [27, p. 8].

Network approaches have been advanced in domains as diverse as biology [31], economics [32], science studies [33], organization science [34], and cognitive science and artificial intelligence [35]. Of particular interest, and central to the topic of this dissertation, are social networks—networks that depict interactions among people. Social networks have been approached not only from a quantitative perspective [36, for a review of social network analysis], but also with respect to the broader cultural and sociological implications that become apparent when investigating a social phenomenon from a network perspective. Examples that fit within the latter body of literature include the work of Manuel Castells on the globalization of a network society and its effect on the economy, labor, and urbanism [37], that of Wellman and Haythornthwaite on the impact of the Internet and networked information on everyday life [38], and that of Mark Granovetter on the sociological significance of interpersonal ties [39]. The computational study of social networks (commonly known as *social network analysis*, or *SNA*) involves the construction and analysis of networks in which human agents represent the vertices and specific types of interaction represent the edges between them. These networks are utilized to perform a number of sociological investigations. For example, one can calculate the degree of a node in a network—the number of edges connecting to it—to measure the relative importance and centrality of an individual in a social network. Degree centrality and other foundational concepts of social network analysis are introduced and discussed in Chapter 3.

1.5 Organization of this dissertation

This dissertation is organized as follows: In the next chapter, I present in detail the subject of study of this dissertation, the Center for Embedded Networked Sensing (CENS), I lay out my problem statement, and discuss the contribution of this dissertation in relation to related literature. A general overview of the research methods and data sources employed throughout this dissertation is included in Chapter 3. These data sources are used to construct three networks of collaboration depicting coauthorship, communication, and acquaintanceship patterns at CENS. The mechanisms by which these networks are constructed, as well as their topological and socio-academic configurations are discussed in Chapter 4. In Chapters 5 and 6, I present the results of the structural and evolutionary analyses of these networks, respectively. These findings are summarized and discussed in Chapter 7. Chapter 8 concludes this dissertation by providing an analysis of the limitations and assets of this research, and delineating possible avenues for future work.

CHAPTER 2

Studying scientific collaboration

This dissertation examines scientific collaboration in the context of a specific research environment: the Center for Embedded Networked Sensing (CENS). In this chapter, I introduce CENS and outline my problem statement, discussing the contribution of this dissertation with respect to related literature on scientific collaboration.

2.1 CENS: Center for Embedded Networked Sensing

The Center for Embedded Networked Sensing (CENS) is a National Science Foundation Science and Technology Center established in 2002, involved in the development and application of wireless sensing systems to critical scientific and societal pursuits. CENS is a multi-institution venture which includes five member universities in California: University of California, Los Angeles (UCLA); University of Southern California (USC); University of California, Riverside (UCR); California Institute of Technology (Caltech); and University of California, Merced (UCM). CENS supports multi-disciplinary collaborations among faculty, students, and staff across disciplines ranging from computer science to biology, with additional partners in arts, architecture, and public health. More than 300 students, faculty, and research staff are associated with CENS. The Center's goals are to develop and implement wireless sensing systems and to apply this technology

to address questions in four scientific areas: habitat ecology, marine microbiology, environmental contaminant transport, and seismology. CENS also has projects concerned with social science issues, ethics and privacy, and citizen science.

CENS features a headquarter base located at UCLA, yet CENS-related work is conducted at all five member institutions, and at remote field-based locations, e.g., the James San Jacinto Mountains Reserve in Southern California. These institutions (and sometimes even departments) are sufficiently distant from one another to prevent continuous physical interactions among scientists: computer-supported communication is at the basis of their collaborative work. The type of research conducted at CENS spans a wide spectrum of disciplines and applications requiring continuous cooperation among individuals that, otherwise, would probably not interact beyond the walls of traditional university departments and faculties. In such a scholarly and scientific environment, distributed collaboration on multi-disciplinary subjects is a defining characteristic of scientific research.

The research work presented in this dissertation builds on a series of previous studies of scientific practice that address questions about the nature of CENS data and how these data are produced and managed. These studies, which incorporate both quantitative and ethnographic techniques, have documented the scientific practices and lifecycle of CENS research [40, 41, 42] and led to the construction of tools and services to assist in scientific data collection, analysis, preservation and sharing [43, 44]. It is through these investigations that I came to realize the complexity of scientific activities performed at CENS and subsequently conceptualize CENS as complex ecology.

From the above description, it is clear that CENS is similar in function and organization to the arrangement of a scientific collaboratory, presented in the previous chapter. It comprises an array of distributed physical and virtual envi-

ronments: physical, such as its headquarter laboratory at UCLA, faculty offices at member institutions, and sensor deployments in the field; and virtual, such as its mailing lists, wikis, digital fori, and other computer-supported communication platforms. Also, CENS research involves collaboration among researchers from different affiliations and disciplines; given the physical distance among member institutions, much research is conducted via computer-supported ICTs.

2.2 A scenario of scientific collaboration at CENS

Scientific activity in a collaboratory is governed by a rich array of interactions. In order to throw some light on such manifold combination of interactions, let us consider a realistic case scenario of CENS research. In sensor network research, a *deployment* is a research activity in which sensors, sensor delivery platforms, or wireless communication systems are taken out into the field to study phenomena of scientific interest. CENS deployments have taken place at numerous locations in California (at various national reserves, lakes, streams, and mountains) and around the world (including Bangladesh, Central and South America). One such deployment, currently carried out at the James San Jacinto Mountains Reserve in Southern California, involves the observation of bird breeding behavior via imaging sensors in a nestbox.

Avian research of this kind focuses on species of birds that nest in tree cavities. For this reason, CENS researchers have constructed wooden nestboxes and supplied them with imaging sensors (cameras). These sensors are located inside the nestbox, pointing downwards. The camera records still images documenting bird behavior during the breeding cycle. A number of other environmental data are recorded alongside images, in the vicinity of the nestbox, such as temperature, humidity, dew point, light intensity and soil moisture. A typical nestbox is

displayed in Figure 2.1; the image include labels for the nestbox's power supply (1), the mote, the sensor node (2) and the antenna (3). Two exemplary pictures produced by this camera are displayed in Figure 2.2.

Figure 2.1: A CENS nestbox

This scenario of scientific collaboration resembles a typical sensor network application in environmental field research. In fact, there are certain research activities that are common to all kinds of applications in environmental sensing: the design and construction of the sensor device, the capture and cleaning of the data, its analysis and the publication of the experimental results [44]. However, one can also perform an in-depth investigation of a deployment's narrative structure — a description of all the events related to a specific deployment and scientific project. One could analyze, for example, whether researchers identi-

Images from a nestbox: a Western Bluebird on the left, and four laid eggs on the right.

Figure 2.2: Images obtained from the imaging sensor in the nestbox.

fied a research problem in the field or in the laboratory, how they located field sites in which hypotheses were tested, how they assessed field sites for appropriate positioning of data collection equipment and sample acquisition, and the ways in which they calibrated equipment in the laboratory and the field. Moreover, what is crucial to the current discussion is the fact that all the described research activities necessarily involve human agency. In particular, the website of the aforementioned avian research project (`http://research.cens.ucla.edu/projects/2007/Terrestrial/AnimalCam`) lists the following people:

- Faculty: Deborah Estrin (UCLA), Michael Hamilton (UCR), John Rotenberry (UCR)
- Staff: Kevin Browne (UC Natural Reserve System), John Hicks (CENS), Jamie King (James Reserve), Mohammed Rahimi (UCLA), Michael Taggart (James Reserve), Tom Unwin (James Reserve)
- Graduate Students: Shaun Ahmadian (UCLA), Sean Askay (UCR), Sharon Coe (UCR)

Moreover, there is a mention of an external collaboration with Cornell University's Laboratory of Ornithology. The configuration for this specific project points to the geographically-distributed nature of CENS as a collaboratory. It follows that all the activities related to this specific deployment involve different configurations of researchers that might be physically based at remote locations and who might not know each other in person. All interactions between them might happen via face-to-face meetings or via computer-supported technologies of various kinds, e.g., email, dedicated electronic mailing lists, and social networking sites.

The researchers listed above were involved in various stages of the deployment. The deployment started with the formulation of the initial research hypothesis by avian researchers who may have documented them in grant proposals and other requests for financial support. These researchers, who may not be affiliated with CENS at all, may have identified in their proposal the use of wireless sensing as a possible solution for the study of bird breeding behavior. The design of the project and system development followed, with the construction of 13 nestboxes and associated micro-climate sensor systems and video cameras. A number of researchers and graduate students were involved in the initial device development and the exploratory data collection. As real data began to be collected, statisticians and computer scientists took part in this project to analyze and refine collected data. For example, computer scientists became involved in the project to develop an image recognition algorithm capable of detecting the bird's presence and the number of eggs present in the nestbox. This specific aspect of the project was documented as a case study in an article that was presented at a specialized conference on distributed smart cameras and imaging sensors [45]. Results of this deployment were also analyzed and summarized in another research paper, recently published, that discusses the use of imagers as biological sensors [46].

The author lists of the two articles written for this project overlap fairly well with the list of individuals presented above, but there are some discrepancies. For example, staff member Michael Taggart of James Reserve and Sean Askay of UCR are listed as collaborators of this project, but do not appear in the author list. The opposite case is also present. For example, Stefano Soatto of UCLA's Computer Vision Lab, involved in the image recognition work for this project, appears in the author list, but not as official member of this project. It follows that, in order to study collaboration patterns of this project, a bibliometric analysis of published articles alone would fail to reveal some important social interactions between graduate students, staff and faculty. CENS collaboration is carried out and manifested in many ways and is best studied by "triangulation" methods that rely on several types of data collected about a single phenomenon [47]. Many aspects relative to this methodology and its use in this dissertation are discussed more in detail in Chapter 3. At this stage, it is sufficient to note that this scenario demonstrates that collaboration at CENS does not necessarily reside in a single physical environment (e.g., the laboratory) or in a single procedure (e.g., the writing of scientific articles).

2.3 Problem statement

This dissertation analyzes scientific collaboration at CENS via manifestations of coauthorship, communication, and acquaintanceship. Why focus on these interactions? Although they are apparent from the scenario presented above, there are also other artifacts and collaborative interactions that encompass CENS research deployments: deployment plans, field notes, contextual data, software code, and raw sensor data, to name a few. Not only are collaborative interactions numerous

and heterogeneous, they are also potentially very different across different deployments, given the multi-disciplinary nature of CENS work. Gathering information about these collaborative interactions, the artifacts generated, and the narrative in which they are situated, would require an in-depth, longitudinal ethnographic study in the form of interviews and participant observation.

For the purpose of this dissertation, however, I limit myself to quantitative techniques of data collection, for two reasons. The first reason has to do with the breadth of my research, which is not limited to exploring in depth collaborative processes in a specific deployment or project. My intent is to capture collaboration patterns at a broader scale, i.e., collaborative interactions that are embedded within the modus operandi of the CENS collaboratory, as a whole. The second reason has to do with my choice of method. As I use a network analytic approach, I am specifically interested in manifestations of collaboration that encompass a form of interaction among people and that can be operationalized in a network format.

The preliminary portion of my dissertation research involved studying the CENS environment to identify collaborative interactions that are not unique to specific projects or work groups, but that are common within the community at large. In other words, I was interested in the aspects of the CENS collaborative culture that permeate single projects, disciplines, institutions, and sites. My preliminary exploratory research questions were: what kind of artifacts are being produced by CENS researchers? Where are they stored? Can I use them to extract information about collaboration? In the early stages of my research, I conducted an audit of CENS data repositories, finding that three categories of artifacts generated across the sensor network lifecycle were stored in specialized databases: contextual data, raw sensor data, and scholarly arti-

cles [43]. All these artifacts have the potential to provide information about collaborative interactions. The contextual data, hosted by the CENS Deployment Center (`http://censdc.cens.ucla.edu`), offer information about deployment teams, i.e., the individuals that participate in different deployments. The raw sensor data, hosted by SensorBase (`http://sensorbase.org/`), contain information about what data are produced and by whom. The bibliographic data, hosted by the eScholarship Repository (`http://escholarship.org/uc/cens`), provide author lists of academic papers. It is worth noting that all these repositories are still in their infancy: they are the result of recent initiatives to make CENS data public and openly accessible. As such, their information may be incomplete and inaccurate. A large portion of the data in SensorBase, for example, is stored for tests and demonstrative purposes only and lacks rich metadata about data authorship. Similarly, contextual data in the CENS Deployment Center does not cover the entire spectrum of CENS deployments. Of these three data repositories, only the scholarly database is updated regularly and is based on information that is directly extracted from official documentation—the CENS Annual Reports.

For these reasons, I choose bibliographic information contained in the official CENS scholarly record as the initial source of data to document collaborative interactions. Despite the variegated nature of artifacts generated across the CENS scientific lifecycle, many projects tend to culminate in some form of intellectual product that is published in the scholarly record: journal articles, conference papers, books, book chapters, posters, and technical reports. Scholarly publication is a crucial vehicle of scientific communication, dissemination, and recognition for researchers across CENS projects. It is also the official means by which CENS reports its accomplishments to funding agencies. As such, the collection of publications authored by CENS researchers represent the collaboratory's most accredited record of collaboration. In this dissertation, I use these bibliographic data to

construct a coauthorship network in which the nodes represent researchers and the edges denote the extent of coauthorship activity between them.

The second interaction analyzed in this study reflects the distributed nature of the CENS collaboratory. I have mentioned that because of the physical distance that separates researchers from different regions, institutions, departments, and workplaces, scientific work at CENS is often carried out via communication technologies. For example, the project discussed in the scenario above—the observation of bird breeding behavior by imaging sensors—involves researchers from institutions and field sites within and outside of Southern California. This set-up implies that research is partially carried out via various means of interpersonal electronic communication, e.g., personal email, dedicated mailing lists, and social networking messaging. Private forms of communication, however, are arduous to obtain because of their privacy and confidentiality. In this dissertation, I collect and analyze mailing list communication. Mailing lists are the principal mode of open communication among researchers at CENS. As the CENS wiki puts it: "CENS lives on mailing lists. This is perhaps the single most important form of communication within CENS"[1]. Using mailing list data logs, I construct a network in which the nodes represent researchers and the edges denote the extent of discussion activity between them on mailing lists.

All the scientific activities discussed thus far involve some form of social interaction between individuals. These interactions, however, may or may not involve physical interpersonal acquaintanceship. For example, researchers from different institutions and departments may have collaborated on several projects, software code, and scholarly papers, never having met in person. As scientific research moves to new organizational paradigms, predicated upon distributed collabora-

[1]From the CENS 3551 wiki, `http://lecs.cs.ucla.edu/wiki/index.php/3551`

tion, it becomes extremely important to investigate the forms of social agency that are at the basis of scholarly coauthorship and electronic communication. For this reason, I supplement my study with information about the role of interpersonal, offline relationships in the accomplishment of scientific work. I run a social survey to collect acquaintanceship data, and I use responses to the survey to construct a network in which the nodes represent researchers and the edges denote the extent of personal acquaintanceship between them.

Clearly, these three interactions do not exhaustively cover the entire spectrum of collaborative dynamics that take place at CENS, as explained more in detail in later chapters. Yet, they are prevalent among researchers of this collaboratory: scholarly coauthorship, communication on mailing lists and personal acquaintanceship are interactions that exist firmly within the modus operandi of CENS engineers, natural scientists, statisticians, computer scientists, sociologists, and life scientists alike. As explained, these interactions can be represented in the form of networks. In this dissertation I study how these networks evolve and interface with each other. My research involves the use of network analysis to study the structure and evolution of these networks with regard to specific aspects of scientific collaboration.

The first part of this dissertation is an analysis of *network structure*. The networks of coauthorship, communication and acquaintanceship briefly introduced above are essentially large structures that group individuals according to specific patterns of interactions. With this portion of research, I study the formation of patterns of activity in these networks, i.e., how the aggregation of individual interactions results in the formation of high-level structures. Via network analytic methods I explore the topology of these networks and their conceptual organization into clusters. In this context, it is interesting to analyze how these detected

cluster formations interface with each other. For example, a comparative analysis between the structure of the coauthorship and acquaintanceship networks can uncover clusters of individuals that, although not publishing articles together, are informally connected via frequent interpersonal relationships. Besides looking at the interactions in a comparative manner, I also record the context in which such interaction takes place (e.g., Who are the authors? What is their affiliation? Who do they know?). In other words, every individual in the networks studied is associated with a social/academic profile that contains information such as institutional and departmental affiliation, academic position, scholarly expertise, and country of origin. A comparative analysis of the detected topological structures with this set of attributes reveals the relationship between collaboration patterns and given organizational, disciplinary, and institutional arrangements of CENS. For example, the level of multi-disciplinary and inter-disciplinary collaboration at CENS can be investigated by analyzing the disciplinary affiliation within communities of frequent collaborators in the coauthorship network. With these considerations in mind, my first research question can be summarized as follows:

> **Research Question #1.** What types of structural communities can be detected in the coauthorship, communication, and acquaintanceship networks of CENS? How do these structures relate to each other and to the disciplinary and institutional arrangements of CENS?

The second part of my research is an analysis of *network evolution*. With this portion of research, I inquire into the scientific collaboration dynamics that take place at CENS. The networks of collaboration studied here — coauthorship, communication, and acquaintanceship — are not static structures: they are all time-dependent. For example, the network of coauthorship represents the act of

collaborative writing of an article. Coauthoring events take place during fixed point in time, "time-stamped" by the date of publication. Similarly, mailing list discussions recorded in the communication network are time-stamped by the email protocol. Finally, acquaintanceship networks are also dynamic, for individuals get to know each other and relationships endure through time. With these notions in mind, I can look at evolutionary patterns in the aforementioned networks using a comparative approach. For example, I can uncover whether specific collaboration patterns emerge in one environment (a mailing list discussion, for example) and then spill into another (the coauthoring of a paper, for example). As for the structural analysis introduced above, I extend my evolutionary analysis to include information such as institutional and departmental affiliation, academic position, and country of origin. This allows me to study the researchers' propensity to collaborate preferentially over time with others with a similar social and academic profile. In turn, this provides an understanding of the ways by which CENS scientific communities are formed and modeled in relation to the social and academic contexts in which they are embedded. My second research question can be summarized as follows:

> **Research Question #2.** What collaboration dynamics can be evinced from the coauthorship, communication, and acquaintanceship networks of CENS? Can specific evolutionary features be explained in terms of changes in the disciplinary and institutional arrangements of collaboration?

2.4 Review of related literature

In the previous section, I outline my problem statement and present my research questions. In order to frame this dissertation in the existing body of literature, I review here previous work which addresses similar research questions to the ones set forth in this dissertation. I review specifically studies of collaboration that analyze coauthorship, communication, and acquaintanceship patterns in science.

2.4.1 Literature on coauthorship and scholarly collaboration

The study of coauthorship falls within the broader research field known as bibliometrics, "the branch of library science concerned with the application of mathematical and statistical analysis to bibliography; the statistical analysis of books, articles, or other publications"[2]. In the context of this dissertation, scholarly output in the form of bibliographic material is a tangible indicator of scientific collaboration and can be conveniently analyzed by the use of bibliometrics. Coauthorship is a prominent indicator of collaboration in scholarship. Authorship, in particular, is of considerable importance both for the public and for researchers. The public is interested in knowing the exact source of novel ideas and research work. In turn, public recognition functions as a lever for researchers who, in publishing their work, become more visible in academic circles, to funding bodies and on the academic market.

It is worth mentioning that bibliometric methods revolve around the study of a number of other indicators of scholarly activity besides coauthorship; these include broad categories of research in citation, co-citation and acknowledgment networks [48, for a review]. Yet, coauthorship patterns are perhaps the most

[2] "bibliometrics, *n. pl.*" The Oxford English Dictionary. 2nd ed. 1989. OED Online. Oxford University Press.

studied scholarly and scientific phenomena. Notable studies of coauthorship have analyzed the literature production within specific domains. Recent work of this kind include investigations of the domains of high energy physics [49], genetic programming [50], neuroscience [51], and nanoscience [52]. Bibliometric analyses are not exclusive to the sciences: domain-specific studies have mined bibliographic databases in fields as diverse as digital library research [53], economics [54], organizational science [55], and psychology and philosophy [56]. These kinds of domain-oriented analyses are also comparative in nature. A noteworthy cross-domain large-scale comparative analysis is presented by Mark Newman, who analyzes large databases of papers in the fields of physics, biology, and mathematics exploring social and normative domain differences of coauthorship behavior [57]. Relevant to this line of work are a number of studies that employ network analysis to study coauthorship patterns in academic and scientific circles. Börner, for example, posits a weighted graph approach to identify the local and global properties of a scientific coauthorship network to document the emergence of a novel field of science [58]. All these network-based analyses have proved viable for a number of visualization studies that employ graphical representations of coauthorship networks to uncover macroscopic patterns that network analysis alone might fail to reveal [59, 60]. Moreover, an increasing number of studies of this kind have accounted for the evolving component of scientific collaboration [61, 62, 63].

Coauthorship patterns have been widely and actively studied from a social network analysis perspective for over two decades [47, 64, 65]. Most social network research involved with coauthorship is based upon this underlying concept: two (or more) individuals are regarded as coauthors if they appear together in the author list of a publication. This study technique works reasonably well to investigate coauthorship patterns in the traditional arrangement of scholarly

publishing. However, authorship models — especially in the context of large collaborations in the physical sciences — have been undergoing a drastic shift marked by a substantial increase in the number of authors per publication—a phenomenon known as "hyperauthorship" [66, 67]. Such increase, coupled with inconsistent authorship practices, makes it impossible to discern the nature and extent of individual contributions to a work or a publication; it is difficult to distinguish between principal authors, research assistants, project advisers, and honorary authors [68]. Hyperauthorship, although not visible in sensor network research (as discussed more in detail in chapter 4) — is a natural consequence of the fact that certain scientific endeavors often require use of large-scale instrumentation that could not be possibly fabricated and employed by few individuals within a small research group. A striking example of this phenomenon is the domain of high energy physics where author lists for a single publication often comprehend tens or even hundreds of researchers [69, 70]. For this reason, a number of recent studies of coauthorship in the physical sciences supplement traditional analytic techniques (i.e., detecting coauthorship from author lists) with more qualitative methods of survey research (i.e., directly asking authors to indicate the real nature of their contributions to a publication) [71, 72].

By performing cross-domain comparative studies, research in bibliometrics has progressively introduced indicators such as multi- and inter-disciplinarity in its body of research. For example, the nature of multi-disciplinary work has been investigated in the fields of information science [73], nanotechnology [52], and the social sciences [69]. Highly interdisciplinary research centers have also become interesting environments to study collaboration. The ensemble of social, academic and demographic characteristics found in these centers offers a convenient platform to study indicators that can enrich the understanding of collaboration patterns. Research interests [74], and academic domain [75] are

examples of characteristics that have been analyzed in bibliometric studies in relation to coauthorship behavior. Academic institution is an obvious indicator to explore how coauthorship patterns are distributed across different affiliations and geographical locations. In a large-scale analysis of coauthorship in physics, Lorigo and Pellacini find that there is a steady growth of inter-institute and cross-country collaborations over a period of three decades [76].

Some preliminary results obtained using earlier versions of the data and methods employed in this dissertation have been published in specialized literature. In particular, I have published (jointly with Marko Rodriguez) a study of inter-disciplinary and inter-institutional scientific collaboration, via a longitudinal analysis of mixing patterns of the coauthorship network [77], and a structural analysis of the same network to reveal the interdependence of structural communities and socio-academic characteristics of scientific collaboration circles [78].

Moreover, it is worth noting that, besides analyzing scientific collaboration as measured by the extent of coauthorship activity, I have become interested in the analysis of the content of the coauthored bibliographic material. In recent work, I employed abstracts and manuscripts authored by researchers at CENS to construct a network of intellectual exchange [79]. Networks of this kind—sometimes referred to as co-word or epistemic networks—link individuals that employ the same topics and knowledge constructs in their scholarly production. A number of studies, both quantitative and qualitative in nature, exist that employ networks of intellectual exchange to mine scientific collaboration. An example is the work of Callon, Law and Rip [80] that builds on the theoretical framework advanced by the Actor Network Theory [33] to investigate the dynamics of scientific and technical production analyzing the specific concepts contained in articles in a qualitative manner. Another example is the work of Leydesdorff

by which he illustrates a qualitative manual mechanism to deconstruct a scientific article in sentences, paragraphs, sections and construct epistemic networks at different levels of aggregation [81]. Qualitative methods of co-word analyses have also been blended with quantitative co-citation techniques as well as survey research. Recently, constructs similar to those proposed by Leydesdorff and Callon have been extracted and used for a number of ad-hoc analyses that often involve semi-automated components. For example, semantic tags (user-assigned keywords) have been employed to perform knowledge discovery and recommendation in large database systems [82], co-citation and co-word analyses have been combined to explore competing scientific paradigms in the real world [83], and quantitative and qualitative bibliometric maps visualizing scientific knowledge domains have been blended to provide new perspectives on science-policy related problems [84]. Despite the recent proliferation of semi-automated techniques, the bulk of research in this domain seeks to find efficient, fully automated procedures of topic extraction from texts. Some methods are grounded in machine learning, using very large corpora of data to train the extraction algorithm, while others rely on basic textual parsing techniques matched with a controlled vocabulary. A recent example is a study of interaction histories of personal email archives, in which Viègas, Golder, and Donath [85] develop an extended version of Salton's term frequency-inverse document frequency algorithm[3] of relative frequency [87].

2.4.2 Literature on online communication

Electronic communication can take place on a number of different online platforms such as emails, wikis, blogs, mailing lists, web fori and newsgroups. Online

[3]The term frequency-inverse document frequency (TF/IDF) weighting algorithm was first developed by Karen Spärck Jones [86]

users contribute to conversations taking place on these digital, interactive environments for a number of reasons: "for debate, to express appreciation or affiliation, to build a sense of community, to provide and receive social support, to collect information, and to provide answers to questions" [88, Introduction]. As anticipated above, my research on communication networks involves the study of interpersonal communication that is performed among CENS individuals using public mailing lists dedicated to CENS research work. Thus, in this section I specifically review literature that investigates email-driven forms of communication, such as analyses of personal emails, mailing lists and newsgroups. Content and network analyses of blogs and wikis are excluded from the current discussion.

With the ongoing proliferation of Internet and communication technologies, electronic communication functions as a mirror of intricate patterns of interpersonal physical and virtual communication. For example, different patterns of email interaction might correspond to different social structures: dense email communication might represent strong, informal, personal ties and less frequent communication might represent more formal connections [89]. Overall, recent research about electronic communication has motivated investigations addressing the formation and convergence of online and offline communities. Notable work in this field includes a study of the impact of the Internet on the "social capital" of virtual and physical communities [90], the shift from physical densely-knit networks of communication to sparsely-knit, geographically-unbound "individual" networks [91], and the effects of informal communication on the productivity of communities of practices in organizational settings [92].

An important characteristic common to all electronic communication data is its threaded structure [88]. Since all these environments (emails, wikis, mailing lists, newsgroups) represent essentially a conversation among different members,

every conversation can be represented as a thread whose nodes are single conversation events. In an email system, for example, a thread is the collection of emails around a certain topic (e.g., same email subject) and the nodes are represented by every email in the thread. Given the structure of the email protocol, it is possible to deduce the characteristics of the node (e.g., the names of senders and receivers from the `to:` and `from:` fields). This threaded structure is common to all other aforementioned electronic environments, and thus enables a number of network analytic studies of electronic communication traces.

A large number of early studies of communication activity on the web revolved around *Usenet*, a general-purpose world-wide distributed Internet discussion system established in the early 1980s and still in use today[4]. In a thorough review of the social structure of Internet discussion platforms, Marc Smith describes Usenet as "a quintessential Internet social phenomenon: it is huge, global, anarchic, and rapidly growing" [93, p. 195]. Due to its distributed, heterogeneous, open nature, the Usenet database has been mined in a number of different studies to reveal a number of different facets of electronic communication. Communication research centered around Usenet newsgroup data includes analyses of racial identity [94], comparisons of behavioral metrics with users' subjective evaluations [95], and an early implementation of collaborative filtering algorithms on large datasets [96]. More recent studies have analyzed social roles on online discussion groups [88] and the variations in hierarchies, newsgroups, authors, and social networks over time [97]. Very often, these analytical studies have been implemented with mapping, visualization and browsing interfaces for Usenet [97, 98, 99].

Research on the communication activity of newsgroups, mailing lists and similar online fori has fostered thanks to the wide availability of data: these data,

[4]The volume of posts on Usenet has increased steadily since its inception. `altopia.com` reports an increase from 4.5 Gigabytes of posts in 1992 to 3.8 Terabytes in 2008 (estimated).

both past and current, are often widely available to the public. Analyses of personal email and privately-owned mailing lists, instead, have been less prominent both for the difficulty to obtain permission to analyze privately owned data and for the ethical implications of performing research on purely personal data. Yet, analyses of personal email logs can be tremendously revealing of social structure. This is for a number of reasons: (i) nowadays, personal email is by far the predominant means of electronic communication, (ii) personal email conversations can range from formal to informal, (iii) personal email data span the widest possible spectrum of topics (unlike dedicated mailing lists or newsgroups), and (iv) email pervades business, social, technical and knowledge exchanges. Analyses of email data have been performed in the context of health care to understand the nature of communication between patients and medical providers [100], in social contexts to identify relationships using the interaction histories [85], in organizational settings to reveal demographic and occupational differences by email signatures [101]. Email-driven communication has also been analyzed in the context of scholarly and scientific environments (the focus of this dissertation): Matzat [102], for example, performs a comparison between the nature of knowledge transfer and social activity on academic Internet Discussion Groups (IDGs), finding that IDGs better support formation of social contacts rather than academic communication.

Particularly relevant to the topic of this dissertation is a study by Tyler, Wilkinson and Huberman that mines the personal email logs of a corporate organization to detect communities of practice among its employees [103]. This study intentionally neglects messages "sent to a list of more than 10 recipients, as these emails were often lab-wide announcements (rather than personal communication), which were not useful in identifying communities of practice" [103, p. 7]. Yet, the study is methodologically similar to the work presented in this dissertation, for it

constructs a network of communication (producing a vertex for every individual and drawing edges between people who corresponded through email) and it identifies clusters of individuals from the network's topological structure. Moreover, this study matches the detected structural communities to a specific characteristic of the network: the corporate hierarchical position of its constituent individuals. This dissertation extends the study by Tyler and colleagues in two ways. First, it seeks to detect coherent topological structures in large-scale discussion-based mailing list logs, rather than purely personal communication traces. Second, it extends the set of studied characteristics beyond hierarchical position (i.e., academic position) to a number of other characteristics, such as scholarly expertise, and departmental and institutional affiliation.

2.4.3 Literature on acquaintanceship and social relationships

Studies that analyze the rules and structures of acquaintanceship patterns have traditionally been found in domains such as sociology, anthropology, and organizational science and management. Acquaintance is the relationship among individuals defined by personal knowledge that is "more than mere recognition, and less than familiarity or intimacy"[5]. An acquaintanceship network is a type of social network in which the vertices of the network represent individuals and the edges represent varying degrees of acquaintanceship. In sum, an acquaintanceship network describes "who knows whom, and how" [104, p. 381] in the environment under study.

In the field of computational sociology a number of social network studies have analyzed the patterns and structure of acquaintance [36, for a review]. The nature

[5] "acquaintance, *a*" The Oxford English Dictionary. 2nd ed. 1989. OED Online. Oxford University Press.

of acquaintanceship has been studied using methods that range from ethnography to questionnaires to semi-structured interviews; examples are business contacts within and among companies [105], friendship [106], sexual relationships [107], and reciprocal relationships [108] in social circles.

An experiment that marked a new trend of studies in sociology is the so-called *small-world* experiment of Stanley Milgram [109, 110] in which individuals in selected U.S. cities were asked to pass a letter to one of their close acquaintances that they thought had the highest probability to pass it to an assigned target individual. This experiment was groundbreaking for it demonstrated that the letters that made it to the target destinations, did so in about six iterations — demonstrating an average path length of six for social networks of people in the United States. This idea is nowadays popularly known as *six degrees of separation*. Since the original experiment by Milgram, a large number of investigations have appeared in the literature. Notable studies directed at the study of small-world networks — networks in which most pairs of vertices are linked by a short path — include analyses of the "trails to Paul Erdös" [111], searching global networks [112], a reversal of the small-world experiment [113] and a critique of the idea itself based on the social nature of acquaintanceship [114]. Watts and Strogatz have found that many real networks, notably social and scientific collaboration networks exhibit small-world properties [115]. In the domain of complex systems, small-world networks have been employed to investigate emergence [116] and to test a typical problem of game theory known as the prisoner's dilemma [117].

With growing availability of online corpora containing social indicators, many social network analyses concerned with acquaintanceship have relocated to the web. Business and scientific contacts, friendship and romantic relationships de-

picted by dedicated social networking sites have been increasingly used for large-scale analyses. Examples include a comparison of online and offline friendship ties [118], an analysis of online friendship in terms of social capital [119], and the structural and temporal evolution of a dating website community [120].

A number of studies on acquaintanceship and interpersonal networks have been carried out in scientific communities. Employing questionnaires and semi-structured interviews, two early studies of this kind analyzed the consequences of using a computer-based conferencing platform [121] and differences in task-based workflow [122] on the structure of interpersonal ties among scientists in various scientific domains. Another study has assessed the impact of scientific meetings on the knowledge flow among scientific contacts [123]. More recent investigations include work on "virtual science" laboratories by Chin, Myers, and Hoyt [124] who explore the central role of personal knowledge exchange, both formal and informal, both work-related and not, in the context of evolving scientific social networks. Hara, Solomon, Kim and Sonnenwald [125] adopt ethnographic methods to capture perception regarding collaboration and work practices in a multi-institutional interdisciplinary research center. Their methods include direct interviews, observations of videoconferences and meetings, and sociometric surveys. In other work, Stokols, Harvey, Gress, Fuqua, and Phillips [126] employ *in vivo* techniques to observe collaboration factors such as personal compatibility, work connections, prior projects, spatial proximity, and face-to-face interaction to depict the extent and impact of interpersonal activity and acquaintanceship on the process of large-scale scientific collaboration.

Relevant to the research presented in this dissertation is the work of Nardi, Whittaker, and Schwarz on *intensional networks* [127]. Based on a qualitative study of collaboration across organizational boundaries, Nardi *et al.* explore the

importance of personal networks for labor management and coordination. They find that collaborative synergies are more often the result of assemblages of people found through personal networks rather than the outcome of organizational planning and structuring. A similar line of work, emerging from site-specific ethnographic observations, also points to the existence of loose arrangements of individuals with given roles, known as *knotworks*, which span organizational and institutional boundaries [128].

Another classic study that is particularly relevant to the work presented here is a network analytic study of the field of biomedicine performed by Lievrouw, Rogers, Lowe, and Nadel [47]. This study uses a methodological strategy known as "triangulation," which involves gathering and employing multiple data sources about a single social phenomenon [129]. Employing quantitative analyses of co-citation, co-word and co-authorship on a database of research grants awarded by the National Institute of Health, Lievrouw *et al.* find that there exist a discrepancy between the communication networks of scientists and the content of the work in which they engage. The work of Lievrouw *et al.* ultimately aims to explain "why and how scientific knowledge [grows] as a function of both formal and informal communication networks" [47, p. 217]. In other words, its objective is to identify large-scale, global structures of interpersonal intellectual exchange, or "invisible colleges" [130, 131]. From a methodological and conceptual perspective, this dissertation draws extensively from this study: both approaches involve the investigation of a scientific environment using different techniques as well as a comparative analysis of the results to explain the growth and structure of social and scholarly networks. My research, however, places the focus on local, small-scale forms of organization within a specific workplace, rather than on the looser global structures of scholarly collaboration.

2.5 Contribution of this dissertation

The intellectual contribution of this dissertation falls within several rubrics. First, as my research examines the topological and structural configuration of CENS collaboration, it complements previous research focused on the topology of scientific networks. As discussed in the previous section, much of this research has found that scientific and social networks exhibit small-world topological properties [61, 132, 133, 115]. Research of this kind is often performed on large, but homogeneous networks, constructed from domain-based bibliographic repositories and well-delimited social circles. The CENS collaboration ecology does not fit these canons. Its research focus, which involves a mosaic of disciplines—from electrical engineering to statistics, from biology to sociology—and its boundary flexibility make it an interesting environment to explore how different scholarly and social practices coalesce in multidisciplinary, multi-sited scientific ventures. This study tests the small-world hypothesis and analyzes preferential attachment rules in the context of such heterogeneous collaborative environment. This contribution is discussed in much detail in the Discussion chapter of this dissertation (§ 7.2).

The second aspect of the contribution has to do with my analysis of the organization and function of CENS in relation to previous research on collaboratories. To date, much research on collaboratories and cyberinfrastructure has been qualitative in nature. The focus of these investigations has been on the relations and interconnections that are at play in cyberinfrastructure [12, 13]. As such, these studies have stressed the importance of *relations* by analyzing the human infrastructure that supports and fuels these initiatives [14]. My research is an attempt to provide a visual and empirical map of these relations by the use of quantitative tools of network analysis. I use network analysis to describe

the scholarly, communication, and social relationships that are at play within the CENS collaboratory. In particular, via a structural analysis, this research explores the local, small-scale organization of the collaboratory: how researchers who perform similar or joint research work organize themselves in small groups. With regard to this, my research explores and substantiates previous work that draws attention to the importance of interpersonal networks and social cohesion in cyberinfrastructure initiatives [127, 128]. This is further discussed in § 7.3.

Third, my research allows me to reflect on the benefits and usefulness of using a complex network approach to study collaboratories. In this dissertation, I conceptualize the CENS collaboratory as a complex system and employ methods of network analysis to explore its collaborative configuration. I outline emergence and boundary flexibility as two salient characteristics for the function of many complex systems. How well do these characteristics become apparent from my study of scientific collaboration? My research allows me to revisit these notions in the light of the obtained results and my personal four-year-long experience with CENS. My reflections on the choice of methods pose as recommendations for researchers that intend to employ a complex system approach for the study of science, as discussed in § 7.4.

Finally, my research brings about a methodological contribution to the field of network analysis. Although my method is essentially quantitative, my research questions, my analytical framework, and the interpretation of my results are guided by considerations that are qualitative and sociological in nature. For this reason, my method is enhanced with qualitative connotations that are normally neglected by studies of scientific networks. Many network-based studies of scientific collaboration are based on large, domain-centric, bibliographic repositories. As a result, these studies rely on a wealth of bibliographic data but only

examine a single manifestation of collaboration. My work puts into action the triangulation strategy posited by Lievrouw *et al.* [47], to explore the multi-faceted nature of scientific collaboration. Moreover, thanks to the manageable size of the CENS network, I am able to collect a wealth of socio-academic information by manually inspecting the personal web pages and biographies of each individual in the network. I employ this information to frame my network analysis results in the broader social and academic landscape in which collaboration takes place. I speculate that the relatively limited size of the constructed network, my familiarity with the underlying data, and my privileged position of information scientist "embedded" in the network, allow me to provide more nuanced interpretations of my results than those obtained in large-scale network analyses. These and other considerations for network researchers and social scientists are discussed in § 7.5.

CHAPTER 3

Research methods, data and instruments

This dissertation is a study of scientific collaboration at the Center for Embedded Networked Sensing (CENS) via an analysis of its coauthorship, communication and acquaintanceship patterns. The research methods, data and instruments employed in this study are discussed in this chapter. In particular, this chapter begins with an overview of the foundational concepts relative to network theory, which is the the overarching methodological framework of this dissertation. Concepts that are discussed in great detail include the community structure and the assortative mixing of networks, which are at the foundation of my study of network structure and evolution. In the following section, I present the data and research instruments. In particular, I introduce the techniques by which coauthorship and communication data were collected and how they were employed to delineate the range and composition of the population under study. In the final section of this chapter, I introduce the social survey instrument that was administered to the CENS population to gather data about personal knowledge patterns and construct an acquaintanceship network.

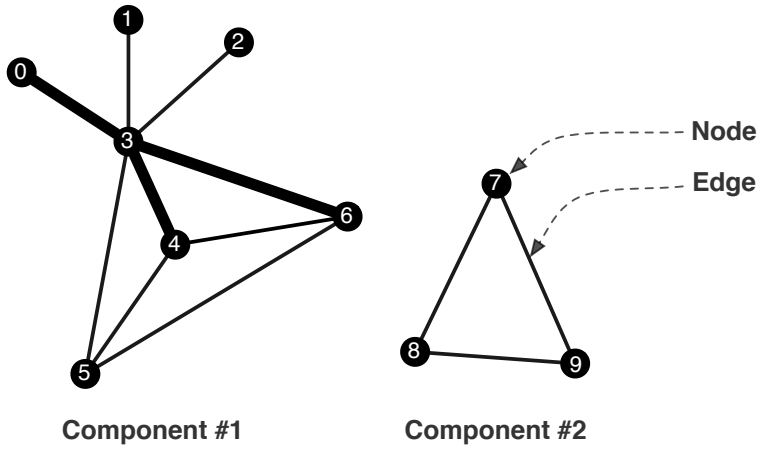
3.1 Foundations of network analysis

As anticipated in Chapter 1, the research presented here employs a network approach to study scientific collaboration. By representing interactions in a network fashion, one can rely on a platform of well established tools and methods of network analysis to depict and study in detail these interactions.

In a comprehensive review on complex networks, Mark Newman divides networks into four loose categories — social, information, technological, and biological networks — based on the general properties of the interactions they represent [16]. *Social* networks represent patterns of social interactions among people. Friendship, acquaintanceship, kinship, business relationships, and sexual relationships are examples of interactions that can be represented by a social network. *Information* networks depict information and knowledge exchange. An example of information network is a scholarly citation network which represents referencing patterns between academic papers. Similarly, the World Wide Web is an information network for it represents linking patterns between web pages. *Technological* networks are man-made networks developed for the distribution of a commodity or resource. The electrical, road, railway, and postal networks are examples of technological, infrastructure networks devised for the distribution of electricity, vehicles, trains and mail, respectively. *Biological* networks represent interactions within or between living organisms. Widely studied biological networks include food webs that depict preying patterns in an ecosystem, and neural networks that depict the structure of the brain.

The networks discussed in this dissertation are all social networks, i.e., they depict interactions between people. In particular, the three interactions covered here are coauthorship of scholarly papers, communication on mailing lists, and personal acquaintanceship. Using a network approach, these interactions can be

described using the same underlying scheme: individuals are represented as *nodes* in the network; nodes are connected to one another by an *edge* if a relationship between them exists. This very simple scheme is depicted in Figure 3.1, for a fictitious network with 10 nodes and 12 edges.



This example network features 10 nodes and 12 edges ($n = 10, m = 12$) in two separate connected components. Nodes 0 *through* 6 *are part of the first component; nodes 7, 8 and 9 are part of the second component. Line width is proportional to edge weight, so that more prominent connections are depicted by wider lines.*

Figure 3.1: A small example network.

3.1.1 Basic properties of networks

Directionality. The network of Figure 3.1 is *undirected*, for the edges connecting nodes do not have directionality (i.e., arrows). Directionality is important in many kinds of networks. A food web (*who eats whom*) is an example of network in which the direction of the edges is fundamental. The networks employed in this dissertation are all regarded as undirected graphs. Scholarly coauthorship and mailing list communication are natively undirected interactions: coauthoring a paper or participating in a discussion are actions that do not involve directionality.

Acquaintanceship, however, is a directed activity: when someone claims to know someone else, this interaction may or may not be reciprocated. For simplicity, and in order to enable a comparative analysis, the acquaintanceship network studied in this dissertation is however regarded as undirected. This matter is discussed in much detail in Chapter 4.

Weight. The network of Figure 3.1 is *weighted*, for its edges are associated with weights. The weight of an edge indicates the intensity or the extent of a given interaction between two nodes. For example, in a communication network, a low-weight edge indicates sporadic communication between two individuals, while a high-weight edge indicates frequent communication. In network visualization, weights are oftentimes represented by edge widths, whereby heavier weights have wider and more marked lines. In Figure 3.1, three edges have higher weights than the others (0—3, 3—4, and 3—6). All networks studied in this dissertation are weighted. The mechanisms used to assign weights are explained in much detail in Chapter 4.

Components. The network of Figure 3.1 has two connected components: the one on the left is composed of 7 nodes, and one on the right is composed of 3 nodes. A connected component is a set of nodes that can be reached by paths running along edges of the network. It is normal for networks to have more than one connected component, i.e., to be partitioned into disconnected groups of nodes. However, many real and artificial networks feature a *giant component*, i.e., a large connected component which is made up by the majority of the graph's nodes.

Diameter. Any two nodes can be connected by different paths running along the edges of a network. A *geodesic path* is the shortest of these paths. The diameter of a network is the length (in number of edges) of the longest geodesic

path between any two nodes, i.e., the distance between the two most remote nodes. As such, the diameter gives an immediate idea of the size and breadth of a network. When a network features multiple components, the diameter is always calculated on the giant (largest) component. In Figure 3.1, the diameter of the graph is 2, since the longest geodesic path in component #1 is only made up of two steps (e.g., the path to get from node 1 to node 5).

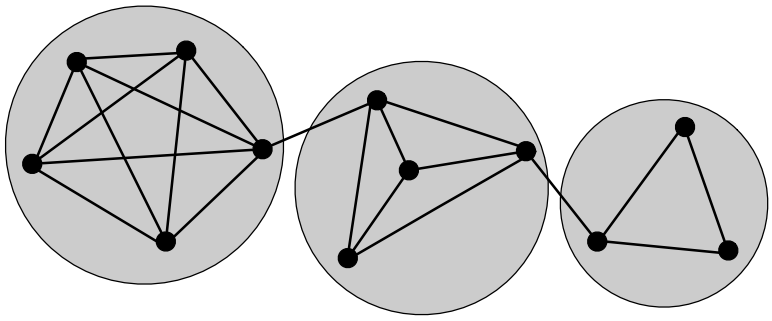
Average path length. The average path length of a network is obtained by computing the average of the shortest path lengths between all possible node pairs. Average path length is an indicator of the efficiency of information transfer in a social network. In the network of Figure 3.1, the average path length (of the giant component) is 1.57, calculated by averaging all shortest paths between nodes 0 through 6.

Node degree and centrality. The degree of a node is the number of edges that connect it to the rest of the network. For example, in Figure 3.1, node 3 has a degree of 6, while node 8 has a degree of 2. Node degree is used to compute the most basic form of network centrality, whereby nodes with the highest degree are considered more central than others. In many network visualizations, the diameter of depicted nodes is adjusted according to centrality scores.

Degree distribution and preferential attachment. A much studied quantity in networks is their degree distribution, i.e., the frequency distribution of degrees of individuals nodes in a network. A degree distribution is normally displayed as a plot of node degrees on the x-axis and their cumulative frequency on the y-axis. Random networks display Gaussian and centered degree distributions, since all connections between nodes are equally probable. Real and natural networks, however, have highly skewed degree distributions, with a majority of nodes of low degree and a small number of nodes with high degree. Many social

networks have been observed to follow an exponentially decaying or a power-law distribution. A network whose degree distribution follows a power-law is commonly known as *scale-free* [134]. Scale-free degree distributions have historically been observed in many types of scholarly networks [135, 61, 132, 63]. It has been widely argued that one of the predominant generative mechanisms behind the formation of power-law degree distributions is *preferential attachment*, i.e., the notion that a network grows (i.e., nodes attach to each other) according to specific preferential rules.

Cliques. Cliques are subsets of a network within which every possible edge exists. For example, in a social network of acquaintanceship, a clique is a group of people wherein all know each other. In a coauthorship network, a clique is a group of authors all of whom collaborate with each other. A clique is maximal if it cannot be extended to a larger clique. Figure 3.1 has two maximal cliques: a 4-node clique in the first component (nodes 3, 4, 5, and 6), and a 3-node clique in the second component (nodes 7, 8, and 9). Some other examples of maximal cliques are shown in Figure 3.2. From a sociological perspective cliques are interesting network structures because they represent tight-knit groups of interconnected individuals who exclusively share specific characteristics and patterns of behavior.



A 5-node clique (left), a 4-node clique (center) and a 3-node clique (right).

Figure 3.2: Three examples of maximal cliques.

Clustering coefficient. The clustering coefficient measures the density of cliques in a network and indicates the extent to which nodes in a network tend to cluster together. In other words, the clustering coefficient gives an indication of how many closed triangles there are in a network. From a sociological perspective, this notion is important since if A knows B and B knows C, there is a probability that A also knows C (and thus a closed triangle is formed). For this reason, network clustering is also knows as network transitivity [36]. The clustering coefficient of a network, C, is computed by the following formula [16]:

$$C = \frac{3\times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}}, \tag{3.1}$$

which for the network of Figure 3.1 is $C = 3 \times 3/18 = 0.5$.

Small-world networks. High clustering coefficient coupled with short average path length indicates that a network exhibits *small-world* properties [16, 115]. A *small-world* is a network in which any two nodes are only a few steps apart, regardless of network size. In a small-world network, individuals are not necessarily all connected to each other, yet they are easily reachable from one another via a short path.

3.1.2 Community structure

Much research in network theory revolves around the study of structure. As discussed in Chapter 1, from a theoretical perspective, structure appears when recurrent small-scale interactions among agents endure into large-scale properties. In network terms, structure refers to high-level topologies that are separate from individual small-scale interactions. An immediate manifestation of structure in a network comes from its natural subdivision into clusters. Clusters are

groups of nodes that are connected with one another to form a separate group. In a social network, a cluster of this kind is a community of individuals related to each other by a high level of interaction. The formation and advancement of communities in the workplace has been studied for decades. In an exploration of situated learning and knowledge acquisition, Lave and Wenger [136] introduced the notion of *community of practice* to delineate groupings of individuals with shared goals and sociocultural practices. More specific to knowledge activities in the scientific workplace is the notion of *epistemic community*, which refers to groups of scientists producing knowledge according to a common framework of conceptual tools, representations, and expertise [137]. These notions have been employed extensively in studies of scientific and non-scientific organization rooted in sociology, information science, and knowledge management [92]. Despite some attempts to align social network topology with community of practice and epistemic community theories [138, 139], most of these studies are qualitative in nature, however. Yet, it is important to stress that both communities of practices and epistemic communities denote an underlying structural component: they imply that researchers coalesce and organize themselves to form a bounded group [140]. For this reason, studying communities of scientific collaboration from a structural perspective is of fundamental importance.

From a computational perspective, there are two broad sets of methods to detect clusters in networks. For unweighted graphs, a widely accepted method is the *K-means* algorithm that clusters network data in a number of given partitions [141]. For weighted graphs, such as the ones discussed in this dissertation, several techniques have appeared in specialized literature in the past two decades. These techniques are based on algorithms that partition a network into *structural communities*, i.e., they reveal the network's *community structure*. Structural communities are "cliquish" sub-graphs composed by groups of vertices that are highly

connected between them, but poorly connected to other vertices [142]. The study of community structure in networks is particularly important because communities might display local properties that differ greatly from the properties of the network as a whole. Even a very detailed analysis of a network at a global level might fail to uncover specific patterns and characteristics that only exist within tight-knit communities and sub-partitions of the network. Figure 3.3 shows a simple network partitioned into three structural communities. It is important to stress the difference between Figures 3.2 and 3.3. While Figure 3.2 displays three cliques, the structural communities of Figure 3.3 are not cliques *per se*: they are cliquish clusters of nodes. It follows that structural communities are not communities in a natural sense, i.e., the members of these communities are not necessarily all connected to each other, as in a clique. Rather, membership to a community is predicated by the overall topological features of the network.

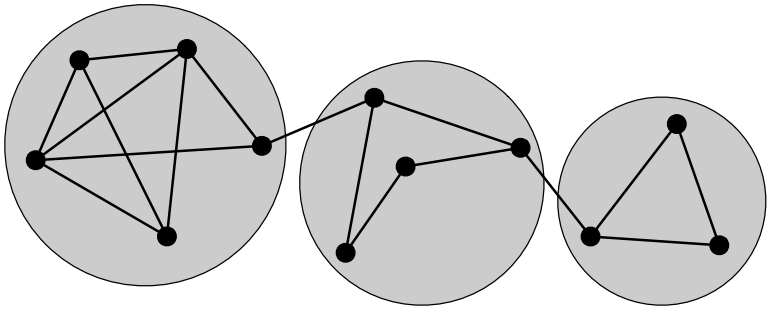


Figure 3.3: A network partitioned into three structural communities.

The natural partitioning of the network of Figure 3.3 is clearly evident to the eye. As Newman notes "the human eye is an analytic tool of remarkable power, and eyeballing pictures of networks is an excellent way to gain an understanding of their structure." [16, p. 169]. As networks grow in size, however, visualization becomes less and less useful as the eye cannot easily discern a network's structure. For this reason, a number of different algorithm to detect community structure

in a computational fashion have been proposed in the literature; they include:

1. *leading eigenvector*, a fast algorithm, based on the definition of the modularity function in terms of the eigenspectrum of matrices [143],
2. *walktrap*, a technique based on random walks [144],
3. *edge betweenness*, the earliest community detection technique, based on vertex betweenness centrality [142]
4. *spinglass*, a technique based on a spin-glass model and simulated annealing [145].

In general, the aim of clustering techniques is to maximize the degree of association between inter-related nodes to thus uncover clusters consisting of nodes with similar features. In previous published work, I compared the performance of these four algorithms to detect structural communities in the CENS coauthorship network [78], finding that the leading eigenvector algorithm is both fast and accurate for the scope of my work. The leading eigenvector method computes the repartition of the network in structural communities based on the vertices' eigenvector centrality [146]. In other words, the algorithm takes a weighted network as an input (e.g., a coauthorship network), and partitions the given network into subgroups (structural communities) based on the topology of the network (e.g., the number of coauthorship connections among authors). This method has been successfully applied to social networks to uncover, for example, the relationship between nationality and collaboration [147], latent communities in large organizations [103], political and organizational structures [148], and to identify communities in networks of collaborating musicians [149, 150].

For the purpose of this dissertation, I tested the efficiency of the four aforementioned algorithms also on the communication and acquaintanceship network.

While the leading eigenvector algorithm performed very well in partitioning the communication network, its application on the acquaintanceship network did not produce similarly accurate results. Only the spinglass algorithm provided valid and accurate results, despite being considerably slower. This finding is very much in line with recent in-depth comparative evaluations of community detection mechanisms that present the superiority of algorithms based on spin-glass models and simulated annealing compared to the leading eigenvector algorithm for both real and artificial networks of medium size and high mixing coefficient [151]. For these reasons, the detection of community structures in all the networks studied in this dissertation was performed using the spinglass algorithm [145].

3.1.3 Homophily and assortative mixing

The network concepts and methods presented thus far enable the study of the topology and structure of networks. However, it is important to remember that networks are not homogeneous entities. Not all nodes of a network are the same. In a social network, for example, nodes may represent individuals of different gender, nationality, race, income, etc.[1] The study of node characteristics can provide insights into the level of *homophily* in a social network, i.e., the tendency of individuals to create ties with similar others [152, for a review]. The homophily principle describes how homogeneous a network is in terms of specific sociodemographic, behavioral, or interpersonal characteristics. For example, a high level of homophily in a friendship network indicates that individuals with

[1]Similarly, not all edges of a network are the same. Besides weight, edges might represent different levels, notions, and shades of interaction. For example, in a social network, acquaintanceship might mean different things to different people. This last point is covered in much detail in Chapter 4.

certain characteristics—such as race, ethnicity, political beliefs, and educational background—tend to make friends with individuals with similar characteristics. Many studies of homophily are grounded in sociology and investigate patterns of homophily as well as their driving forces and their implications.

An established method to measure mathematically the level of homophily in a network is by computing its *assortative mixing*, or *assortativity*, i.e., the extent of mixing between similar nodes in a network [153]. In a social network, assortativity can be defined as the tendency for individuals to establish connections preferentially to other individuals with similar characteristics. While many different components of similarity can be investigated, the vast majority of large-scale studies of networks look at the mixing of node degree, i.e., how nodes with similar degree preferentially attach to one another. In a coauthorship network, for example, degree assortativity indicates the tendency for individuals to write papers with others with a similar number of collaborators. In other words, a high degree assortativity means that very productive authors collaborate with other very productive authors, while low-degree authors (i.e., authors that do not collaborate very much) collaborate with other low-degree authors. In this dissertation degree assortativity is measured computing the Pearson correlation coefficient of the node degrees found at the ends of every edge, using the following formula [154]:

$$r = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}, \tag{3.2}$$

where r is the degree assortativity coefficient, M is the total number of edges, j_i, k_i are the degrees of the nodes at the ends of the ith edge, with $i = 1 \ldots M$. This formula returns a coefficient, r in the range $-1 \leq r \leq 1$, where $r = 1$

indicates perfect assortativity, $r = 0$ indicates no assortativity, and $r = -1$ indicates perfect disassortativity.

Mixing patterns, however, can also be calculated based on discrete node-specific characteristics. In other words, one can study whether individuals with certain characteristics associate preferentially with similar others. In studies of scholarly and scientific collaboration networks, examples of characteristics that have been investigated in this manner include: research interests [74], academic domain [75], geographical location [76], age group [155], and country of origin [78]. These studies offer insights into the mechanisms by which disciplinary, institutional, and spatial arrangements shape, and are shaped by, collaboration patterns. For nominal parameters, such as race, affiliation, gender, etc., the discrete assortativity coefficient, r, can be computed using the following formula [153]:

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} \tag{3.3}$$

where e_{ij} is the fraction of edges in a network that connect a node of type i to one of type j, a_i is the fraction of edges that have a node of type i on the head of the edge, and b_i is the fraction of edges that have a node of type i on the tail of the edge. As for degree assortativity, $r = 1$ when there is perfect assortative mixing, $r = 0$ when there is no assortative mixing, and $r = -1$ when there is perfect disassortative mixing. In other words, the discrete assortativity coefficient, r, indicates the level of homophily of the network for a certain parameter. For example, if in a coauthorship network, r for academic affiliation is 1.0, this means that individuals in the network only write papers with other individuals with same institutional affiliation. In this kind of network, there are no multi-institutional collaborations. On the other side of the spectrum, we can imagine a completely disassortative network ($r = -1$) in which every single collaboration (i.e., paper)

in the network is authored by individuals that belong to different institutions.

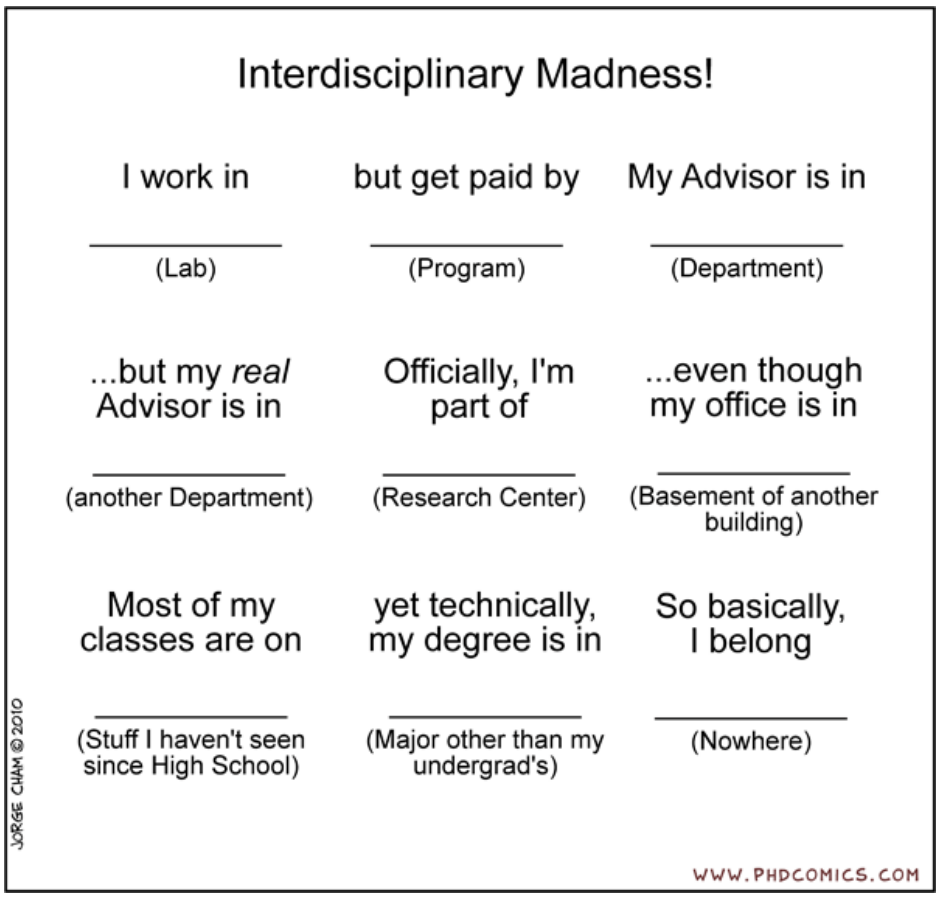
3.2 Data and instruments

In this dissertation, I employ a number of data sources to construct networks of coauthorship, communication and acquaintanceship. I gather information about scholarly coauthorship by inspection of bibliographic records available at the CENS institutional repository; about electronic communication by analysis of official CENS mailing list archives; and about acquaintanceship patterns by administering an online social network survey instrument. The remainder of this section discusses the collection techniques and instruments by which these data are collected.

3.2.1 The CENS bibliographic record

In Chapter 2, I note that the set of scientific activities and practices of CENS generate various scholarly artifacts, such as journal articles, conference papers, technical reports, and posters. Building a collaboration network around coauthorship activity requires understanding what the CENS bibliographic record consists of. The question at hand is: what scholarly publications are part of the CENS bibliographic record? Clearly, answering this question has immediate repercussions on the size, configuration and composition of the population under study. As outlined in Chapter 1, one major obstacle that is often encountered when studying modern science collaboratories is their inherent boundary flexibility. CENS comprises researchers from multiple institutions and disciplines. Researchers affiliated with CENS may also be affiliated with other laboratories and perform interdisciplinary work on other projects and under different affili-

ations. The comic of Figure 3.4 illustrates the boundary problem better than any description. To add complexity to an already complex scenario, many CENS collaborations include researchers that are not affiliated with CENS at all. In this context, how do you discriminate between a CENS publication and a non-CENS one?



Comic published 3/22/2010 on "Piled Higher and Deeper" by Jorge Cham. www.phdcomics.com. Reprinted with permission.

Figure 3.4: Comic: Interdisciplinary Madness

Previous environment-specific studies of coauthorship delineate the population under study by relying on data contained in an institutional repository [55] or domain-specific bibliographic databases [53] to mine patterns of coauthorship

that take place within a given institution or academic domain, respectively. To establish the population that constitutes CENS' coauthorship network, I employ a similar mechanism. As alluded to in the previous chapter, I assemble the scholarly items included in the CENS Annual Reports, the official documents published by CENS every year to report its progress to the National Science Foundation and other funding agencies.

At the time of writing, seven CENS annual reports are available (2003-2009), describing the progress of the Center since its inception to date (`http://research.cens.ucla.edu/about/annual_reports/`). The annual reports, published at the end of every fiscal year, describe the Center's *state of affairs* during its preceding 12 months: its research goals and objectives, policies, organization charts, budget summaries, faculty biographies, list of members, and list of scholarly contributions. Annual reports are compiled by CENS administrative staff. Every year, staff members ask project leaders to provide descriptions of their project activities and a list of related publications. In turn, project leaders may ask members of the projects they lead to provide description of their individual activities and a list of their personal scholarly contributions. As such, the list of publications is constructed incrementally, by aggregating individual scholarly articles that researchers deem to be contributions to CENS research. Clearly, this procedure is not without error. Some researchers might overlook their personal bibliographic record and fail to submit important CENS publications. Others might submit publications that are not entirely related to CENS research. Also, over the years researchers might be asked to provide slightly different segments of their work, and thus the annual reports might not be entirely consistent over time. Despite these minor inconsistencies, by a mechanism of distributed self-reporting, the Annual Report is the most comprehensive bibliographic record of CENS activities. Every year, bibliographic metadata (authors, title, year,

publication venue, etc.) for every publication in the annual report are also uploaded at a dedicated site of the California Digital Library eScholarship repository (`http://repositories.cdlib.org/cens`).

It is important to note that the publication list contained in the Annual Reports only lists books, book chapters, journal articles and papers published in conference proceedings. It excludes other scholarly materials such as posters and technical reports. In this dissertation, I choose not to include posters and technical reports in order to avoid repetition of material and because they are not part of an official reporting mechanism. Although not documented in the annual reports, posters and technical reports are listed and stored in dedicated sections of the eScholarship repository. At the time of writing, the CENS eScholarship repository contains 369 posters and 68 technical reports. Posters and technical reports at CENS are important vehicles of scientific dissemination. Posters, especially, are an efficient and compact means to present the latest achievements of a working group during CENS events. The contents, titles, and author lists of posters often overlap with those of related journal and conference papers. Yet, author lists of posters tend to be more inclusive than those of scholarly articles. This has a disadvantage: posters do not always reflect the true arrangement of a given collaboration; oftentimes, all members of a project or team are indicated as authors of a poster. However, this also brings about an advantage: some posters might include researchers such as software developers and technical staff who do not appear as authors in published articles, but whose work is crucial to collaboration. This issue is discussed further among the limitations and future work of this dissertation, in Chapter 8.

Using bibliographic information about books, book chapters, journal articles and conference papers available in the Annual Reports, I assemble a publica-

tion database, consisting of 608 papers published over a period of ten years (2000–2009). Table 3.1 summarizes some important statistics relative to the collected bibliographic data: paper distribution by publication type, publication year, number of authors, and publication venue.

A quick analysis of Table 3.1 reveals some important properties relative to the nature of publication practices at CENS. The distribution of papers by publication type, for example, shows that about two-thirds of publications are papers in conference proceedings, while journal articles take up the other third of the volume of publications. This is not a surprising result, given the fact that many technical disciplines rely on conferences rather than journals for scientific communication and knowledge dissemination. In fact, it has been noted that in the field of computer science, which is a core discipline at CENS, there is a strong publication culture that favors conference papers over journal articles [156]. The year of publication of the articles in the bibliographic database shows that the number of publications by CENS authors increased sharply a couple of years after the inception of the center in 2002 and then stabilized at a rate of about 80 publications per year. The distribution of items per number of authors reveals that about half of all publications are authored by two or three individuals. This is perfectly in agreement with recent findings that report frequent coauthoring team sizes of two to three members in the computer sciences [68]. Publications by four and five authors are not rare at CENS, however, making up together about a quarter of all publications. Author lists rarely exceed six authors and this is a confirmation of the hypothesis anticipated in § 2.4.1 — that hyperauthorship is not common in CENS research. It is also worth noting that the publication database includes 59 sole authored documents (roughly 10% of all publications). These publications do not directly contribute to the construction of the coauthorship network: they are not manifestation of collaboration and thus cannot be used to

Paper type	n = **608**
Conference proceedings	400
Journal article	189
Book chapter	18
Book	1
Year of publication	n = **608**
2000	6
2001	20
2002	41
2003	94
2004	116
2005	105
2006	71
2007	64
2008	68
2009	23
Number of authors	n = **608**
1	59
2	155
3	158
4	94
5	59
6	32
7	19
8	13
9	5
10+ (where 14 is the maximum number of authors found)	14
Venue of publication	n = **608**
International Conference on Robotics and Automation (ICRA)	32
International Conference on Information Processing in Sensor Networks (IPSN)	15
Conference on Embedded Networked Sensor Systems (Sensys)	11
International Conference on Intelligent RObots and Systems (IROS)	11
American Geophysical Union Meetings (AGU)	11
International Conference on Micro Electro Mechanical Systems (MEMS)	11
IEEE Transactions on Mobile Computing	6
European Conference on Computer Vision (EECV)	5
Applied Physics Letters	4
Earthquake Spectra	4

Bibliographic data statistics: paper distribution by publication type, publication year, number of authors, and publication venue.

Table 3.1: Basic statistics for the collected bibliographic data.

generate any edges between the nodes. Yet, they constitute important scholarly output and are therefore used for the calculation of specific network metrics, e.g., authors' centrality can be assessed summing coauthored and sole authored items. Finally, the distribution by venue shows that the majority of publications appear in technical conferences that specialize in sensor network and wireless sensing research (e.g., ICRA, IPSN, Sensys, MEMS) with a smaller proportion appearing in journals that cover CENS' application domains (e.g., Applied Physics Letters and Earthquake Spectra).

3.2.2 The CENS mailing list archive

As discussed throughout this dissertation, much of communication at CENS takes place, as in most collaboratory research, via online electronic platforms. In this dissertation, I specifically analyze communication activities traceable from a set of electronic mailing lists maintained by CENS. I construct a communication network measuring the extent of online communication among researchers extracted from mailing list logs.

An electronic mailing list consists of a *reflector*, an email address that, when used as email recipient, distributes a copy of the email to all subscribers of the list. Mailing lists can be private or public. Administrators may decide to make available the archive of past discussions and the list of subscribers to anyone or to subscribers only. Administrators also decide how users subscribe to the list (by email, by web interface, by invitation, etc.) and whether to allow message moderation. To initiate a discussion, subscribers send an email to the list (the reflector).

When replying to a topic of discussion, a *thread* is formed, i.e., emails are grouped together according to email subjects to discriminate among different

topics of discussions. I assume for the purpose of this dissertation that individuals involved in mailing list discussions do not modify the email subject of the discussion. This assumption has its flaws. As for most email communication, people might, in fact, vary the email subject when replying to a mailing list discussion that initially started under a different heading — the email with the new subject is a *false positive*, which results from detecting a difference that does not in fact exist (Type I error). In a similar vein, the discussion taking place in a thread might diverge into new topics, but the email subject might stay unvaried. In mailing list lingo, this is called *hijacking a thread* and is a case of *false negative*, i.e., failing to observe a difference that is, in fact, true (Type II error). Based on this illustration, I consider communication activity the interaction among individuals on an electronic mailing list around a certain topic.

CENS currently operates 100 different mailing lists (`http://www.cens.ucla.edu/mailman/listinfo`). Many of them (87 out of 100) are unmoderated public lists to which anyone can subscribe; the remaining 13 are private lists and are thus excluded from this study. Subscription to CENS mailing lists is subject to the approval of a technical administrator. Upon successful subscription, subscribers can post to the mailing lists and browse the archive of past discussions. CENS mailing lists vary in volume, subscribers and function. They are used for a number of different purposes: to discuss past and ongoing sensor deployments, to advertise events, or merely to organize social activities. Examples of mailing lists currently in use at CENS are `us-general`, a general discussion mailing list on topics concerning urban sensing projects; `Integrity`, a list dedicated to discussions around data integrity issues, and `3551`, a low-volume, high-importance list for occupants of Boelter Hall office 3551 - the CENS headquarters at UCLA.

CENS mailing lists are handled using Mailman, a Unix-based application dis-

tributed under the GNU General Public License (`http://www.gnu.org/software/mailman`). It is important to note that versions of the Mailman program prior to 2.1.5 (released in May 2004) did not log detailed information about threads. The CENS mailing list system was updated to this version of Mailman in May 2005. As a consequence, mailing list data prior to April 2005 could not be decoded in a threaded format because not enough identification information and description are provided in the system logs of CENS. Thus, the dataset analyzed in this dissertation includes all emails sent on the 87 public mailing lists of CENS from May 2005 to April 2009. A total of 1454 threads were identified in this dataset. Following the definition of communication activity (presented above) and drawing from a number of previous investigations of communication activity [93, 95, 96, 103, for a review], I exclude from this study threads that involve less than three individuals (i.e., threads with one or two discussants only). This is not only because unreplied and one-to-one emails are uninteresting from a network perspective but also because an exploratory analysis of these low-involvement messages revealed that most of them are announcements of events and not discussions *per se*. Some basic statistics relative to the collected mailing list data are presented in Table 3.2: distribution of threads per year, distribution of discussants per thread, and mailing lists with the highest number of emails.

The values at the top of Table 3.2 show the distribution of threads by year. As explained above, complete data are only available for years 2006 through 2008, and as the thread distribution shows, the average amount of threads in these years is roughly 400. The second set of values shows the distribution of discussants per thread. About one third of threads involve three individuals. About one fifth involve four discussants. The amount of threads decreases as the number of discussants increases. It is interesting to note that conversations among more than fifteen individuals are not many but make however about 5%

Year	n = **1454**
2005	255
2006	468
2007	339
2008	375
2009	17

Discussants per thread	n = **1454**
3	493
4	299
5	189
6	144
7	80
8	53
9	44
10	36
11	25
12	20
15+ (maximum is 26)	71

Mailing list name	n = **30 671**
cens-seismic	2502
dgroup	2088
peir	1888
us-internal (formerly: urbansensing-internal)	1704
stargate-users	1388
cens-rec	734
sysadmin	710
us-general	708
tenet	629
emstar-users	553
jr-systems	543
metwi	376
ess2	375
urbansensing	359
kaiserlab	345
emissary	291
integrity	280
education	258
nims	246
peir-info	221

Mailing list data statistics: distribution of threads per year, distribution of discussants per thread, and mailing lists with the highest number of emails.

Table 3.2: Basic statistics for the collected mailing list logs.

of the dataset. The set of values at the bottom of Table 3.2 displays the most-emailed mailing lists. A number of CENS application areas make extensive use of mailing lists, such as `cens-seismic` of the seismology group, and `us-general` of the urban sensing group. Other high volume lists are project-specific, e.g., `peir`, for the Personal Environmental Impact Report project, or dedicated to users of specific sensor network technologies, e.g., `emstar-users`.

3.2.3 The social network survey instrument

The third interaction analyzed in this dissertation is acquaintanceship. As discussed in Chapter 2, acquaintanceship involves a form of personal knowledge that is stronger than mere recognition, but is less prominent than a familiar and intimate relationship. How can one collect acquaintanceship data? While the two interactions discussed above (scholarly coauthorship and electronic communication) measure tangible indicators of collaboration directly extracted from artifacts (bibliographic records and mailing list logs), in order to measure acquaintanceship I rely upon methods of survey research. There are many ways to collect acquaintanceship data — see Chapter 2 for an extensive review — but most of them essentially involve asking respondents to name or indicate who they know via surveys, questionnaires, and interviews.

The measurement of acquaintanceship is inevitably an error-prone procedure. Mark Newman notes that the use of qualitative methods for the study of acquaintanceship is subject to two major issues [157]. First, the data collection process is elaborate and thus the quantity of data returned is necessarily limited to small samples — "most data sets contain no more than a few tens or hundreds of actors" [157, p. 338]. This, in turn, affects statistical accuracy for large-scale investigations [158]. Second, Newman notes that the subjective nature

of the respondents' perception of acquaintance introduce uncontrolled statistical errors — "what one respondent considers to be a friendship or acquaintance, for example, may be completely different from what another respondent does" [157, p. 338]. For these reasons, Newman and many other researchers in similar approaches have operationalized acquaintance through more tangible indicators, better suited for large-scale quantitative analyses. For example, in *Who is the best connected scientist?* [157], Newman constructs a coauthorship network and employs it as an acquaintanceship network, assuming that "it is probably fair to say that most people who have written a paper together are genuinely acquainted with one another" [157, p. 339]. Using coauthorship networks as proxies to social networks can be a fair assumption for the study of very large collaborative environments, for which collecting data via interviews and surveys would be an arduous task. At CENS, however, the population under study is relatively small (a few hundred individuals) compared to large-scale social network analyses that suffer from these sampling errors. For this reason, I develop and employ a survey instrument that asks all respondents in the population to directly indicate their acquaintances[2]. The rosters extracted from the bibliographic record and the mailing list archives, i.e., the list of coauthors on scholarly articles and discussants on mailing lists, are conjoined to form the survey roster.

At the time that this research was conducted, the survey roster consisted of a total of 388 individuals, all of whom were invited to take part in the survey via a recruitment letter, sent via email on August 17, 2009 (included in the Appendix, A.1). The email includes a hyperlink pointing to a customized online questionnaire. Delivery to some recipients failed due to a number of technical reasons, such as non existing address or invalid domain. This was due to the

[2]The survey instrument employed in this research was certified exempt by the Institutional Review Board (IRB), Office for Protection of Research Subjects, University of California, Los Angeles on December 8, 2008 (Protocol no. `08-471`).

fact that some of the participants have changed affiliation and/or email address since the time of compilation of the email database. I dealt with each invalid email address individually, trying to gather the latest contact information about unreachable individuals. One month into the data collection, on September 14, I sent a reminder (again, via electronic mail) to participants that had not yet responded to the survey. It is worth noting that the Institutional Review Board (IRB), who certified this research as exempt, only allowed me send a total of one recruitment letter and one reminder to the survey population.

The social network survey instrument employed in this research, that I developed from scratch using HTML and PHP, is hosted on a dedicated server at CENS (more details about this in the Appendix, § A.5). The survey is structured as follows. The questionnaire begins with an informed consent form in which respondents are prompted with basic information about the study: the names and contact details of the principal investigators, the context of the research and its goals, and the details of participation. The text used in the informed consent form is included in the Appendix, § A.2. The questionnaire follows, divided into two parts. In the first part, respondents are asked to select their acquaintances from a list. In the second part, they are then asked to indicate the nature and length of the relationship with their acquaintances. Screen-shots of the first and second parts of the questionnaire are included in the Appendix, Figures A.1 and A.2.

The first part of the questionnaire asks the question: "Who do you know?". Respondents are asked to select individuals that they are acquainted with from the roster. In particular, in order to clarify the notion of acquaintanceship, respondents are reminded that:

> For the context of this survey, an acquaintance is someone that you

know in person and that you would say "hi" to if you bumped into them in the hallway.

This description is aimed at informing respondents that they are asked to indicate real, offline acquaintances with whom a personal form of knowledge exists. This question is followed by a list of all of the individuals in the survey roster. Individuals are identified by name, last name and a thumbnail picture, if available. Pictures are obtained from a public roster held by the CENS administrative office. In order to aid recognition, individuals are grouped together by department and affiliation (Figure A.1, in the Appendix). In the first page of the questionnaire, respondents select their acquaintances and after submitting the data are taken to the second part of the survey.

The second part of the questionnaire asks the question: "How do you know them?", i.e., it asks respondents to describe the nature and length of the relationship with their acquaintances. In this page, respondents are prompted with the list of people that they selected as acquaintances in the previous page. For each acquaintance, the following two questions are asked (see Figure A.2, in the Appendix):

1. When did you first meet?
   (a) 2001 or earlier
   (b) 2002
   (c) 2003
   (d) 2004
   (e) 2005
   (f) 2006
   (g) 2007
   (h) 2008

(i) This year

2. How often are you in touch?

   (a) At least once/week

   (b) At least once/month

   (c) Occasionally

   (d) Rarely or never

Answers to these questions are optional and respondents are asked to leave answers blank if they do not know or cannot remember these details. Submitted data are saved to a database hosted on the CENS server. A total of 191 responses were collected over a period of about two months, from August 17 to October 7, 2009. The distribution of survey responses over the data collection period is presented in Figure 3.5 as a time-series chart. The two spikes in the figure clearly correspond to the day the first email invitation and the reminder email were sent — August 17 and September 14, respectively.

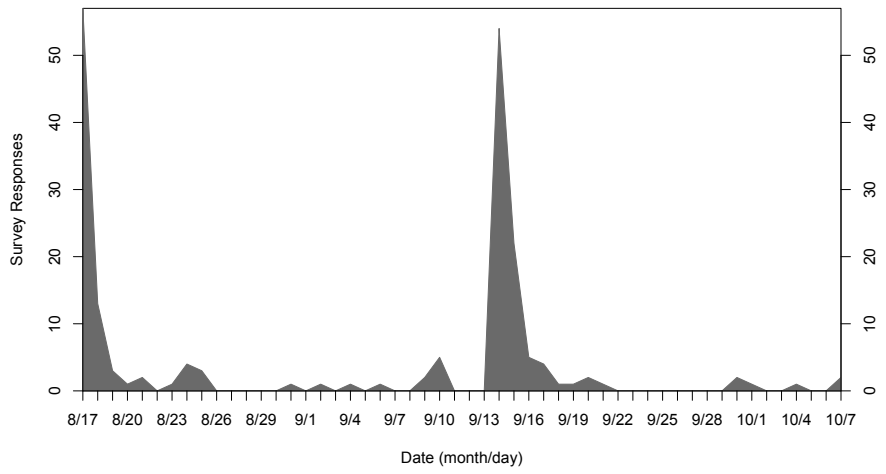


Figure 3.5: Temporal distribution of survey responses.

Some basic statistics relative to the data collected via the social network survey are summarized in Table 3.3. Nearly half of respondents invited to fill in the survey (49%) participated in the study. The rest were either not reachable (4%), or did not respond to the survey by the end of data collection (47%). About one third of respondents (39%) only completed the first part of the survey. The vast majority of respondents indicated a number of acquaintances ranging between 5 and 40.

Survey response	$n = 388$
Respondents	191 (49 %)
Non-respondents	182 (47 %)
Unreachable	15 (4 %)
Portion of survey completed	$n = 191$
Full survey	116 (61 %)
Only part one	75 (39 %)
Number of acquaintances	$n = 191$
1-5	12
5-10	21
10-20	47
20-30	38
30-40	19
40-50	12
50-60	16
60-70	9
70-100	10
100+ (maximum is 197)	7

Table 3.3: Basic statistics for the collected social survey data.

3.3 Summary

This chapter is divided in two parts. The first part (§ 3.1) includes an overview of the research methods used in this dissertation. As network theory is the

overarching analytic framework employed here, this part provides an explanation of concepts and topics relative to the structure and function of networks. The second part (§ 3.2) includes an overview of the data employed in this research, and the methods and instruments of data collection. In the next chapter, I discuss how these data — a bibliographic record, a mailing list archive, and a record of personal knowledge patterns — are converted to networks of coauthorship, communication and acquaintanceship.

CHAPTER 4

Results: Network topology and socio-academic configuration

In the previous chapter, the three data sources employed in this study are introduced: a bibliographic record of scholarly publications, extracted from CENS' annual reports; an archive of threaded discussions derived from mailing list archives; and a record of acquaintanceship obtained by collecting the responses to a survey questionnaire. In this chapter, I convert these datasets to a graph-based format. I illustrate the construction of these networks and study their basic topology.

4.1 Coauthorship network

The bibliographic database of CENS publications, consisting of 608 metadata records, was gathered and managed in BibTEX format (`http://www.bibtex.org/`). In order to illustrate the procedure by which I construct the coauthorship network, consider the BibTEX entry below, relative to a recent CENS publication. The entry is of type `article` (journal article); conference papers are indicated in BibTEX as `inproceedings`. The entry contains a unique identifier and fields for authors, title, year, etc. Authors are separated by the keyword `and`.

```
@article{Pepe_Rodriguez:2010,
  Author = {Alberto Pepe and Marko A. Rodriguez},
  Title = {Collaboration in sensor network research:
           an in-depth longitudinal analysis of assortative
           mixing patterns},
  Abstract = {Many investigations of scientific collaboration
              are based on statistical analyses of large networks
              constructed from bibliographic repositories. These
              investigations [...] },
  Year = {2010}},
  Journal = {Scientometrics}
}
```

In order to construct a coauthorship network, the crucial information to be extracted from each publication is the author list. Even when dealing with relatively small datasets (as in this case, $n = 391$), constructing a reliable coauthorship network, requires disambiguating author names. Authors might have identical names and last names, and their names might be spelled differently or incorrectly in the bibliographic metadata. In order to overcome this problem, I allocate unique identifiers (a string composed of the name initial followed by a dot and the last name) to every author in the bibliographic database. Whenever identifiers are already taken, the middle name initial is introduced, or the full name. As BibTEX handles ASCII only, names containing non-ASCII characters, such as cedillas (ç) and umlauts (ö) are converted manually into arbitrary identifiers. The author database was de-duplicated, checked and curated manually. For example, the BibTEXentry above is coded as follows:

```
  Author_ID = {a.pepe},
  Author_ID = {m.rodriguez},
```

These identifiers are used to construct the edges of the coauthorship network — a network in which vertices represent authors and edges represent the extent

of coauthorship activity. The network is weighted and the edge weights are established by partitioning a set value for every publication. In order to determine the weights between nodes, i.e., the strength of collaboration among coauthors, I use a weighting mechanism proposed by Newman [159] by which the weight of the edge between nodes i and j is:

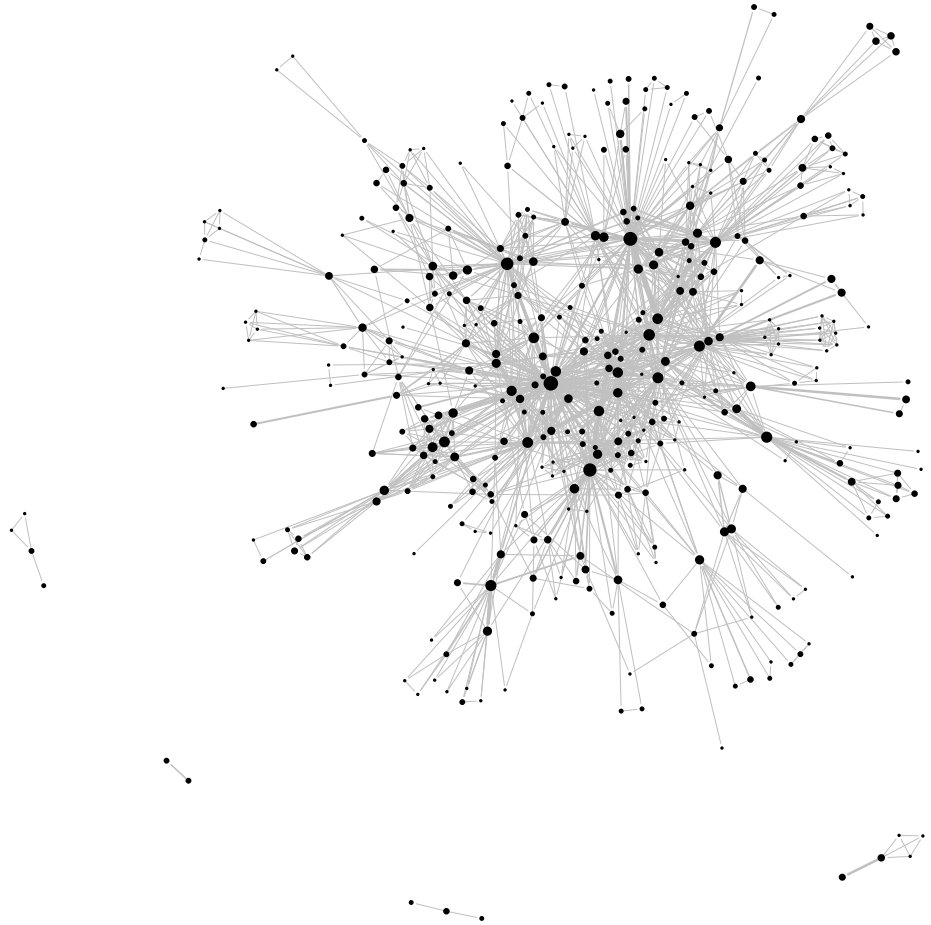
$$w_{ij} = \sum_{k} \frac{\delta_i^k \delta_j^k}{n_k - 1}, \tag{4.1}$$

where δ_i^k is 1 if author i collaborated on paper k (and zero otherwise) and n_k is the number of coauthors of paper k. For the example above, the edge between authors Pepe and Rodriguez have $w_{ij} = 1$, or in `ncol` format [16]:

```
a.pepe m.rodriguez 1.0
```

An article written by three authors (e.g., Pepe, Rodriguez, and Bollen) would result in three edges (Pepe-Rodriguez, Pepe-Bollen, and Rodriguez-Bollen), each one with $w_{ij} = 0.5$. And so on. As such, this weighting mechanism confers more weight to small and frequent collaborations, based on the assumptions that: i) publications authored by a small number of individuals involve stronger interpersonal collaboration than multi-authored publications, and ii) authors that have authored multiple papers together know each other better on average and thus collaborate more strongly than occasional coauthors [159]. Using this weighting mechanism, a network of coauthorship is constructed. It is depicted in Figure 4.1.

Some statistics relative to the topology of the constructed coauthorship network (depicted in Figure 4.1) are presented in Table 4.1. A descriptive summary



The coauthorship network ($n = 391$, $m = 1\,747$) diagrammed according to the Fruchterman-Reingold network layout algorithm [160]. Line width is proportional to edge weight, where more intense collaborations have wider and more marked lines; the diameter of the nodes is proportional to the weighted centrality score on a logarithmic scale, or strength [161], where more central nodes have larger diameters.

Figure 4.1: Weighted coauthorship network.

of the code and tools developed to perform the analyses presented in dissertation is provided in the Appendix, § A.5.

Topological property	Value
Number of nodes (individuals), n	391
Number of edges (collaborations), m	1 747
Connected components	5 (377, 5, 4, 3, 2)
Diameter	6
Average path length, ℓ	2.952
Maximal cliques	291
Largest clique	14
Clustering coefficient, C	0.301

Table 4.1: Topological properties of the coauthorship network.

An analysis of the statistics of Table 4.1 provides insights into the topology of the coauthorship network. The bibliographic dataset includes a total of 1747 scholarly collaborations among 391 authors, which in network terms are expressed as edges and nodes, respectively. The analysis of the network's configuration shows that the network is partitioned into 5 different connected components. This finding is also evident from Figure 4.1, which depicts five separate clusters. Most of the network's nodes, however, are grouped within the largest component, i.e., the cluster with the maximum number of nodes, which in this case includes 96% of nodes (377 out of 391). This means that the vast majority of the network is connected, i.e., one node can be reached from any other one by following simple paths in the network. The network diameter indicates that these simple paths (to connect any two nodes in the largest connected component) are at most 6 steps long and, on average, 2.952. This last value — the average path length, ℓ, indicates that on average, just under three steps are necessary to connect any two individuals in the coauthorship network. This means that between two and three steps is enough to reach any other coauthor in the network. Thus, if we take two random individuals in the network, they are very likely to have collaborated with a common coauthor (path length of 2). The coauthorship network features 291 maximal cliques, with the largest consisting of 14 nodes. The clustering

coefficient, which is 0.301 in this case, indicates that the network is not highly clustered, i.e., the density of cliques in this network is not very high.

4.2 Communication network

The archive of electronic communication derived from 100 CENS-managed mailing lists was obtained in mbox (`http://www.qmail.org/man/man5/mbox.html`), a file format for holding collections of electronic mail messages. As an example, consider the following discussion thread called "beta.sensorbase.org down" that was initiated by a CENS staff member on July 1, 2008 on mailing list `3551`, upon realizing that the server holding Sensorbase, CENS' sensor data sharing platform, was inaccessible (email and name of sender, message ID and body of the email intentionally removed):

```
From: [sender1]@ucla.edu ([sender name])
Date: Tue, 1 Jul 2008 12:43:00 -0700 (PDT)
Subject: [3551] beta.sensorbase.org down
Message-ID: [some message ID]

[body of the email]
```

The snippet above displays an email sent by `[sender1]` on the date shown (1 Jul 2008), with subject "beta.sensorbase.org down" (anonymized). Analyzing the mbox archive around the given subject (`beta.sensorbase.org down`), one finds that this particular email was followed by five responses within the next 24 hours by two other researchers of UCLA: `[sender2]@ucla.edu` and `[sender3]@ucla.edu`. In particular, the communication pattern relative to this specific email thread can be summarized as follows:

```
sender1 sender 2 sender1 sender2 sender3 sender1
```

This email list represents the sequence and extent of electronic communication between `[sender1]`, `[sender2]`, and `[sender3]`. These email addresses are associated to the respective individuals automatically, using a lookup table. In order to enable comparisons among different networks consisting of the same individuals, I use the same unique identifiers already utilized in the coauthorship and acquaintanceship networks. For the purpose of the current presentation, let us suppose that `sender1` is `a.pepe`, `sender2` is `m.rodriguez`, and `sender3` is `j.bollen`. The mailing list thread presented above can be then expressed as:

```
a.pepe m.rodriguez a.pepe m.rodriguez j.bollen a.pepe
```

Using this list of unique identifiers, it is possible to construct a network in which vertices represent discussants and edges represent the extent of the communication activity. In the case presented above, for example, `a.pepe`, `m.rodriguez`, and `j.bollen` should be connected by edges. There are many different schemes by which edge weights can be assigned. For example, one might decide to confer more weight to individuals that have communicated more in a single thread (e.g., `a.pepe` in the above example). However, emails can be short or long and since I am not analyzing email content in my study, it is not fair to assume that more emails sent constitute necessarily more involvement in a discussion. Thus, I distribute weights equally to all individuals involved in an email thread. The thread presented above is reduced to the following list (duplicates removed):

```
a.pepe m.rodriguez j.bollen
```

This list can now be easily converted to a network format, such as `ncol`. For the coauthorship network, presented in the previous chapter, I assign edge weights using a weighting mechanism by Newman that confers more weight to small and frequent collaborations [159]. While this weighting scheme is appropriate for scholarly coauthorship — it is fair to assume that smaller collaborations involve stronger interpersonal collaboration — it does not fit the nature of collaboration found in communication network. For this reason, I modify Newman's formula (1) by calculating the weight of an edge between nodes i and j as:

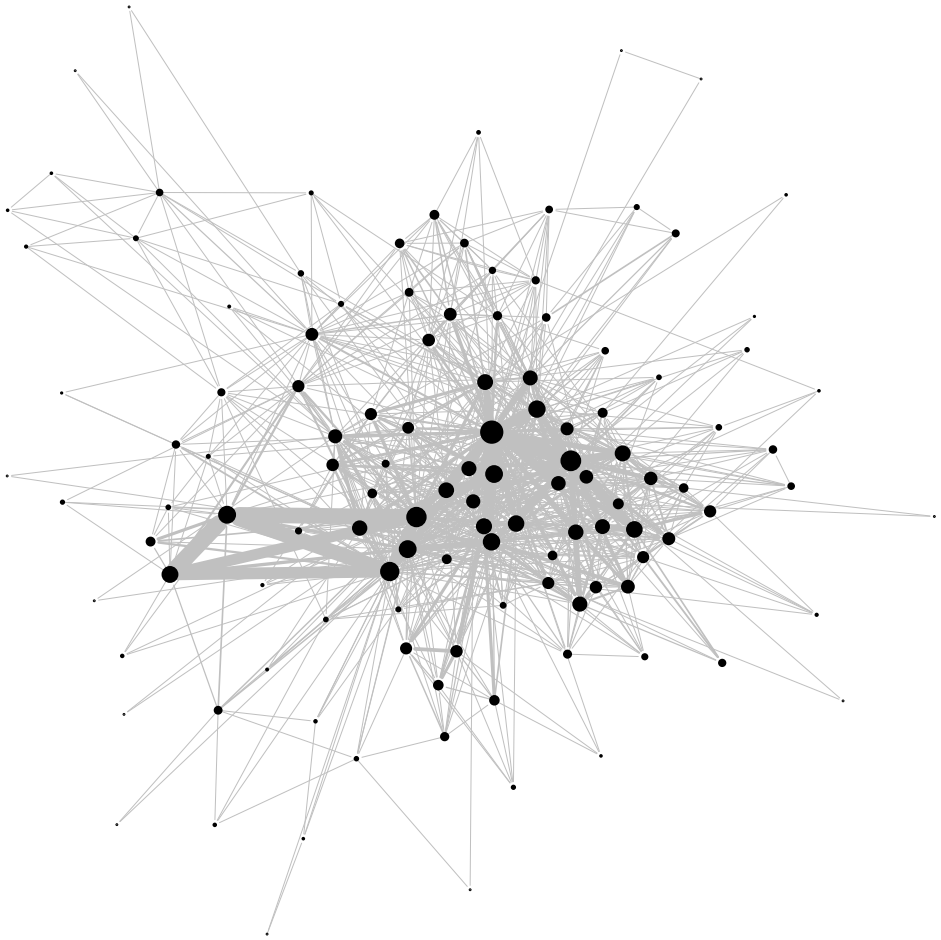
$$w_{ij} = \sum_{k} \delta_i^k \delta_j^k, \tag{4.2}$$

where δ_i^k is 1 if author i is a discussant in thread k (and zero otherwise). As such, the scheme confers weight equally to all discussants in a thread and gives more weight to frequent communication between individuals. For the example above, the edges between individuals Pepe, Rodriguez and Bollen would be:

```
a.pepe m.rodriguez 1.0
a.pepe j. bollen 1.0
m.rodriguez j.bollen 1.0
```

Using this weighting mechanism, a communication network representing mailing list activity is constructed. It is depicted in Figure 4.2. Some statistics relative to the topology of the communication network are presented in Table 4.2.

Looking at Table 4.2, it is evident that the communication network is much smaller than the coauthorship network (Table 4.1), consisting of only 119 nodes and 994 edges, all assembled in a unique connected component. The diameter of the network is also smaller (4) as well as the average path length (2.095): only two steps are required to reach any two nodes in the communication network.



The communication network ($n = 119$, $m = 994$) diagrammed according to the Fruchterman-Reingold network layout algorithm. In the Figure, line width is proportional to edge weight, where more intense communication activities are represented by wider and more marked lines; the diameter of the nodes is proportional to the node strength, where more central nodes have larger diameters.

Figure 4.2: Weighted communication network.

This network also features a higher number of maximal cliques and a higher clustering coefficient (0.461) indicating a more dense concentration of tight-knit circles, compared to the coauthorship network, in which collaboration is more

Topological property	Value
Number of nodes (individuals), n	119
Number of edges (discussions), m	994
Connected components	1 (119)
Diameter	4
Average path length, ℓ	2.095
Maximal cliques	368
Largest clique	14
Clustering coefficient, C	0.461

Table 4.2: Topological properties of the communication network.

sparse. This means that CENS individuals tend to write papers with a diverse group of collaborators, but when it comes to electronic communication, most discussions happen with the usual restricted circle of collaborators: communication patterns are more cliquish. This finding is probably linked to the fact that people only subscribe to the mailing lists that most specifically match their interests and research areas. Thus, the organization of mailing lists naturally restrain cross-fertilization of ideas and novel communication patterns among previously disconnected individuals.

4.3 Acquaintanceship network

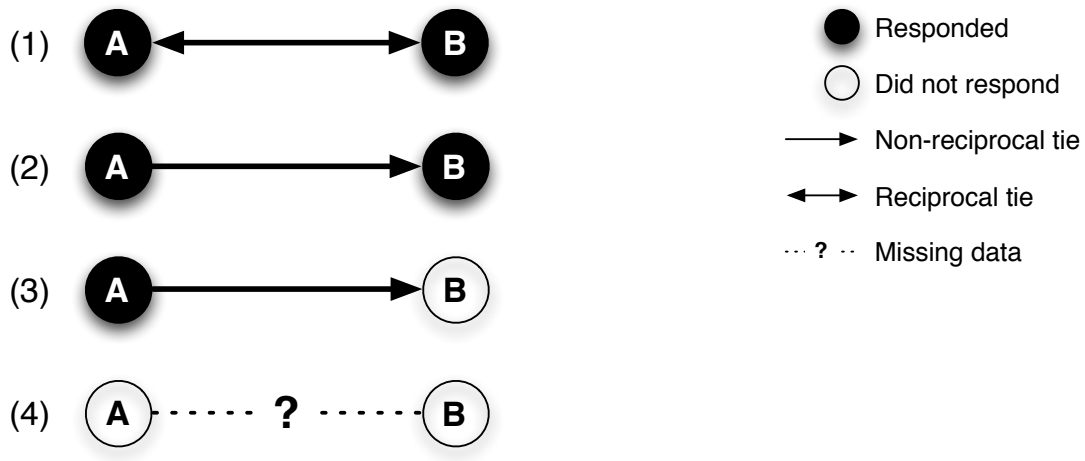
As anticipated in the previous chapter, the social network survey research administered to 388 participants had a response rate of 49%. Although in most social science research such a response rate would be deemed more than sufficient, when using survey data to construct social networks, missing data becomes of crucial importance. This is because in a social network survey, participants are asked to describe their relationship to one another. When information is missing, it cannot always be reconstructed or inferred from the topology of the rest of the network. Missing data has been notoriously noted as a "curse" to social network

research because network analysis is especially sensitive to missing data [162]. In the case of large scale *who-knows-whom* networks, such as the one presented here, missing data might result in large holes in the adjacency matrix, thus distorting the overall network structure. It is thus important to consider various techniques to deal with missing and incomplete data before constructing a social network from the collected data.

In the social network survey, each participant is asked to indicate who they know in the entire population ($N = 388$). Responses obtained are independent from one another and might result in either reciprocal or non-reciprocal ties. The response scenarios that are relevant here are presented in Figure 4.3. Filled (black) nodes represent individuals that took the survey, while blank (white) nodes depict individuals that did not respond. The first case represents the simplest case: both respondents A and B took the survey and indicated one another as acquaintances. In the second case, data is also complete, i.e., both A and B responded, however A indicated B as an acquaintance, but B did not. In the third case, data is incomplete: A took the survey and indicated B as an acquaintance but B did not take the survey. Finally, the fourth case depicts a case of missing data: a tie might or might not exist between A and B but we cannot know about it because neither one of them took the survey.

The typology presented in Figure 4.3 includes both reciprocal ties (represented in a directed network by bidirectional edges) and non-reciprocal ties (unidirectional edges). Using the data collected in the survey, I construct a preliminary directed network, drawing bidirectional edges for case (1), and unidirectional edges for cases (2) and (3). Case (4) ties are not considered. This directed network is presented in Figure 4.4.

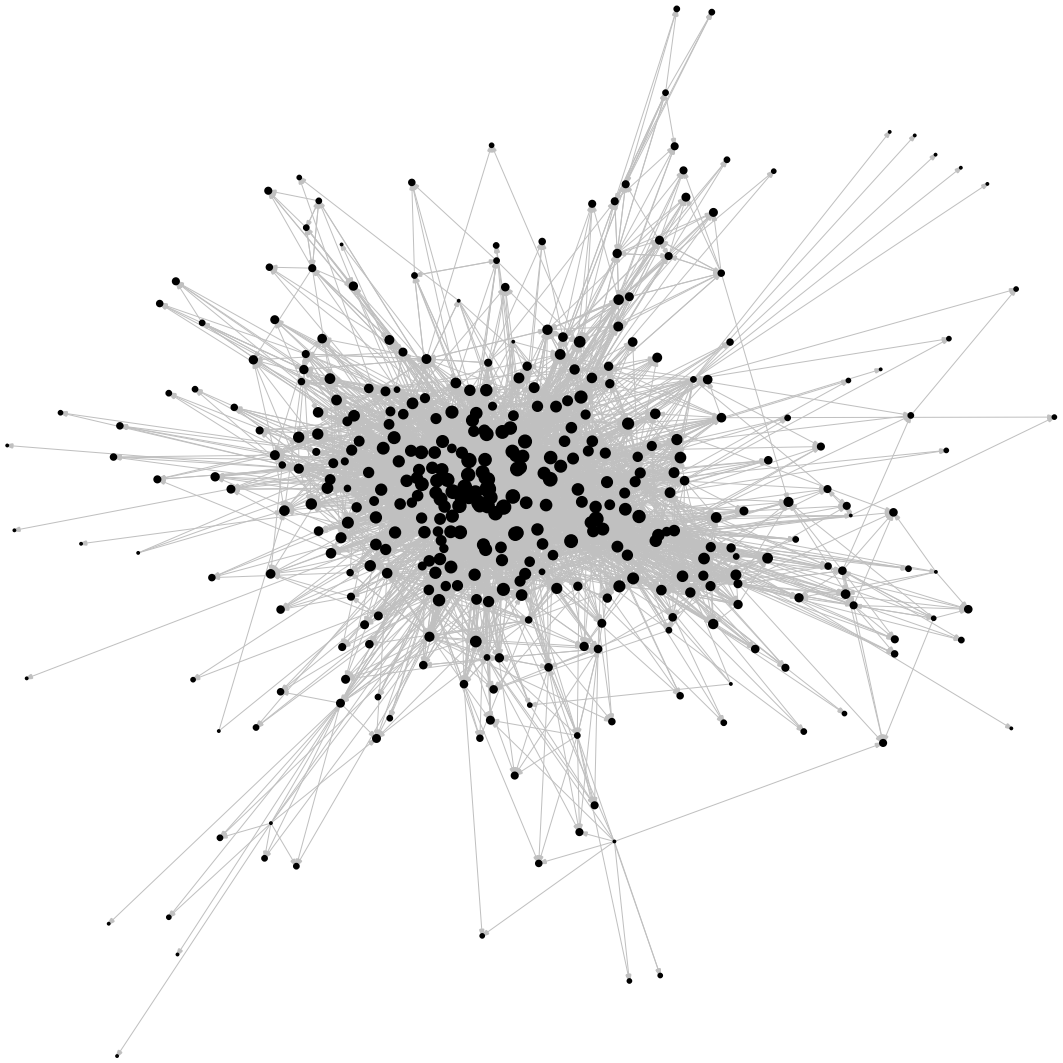
Many network studies ignore directional information by converting directed



Survey responses give rise to four possible classes of acquaintanceship ties between surveyed individuals: (1) complete data, reciprocal tie, (2) complete data, non-reciprocal tie, (3) incomplete data, non-reciprocal tie, and (4) missing data

Figure 4.3: Four possible classes of acquaintanceship ties.

networks to undirected. This conversion necessarily results in information loss, as both mutual and non-mutual connections are converted to undirected ties. As mentioned in Chapter 4, directional information is of fundamental importance for a number of networks. For example, a citation network (*who cites whom*) is useless without directional information. In an acquaintanceship network, such as the one studied here, directionality might also be useful: if studying social norms or subjective connotations of friendship, for example. For the purpose of this dissertation, however, I am exclusively interested in overall large-scale structures of network ties and their evolution. Also, as explained earlier in this chapter, the coauthorship and communication networks are natively undirected (article coauthoring and mailing list discussion are interactions without direction), so that maintaining information at a finer granularity does not enable additional analyses and comparisons between networks. For these reasons, I convert the acquaintanceship network to an undirected network.



The directed acquaintanceship network ($n = 385$, $m = 6,183$) diagrammed according to the Fruchterman-Reingold network layout algorithm. In the Figure, line width is fixed and node diameter is proportional to in-degree centrality, i.e., the more an individual has been indicated as an acquaintance, the larger its node diameter.

Figure 4.4: Weighted acquaintanceship network (directed).

In general, there are three different approaches one can take to convert the undirected tie typology presented in Figure 4.3 to a simple directed network: *complete-case analysis*, *available-case analysis*, and *imputation* [163]. Complete-

case analysis is the simplest approach: it only considers reciprocal ties with complete descriptions and it discards all incomplete data and non-reciprocal ties. In other words, if a complete-case approach is used, then edges are drawn between nodes only in the case (1) of Figure 4.3. An available-case approach allows more flexibility by including both complete and incomplete cases with reciprocal and non-reciprocal ties. In other words, when adopting an available-case approach, one would draw edges for cases (1), (2) and (3). The third approach, imputation, uses statistical techniques to replace missing data with expected values, i.e., it computes missing ties based on the topology of the rest of the network. When using imputation, one uses all available data — cases (1), (2) and (3) — as well as missing data — case (4).

For the purpose of this study, given the relatively high response rate (49%), I choose to employ an available-case analysis, i.e., include all reciprocal and non-reciprocal ties with both complete and partial descriptions—or, cases (1), (2) and (3), in Figure 4.3. Case (1)—reciprocal tie descriptions—are the simplest case: an edge between A and B is drawn for every reciprocal tie encountered. Cases (2) and (3)—non-reciprocal ties—are more complex to deal with. To draw these edges, I employ an approach called *reconstruction*, which is based upon the following assumption: if A describes a relationship with B, then a tie between A and B exists, regardless of the response provided by B. In statistics literature, it has been noted that reconstruction is a reliable technique of network data manipulation as long as it is appropriately justified [164]. Stork & Richards suggest two criteria for justification. The first is that the population of respondents should not be systematically different from the larger population. The second is that incomplete data should be manually inspected and checked for reliability. In order to justify reconstruction of ties based on non-reciprocal descriptions, I perform the following two tests.

First, I check that respondents' characteristics match closely those of the population at large, with no systematic differences. It might be the case that only a specific subgroup of the population responded to the survey (e.g., graduate students, or individuals from a certain department) causing the results, regardless of the obtained response rate, to be skewed and non representative of the population at large. For this reason, I compare academic attributes of respondents to those of the overall population. Table 4.3 lists counts for selected academic properties (academic affiliation, department, and position) of both the population at large (all) and the survey respondents (resp). Results of a Pearson product-moment correlation show that these sets of values are highly correlated (r is very close to 1 and p-value is low.), indicating that respondents' characteristics match the overall population. These populations counts are also plotted as pie charts, in Figure 4.5. From a quick visual analysis of the charts, it is clear that not only is the academic profile of respondents very diversified (pie charts on the left), but it also matches very well that of the broader population (pie charts on the right).

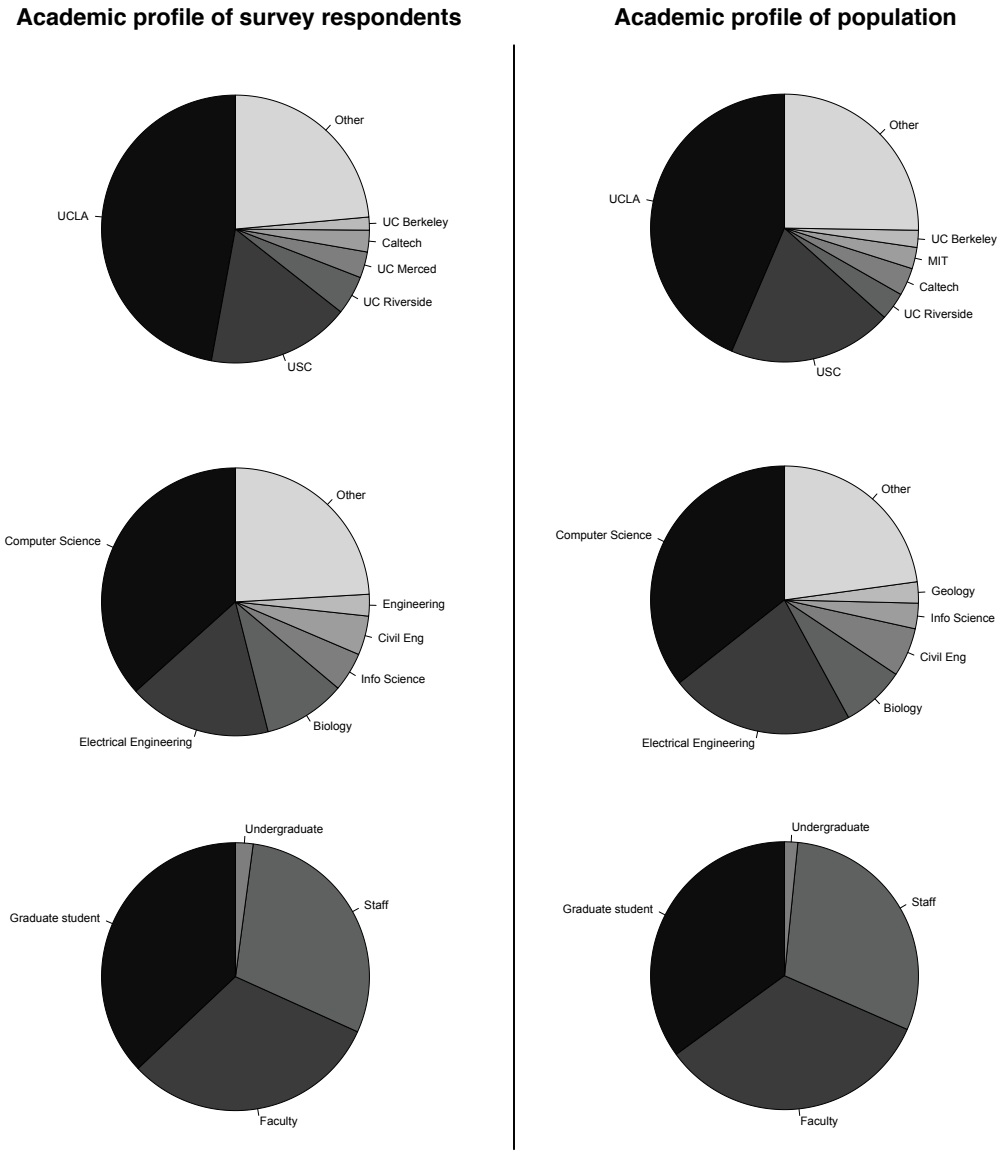
In order to confirm this finding (that the profile of respondents match that of the population at large), I also present in Figure 4.6, the coauthorship degree distribution of respondents and compare it to that of the broader population. The coauthorship degree distribution displays the number of times (frequency) that individuals that have authored the same number of papers (degree) appear in the bibliographic corpus. The distribution curves for survey respondents (left) and the entire population (right) presented in Figure 4.6 are very similar. They show that the majority of individuals among both respondents and the entire population authored only one paper (degree = 1). The frequency rapidly drops as a function of increasing degree. This suggests that survey respondents' centrality measures in the coauthorship network are a fair representation of the entire population.

	resp	all
Academic affiliation, $r = 0.995^{\dagger}$		
University of California, Los Angeles (UCLA)	90	169
University of Southern California (USC)	33	77
University of California, Riverside (UC Riverside)	9	13
California Institute of Technology (Caltech)	6	13
Massachusetts Institute of Technology (MIT)	5	10
University of California, Berkeley (UC Berkeley)	3	8
Academic department, $r = 0.981^{\dagger}$		
Computer Science	70	147
Electrical Engineering	33	91
Biology	19	31
Civil Engineering	9	24
Geology	9	14
Information Studies/Sciences	5	13
Academic position, $r = 0.991^{\dagger}$		
Graduate student	70	134
Faculty	59	128
Staff / Postdoc	56	106
Undergraduate Student	4	7

Population counts for selected socio-academic properties (academic affiliation, department, and position) of the individuals in the entire survey population (all) and survey respondents (resp), and associated Pearson correlation results, r. The † symbol indicates that a correlation has p-value < 0.05.

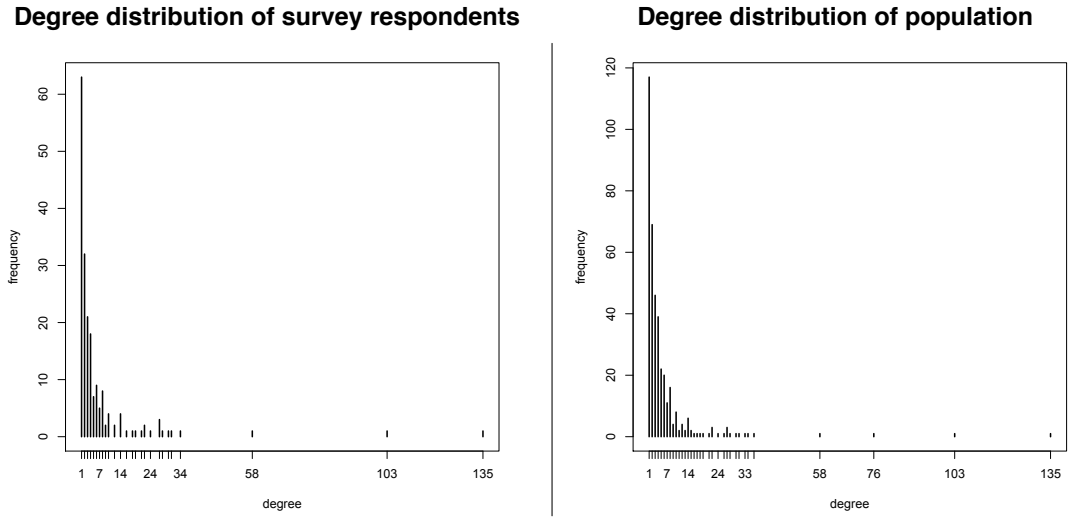
Table 4.3: Socio-academic profile of survey population and respondents

Second, I check collected data for reliability. As explained above, data collected from respondents can fall into one of three categories—cases (1), (2) and (3)—based on whether indicated acquaintance relationships are complete and/or reciprocal, as shown in Figure 4.3. The 191 respondents to the survey indicated a total of 6183 acquaintance relationships, i.e., possible edges in the network. The breakdown of the obtained results is presented in Table 4.4. The majority of ties indicated by respondents fall within case (1): both respondents (say, A and B) indicated to know each other in the survey. As discussed above, this is the simplest case. I can safely deem data reliable and their acquaintanceship tie valid and



Academic profile (institutional affiliation, department and academic position) of survey respondents (left) and broader population (right).

Figure 4.5: Academic profile of survey respondents and entire population.



Coauthorship degree distribution of survey respondents (left) and broader population (right).

Figure 4.6: Coauthorship degree distribution of survey respondents and entire population

thus construct an edge in the network between, say, A and B. From Table 4.4, a small portion of the ties indicated (about 16%) fall within case (2), which means that two survey respondents expressed diverging opinions about each other, e.g., respondent A indicated B as an acquaintance, but B did not. Clearly, some of these inconsistencies might be related to the fuzzy and subjective nature of the concept of acquaintanceship. Even though I specifically ask survey respondents to indicate individuals whom "they would say 'hi' to if they bumped into them", many individuals might still have different practices of greeting colleagues and acquaintances. Also, there are infinite levels and modalities of "knowing a person" and thus acquaintanceship relationships are not necessarily replicated, e.g., A might consider B an acquaintance, but B might think otherwise. However, manually inspecting the network of non-reciprocated acquaintanceship ties, I note that highly-connected individuals are responsible for the vast majority of

missed connections, i.e., people that had a lot of acquaintances failed to reciprocate ties the most. This finding suggests that many non-reciprocated ties might be due solely to a response error, i.e., individuals that had a large number of acquaintances to indicate in the survey form overlooked some of them. Based on this finding, I assume that all case (2) ties are non-reciprocated because of respondents' oversight. Thus, I consider them valid and, by reconstruction, I can include them as reciprocal ties in the network. Finally, a significant portion of the collected data falls within the third category—case (3). This category includes ties that are non-reciprocal, but for which only partial information is available, i.e., respondent A indicated B as an acquaintance but B did not respond to the survey. In order to justify the reconstruction of these ties, I look at the percentage of non-reciprocal ties in the complete dataset. Complete data (cases 1 and 2) makes up a total of 3732 ties, of which only 976 (26%) are non-reciprocal. Extending this finding to incomplete data, I set forth the following assumption: a case (3) tie between A and B is non-reciprocated only because B did not take the survey. Had B taken the survey, they would have indicated A as an acquaintance. Based on this assumption, I can then consider case (3) ties reliable for data reconstruction.

	Data	Type of tie	# ties
Case (1)	Complete	Reciprocal	**2756** (45%)
Case (2)	Complete	Non-reciprocal	**976** (16 %)
Case (3)	Incomplete	Non-reciprocal	**2451** (39 %)
Totals			**6183** (100%)

Table 4.4: A summary of the data collected in the social network survey.

Based on the above justification, I can safely reconstruct case (1), (2) and (3) ties. In other words, I draw undirected acquaintanceship ties between nodes whether they are reciprocated or not, both with complete and partial data. After

removing directionality, the resulting undirected network consists of the same number of nodes ($N = 385$), but the number of edges between them drops from 6183 to 4805, since bidirectional edges (two arrows) are collapsed to a single non-directional edge (no arrows).

At this point, the last step of the data processing involves dealing with the additional data collected in survey responses. In the second part of the survey, I ask respondents to indicate how long they have known their acquaintances for and how frequently they are in touch with them. Although answers to these questions were optional, many respondents provided this information. Table 4.5 summarizes the data collected in this portion of the study.

As shown in Table 4.5, respondents provided data relative to the frequency of communication for about three quarters of the total number of acquaintances indicated. Moreover, a quick analysis of the distribution of responses reveals that nearly half of all acquaintances (2245 out of 4621) communicate rarely or never. About a third of all ties (1539 out of 4621) are based on occasional communication. Only about a fifth of all ties relies on frequent communications (once a month and once a week). This information is used to assign a weight to the edges connecting acquaintances. Frequent communication are given a higher weight, based on the assumption that frequent communication involves a higher degree of cognizance among individuals, regardless of the nature of the relationship (formal or informal). If respondents indicate to communicate once a week, the edge among them is weighted 1.0; acquaintances communicating at least once a month are assigned with a weight of 0.75. Occasional communication is weighted 0.5 and even less frequent communication is weighted 0.25.

Similar to cases (1)-(4) discussed above, data relative to the frequency of communication among respondents might be non-reciprocal and incomplete. In line

Frequency of communication *How often do you communicate with [name]?*	
Did not respond (no data available)	1668 (26%)
Responded:	4621 (74%)
.....*Rarely or never (0.25)*	2245
.....*Occasionally (0.50)*	1539
.....*At least once/month (0.75)*	421
.....*At least once/week (1.0)*	416

Length of acquaintanceship *When did you first meet [name]?*	
Did not respond (no data available)	2454 (39%)
Responded:	3835 (61%)
.....*2001 or earlier*	933
.....*2002*	292
.....*2003*	363
.....*2004*	430
.....*2005*	580
.....*2006*	516
.....*2007*	467
.....*2008*	191
.....*2009 (this year)*	63

Table 4.5: A summary of the data collected in the second part of the social network survey.

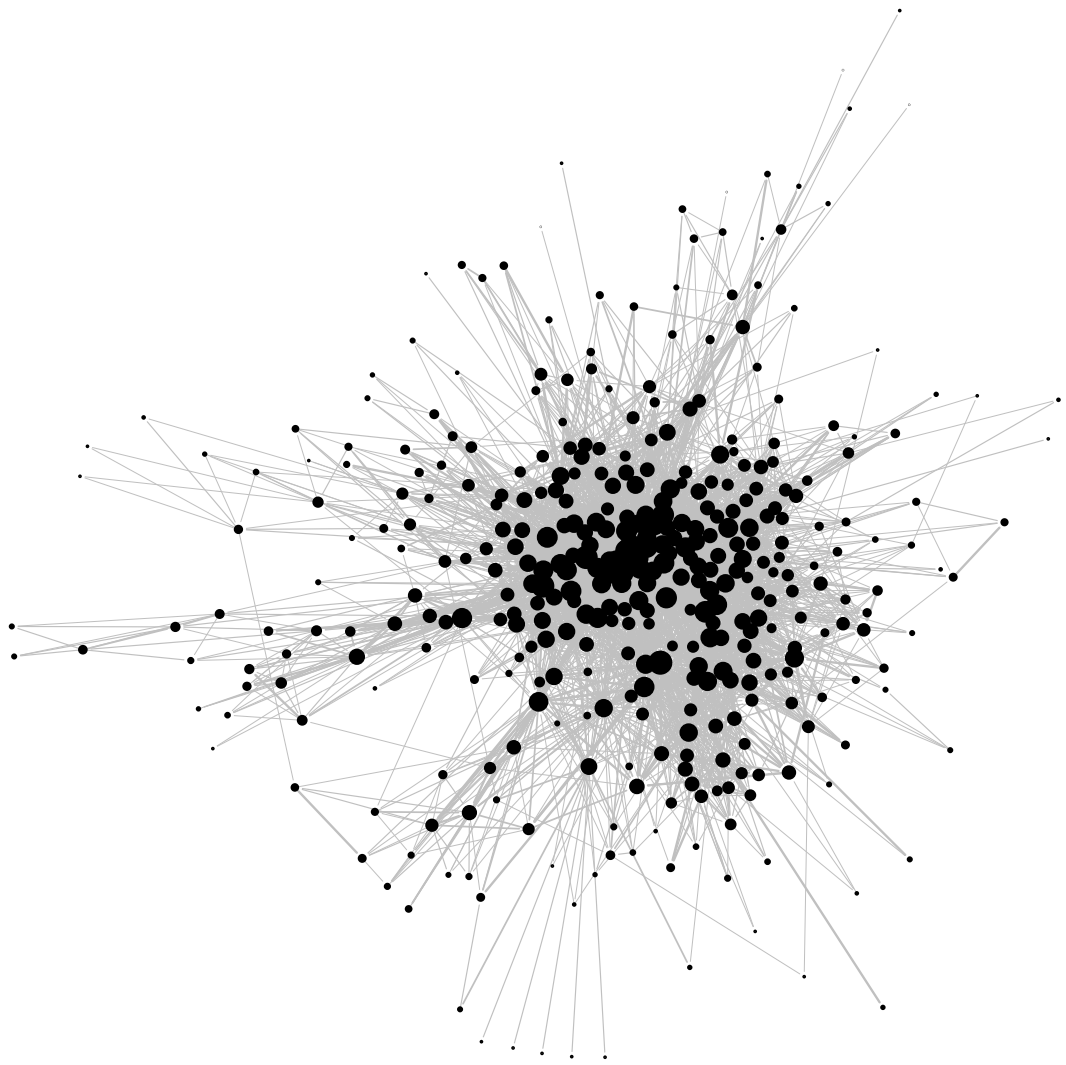
with my decisions made for the construction of the weighted network, discussed above, I preserve all collected data (including partial and non-reciprocal data) to construct weighted edges. So, for example, if respondent A indicates to communicate occasionally with respondent B, I assign a weight of 0.5 between node A and B, even if B did not indicate a frequency of communication with A. One problem, however, is to deal with incongruent responses—for example, A indicates occasional communication with B (0.5), but B indicates rare communication (0.25) with A. In these cases, the highest available weight is chosen (in the example, the edge between A and B would be given weight 0.5). Moreover, when weight data

are missing entirely (i.e., A and B are connected by an edge but neither of them indicated frequency of communication), then a default weight of 0.25 is assigned to edges. This is a fair assumption, given the fact that the vast majority of ties indicated is, in fact, predicated by rare communication.

The other set of information provided by respondents in the second portion of the survey is the length of acquaintanceship. As shown in Table 4.5, this information was obtained only for 61% of all ties. Moreover, a quick investigation of the responses reveals that most respondents have known each other for a relatively long time (at least since 2001). The vast majority of all other acquaintances has begun in the period between 2004 and 2007. This information is not *per se* useful for the construction of the overall acquaintanceship network, but it is necessary to study its evolution over time, presented in much detail later in this dissertation, in Chapter 7. At this point, it is worth mentioning that these data can be handled using exactly the same method presented above: all available data are used, including partial and non-reciprocal data, to reconstruct the length of acquaintanceship relationships. When incongruent data are detected, the oldest available value is recorded. For example, if A indicates that has known B since 2002, while B indicates that has known A since 2003, the edge between A and B is recorded to exist since 2002. Missing data are not reconstructed.

Using the information provided in the survey, computed as discussed above, a weighted network of acquaintanceship is constructed. It is depicted in Figure 4.7. In the figure, line width represents edge weight, where more frequent communication activities are represented by wider and more marked lines; the diameter of the nodes is proportional to the node strength, where more central nodes (i.e., individuals that both know and are known by more people) have larger diameters. Some statistics relative to the topology of the acquaintanceship network are

included in Table 4.6.



The undirected acquaintanceship network ($n = 385$, $m = 4,805$) diagrammed according to the Fruchterman-Reingold network layout algorithm. In the Figure, line width is fixed and node diameter is proportional to in-degree centrality, i.e., the more an individual has been indicated as an acquaintance, the larger its node diameter.

Figure 4.7: Weighted acquaintanceship network (undirected).

An analysis of Table 4.6 reveals that the acquaintanceship network, with 385

Topological property	Value
Nodes (individuals), n	385
Edges (recorded acquaintances), m	4 805
Connected components	1 (385)
Diameter	5
Average path length, ℓ	2.427
Maximal cliques	5 925
Largest clique	20
Clustering coefficient, C	0.359

Table 4.6: Topological properties of the acquaintanceship network.

nodes, is comparable in size to the coauthorship network (Table 4.1). What differs greatly, however, is the number of edges in the acquaintanceship network, which is almost three times bigger than that of the coauthorship network. This value alone indicates that, in general, researchers have more acquaintances than coauthors. All the nodes in the acquaintanceship network form a single connected component with relatively short diameter (5) and average path length (2.427). This means that any member in the network of acquaintanceship is easily accessible within few hops in the network. The values related to the topological features—very high number of maximal cliques and moderate clustering coefficient—expose a very dense social environment in which "everyone knows everyone". Among the three studied networks, the acquaintanceship network is that one with the most tangible small-world properties—short average path length (accessibility) and moderately high clustering coefficient (density).

4.4 Socio-academic configuration

The networks of collaboration constructed in this chapter are based on the observation of specific interactions among CENS researchers, i.e., their coauthorship, communication and acquaintanceship patterns. However, as anticipated in Chap-

ters 1 and 2, this dissertation also aims to explore the social and academic landscape in which these collaboration patterns take place, and specifically how the structure and evolution of these networks relate to organizational, disciplinary, institutional and international arrangements of collaboration at CENS. In order to support this portion of the study, I collect additional information about every individual in the population under study, namely a) academic affiliation, b) academic department, c) academic position, and d) country of origin. A summary of the population counts for these parameters, and for each network under study, are presented in Table 4.7.

These socio-academic data are collected via manual techniques, i.e., gathering required information on the authors' personal web pages, curriculum vitae, and consulting online directories from university and departmental web sites. The data presented in Table 4.7 summarizes the latest available socio-academic data (year 2009). It is worth noting, however, that all the parameters collected (except for country of origin) are subject to change over time. Scholars are likely to change academic institution, department and position over the period studied here (2001-2009), e.g., a graduate student may become a Postdoctoral Researcher and later an Assistant Professor. For this reason, these parameters are also recorded historically by inspection of researchers' curriculum vitae and biographies. Curriculum vitae are also useful to collect the country of origin of researchers, which, for the purpose of this study, I consider to be the country of principal citizenship, if available, or the country in which researchers pursued their lowest recorded level of education. It is worth noting that an additional characteristic—workplace location—was collected for occupants of the CENS headquarters (3551 Boelter Hall). This information is discussed in detail with the analysis of physical proximity, in Chapter 6, § 6.7.

	Node property	(a)	(b)	(c)
Academic affiliation	**University of California, Los Angeles (UCLA)**	169	90	169
	University of Southern California (USC)	77	11	73
	University of California, Riverside (UC Riverside)	13	3	13
	California Institute of Technology (Caltech)	13	3	13
	Massachusetts Institute of Technology (MIT)	10	2	10
	University of California, Berkeley (UC Berkeley)	8	1	8
	University of California, Merced (UC Merced)	7	4	7
	University of Illinois at Urbana-Champaign (UIUC)	4	-	4
	Stanford University	4	-	4
	State University of New York at Stony Brook (SUNYSB)	4	-	4
	Carnegie Mellon University (CMU)	4	-	4
	None	-	3	-
Academic department	Computer Science	147	54	146
	Electrical Engineering	91	28	86
	Biology	31	5	30
	Civil Engineering	24	2	24
	Geology	14	2	14
	Information Studies/Sciences	13	2	13
	Environmental Sciences	11	2	11
	Engineering (others)	7	1	7
	Education	7	-	7
	Marine Biology	6	1	6
	Film, media, arts	5	5	5
	Statistics	4	5	4
Academic position	Graduate Student	138	48	134
	Staff / Research Associate (Staff)	90	32	89
	Full Professor (Professor)	66	15	65
	Assistant Professor	35	3	35
	Associate Professor	27	4	27
	Postdoctoral Student (PostDoc)	26	3	25
	Undergraduate Student	7	7	7
	Lecturer	4	1	3
	CENS Admins	-	3	-
Country of origin	United States of America (USA)	191	78	191
	India	48	15	48
	China	26	3	22
	South Korea (Korea)	12	2	11
	Italy	11	-	11
	Australia	6	2	6
	Mexico	5	-	5
	Iran	5	1	5
	Brazil	5	2	5
	Taiwan	4	1	4
	Greece	4	2	4
	Totals (n)	**391**	**119**	**385**

Population counts for selected socio-academic properties (academic affiliation, department, position and country of origin) of the individuals in the (a) coauthorship, (b) communication and (c) acquaintanceship networks of collaboration. Entries sorted by frequency in the coauthorship network. Abbreviations listed in brackets, when available. CENS official member institutions are indicated in bold.

Table 4.7: Socio-academic profile of the coauthorship, communication and acquaintanceship networks of collaboration

A quick analysis of Table 4.7 reveals that, overall, the three collaboration networks do not differ very much in their social and academic configurations. In particular, since the population of the acquaintanceship study is derived directly from the coauthorship network, and the survey results cover nearly the entire population set ($n = 391$ and $n = 385$, in the coauthorship and acquaintanceship networks, respectively), their scores hardly differ. The results obtained for the communication network ($n = 119$) also present a population distribution that is similar to that of the coauthorship network. The population distribution of the communication network, however, presents some minor, yet interesting differences. For example, only in this communication network does one find individuals that are not affiliated with any academic institution, and CENS administrative staff. This shows that some individuals that do not normally take part in the authoring of papers might however be involved in open discussions on dedicated mailing lists.

Overall, Table 4.7 displays a population scenario dominated by the presence of UCLA scholars, which make up almost half of the population. The other member universities affiliated with CENS (USC, UC Riverside, Caltech and Merced) account, altogether, for about a quarter of the population. The population count by department is vastly taken up by two disciplines: Computer Science and Electrical Engineering. Top 'science' disciplines are Biology, Geology and the Environmental and Marine Sciences. The repartition by academic position shows doctoral students and staff researchers making up the vast majority of the population, the remainder being composed by a balanced mix of professors (at all levels) and postdoctoral students. Finally, about half of the individuals in the collaboration networks come from the United States. Researchers from India and China make up about a quarter of the population. Overall, this scenario is not very surprising, considering that a) UCLA is the central institution behind CENS and the

location of its headquarters, b) Computer Science and Electrical Engineering are the core 'technology' disciplines in the domain of sensor network research, and c) CENS is a scientific enterprise largely funded by agencies from the United States. What is interesting, from this preliminary analysis, is the large involvement of doctoral students in CENS research. These findings are supplemented by further data, analyzed and discussed in detail in the following chapters.

4.5 Summary

In this chapter, I demonstrate how different manifestations of collaborative activity can be represented as networks. Three data sources (a list of publications, an archive of mailing list emails, and the results of a social survey) are converted to dedicated networks of CENS collaboration: coauthorship, communication, and social networks.

A topological analysis of these networks reveals that all have a peculiar configuration. The coauthorship network features nearly 400 authors and 2000 connections among them, in five separate components. The vast majority of authors, however, are part of a giant component, indicating that the CENS has a solid, connected core of collaborating researchers. The giant component has a low clustering coefficient, indicating that collaborations are sparse, i.e., CENS authors connect on paper with a relatively large number of other authors.

The communication network is much smaller in size, compared to the other networks, with just over 100 individuals involved in nearly 1000 discussions on mailing lists. Communication patterns are centered around a single connected component, and are very clustered, i.e., individuals always tend to communicate in online discussions with the same circles of contacts.

The population making up the acquaintanceship network is comparable in size to the coauthorship network, but there are many more connections among individuals: nearly 5000 knowledge relationships among CENS researchers are recorded. This network is solidly connected in a single, dense component.

In the chapters that follow, the structure and dynamics of these networks are analyzed and discussed.

CHAPTER 5

Results: Structural analysis

The previous chapter details the construction of three networks that embody collaboration activities at CENS: how researchers write papers, how they communicate on mailing lists, and how they are acquainted with one another. This chapter presents the results of a comparative structural analysis of these networks. This analysis addresses the first research question of this dissertation, restated here:

> **Research Question #1.** What types of structural communities can be detected in the coauthorship, communication, and acquaintanceship networks of CENS? How do these structures relate to each other and to the disciplinary and institutional arrangements of CENS?

This chapter explores the community structure of the networks of collaboration at CENS, i.e., their topological repartition into clusters. By performing comparative analyses of community structure, I expose how CENS researchers organize themselves in scholarly, social, and communication circles, and how these structures relate to the disciplinary and institutional arrangements of CENS.

5.1 Detection of community structure

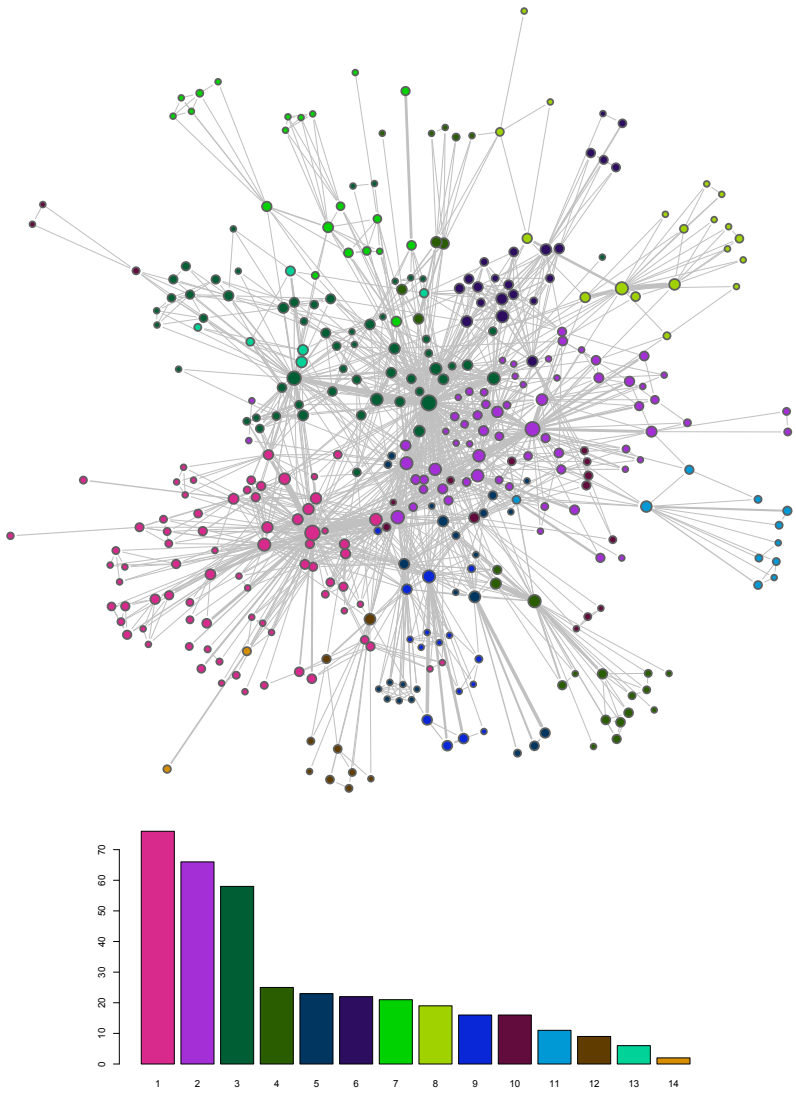
While the analysis of the topological properties presented in the previous chapter gives a general idea about the structural configuration of the networks under study, further analysis is required to understand more in detail the actual arrangement of CENS collaboration circles. Studying a network's clique structure, for example, can throw light on its mechanisms of information exchange, homophily and other forms of social seclusion. But clique analysis is only convenient for networks composed of only a few actors. The networks studied here feature so many overlapping maximal cliques (see Table 4.1) that an in-depth clique analysis is impossible. For this reason, I analyze another structural property of the network, known as community structure, and introduced in § 3.1.2. A study of community structure enables a network to be partitioned in clusters and comparatively analyze how different configurations of collaboration patterns overlap with each other.

As discussed in Chapter 3, the community structure of a network reveals its underlying "cliquish" groupings of nodes that are highly connected between them, but poorly connected to other vertices. In the networks discussed in this dissertation, the clusters detected via a structural analysis correspond to communities of collaborating researchers that write papers together, communicate over email, and know each other. The community detection method used in this dissertation is the spinglass algorithm [145]. This method relies on an analogy between the statistical mechanics of networks and physical spin glass models to deconstruct a network into communities. In doing so, it assigns a community membership value to each node. Thus, individuals that are in the same structural community are given the same membership value. It is worth noting that the membership value is a nominal value identifying distinction, not relative similarity between identi-

fied communities. For example, the node that represents myself (`id = a.pepe`) in the CENS collaboration networks can be associated with information regarding my membership to different communities (Figure 5.1). In my case, I belong to community #3 in the coauthorship network, community #2 in the acquaintanceship network, and I am not a member of any community in the communication network (because I have not participated in discussions on mailing lists).

Figure 5.1: Community membership as node metadata.

The structural communities found in the CENS collaboration networks via the spinglass algorithm are diagrammed in Figures 5.2, 5.3, and 5.4. Each Figure presents a network with nodes colored according to the structural community that they belong to. Each community is represented using a different color (or shade). Node diameter represents the betweenness centrality score of nodes, where more central vertices have larger diameters. The histogram associated with each Figure describes the frequency distribution of each community, i.e., the number of scholars in each identified structural community. It is worth noting that structural communities are computed only on the giant connected component. The communication and acquaintanceship networks feature only one connected component, so the entire networks were employed to detect structural communities. The coauthorship network, however, features 5 connected components (see Table 4.1), thus structural communities were only computed on the giant component, consisting of 377 vertices.



Structural communities in the CENS coauthorship network detected according to the spinglass algorithm. Node color represents structural community membership. Node diameter represents betweenness centrality. Associated histogram describes the frequency distribution of each community.

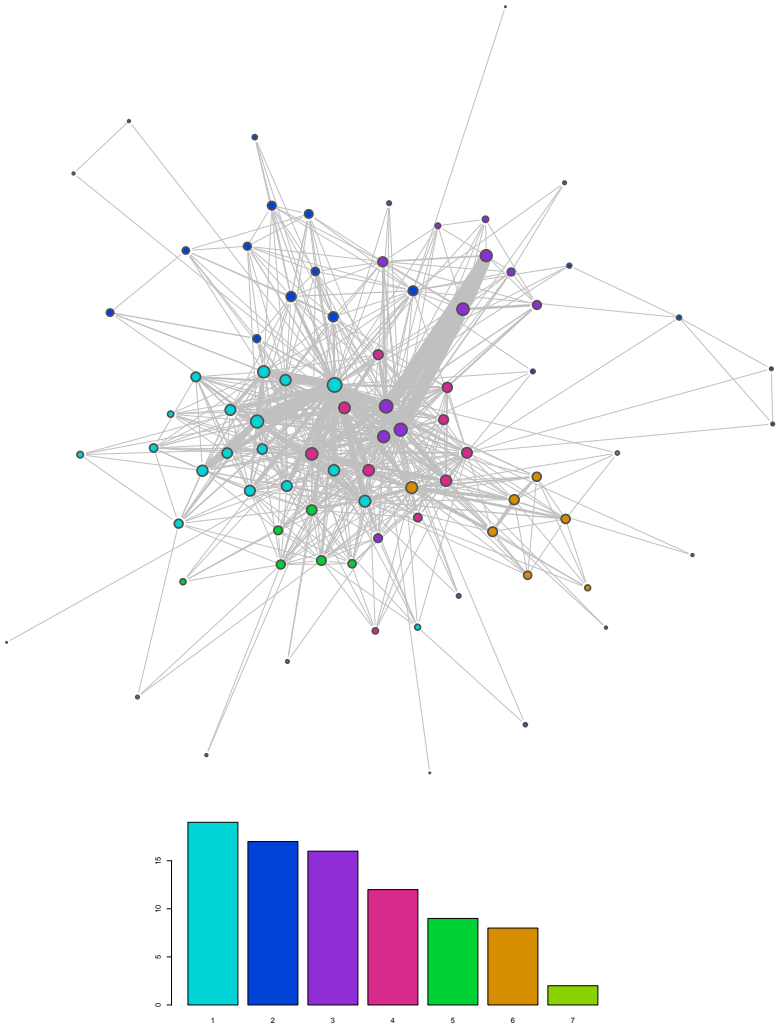
Figure 5.2: Coauthorship network: detected community structure.

A total of 14 structural communities were found in the CENS coauthorship network (Figure 5.2). The population distribution histogram shows that three

single communities (with populations 76, 66, and 58) cover about half of the entire CENS coauthorship population. The remaining communities are smaller in size (with an average of 15 members). At a first glance, this indicates that three large-scale coauthorship circles exist, possibly corresponding with CENS' main application or system areas. The other communities, more limited in size, correspond with more specialized domains.

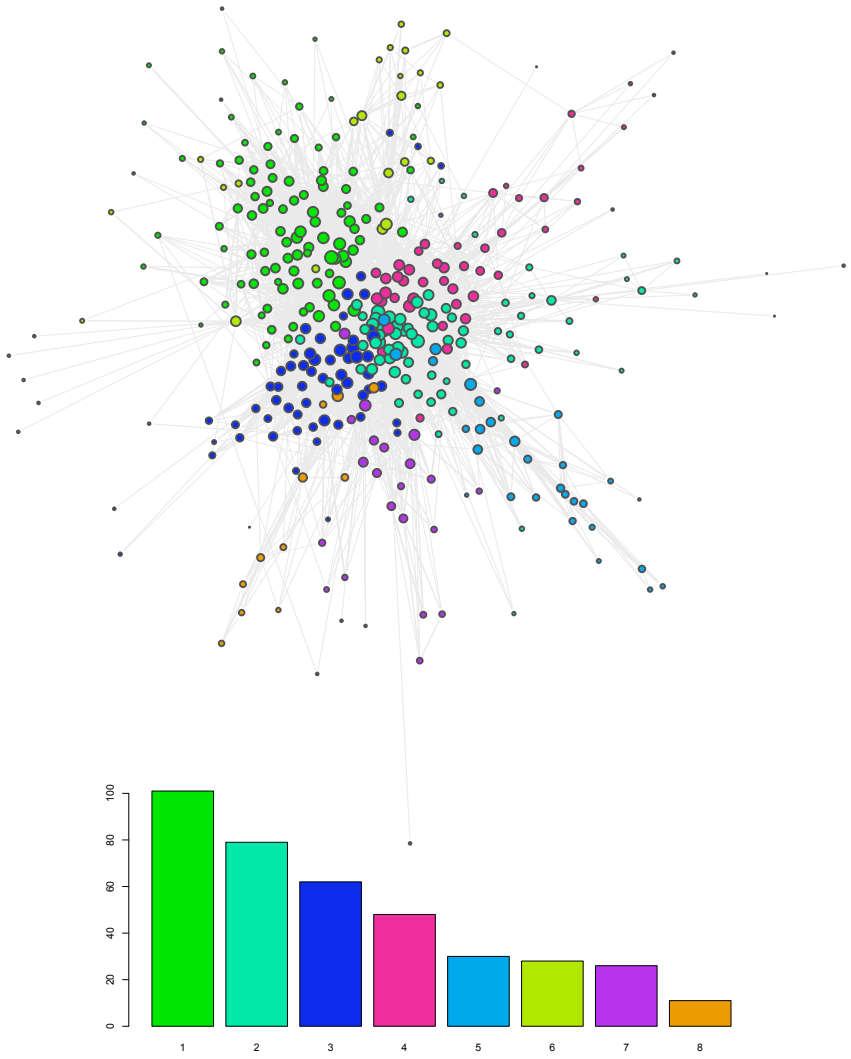
A total of 7 structural communities were found in the communication network (Figure 5.3). As the Figure shows, the population is distributed in the communities more homogeneously with respect to the coauthorship network. Six out of the seven communities are made up by a similar number of members, ranging between 9 and 16. Only one community is much smaller with two members only.

The acquaintanceship network was partitioned into 8 communities (Figure 5.4). The population distribution shows that there are three communities that are highly populated (with populations 101, 79, and 62) and the remainder of the nodes more or less evenly distributed in the remaining five communities.



Structural communities in the CENS communication network detected according to the spinglass algorithm. Node color represents structural community membership. Node diameter represents betweenness centrality. Associated histogram describes the frequency distribution of each community.

Figure 5.3: Communication network: detected community structure.



Structural communities in the CENS acquaintanceship network detected according to the spinglass algorithm. Node color represents structural community membership. Node diameter represents betweenness centrality. Associated histogram describes the frequency distribution of each community.

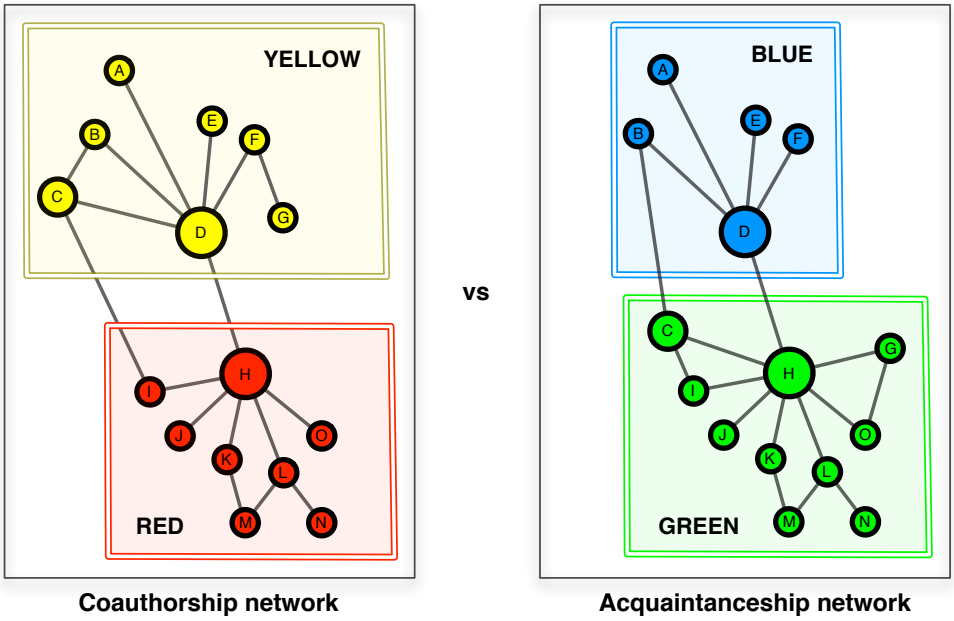
Figure 5.4: Acquaintanceship network: detected community structure.

5.2 Comparative analysis of community structure

The community structure analysis presented above shows the repartition of the CENS networks into clusters of collaboration. In order to reveal how these arrangements of collaboration differ with each other, i.e., how researchers group into communities in their scholarly, communication, and social networks, I perform here a statistical comparison of the detected communities, using two statistical independence tests: Pearson's χ^2 and Fisher's exact tests. These comparisons are aimed at understanding whether certain individuals tend to group with the same individuals across different manifestations of collaboration, i.e., coauthorship, communication and acquaintanceship. In other words, these independence tests address this question: how well do communities of coauthors, discussants, and acquaintances overlap with one another?

5.2.1 Tests for statistical independence

Pearson's χ^2 (chi-square) test is one of the most widely used statistical tests of independence. As an example consider the two networks depicted in Figure 5.5. The two networks diagram the same 15 nodes (A-O) in two different networks. Let us suppose that the network on the left represents coauthorship, and the network on the right represents acquaintanceship. These networks are very similar. The only difference between them is that the nodes C and G are more prominently connected to the hub D in the coauthorship network, and hub H in the acquaintanceship network. From a visual analysis, one can say that coauthorship and communication patterns in these two networks are very similar and thus, individuals that write papers together, also know each other. When analyzing larger networks, however, a visual investigation is not feasible.



Two fictitious networks, each one composed of 15 nodes (A-N). The nodes represented in each network are the same, but the interactions among them are different. The community detection algorithm partitions these two networks in different ways: the network on the left is partitioned in communities YELLOW and RED, while the network on the right is partitioned in communities BLUE and GREEN. Community membership is depicted by node color and an enclosing box.

Figure 5.5: Two fictitious networks and detected community structure.

One solution is to detect the community structure in each network and compare them statistically. Let us suppose that a community detection mechanism subdivides each network into two communities. In particular, the coauthorship network is divided into communities YELLOW (composed of nodes A through G) and RED (composed of nodes H through N). The communication network is divided into communities BLUE (A, B, D, E, F) and GREEN, $(H - N, C, G)$. In this simple case, it is fairly easy to deduce from the network visualization that community pairs YELLOW—BLUE and RED—GREEN overlap. But for larger and more complex networks, visual analysis might not be sufficient and a

statistical test of independence might be necessary. One can test for statistical independence between two groups (in this case, the detected communities in the two different networks), by conducting a Pearson's χ^2 analysis.

The community membership values depicted in Figure 5.5 can be expressed as a contingency table. In a contingency table, the x-axis elements represent a distinct community membership value in a network (e.g., YELLOW and RED) and the y-axis represents a distinct community membership value in the other network (e.g., BLUE and GREEN). Cell values in each contingency table identify the number of observed occurrences of an x/y relationship. As an example, the contingency table summarizing the community membership of the two networks of Figure 5.5 is presented in Table 5.1. The table displays how the population of a community is decomposed in the other network. For example, a total of seven nodes makes up the YELLOW community of the coauthorship network. Of these, five are members of the BLUE community in the acquaintanceship network (nodes A, B, D, E, F), and two of them belong to the GREEN community (nodes C, G).

Membership	YELLOW	RED	**Totals**
BLUE	5	0	5
GREEN	2	8	10
Totals	7	8	**15**

Table 5.1: Contingency table displaying the community membership distribution of two fictitious networks.

Subjecting this contingency table to a Pearson's χ^2 test one can determine whether community membership in one network is dependent or independent on membership in the other network. In other words, the χ^2 test calculates the values we would expect for a contingency table in case of independence of the variables and then computes deviations between expected and actual values. The χ^2 score for the contingency table presented in Table 5.1 is 5.65, a value suffi-

ciently high to indicate variability in the actual data, compared to the expected values. The test also returns a p-value of 0.02 which confirms that the community membership in the two networks are statistically related. The p-value denotes the probability that an association is random (i.e. a p-value greater than 0.05 is generally considered statistically independent).

Pearson's χ^2 is a reliable test of independence when dealing with samples of sufficiently large size. When samples are small, however, contingency tables have sparse data, i.e., many cells with expected values below 5. It has been noted that, as a rule of thumb, results of a χ^2 test should be considered suspicious or invalid if more than one fifth of its cell expectations are below 5 [165]. As demonstrated in the next section, some contingency tables studied in this dissertation have sparse data. In order to remedy this situation, statisticians can employ different techniques.

One obvious technique involves directly manipulating the data in the table. One manipulation technique, known as "re-binning" involves grouping together columns and rows of sparsely populated contingency tables. For example, if a table presents Likert-scale responses ("Strongly Agree", "Agree", "Neither agree or disagree", etc.), it is possible to collapse responses to "Strongly Agree" and "Agree" into a single class, to augment cell frequency. Another technique of data manipulation is to remove rows and columns with low frequency counts.

Other techniques to deal with table sparseness involve using alternative tests of independence. For example, one viable alternative when dealing with simulated data is to run a variation of the traditional χ^2 with a Monte Carlo simulation, which computes the p-value for a Monte Carlo test with a number of replicates [166, 167]. Another alternative is to replace the χ^2 test with a Fisher's exact test of independence [168]. This test is particularly convenient for contingency

tables with small samples and sparse data. The advantage of Fisher's test is that it calculates statistical independence exactly and not based on an approximation; this advantage, however, comes at a cost: Fisher's test can be extremely computation-intensive and resource-consuming, especially for large tables (larger than 6×6).

In this dissertation, in order to deal with table sparseness I use both direct manipulation of table data (re-binning and removal techniques) and computation of different statistical significance tests (both Pearson's χ^2 and Fisher's exact tests).

5.2.2 Community structure across networks

In this section, I present the results of a comparative analysis of structure among the networks of collaboration. In other words, I run the tests of independence presented above (Pearson's χ^2 and Fisher's exact tests) on the coauthorship, communication and acquaintanceship networks to determine whether community membership in one network is dependent or independent on membership in the other network. The community structures describing how researchers organize themselves in scholarly, communication, and social circles, diagrammed in Figures 5.2, 5.3, and 5.4, are converted to contingency tables in order to allow data manipulation and statistical analysis. These contingency tables are included in Tables 5.2, 5.3, and 5.4.

Each one of these tables is a numerical representation of the overlap between community membership in two networks. For example, Table 5.2 displays the association between the communication and coauthorship networks, i.e. between circles of collaboration depicted in Figures 5.2 and 5.3. Columns in this table (the x-axis) list community membership values in the communication network

and rows (y-axis) in the coauthorship network. The description under Table 5.2 discusses how to interpret this and following contingency tables.

		Communication network							
		1	2	3	4	5	6	7	**T**
Coauthorship network	1	17	3	3	2	3			28
	2	2		1	7		8		18
	3			8				1	9
	4		3	1	2	3			9
	5								0
	6		2						2
	7								0
	8		6		1				7
	9							1	1
	10		2						2
	11					1			1
	12			3					3
	13					2			2
	14		1						1
	T	19	17	16	12	9	8	2	**83**

Community membership association between the communication and coauthorship networks. "T" stands for "totals". Columns present community membership values in the communication network and rows in the coauthorship network. This table can be read as follows. There are a total of 83 nodes that are found both in the communication and coauthorship networks (i.e., the communication and coauthorship networks overlap by 83 nodes). In the communication network, these 83 nodes are partitioned into 7 different communities, with populations [19, 17, 16, 12, 9, 8, 2] (bottom row of the table). The columns display how these communication-based communities overlap with coauthorship-based communities. For example, looking at the first column, one can see that the majority of nodes (17 out of 19) composing community #1 of the communication network are found in community #1 of the coauthorship network. This indicates that there is a strong correlation between community #1 in the communication network and community #1 in the coauthorship network, i.e., most of the people that compose community #1 in the communication network are found to connect both on mailing list platforms and on scholarly papers.

Table 5.2: Contingency table: communication vs. coauthorship networks.

		Communication network							
		1	2	3	4	5	6	7	**T**
Acquaintanceship	1		2	1	2	1	7		13
	2	15	2	10	5	4	1		37
	3	4	3	2	2	2			13
	4		6	3	1			1	11
	5		3		1				4
	6		1					1	2
	7				1	1			2
	8					1			1
	T	19	17	16	12	9	8	2	**83**

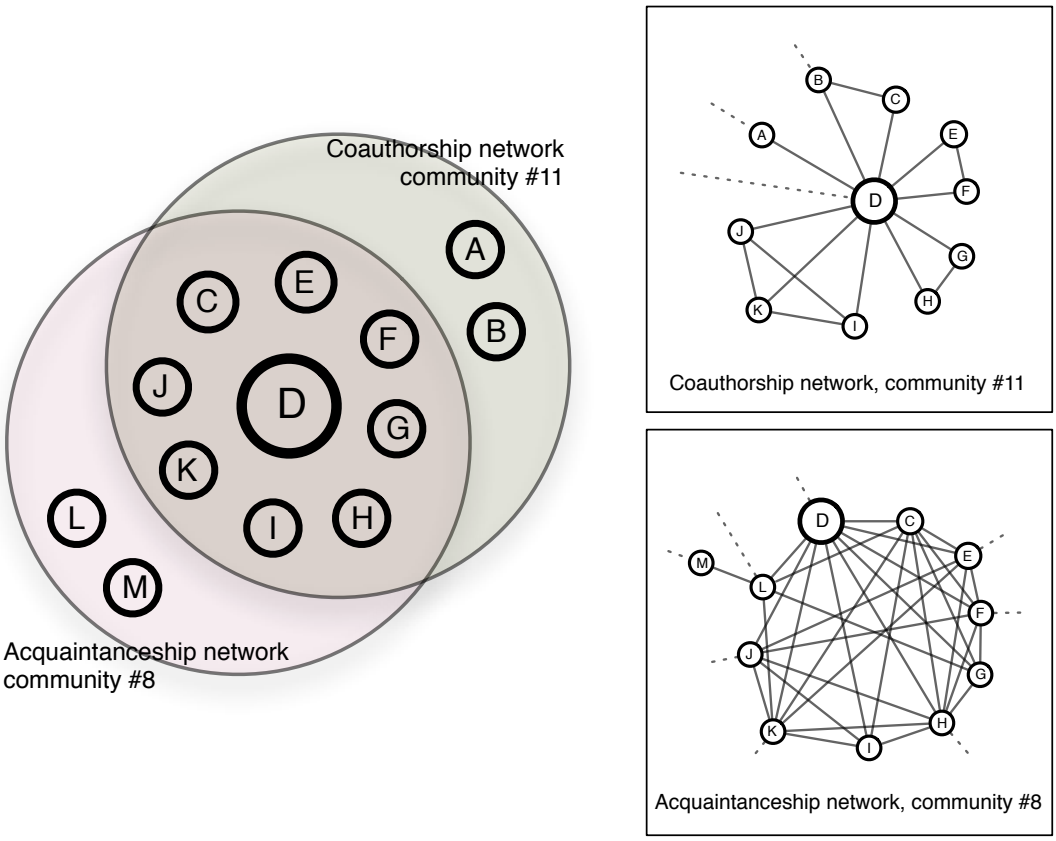
Table 5.3: Contingency table: communication vs. acquaintanceship.

		Coauthorship network														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	**T**
Acquaintanceship	1	50	7	37				1	2	2					2	101
	2	1	23	10	8	8	1	18	3	1		1				74
	3		34	5			1		14	1		1				56
	4	6	2	2	1			1		12	15			6		45
	5	1			16	13										30
	6	18											9			27
	7			4		1	20				1					26
	8					1		1				9				11
	T	76	66	58	25	23	22	21	19	16	16	11	9	6	2	**370**

Table 5.4: Contingency table: coauthorship vs. acquaintanceship.

Analyzing the composition of the contingency tables (Tables 5.2, 5.3, and 5.4), it can be seen that some communities overlap nearly perfectly, while others are highly partitioned. For example, looking at Table 5.4, it can be seen that community # 11 in the coauthorship network and community # 8 in the acquaintanceship network overlap almost completely. Both communities are composed of 11 individuals and 9 of them are found to be part of the same communities of coauthorship and acquaintanceship. A closer look at the composition of these communities reveals that they represent a collaboration circle centered around the Computer Vision Laboratory of UCLA. Let us employ this collaboration circle as an anecdotal example to explore more in detail the overlap between communities of coauthors and acquaintances. Figure 5.6 depicts such anecdotal example.

Coauthorship community # 11 and acquaintanceship community # 8 are diagrammed in the inset images of Figure 5.6. The main image of Figure 5.6 shows the overlap between these communities, with nine members ($C - K$) common to both communities. The inset image at the top depicts community # 11 in the coauthorship network. This coauthorship community is tightly centered around one individual (node D) whom collaborates separately with different groups (e.g., with B and C, with I, J, and K, etc.). It is interesting to note that this community is very marginalized from the rest of the coauthorship network: only nodes A, B, and D have connections with outside this community (dashed edges). The inset image at the bottom is community # 8 in the acquaintanceship network. This community is composed of roughly the same nodes, but the relationships among them are more frequent and dense: nearly everyone is connected to everyone else in the acquaintanceship network. This indicates that even though not all members of a coauthoring community collaborate directly with each other, they are likely to know each other. Also, while the members of this community are separated from the rest of the coauthorship network (i.e., they only write papers



The inset image in the top right corner shows community # 11 in the coauthorship network ($n = 11, m = 16$). The inset image at the bottom right corner shows community # 8 in the acquaintanceship network ($n = 11, m = 29$). The main image shows the overlap between these two communities.

Figure 5.6: Anecdotal example: overlap of coauthorship and acquaintanceship communities.

with each other), they are considerably more integrated with the social network (dashed lines represent social relationships with the outside).

This anecdotal example provides a precise understanding of the scholarly and social organization of the Computer Vision Lab at UCLA. This analysis is limited to a specific case study. In order to determine statistically the level of independence between community membership at a broader level, it is necessary to run

a significance test on Tables 5.2, 5.3, and 5.4. These tables, however, are not suitable for a Pearson's χ^2 or a Fisher's exact test, because of data sparseness. It is natural for these tables to be sparsely populated, i.e., it is normal that a high proportion of the cells are low (or zero). This is because, across all studied networks, individuals tend to cluster together with the same individuals across all networks. The Fisher exact test would be a viable alternative to remedy table sparseness, but its computation is too resource-consuming and thus unfeasible for tables of this size.

In order to allow both Pearson's and Fisher tests, I produce reduced versions of all contingency tables by removal of rows and columns with low frequency counts. The removal is performed manually, in iterations. At every iteration, I remove the row or column with the lowest count value. I follow this procedure until less than one fifth of the cells have values below 5. For example, for Table 5.2, I begin the removal with one of the rows with total count of 1, i.e., rows 5, 9, or 14. By removal of few rows and columns with low counts, table sparseness is greatly reduced. Moreover, the overall population and composition of the contingency table is not affected greatly. The original population sizes of all three contingency tables (83, 83, 370) are only reduced by about a third or less (49, 58, 224). The three reduced contingency tables are then subjected to a Pearson's χ^2 test and a Fisher's exact test. Results from the tests are summarized in Table 5.5.

A first glance at Table 5.5 reveals that the p-values for the χ^2 and Fisher's tests are very similar for every association. The highest χ^2 score obtained is for the contingency table associating coauthorship and acquaintanceship networks (301.45), indicating a strong correlation between community membership in these two networks. Moreover, the p-value for this test is the lowest recorded (well below 0.001) indicating that the results are statistically significant. This

Contingency table	χ^2	p-value	Fisher p-value
Communication-Coauthorship	46.66	4.54×10^{-7}	1.14×10^{-7}
Communication-Acquaintanceship	20.24	1.6×10^{-2}	0.4×10^{-2}
Coauthorship-Acquaintanceship	301.45	2.2×10^{-16}	2.2×10^{-16}

Pearson's χ^2 and Fisher's exact tests on the contingency tables associating community membership in the coauthorship, communication and acquaintanceship networks (Tables 5.2, 5.3, and 5.4). All tests performed on statistically reduced tables.

Table 5.5: Independence tests between collaboration networks.

result means that scholarly and social circles overlap very well. In other words, coauthors of scholarly papers are very likely to know each other, as they are part of the same communities of collaboration.

The other two tests performed result in lower χ^2 scores and higher p-values. The second highest score (46.66) was obtained for the contingency table associating communication and coauthorship networks. The p-values recorded both for the χ^2 and Fisher tests are again very low. This indicates that communities of usual online contacts on CENS mailing lists overlap fairly well with communities of coauthors. In other words, individuals that communicate frequently on mailing lists also write papers together.

Finally, the lowest recorded χ^2 score (20.24) is obtained for the contingency table associating communication and acquaintanceship networks. Recorded p-values for χ^2 and Fisher tests are well above the other ones, but are are however within significance levels (below 0.05). This result indicates that communities in the communication and acquaintanceship network do overlap slightly but not as much as the overlap detected in the other associations. In other words, it is less probable that frequent contacts on mailing lists actually know each other.

5.2.3 Community structure and socio-academic configuration

The comparative analysis of the previous section is external — it measures the overlap between structural communities detected in different networks. If the socio-academic configuration of a network is known, a network can also be analyzed internally — by comparing its community structure to the social and academic arrangement of its constituents. In other words, I look at how scholarly, communication, and social circles are variegated in their academic configurations.

In the previous chapter, in § 4.4, I introduce the socio-academic configuration of the population under study: how individuals are distributed across different academic affiliations, academic departments, academic positions, and countries of origin. Similar to the way that membership to a structural community is included in a node metadata (Figure 5.1), socio-academic information also can be included as metadata information for every node. For example, the node that represents myself (`id = a.pepe`) in the CENS collaboration networks can be associated with information about my affiliation, department, position, and country of origin, in addition to the previously recorded community membership information, as shown in Figure 5.7.

Figure 5.7: Socio-academic information as node metadata.

From a network perspective, these node-based metadata can be used to group

together individuals that belong to the same institution, department, etc. into a community, i.e., similar to the structural community presented earlier in this chapter. This type of community, however, is not structural, i.e., it is not made salient by the network topology; rather, it is based on socio-academic information that is external to the structure of the network. Thus, for example, if I am a frequent collaborator with another student in the Department of Information Studies, we are part of the same scholarly collaboration circle, i.e. of the same structural community in the coauthorship network. If I am not a frequent collaborator with her/him, we will probably not be part of the same structural community, but we will however be part of the same socio-academic community, for we both belong to the same department.

This enables an interesting comparison between the repartition of the network into *structural* communities, based on topological properties of the network (e.g., the extent of scholarly collaboration), and *socio-academic* communities, made salient by selected social and academic properties of the individuals in the network (e.g., the academic department to which one is affiliated). Thus, the node representing me in the network (`id = a.pepe`) belongs to three different structural communities, based on my coauthorship, communication, and acquaintanceship patterns. I also belong to four different socio-academic communities, based on my socio-academic profile of Figure 5.7. Thus, for example, since I am part of the Department of Information Studies, I belong to a socio-academic community that groups together all Information Studies scholars in the network.

Similarly to previous analyses, the comparison between structural and socio-academic communities is made by constructing a contingency table. In this case, the columns of the contingency table list structural community membership values (i.e., the scholarly, communication, and social community to which a person

belongs), while rows represent socio-academic community membership (i.e., the institution, department, position, and country of origin of that person). These contingency tables are presented in Tables 5.7, 5.8, and 5.9, for the coauthorship, communication, and acquaintanceship networks, respectively. It is worth noting that these contingency tables are "re-binned", as described in § 5.2.1. This involves collapsing together similar categories under a single denomination. Re-binning similar categories greatly reduces table sparseness and improves the reliability of the statistical independence tests. However, re-binning also leads to information loss, as categories lose granularity. In order to reduce table sparseness, but retain as much information as possible, I group categories together only when (i) they are similar, (iii) re-binning brings a significant reduction of table sparseness, and (iii) studying the interaction among them is not of crucial importance for this study. The categories that I re-bin in this dissertation are included in Table 5.6. For the most part, I re-bin similar academic departments. For example, the categories "Statistics" and "Mathematics" are folded into one: "Stats/Math". The reason for this is that there are only 4 members of the Department of Statistics and 1 from the Department of Mathematics in the population. Statistics and Mathematics are similar enough and the interaction among them (e.g. the scholarly collaborations between statisticians and mathematicians) are not of fundamental importance for this study. By grouping together these categories, the cells in the contingency tables become more populated allowing reliable statistical tests. The most substantial re-binning I perform in Table 5.6 is that of "Electrical Engineering" and "Computer Science" into "EECS". The reason for bundling these crucial categories together can be found in the fuzziness by which these departments are separated from one another in many institutions. In the case of CENS, in particular, belonging to one or another department is just nominal, as most scholars perform work that clearly

bridges the Computer Sciences with Electrical Engineering. Please note that I could have performed much more extensive re-binning. For example, I could have grouped together related institutions (e.g., all University of California campuses) and country of origins by their geographical location (e.g., all countries of the North American continent). I explicitly choose not to re-bin these categories to avoid loss of information.

New category	**Old categories**
Stats/Math	Statistics Mathematics
EECS	Electrical Engineering Computer Science
Engineering	Mechanical Engineering Chemical Engineering
Environment	Environmental Science Ecology Earth & Space Botany Meteorology Geology
Civil Engineering	Civil & Environmental Engineering Urban Planning
Faculty	Assistant Professor Associate Professor Professor Lecturer

Table 5.6: Collapsed categories (academic departments and positions).

Each one of these tables (Tables 5.7, 5.8, and 5.9) consists of four contingency tables that indicate the association between a given structural community membership and socio-academic community membership. For example, the top contingency table of Table 5.7 presents the association between coauthorship community membership and academic affiliation membership. In other words, the columns in this table refer to the 14 different communities of coauthors; each

one of these communities is decomposed according the affiliation of its members (rows). The description text shown below Table 5.7 provides guidance on how to read this contingency table, as well as Tables 5.8 and 5.9.

Starting with the contingency tables displaying the association between coauthorship community membership and socio-academic community membership (Table 5.7), it is evident that the coauthorship network has two large, distinct structural communities (#1, and #2) that include almost exclusively researchers from either USC or UCLA. Remarkably, these two communities are fairly homogeneous along the other socio-academic components (they are composed mostly of graduate students, and equal parts of faculty and staff, from the departments of Computer Science and Electrical Engineering, and largely from the United States, and India). This indicates that there are two major scholarly communities of collaboration almost identical in size, domain, position, and origin distribution, but they are based at different universities. Community #3 is also interesting for it is evenly split between UCLA and USC, but is then homogeneous along the other components. Thus, while communities #1 and #2 are centered exclusively around USC and UCLA, respectively, community #3 is a community of multi-institutional collaboration among the two universities. All these three communities are largely composed of members of CS and EE departments. Among the other communities, community #4 and #5 are interesting because they are mostly composed of biologists and environmental scientists from UCLA and UC Riverside. Also, in these communities faculty and staff are more prevalent than graduate students, and there are only U.S. domestic collaborations.

A look at Table 5.7, which portrays the association between communication and socio-academic community membership, reveals a different scenario. The most populated community (#1) is entirely made up of UCLA individuals, but it

	Coauthorship network														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	**T**
Affiliation															
UCLA	8	55	20	6	11	18	9	8	3	7	8			6	159
USC	42	2	23							2		1	2		72
UC Riverside				8	5										13
Caltech	1		1				4					7			13
MIT	2		3							4					9
UC Berkeley		1	4					3							8
UC Merced			1						6						7
UIUC		1						1	1	1					4
SUNYSB	3							1							4
Department															
EECS	55	51	50	1	7	16	6	14	1	4	11	7		2	225
Environment	2			8	4	3	15		5	1					38
Civil Eng		3	3						5	11			6		28
Biology	9			8	12							1			30
Engineering	4							1	5			1			11
Info Stu		2	4	7											13
Education				1				3							4
Marine	6														6
Stats/Math		4	1												5
Media/arts		5													5
Position															
Graduate	26	33	21	4	4	8	5	6	1	4	6	5	1	1	125
Faculty	22	12	16	17	8	7	9	9	7	7	4	2	3	1	124
Staff	20	16	14	3	9	6	7	4	1	4	1	1	1		87
Postdoc	6	2	6	1		1			4	1		1	1		23
Undergrad	1	3	1		1				1						7
Origin															
USA	40	32	29	18	21	12	10	12	13	12	4	4	5	1	213
India	15	15	10					2		1	1			1	45
China	1	3	7			1	1	1		2	1	4			21
Korea	3	4	3				1								11
Italy	1			1		3		1			4				10
Australia	4			1			1								6
Iran		1				1		2		1					5

Columns present community membership values in the coauthorship network and rows present selected socio-academic values. Each column indicates how the members of the same coauthorship community are distributed across academic affiliation, department, position and country of origin. For example, community #1 is heavily populated with scholars affiliated with USC (42 individuals), while community #2 is largely composed by scholars affiliated with UCLA (55 individuals). Both communities are composed mostly of scholars in the departments of Computer Science and Electrical Engineering. Also, the distribution of academic positions and country of origin is very similar in both communities, with a large presence of graduate students, faculty, and researchers, from the U.S.A. and India.

Table 5.7: Contingency tables: coauthorship vs. socio-academic community membership.

	Communication network							
	1	2	3	4	5	6	7	**T**
Affiliation								
UCLA	19	11	14	8	7	2	1	62
USC		3		1	1	6		11
UC Riverside		2		1				3
Caltech			1				1	2
UC Merced		1		1				2
Department								
EECS	9	10	8	11	9	8	2	57
Environment		1	5					6
Civil Eng	2		3					5
Biology		4		1				5
Stats/Math	4							4
Media/arts	3							3
Info Stu	1	1						2
Position								
Graduate	9	4	5	4	6	4		32
Staff	6	5	8	5	2	1		27
Faculty	4	8	3	1	1	3	2	22
Postdoc				2				2
Origin								
US	11	14	12	10	5	4	1	57
India	5	3	2		1	2		13
Korea	1					1		2

Table 5.8: Contingency tables: communication vs. socio-academic community membership.

is very diversified by department, position and origin. This indicates that members of this community are involved in discussions about large, multi-disciplinary, UCLA-based projects. All the other communities are all prevalently composed of UCLA affiliates of CS and EE, but are overall more diversified. In particular, community #2 and #3 involve discussions by a discrete amount of biologists and environmental scientists, respectively.

Finally, Table 5.9 portrays the association between acquaintanceship and socio-academic community membership, revealing yet another scenario. Communities of acquaintances are very much dependent on academic affiliation: acquaintanceship community #1 is composed of USC affiliates, communities #2, #3, and #4 have mostly UCLA scholars, community #5 has affiliates of UC

	Acquaintanceship network								
	1	2	3	4	5	6	7	8	**T**
Affiliation									
UCLA	6	61	46	27	1	1	18	9	169
USC	58	1	1	1		12			73
UC Riverside					13				13
Caltech	2	4				7			13
MIT	4			5			1		10
UC Berkeley	4		3				1		8
UC Merced	1			6					7
UIUC	2		1	1					4
SUNYSB			1			3			4
CMU	3				1				4
Department									
EECS	93	32	59	9	2	8	18	11	232
Environment		16		6	12	1	3		38
Civil Eng	3	3		25					31
Biology		3			16	10	1		30
Engineering	2	1	2	6		3			14
Info Stu	2	7	1	2			1		13
Education		7							7
Marine						6			6
Stats/Math	1	4							5
Media/arts		5							5
Position									
Graduate	40	27	28	8	3	7	10	7	130
Faculty	29	20	16	20	20	13	9	3	130
Staff	25	26	13	9	6	4	5	1	89
Postdoc	6	2	5	6		4	2		25
Undergrad	1	4		3					8
Masters	1			2	1				4
Origin									
USA	37	56	25	40	26	18	15	5	222
India	25	3	18	2					48
China	8	2	2			6	2	1	21
Korea	6	1	3					1	11
Italy	2	1	1				3	4	11
Australia	4	2							6
Iran	2		2				1		5
Mexico					2		2		4

Table 5.9: Contingency tables: acquaintanceship vs. socio-academic community membership.

Riverside, and so on. The other socio-academic components, however, display a high degree of diversification. Except for communities #1 and #3 (that are made up of CS and EE scholars), all the other communities are multi-disciplinary in

their population.

While this analysis provides a sense of the overlap between CENS structural and socio-academic communities, the presented findings need to be validated by a statistical comparison. Table 5.10 summarizes the results of a Pearson's χ^2 test of independence on these data. These results can be interpreted as follows.

Contingency table	Affiliation	Department	Position	Origin
Coauthorship	2.2×10^{-16}	2.2×10^{-16}	0.024	0.0023
Communication	0.070	0.081	0.20	0.23
Acquaintanceship	2.2×10^{-16}	2.2×10^{-16}	7.3×10^{-4}	6.22×10^{-8}

Results of Pearson's χ^2 test of independence (p-values) on contingency tables that display association between structural and socio-academic community membership. All tests performed on statistically reduced tables.

Table 5.10: Independence tests: collaboration networks and socio-academic profile.

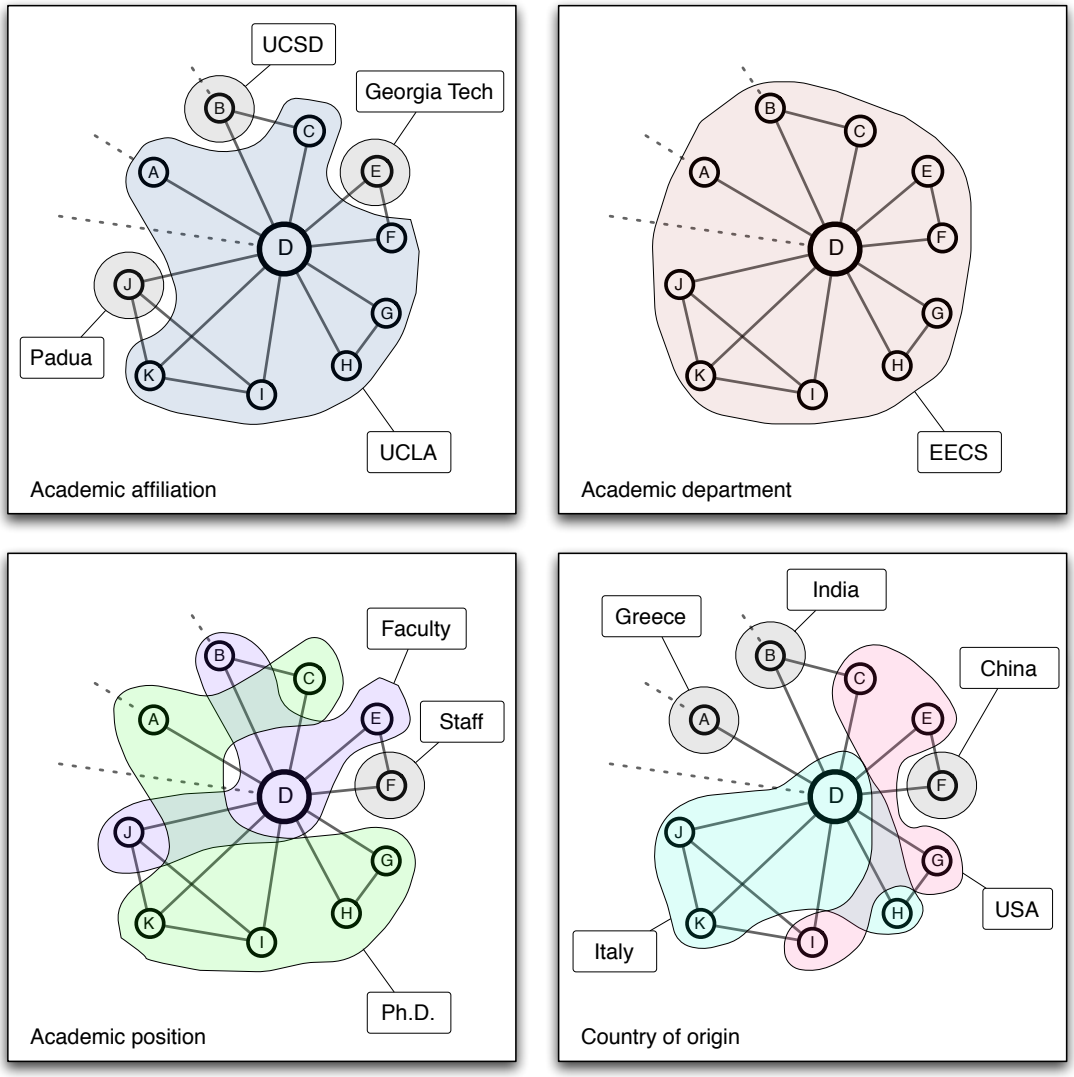
Starting with the coauthorship network (first row), it can be noted that two very different groups of p-value scores are obtained. These value show that, on one side of the spectrum, scholars' membership in a coauthorship community is dependent on scholars' department and affiliation (first two columns, very low p-values). On the other side of the spectrum, community membership is much less dependent on academic position and country of origin (last two columns, higher p-values). Both affiliation and department have extremely low p-values: communities of coauthors overlap very well with communities of scholars of the same institution and department. This means that tight-knit communities of coauthors are normally made up of scholars that belong to the same institution and department. In turn, this result indicates that coauthorship patterns at CENS do not involve significant inter-institutional and inter-departmental efforts. Academic position and country of origin have much higher p-values (0.024 and 0.0023, respectively). Although both these value are statistically significant

(below 0.05), they indicate that academic position and country of origin are much less dependent of coauthorship community membership, i.e., scholars at all levels and positions (Ph.D. students, professors, etc.) and of different nationalities coauthor papers together.

The communication network (second row in Table 5.10) presents a much more variegated scenario. All p-values obtained via the Pearson's χ^2 test on the communication network indicate independence between membership in structural and socio-academic community. In other words, these values show that individuals participate in discussions on CENS mailing lists regardless of their affiliations, departments, positions and country of origin. Communities of discussants are very variegated in their socio-academic composition.

The acquaintanceship network presents the least variegated scenario of the three. All recorded p-values are well below significance level (0.05). The p-values recorded for academic affiliation and department are the lowest, indicating that scholars really know mostly people in their own institution and in their own department. Academic position and country of origin are less dependent, but are also low, indicating that acquaintances are made within one's academic position (e.g., Ph.D. students are acquainted with other Ph.D. students) and country of origin (e.g., Italian scholars are acquainted with other Italians).

Let us discuss this finding in terms of the same anecdotal example presented above (the Computer Vision Lab at UCLA). Figure 5.8 depicts coauthorship community #11, annotated according to the socio-academic characteristics of its constituents. In other words, Figure 5.8 is a pictorial representation of column #11 in Table 5.7. The image in the top left corner displays the institutional composition of this community. Four different academic affiliation are present, although the community is largely made up by UCLA scholars. The departmen-



Coauthorship community #11 ($n = 11, m = 29$) annotated according to the socio-academic characteristics of its constituent nodes.

Figure 5.8: Anecdotal example: overlap of coauthorship community and socio-academic profile.

tal decomposition, in the top right corner, shows that all the individuals come from departments of Electrical Engineering and Computer Science (EECS). In the bottom left corner, academic position of this community is displayed. The community is almost evenly split between graduate students and faculty. Fi-

nally, in the bottom right corner, the decomposition by country of origin shows a very international environment with five different countries (most represented countries: United States and Italy).

This case study—although limited to a single community of coauthorship—is representative of the population at large. It shows that scholarly communities internally exhibit little (or no) disciplinary and institutional variation. When zooming into tight-knit circles of scholarly collaboration, a local ecology emerges, homogeneous in terms of academic affiliation and department. The same can be said of acquaintanceship communities: circles of acquaintances tend to be mono-institutional and mono-disciplinary. This finding does not indicate that CENS collaboration networks, at large, are mono-institutional and mono-disciplinary. As noted throughout this dissertation, CENS is populated by a very variegated array of researchers from different institutions and departments. What this finding indicates is that the mélange between disciplines and institutions does not happen at a community level, but via the bridging action of community hubs (e.g., node D in Figure 5.8).

5.3 Summary

This chapter describes the findings of a structural analysis of CENS collaboration networks. I find that the coauthorship, communication, and acquaintanceship networks are composed of 14, 7, and 8 structural communities, respectively. A comparative analysis of these structures reveals the following scenario: communities of coauthoring researchers at CENS overlap very well with communities of acquaintanceship, and relatively well with communities of discussants on electronic platforms; these communities of discussants, however, overlap only slightly

with communities of acquaintances. These results substantiate the theory that scientists that write papers together know and communicate with each other, but communication alone does not necessarily indicate acquaintanceship. As such, these results present a fairly homogeneous scenario of collaboration in which researchers are divided in fairly separated communities of collaboration, the members of which write papers together, communicate on mailing lists and know each other.

This research finding is further supplemented by a comparative analysis of community structure and socio-academic configuration. This analysis reveals how topological structures relate to the organizational, disciplinary, institutional and international arrangements of collaborations at CENS. Findings show that structural communities in the coauthorship network overlap very well with academic affiliation and department, i.e., communities of scholarly collaborators tend to be populated with individuals working in the same institution and domain. Comparison of communities of online discussants and their socio-academic configuration shows that there is no significant dependence among them, i.e., individuals connect online regardless of their affiliations, departments, positions and country of origin. Finally, the comparative analysis of acquaintanceship shows a high level dependence between communities made up of individuals that all know each other and socio-academic configuration. This indicates that, even though the acquaintanceship network is dense and sparse ("everyone knows everyone"), social circles tend to be populated, at large, with people of the same institution, department, position, and country of origin.

CHAPTER 6

Results: Evolutionary analysis

In the previous chapter, I report a comparative structural analysis of three different networks of scientific collaboration, looking at how researchers organize themselves in communities of scholarly, communicative, and social interaction. In this chapter, I shift my analysis to the evolution of these networks, in order to address the second research question, restated here:

> **Research Question #2.** What collaboration dynamics can be evinced from the coauthorship, communication, and acquaintanceship networks of CENS? Can specific evolutionary features be explained in terms of changes in the disciplinary and institutional arrangements of collaboration?

6.1 Slicing the networks by their temporal component

The structural analysis of the CENS collaboration networks, presented in the previous chapter, is based on the most recently available data (data collection was concluded in October 2009). Thus, the networks analyzed thus far represent the cumulative volume of scholarly coauthorship, electronic communication, and acquaintanceship patterns. In order to allow a historical or longitudinal network

analysis, it is necessary to "slice" these cumulative networks along their temporal component. As discussed in previous chapters, all data studied in this dissertation contain some temporal information: the bibliographic record contains the date of publication of papers; the communication logs contain the date at which a discussion in a mailing list thread took place; and with the administered social network survey, I gathered information about the length of acquaintanceship patterns. There are, however, several different ways in which this temporal information can be utilized. I use different techniques for each different network, in order to reflect the different aspects of collaboration that these networks represent.

Slicing the coauthorship network. The publication date of papers in the CENS bibliographic record represent the date at which an article is published in a journal or in conference proceedings. Many scientific collaboration networks are constructed cumulatively, i.e., they sum the coauthorship contributions for every year, based on the assumption that a coauthorship relationship, once established, can increase in intensity but cannot decay. This assumption, however, does not genuinely reflect the writing lifespan of an article. In fact, due to the lengthy processes of scholarly peer-reviewing, revision, proof-reading, and editing, the publication date of an article very rarely represent the year in which it was written. Of course, the timescales of these processes vary depending on a number of factors: the promptness of peer-reviewers and editors to return reviews, the efficiency of coauthors to produce revisions, the nature of publication (journal article, conference proceedings, book chapter, etc.) and the scholarly domain in which an article is published (some domains have faster publication workflows than others). Some recent research has addressed this issue by proposing a network model in which ties have a lifespan, i.e., they decay with time [169, 170]. For the purpose of this research, I employ a similar decaying model. I assume that the body of literature studied here is subject to the following publication

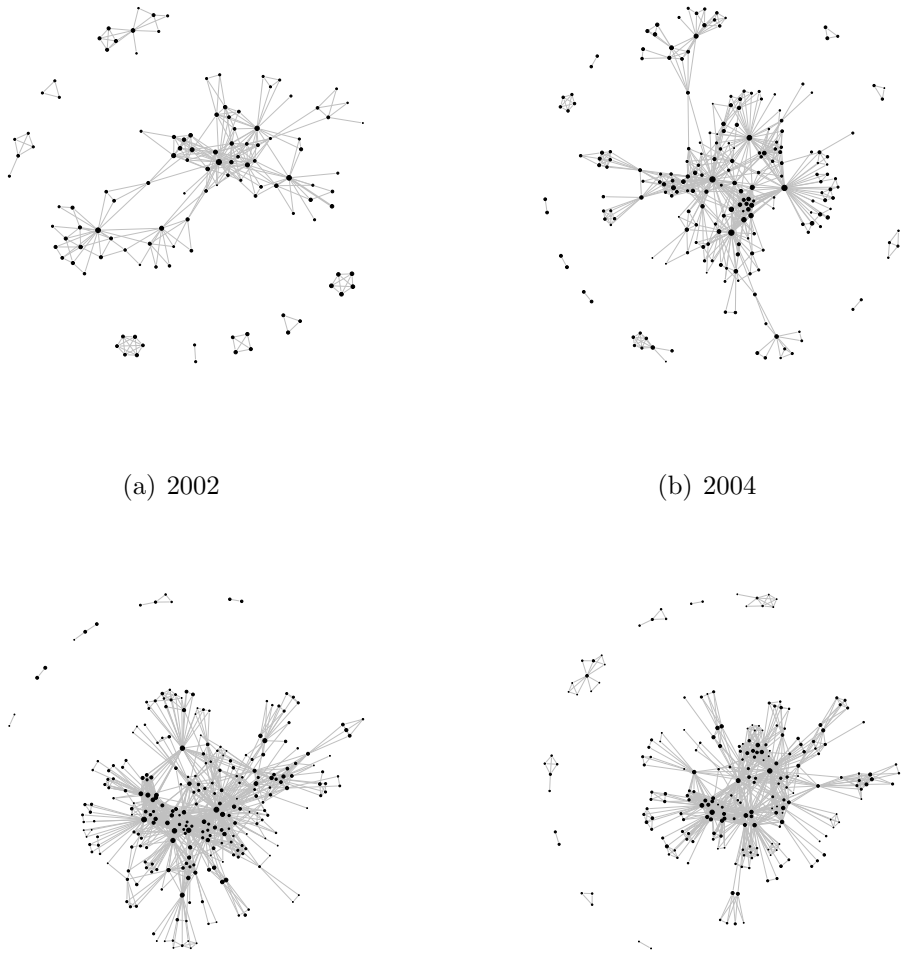
timeline: most of the collaborative activity to produce a paper takes place the year prior to the publication date; the year in which a paper is published, collaboration is less prominent, but a fair amount of research and collaborative activities take place (reviews, edits, revisions); the year after a paper is published research and editorial activities are at a minimum, yet coauthors may still collaborate on lateral related activities (disseminating the paper, making related material publicly available, etc.). In other words, I posit that publications have a decaying lifespan of 3 years. Assuming this publication timeline, I can slice the coauthorship network, as follows. The year prior to a publication, coauthorship activities are computed in full, i.e., I use Formula 4.1 to calculate edge weights. The year in which a paper is published, I reduce the original edge weights by half. In the year following publication, I further reduce edge weights by a factor of 2. For example, an article authored by Alberto Pepe and Marko Rodriguez published in 2007 would appear as an edge between the two in the coauthorship networks of years 2006, 2007, and 2008, with weights 1.0, 0.5, and 0.25, respectively (cumulative, decaying weight over a 3-year period). Four network snapshots (years 2002, 2004, 2006, and 2008), depicting the evolution of coauthorship activity at CENS in these years, are included in Figure 6.1 (a through d).

Slicing the communication network. The communication network is based on mailing list logs that are time-stamped. Thus, every discussion activity on a thread has a specific date and the communication network can be sliced accordingly. Whereas the authoring of a paper is a collaborative process that potentially extends over a period of several years, electronic communication on mailing lists is predicated upon narrow windows of time: an analysis of CENS mailing lists reveals that threads are only active for short periods, never exceeding one or two weeks in time. For this reason, the amount of communication among two given discussants (i.e., edge weights) can be computed for a given year by

summing individual interactions (calculated using Formula 4.2). Communication activity is not considered cumulative over the years. The evolution of communication activity at CENS for the available timeframe (years 2005 through 2008) is depicted in the networks included in Figure 6.2 (a through d).

Slicing the acquaintanceship network. In the social network survey administered to gather information about personal knowledge relationships, respondents are asked to provide temporal information regarding every individual they indicate as an acquaintance (see Figure A.2, in Appendix A). More specifically, respondents indicate the date (year) when they first met an acquaintance. It is fair to assume, in this context, that respondents that provide this information are still acquainted (to date) with the individuals they indicate. For this reason, acquaintanceship is sliced by year of first acquaintance, and cumulatively. For example, if Marko indicates in the survey to have known me (Alberto) since 2006, and that we communicate with weekly frequency, then Marko and I would be connected by an edge of weight 1.0 (maximum weight) in year 2006, and subsequent years (i.e., 2007, 2008, and 2009). As such, the acquaintanceship network is cumulative, and non-decaying. Four network snapshots (years 2003, 2005, 2007, and 2009), depicting the evolution of acquaintanceship at CENS in these years, are included in Figure 6.3 (a through d).

All the evolutionary analyses presented in this chapter are based upon the networks sliced according to the criteria presented above.



(a) 2002

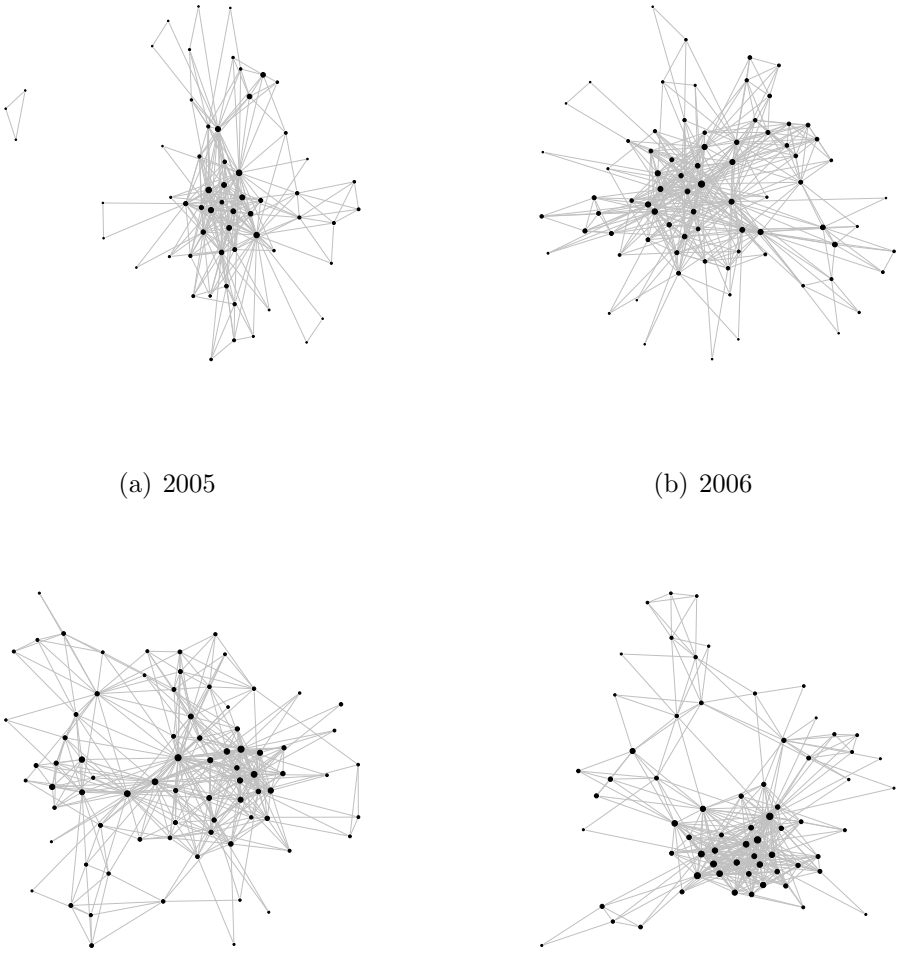
(b) 2004

(c) 2006

(d) 2008

The coauthorship networks in 2002, 2004, 2006, and 2008, diagrammed according to the Fruchterman-Reingold network layout algorithm [160]. Line width is proportional to edge weight, where more intense collaborations have wider and more marked lines; the diameter of the nodes is proportional to the weighted centrality score on a logarithmic scale, or strength [161], where more central nodes have larger diameters.

Figure 6.1: Evolution of the coauthorship network.



(a) 2005

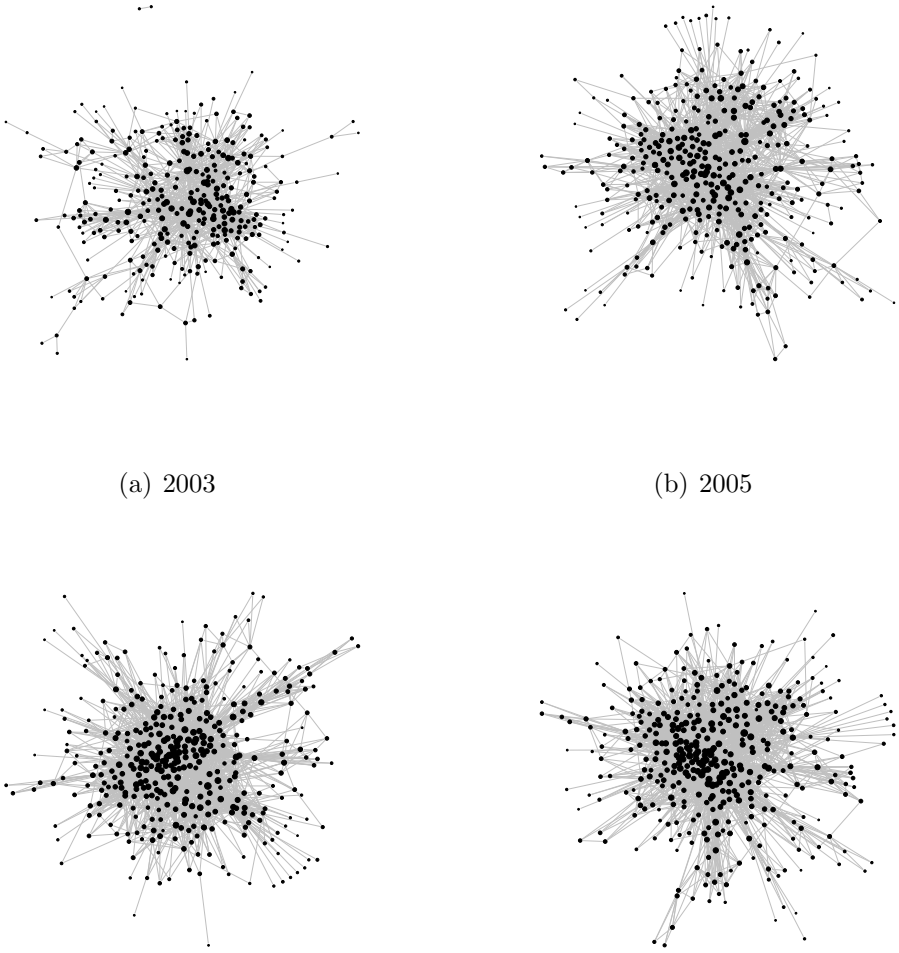
(b) 2006

(c) 2007

(d) 2008

The communication networks in 2005, 2006, 2007, and 2008, diagrammed according to the Fruchterman-Reingold network layout algorithm [160]. Line width is proportional to edge weight, where more intense collaborations have wider and more marked lines; the diameter of the nodes is proportional to the weighted centrality score on a logarithmic scale, or strength [161], where more central nodes have larger diameters.

Figure 6.2: Evolution of the communication network.



(a) 2003

(b) 2005

(c) 2007

(d) 2009

The acquaintanceship networks in 2005, 2006, 2007, and 2008, diagrammed according to the Fruchterman-Reingold network layout algorithm [160]. Line width is proportional to edge weight, where more intense collaborations have wider and more marked lines; the diameter of the nodes is proportional to the weighted centrality score on a logarithmic scale, or strength [161], where more central nodes have larger diameters.

Figure 6.3: Evolution of the acquaintanceship network.

6.2 Evolution of network topology

In Chapter 4, I compute some fundamental statistics relative to the topology of the networks of CENS collaboration. In this section, I examine how these topological properties evolved over time. I present in Table 6.1, the following network statistics: number of nodes, edges, number of connected components, diameter of the largest connected component, average path length, clustering coefficient and degree assortativity. These measures are computed on each collaboration network sliced temporally according to the mechanisms discussed above. An analysis of the statistics presented in Table 6.1 provides insights into the evolution of the CENS networks over time.

Coauthorship network. The first three rows of Table 6.1 present the number of publications, nodes (authors), and edges (collaborations) in the coauthorship network sliced along its temporal component. When analyzed over time, it can be seen that CENS scholarly collaboration is at its most active period during years 2004 through 2008. Although the number of publications peaks in 2004, the number of authors and the number of collaborations reach at maximum in 2006. This result alone suggests that the number of authors per paper has increased over time. Also, looking at the overall range of the network (number of nodes and edges), two distinct time periods can be discerned: a first term (2001-2005) during which number of nodes and edges increase, and a second term (2006-2009), during which the growth slightly slows down. In particular, the author count values indicate that CENS quickly became large and diversified in its population in the first term reaching a solid population base of collaborators by the year 2004. In the second term, the number of published works and collaborations maintains a regular growth (with nearly 1000 collaborations in 2008), but the author base shrinks to a solid core of about 200 individuals.

	quantity/year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Coauthorship	Publications		6	20	41	94	116	105	71	64	68	23	
	Nodes	10	33	66	117	195	216	248	257	223	163	65	
	Edges	16	86	168	308	561	708	868	1088	1053	921	597	164
	Components	2	4	5	9	14	11	7	6	6	9	8	8
	Diameter	2	3	4	6	7	7	7	7	6	6	6	3
	Av. path length, ℓ	1.448	1.543	2.231	2.915	2.847	3.213	3.171	2.906	3.022	2.813	2.967	1.903
	Clust. coefficient, C	0.720	0.648	0.547	0.447	0.395	0.392	0.356	0.357	0.389	0.436	0.534	0.762
	Degree assortativity, r	-0.045	0.217	0.055	0.035	-0.043	-0.050	-0.068	-0.084	-0.052	0.013	0.101	0.240
Communication	Threads							255	468	339	392		
	Nodes							60	77	77	67		
	Edges							277	424	406	384		
	Components							2	1	1	1		
	Diameter							4	4	4	6		
	Av. path length, ℓ							2.115	2.149	2.235	2.276		
	Clust. coefficient, C							0.519	0.443	0.451	0.560		
	Degree assortativity, r							-0.106	-0.138	-0.067	0.046		
Acquaintance	Nodes			270	294	318	339	360	371	377	381	381	
	Edges			785	1041	1370	1746	2246	2679	3058	3208	3261	
	Components			9	8	2	1	1	1	1	1	1	
	Diameter			9	9	9	6	6	6	6	6	6	
	Av. path length, ℓ			3.506	3.319	3.195	2.982	2.815	2.746	2.678	2.662	2.655	
	Clust. coefficient, C			0.201	0.213	0.216	0.232	0.243	0.258	0.276	0.281	0.283	
	Degree assortativity, r			-0.171	-0.132	-0.118	-0.099	-0.086	-0.077	-0.067	-0.061	-0.060	

Number of nodes and edges, number of connected components, diameter of the largest connected component, average path length, clustering coefficient and degree assortativity. Blank cells indicate unavailable data.

Table 6.1: Fundamental network statistics of the collaboration networks, 1999-2010

This finding is confirmed by an analysis of the network's configuration (number of components and diameter). The number of connected components grows from 2, in 1999, to 14, in 2003, indicating that the network becomes more fragmented in the first term, even if collaboration is overall increasing. In the second term, however, the number of connected components begins to drop and the network quickly solidifies into a giant component, which indicates a solid base of strong collaboration. By looking at the the network diameter, the formation of the giant component, starting in 2004, is evident. This is further reinforced by a quick analysis of Table 6.2, which lists component populations by year, and shows that the number of components drops between year 2003 and 2005 resulting in a giant component of over 200 nodes.

year	#	population
1999	2	2 8
2000	4	18 6 4 5
2001	5	36 8 16 4 2
2002	9	79 10 6 5 5 4 3 3 2
2003	14	131 10 8 8 7 7 5 4 2 3 3 2 3 2
2004	11	183 8 5 4 3 3 2 2 2 2 2
2005	7	230 5 4 3 2 2 2
2006	6	255 4 3 2 2 2
2007	6	241 5 4 3 2 2
2008	9	188 10 7 5 4 3 2 2 2
2009	9	137 7 4 4 3 3 3 2

Table 6.2: Components of the coauthorship network (by year).

The last three sets of values presented in Table 6.1 (average path length, clustering coefficient, and degree assortativity) can be investigated to provide an understanding of network topology. The average path length, ℓ, is about 1.5 in 1999; it grows steadily in the first term, reaching a value of about 3.2 in 2004, which stays roughly constant (or slightly shrinks) throughout the second term. This indicates that once a CENS authoring base is formed, an average of 3 steps are necessary to transfer information among any two pairs of nodes.

The clustering coefficient, C, decreases steadily over time. It halves in the first period, from an initial value of about 0.7 in 1999, to about 0.35 in 2005, and then remains constant. This suggests that the network becomes less cliquish and collaboration patterns becomes more uniform across the network over time.

A final indicator of network topology presented here is degree assortativity. As explained in § 3.1.3, assortativity can be defined as the tendency for individuals (nodes) in a social network to establish connections preferentially to other individuals with similar characteristics. The measure of assortativity presented here is computed based on the individuals' degree centrality, which indicates the tendency for CENS researchers to write papers with others with a similar number of collaborators. In other words, a high degree assortativity means that very productive authors collaborate with other very productive authors, while low-degree authors (i.e., authors that do not collaborate very much) collaborate with other low-degree authors. The degree assortativity coefficient, r, calculated using Formula 3.2 returns a value in the range $-1 \leq r \leq 1$, where $r = 1$ indicates perfect assortativity, $r = 0$ indicates no assortativity, and $r = -1$ indicates perfect disassortativity. In the coauthorship network, degree assortativity has a declining trend: in year 2001, the network is slightly assortative ($r = 0.217$), but very soon the coefficient drops to zero or just below zero, indicating no significant level of degree assortativity or disassortativity. In other words, individuals collaborate (with no preference) with other individuals of any degree.

It is interesting to note that, the decline of the degree assortativity measure follows very closely that of the clustering coefficient — a Spearman correlation between the two gives $\rho = 0.975$ (p-value < 0.005). This means that as collaboration patterns in the network become more sparse and uniform (decreasing C), they also become more mixed (decreasing r), i.e., highly-connected individuals

begin to collaborate with lowly-connected ones.

Communication network. The mailing list logs upon which the communication network is based contain data for the years 2005 to 2009. Moreover, 2005 and 2009 data are partly incomplete. In particular, 2009 data are only limited to the month of January and February, and they were folded into the logs of year 2008, for the purpose of the evolutionary analysis. Overall, the communication network does not present a variation comparable to that of the coauthorship network. From 2005 to 2008, the number of mailing list threads varies only slightly. Every year, roughly 400 threads contain conversational communication by more than three individuals. The total number of individuals involved in CENS mailing list discussions is also roughly the same every year, with about 70 discussants per year. With the exception of year 2005 (that has 2 components with populations 57 and 3), all the individuals that are part of the communication network belong to a single connected component through time. The diameter is also stable, it takes 4 to 6 steps to connect the most remote nodes in the communication network. Similarly, the average path length is almost unchanged: an average number of roughly 2 steps are needed to connect any two discussants in the network at any point in time.

More interesting are the results obtained for the clustering and degree assortativity coefficients, C and r. Although these coefficients only increase slightly, they follow roughly the same pattern of growth of the respective measures in the coauthorship network. A Spearman correlation between the two gives $\rho = 0.8$ (p-value < 0.05). Moreover, the range of change of C and r for the communication and coauthorship networks are comparable. This relationship suggests that the practices by which individuals connect to others when authoring papers and communicating online follow similar dynamics. Thus, for example, high-

degree individuals attach preferentially both to other high-degree authors and high-degree mailing list discussants.

Acquaintanceship network. The temporal slicing of the acquaintanceship network results in nine networks (2001-2009). Since these networks are constructed from cumulative data, the number of nodes and edges naturally increases over time. Yet, these measures follow different trends of growth. At the outset, in 2001, 270 individuals are already in the network, i.e., they are already acquainted with someone who is part of the acquaintanceship network. In the first term, up to year 2005, the number of nodes increases by about 100 nodes, and it then stays roughly constant up to 2009. The number of edges is fairly low in 2001: only about 800 acquaintanceship relationships predate 2002. This number increases steadily over time: by 2009 the number of personal relationships quadruples. This indicates that the network of acquaintanceship includes many individuals but few connections among them at the outset, but acquaintanceship patterns become more and more dense over time. The analysis of components and diameter reinforces this finding and demonstrates that the network becomes solidly interconnected in year 2004, when all nodes can be found within a single connected component and there are 6 degrees of separation between the most remote nodes (diameter).

The temporal analysis of average path length, clustering coefficient and degree assortativity point to a linear evolution. The average path length decreases steadily from 3.5 in 2001 to 2.6 in 2009: over time, the network becomes more connected and the average number of steps required to travel from any two individuals diminishes. The clustering coefficient increases throughout this period, from 0.2 to 0.28, as the network becomes progressively more clustered. Finally, the network is slightly disassortative in 2001 ($r = 0.17$) but over time assortativ-

ity comes close to zero ($r = -0.06$ in 2009), i.e., at the outset, individuals with similar degree attach preferentially with each other, but later acquaintanceship patterns become more sparse and variegated. Two remarks can be advanced when comparing these measures to those found in the coauthorship network. First, it is interesting to note that even though the coauthorship network is overall more clustered than the acquaintanceship network, its clustering coefficient diminishes over time, i.e., scholarly collaboration patterns become more sparse. Second, both the coauthorship and acquaintanceship network exhibit very poor assortativity, i.e., both scholarly collaboration and personal knowledge are not driven by preferential attachment with individuals of similar degree. These findings are further analyzed in the next section, below.

6.3 The dynamics of preferential attachment

The dynamics by which individuals connect with each other in the networks of coauthorship, communication and acquaintanceship are best understood by an analysis of three degree distributions: the degree frequency distribution, the degree clustering distribution, and the degree assortativity distribution. These distributions are plotted for the coauthorship, communication, and acquaintanceship networks in the block images of Figures 6.4, 6.5, and 6.6 respectively. Each row in these block images represents a different year, i.e., a different evolutionary stage of the network.

Coauthorship network. Figure 6.4 presents degree distributions of the coauthorship network in years 2002, 2004, 2006, and 2008. The first column from the left displays a plot of degree frequency distribution, i.e., it depicts the frequency with which nodes of a certain degree occur. The x-axis presents node

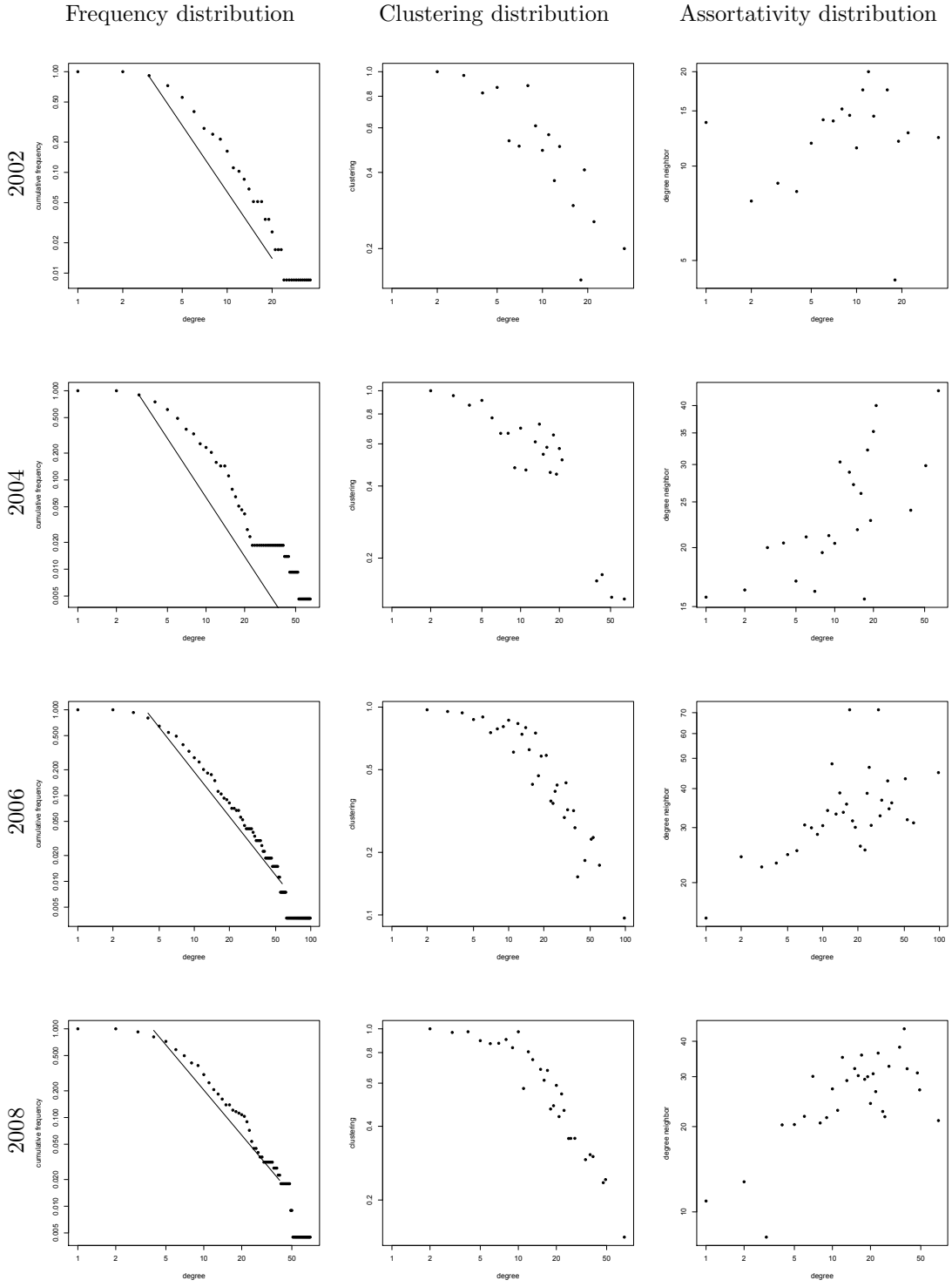


Figure 6.4: Degree distributions of the coauthorship network

degrees and the y-axis their cumulative frequency. Low-degree nodes (i.e., individuals with very few publications) lay on the left portion of these plots while high-degree nodes (i.e., well published authors) lay on the right side of the plot. A visual analysis of these plots over the years reveals that the degree distribution has a power-law tail (a power-law fit line is plotted) and thus the coauthorship network is *scale-free*. It is important to note that the degree distribution evolves into a perfect power law tail with time: in years 2001 and 2004, degree and power law distributions are not perfectly aligned; in later years, they overlap completely. As discussed in Chapter 4, power-law degree distributions and scale-free effects are normally considered a common feature of coauthorship networks. It is interesting to note that the power law fits particularly well the middle part of the degree distribution with a flat "hook" on the top left portion of the plot and a flat "tail" on the bottom right portion of the plots. It has been argued that the hook represents young newcomers in the collaboration network (with very few publications on their record), while the tail represents older, well-established scientists (with many publications) [63]. The coauthorship network analyzed here is not mature enough for this assumption to hold. However, the hook and the tail can be clearly seen in all the time-based degree distributions, suggesting that the CENS collaboration network is subject to a constant influx of newcomers that occupy the top left portion of these plots. Moreover, the flat tail has a visible offset in year 2004. This offset indicates that during this year a significant number of well-established researchers authored more CENS publications than usual. This validates the finding discussed above that the network solidifies into a giant component in years 2004/2005, with well-established researchers becoming major hubs.

The second column in Figure 6.4 shows the degree clustering distribution of the CENS coauthorship network over time. In these diagrams, node degrees are

plotted on the x-axis and their clustering coefficient on the y-axis. Overall, in all time frames analyzed, clustering coefficient nearly follows a power-law distribution with clustering decreasing as the node degree increases. This means that low-degree degree nodes belong to highly clustered portions of the network that are connected to each other via hubs (the high-degree nodes, which sit in less clustered portions of the network). The formation of four main hubs is evident in the clustering degree distribution plot of year 2004, in which a small number of nodes with high-degree and low-clustering is separated by the rest of the degree population (low-degree and high-clustering).

The third column presented in Figure 6.4 shows the degree assortativity distribution of the CENS coauthorship network over time. In these plots, the x-axis is the node degree and the y-axis is the average degree of the neighboring nodes. As such, this plot gives the distribution of how nodes of a given degrees connect to others (i.e., assortativity measure). This distribution, especially in years 2002 and 2004, is not as linear as those observed thus far. Yet in year 2006, a trend emerges in which neighboring degree increases as a function of node degree, i.e., the higher the degree of an author, the higher the degree of a collaborating author. This indicates that very prolific, centrally-located authors tend to collaborate together, while newcomers (low-degree nodes) tend to attach preferentially to higher degree, but marginal authors.

Communication network. Figure 6.5 presents degree distributions of the communication network in years 2005, 2006, 2007, and 2008: degree frequency, degree clustering and degree assortativity distributions. The first column shows the degree frequency distribution of the communication network in years 2005 though 2008. The power law fit line is plotted in each diagram. Even though the communication network is not constructed from cumulative data, degree fre-

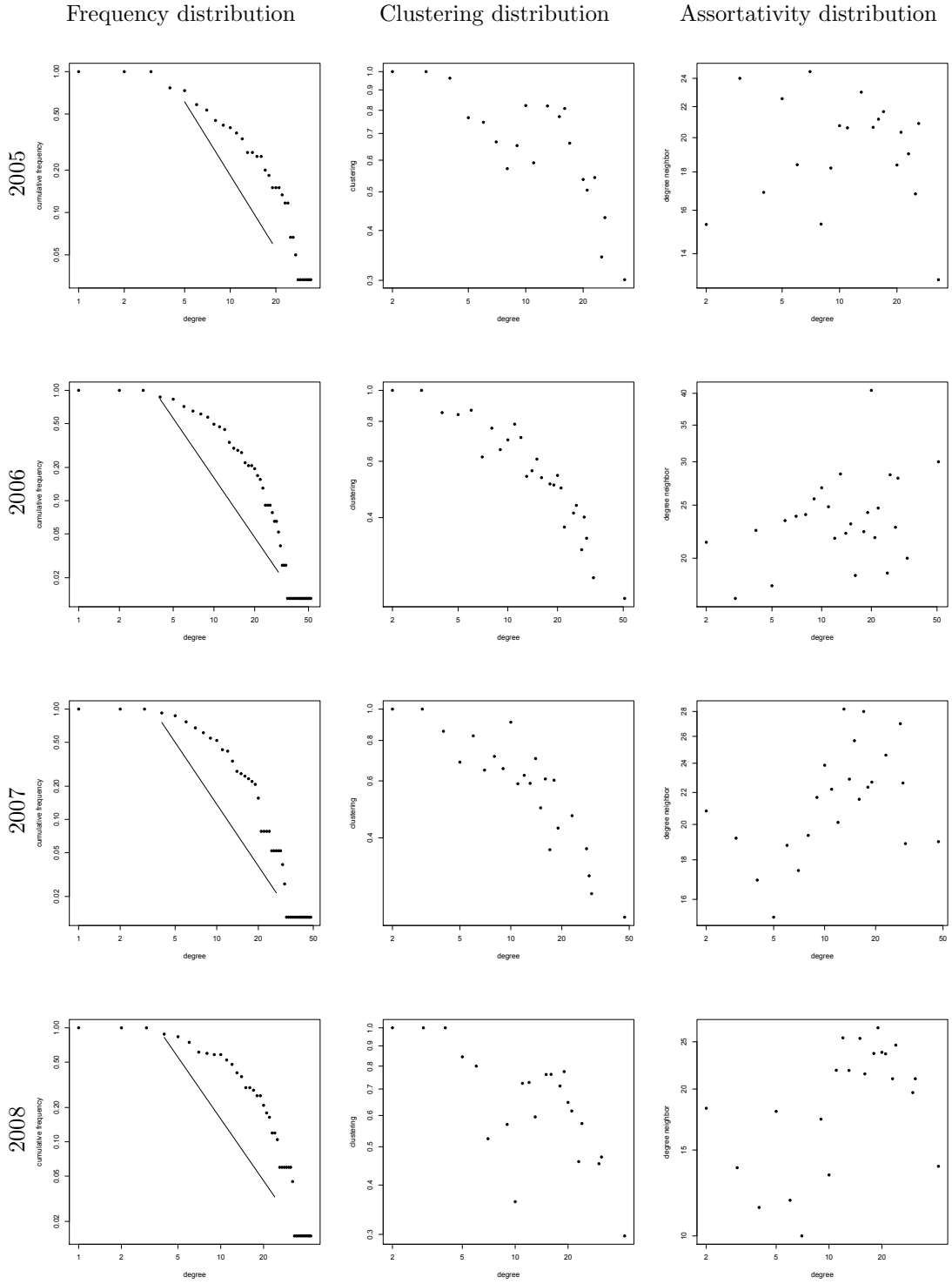


Figure 6.5: Degree distributions of the communication network.

quency has essentially an identical distribution every year. This distribution presents a top left hook and a bottom right tail, similar to those found in the coauthorship networks, however, the middle distribution is skewed to the right. This indicates that there is a much more distinct gap between individuals that communicate on mailing lists frequently and those who do not.

The second column in Figure 6.5 shows the clustering distribution. While the plot for year 2006 has a distribution that seems to follow a power law, i.e., with clustering decreasing as the node degree increases, all the other years have distributions that cannot be easily fit to a linear trend. This suggests that clustering and degree are not significantly related in the communication network, i.e., communication patterns are fairly sparse and clustering is independent of node degree.

The same remark can be advanced for the assortativity degree distribution, shown in the third column of Figure 6.5: there is no visible dependence between node degree and average degree of neighboring nodes. This indicates that there is no specific rule of preferential attachment to govern communication patterns.

Acquaintanceship network. Figure 6.6 presents degree distributions of the acquaintanceship network in years 2003, 2005, 2007, and 2009. The first column depicts the degree frequency distribution. Similarly to the previous two networks, also the acquaintanceship network follows a power-law degree distribution, especially in the first years, up to 2005. From 2005 onwards, the top left hook flattens and the bottom right tail drops, to form a nearly straight vertical line. Only the middle portion of the distribution remains aligned with the power-law line. The flattening of the hook means that the volume of low-degree nodes in the acquaintanceship network increases over time, i.e., every year there are newcomers who progressively become acquainted with the CENS population,

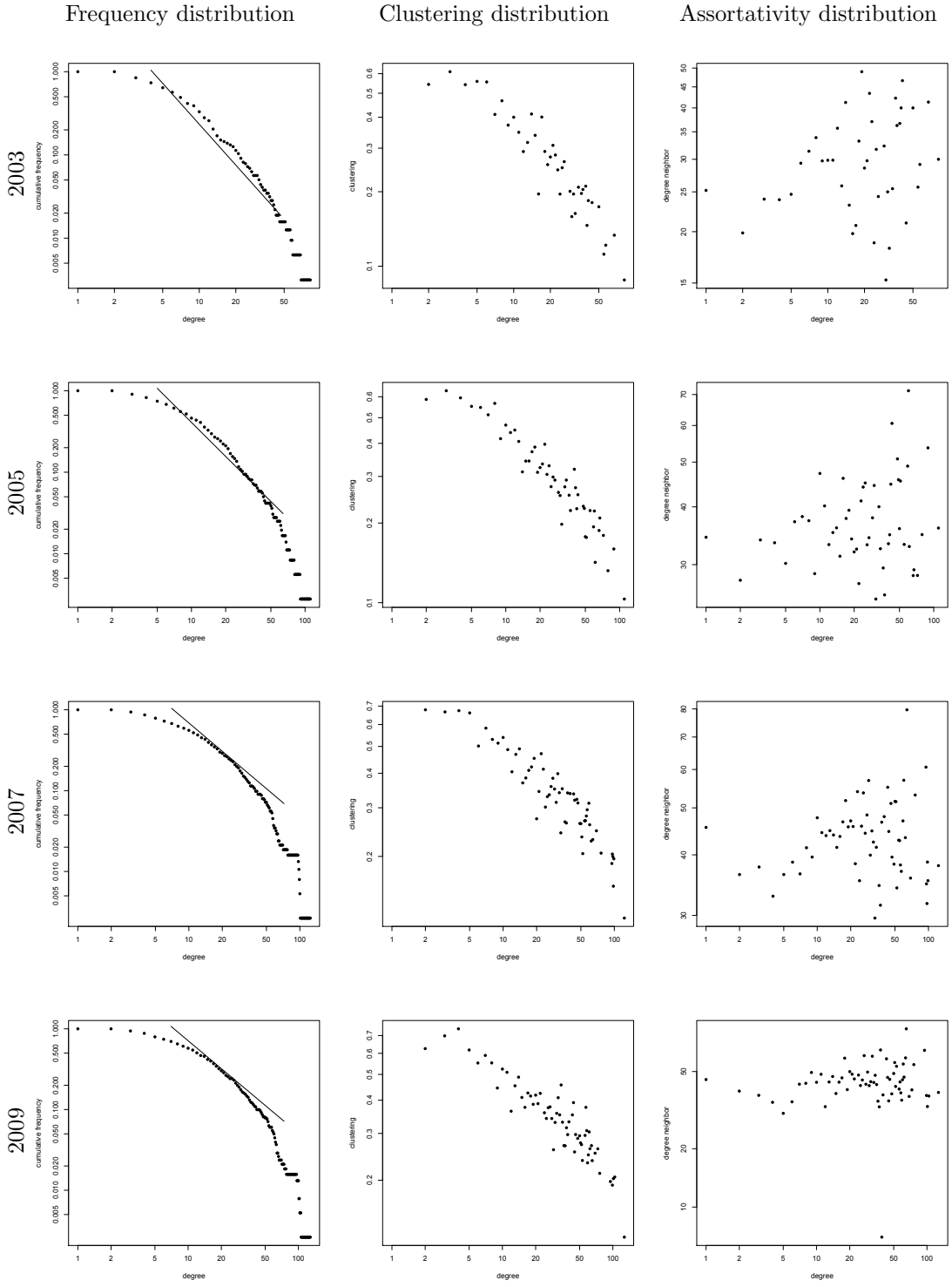


Figure 6.6: Degree distributions of the acquaintanceship network.

moving to the middle of the distribution. The absence of a horizontal tail signifies that the network does not have a small "elite" group of individuals that act like "hubs", connecting different communities. Rather, there are many high-degree nodes, i.e., individuals acquainted with many other individuals.

The second column in Figure 6.6 shows the degree clustering distribution. The acquaintanceship network evolves into a nearly perfect power-law with time, indicating that high-degree nodes are indeed connected to nodes with low clustering coefficient. This confirms that hubs do exist, but as explained above, the volume of hubs is much higher than that found in the coauthorship and communication networks.

Finally, the degree assortativity distribution of the acquaintanceship network, presented in the third column of Figure 6.6 presents a curious case. The plots for years 2003 through 2007 do not exhibit an easily identifiable linear dependence. However, the network in its mature stage, in year 2009, shows that assortativity is nearly perfectly independent of degree. The plot shows a nearly horizontal linear distribution, indicating that no matter what degree a node has, the degree of its neighbor is high. In other words, the acquaintanceship network in 2009 is so densely connected that most nodes are connected to many other nodes ("everybody knows everybody") and thus every incoming node to the network will have to necessarily (rather than preferentially) connect to a high-degree node.

6.4 Evolution of the socio-academic configuration

The analysis of network evolution presented thus far illustrates the dynamics by which CENS scholars connect on different collaboration networks over time and the mechanisms of preferential attachment. This analysis is based solely

on selected topological features of the network: average path length, clustering coefficient, degree assortativity coefficient, and various degree distributions. A more nuanced interpretation of these findings can be obtained by placing them in the socio-academic context in which CENS collaboration interactions take place.

For example, one of the findings reported above, in § 6.2, indicates a strong correlation between clustering coefficient and degree assortativity of the coauthorship network. The analysis indicates that there exists a solid correlation between these two patterns: as the network becomes more sparse and uniform with time (decreasing C), scholarly collaboration patterns become more mixed (decreasing r), i.e., highly-connected individuals start collaborating with lowly-connected ones. This finding, however, is restricted to degree assortativity only; it ignores other types of mixing patterns that might have contributed to the decrease in network clustering over time. For example, is it possible to speculate that the network becoming more sparse is indicative of higher interdisciplinary collaboration and/or higher collaboration across different institutions? In this context, the question that I would like to address is: what specific mixing patterns are accountable for the decrease in the network's clustering coefficient?

In the remainder of this chapter, I extend the evolutionary analysis presented above to a set of node-based social and academic characteristics. The aim of this analysis is to address the second part of the second research question, i.e., can specific evolutionary features in the CENS collaboration networks be explained in terms of changing configurations of organizational, disciplinary, institutional, and international nature? I address this question in the next section, via a detailed analysis of the dynamics of preferential attachment in terms of specific socio-academic characteristics.

In this section, I look at the overall evolution of the socio-academic landscape

of CENS. Leaving aside collaboration interactions for a moment, I present below some statistics about the history of CENS in terms of number of participating researchers, their affiliation, their department, and their academic rank. The results presented in this section are not based on network data, i.e., they do not illustrate any form of collaborative activity. Data presented here are statistics obtained from the CENS annual reports. As explained in § 3.2.1, the annual reports are the official reporting documents published yearly by CENS. They are published in May and they report the progress of the center for the preceding fiscal year. The first CENS annual report was published on May 1, 2003 and contains crucial information regarding this nascent NSF Science and Technology Center. The latest available annual report described in this dissertation is the one published on April 30, 2009.

Although the exact structure of annual reports has changed over time, they all have a similar overall organization. An annual report begins with an executive summary, followed by a general description of the research objectives and areas. The bulk of the annual report is divided into two major sections. The first section describes in detail the progress of all technology and application areas of CENS. The second section describes education, knowledge transfer, and diversity activities of CENS, together with a management plan, budget information and a summary of outputs and impacts. Annual reports also include a list of scholarly materials published by CENS members in the reporting period, biographical sketches of new CENS faculty, organizational charts, and press materials. Up to year 2007, CENS annual reports also contain a comprehensive list of official CENS participants, which includes names of participants, academic position, affiliation, department, gender, ethnicity, and citizenship. As explained in § 4.4, I use these data (together with data gathered on personal websites, curriculum vitae, and online directories) to construct the socio-academic profile of nodes in the

collaboration networks. While the CENS socio-academic configuration presented in § 4.4 refers specifically to the academic configuration of the collaboration networks, it is interesting to look at the official socio-academic distribution of CENS participants. Some statistics collected from the annual reports are presented in Table 6.3.

	Quantity	2003	2004	2005	2006	2007
	# Participants	57	198	321	305	277
	# Graduating students	0	14	18	27	9
Affiliation	UCLA		135	235	213	204
	USC		32	41	27	26
	UC Riverside		5	21	30	28
	Caltech		4	7	9	5
	UC Merced		0	4	7	6
Department	CENS†		95	199	121	48
	Computer Science		65	43	49	68
	Electrical Engineering		38	23	29	43
	Biology		3	16	23	27
	Civil Engineering		0	6	16	14
	Information Studies		3	3	2	4
	Environmental Sciences		0	8	24	20
	Education		3	3	8	15
	Film, Media, and Design		2	2	2	3
	Statistics		2	1	4	6
Position	Faculty	27	42	52	47	44
	Graduate student	12	95	150	120	123
	Admin Staff	9	9	13	25	37
	Research Staff	6	6	26	37	30
	Postdoctoral	2	2	6	10	3
	Undergraduate	0	41	72	63	73

Data obtained from official Annual Reports. Blank cells indicate unavailable data.
† Under the heading "Department" of the Annual Reports, university departments, research laboratories, field locations, and research centers are listed. Many individuals listed as "CENS" are likely to be affiliated with a university department, but this information is not available.

Table 6.3: Evolution of the CENS socio-academic configuration, 2003-2007

Table 6.3 is based on the annual reports published by CENS since its inception up to 2007 (annual reports published in 2008 and 2009 do not include a

comprehensive list of participants). The number of CENS participating members has increased steadily over time from 2003 to 2005, from about 50 to just over 300 individuals. This population slightly shrinks in 2006 and 2007. The number of CENS graduating students also has an increasing trend, and it oscillates between zero in the first year, and 27 in 2006.

About two-thirds of all participants are affiliated with the lead university (UCLA) at any time. Participants affiliated with the major participating university (USC) are also constant through time, at about 10-15% of the yearly population. Interestingly, UC Riverside grows into a major collaborating university: in year 2004, it has only five participating members, but this number grows to nearly 30 in 2007. Caltech and UC Merced only have a handful researchers working on CENS projects at any time.

The distribution by department presented in Table 6.3 is likely to be inaccurate. This is because annual reports conflate university departments, research laboratories, field locations, and research centers into one heading, so that, for example, CENS graduate student researchers from the Computer Science department might be listed as "CENS" or "Computer Science". Despite these inaccuracies, it is possible to see a general trend in which the number of participants from application areas (Biology, Civil Engineering, Environmental Science) increases over time, while technology research areas (Computer Science and Electrical Engineering) remain constant.

Finally, Table 6.3 presents participant distribution by academic position. While nearly all categories increase in the presented time period, the following trends are clearly evident from the data: the number of faculty only increase slightly, from 27 in 2003 to 43 in 2007; postdoctoral, graduate and undergraduate students, however, increase sharply in the first three years (graduate students, for

example, rise from a mere 12 in 2003 to 150 in 2006). They decrease in number in following years, but this decline might be due to an increase in the number of graduating students (see above). CENS research and administrative staff increase modestly over time.

Given the limited number of participating faculty, and their important influence in setting the research agenda and objectives, I can analyze more in detail the dynamics in the population of participating faculty. The overall trend, presented in Table 6.3, is that the number of faculty almost doubles in the first two years (from 2003 to 2005), but then slows down considerably. It is interesting to look at what exact dynamics occurred during this time, i.e., what faculty entered and exited the CENS list of participants? This information can be gathered by inspecting Table 6.4, generated using information from the annual reports. Faculty listed as a CENS participants for a given year are marked by a bullet (•). Prior participants that are dropped out of the list are marked by a cross (×). Faculty names in bold typeface indicate CENS participants as of 2007.

Table 6.4 lists details of faculty participants (name, affiliation, and department) for years 2003 through 2007. Faculty are separated by horizontal lines based on the year of joining CENS, and organized in alphabetical order. The table shows that of the 27 faculty listed in the 2003 report, the majority are from UCLA and from the departments of Computer Science and Electrical Engineering. However, there is sufficient institutional and departmental diversity already at the outset of CENS, with faculty from other member universities (USC, UCR, UC Merced, Caltech) and from a wide mosaic of domains (from Atmospheric Science to Civil Engineering, to Education). Nearly all faculty that "jump-start" CENS remain part of the center for the first few years, and are still members today. About six of them leave in year 2006 or 2007.

Faculty name and affiliation	2003	2004	2005	2006	2007
Alkalai, Leon (Jet Propulsion Lab)	•	•	•	•	×
Allen, Michael (UCR, Biology)	•	•	•	•	•
Borgman, Christine (UCLA, IS)	•	•	•	•	•
Caron, David (USC, Biology)	•	•	•	•	•
Daneshgaran, Fred (CSULA, EE)	•	•	•	•	×
Davis, Paul (UCLA, Earth & Space)	•	•	•	•	•
Estrin, Deborah (UCLA, CS)	•	•	•	•	•
Govindan, Ramesh (USC, CS)	•	•	•	•	•
Hamilton, Michael (UCR, Biology)	•	•	•	•	•
Harmon, Tom (UC Merced, Engineering)	•	•	•	•	•
Ho, Chih-Ming (UCLA, Mech Eng)	•	•	•	•	•
Judy, Jack (UCLA, EE)	•	•	•	•	•
Muntz, Richard (UCLA, CS)	•	•	•	×	×
Potkonjak, Miodrag (UCLA, CS)	•	•	•	•	•
Pottie, Greg (UCLA, EE)	•	•	•	•	•
Requicha, Ari (USC, CS)	•	•	•	•	×
Rotenberry, John (UCR, Biology)	•	•	•	•	•
Rundel, Philip (UCLA, Biology))	•	•	•	•	•
Sandoval, William (UCLA, Education)	•	•	•	•	•
Soatto, Stefano (UCLA, CS)	•	•	•	•	•
Srivastava, Mani (UCLA, EE)	•	•	•	•	•
Sukhatme, Gaurav (USC, CS)	•	•	•	•	•
Tai, Y.C. (Caltech, EE)	•	•	•	•	•
Taylor, Charles (UCLA, Ecology)	•	•	•	•	•
Turco, Richard (UCLA, Atmo Sci)	•	•	•	•	×
Wallace, John (UCLA, Civil & Env Eng)	•	•	•	•	•
Yao, Kung (UCLA, EE)	•	•	•	•	•
Zhou, Chong-Wu (USC, EE)	•	•	•	•	×
Burke, Jeffrey (UCLA, Film)		•	•	•	•
Chu, Wesley (UCLA, CS)		•	•	×	×
Cuff, Dana (UCLA, Architecture)		•	•	•	×
Furner, Jonathan (UCLA, IS)		•	•	×	×
Hansen, Mark (UCLA, Statistics)		•	•	•	•
Heidemann, John (USC, Info Systems)		•	•	•	•
Jay, Jenny (UCLA, CEE)		•	•	•	•
Kaiser, William (UCLA, EE)		•	•	•	•
Kohler, Edward (UCLA, CS)		•	•	•	•
Millstein, Todd (UCLA, CS)		•	•	•	•
Palsberg, Jens (UCLA, CS)		•	•	•	•
Sax, Linda (UCLA, Education)		•	•	•	•
Urbashi, Mitra (USC, EE)		•	•	×	×
Ambrose, Richard (UCLA, Environmental Sci)			•	•	•
Enyedy, Noel (UCLA, Education)			•	×	×
Fitz, Michael (UCLA, EE)			•	×	×
Goldberg, Ira (Rockwell, Engineering)			•	×	×
Kohler, Monica (UCLA, Earth & Space)			•	•	•
Mishler, Brent (UC Berkeley)			•	•	×
Saez, Jose (LMU, Civil & Env Eng)			•	•	•
Stabler, Edward (UCLA, Linguistics)			•	•	•
Vallejo, Edgar (UCLA, Biology)			•	•	×
Villasenor, John (UCLA, EE)			•	•	•
Ye, Wei (USC, CS)			•	×	×
Allen, Edith (UCR, Botany)				•	•
Cody, Martin (UCLA, Biology)				•	•
Heaton, Tom (Caltech, Civil Eng)				•	×
Majumdar, Rupak (UCLA, CS)				•	•
Margulis, Steve (UCLA, Civil & Env Eng)				•	•
Taciroglu, Ertugrul (UCLA, Civil & Env Eng)				•	•
Blumstein, Daniel (UCLA, Ecology)					•
Golubchik, Leana (USC, CS)					•
Sabol, Tom (UCLA, Civil & Env Eng)					•

• = present; × = absent; **bold names** = 2007 participants

Table 6.4: CENS Faculty dynamics, 2003-2007

In year 2004, fifteen new faculty become CENS participants. Almost all of them are from UCLA. The two non-UCLA members that join this year are from technical departments at USC (John Heidemann and Mitra Urbashi). Despite being all from UCLA, the incoming faculty of 2004 present a diverse array of specializations. While half of them are from the technical side of the spectrum (CS and EE), the other half includes disciplines such as Architecture, Film/media, Education, Information Studies, and Statistics. Most faculty that joined in year 2004 remain CENS participants until 2007. Many of them are principal investigators and research area leaders at the time of writing.

In 2005, eleven new faculty join CENS collaboration. Again, most of the incoming faculty are from UCLA, but the array of disciplines is less technical, with faculty joining from departments of Education, Linguistics, Biology, Earth & Space, Environmental Science. More than half of the faculty joining in 2005 will discontinue their participation with the Center by 2007.

In 2006, growth slows down for the first time. Ten faculty leave, but six new faculty (from UCLA, UCR and Caltech) join CENS activities. Interestingly, nearly all of the outgoing faculty are from technical disciplines, while incoming faculty are from scientific and applied disciplines (Botany, Biology, Civil and Environmental Engineering). This reorganization of participants, points to a broader pattern of disciplinary readjustment of technology vs. science populations.

Finally, in year 2007, three applied scientists from member universities join (Ecology, Civil and Environmental Engineering), while few others leave, bringing the faculty population down to 43.

In the next section, the institutional, organizational, and departmental arrangements that make up the socio-academic configuration of CENS are explored

in conjunction with the evolution of the coauthorship, communication, and acquaintanceship networks.

6.5 Network evolution and socio-academic configuration

The evolutionary analysis of degree-based assortativity presented in this chapter elucidates the mechanisms of preferential attachment, i.e., the dynamics by which scientists connect with each other based on their position in the network. As explained in the previous section, this analysis is limited to the topology of the network only: it ignores the socio-academic standing of the individuals under study, and how this might have affected the mechanisms of attachment.

In this section, I extend the calculation of assortativity to a set of socio-academic properties, by computing discrete assortativity coefficients. While degree assortativity measures the extent by which nodes with similar degree (i.e., with similar centrality and position in the network) attach to each other, discrete assortativity measures the extent by which nodes with similar characteristics attach to each other. The characteristics analyzed here are the same already employed in the structural analysis of the previous chapter (§ 5.2.3): academic affiliation, department, position and country of origin.

Table 6.5 presents the evolution of discrete assortativity coefficients for the coauthorship, communication, and acquaintanceship networks through the studied time periods, based on nodes' academic affiliation, department, position and country of origin. Discrete assortativity coefficients are calculated using Formula 3.3, which returns $r = 1$ when there is perfect assortative mixing, $r = 0$ when there is no assortative mixing, and $r = -1$ when there is perfect disassortative mixing. By analyzing these results, it is possible to understand how discrete

	property/yr	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Coauth	Affiliation	0.182	0.188	0.302	0.345	0.417	0.446	0.482	0.373	0.326	0.301	0.399	0.436
	Department	0.238	0.196	0.375	0.490	0.462	0.455	0.408	0.293	0.243	0.236	0.281	0.286
	Position	0.084	-0.055	-0.043	-0.039	-0.036	-0.026	-0.015	0.005	0.013	0.019	-0.008	0.012
	Origin	-0.178	-0.050	0.010	0.180	0.143	0.139	0.131	0.043	0.082	0.071	0.078	0.073
Commun	Affiliation							0.014	0.051	0.082	0.047		
	Department							-0.018	0.146	0.122	0.113		
	Position							0.001	0.051	0.041	0.038		
	Origin							0.011	-0.008	0.012	-0.019		
Acquaint	Affiliation			0.262	0.293	0.328	0.344	0.353	0.334	0.319	0.319	0.319	
	Department			0.382	0.365	0.359	0.345	0.318	0.291	0.261	0.251	0.252	
	Position			0.134	0.127	0.108	0.098	0.099	0.092	0.087	0.086	0.086	
	Origin			0.176	0.171	0.156	0.156	0.158	0.148	0.140	0.136	0.135	

Discrete assortativity coefficient, r, computed using formula 3.3, for the coauthorship, communication and acquaintanceship networks (years 1999 through 2010) based on individuals' academic affiliation, department, position and country of origin. Blanks indicate unavailable data.

Table 6.5: Discrete assortativity coefficients of the collaboration networks.

assortativity coefficients have changed in the three collaboration networks, i.e., how scientists have connected with others from similar affiliation, department, position, and country of origin over time. In the remainder of this chapter, I present an in-depth analysis of these results.

6.5.1 Assortative mixing in the coauthorship network

The discrete assortativity coefficients relative to the evolution of the coauthorship network (Table 6.5) are plotted in a graph (Figure 6.7), to aid interpretation and discussion of the results. Figure 6.7 shows how discrete assortativity has changed in the coauthorship network over time, i.e., the extent to which individuals from the same academic affiliation, department, position and country of origin have coauthored papers with each other. As explained earlier, a positive coefficient (approaching $r = 1$) indicates perfect assortativity, i.e. homophilious collaborations; a near zero coefficient ($r = 0$) indicates no assortativity; and a negative coefficient (approaching $r = -1$) indicates perfect disassortativity, i.e., highly variegated patterns of collaboration.

The coauthorship network has a moderately high variation in assortativity mixing over time: as Figure 6.7 shows, coefficients range between a minimum of -0.2 and a maximum of 0.5 between 1999 and 2010. Overall, the coauthorship network is more assortative by academic affiliation and department, and less assortative by academic position and country of origin. In particular, during the first period, assortativity by affiliation and department rises sharply: by year 2005, CENS publications are largely authored between individuals working in the same institutions and departments. In the second period, these two coefficients decrease, indicating that coauthoring patterns become more inter-disciplinary and inter-institutional. The rise in assortativity registered at the very end of

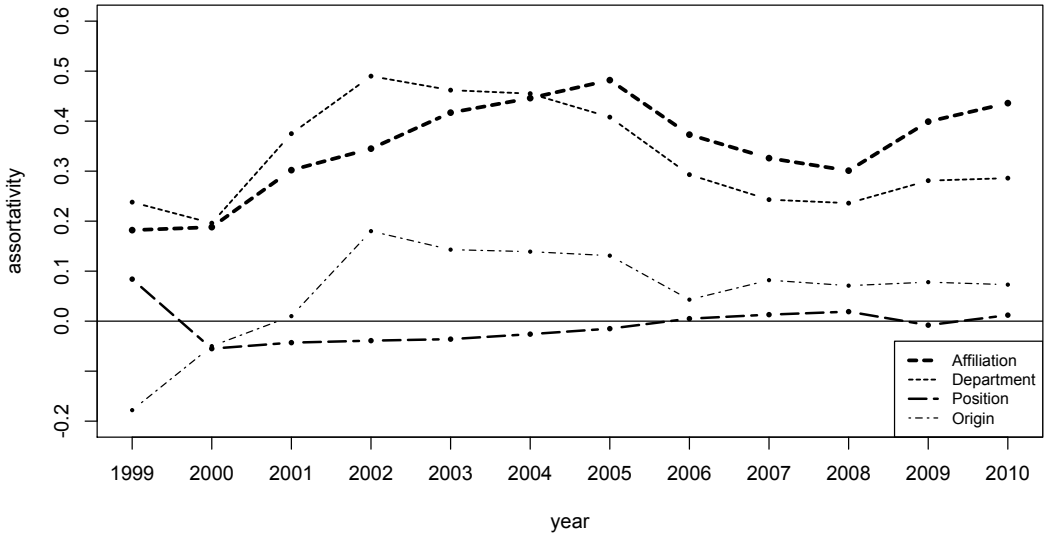


Figure 6.7: Evolution of discrete assortativity mixing in the coauthorship network

the time period (i.e., years 2009-2010) can be attributed to the network being much smaller in size in those years. Country of origin follows a trend similar to that of academic affiliation and department (increasing in the first period, decreasing in the second period), but both its coefficient and fluctuation are much smaller: the coauthorship network only becomes slightly assortative by year 2003, but then drops back to a near-zero value, indicating neither a significant intranational nor international pattern of collaboration among CENS authors. In the studied period, academic position has an essentially unchanged, null assortativity measure. This means that individuals of all academic ranks coauthor papers with others of any rank, without following specific preferential mechanisms.

This analysis of the extent and nature of assortative mixing patterns can be pushed even further to reveal the specific components that contribute to the network becoming more or less assortative. For example, the assortativity mixing by

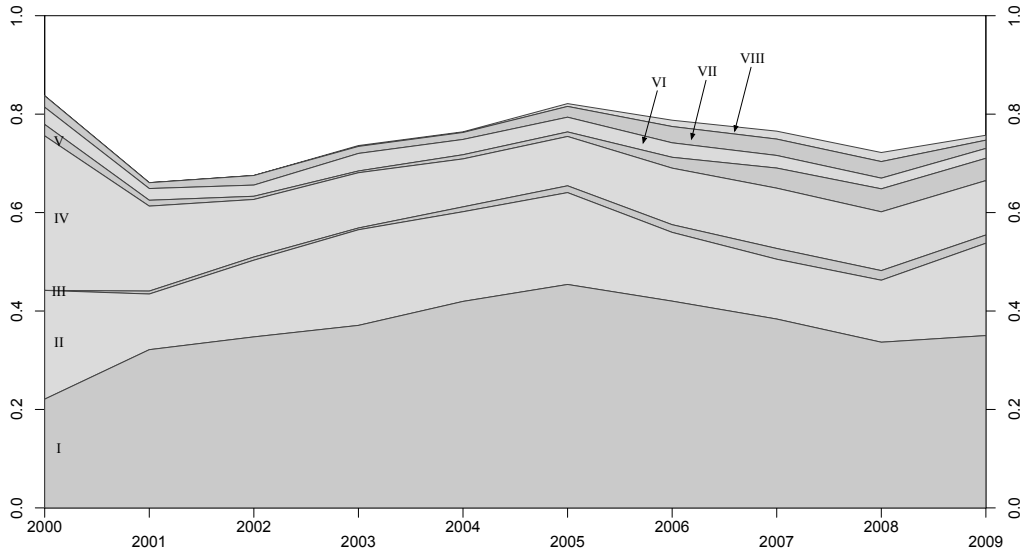
academic department in the coauthorship network, discussed above, shows an increase in assortativity in the first period, and a decrease in the second period, i.e., CENS publications are initially authored by individuals from the same domain, but in the second period, they become more and more inter-disciplinary. In this context, the question at hand is: *what specific collaborations are most responsible for the decrease and the increase in inter-disciplinarity?* In the remainder of this chapter, I address this question for every studied socio-academic property. For each property, I inspect the collaboration pairs that are most responsible for the observed assortativity trend. This analysis is useful both to validate and to justify the results presented thus far.

Academic affiliation. In Figure 6.7, the assortativity coefficient in the coauthorship network based on academic affiliation grows steadily over time, until 2005 and then drops in the second period. This indicates that, coauthorship patterns are intra-institutional at the outset, but then become more variegated. Yet, in the latest snapshot studied here (year 2010) academic affiliation is the single most assortative characteristic in the coauthorship network, suggesting that CENS authors collaborate preferentially with individuals of their institution.

In order to investigate this finding further, I analyze the specific intra- and inter-institutional collaborations that contribute to this assortativity trend. I inspect the most prominent yearly mixing patterns, i.e., the institutional pairs that make up the majority of collaboration volume in every year. These values (raw counts) are presented in Table 6.6, accompanied by a stacked plot which depicts the same values normalized by yearly volume.

Each row in Table 6.6 presents the volume of scholarly collaboration among institutions. The top five rows in this table (IV-VIII), present the pairs that contribute to inter-institutional collaboration. Clearly, as the coauthorship network

	affiliation pair	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09
VIII	UCLA- Caltech	0	0	0	1	1	5	14	17	17	6
VII	UCLA - UC Merced	2	2	6	8	10	19	36	35	31	10
VI	UCLA - MIT	3	4	7	20	22	26	32	27	20	12
V	UCLA - UCR	2	2	2	2	6	8	24	43	43	27
IV	UCLA - USC	27	29	36	63	69	87	125	129	110	66
III	UCR - UCR	0	1	2	2	7	12	17	23	18	10
II	USC - USC	19	19	48	109	129	162	152	128	116	112
I	UCLA - UCLA	19	54	107	208	297	394	457	404	310	209
	Totals	86	168	308	561	708	868	1088	1053	921	597



The table at the top presents raw counts, i.e. the volume of coauthorship connections that exist among scholars of different or same institution. The stacked plot at the bottom depicts the same values as proportion of the totals, i.e., normalized by yearly volume.

Table 6.6: Academic affiliation pairs in the coauthorship network.

is undirected, the order of the pairs is not relevant (e.g., Caltech-UCLA is the same as UCLA-Caltech). The bottom three rows (I-III), present the pairs that contribute to intra-institutional collaboration. For example, in year 2000 there are 19 coauthorship activities among individuals affiliated with UCLA. From a network perspective, this means that in the year 2000, 19 out of 86 total edges

connect nodes of affiliation type "UCLA". In turn, $19/86 = 0.22$ and thus, the stacked plot below the table shows that about 20% of the total volume of collaborations in 2000 are UCLA-UCLA (I). Please note that years 1999 and 2010 are not shown in the table and stacked plot because data are scarce in these years.

A visual analysis of academic affiliation pairs in the stacked plot of Table 6.6 reveals the following scenario. The increase in assortativity registered in the first period can be attributed to the increase in intra-institutional collaboration at UCLA (I): by year 2005, nearly half of the recorded scholarly collaborations are between UCLA scholars. Intra-institutional collaborations at USC (II) increase slightly during the same period. Also, the collaborations between UCLA and USC scholars (IV) become less prominent from 2000 to 2005. These dynamics are responsible for the coauthorship network becoming more intra-institutional in the first period.

In the second period, assortativity mixing by affiliation slows down, i.e., the coauthorship network becomes slightly more inter-institutional. The pairs that fluctuate the most in this direction during the same period are the inter-institutional collaborations between UCLA and UC Riverside (V) and UC Merced (VII): both were negligible in the first period, but become more prominent in the second. Also, after 2005, the growth of intra-institutional collaborations within UCLA (I) and USC (II) slow down considerably, making the network more variegated in terms of institutional cross-fertilization.

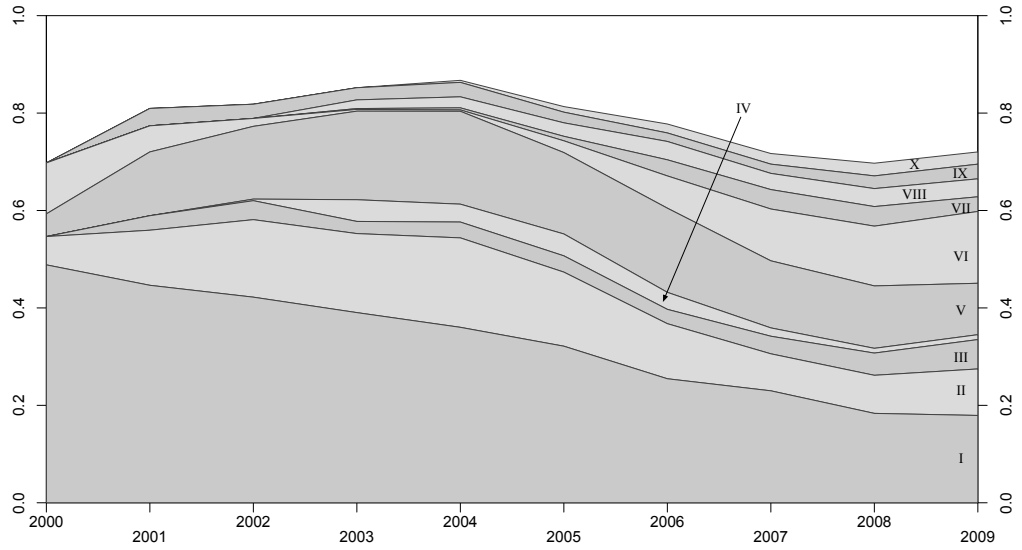
In sum, the scholarly collaboration scenario is initially dominated by publications authored within UCLA and USC. The inception of CENS, in 2002, does not provide an immediate boost of inter-institutional activity. However, by 2005, intra-institutional collaboration starts dropping and new inter-institutional collaborations with partnering universities UC Riverside and UC Merced become

more and more prominent. Despite this increase in inter-institutional collaboration, affiliation is the single most assortative socio-academic property in 2009: CENS authors write papers preferentially with others within their own institution.

Academic department. From Figure 6.7, the evolution of discrete assortativity mixing based on academic department follows a trend very similar to that based on academic affiliation: in the first period, up to about 2004, it grows steadily; in the second period it drops considerably. Thus, coauthorship patterns are very intra-disciplinary at the outset (scientists collaborate within their own department) but then become more inter-disciplinary. It is interesting to note, from Figure 6.7, that the growth of assortativity mixing by department predates that by affiliation: department reaches a peak in year 2002, just prior to the constitution of CENS; affiliation reaches a peak in 2005, when CENS is already a consolidated research center. This result alone suggests that the inception of CENS did indeed contribute in making scholarly collaboration patterns more inter-disciplinary, but authors hardly collaborated outside the walls of their own institutions. I explore more in detail assortativity mixing by department, by looking at Table 6.7, which presents prominent academic department pairs in the CENS coauthorship network.

A quick visual inspection of Table 6.7 reveals that the dynamics of scholarly collaboration at CENS in the period under study are far from stable. Intra-departmental collaborations in Computer Science (I) make up about 50% of all coauthorship volume in 2000; by 2009, they are halved. The strong presence of intra-departmental collaborations in Computer Science is telling of the nature of research being performed at CENS. The domain of networked sensing emerges historically from computer network research and is thus, normally located as a

	department pair	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09
X	EECS - Film	0	0	0	0	3	10	20	23	24	15
IX	Biology - Env Sci	0	6	9	14	21	19	19	20	24	18
VIII	EECS - Statistics	0	0	0	10	16	24	41	35	34	22
VII	EECS - Env Sci	0	0	0	1	3	8	36	42	37	18
VI	EECS - Biology	9	9	5	2	2	21	72	112	113	88
V	Comp Sci - Elec Eng	4	22	46	102	135	145	188	145	118	63
IV	Civil Eng - Civil Eng	0	0	1	25	26	39	38	18	9	6
III	Biology - Biology	0	5	12	14	23	29	32	38	42	36
II	Elec Eng - Elec Eng	5	19	49	91	130	132	123	80	72	57
I	Comp Sci - Comp Sci	42	75	130	219	255	279	277	242	169	107
	Totals	86	168	308	561	708	868	1088	1053	921	597



The table at the top presents raw counts, i.e. the volume of coauthorship connections that exist among scholars of different or same department. The stacked plot at the bottom depicts the same values as proportion of the totals, i.e., normalized by yearly volume.

Table 6.7: Academic department pairs in the coauthorship network.

branch in departments of Computer Science. Sensor network technologies, however, require the design and construction of wireless sensors, and, in turn, interaction with engineering disciplines follows necessarily. The increasing incidence of Electrical Engineering in the CENS coauthorship network can be seen both in

intra-departmental (II) and inter-departmental (V) patterns between 2000 until 2006, when they both settle down in volume.

Besides the network becoming less centered around intra-departmental collaborations in Computer Science, another factor that greatly contributes to its increase in inter-disciplinarity in the second term (starting from 2003) is the appearance of new collaboration between the technology disciplines (CS and EE) and applied natural and social sciences. Just one year after the inception of CENS, many inter-disciplinary collaborations begin flourishing: between EECS and Biology (VI), Environmental Sciences (VII), and Film and Media (X). These collaborations directly reflect the evolution of the center's research agenda and application areas.

In sum, looking at paper collaborations, CENS has progressively detached itself from a CS-centric scholarly record, to become more inter-disciplinary over time. The increase in inter-disciplinarity can be attributed to CENS' need to develop sensor network technologies (Electrical Engineering), apply and deploy them in field environments (Biology and Civil Engineering), and concurrently deal with data analysis issues (Statistics).

It is important to mention that in Table 6.7, I list together both binned and un-binned categories, e.g., "EECS" and "Comp Sci - Elec Eng". This does not influence the reliability of the results presented. The row indicated as "Comp Sci - Comp Sci" indicates the volume of coauthorship among computer scientists only. The row indicated as "Comp Sci - Elec Eng" refers to the volume of coauthorship events among computer scientists and electrical engineers. These results are interesting to be displayed separately. However, when reporting the extent of collaboration of biologists with both computer scientists and electrical engineers, I sum these values and include the results as "EECS - Biology". This

allows me to discuss the interaction among natural scientists (biologists) and technologists (computer scientists and electrical engineers) more conveniently, without affecting the indicated totals and proportions.

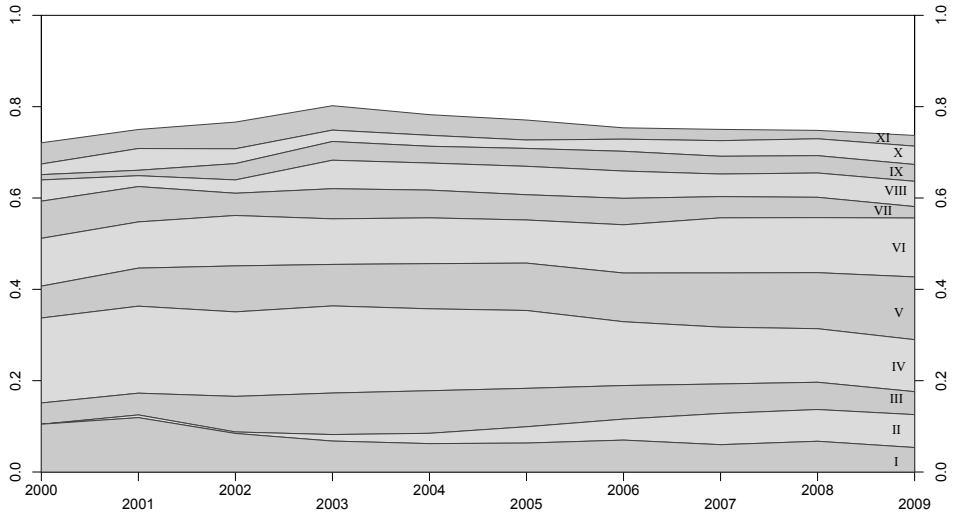
Academic position. It is clear, from Figure 6.7, that academic position has a null assortativity mixing which stays unchanged through time. A look at the academic position pairs of the coauthorship network, shown in Table 6.8, confirms this finding.

The stacked plot in Table 6.8 shows a very linear, nearly unchanged volume of collaboration among different academic ranks in the studied period. None of the analyzed pairs stands out considerably. Some minor fluctuations are nevertheless observed. For example, collaborations among full professors (I) shrink slightly, while coauthorship between research staff (II) and graduate students (III) increases, in the same period. This result suggests that CENS research was "bootstrapped" by faculty members, but later, as the center grew in size, more and more researchers and graduate students became involved and patterns of collaboration among them became more prominent (V).

In sum, the coauthorship network is weakly assortative with respect to academic position, i.e., coauthorship activities involve scholars of all ranks without significant preferential attachment mechanisms. A decomposition of the prominent academic position pairs reveals that, although collaboration pairs are overall highly mixed according to academic position, full professors contributed to bootstrapping scholarly collaboration at the outset of CENS. Graduate students and staff researchers become major players in CENS scholarly publications as the network and the research center become more mature.

Country of origin. As explained above and shown in Figure 6.7, the evolution of assortativity by country follows a trend which is similar to that of affiliation

	position pair	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09
XI	Assistant Prof - Grad	4	7	18	30	32	38	27	26	17	14
X	Postdoc - Staff	2	8	10	14	17	16	29	36	34	24
IX	Assoc Prof - Staff	1	2	11	23	26	34	47	41	35	22
VIII	Assoc Prof - Grad	4	4	9	35	42	54	65	52	49	33
VII	Assoc Prof - Prof	7	13	15	37	43	48	63	49	41	15
VI	Professor - Staff	9	17	34	56	71	82	115	127	111	77
V	Grad - Staff	6	14	31	51	70	90	116	125	113	82
IV	Grad - Professor	16	32	57	107	127	148	152	131	108	68
III	Grad - Grad	4	8	24	51	66	73	80	68	55	30
II	Staff - Staff	0	1	1	8	16	31	50	72	64	43
I	Professor - Professor	9	20	26	38	44	55	76	63	62	32
	Totals	86	168	308	561	708	868	1088	1053	921	597



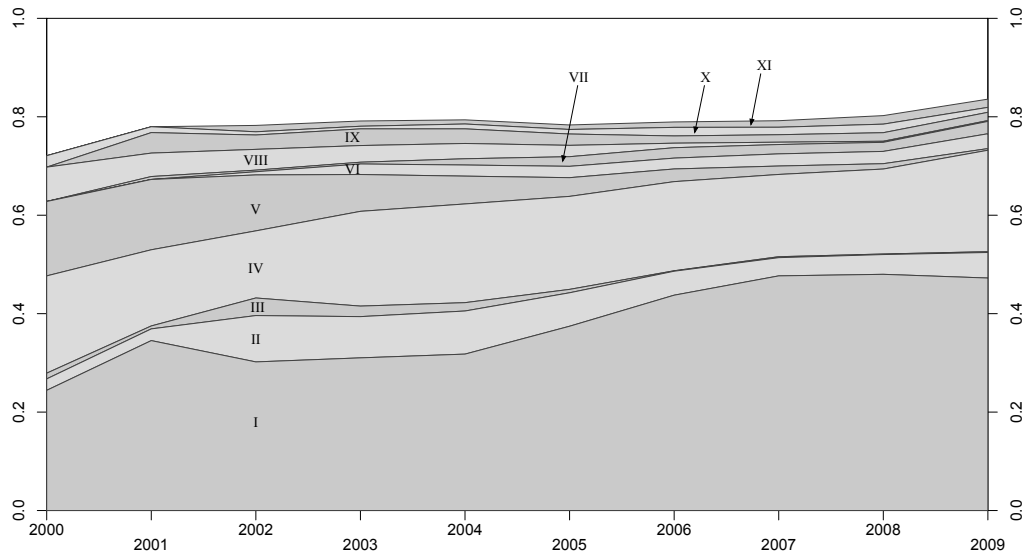
The table at the top presents raw counts, i.e. the volume of coauthorship connections that exist among scholars of different or same academic position. The stacked plot at the bottom depicts the same values as proportion of the totals, i.e., normalized by yearly volume.

Table 6.8: Academic position pairs in the coauthorship network.

and department — with an increase in the first term, followed by a decrease in the second term. However, assortativity measures for country of origin are much closer to zero (i.e., no assortativity).

Table 6.9 shows the specific intra-national (I-III) and inter-national (IV-XI) collaborations that account for assortativity mixing by country of origin. At

	origin pair	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09
XI	Australia - USA	0	0	4	6	6	8	12	14	16	10
X	Taiwan - USA	2	2	2	3	7	8	19	16	16	6
IX	Italy - USA	0	7	9	19	21	20	16	16	16	10
VIII	China - India	6	8	13	19	22	20	10	5	2	1
VII	South Korea - USA	0	1	1	2	9	17	23	20	17	15
VI	Iran - USA	0	0	2	12	16	20	24	26	23	18
V	China - USA	13	24	35	42	40	33	28	18	10	2
IV	India - USA	17	26	42	108	142	164	197	176	159	123
III	China - China	1	1	11	12	12	6	1	2	1	1
II	India - India	2	4	29	47	62	59	53	39	37	31
I	USA - USA	21	58	93	174	225	325	476	502	442	282
	Totals	86	168	308	561	708	868	1088	1053	921	597



The table at the top presents raw counts, i.e. the volume of coauthorship connections that exist among scholars of different or same country of origin. The stacked plot at the bottom depicts the same values as proportion of the totals, i.e., normalized by yearly volume.

Table 6.9: Country of origin pairs in coauthorship network.

the network's outset, the vast majority of collaborations is among Americans (I) and between Americans and Indian (IV) and Chinese (V) researchers. Between years 2001 and 2006, intra-national collaborations flourish within Indian

(II) and Chinese (III) researchers, while China-US collaborations (V) shrink considerably. These patterns are responsible for the increase in assortativity mixing by country recorded from 2002 to 2006, visible in Figure 6.7. By year 2007, the coauthorship scenario becomes regular again: it is dominated by collaborations among American researchers (I) and inter-national collaborations between USA and India (IV).

6.5.2 Assortative mixing in the communication network

Compared to the coauthorship network, discussed above, the communication network presents a much more regular scenario of assortativity mixing patterns. As shown in Figure 6.8, which is a pictorial representation of the values presented in Table 6.5, it is evident that all recorded assortativity measures change very slightly, or not at all, between 2005 and 2008. Moreover, all measures are very low and largely remain between 0 and 0.1 levels. Overall, none of the recorded assortativity measures stands out: online communication patters do not follow any preferential attachment rule, as CENS individuals discuss with others regardless of affiliation, department, position, and origin.

The biggest variation is recorded between 2005 and 2006: the network becomes slightly assortative by academic department and this value stays constant throughout the end of 2008. This indicates that academic department is the only socio-academic property to have a slight influence over online discussion patterns. Since the study of assortativity mixing by academic affiliation, position, and origin in the communication network does not reveal any interesting fluctuations, only mixing by academic department is further investigated here.

Academic department. The relatively high assortativity coefficient by academic department is entirely expected in a communication network: on mailing

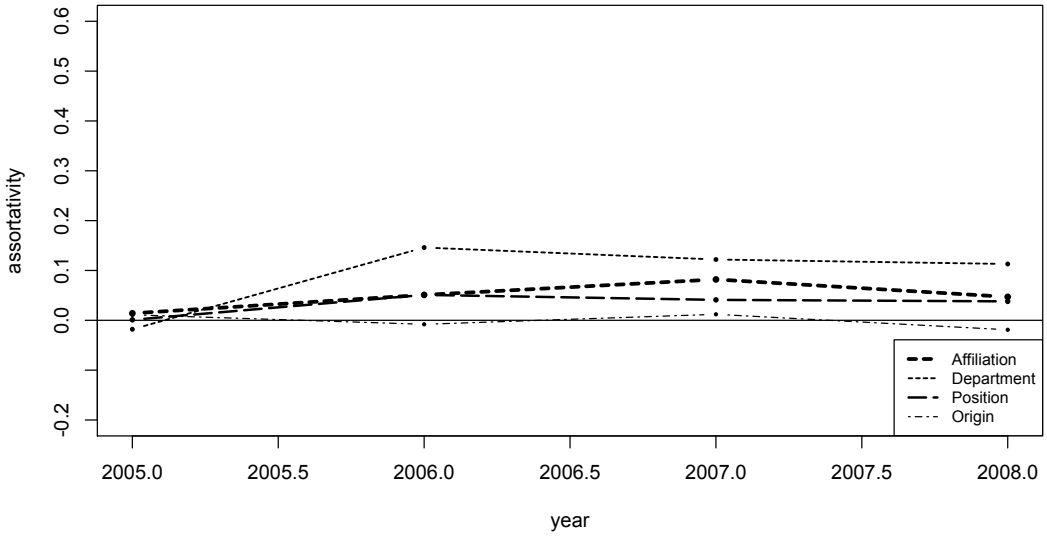
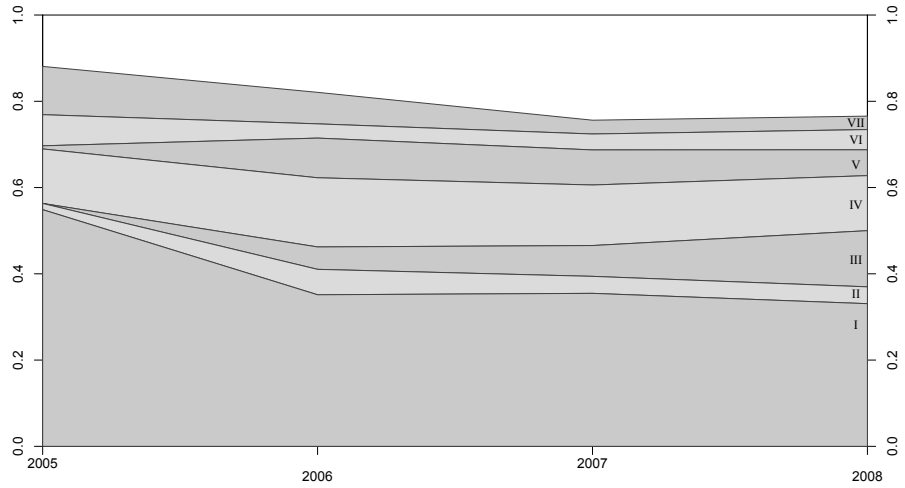


Figure 6.8: Evolution of discrete assortativity mixing in the communication network

lists, it is normal for individuals to connect with others that work in their own area of specialization. In fact, one would expect even a much higher assortativity coefficient by department in this sort of network. This value can be better understood with the aid of Table 6.10, which lists prominent department pairs of mailing lists discussants. The bottom two rows in the table (I-II) present intra-departmental communication within the departments of Electrical Engineering and Computer Science. The top five rows present inter-departmental communication pairs (III-VII). A visual analysis of the stacked plot reveals that the fluctuation recorded in Figure 6.8 between 2005 and 2006 is linked to a number of minor adjustments: online communications between computer scientists stay roughly the same (I), but those between electrical engineers increase considerably (II). Between 2005 and 2006, a number of inter-disciplinary communication trends are also established: members of Film, Theatre, and Media departments

(III) and Statistics (V) begin using mailing lists on a regular basis. In the same time period, biologists (VII) and environmental scientists (VI) slow down their communication with computer scientists and engineers.

	department pair	**'05**	**'06**	**'07**	**'08**
VII	EECS - Biology	31	31	13	12
VI	EECS - Environmental Sciences	20	14	15	18
V	EECS - Statistics	2	39	33	23
IV	Computer Science - Electrical Engineering	35	68	57	49
III	EECS - Film, Theatre & Media	0	22	29	50
II	Electrical Engineering - Electrical Engineering	4	25	16	15
I	Computer Science - Computer Science	152	149	144	127
	Totals	255	468	339	392



The table at the top presents raw counts, i.e. the volume of discussions among scholars of different or same department. The stacked plot at the bottom depicts the same values as proportion of the totals, i.e., normalized by yearly volume.

Table 6.10: Academic department pairs in the communication network.

Overall, this analysis shows that CENS mailing list activity is dominated by interactions between technologists — computer scientists and electrical engineers. Different disciplines have different collaborative practices: technologists rely heavily on mailing lists to discuss their work, report on project progress,

ask for help, etc. Their communication traces and published content are openly available to all mailing list subscribers. Other disciplines rely on mailing lists much less. Contribution from the scientific side of the collaborative spectrum is only marginal (biologists and environmental scientists are only initially involved in mailing list activity; their communication slows down considerably in later years). Moreover, with the exception of Film, Theatre and Media, many other non-technical disciplines are absent from the communication network. Disciplines such as Civil Engineering and Education, largely represented in the coauthorship network, do not rely on mailing lists for electronic communication.

6.5.3 Assortative mixing in the acquaintanceship network

A study of discrete assortativity in the acquaintanceship network (Figure 6.9) does not reveal any major fluctuations in the time period studied. However, both acquaintanceship and coauthorship networks display the same overall configuration of discrete assortative mixing, i.e., with academic affiliation and department more assortative than academic position and country of origin. In the first term, up to 2005, affiliation and department follow inverse trends, with assortativity by affiliation increasing, and assortativity by department decreasing. In the second term, the acquaintanceship network becomes less assortative by department and assortativity by affiliation remains constant. Thus, while in the first term CENS scholars are acquainted mostly with others within their own department, acquaintanceship patterns become more interdisciplinary over time, yet intra-institutional: scholars increasingly connect with others working in other disciplines, but in their own institution. In the time period under study the acquaintanceship network is only slightly assortative by country of origin and academic position: there is only a minimal preference for individuals to be ac-

quainted with others of their own country and of similar academic rank. Due to their high coefficient values and their variation over time, assortativity mixing measures by academic affiliation and department deserve a closer look.

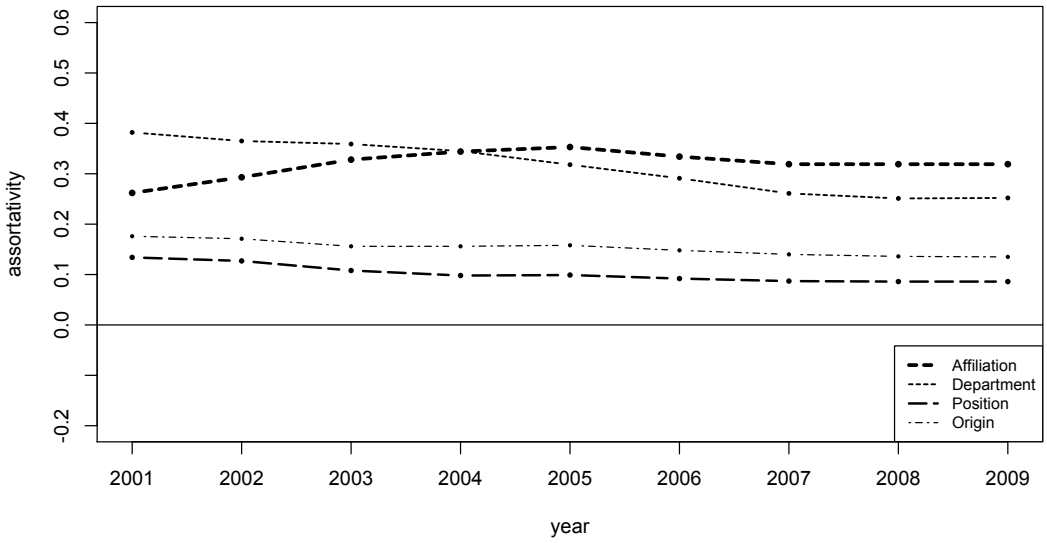
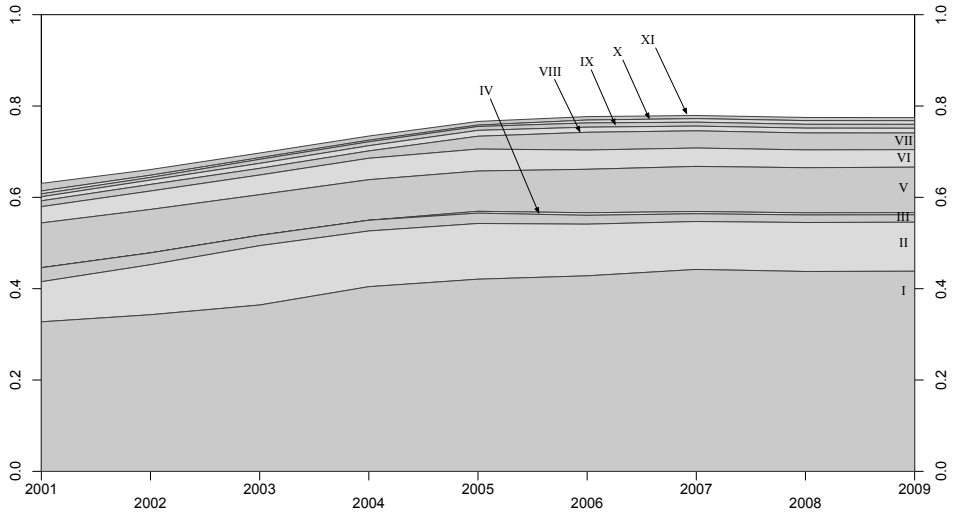


Figure 6.9: Evolution of discrete assortativity mixing in the acquaintanceship network

Academic affiliation. Prominent academic affiliation pairs in the acquaintanceship network are presented in Table 6.11. The rows of this table show the number of scholars from different institutions that became acquainted with each other in a certain year. Please note that year 2001 refers to both year 2001 and earlier years. The bottom four rows in the table (I-IV) display intra-institutional collaboration, while the top seven rows (V-XI) present inter-institutional acquaintanceship.

The values and diagram of Table 6.11 show that, at the outset, the network is already richly populated by scholars of UCLA, USC, UCR, and Caltech. About 20% of both UCLA and USC scholars who know each other in 2009 met in

	affiliation pair	'01	'02	'03	'04	'05	'06	'07	'08	'09
XI	UCLA - Caltech	13	13	14	16	18	20	21	22	22
X	USC - UCM	5	5	5	6	7	17	24	26	26
IX	USC - UCR	5	6	12	15	19	25	26	27	28
VIII	UCLA - MIT	7	10	15	20	29	30	32	34	34
VII	UCLA - UCM	10	15	20	28	63	104	115	119	119
VI	UCLA - UCR	28	42	59	82	108	113	123	125	125
V	UCLA - USC	77	99	122	155	198	255	302	317	325
IV	UCM - UCM	0	0	0	0	9	15	15	15	15
III	UCR - UCR	24	27	31	41	51	52	52	53	53
II	USC - USC	69	114	178	213	274	303	321	344	350
I	UCLA - UCLA	257	357	499	706	945	1147	1352	1404	1429
	Totals	785	1041	1370	1746	2246	2679	3058	3208	3261



The table at the top presents raw counts, i.e. the volume of acquaintanceship relationships (social connections) that exist among scholars of different or same institution. The stacked plot at the bottom depicts the same values as proportion of the totals, i.e., normalized by yearly volume.

Table 6.11: Academic affiliation pairs in the acquaintanceship network.

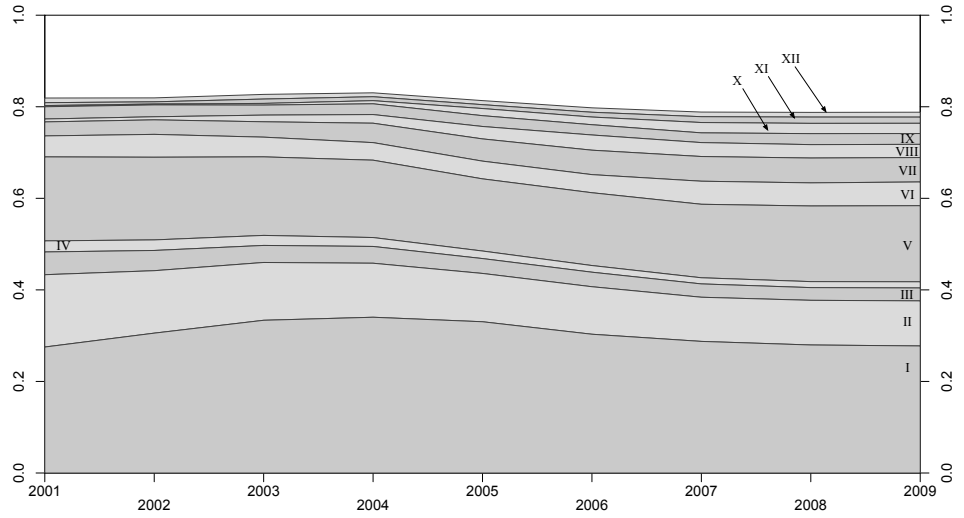
(or prior to) 2001. This indicates that many social, and necessarily academic connections that are at the foundation of CENS predate even the earliest recorded publications: *acquaintanceship precedes coauthorship.* This finding is investigated further below, in § 6.6.

Acquaintanceship between the affiliation pairs in Table 6.11 grows steadily and uniformly over time: from 2001 to 2009, all the recorded pairs grow roughly by a factor of five. It is clear that the increase in the assortativity coefficient observed in Figure 6.9 between 2001 and 2005 can be attributed almost exclusively to the increase in internal connections between UCLA scholars (I): this value grows faster than the rest during this time. Other remarkable fluctuations can be observed for CENS partner institutions UC Riverside an Merced. The personal connections between UC Riverside scholars (III) solidify before the inception of CENS (they reach 51 by year 2005), while those between UC Merced scholars (IV) only begin after CENS is born (they are null, up to year 2004). This can certainly be attributed to the fact that UC Merced is a much younger university (only established in 2005), but also to the fact that UC Riverside played a bigger role in the establishment of the CENS social landscape. Also, a look at the growth of social interactions between these two institutions and UCLA and USC (rows VI, VII, IX, X) reveals the central and growing role of UCR and UCM in the CENS collaboration.

Academic department. In the acquaintanceship network, assortativity mixing by academic department can be decomposed along the pairs presented in Table 6.12. The bottom four rows in the table (I-IV) display intra-departmental collaboration, while the top eight rows (V-XII) display inter-departmental collaboration. There are a number of fluctuations that deserve to be discussed. First of all, all intra-departmental knowledge connections slow down with time: computer scientists (I), electrical engineers (II), biologists (III), and civil engineers (IV), initially all mainly acquainted with other in their own domain, begin to build up inter-disciplinary connections. Technologists (computer scientists and electrical engineers) have strong, constantly growing acquaintanceship connections from the very beginning (V), and they increasingly make connections

with other disciplines: Biology (VII), Environmental Science (VIII), Statistics (X), and Film, Media & Theatre (XI). A combination of these dynamics is responsible for the overall assortativity pattern of Figure 6.9 — a slowly decaying assortative coefficient by academic department, indicating the acquaintanceship network becoming more inter-disciplinary over time.

	department pair	**'01**	**'02**	**'03**	**'04**	**'05**	**'06**	**'07**	**'08**	**'09**
XII	EECS - Civil Eng	8	9	14	15	20	26	31	33	35
XI	EECS - Film	5	5	13	15	19	28	39	44	44
X	EECS - Statistics	2	2	5	12	35	45	69	72	73
IX	Biology - Env Sci	21	27	30	41	53	60	65	77	77
VIII	EECS - Env Sci	5	7	20	33	61	89	93	94	94
VII	EECS - Biology	24	33	46	74	109	143	165	174	174
VI	EECS - Info Studies	36	52	59	67	87	106	154	162	169
V	Comp Sci - Electr Eng	144	188	235	295	354	426	491	530	542
IV	Civil Eng - Civil Eng	19	24	30	34	37	39	41	42	44
III	Biology - Biology	39	46	51	64	73	85	89	89	92
II	Electr Eng - Electr Eng	124	142	173	206	237	278	295	313	321
I	Comp Sci - Comp Sci	216	318	457	594	742	812	879	897	905
	Totals	785	1041	1370	1746	2246	2679	3058	3208	3261



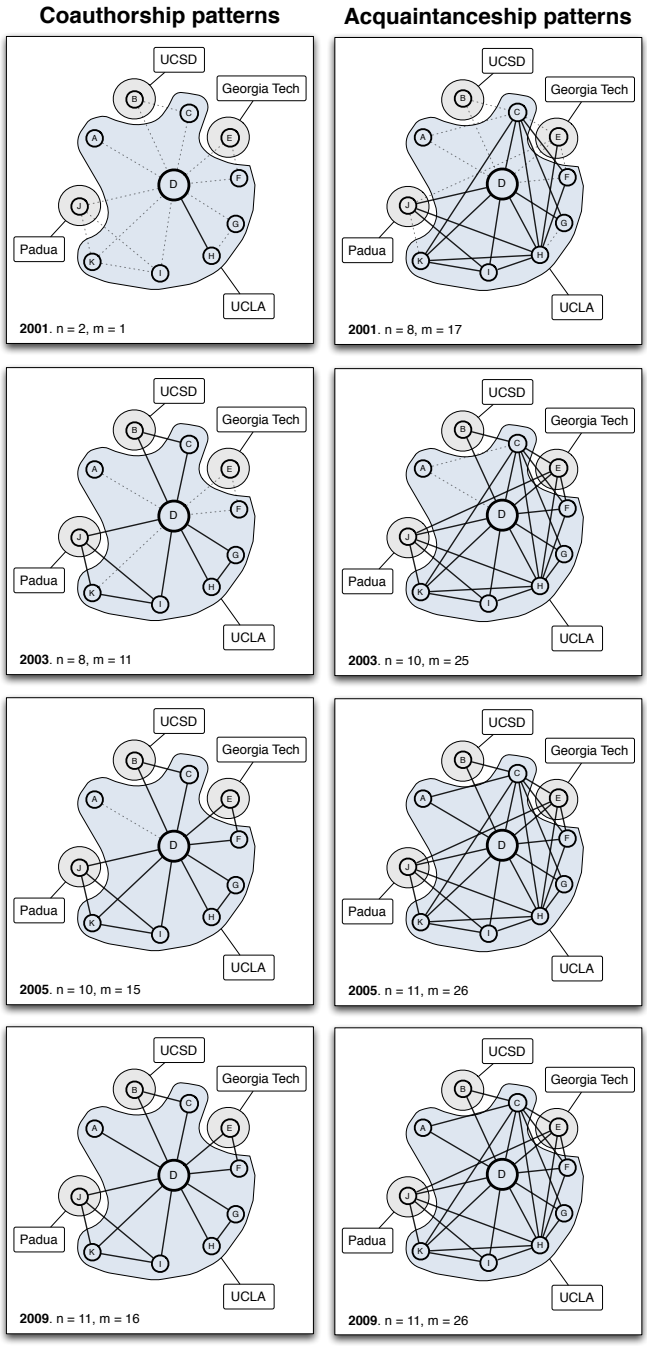
The table at the top presents raw counts, i.e. the volume of acquaintanceship relationships (social connections) that exist among scholars of different or same department. The stacked plot at the bottom depicts the same values as proportion of the totals, i.e., normalized by yearly volume.

Table 6.12: Academic department pairs in the acquaintanceship network.

6.6 Sequential relationship: acquaintanceship and coauthorship

The results presented above point to interesting dynamics between the acquaintanceship and coauthorship networks. A quick comparison between Tables 6.6 and 6.11 shows that scholarly collaborations among and within UCLA and USC scholars in 2001 are just above one hundred (29 + 19 + 54); the corresponding figure in the acquaintanceship network is about four hundred (77 + 69 + 257). This means that while almost four hundred personal connections existed in 2001, only one fourth of them were recorded on paper in the form of scholarly collaboration. This finding points to a sequential relationship between acquaintanceship and coauthorship ties. I corroborate this finding by turning back to the case study discussed earlier on. Figure 6.10 shows the evolution of scholarly and social patterns in the UCLA Computer Vision Lab of UCLA (coauthorship community #11). The figure shows these networks at time intervals 2001, 2003, 2005, and 2009. To put findings into context, the networks are annotated with the academic affiliation of collaborating researchers.

Looking at Figure 6.10, the sequential relationship between acquaintanceship and coauthorship becomes clear. In 2001, only one scholarly collaboration exists, between nodes D and H. In the same year, however, 17 social relationships (out of a total of 26) are already established, between 8 individuals in the community (out of a total of 11). In other words, even though only one CENS-related paper was being written in 2001 by two members of this community, most its members were already acquainted with each other at this time. By year 2003, scholarly collaboration increases, as more papers are being written and published by this community. Also, by year 2003, nearly all members of this community know each



Interactions are indicated as dashed lines (missing) or solid (existing)

Figure 6.10: Anecdotal example: evolution of coauthorship and acquaintanceship.

other. The one remaining acquaintanceship interaction (between nodes A and D) is completed by year 2005. By year 2005, also the collaboration network is nearly complete. The images at the bottom of Figure 6.10 show the coauthorship and acquaintanceship networks at their latest recorded stage.

This case study suggests that social patterns anticipate scholarly collaboration. This does not mean that coauthorship is a direct consequence of acquaintanceship: the act of befriending does not necessarily imply consequent scholarly collaboration. Yet, I find that coauthorship activities take place along existing social paths. For example, from Figure 6.10, it is possible to see that the new collaborations observed between the 2003 and 2005 scholarly networks (e.g., between nodes D, E, and F), emerge along social paths that were established in the years between 2001 and 2003. Clearly, this analysis is limited by the fact that the roster employed for the survey of acquaintanceship is based on the bibliographic record. Thus, my analysis is only capable of identifying temporal relationships between scholarly and social interactions that take place within the boundaries of coauthorship activities. As such, it fails to explain how social dynamics that operate beyond scholarly circles potentially influence future coauthorship activities.

The sequential relationship between acquaintanceship and coauthorship can also be investigated at a broader level, by studying the relationship between scholarly and interpersonal relationships both in terms of academic affiliation and department. Table 6.13 summarizes and compares selected values from Tables 6.6 and 6.11. In particular, it lists academic affiliation pairs within and between UCLA, USC, UC Riverside (UCR) and UC Merced (UCM) in both the coauthorship and the acquaintanceship network in different years. These values are also associated with a Pearson product-moment correlation coefficient, r, which is a measure of the dependence between them ($r = 1$ for linear dependence).

affiliation pair	net†	'01	'02	'03	'04	'05	'06	'07	'08	'09	r‡
USC - UCM	acq	5	5	5	6	7	17	24	26	26	0.446
	coauth	1	2	2	2	3	4	4	4	1	
USC - UCR	acq	5	6	12	15	19	25	26	27	28	0.651
	coauth	1	1	0	0	0	0	6	6	6	
UCLA - UCM	acq	10	15	20	28	63	104	115	119	119	0.787
	coauth	2	6	8	10	19	36	35	31	10	
UCLA - UCR	acq	28	42	59	82	108	113	123	125	125	0.824
	coauth	2	2	2	6	8	24	43	43	27	
UCLA - USC	acq	77	99	122	155	198	255	302	317	325	0.771
	coauth	29	36	63	69	87	125	129	110	66	
UCM - UCM	acq	0	0	0	0	9	15	15	15	15	0.897
	coauth	0	0	0	0	4	4	12	11	10	
UCR - UCR	acq	24	27	31	41	51	52	52	53	53	0.875
	coauth	1	2	2	7	12	17	23	18	10	
USC - USC	acq	69	114	178	213	274	303	321	344	350	0.776
	coauth	19	48	109	129	162	152	128	116	112	
UCLA - UCLA	acq	257	357	499	706	945	1147	1352	1404	1429	0.682
	coauth	54	107	208	297	394	457	404	310	209	

† *"acq" and "coauth" indicate acquaintanceship and coauthorship networks.* ‡ *r is the Pearson product-moment correlation among the two list of values (acq and coauth) for each affiliation pair (the more correlated the values, the closer r is to 1). All correlations have a p-value* < 0.5.

Table 6.13: Affiliation pairs in the coauthorship and acquaintanceship networks: summary and statistical comparison

Glancing at the values of Table 6.13 for different affiliation pairs, it can be noted that the volumes of collaborations in the coauthorship and acquaintanceship networks follow a very similar growth over the years. In particular, it can be seen that the volume of personal connections (acquaintanceship) always exceeds that of scholarly connections (coauthorship) in any year and for every pair. This indicates that, in the CENS community, coauthorship activity is rooted in interpersonal knowledge relationships: acquaintanceship precedes coauthorship.

Moreover, the values of Table 6.13 suggest that the growth in coauthorship volume is linked to that of acquaintanceship, i.e., as more people become acquainted with each other, they also start writing papers together. This finding is validated by a Pearson product-moment correlation. It can be noted that the Pearson coefficient is high (i.e., close to 1) for nearly every affiliation pair. Only three pairs

have $r < 0.75$: USC-UCM, USC-UCR and UCLA-UCLA. The first two can be attributed to the overall small volume of articles written by USC scholars with UCR and UCM scholars. Even though acquaintanceship relationships build up quickly after the inception of CENS, through 2005, scholarly collaboration grows much more slowly. The relatively low Pearson coefficient recorded for the UCLA-UCLA pair can be attributed to the inverse phenomenon. Acquaintanceship ties triple in volume in four years — from 2003 to 2008 — a sudden rise that cannot be possibly replicated in coauthorship activity (coauthorship only doubles during this time). In other words, the inception of CENS and the construction of its headquarters laboratory at UCLA in 2005 greatly fosters social relationships and acquaintanceship among researchers. Such increase in personal connections is reflected in coauthorship activity but to a lesser extent. The physical proximity of the CENS laboratory does increase both scholarly collaboration and social non-academic interactions. This is analyzed in more detail in the next section.

In a similar way, I investigate the relationship (or the lack thereof) between interpersonal knowledge patterns and scholarly collaboration in terms of academic domain. Table 6.14 lists department pairs in the coauthorship and acquaintanceship networks along with a Pearson correlation coefficient, for comparison.

A visual quick analysis of Table 6.14 reveals that one of the key characteristics found in Table 6.13, i.e., that the volume of acquaintanceship ties in affiliation pairs is greater than that of coauthorship ties, is not valid for every department pair. Between years 2001 and 2004, for example, some coauthorship ties exceed acquaintanceship ties between statisticians and technologists (EECS). The same is also true of collaborations between civil engineers: between years 2004 and 2006, coauthorship activity surpasses acquaintanceship. These are, however, two isolated cases. In all the other pairs studied here, it can be seen that, by and large,

department pair	net†	'01	'02	'03	'04	'05	'06	'07	'08	'09	r‡
Biology - Env Sci	acq	21	27	30	41	53	60	65	77	77	0.812
	coauth	6	9	14	21	19	19	20	24	18	
EECS - Film	acq	5	5	13	15	19	28	39	44	44	0.907
	coauth	0	0	0	3	10	20	23	24	15	
EECS - Statistics	acq	2	2	5	12	35	45	69	72	73	0.814
	coauth	0	0	10	16	24	41	35	34	22	
EECS - Environmental Sci	acq	5	7	20	33	61	89	93	94	94	0.888
	coauth	0	0	1	3	8	36	42	37	18	
EECS - Biology	acq	24	33	46	74	109	143	165	174	174	0.925
	coauth	9	5	2	2	21	72	112	113	88	
Comp Sci - Electr Eng	acq	144	188	235	295	354	426	491	530	542	0.489
	coauth	22	46	102	135	145	188	145	118	63	
Civil Eng - Civil Eng	acq	19	24	30	34	37	39	41	42	44	0.373
	coauth	0	1	25	26	39	38	18	9	6	
Biology - Biology	acq	39	46	51	64	73	85	89	89	92	0.982
	coauth	5	12	14	23	29	32	38	42	36	
Electr Eng - Electr Eng	acq	124	142	173	206	237	278	295	313	321	0.297
	coauth	19	49	91	130	132	123	80	72	57	
Comp Sci - Comp Sci	acq	216	318	457	594	742	812	879	897	905	0.424
	coauth	75	130	219	255	279	277	242	169	107	

† *"acq" and "coauth" indicate acquaintanceship and coauthorship networks.* ‡ *r is the Pearson product-moment correlation among the two list of values (acq and coauth) for each affiliation pair (the more correlated the values, the closer r is to 1). All correlations have a p-value* < 0.5.

Table 6.14: Department pairs in the coauthorship and acquaintanceship networks: summary and statistical comparison

scholarly collaboration is well rooted in personal relationships, i.e., coauthors actually know each other in person, across all departmental pairs inspected.

What differs significantly from the previous analysis is the rate of growth of acquaintanceship and coauthorship networks. While in Table 6.13 there is a linear dependence between them (i.e., as more people become acquainted with each other, they also increase their coauthorship activities), Table 6.14 reveals a different scenario: technology-based disciplines and Civil Engineering score low dependency, while natural sciences, statistics and film score very high. This points to different social and scholarly practices by which technologists and other scholars operate when collaborating within and across their disciplines. At one end of the spectrum, technologists make so many personal connections with each other so that it is hard to find a direct incidence of acquaintanceship on coauthor-

ship: technologists at CENS have large social networks that grow independently from their coauthorship networks. At the other end of the spectrum, scholars from non-technical disciplines have fewer acquaintances in the CENS network, but their authorship collaboration with technical scholars is directly dependent on those acquaintances.

Thus, in the context of CENS research, the social networks of technical scholars (computer scientists and electrical engineers) are extensive and go well beyond their coauthorship patterns. The social networks of other CENS scholars (biologists, environmental scientists, statisticians, etc.) are smaller, but they are instrumental to their coauthorship activity.

6.7 Physical proximity at the CENS headquarters

The results presented in this chapter point to the social nature of the CENS collaboration networks and the crucial role of interpersonal networks. In discussing my results, I often refer to the importance of physical proximity in stimulating social and scholarly interactions. My interpretation, however, is only based on observation of institutional and departmental affiliations. Although these two academic characteristics are telling of the approximate position of researchers—the institution, and the department within an institution in which they are based—they do not exactly designate their workplace, i.e., the principal place where scientific work is conducted.

Collecting the geographical position of researchers' workplace was among my initial intentions, for this dissertation research. In fact, I did begin to collect this information for every researcher in the population, but in the midst of data collection, I realized that this process is hindered by two problems. First, it is very

hard to find former workplace information post factum, i.e., researchers rarely make available on their curricula and personal web pages the location of their previous workplace. They do indicate previous job positions, affiliations, and job descriptions, but rarely mention the exact location of their office or laboratory. Second, even when such information is available, it needs to be decoded spatially, i.e., the site plan of an institution is needed to interpret the geographical location of a room name or a seating booth number. It turns out that it is very hard to obtain the seat maps and site plans of office and laboratories where these researchers are based. The seating map of the CENS headquarter laboratory, at Boelter Hall 3551, is however publicly available. The latest available version, compiled in March 2010, is reproduced in the Appendix, § A.3.

The CENS seating map shown in Figure § A.3 covers the entire space shared by researchers, staff, and faculty in the CENS workspace. It excludes administrative offices, director's office, and conference rooms. 3551 Boelter Hall is an open-space laboratory with seating booths organized in seven rows (G, H, J, K, L, M, N), three semi-private booths (T, S, R), and four semi-open office spaces. For the purpose of the present discussion and to aid the analysis of proximity, every row is regarded separately and denoted by its letter, whereas all the detached spaces are grouped together under the letter 'A'. Figure 6.11 presents the coauthorship and acquaintanceship networks of Boelter Hall 3551, depicting how occupants of the CENS headquarters are connected to one another by coauthoring and acquaintanceship relationships, respectively.

A visual analysis of Figure 6.11 reveals the same conceptual arrangement already observed throughout this dissertation: acquaintanceship patterns greatly exceed coauthorship patterns. Within the CENS headquarters there are five times more acquaintanceship than coauthorship interactions. This proportion

is even greater than that observed in the networks as a whole (see Table 6.16), revealing that social interactions are more marked within the CENS headquarters. This result alone reinforces the finding that physical proximity fuels social and scholarly interactions. But it is possible to push this finding further by analyzing a finer granularity of proximity: at the level of seating rows. Table 6.15 lists coauthorship and acquaintanceship interactions among individuals that sit in different rows of Boelter Hall 3551. It shows that the coauthorship network of 3551 is lowly populated and neighbors tend to collaborate more prominently (e.g., individuals in rows G, K, and L coauthor within their rows). The acquaintanceship network is much more populated and variegated: individuals of all rows make acquaintances with one another.

	Row	**A**	**G**	**H**	**K**	**L**
coauth	**A**	0	0	1	3	4
	G	-	3	0	0	3
	H	-	-	1	1	1
	K	-	-	-	6	4
	L	-	-	-	-	5
		A	**G**	**H**	**K**	**L**
acquaint	**A**	6	10	5	34	14
	G	-	3	0	11	7
	H	-	-	1	12	8
	K	-	-	-	28	32
	L	-	-	-	-	6

The letters denote different seating rows of CENS headquarters, Boelter Hall 3551, at UCLA. The number in the cells denote the volume of coauthorship and acquaintanceship interactions between individuals that sit in these rows.

Table 6.15: Physical proximity pairs in the coauthorship and acquaintanceship networks of Boelter Hall 3551.

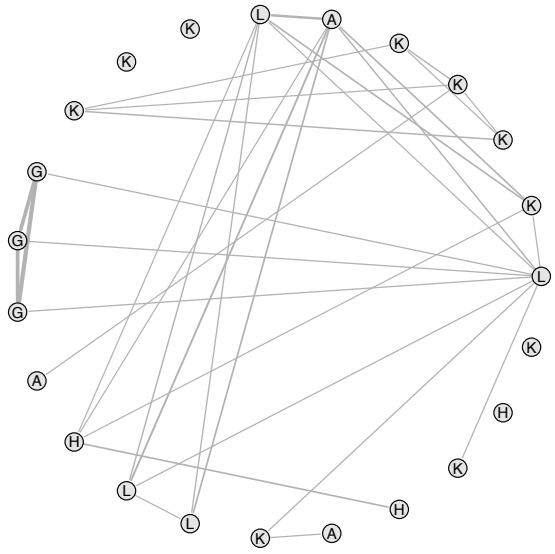
This finding is confirmed by an analysis of discrete assortativity, summarized in Table 6.16. The coauthorship network of Boelter Hall 3551 has a discrete as-

sortativity coefficient by row seat of 0.332, very similar to assortativity levels by affiliation and department in the entire CENS network. This suggests that looking at coauthorship alone, the CENS laboratory can be considered a microcosm of the entire collaboratory. The attachment rules that govern coauthorship at the institutional and departmental level are replicated at a much smaller scale, at the level of physical proximity in seating arrangements. This is also because the specific location of individuals in 3551 loosely reflects departmental subdivisions. The acquaintanceship network, however, presents a different scenario. Assortativity by physical proximity in the CENS headquarter lab is very minimal (0.058), suggesting that although there is a minimal preferential rule to attach with others within the same seating area, acquaintanceship patterns in Boelter Hall 3551 pervade single seating rows.

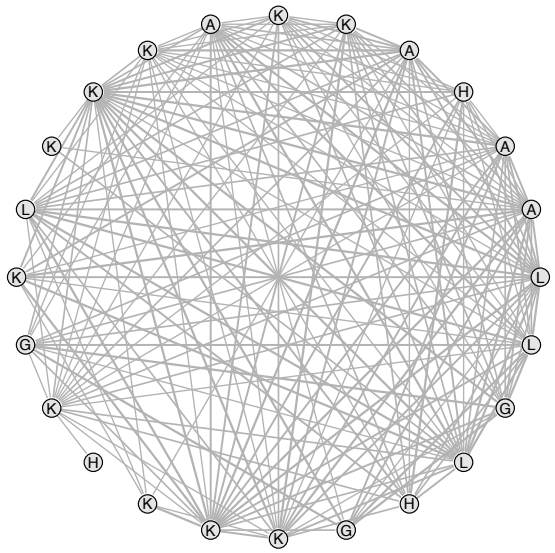
	Coauthorship	Acquaintanceship
CENS (total)		
Nodes (individuals), n	391	385
Edges (collaborations), m	1 747	4 805
Assortativity, r		
....by affiliation	0.301	0.319
....by department	0.236	0.251
CENS (Boelter Hall 3551)		
Nodes (individuals), n	24	25
Edges (collaborations), m	32	177
Assortativity, r		
....by seating row	0.332	0.058

Table 6.16: Assortativity by workplace location (seating row).

Overall, this analysis is indicative of the prominence of interpersonal networks at CENS. While acquaintanceship interactions are pervasive at the CENS collaboratory, as a whole, they are even more pronounced in the context of its headquarter: acquaintanceship patterns diffuse widely within Boelter Hall 3551 regardless of geographical location or academic association.



(a) Coauthorship network, $n = 23, m = 32$



(b) Acquaintanceship network, $n = 24, m = 177$

Figure 6.11: The coauthorship and acquaintanceship networks of Boelter Hall 3551.

6.8 Summary

This chapter describes the findings of an evolutionary analysis of collaboration networks at CENS. Three networks: of scholarly authorship, communication on mailing lists, and social acquaintanceship are analyzed across their temporal components to reveal the collaboration dynamics by which researchers connect on different platforms.

I begin with an evolutionary analysis of network topology and the dynamics of preferential attachment. The coauthorship network is found to be small and very fragmented at the outset, but it later grows into a more extensive, connected network, with a solid CENS author base. Over time, the coauthorship network becomes less cliquish, and more uniform. An analysis of degree assortativity reveals that the coauthorship network is poorly assortative, i.e., the standing of an author is not a major driver of their collaboration patterns. The network based on mailing list communication presents a much more homogeneous, hardly-varying scenario of collaboration. It is small and centered around one connected component. Although communication hubs do exist, the network is fairly sparse and is not governed by any preferential attachment rules: researchers communicate with others regardless of centrality and prestige. The social network of acquaintanceship is the most dense of all studied networks. Many acquaintanceship relationships exist at its outset and they increase significantly over time. At the latest recorded snapshot, the CENS social network is highly clustered and connected — a small-world. Although preferential attachment mechanisms exist at the outset, they slow down with time, and the network becomes so densely connected that most nodes are connected to many others ("everyone knows everyone").

In the second portion of this chapter, I connect the above findings to the

evolution of the CENS socio-academic configuration. By a content analysis of CENS annual reports, I describe the evolution of the center's socio-academic landscape in terms of its changing human composition. This analysis uncovers the formation of a growing core of faculty members. Initially populated by faculty from technical disciplines, this community becomes more inter-disciplinary with time.

This socio-academic information is used to extend the analysis of assortativity to a number of discrete characteristics to uncover how researchers with similar socio-academic profiles attach preferentially with each other. Among the studied socio-academic characteristics, the communication network does not present any major attachment rules. The coauthorship and acquaintanceship networks are found not to be assortative by country of origin and academic position, i.e., these properties do not have a direct influence on collaboration patterns. However, they are moderately assortative by academic affiliation and department, i.e., individuals that belong to the same institution and department tend to write papers with and befriend preferentially others within their own institution and their own academic specialization. Both networks become more inter-disciplinary with time, i.e., over the years, researchers tend to connect more, both on paper and in person, with others from different departments. Affiliation, however, becomes more assortative; over time, researchers favor scholarly and personal connections within their own institutions.

In the concluding two sections of this chapter, I employ the coauthorship and acquaintanceship networks to perform two specific analyses. First, I analyze the temporal relationship between them, finding a sequential relationship: acquaintanceship precedes coauthorship. Second, I perform a site-specific analysis of physical proximity. I find that the CENS headquarter office at Boelter Hall

3551 is a microcosm of the broader CENS ecology: acquaintanceship patterns are largely more prominent than coauthorship patterns and diffuse widely within the laboratory regardless of exact workplace location (seating row).

CHAPTER 7

Discussion

The results presented thus far illustrate the configuration and evolution of the CENS collaborative ecology in terms of selected manifestations of collaboration: scholarly coauthorship, communication on mailing lists, and personal acquaintanceship. In this chapter, I distill and interpret these results, and frame them in the context of related literature. I begin, in the next section, by presenting a summary of the findings of this dissertation. The sections that follow discuss the methodological and theoretical implications of my research for related studies of scientific networks and cyberinfrastructure.

7.1 Summary of the results

This dissertation examines the topology, structure, and evolution of scientific collaboration networks in a modern research collaboratory. Many of the results presented in this dissertation cannot be easily summarized in a compact tabular format, especially those relative to network evolution. However, for simplicity of reference, I have compiled, in Table 7.1, a list of the key topological and structural properties of the coauthorship, mailing list communication, and acquaintanceship networks.

Property	Coauthorship	Communication	Acquaintanceship
Nodes (individuals), n	391	119	385
Edges (collaborations), m	1 747	994	4 805
Connected components	5 (377, 5, 4, 3, 2)	1 (119)	1(385)
Diameter (largest distance)	6	4	5
Average path length, ℓ	2.952	2.095	2.427
Maximal cliques	291	368	5 925
Largest clique	14	14	20
Clustering coefficient, C	0.301	0.461	0.359
Structural communities	14	7	8
Comm. overlap, χ^2 (p-value)			
....coauthorship	-	4.54×10^{-7}	2.2×10^{-16}
....communication	4.54×10^{-7}	-	1.6×10^{-2}
....acquaintanceship	2.2×10^{-16}	1.6×10^{-2}	-
....affiliation	2.2×10^{-16}	0.070	2.2×10^{-16}
....department	2.2×10^{-16}	0.081	2.2×10^{-16}
....position	0.024	0.20	7.3×10^{-4}
....origin	0.0023	0.23	6.22×10^{-8}
Assortativity, r			
....by degree	0.013	-0.057	-0.060
....by affiliation	0.301	0.047	0.319
....by department	0.236	0.113	0.251
....by position	0.019	0.038	0.086
....by origin	0.071	-0.019	0.136
....by seating row (Boelter Hall)	0.332	-	0.058

This table summarizes some basic topological and structural properties of the coauthorship, communication, and acquaintanceship networks: number of nodes (n) and edges (m), number of connected components, diameter of the largest connected component, average path length (ℓ), number of maximal cliques, size of the largest clique, clustering coefficient (C), number of detected structural communities, p-value of the test of independence (χ^2) between structural and socio-academic communities, and assortativity by degree and socioacademic properties (2008 values).

Table 7.1: Summary of the main results found for the coauthorship, communication, and acquaintanceship networks.

The remainder of this section reports the main results of this dissertation. I begin by summarizing the results of the topological analysis (points 1-3), followed by the results of the structural analysis (points 4 and 5). Later, I report the findings of assortativity and preferential attachment analyses (points 6 and 7) and the study of network evolution (points 8 and 9). The final item in the list covers separately some key findings relative to the communication network built from mailing list data (point 10).

1. Acquaintanceship patterns extend beyond coauthorship circles. The overall configuration of the networks of collaboration, introduced in Chapter 4, and summarized in Table 7.1, is useful to draw some conclusions relative to the fundamental modes of collaboration at CENS. I found that CENS research involves nearly 400 individuals. Over a period of 10 years (2000-2009), these individuals author over 600 publications, for a total of 1747 coauthoring relationships (coauthorship edges). The same population is internally linked by 4805 personal relationships (acquaintanceship edges). This result alone suggests that acquaintanceship ties are far more numerous than are coauthoring ties (4805 vs. 1747). This indicates that there are more researchers connected to each other via personal relationships than by coauthorship. This finding, which is even more pronounced within the CENS headquarters office (at Boelter Hall), points to the inherent social nature of the CENS collaboration ecology.

2. All networks of collaboration are well-connected: average path length is low. With the analysis of configuration and topology, presented in Chapter 4, and also summarized in Table 7.1, I find that all the collaboration networks feature a giant component, i.e., a connected portion of the graph that includes the majority of nodes. Nodes in the giant component are accessible by simple paths. In the networks analyzed in this dissertation, these giant compo-

nents have diameters of length six or less. This means that six steps, at most, are required to connect the two most remote nodes in the networks. I also find that the average path length of these networks is very low—between two and three. For example, the CENS coauthorship network has an average path length of 2.952 (from Table 7.1), indicating that, on average, any node in this network can be reached from any other by only performing just under three steps. Short diameter and average path length connote a well-connected network in which information can transfer easily between nodes. Well-connected networks are crucial to productivity, for they enable knowledge exchange and cross-fertilization of ideas. For example, in the CENS acquaintanceship network, an average path length of two means that it takes only two steps, on average, to transfer information from a node to another. This means that even if I (Alberto) have only 30 acquaintances at CENS (accessible via path length = 1), I can potentially reach the entire CENS population by only performing one step outside my social circle, i.e., by asking my acquaintances to introduce me to theirs (path length = 2).

3. All networks of collaboration are sparse: clustering coefficient is low. The clustering analysis presented in Chapter 4 reveals that all networks of CENS collaboration have low clustering coefficients. For example, the CENS coauthorship network has a clustering coefficient of 0.301 (from Table 7.1). This means that the coauthorship network is not very dense, i.e., it does not have a high number of cliques. In other words, although authors do organize themselves in communities of collaboration, they tend to write papers both within and outside of their closest circle of collaboration. All networks of CENS collaboration have a similar sparse topology. Similar clustering coefficients are observed across all studied networks, suggesting that the modalities by which CENS researchers cluster together do not change significantly based on the platform of collaboration: whether it is scholarly papers, mailing lists, or interpersonal knowledge

relationships, researchers form communities that result in similar topologies.

4. Coauthorship and acquaintanceship communities overlap significantly. In Chapter 5, I present a comparative analysis of community structure at CENS. I subdivide the networks of collaboration in structural communities, i.e., groupings of researchers that are highly connected in these networks. Using a community detection algorithm, I find that these networks are composed of 14, 7 and 8 structural communities, respectively (Table 7.1). A comparative analysis of these communities reveals the following relationships (also summarized in Table 7.1): (a) coauthorship and acquaintanceship communities overlap significantly; (b) communities of coauthors and mailing list discussants overlap slightly; (c) communities of discussants and acquaintances do not overlap significantly. These findings are telling of the modalities by which researchers collaborate across different platforms. The first finding, in particular, elucidates the mechanisms of social and scholarly interaction. It shows that in the field of sensor network research, and in the context of the collaboratory studied here, some form of personal relationship exists between coauthors. I find that tight-knit communities of coauthors overlap fairly well with communities of acquaintances, i.e., groups of researchers that actually know each other.

5. Coauthorship and acquaintanceship communities are internally mono-institutional and mono-disciplinary. In Chapter 5, I push further the comparative structural analysis discussed above, to investigate how topological structures relate to the organizational, disciplinary, institutional and international arrangements of CENS collaborations. Findings from a comparison between structural and socio-academic communities across all studied networks are summarized in Table 7.1. They show significant overlap between scholarly communities and academic affiliation and department. This means that commu-

nities of coauthors tend to be populated with individuals working in the same institution and domain. A similar finding is obtained for the acquaintanceship network, indicating that communities of people who know each other are mono-institutional and mono-disciplinary. These communities are also found to be slightly homophilious in terms of academic position and country of origin. Finally, no dependence is found between communities of mailing list discussants and their socio-academic composition: discussion groups are variegated in terms of their disciplinary and institutional components.

6. Network centrality of researchers has very little effect on their collaboration patterns. In Chapter 6, I look at the evolution of assortative mixing patterns in the CENS collaboration networks. I begin by looking at the simplest form of assortativity mixing, by degree. This measure illustrates how individuals with different degree ranks, i.e., of different centrality in the network, connect with others. For example, is it fair to speculate that prolific authors in this network are more likely to collaborate with other prolific authors? My analysis of degree assortativity reveals that all the analyzed networks at CENS show very little (or no) preferential patterns of this kind (Table 7.1). An historical analysis of assortativity measures (Table 6.1 in Chapter 6), reveals that in the years prior to the inception of CENS, the coauthorship network is highly fragmented and cliquish. At this time, the network is slightly assortative—authors tend to connect with other authors with similar standing. In later years, the network quickly becomes more uniform, a solid core of scholarly collaboration emerges, and assortativity drops to zero or near-zero values, suggesting that authors' prestige has no direct incidence on collaboration patterns. The communication and acquaintanceship networks exhibit very poor assortativity from the very beginning and their degree assortativity levels are essentially unchanged throughout the time period studied: communication on mailing lists and acquaintanceship

patterns are not influenced by standing in the network, i.e., by how much someone interacts on a mailing list, and by how well they are known in the community.

7. Researchers write papers and make acquaintances preferentially with others within their own institution and departments. In the second portion of Chapter 6, I extend the evolutionary study of assortativity to a set of socio-academic characteristics, to reveal how different academic configurations of collaboration govern the dynamics of attachment. In other words, I address the question: how do individuals with a similar socio-academic profile connect with one another? From this analysis the following scenario emerges. The social and scholarly networks are found to be moderately assortative by academic affiliation and department, i.e., individuals from the same institution and the same department tend to preferentially collaborate and to know each other. This finding validates the results of the structural analysis: that coauthorship and acquaintanceship communities are internally mono-institutional and mono-disciplinary. It indicates that besides the mechanisms of attachment that exist at the local level, i.e., within structural communities, global patterns also enable collaboration with peers in their own institution and department. None of the studied networks are assortative by academic position. This indicates that researchers of all ranks (professors, staff researchers, graduate students, etc.) connect on paper, on mailing lists, and in person without any attachment preference. This finding corroborates the above remark about the absence of prestige effects in sensor network research collaboration. Besides being highly mixed in terms of academic position, CENS collaboration networks are also found to be highly mixed in terms of country of origin, i.e., the nationality of researchers does not have a direct influence on the dynamics of collaboration. Finally, an analysis of physical proximity in the workplace (limited to Boelter Hall 3551) reveals that neighboring researchers coauthor preferentially with neighbors (i.e., researchers seating

in their vicinity). No direct relationship is observed between physical proximity and acquaintanceship: social patterns at Boelter Hall diffuse throughout the laboratory regardless of seating location.

8. The coauthorship and acquaintanceship networks become both more intra-institutional and inter-disciplinary over time. The temporal analysis of assortativity mixing presented in Chapter 6 illustrates the attachment mechanisms observed in the CENS collaboration networks. As noted above, in the coauthorship and acquaintanceship networks, assortativity coefficients by affiliation and department are high. A temporal investigation of these coefficients reveals, however, that these properties evolve according to different dynamics. While assortativity by affiliation increases over the time period under study, affiliation by department decreases. The first portion of this finding indicates that not only are researchers found to connect preferentially within their institutional boundaries (both on paper and in person); they also do so at an increasing rate, making the social and scholarly networks more intra-institutional over time. My in-depth analysis of assortativity patterns reveals that the lack of interdisciplinary growth is due to the high and increasing volume of collaborations within UCLA and USC: most of CENS work is performed within the walls of these institutions. Although some inter-institutional collaborations appear as CENS matures, notably with partnering universities UC Riverside and UC Merced, the coauthorship and acquaintanceship networks remain remarkably intra-institutional. The second portion of the finding indicates that, although researchers connect preferentially with others within their own discipline, in the long run, both scholarly and social networks become more variegated in their disciplinary configuration. A temporal decomposition of department pairs reveals a drastic decline of scholarly collaboration between computer scientists and between electrical engineers in the period under study, and an increasing volume of collaborations between

technologists and scholars from natural sciences, social sciences, and media scholars.

9. Acquaintanceship precedes coauthorship. The temporal analysis of attachment dynamics presented in Chapter 6 exposes yet another characteristic about the CENS collaboration ecology: scholarly interactions are found to be preceded by acquaintanceship. I find that some fundamental social connections among CENS researchers predate the oldest available publications: researchers indicate that they have known each other in person prior to becoming scholarly collaborators. This finding confirms the crucial role of social cohesion in this scientific research center.

10. Network centrality and socio-academic profile of researchers have no effect on their patterns of communication on mailing lists. While the social and scholarly arenas are governed by institutional and disciplinary attachment rules, as well as minor prestige-based mechanisms, I find no direct dependence between socio-academic profile and standing of researchers and their communication activity on mailing lists. This means that these interactions are highly mixed and involve both frequent mailing list users and occasional ones, from a variegated mosaic of disciplines and institutions. The lack of attachment rules can be attributed to data sparseness and to the very open nature of mailing list platforms.

This summary of results is a review of the mechanisms of collaboration that operate at CENS, based on the structure and evolution of its scholarly, communication, and social networks. In the following sections, I reflect on these results and discuss how they compare and complement related research on scientific collaboration.

7.2 A fluid, non-cliquish small-world

The analysis of network topology presented in this dissertation indicates that all CENS collaboration networks share a peculiar topological configuration. I have found that both average path length and clustering coefficient are fairly similar across networks of coauthorship, communication, and acquaintanceship. Average path length and clustering coefficient are important network measures as they are telling of the small-world nature of a network. The average path length of a network is simply the average of all paths between all of its nodes. As explained throughout this dissertation, average path length also gives an idea of the information transfer of a network: the smaller the average path length, the more connected is the network, i.e., the easier it is to transfer information around it. Clustering coefficient is a measure of clique density in a network. The higher the clustering coefficient, the more clustered and cliquish is the network. When a network has high clustering coefficient and short average path length, it exhibits *small-world* properties [115]. A *small-world* is a network in which any two nodes are only a few steps apart, regardless of network size. In a small-world network, individuals are not necessarily all connected to each other, yet they are easily reachable from one another via a short path. My results indicate that the CENS collaboration networks are hybrid variants of small-worlds: they have very low average path length, but not a significantly high clustering coefficient. This makes the CENS topological configuration worth discussing further, especially in comparison to other collaboration networks.

Small-world effects have been noted in a number of scholarly networks[1]. For example, coauthorship networks in biology [132, 133] and neuroscience [61], have

[1]Table A.1 in the Appendix presents average path length (ℓ) and clustering coefficient (C) for a number of published networks, alongside the networks of CENS collaboration. The table is subdivided according to the network type: bibliographic, communication, and social.

fairly low average path lengths ($\ell = 4.92$ and $\ell = 5.7$, respectively) and high clustering coefficients ($C = 0.60$ and $C = 0.76$, respectively). Coauthorship networks in physics have been found to have subtler, but still tangible, small-world effects, as they have higher average path length ($\ell = 6.19$) and slightly lower clustering coefficient ($C = 0.56$) [132, 133]. In sum, the coauthorship networks constructed from biology, neuroscience, and physics papers approximate very well the typical properties of small-worlds.

Small-world effects have also been found in social networks. For example, Watts and Strogatz [115] analyzed the topological properties of a Hollywood actor network, in which two actors are joined to each other if they have acted in a movie together, uncovering strong small-world effects ($\ell = 3.48$ and $C = 0.78$). A very similar configuration was found in the network of American corporate company directors ($\ell = 4.60$ and $C = 0.88$) [171]. These results demonstrate that both the social networks of Hollywood actors and company directors are true small-worlds: both actors and directors form cliquish circles of acquaintanceship (high clustering coefficient) and are reachable within very few hops in the network (short average path length).

The CENS coauthorship and acquaintanceship networks deviate from this small-world model[2]. They have a hybrid configuration, with very low average

[2]I have intentionally excluded the communication network from my discussion on small-world topology and preferential attachment. This is not because the topology of the communication network differs from that of the coauthorship and acquaintanceship networks. In fact, it presents very similar properties: low average path length, low clustering coefficient, and near-zero assortativity. I do not discuss it for two reasons. First, my mailing list data are fairly scarce: only four years (as opposed to ten years for the other networks), and one fourth of the entire population (119 individuals out of 391). Data scarcity makes the communication network not very appropriate for comparative analyses. Second, there are not many other networks to compare these data to. While many studies of communication networks exist, not many of them look at semi-public mailing lists, like the one I have. For example, networks of email correspondence [172] and peer-to-peer file exchange [173] have both been shown to exhibit poor small-world properties. However, these networks are based on very different communication traces and comparing them to the CENS mailing list network is not useful or informative.

path length and low clustering coefficient (see Table 7.1 for details). What reasons can be found behind this hybrid form of small-world configuration? Why is it that individuals in the CENS networks are very close to each other (short average path length), yet they do not form cliquish clusters (low clustering coefficient), as it is observed in many similar networks? The peculiar "small-worldish" configuration of the CENS coauthorship and acquaintanceship networks can be explained by analysis of their attachment patterns. My analysis of degree assortativity has shown that the network centrality of researchers has very little effect on their collaboration patterns. The degree assortativity of a network measures the extent to which nodes with similar degree preferentially attach to one another. High degree assortativity in a network suggests the existence of prestige-based mechanism, as individuals with high degree attach to others with high degree.

Network studies of scientific coauthorship networks have shown moderate degree assortativity coefficients[3]. For example, the network of coauthorship in physics has a significantly high assortativity coefficient ($r = 0.363$) indicating that very prolific physicists tend to coauthor papers with other physicists of similar high standing [132, 133]. The coauthorship networks of biologists [132, 133] and mathematicians [111, 174] have been found to have lower coefficients ($r = 0.127$ and $r = 0.120$, respectively). Yet, all these values point to the existence of a moderate form of prestige-based mechanisms in coauthorship patterns. The near-zero assortativity coefficient of the CENS coauthorship network ($r = 0.013$), as well as its evolution pattern, indicate a contrasting scenario: prestige has essentially no influence on the way that coauthoring relationships are established. This is the main reason behind the small-worldish topology of the CENS coauthorship network. Its low degree assortativity means that as new members enter the CENS

[3]Table A.1 in the Appendix presents degree assortativity (r) for a number of published networks, alongside the networks of CENS collaboration.

collaborative ecology (with low degree) they begin to collaborate both with prolific and with marginal authors. This constant mixing between authors of all degrees keeps the network from becoming cliquish, i.e., very prolific authors do not isolate themselves from the rest of the network.

Exactly the same observation can be made for the network of CENS acquaintanceship. The social networks presented above—of movie actors [115] and company directors [171]—have both been found to exhibit some form of assortativity ($r = 0.208$ and $r = 0.276$, respectively). This means that both the world of entertainment and that of corporations are ecologies in which prestige and popularity matter. The CENS acquaintanceship network has a different configuration. With a near-zero assortativity coefficient ($r = -0.060$), it is an open, inclusive ecology, not subject to prestige-based mechanisms.

The lack of prestige-based mechanisms accounts for the "small-worldish" configuration of the CENS coauthorship and acquaintanceship networks. These networks have a remarkably similar topological configuration: they are small-worlds, for the distance between individuals is very small (low average path length); they are open and not cliquish (low clustering coefficient); they are fluid and inclusive, for individuals of all degrees collaborate and know each other regardless of standing (near-zero degree assortativity coefficient). This finding is telling of both social practices and authorship conventions at CENS.

Regarding social practices and norms, my finding that the configuration of the CENS acquaintanceship network is one in which "everyone knows everyone" substantiates the fact that CENS is a fluid social ecology in which befriending patterns are ubiquitous and are not influenced by standing. The analysis of physical proximity patterns at Boelter Hall 3551 (the CENS headquarters) presented in § 6.7 corroborates this view: the physical and organizational configuration of

the CENS laboratory is one that facilitates social interaction. This open social configuration not only benefits young students and newcomers, but it also contributes to making CENS a more diverse and inclusive collaborative environment.

It is also interesting to employ results of this research to reflect on the nature of coauthorship conventions in this research community. My study offers a portrait of the social, disciplinary, institutional, international arrangements that compose CENS coauthorship. My analysis of assortativity, in particular, provides clues as to the preferential attachment patterns of authors, and by extension, to the nature of coauthorship in multi-disciplinary collaboratory research. I found that researchers coauthor preferentially with others within their department and institution. Yet, scholarly collaboration is largely inclusive: scholars of different ranks (faculty, graduate students, staff, etc.) and of different network degrees (both prolific authors and non-prolific authors) are observed to collaborate on scholarly papers without specific prestige-based attachment rules. This is not an isolated finding of my network analysis; it is consonant with common CENS authorship norms that support inclusion of scholars of all ranks in the preparation and publication of articles: Master's students, Undergraduates, Summer Program visitors, as well as other participating researchers at early stages of their career.

My use of assortativity analysis as an observational lens to peer into the composition of author lists confirms the view of an heterogeneous, inclusive authoring environment, but signals new questions about the means by which researchers get credit for their work. Since successful scholarship is the means to get tenure, promotions, and community recognition, one would expect more monographs, or at least a more conservative, elitist authorship attitude, i.e., a preference to author alone or with few prestigious scholars. However, my results show an inclu-

sive, non-elitist authorship scheme, implying that different recognition and credit mechanisms may be at play in such an inclusive research environment. Not much research on collaboratories and cyberinfrastructure has yet addressed the influence of prestige and elitism on the advancement of scientific collaboration. A study by Finholt and Birnholtz [175] puts this problem forward, by providing a preliminary review of differences in professional culture among three distinct U.S. National Science Foundation cyberinfrastructure initiatives. Related research on invisible colleges and similar high-level interpersonal structures of scientific collaboration has historically addressed the issue. In a 1971 article, de Solla Price posits that it is via mechanisms of inequality, elitism and close-knit connection that the hierarchy of invisible colleges has emerged [176]. The existence of elitist structures has been corroborated in more recent research. In reformulating the notion of the invisible college, Wagner describes a "new Invisible College" that circumvents traditional scholarly, bureaucratic, and political power structures in favor of a social form of elitism: "The more elite the scientist, the more likely it is that he or she will be an active member of the global invisible college" [177, p. 15]. With regard to this research, my study presents a contrasting scenario. My results, grounded in network analyses, suggest that modern collaboratory research circumvents both forms of social and scholarly prestige-based attachment. In the context of a small-scale collaboratory, I find a well connected, uniformly distributed, fluid collaboration network.

7.3 The role of interpersonal networks

In Chapter 1, I defined a collaboratory as a blend of physical and virtual environments in which a multi-disciplinary mix of geographically distributed researchers

use computer-supported technologies to produce scientific knowledge interacting formally and informally, solving problems, sharing data, resources, and ideas. In this dissertation I conceptualized CENS as a research collaboratory to emphasize its loose geographical and organizational structure, its reliance on electronic platforms, its highly collaborative nature, as well as its disciplinary and institutional variety. My conceptualization, however, was done *a priori*, based on my initial understanding of this research center. It is now appropriate to look back at previous and current research on collaboratories and frame my findings within this literature.

A distillation of my results indicates that scholarly and social networks at CENS are mono-disciplinary and mono-institutional at a local level, but at a larger scale, they become more inter-disciplinary over time. The local level that I refer to is that of structural communities—loosely connected, cliquish clusters of individuals grouped with each other based on the topological properties of the network. Turning back to related literature on cyberinfrastructure initiatives, the configuration that emerges from my structural analysis resembles—at least at a conceptual level—the *intensional networks* [127] and the *knotworks* [128] discussed earlier on, in § 2.4.3. Whether one calls them structural communities, intensional networks, or knotworks, my research, in agreement with previous work, points to the importance of small-scale, local, personal networks for accomplishing work in cyberinfrastructure initiatives. As such, this dissertation validates previous ethnographic observations by providing a quantitative map of the social and scholarly infrastructure that supports scientific and technological practices in modern collaboratory research.

Although both my work and that emerging from qualitative studies of cyberinfrastructure signal the importance of personal networks in collaboratory

research, my research however differs in an important way. Nardi *et al.* [127] and Engeström *et al.* [128] find that personal networks are loose assemblages of individuals that span organizational and institutional boundaries. I find the contrary: that structural communities of coauthors and acquaintances are internally composed by researchers of the same institution and the same department. This is in contrast with the perception that modern collaboratories are inherently inter-institutional and inter-disciplinary. It is true that, overall, CENS is composed of a variegated collection of institutions and disciplines: CENS is indeed, like many other collaboratories, multi-institutional and multi-disciplinary. However, I find that, for the most part, inter-disciplinary and inter-institutional interactions do not take place at a local, community-based level. Rather, I find that within small communities, collaboration interactions are fairly homogeneous: they involve scholars of the same institution and department. Focusing on disciplinary diversity, the following scenario emerges: CENS communities are internally mono-disciplinary, but CENS is overall inter-disciplinary. This scenario demonstrates that inter-disciplinary work in a collaboratory is not necessarily carried out at the community level; rather, it is the result of the bridging action of community hubs.

While my findings indicate a general growth of inter-disciplinarity, I do not find a corresponding increase in inter-institutional collaboration. One of the core promises (and premises) of cyberinfrastructure is exactly to provide simultaneous and remote access to data, resources, and tools to relax the distance constraints imposed by traditional laboratory research. My findings, however, do not point in this direction. I find that the volume of inter-institutional coauthorship patterns decreases over time. I wonder then, whether the use of computer-supported collaborative technologies can actually relax physical distance constraints and enable inter-institutional collaboration. Recent research on this matter has found that

even though inter-institutional collaboration has increased over time, its increase is not directly linked to increased use of dedicated collaborative technologies. Lorigo and Pellacini have found that inter-institutional and cross-country collaborations in the past 30 years of physics collaborations, has increased at the same rate both before and after introduction of such technologies [76]. With regard to this, I find my research to be in line with the work of Cummings and Kiesler [15] who warn against the inefficiency of cyberinfrastructure programs to foster collaboration between scholars from different institutions. What my results show, overall, is that it is via social, interpersonal relationships that scientific collaboration nurtures. I find that CENS collaboration has a strong social component and its acquaintanceship patterns are pervasive: researchers know each other well within their coauthorship circle and beyond. In particular, I detect a sequential relationship between acquaintanceship and coauthorship: collaborative interactions are seen to take place along existing social paths. These findings point, again, to the importance of personal, local networks in the processes of scientific collaboration.

In sum, my findings portray a form of scientific collaboration driven by social cohesion: interpersonal knowledge is the glue that holds the CENS scientific collaboratory together. It is this form of social cohesion that makes collaboratory research particularly peculiar: even though the collaboratory is by definition predicated upon notions of remote collaboration and computer-supported communication, my research shows that social, face-to-face relationships are at the core of collaboration activity. Given the crucial role of social relationships for the advancement of scholarly collaboration, one wonders whether the cyberinfrastructure vision of a fluid, distributed, multi-sited science, agnostic to geographical and physical constraints, can ever be attained. In this context, my work reinforces previous recommendations to consider the spatial, social, and human

arrangements that drive scientific advancement and collaboration, and how they differ across different disciplines and organizational settings [178]. Bringing this recommendation to the attention of policy makers and funding agencies has the potential to shape the direction and form of future investments and efforts in cyberinfrastructure.

7.4 Reflections on the notion of complexity

I began this dissertation by framing modern scientific collaboratories, and CENS, in the realm of complex systems. In particular, I outlined some key defining characteristics of complexity—emergence, self-organization, and boundary flexibility—and I discussed how they apply to the organization and function of collaboratory research. With results of my research in hand, and my personal experience working close to CENS for nearly four years, I now find myself at a favorable position to discuss whether complexity science is an appropriate methodological platform to investigate scientific work, especially in the context of collaboratories and related cyberinfrastructure initiatives.

One level of complexity that comes to the surface with my research is the boundary flexibility of the collaboratory. As thoroughly discussed in Chapters 1 and 3, delineating the population under study was not a trivial task. Although CENS maintains and updates an official list of participants, the multi-disciplinary nature of CENS work necessarily involves collaborations that span its organizational boundaries. Throughout my research, I ran into *boundary objects* [28], of many different kinds. Although I did not employ qualitative observation techniques to detect and document their existence, I can provide here some examples that directly emerge from my analyses of data sources. In this dissertation, I

used the official bibliographic record and mailing list logs of CENS to construct a survey roster. Manual analysis of these data sources revealed papers, emails, and people that were clearly at the intersection between two or more disciplines, understandings, and scientific practices. Some papers in the bibliographic record, for example, were only marginally related to CENS research. An example, is the paper that I employ in some of the examples in Chapter 4 [77]. This article, published in the journal *Scientometrics*, is a study of coauthorship practices at CENS. Is this a product of CENS research? Clearly, it is not a paper about the development and application of sensor network technologies, but it is about CENS, and it comes out of the CENS collaboration. As such, this paper can be regarded as a boundary object, in the definition of Bowker and Star [29]: it inhabits different communities of practices (CENS, the field of bibliometrics, the information sciences, and research policy on sensor networks), satisfying the informational requirements of each of them. There are many more papers in the bibliographic record I analyzed that were in a similar boundary location. The same can be said of the survey population. As the survey roster was directly derived from the bibliographic record and mailing list logs, many people in the roster were situated at the intersection of many domains, practices, and laboratories. For example, some individuals that I invited to participate in the survey, wrote back to me saying that they were not sure whether their input would be useful, as their work with CENS was on an occasional and short-term basis. Others wrote back saying that they had been invited to the survey by accident, as they were not officially affiliated with (and paid by) CENS, even though the bibliographic records showed that they were frequent collaborators on CENS articles.

The scenarios above reinforce the observation, made by many complex system researchers, that complex ecologies have inherently flexible boundaries [26]. When studying scientific work via empirical methods, however, one needs to de-

termine a cutoff point. How should one go about drawing these boundaries? There is a binary tension at play. At one end of the spectrum, qualitative studies of science rooted in sociology and cooperative work research can deal with loose, fuzzy, flexible boundaries by employing qualitative techniques of observation. These approaches are able to reveal and interpret the nuances of scientific work and collaboration, but probably fail to see high-level patterns. At the other end of the spectrum, computational and quantitative studies can use network analysis to reveal broad patterns of scientific organization and function, but in order to do so, they need a rigid, well-defined cutoff point. The first approach limits the possibilities of employing network analytic techniques, while the second compromises fluid, open-ended techniques to study the boundaries and the margins of the network. With my research, I tried to play around this methodological tension. First, I mined the topology and configuration of the coauthorship network to construct a population set with rigid boundaries, but as inclusive as possible. Then, I relaxed those boundaries, by allowing individuals to define their own communities within the broader population. Despite being subject to limitations, detailed later in this and in the next chapter, my approach is instrumental in making apparent the breadth of the CENS collaboration ecology. In fact, I argue that, at the least possible level, by administering a social survey listing names and pictures of people related with CENS research, I raised awareness about the existence of a CENS community, delimited by boundaries, blurry as they may be.

While I did find that the notion of boundary flexibility applies very well to modern collaboratory research, I did not find an equally fitting conceptualization for the notions of emergence and self-organization. As anticipated in Chapter 1, the emergent and self-organizational components of collaboratories pertain specifically to their functional arrangement which is not fixed and guided by a master

plan, but rather by adaptive and dynamic interactions among its constituent components [24, 9]. My findings partially corroborate this view.

My research indicates that interpersonal relationships and small-scale dynamics at the level of communities play a preeminent role in collaborative interactions, as discussed in the previous section; these findings, in turn, point to a bottom-up organizational scheme, as those found in many emerging systems. In other words, the small-scale interactions that govern collaborative work within individual communities pose an ideal prerequisite for emergent growth and self-organization. It is at this scale that individual interactions of coauthorship, communication, and acquaintanceship are performed.

At a higher level, the aggregated ensemble of small-scale interactions solidifies into the structure of a system—its top-down configuration which manifests itself in the form of rules, values, ethics, morals, and large-scale patterns. I have mentioned, in Chapter 1, that the structure of a system both constrains and enables small-scale interactions: structural properties can act as barriers limiting the scope of action of individuals, but at the same time they also provide them with a potential framework for action. In this research, I unveil CENS's structural configuration based on its network topology and compare it to its socio-academic arrangement. My major finding in this context, as discussed above, is that the structure of the collaboratory—its repartition into communities of collaboration—reflects very closely its academic top-down organization—its repartition into institutions and departments. From the perspective of complex system science, this finding suggests that the academic organization of the collaboratory acts as a constraint on the extent and breadth of individual interaction. In other words, the adaptive, emergent, dynamic properties of modern collaboratory research are inhibited by the overarching, top-down academic structure in

which the collaboratory is embedded.

This finding brings about two immediate considerations for complex systems research. First, it points to the difficulty of detecting emergent and self-organizing forms of collaboration in scientific research environments with such strong organizational constraints. What kind of methods and tools are to be used to discern emergent bottom-up agency when the behavior of a system is heavily dependent on its top-down structure? Second, it demonstrates that my research elucidates only one of the myriad relationships that define the dialectic between agency and structure; it ignores other forms of structural factors such as political interests, financial constraints, and cultural values that have unequivocal capacity to inhibit or enable interaction. These problems—the lack of powerful, well-established techniques to discern emergence and the inability to account exhaustively for the multi-dimensionality of complex ecologies—are the disputed aspects of the science of complex systems. My criticism is by no means unprecedented; it has been advanced in the past both from within and outside of the field of complex systems [21, for a review]. With regard to this, in his seminal work on the complexity of social systems, Niklas Luhmann came to notice that, paradoxically, "complexity cannot be observed" [17, p. xviii]. A related criticism comes from scholars in science studies that have warned about the risks of borrowing a complex systems approach to explain emergent phenomena in cyberinfrastructure. In recent work, Jackson, Edwards, Bowker, and Knobel posit that in order to move between social organization and technical infrastructure, "what is needed are not rigid maps, but flexible and creative principles of navigation" [24]. I argue that a major challenge for contemporary complex network research is finding a strategy to accommodate such methodological flexibility. Integrating computational studies of networks with field-based observations, ethnography, and other qualitative methods, as discussed in the next section, is an obvious step in this direction.

Only then will sociological theory be able to embrace the notions of emergence and self-organization empirically, and not solely as a mere metaphorical tool.

7.5 Lessons learned amid two modes of studying science

The work presented in this dissertation is essentially quantitative in nature. I employ network analysis and survey research to collect and analyze tangible indicators of scientific interaction. Yet, in many parts of this dissertation I have discussed how the framing of my research, my choice of methods, my analytical investigations, and the interpretation of my results were guided by considerations triggered by my interest in studies of science grounded in qualitative methods and sociological theory. I have discussed that these two modes of studying science—one quantitative, the other qualitative—are not easy to reconcile. In this context, while covering the boundary flexibility problem, above, I have referred to a "methodological tension" between computational and ethnographic studies of scientific collaboration. This research is an attempt to explore and loosen up this tension. Having performed a quantitative analysis of scientific collaboration giving thought to the socio-academic landscape of CENS for the interpretation of my results, I can now reflect on my research process, and the validity of my method, with the aim to inform future studies of this kind. Based on my experience and the lessons I learned throughout my research, I provide in this section some considerations regarding the benefits and shortcomings of using the method advanced in this dissertation to study scientific work in different contexts and environments.

First of all, using a quantitative framework of network analysis as the methodological foundation of my research has an important ramification: I was able to

render and scrutinize the "big picture" of scientific collaboration at CENS. Quantitative analyses of networks make apparent phenomena that are not visible to the "naked eye": they provide a far-reaching perspective, a *bird's eye view* of the dynamics of collaboration. This has its advantages and disadvantages. On one hand, it misses important minutiae and microscopic manifestations of scientific collaboration that can only be revealed by personal interviews and other ethnographic methods. On the other hand, it provides a powerful observational lens to detect and make sense of high level topological, structural, and evolutionary features of collaboration. A scientific environment as a whole might exhibit high-level patterns that are not necessarily apparent at the local level. An example of this is my analysis of network topology: average distance, density and assortativity of a network are high-level features of a collaborative environment that could not be studied using ethnographic methods only. These topological features are important as they provide a standardized, systematic procedure to compare scientific collaboration endeavors with one another. Moreover, only such a high-level analysis can provide insights into aggregate patterns of collaboration. For example, my analysis of discrete assortativity reveals the evolution of inter-disciplinary and inter-institutional work and the specific academic pairs that most contribute to those evolutionary patterns. These metrics are not only useful from a science studies perspective, i.e., to study the development of scientific collaboration in relation to its institutional and disciplinary organization. They are also the yardstick by which science policy and funding agencies review the extent and composition of inter-disciplinary and inter-institutional work in a research center.

There is yet another benefit of using a quantitative approach for the study scientific collaboration: research reusability and reproducibility. My network analysis is based upon tangible indicators of collaboration and it produces re-

sults that can be discerned by computational analysis. As covered in § A.5 of the Appendix, the code developed and the analytical tools employed in this research are built on open source platforms. The coauthorship network is built from a public and openly accessible bibliographic record. The communication network is built from semi-public mailing list interactions available upon subscription. Only the acquaintanceship data, which I collected via a social survey, are not made publicly available, to protect responders' privacy and comply to the ruling of the Institutional Review Board. As such, with the exception of acquaintanceship data, my research method is fully reusable and my results are reproducible. As recently noted by Stodden [179], data reusability and reproducibility of results are crucial to modern scientific communication, especially as scientific inquiry progressively becomes more dependent on scientific computation. Besides reproducibility, the availability of data and tools also ensures that further research on the same or other data sources can be conducted. As new tools for the study of network structure and evolution appear in the literature, the data of this study could be analyzed in novel ways to both improve existing results and provide novel insights. For example, as discussed further in the next Chapter (§ 8.2), the multi-faceted nature of my data sources coupled with the richness of available socio-academic metadata, would make them a convenient platform to develop and test scientific recommendation algorithms. Moreover, my analytical tools, that are specifically suited for the study of time-dependent, multi-faceted, richly annotated networks, could be applied to larger and more complex networks of interaction to validate their potential beyond the scope of scientific collaboration.

While the use of quantitative techniques allowed me to provide a portrait of high-level patterns of collaboration at CENS, supplementing my network analysis with an in-depth analysis of the socio-academic landscape in which those collab-

oration networks operate was instrumental in elucidating my findings. By and large, network analyses of scientific collaboration are based on large data about one specific phenomenon. Studies of scientific collaboration networks, such as those reviewed in Chapter 2, are abundant in data—networks typically range between tens and hundreds of thousands of nodes—but examine only a single manifestation of collaboration, e.g., patterns of coauthorship, co-citation, or co-word use. As such, this body of literature ignores the advantages of triangulation, i.e., the benefit of capturing, investigating, and validating phenomena using multiple data sources. Triangulation techniques are not new. They were introduced by Webb [129], and adapted to social network analysis by Lievrouw *et al.* [47] over two decades ago. Yet, the bulk of computational research on social networks has been slow in embracing them, or has ignored them altogether. This is very likely because capturing information about multiple relationships can be expensive and inconvenient, especially for large-scale studies.

For my dissertation research, I worked with networks of manageable size and I was able to supplement purely quantitative techniques of network analysis with an in-depth inquiry of CENS's socio-academic milieu. The collaboratory ecology that I study is composed of roughly four hundred researchers. Given the relatively small breadth of this collaboratory, I was able to collect information about collaboration patterns from three distinct data sources, rather than just one. I collected information about coauthorship using a bibliographic record, about electronic communication using mailing list logs, and about acquaintanceship by administering a social survey. This allowed me to perform comparative analyses between different manifestations of collaboration. Unfortunately, the collected mailing list data and the resulting communication network are not rich and ample enough to explore in detail communication practices. However, the wealth of collected bibliographic and social data allowed me to perform detailed

comparative analyses between coauthorship and acquaintanceship patterns. This relationship—between coauthorship and acquaintanceship—is important to scientific communication, but it is rarely studied in depth; it is often based on assumptions. As already discussed in Chapter 3, Newman states that "it is probably fair to say that most people who have written a paper together are genuinely acquainted with one another" [157, p. 339]. Many network researchers make assumptions of this kind, without supporting them with data. My dissertation research provides data to validate this assumption.

My results demonstrate that in the context of a small multi-disciplinary collaboratory, circles of coauthors overlap very well with natural communities of acquaintances, thus validating Newman's assumption. My research examines both the extent of social cohesion and its temporal nature in the process of scientific collaboration. The results, as explained in the previous section, are important to policymakers and researchers of cyberinfrastructure initiatives. My hope is that my dissertation brings the importance of this kind of multi-faceted network analysis to the attention of network researchers. As scientific work and scholarly communication move to new, digital paradigms, exploring a single manifestation of connection among scientists would fail to reveal the rich web of interactions that scientists engage in, on electronic platforms, in person, and on social media.

Another advantage of working with networks of manageable size has to do with the wealth of additional data that I could gather for every individual in the network. It is rare that bibliographic repositories, mailing list logs, and similar data sources used for network research make available data about the academic affiliation, department, position and country of origin of people in the database. Even when such data are made available, many studies ignore them, as they are not offered in a structured format, and are thus unusable. Moreover, these data are

often not recorded historically, so that only latest available data are used in network studies, hindering detailed evolutionary analyses. Capturing and analyzing historical socio-academic information about individuals in the CENS population allowed me to perform detailed comparisons between the community structure of collaboration networks and their organizational arrangements. In other words, I was able to test the capability of computational techniques of community detection to describe social and academic configurations of scientific collaboration. My analysis points, once again, to the methodological tension between quantitative and qualitative approaches to the study of science. It is in the interest of both approaches to understand how scientists coalesce into communities. However, these approaches seem to grow in two separate directions. On the one hand, qualitative studies struggle to both provide a framework to quantify the structural components of communities of practice [140] and to align their ethnographic approaches with social network theory [138]. On the other hand, network researchers continue to develop algorithms for the detection of structural communities without considering the sociological ramifications of their methods. All the techniques to detect structural communities in networks are based solely on quantitative properties of the networks. Paradoxically, their efficiency and quality are validated using benchmarks that are also based on quantitative properties [151], leaving one to wonder: what are structural communities representative of? Network researchers have the ability to develop large-scale automated mechanisms for the detection of communities and social structure in networks. Sociologists of science have the ability to make sense and validate the composition of those structures. Unfortunately, these two abilities rarely meet.

In my dissertation research, I have scratched the surface of this problem by comparing the results of a quantitative analysis of community structure with a more qualitative analysis of the institutional and disciplinary settings that those

communities represent. I can argue that the level of familiarity that I reached with the network and the underlying data had its benefits. First, all the data I collected—names, affiliations, coauthorship relationships, communication traces, etc.—were filtered and validated by manual inspection. In large-scale studies of scientific networks, the reliability of collected data is a significant source of error. Name ambiguity in bibliographic databases, for example, negatively affects results by conflating and/or failing to combine names. Even the most reliable techniques of name disambiguation on average only resolve 85% of ambiguities [180]. Given the size of the CENS network and the amount of work I dedicated to maintaining it, I can safely assume that my study does not suffer from name ambiguity issues and other data collection errors. Moreover, by working so closely to the data, I developed an extensive knowledge of the individuals in the database and their scientific and social interactions. By close inspection of their personal web pages, biographies, and curricula, I was able reconstruct the narratives behind their scholarly production, their academic career, and their scientific collaboration patterns. I do not claim that my level of inspection can compete with in-depth ethnographic observations of science. However, this level of familiarity with data is unheard of in large-scale studies of scientific networks. I argue that complex network researchers would gain great advantage from performing manual, close investigations of the data they produce. By connecting more "intimately" with their data, they would improve statistical errors and carry out more nuanced interpretations of their results. The future of complex network research lies at the intersection of the two modes of studying science. Only by supplementing computational and algorithmic techniques with an interpretive approach, will quantitative studies of science be able to reconstruct the narrative fragments that lie behind the patterns of interactions, the topology, and structure of a scientific collaboration network.

CHAPTER 8

Conclusion

This dissertation is a study of collaboration in the Center for Embedded Networked Sensing (CENS), a National Science Foundation Science and Technology Center involved in sensor network research. CENS is a collaboratory embedded in a larger cyberinfrastructure; it is a modern, multi-disciplinary, distributed laboratory. In this research, I have examined its collaboration patterns by using network analytic methods and studying its collaborative ecology in terms of three networks of interaction: coauthorship of scholarly publications, communication on mailing list platforms, and interpersonal acquaintanceship. Results from a topological analysis of these networks indicate that *(i)* acquaintanceship patterns at CENS diffuse beyond coauthorship circles; *(ii)* average path length is low: only two steps are needed to connect any two individuals in the coauthorship, communication, and acquaintanceship networks; *(iii)* all collaboration networks have low clustering coefficient. My structural analysis indicates that *(iv)* individuals who are part of the same coauthorship community are well acquainted with each other; *(v)* coauthorship and acquaintanceship communities are mono-institutional and mono-disciplinary. Looking at preferential attachment mechanisms in the CENS collaboration networks, I find that *(vi)* researchers' network centrality has no direct influence on any collaboration pattern; *(vii)* on scholarly coauthorship and in interpersonal relationships, researchers tend to connect

preferentially with others within their own institution and department. My evolutionary analysis reveals that *(viii)* coauthorship and acquaintanceship networks become both more intra-institutional and more inter-disciplinary over time; *(ix)* researchers indicate to have known their coauthors prior to the beginning of their coauthoring relationship; *(x)* the communication network built from mailing list data is not affected by any preferential attachment patterns. In the previous chapter, I discussed these results in the context of related literature on collaboratories and cyberinfrastructure initiatives. This chapter draws this dissertation to a close by discussing the strengths and weaknesses of this study, and sketching out possible avenues for future research.

8.1 Limitations and assets of this study

While the research presented in this dissertation successfully elucidates the dynamics and the configuration of scientific collaboration at CENS, and its repercussions for modern collaboratory research, my methodological approach is not without its limitations. I identify here three important shortcomings of my research. This list is not intended to be exhaustive. Also, these shortcomings are not only intended to highlight potential fallacies of this study; they are also intended as clues on how this and related studies may be improved and extended in future work, as detailed in the next section.

The boundary problem. As alluded to in many parts of this dissertation, it is unavoidable for studies of science not to encounter problems of boundary definition. My efforts to address this issue only scratch the surface of this problem. Collaboration in modern scientific and scholarly endeavors permeates national, institutional, and disciplinary boundaries. While a number of methods

can be employed to define the boundaries of a scientific environment—the people, artifacts, relationships, practices that are to be included in the research—it is important to remember that regardless of the method used and its efficiency, the choice of a method and the resulting sample greatly affect the outcome of a study. In this dissertation, I identify the population of interest by including all individuals indicated as authors on scholarly items submitted to the annual reports, the official reporting documents published by CENS every year. This list is then supplemented with individuals found in mailing list logs that are not part of the coauthorship population, and used as the roster for the social network survey. Thus, I use exclusively recorded interactions in a bibliographic record and a mailing list log to delineate the population under study. As such, my methods fail to include individuals that are possibly directly involved in CENS research, but whose work does not appear in scholarly and communication interactions. This group might include members involved with administrative work, knowledge transfer and technical support groups. These individuals are physically present and actively involved in the CENS collaboratory environment. Despite being absent from bibliographic and mailing list records, they may be instrumental to collaboration, in their pivotal role of social routers and catalysts in the collaboration ecology. In my work, I could have included these individuals in the population via a number of methods, e.g., via consulting the narratives present in the project descriptions of the CENS annual reports, via a census survey, via *in situ* monitoring, and similar participant observation techniques.

Undocumented interactions and characteristics. This research is based upon three manifestations of scientific collaboration: coauthorship of scholarly papers, communication activity on mailing lists, and interpersonal acquaintanceship. As such, it is strictly limited to manifestations of collaboration that are evident from analyses of these interactions. For example, by restricting my analysis

of coauthorship to scholarly articles only (journal articles, conference papers, and book chapters), I fail to analyze collaboration activities documented in posters. As mentioned in Chapter 3, § 3.2.1, posters might include in their author lists researchers who do not appear as authors in published articles. These include software developers and technical staff whose work is instrumental to scientific work and collaboration. Besides being limited to a set of interactions, this research is also limited to a given set of characteristics: for every individual in the population under study, I collect information about their academic affiliation, academic department, academic position, and country of origin. There are a number of other node-based characteristics that I could have included in this study, to provide a richer comparative analysis and interpretation of collaboration at CENS. Some examples are given below, in the next section, covering future work.

Lack of exhaustive interpretation. In this research, I interpret the findings of my network research using my knowledge of the CENS community, its history, and its socio-academic environment. Yet, my interpretation lacks the fine granularity of most ethnographic studies of science. This is because my research does not include a comprehensive investigation of certain organizational, political, and financial events that affect the shape and dynamics of the collaboration networks. My network analysis results would have benefited enormously from a parallel ethnographic study, in the format of personal interviews, focus groups, and/or in-depth content analyses. By simply asking members of the population to describe the network maps of scientific collaboration, or by manually inspecting the content of CENS Annual Reports—the project descriptions, the research goals, the funding and management plans—I could have produced more genuine and exhaustive interpretations of my results.

Besides being subject to the aforementioned limitations, the methodological framework developed and utilized in this study also yield clear advantages, when compared to similar approaches. While the strengths of this research have already been discussed in the previous chapter, I summarize here two important methodological assets of this study.

Small network analysis. A great deal of research on scientific collaboration is performed on large networks, constructed from large bibliographic datasets harvested from domain-based and institutional document repositories. These studies document the organization and growth of large-scale collaborations, such as astronomical surveys, international high-energy physics experiments, and similar "Big Science" endeavors. Because of this, there is a tendency to think that network analyses are especially useful to investigate the dynamics of large-scale collaborative efforts only. This research demonstrates that the opposite is also true: network analysis is a convenient platform to document and examine "little sciences" and research collaborations that operate at small scales and levels. Moreover, I have argued that working with small-scale networks presents many benefits. For example, studies of scientific collaboration based on networks of manageable size can be easily complemented and elucidated by a number of ethnographic and qualitative methods, rooted in social science research. The relatively small size of the CENS coauthorship network allowed me to run a social network survey. Clearly, the use of similar social survey methods would be unfeasible for Big Science collaborations that involve tens of thousands of researchers.

Embedded social network analysis. As noted in the previous chapter, many large-scale investigations of scientific collaboration rely on great quantities of data to study the structure, evolution and similar macroscopic features of scientific collaboration patterns, but often ignore certain contextual and microscopic

factors, such as the social and academic arrangements in which collaboration takes place. This is because the bibliographic datasets upon which these studies are conducted contain detailed publication metadata, but very little or poorly structured data about the authors writing those publications. In other words, these datasets contain a lot of information about the links between the nodes, but no information about the nodes themselves. Thanks to the manageable size of the networks studied in this dissertation, I was able to research the personal web pages, biographies, and curricula of the individuals in the population and collect additional information about a number of socio-academic characterstics. Working with this level of data granularity not only allowed me to extend and compare the results of my network analysis to the disciplinary and institutional arrangements of CENS. I also had advantage of being able to directly manipulate and validate my data. Manual techniques of data processing and cleaning allow researchers to gain a deeper understanding of the networks they study. Moreover, being personally affiliated with CENS, I was myself a node embedded in the networks I studied. This privileged position allowed me to gain in-depth knowledge about the minutiae of scientific collaboration at CENS, and acquire familiarity with the narratives behind certain collaboration traces. This extensive knowledge of the CENS human infrastructure was instrumental for the interpretation of my results.

8.2 Future work

An obvious avenue for potential future research is to deal with the limitations outlined in the previous section. The second limitation discussed above, in particular, points to many possibilities for future research. For my dissertation

research, I could only collect data relative to three interactions. What other collaborative activities can be documented to complement and inform the study of scientific collaboration? The list is potentially endless. Some examples include citation patterns, epistemic connection, collaborative software coding, and private email communication. Citation networks can be constructed in which authors are connected to each other based on the authors they cite in their scholarly work. Epistemic connection and co-word networks can also be constructed to depict the intellectual connection among scientists based on the full text of their scholarly work. In a technology-driven environment like CENS, collaboration is not limited to traditional scholarly artifacts, however: much work goes into the development of software to run and collect data from sensor devices. Many software coding projects at CENS are collaborative and are handled using version control systems. Gathering and analyzing these data could provide insights into software coding collaborative practices. Finally, private email exchange is, naturally, the most revealing form of communication, but also the hardest to obtain, because of privacy and confidentiality issues. An analysis of email traces could expose very accurately the effect of email communication on the propagation of scientific ideas and scientific collaboration.

A similar direction for future research can be postulated with regard to undocumented node characteristics. For this research, I only had resources and time to collect four characterstics for each node in the collaboration network: academic affiliation, academic department, academic position, and country of origin. This collection of characterstics was crucial for understanding the academic arrangement of CENS researchers in their collaborative activities. But there are many more social, academic, and geographical characteristics that would have improved my study greatly. One of them is the geographical position of researchers—the exact location of their workplace. Although a partial analysis of this kind is

presented in Chapter 6 (§ 6.7), it is limited to Boelter Hall only, and as such it fails to provide a comprehensive scenario of the influence of physical distance on the processes of collaboration. Gender, ethnicity, age, and background of the researchers are examples of other characteristics that could be included in this, or related studies, to examine the impact of socio-cultural attributes on scientific work.

From the above discussion, it becomes clear that studies of scientific networks would benefit enormously from richly annotated datasets. Network researchers that avail themselves of these data have the potential to provide a more nuanced interpretation of their results and frame them in the context of the environment in which the networks operate. The major barrier to constructing these rich networks, however, is acquiring the data in the first place: collecting and making sense of these data can be a tedious and expensive task. Many network researchers are turning to the Internet, constructing networks from openly available data collected from the web. Much recent research on scientific collaboration, for example, is performed on networks extracted from digital library collections and online repositories. Similarly, much contemporary Internet research employs data harvested from social networking sites, such as data about online friendship and communication patterns. There is immense value in aggregating, correlating, and making sense of these and related online data. Not only would these data enable a new trend of studies of science, surveying scientific collaboration and knowledge production over multiple sites and platforms. It would also serve back scientific communities by providing them with personalized scientific recommendation services. Constructing such an open distributed platform for the study of scientific work is the mission of the next generation of information scientists working with cyberinfrastructure initiatives.

APPENDIX A

Appendix

A.1 Survey invitation letter

Dear [full name],

My name is Alberto Pepe. I am a member of the Statistics and Data Practices Team at the Center for Embedded Networked Sensing (CENS). I would like to invite you to take part in an online survey.

The aim of this survey is to collect acquaintanceship data for the purpose of modeling a social network of scientific collaboration. This is part of my doctoral research at UCLA. Your input would be greatly appreciated.

To fill in the online survey, please click on the following link:

`http://www.lecs.cs.ucla.edu/ apepe/form.php?id=[id]`

Completing the survey should take between 5 and 20 minutes of your time. If you prefer to take this survey on paper or through a personal interview, please contact me at the address below.

Thank you for your time.

Alberto Pepe

A.2 Text of the informed consent form

You are asked to participate in a research study conducted by Alberto Pepe, M.Sc., and Christine Borgman, Ph.D., from the Department of Information Studies at the University of California, Los Angeles. You were selected as a possible participant in this study because you appear as a participant in current or past projects at the Center for Embedded Networked Sensing (CENS). Your participation in this research study is voluntary.

The aim of this project is to study historical evolution and topological structure of selected small-scale interactions that mark the process of scientific knowledge production within the CENS collaboratory. In particular, the present survey measures acquaintanceship interactions, to mine true social contacts of personal, electronic and formal interaction.

If you volunteer to participate in this study, we would ask you to indicate individuals at CENS that you are acquainted with and to briefly describe your relationship to them.

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law. In particular, the data you provide in this study will only be used to investigate high-level patterns of relationships among acquaintances, such as structural differences among the acquaintanceship and other related networks. This will result in the publication of unidentifiable data (numerical codes) in the form of tables, charts and network diagrams. Confidential and identifiable data will not be published and will be destroyed upon termination of this dissertation research.

You can choose whether to be in this study or not. If you volunteer to be in

this study, you may exit the online questionnaire or interview at any time without consequences of any kind. You will receive no payment for your participation.

If you have any questions or concerns about the research, please feel free to contact:

Alberto Pepe, 3551 Boelter Hall, UCLA, apepe@ucla.edu

or

Christine Borgman, 235 GSE&IS Building, borgman@gseis.ucla.edu

If you have questions regarding your rights as a research subject, contact the Office for Protection of Research Subjects, UCLA, 11000 Kinross Avenue, Suite 102, Box 951694, Los Angeles, CA 90095-1694, (310) 825-8714.

A.3 Online survey instrument

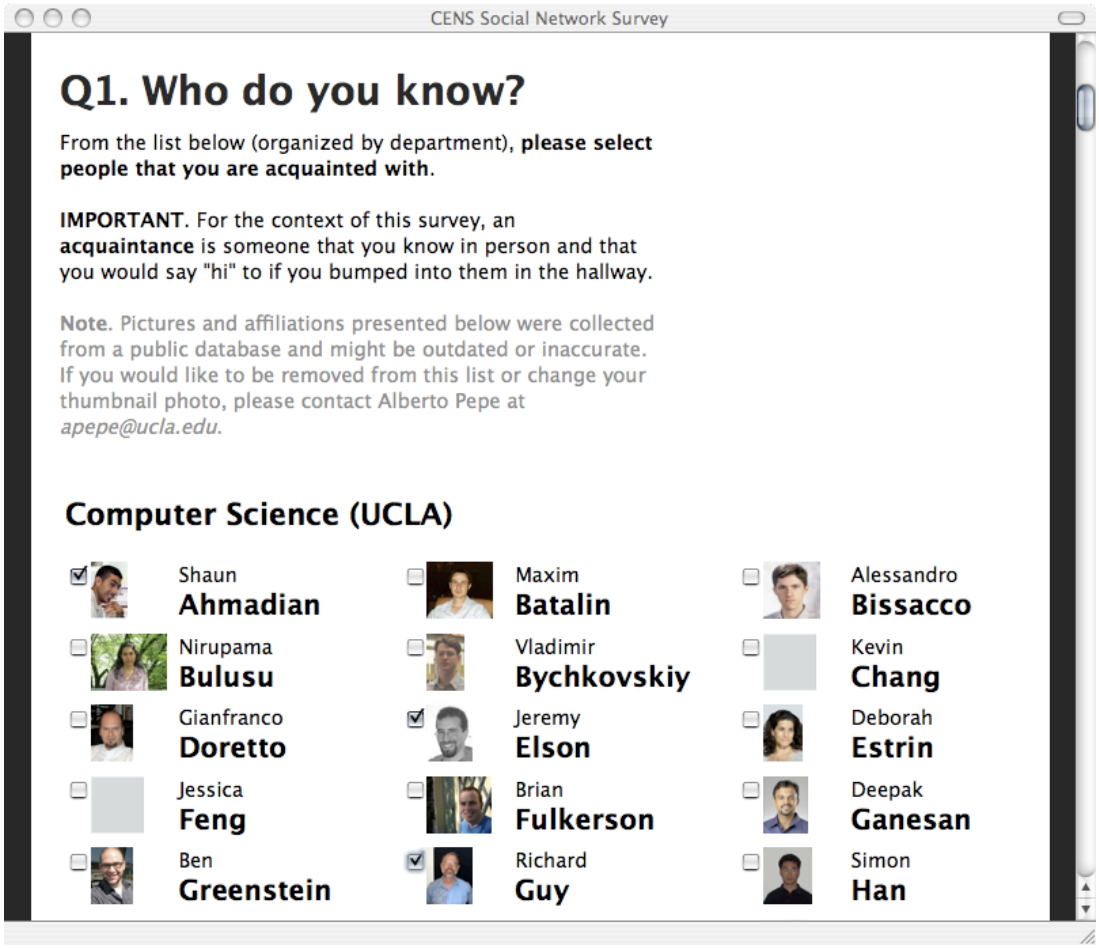


Figure A.1: Screenshot of the first page of the survey instrument.

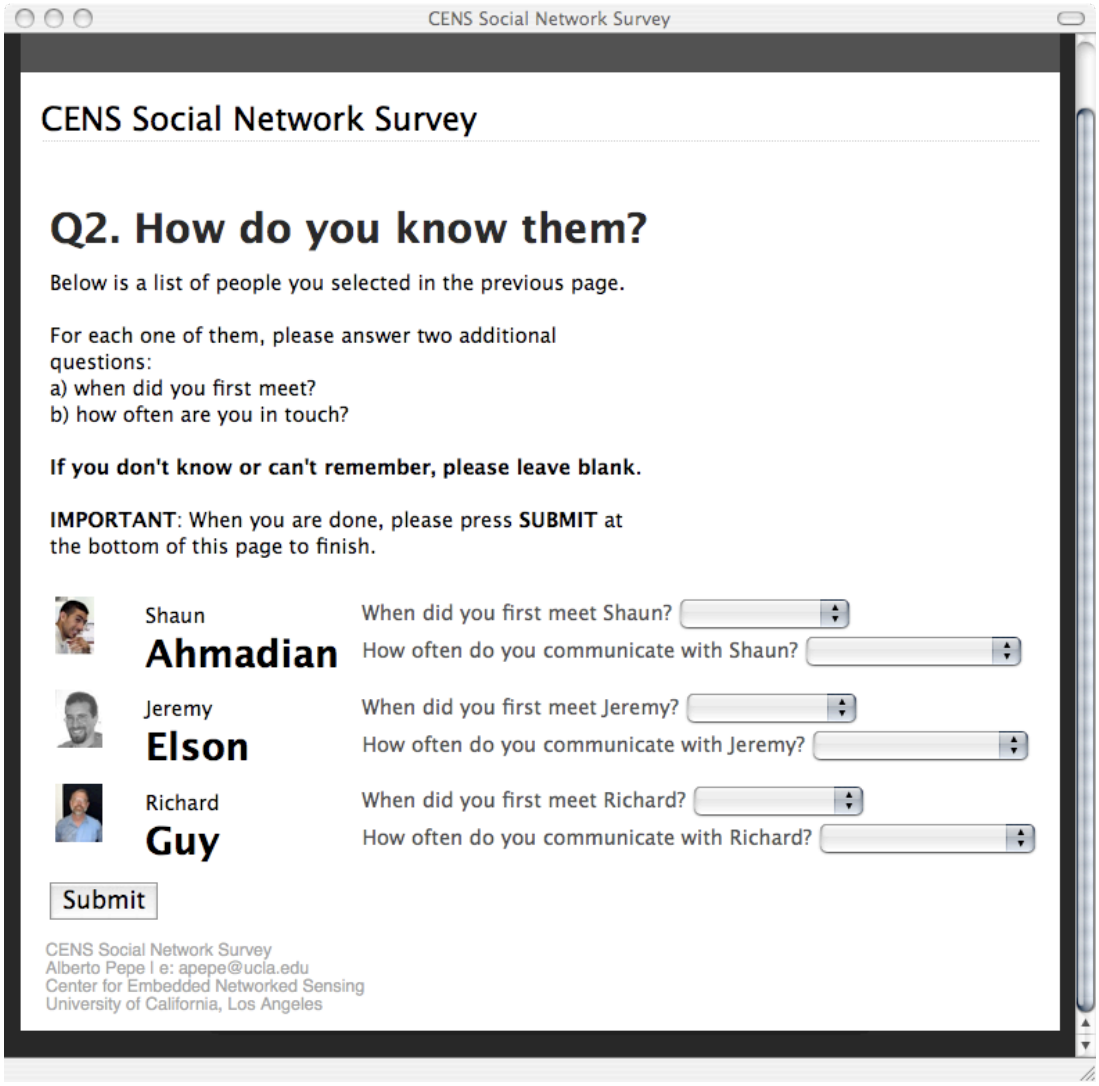


Figure A.2: Screenshot of the second page of the survey instrument.

A.4 Comparative network data

	Network	n	m	ℓ	C	r	Ref
bibliographic	Biology	1 520 251	11 803 064	4.92	0.60	0.127	[132, 133]
	Mathematics	253 339	496 489	7.57	0.34	0.120	[111, 174]
	Neuroscience	209 293		5.7	0.76		[61]
	Physics	52 909	245 300	6.19	0.56	0.363	[132, 133]
	Astronomy	16 706		4.66	0.41		[157]
	Computer Science	11 994		9.7	0.49		[157]
	Digital library research	1 567	3 401	6.6	0.89		[53]
	CENS coauthorship	**391**	**1747**	**2.950**	**0.31**	**0.013**	—
comm	email messages	59 912	86 300	4.95	0.16		[172]
	email address books	16 881	57 029	5.22	0.13	0.092	[181]
	peer-to-peer	880	1 296	4.28	0.011	−0.366	[173]
	CENS communication	**119**	**994**	**2.095**	**0.461**	**-0.057**	—
social	Hollywood actors	449 913	25 516 482	3.48	0.78	0.208	[115]
	company directors	7 673	55 392	4.60	0.88	0.276	[171]
	CENS acquaintance	**385**	**4 805**	**2.427**	**0.359**	**-0.060**	—

Columns display type of network, total number of nodes n; total number of edges m; average path length ℓ; clustering coefficient C; and degree assortativity coefficient, r. The last column gives the citation(s) for the network in the bibliography. Entries ordered by descending size (number of nodes). Blank entries indicate unavailable data. Entries corresponding to the CENS coauthorship, communication, and acquaintanceship networks are highlighted, in bold typeface.

Table A.1: Basic statistics for a number of published bibliographic, communication, and social networks.

A.5 Description of software code and tools

In order to perform the data collection, manipulations, and analysis presented in this dissertation, I made extensive use of two programming languages: Python and R.

Python (`http://www.python.org/`) is a free, open source high-level programming language. I used Python to perform all the data collection and manipulation processes described in this dissertation. For example, I developed Python scripts to convert the author lists contained in the bibliographic database, initially harvested in BibTEX format, to a graph-based format (`ncol`). The entire workflow of network construction and edge weighting, described in detail in § 4.1 and subsequent sections, was implemented using custom Python scripts. All the other processes of data collection, filtering, and manipulation were also implemented using Python: from the extraction of the threaded structure from mailing list logs, to the conversion of socio-academic properties (academic affiliation, department, etc.) to node-based properties in a graph format. The only portion of data collection that was not performed using Python is the social survey, that was developed from scratch using simple HTML with a PHP backend. The Python scripts and HTML code are too long to be included here. However, they are available upon request, and I plan to bundle them in a suite for social survey research and network analysis that I will include as part of the supplemental material when the text of this dissertation is made available online.

R (`http://www.r-project.org/`) is an open source environment for statistical computing and graphics, available under the GNU General Public License. I use the R environment to perform all the network analysis work presented in this dissertation. In particular, I use the `igraph` library of R, a package for creating, manipulating, and visualizing networks. The `igraph` library includes a number

of pre-implemented functions for graph theory problems. These include functions to calculate network topology (clustering coefficient, average path length, cliques, diameter). The library also includes functions to detect community structure of a network. In this dissertation, I used spinglass (`spinglass.community`) and eigenvector (`leading.eigenvector.community`) algorithms of community detection. The `igraph` library lacks, however, functions for some network theory calculations that are used in this dissertation, especially those involving node-based socio-academic information. I implemented these missing functions from scratch and they will be contributed to the `igraph` library. Some sample R code that I developed to implement assortativity measures (degree assortativity and discrete assortativity) is included below.

```
degree.assortativity <- function(graph){
   deg <- degree(graph)
   deg.sq <- deg^2
   m <- ecount(graph)
   num1 <- 0
   num2 <- 0
   den <- 0
   x <- NULL
   y <- NULL
   edges <- get.edgelist(graph)

   for (i in 1:m) {
     x <- append(x,deg[V(graph)$name==edges[,1][i]])
   }

   for (i in 1:m) {
     y <- append(y,deg[V(graph)$name==edges[,2][i]])
   }

   num1 <- sum (x*y) / m
   num2 <- (sum((x+y)/2) / m)^2
   den <- sum((x^2 + y^2)/2) / m

   return((num1-num2)/(den-num2))
   }
```

R code to compute degree assortativity of a graph, via eq. A.1

$$r = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}, \tag{A.1}$$

```
discrete.assortativity <- function(filename){
        tab <- read.csv(filename, header = FALSE, sep = " ")

        if (length(levels(tab$V1))>length(levels(tab$V2))){
          tab <- table(factor(tab$V1,levels(tab$V1)),
                       factor(tab$V2,levels(tab$V1)))
        }
        else if (length(levels(tab$V2))>length(levels(tab$V1))){
          tab <- table(factor(tab$V1,levels(tab$V2)),
                       factor(tab$V2,levels(tab$V2)))
        }
        else {
          tab <- table(tab$V1,tab$V2)
        }

        ptab <- prop.table(tab)
        num1 <- 0
        num2 <- 0

        for (i in 1:length(ptab[,1])) {
          num1<-num1+ptab[i,i]
        }

        num2 <- sum(margin.table(ptab,1)*margin.table(ptab,2))

    return((num1-num2)/(1-num2))
    }
```

R code to compute discrete assortativity of an edge list, via eq. A.2

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} \tag{A.2}$$

A.6 Seating chart of 3551 Boelter Hall

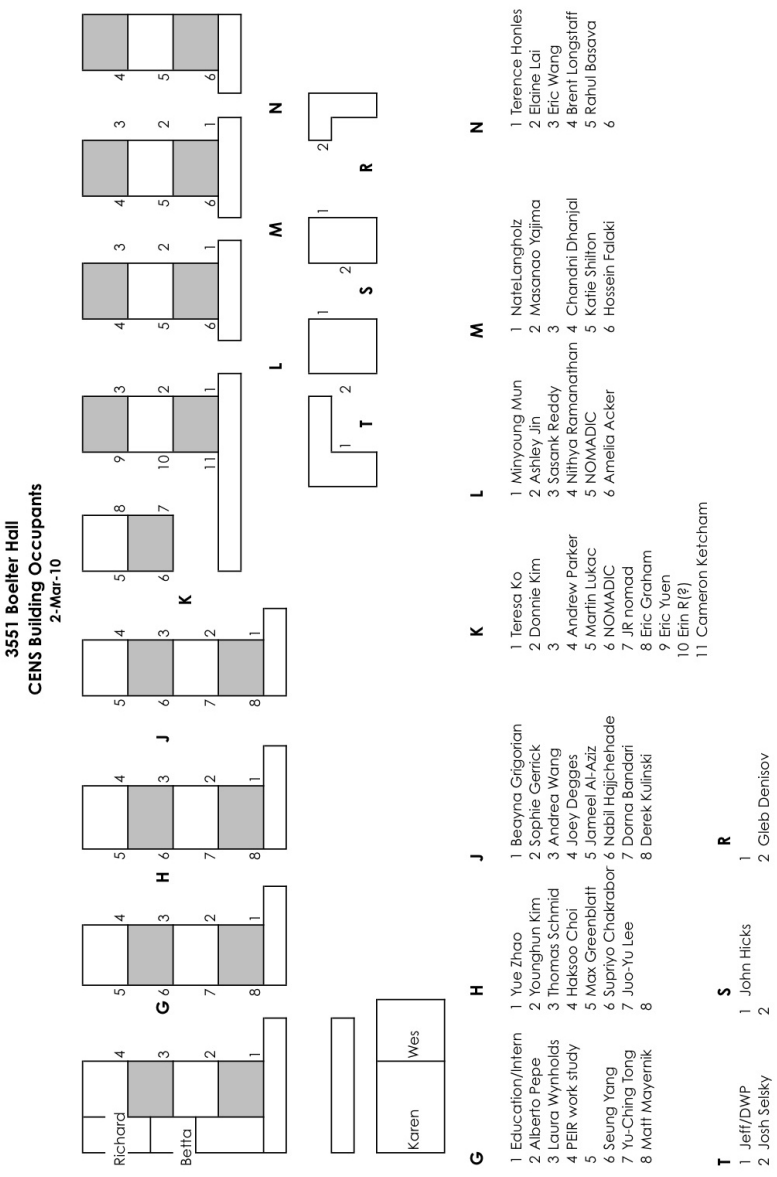


Figure A.3: Seating chart of the CENS headquarters (3551 Boelter Hall, UCLA) as of March 2010.

References

[1] K. Mees, "The production of scientific knowledge," *Nature*, vol. 100, pp. 355–358, 1918.

[2] D. J. Hess, *Science Studies. An advanced introduction.* New York University Press, 1997.

[3] K. Dunbar, "Scientific thinking and reasoning," in *The Cambridge Handbook of Thinking and Reasoning* (K. J. Holyoak and R. Morrison, eds.), Cambridge University Press, 2005.

[4] P. C. Wason, "Reasoning about a rule," *The Quarterly Journal of Experimental Psychology*, vol. 12, pp. 273–281, 1968.

[5] D. J. Schiano, L. A. Cooper, R. Glaser, and H. C. Zhang, "Highs are to lows as experts are to novices: Individual differences in the representation and solution of standardized figural analogies," *Human Performance*, vol. 2, no. 4, pp. 225–248, 1989.

[6] D. K. Simonton, *Creativity in Science: Chance, Logic, Genius, and Zeitgeist.* Cambridge, UK: Cambridge University Press, 2004.

[7] B. Latour and S. Woolgar, *Laboratory Life: the social construction of scientific facts.* Sage Publications, Ltd., 1979.

[8] W. Wulf, "The collaboratory opportunity," *Science*, vol. 261, no. 5123, pp. 854–855, 1993.

[9] T. A. Finholt, "Collaboratories.," in *Annual Review of Information Science & Technology* (B. Cronin, ed.), vol. 36, pp. 73–107, Information Today, 2002.

[10] C. L. Borgman, *Scholarship in the digital age: Information, Infrastructure, and the Internet.* Cambridge, MA: MIT Press, 2007.

[11] N. C. Council, "Cyberinfrastructure vision for 21st century discovery," tech. rep., National Science Foundation, 2007. URL: www.nsf.gov/od/oci/ci_v5.pdf (last visited July 25, 2008).

[12] S. L. Star and K. Ruhleder, "Steps towards an ecology of infrastructure: Complex problems in design and access for large-scale collaborative systems," in *Proceedings of the 1994 ACM conference on Computer supported cooperative work* (J. B. Smith, F. D. Smith, and T. W. Malone, eds.), pp. 253–264, ACM Press, 1994.

[13] G. C. Bowker, "Information mythology and infrastructure," in *Information Acumen: The Understanding and Use of Knowledge in Modern Business* (L. Bud, ed.), pp. 231–247, London: Routledge, 1994.

[14] C. Lee, P. Dourish, and G. Mark, "The human infrastructure of cyberinfrastructure," in *Proceedings of CSCW: Conference on Computer-Supported Cooperative Work*, pp. 483–492, ACM, 2006.

[15] J. N. Cummings and S. Kiesler, "Collaborative research across disciplinary and organizational boundaries," *Social Studies of Science*, vol. 35, no. 5, pp. 703–722, 2005.

[16] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.

[17] N. Luhmann, *Social Systems*. Stanford University Press, 1995.

[18] C. Fuchs and W. Hofkirchner, "The dialectic of bottom-up and top-down emergence in social systems," *tripleC (Cognition, Communication, Cooperation)*, vol. 1, no. 1, pp. 28–50, 2005.

[19] L. A. Amaral and J. M. Ottino, "Complex networks: Augmenting the framework for the study of complex systems," *The European Physical Journal B*, vol. 38, no. 2, pp. 147–162, 2004.

[20] M. Gell-Mann, "What is complexity?," *Complexity*, vol. 1, no. 1, 1995.

[21] K. A. Richardson and P. Cilliers, eds., *Emergence. Special issue: What is Complexity Science?*, vol. 3, no. 1. ISCE Publishing, 2001.

[22] J. Goldstein, "Emergence as a construct: History and issues," *Emergence*, vol. 1, no. 1, pp. 49–72, 1999.

[23] F. Heylighen, "Self-organization, emergence and the architecture of complexity," in *Proceedings of the European Congress on System Science,*, (Paris), pp. 23–32, 1989.

[24] S. J. Jackson, P. N. Edwards, G. C. Bowker, and C. P. Knobel, "Understanding infrastructure: History, heuristics, and cyberinfrastructure policy," *First Monday*, vol. 12, no. 6, 2007.

[25] D. Ribes and T. A. Finholt, "Tensions across the scales: planning infrastructure for the long-term," in *GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work*, (New York, NY, USA), pp. 229–238, ACM, 2007.

[26] P. Cilliers, "Boundaries, hierarchies and networks in complex systems," *International Journal of Innovation Management*, vol. 5, no. 2, pp. 135–147, 2001.

[27] F. Capra, *The Hidden Connections: Integrating The Biological, Cognitive, And Social Dimensions Of Life Into A Science Of Sustainability*. Doubleday, 2002.

[28] S. L. Star and J. R. Griesemer, "Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in berkeley's museum of vertebrate zoology, 1907-39," *Social Studies of Science*, vol. 19, no. 3, pp. 387–420, 1989.

[29] G. C. Bowker and S. L. Star, *Sorting Things Out: Classification and Its Consequences (Inside Technology)*. The MIT Press, 1999.

[30] C. Lee, "Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing chaos in collaborative work," *Journal of Computer Supported Cooperative Work*, vol. 16, pp. 307–339, 06 2007.

[31] F. Chung, L. Lu, T. G. Dewey, and D. J. Galas, "Duplication models for biological networks," *Journal of Computational Biology*, vol. 10, no. 5, pp. 677–687, 2003.

[32] G. Demange and M. Wooders, *Group Formation in Economics: Networks, Clubs, and Coalitions*. Cambridge University Press, 2005.

[33] B. Latour, *Reassembling the Social: An Introduction to Actor-network-theory (Clarendon Lectures in Management Studies)*. Oxford University Press, 2005.

[34] S. P. Borgatti and P. C. Foster, "The Network Paradigm in Organizational Research: A Review and Typology," *Journal of Management*, vol. 29, no. 6, pp. 991–1013, 2003.

[35] S. J. Hanson and M. Negishi, "On the Emergence of Rules in Neural Networks," *Neural Computation*, vol. 14, no. 9, pp. 2245–2268, 2002.

[36] S. Wasserman and K. Faust, *Social network analysis*. Cambridge: Cambridge University Press, 1994.

[37] M. Castells, *The Rise of the Network Society*. Blackwell Publishing, 1996.

[38] B. Wellman and C. Haythornthweait, *The Internet in Everyday Life*. Oxford: Blackwell, 2002.

[39] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, pp. 1360–1380, 1973.

[40] C. L. Borgman, J. C. Wallis, M. Mayernik, and A. Pepe, "Drowning in data: Digital library architecture to support scientific use of embedded sensor networks," in *Proceedings of the 7th Joint Conference on Digital Libraries*, pp. 269–277, ACM, 2007.

[41] J. C. Wallis, C. L. Borgman, M. Mayernik, A. Pepe, N. Ramanathan, and M. Hansen, "Know thy sensor: Cens as a case study of the relationship between data integrity, metadata, and data interpretation," in *Proceedings of 11th European Conference on Research and Advanced Technology for Digital Libraries*, vol. 4675, Springer, 2007.

[42] J. C. Wallis, C. L. Borgman, M. Mayernik, and A. Pepe, "Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative field research," *International Journal of Digital Curation*, vol. 3, no. 1, 2008.

[43] A. Pepe, C. L. Borgman, J. C. Wallis, and M. Mayernik, "Knitting a fabric of sensor data resources," in *Proceedings of the ACM IEEE International Conference on Information Processing in Sensor Networks*, 2007.

[44] A. Pepe, M. Mayernik, C. L. Borgman, and H. Van de Sompel, "From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 3, pp. 567–582, 2010.

[45] T. Ko, Z. Charbiwala, S. Ahmadian, M. Rahimi, M. Srivastava, S. Soatto, and D. Estrin, "Exploring tradeoffs in accuracy, energy and latency of scale invariant feature transform in wireless camera networks," in *ICDSC: Proceedings of the first ACM/IEEE International Conference on Distributed Smart Cameras*, pp. 313–320, 2007.

[46] S. Ahmadian, T. Ko, S. Coe, M. P. Hamilton, M. Rahimi, S. Soatto, and D. Estrin, "Heartbeat of a nest: Using imagers as biological sensors," *ACM Transactions on Sensor Networks*, vol. 6, no. 3, 2010.

[47] L. A. Lievrouw, E. M. Rogers, C. U. Lowe, and E. Nadel, "Triangulation as a research strategy for identifying invisible colleges among biomedical scientists," *Social Networks*, vol. 9, pp. 217–248, 1987.

[48] C. L. Borgman and J. Furner, "Scholarly communication and bibliometrics," *Annual Review of Information Science & Technology*, vol. 36, no. 1, pp. 2–72, 2002.

[49] S. Mele, D. Dallman, J. Vigen, and J. Yeomans, "Quantitative analysis of the publishing landscape in high-energy physics," *Journal of High Energy Physics*, vol. 12, 2006.

[50] M. Tomassini, L. Luthi, M. Giacobini, and W. B. Langdon, "The structure of the genetic programming collaboration network," *Genetic Programming and Evolvable Machines*, vol. 8, no. 1, pp. 97–103, 2007.

[51] T. Braun, W. Glanzel, and A. Schubert, "Publication and cooperation patterns of the authors of neuroscience journals," *Scientometrics*, vol. 51, no. 3, pp. 499–510, 2001.

[52] J. Schummer, "Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology," *Scientometrics*, vol. 59, no. 3, pp. 425–465, 2004.

[53] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel, "Co-authorship networks in the digital library research community," *Information Processing & Management*, vol. 41, no. 6, pp. 1462–1480, 2005.

[54] A. Hollis, "Co-authorship and the output of academic economists," *Labour Economics*, vol. 8, no. 28, pp. 503–530, 2001.

[55] F. Acedo, C. Barroso, C. Casanueva, and L. Galan, "Co-authorship in management and organizational studies: An empirical and network analysis," *Journal of Management Studies*, vol. 43, no. 5, pp. 957–983, 2006.

[56] B. Cronin, D. Shaw, and K. L. Barre, "A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy," *Journal of the American Society for Information Science & Technology*, vol. 54, no. 9, pp. 855–871, 2003.

[57] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration.," *Proceedings of the National Academy of Sciences*, vol. 101 Suppl 1, pp. 5200–5205, 2004.

[58] K. Börner, L. Dall'Asta, W. Ke, and A. Vespignani, "Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams: Research articles," *Complexity*, vol. 10, no. 4, pp. 57–67, 2005.

[59] K. Börner, C. Chen, and K. Boyack, "Visualizing knowledge domains," *Annual Review of Information Science & Technology*, vol. 37, no. 1, 2005.

[60] S. M. Douglas, G. T. Montelione, and M. Gerstein, "Pubnet: a flexible system for visualizing literature derived networks.," *Genome Biology*, vol. 6, no. 9, 2005.

[61] A. L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A*, vol. 311 (3-4), 2002.

[62] M. Catanzaro, G. Caldarelli, and L. Pietronero, "Social network growth with assortative mixing," *Physica A Statistical Mechanics and its Applications*, vol. 338, pp. 119–124, 2004.

[63] C. S. Wagner and L. Leydesdorff, "Network structure, self-organization, and the growth of international collaboration in science," *Research Policy*, vol. 34, no. 10, pp. 1608–1618, 2005.

[64] S. Liberman and K. B. Wolf, "Bonding number in scientific disciplines," *Social Networks*, vol. 20, no. 3, pp. 239 – 246, 1998.

[65] T. Fenner, M. Levene, and G. Loizou, "A model for collaboration networks giving rise to a power-law distribution with an exponential cutoff," *Social Networks*, vol. 29, no. 1, pp. 70 – 80, 2007.

[66] B. Cronin, "Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices?," *Journal of the American Society for Information Science & Technology*, vol. 52, no. 7, pp. 558–569, 2001.

[67] B. Cronin, *The Hand of Science*. Scarecrow Press, 2005.

[68] J. Solomon, "Programmers, professors, and parasites: Credit and co-authorship in computer science," *Science and Engineering Ethics*, vol. 15, pp. 467–489, 12 2009.

[69] T. Braun and A. Schubert, "The growth of research on inter-and multidisciplinarity in science and social science papers, 1975-2006," *Scientometrics*, vol. 73, no. 3, pp. 345–351, 2007.

[70] S. Traweek, *Beamtimes and lifetimes: The world of high energy physicists.* Cambridge, MA: Harvard University Press, 1992.

[71] E. Tarnow, "Coauthorship in physics," *Science and Engineering Ethics*, vol. 8, no. 2, pp. 175–190, 2002.

[72] J. P. Birnholtz, "What does it mean to be an author? The intersection of credit, contribution, and collaboration in science," *Journal of the American Society for Information Science & Technology*, vol. 57, no. 13, pp. 1758–1770, 2006.

[73] A. Y. Chua and C. C. Yang, "The shift towards multi-disciplinarity in information science," *Journal of the American Society for Information Science & Technology*, vol. 59, no. 13, pp. 2156–2170, 2008.

[74] F. Havemann, M. Heinz, and H. Kretschmer, "Collaboration and distances between german immunological institutes - a trend analysis," *Journal of Biomedical Discovery and Collaboration*, vol. 1, no. 1, p. 6, 2006.

[75] J. Moody, "The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999," *American Sociological Review*, vol. 69, no. 2, pp. 213–238, 2004.

[76] L. Lorigo and F. Pellacini, "Frequency and structure of long distance scholarly collaborations in a physics community," *Journal of the American Society for Information Science & Technology*, vol. 58, no. 10, pp. 1497–1502, 2007.

[77] A. Pepe and M. A. Rodriguez, "Collaboration in sensor network research: an in-depth longitudinal analysis of assortative mixing patterns," *Scientometrics*, 2010 (In press).

[78] M. A. Rodriguez and A. Pepe, "On the relationship between the structural and socioacademic communities of a coauthorship network," *Journal of Informetrics*, vol. 2, no. 3, pp. 195–201, 2008.

[79] A. Pepe, "Socio-epistemic analysis of scientific knowledge production in little science research," *tripleC (Cognition, Communication, Co-operation)*, vol. 6, no. 2, pp. 134–145, 2009.

[80] M. Callon, J. Law, and A. Rip, "Qualitative scientometrics," in *Mapping the dynamics of science and technology* (M. Callon, J. Law, and A. Rip, eds.), London: Macmillan., 1986.

[81] L. Leydesdorff, "In search of epistemic networks," *Social Studies of Science*, vol. 21, no. 1, pp. 75–110, 1991.

[82] L. M. Rocha, "Semi-metric behavior in document networks and its application to recommendation systems," in *Soft Computing Agents: A New Perspective for Dynamic Information Systems* (V. Loia, ed.), IOS Press, 2002.

[83] C. Chen, T. Cribbin, R. Macredie, and S. Morar, "Visualizing and tracking the growth of competing paradigms: Two case studies," *Journal of the American Society for Information Science & Technology*, vol. 53, no. 8, pp. 678–689, 2002.

[84] R. K. Buter, E. C. M. Noyons, and A. F. J. V. Raan, "A combination of quantitative and qualitative maps in an evaluative bibliometric context," in *IV '04: Proceedings of the Information Visualisation, Eighth International Conference*, (Washington, DC, USA), pp. 978–982, IEEE Computer Society, 2004.

[85] F. B. Viégas, S. Golder, and J. Donath, "Visualizing email content: portraying relationships from conversational histories," in *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, (New York, NY, USA), pp. 979–988, ACM, 2006.

[86] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," 1972.

[87] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[88] H. T. Welser, E. Gleave, D. Fisher, and M. Smith, "Visualizing the signatures of social roles in online discussion groups," *Journal of Social Structure*, vol. 8, no. 2, 2007.

[89] M. F. Schwartz and D. C. M. Wood, "Discovering shared interests using graph analysis," *Communications of the ACM*, vol. 36, no. 8, pp. 78–89, 1993.

[90] B. Wellman, A. Q. Haase, J. Witte, and K. Hampton, "Does the Internet Increase, Decrease, or Supplement Social Capital?: Social Networks, Participation, and Community Commitment," *American Behavioral Scientist*, vol. 45, no. 3, pp. 436–455, 2001.

[91] B. Wellman, "Little boxes, glocalization, and networked individualism," *Digital Cities II: Computational and Sociological Approaches*, pp. 337–343, 2002.

[92] E. Davenport and H. Hall, "Organizational knowledge and communities of practice.," *Annual Review of Information Science & Technology*, vol. 36, pp. 171–227, 2002.

[93] M. A. Smith, "Invisible crowds in cyberspace," in *Communities in Cyberspace* (P. Kollock, ed.), pp. 195–218, Routledge, 1998.

[94] B. Burkhalter, "Reading race online," in *Communities in Cyberspace* (P. Kollock, ed.), ch. 3, Routledge, 1998.

[95] A. T. Fiore, S. L. Tiernan, and M. A. Smith, "Observed behavior and perceived value of authors in usenet newsgroups: bridging the gap," in *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, (New York, NY, USA), pp. 323–330, ACM, 2002.

[96] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "Grouplens: applying collaborative filtering to usenet news," *Communications of the ACM*, vol. 40, no. 3, pp. 77–87, 1997.

[97] T. C. Turner, M. A. Smith, D. Fisher, and H. T. Welser, "Picturing usenet: Mapping computer-mediated collective action," *Journal of Computer-Mediated Communication*, vol. 10, no. 4, 2005.

[98] D. Boyd, H. Lee, D. Ramage, and J. Donath, "Developing legible visualizations for online social spaces," in *Proceedings of the 35th Hawaii International Conference on System Sciences*, 2002.

[99] W. Sack, "Conversation map: a content-based usenet newsgroup browser," in *IUI '00: Proceedings of the 5th international conference on Intelligent user interfaces*, (New York, NY, USA), pp. 233–240, ACM, 2000.

[100] C. B. White, C. A. Moyer, D. T. Stern, and S. J. Katz, "A content analysis of e-mail communication between patients and their providers: patients get the message," *Journal of the American Medical Informatics Association*, vol. 11, no. 4, pp. 260–267, 2004.

[101] A. M. Young, "A sign of the times: An analysis of organizational members email signatures," *Journal of Computer-Mediated Communication*, vol. 11, pp. 1046–1061(16), 2006.

[102] U. Matzat, "Academic communication and internet discussion groups: transfer of information or creation of social contacts?," *Social Networks*, vol. 26, no. 3, pp. 221–255, 2004.

[103] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, "Email as spectroscopy: automated discovery of community structure within organizations," *The Information Society*, vol. 21, no. 2, pp. 143–153, 2005.

[104] P. D. Killworth, H. R. Bernard, C. McCarty, P. Doreian, S. Goldenberg, C. Underwood, P. Harries-Jones, R. M. Keesing, J. Skvoretz, and M. V. S. Wemegah, "Measuring patterns of acquaintanceship [and comments and reply]," *Current Anthropology*, vol. 25, no. 4, pp. 381–397, 1984.

[105] H. Hoang and B. Antoncic, "Network-based research in entrepreneurship: A critical review," *Journal of Business Venturing*, vol. 18, no. 2, pp. 165–187, 2003.

[106] E. J. Aries and F. L. Johnson, "Close friendship in adulthood: Conversational content between same-sex friends," *Sex Roles*, vol. 9, no. 12, pp. 1183–1196, 1983.

[107] F. Lilijeros, C. Edling, L. Amaral, E. Stanley, and Y. Aberg, "The web of human sexual contacts," *Nature*, vol. 411, pp. 907–908, 2001.

[108] G. Plickert, R. R. Côté, and B. Wellman, "It's not who you know, it's how you know them: Who exchanges what with whom?," *Social Networks*, vol. 29, no. 3, pp. 405 –429, 2007.

[109] S. Milgram, "The small world problem," *Psychology Today*, vol. 2, pp. 60–67, 1967.

[110] J. Travers and S. Milgram, "An experimental study of the small world problem," *Sociometry*, vol. 32, no. 4, pp. 425–443, 1969.

[111] R. D. Castro and J. W. Grossman, "Famous Trails to Paul Erdös," *The Mathematical Intelligencer*, vol. 21, 1999.

[112] P. S. Dodds, R. Muhamad, and D. J. Watts, "An experimental study of search in global social networks," *Science*, vol. 301, no. 5634, pp. 827–829, 2003.

[113] P. D. Killworth and H. R. Bernard, "The reversal small-world experiment," *Social Networks*, vol. 1, no. 2, pp. 159–192, 1979.

[114] J. S. Kleinfeld, "Could it be a big world after all? The "six degrees of separation" myth," *Society*, 2002.

[115] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks.," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[116] J. Davidsen, H. Ebel, and S. Bornholdt, "Emergence of a small world from local interactions: Modeling acquaintance networks," *Physical Review Letters*, vol. 88, p. 128701, 2002.

[117] P. Holme, A. Trusina, B. J. Kim, and P. Minnhagen, "Prisoners' dilemma in real-world acquaintance networks," *Physical Review E*, vol. 68, no. 3, p. 030901, 2003.

[118] S. T. Tong, B. Van Der Heide, L. Langwell, and J. B. Walther, "Too much of a good thing? the relationship between number of friends and interpersonal impressions on Facebook," *Journal of Computer-Mediated Communication*, vol. 13, no. 3, pp. 531–549, 2008.

[119] N. B. Ellison, C. Steinfield, and C. Lampe, "The benefits of facebook "friends:" social capital and college students use of online social network sites," *Journal of Computer-Mediated Communication*, vol. 12, no. 4, pp. 1143–1168, 2007.

[120] P. Holme, "Structure and time evolution of an internet dating community," *Social Networks*, vol. 26, no. 2, pp. 155–174, 2004.

[121] L. C. Freeman, "The impact of computer based communication on the social structure of an emerging scientific specialty," *Social Networks*, pp. 201–221, 1984.

[122] R. E. Kraut, J. Galegherb, and C. Egidoa, "Relationships and tasks in scientific research collaboration," *Human-Computer Interaction*, vol. 3, no. 1, 1987.

[123] S. Liberman and K. B. Wolf, "The flow of knowledge: Scientific contacts in formal meetings," *Social Networks*, vol. 19, no. 3, pp. 271 –283, 1997.

[124] G. Chin, J. Myers, and D. Hoyt, "Social networks in the virtual science laboratory," *Communications of the ACM*, vol. 45, no. 8, pp. 87–92, 2002.

[125] N. Hara, P. Solomon, S.-L. Kim, and D. H. Sonnenwald, "An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration," *Journal of the American Society for Information Science & Technology*, vol. 54, no. 10, pp. 952–965, 2003.

[126] D. Stokols, R. Harvey, J. Gress, J. Fuqua, and K. Phillips, "In vivo studies of transdisciplinary scientific collaboration: Lessons learned and implications for active living research," *American Journal of Preventive Medicine*, vol. 28, no. 2, Supplement 2, pp. 202–213, 2005.

[127] B. A. Nardi, S. Whittaker, and H. Schwarz, "Networkers and their activity in intensionalnetworks," *Journal of Computer Supported Cooperative Work*, vol. 11, no. 1-2, pp. 205–242, 2002.

[128] Y. Engeström, R. Engeström, and T. Vähäaho, "When the center doesn't hold: The importance of knotworking.," in *Activity Theory and Social Practice* (S. Chaiklin, M. Hedegaard, and U. Jensen, eds.), Aarhus University Press, 1999.

[129] E. J. Webb, D. Campbell, R. Schwartz, L. Sechrest, and J. Grove, *Nonreactive Measures in the Social Sciences,*. Boston, MA: Houghton Mifflin, 2nd edition ed., 1981.

[130] D. Crane, "Social structure in a group of scientists: A test of the invisible college hypothesis," *American Sociological Review*, vol. 34, no. 3, pp. 335–352, 1969.

[131] D. Crane, *Invisible colleges. Diffusion of knowledge in scientific communities*. University of Chicago Press, 1972.

[132] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 404–409, 2001.

[133] M. E. J. Newman, "Scientific collaboration networks: I. network construction and fundamental results," *Physical Review E*, vol. 64, no. 1, p. 016131, 2001.

[134] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[135] D. J. de Solla Price, "Networks of scientific papers," *Science*, vol. 149, no. 3683, pp. 510–515, 1965.

[136] J. Lave and E. Wenger, *Situated learning. Legitimate peripheral participation.* Cambridge University Press, 1991.

[137] P. Haas, "Epistemic communities and international policy coordination," *International Organization*, vol. 46, no. 1, pp. 1–35, 1992.

[138] J. T. Alatta, "Structural analysis of communities of practice: an investigation of job title, location, and management intention," in *Communities and technologies*, pp. 23–42, Deventer, The Netherlands: Kluwer, B.V, 2003.

[139] C. Roth and P. Bourgine, "Epistemic communities: Description and hierarchic categorization," *Mathematical Population Studies*, vol. 12, pp. 107–130, 2005.

[140] M. Thompson, "Structural and epistemic parameters in communities of practice," *Organization Science*, vol. 16, no. 2, pp. 151–164, 2005.

[141] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability* (L. M. L. Cam and J. Neyman, eds.), vol. 1, pp. 281–297, University of California Press, 1967.

[142] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, p. 7821, 2002.

[143] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, 2006.

[144] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *Journal of Graph Algorithms and Applications*, vol. 10, no. 2, 2006.

[145] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical Review E*, vol. 74, no. 016110, 2006.

[146] P. Bonacich, "Power and centrality: A family of measures.," *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[147] S. Lozano, J. Duch, and A. Arenas, "Analysis of large social datasets by community detection," *The European Physical Journal Special Topics*, vol. 143, no. 1, pp. 257–259, 2007.

[148] M. A. Porter, P. J. Mucha, M. E. J. Newman, and A. J. Friend, "Community structure in the United States House of Representatives," *Physica A*, vol. 386, 2007.

[149] P. Gleiser and L. Danon, "Community structure in jazz," *Advances in Complex Systems*, vol. 6, p. 565, 2003.

[150] R. D. Smith, "The network of collaboration among rappers and its community structure," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2006, no. 2, p. P02006, 2006.

[151] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 9, 2005.

[152] M. McPherson, L. Smith-Lovin, and J. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, pp. 415–444, 2001.

[153] M. E. J. Newman, "Mixing patterns in networks," *Physical Review E*, vol. 67, no. 2, p. 026126, 2003.

[154] M. E. J. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, 2002.

[155] A. Bonaccorsi and C. Daraio, "Age effects in scientific productivity," *Scientometrics*, vol. 58, no. 1, pp. 49–90, 2003.

[156] M. Y. Vardi, "Revisiting the publication culture in computing research," *Communications of the ACM*, vol. 53, no. 3, p. 5, 2010.

[157] M. E. J. Newman, "Who is the best connected scientist? A study of scientific coauthorship networks," in *Complex Networks* (E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai, eds.), pp. 337–370, Springer, 2004.

[158] P. V. Marsden, "Network data and measurement," *Annual Review of Sociology*, vol. 16, pp. 435–463, 1990.

[159] M. E. J. Newman, "Scientific collaboration networks: II. shortest paths, weighted networks, and centrality," *Physical Review E*, vol. 64, no. 1, p. 016132, 2001.

[160] T. Fruchterman and E. Reingold, "Graph drawing by force-directed placement," *Software Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[161] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences*, vol. 101, no. 11, pp. 3747–3752, 2004.

[162] R. S. Burt, "A note on missing network data in the general social survey," *Social Networks*, vol. 9, no. 1, pp. 63–73, 1987.

[163] R. Little and D. Rubin, "The analysis of social science data with missing values," *Sociological Methods and Research*, vol. 18, pp. 292–326, 1990.

[164] D. Stork and W. D. Richards, "Nonrespondents in communication network studies: Problems and possibilities," *Group Organization Management*, vol. 17, no. 193, 1992.

[165] W. G. Cochran, "Some methods for strengthening the common chi-square tests," *Biometrics*, vol. 10, no. 4, pp. 417–451, 1954.

[166] A. Agresti, "A survey of exact inference for contingency tables," *Statistical Science*, vol. 7, no. 1, 1992.

[167] M. Reiser and Y. Lin, "A goodness-of-fit test for the latent class model when expected frequencies are small," *Sociological Methodology*, vol. 29, pp. 81–111, 1999.

[168] R. A. Fisher, "On the interpretation of chi-squared from contingency tables, and the calculation of p," *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87–94, 1922.

[169] G. Kossinets and D. J. Watts, "Empirical Analysis of an Evolving Social Network," *Science*, vol. 311, no. 5757, pp. 88–90, 2006.

[170] P. Panzarasa, T. Opsahl, and K. M. Carley, "Patterns and dynamics of users' behavior and interaction: Network analysis of an online community," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 5, pp. 911–932, 2009.

[171] G. F. Davis, M. Yoo, and W. E. Baker, "The small world of the american corporate elite, 1982-2001," *Strategic Organization*, vol. 1, no. 3, pp. 301–326, 2003.

[172] H. Ebel, L.-I. Mielsch, and S. Bornholdt, "Scale-free topology of e-mail networks," *Physical Review E*, vol. 66, no. 3, p. 035103, 2002.

[173] M. Ripeanu, I. Foster, and A. Iamnitchi, "Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system," *IEEE Internet Computing Journal*, vol. 6, p. 2002, 2002.

[174] J. W. Grossman and P. D. F. Ion, "On a portion of the well-known collaboration graph," *Congressus Numerantium*, vol. 108, pp. 129–131, 1995.

[175] T. Finholt and J. Birnholtz, "If We Build It, Will They Come? The Cultural Challenges of Cyberinfrastructure Development," *Managing nano-bio-info-cogno innovations*, pp. 89–101, 2006.

[176] D. J. D. S. Price, "Some remarks on elitism in information and the invisible college phenomenon in science," *Journal of the American Society for Information Science*, vol. 22, no. 2, pp. 74–75, 1971.

[177] C. S. Wagner, *The New Invisible College: Science for Development*. Brookings Institution Press, 2008.

[178] G. M. Olson and J. S. Olson, "Distance matters," *Human-Computer Interaction*, vol. 15, no. 2, pp. 139–178, 2000.

[179] V. Stodden, "Enabling Reproducible Research: Open Licensing For Scientific Innovation," *International Journal of Communications Law and Policy*, vol. 13, 2009.

[180] I.-S. Kang, S.-H. Na, S. Lee, H. Jung, P. Kim, W.-K. Sung, and J.-H. Lee, "On co-authorship for author disambiguation," *Information Processing & Management*, vol. 45, no. 1, pp. 84 – 97, 2009.

[181] M. E. J. Newman, S. Forrest, and J. Balthrop, "Email networks and the spread of computer viruses," *Physical Review E*, vol. 66, no. 3, p. 035101, 2002.