Title

Impacts of Computer-Mediated Communication on Travel and Communication Patterns: The Davis Community Network Study

Author

Balepur, Prashant Narayan

Publication Date

1998

CALIFORNIA PATH PROGRAM
INSTITUTE OF TRANSPORTATION STUDIES
UNIVERSITY OF CALIFORNIA, BERKELEY

Impacts of Computer-Mediated Communication on Travel and Communication Patterns: The Davis Community Network Study

Prashant Narayan Balepur
University of California, Davis

California PATH Working Paper
UCB-ITS-PWP-98-11

This work was performed as part of the California PATH Program of the University of California, in cooperation with the State of California Business, Transportation, and Housing Agency, Department of Transportation; and the United States Department Transportation, Federal Highway Administration.

Report for MOU 279

May 1998

ISSN 1055-1417

Impacts of Computer-Mediated Communication on Travel and Communication Patterns:

The Davis Community Network Study

By

Prashant Narayan Balepur
B. Tech. (Indian Institute of Technology, Madras) 1994

THESIS

Submitted in Partial Satisfaction of the Requirements for the Degree of

MASTER OF SCIENCE

in

CIVIL ENGINEERING

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Thesis Advisor: Patricia L. Mokhtarian

ACKNOWLEDGMENTS

I would like to thank a number of people for their guidance, support and assistance during the course of this thesis:

- Prof. Patricia Mokhtarian, my thesis advisor, for getting me involved with the DCN project, for suggesting an interesting and uncharted topic to work on and for her keen understanding of the issues involved. It's been a pleasure knowing and working with you Pat!
- Profs. Dan Sperling and Debbie Niemeier, members, thesis committee, for reviewing this thesis and for providing critical input.
- Prof. Ilan Salomon, visiting faculty, for his seminal work in the area of travel and telecommunications interactions and for reviewing many of the initial analysis.
- Mike Bagley, David Stanek and Krishna Varma, fellow ITS researchers and members of the DCN evaluation project and everyone else who helped out on the project, for doing the grunt work that made this study possible.
- DCN system users, particularly those who were kind enough to turn in their surveys, for providing us with data that forms the basis for this study.
- Niranjan Avadhanam of the Electrical Engineering department and other unnamed denizens of the sci.stat.math newsgroup for discussing arcane cluster and similarity analysis issues.

Impacts of Computer-Mediated Communication on Travel and Communication Patterns: The Davis Community Network Study

Abstract:

This report examines the interactions among different forms of communication, where travel is also considered a form of communication. Data for this study comes from 148 respondents to the "Activity Diary" survey instrument of the Davis Community Network (DCN) project, which obtained information on 636 uses of DCN. Generation, elimination and modification were considered to be the three major potential results of the present DCN communication and the five major types of communication considered were: in-person, physical object, electronic, in-person with travel, and physical object with travel. New activities were generated due to the use of DCN, a Computer-mediated Communication (CmC) system. Cluster analysis performed on the data to identify patterns of outcomes of the current communication yielded 20 clusters, of which the largest comprised neutral outcomes (no impact on future communication) and the second largest involved generation of electronic communication. A cluster analysis on the media alternatives considered by the respondents yielded 17 clusters of which the largest one was characterized by the "no alternative to DCN" response. Among media characteristics, economy, ease of use, speed and non-disruptiveness were found to be the most significant factors influencing the media choice of individuals. DCN was considered to be faster and cheaper than other forms of communication. Work-related and social/entertainment were the two purposes most frequently offered for the current DCN activity. Impacts of individual characteristics such as age, gender, income level and computer experience on DCN activities were also studied. Recommendations for future research are provided.

Keywords: Communication Networks, Communications, Computer Networks, Telecommunication, Travel Behavior.

…

Summary:

The major aims of this study were to answer the following two questions:

- How does the availability and usage of one medium of communication affect other communication activities conducted via other media?
- What are the major factors that influence the choice of one communication medium over another?

A Computer-mediated Communication (CmC) is a form of communication conducted using the computer and telecommunications networks. The specific CmC studied in this report was the Davis Community Network (DCN), which provided electronic mail and Internet access capabilities to its users. Aggregate analysis of 636 DCN communications (from 148 users) showed that new activities were generated due to the use of the DCN system. Some substitution of travel was also seen but that effect was relatively small and was outweighed by increases in electronic communication.

For each use of DCN, the respondent indicated whether that use was expected to generate, eliminate, or modify an in-person, physical object, or electronic form of communication. Multiple outcomes were possible. These responses formed a vector associated with each DCN use. Cluster analysis was performed on the set of 636 vectors to identify similar patterns of outcomes and their prevalence. Twenty clusters captured 95% of the cases. The largest cluster, containing approximately 31% of the cases, comprised neutral outcomes which neither generated nor eliminated future communication. The second largest cluster comprised outcomes which involved generation of electronic communication.

The alternatives to DCN for the current communication were identified. Each communication description was coded as a 9-bit vector corresponding to the presence or absence of a particular alternative medium in the choice set. A cluster analysis was performed to identify patterns of perceived alternatives, and 17 major clusters were identified (comprising 95% of the cases). The largest cluster (comprising more than 44% of the cases) was the one for which the respondents replied that there were "no alternatives to DCN ". Among media characteristics, economy, ease of use, speed and non-disruptiveness were found to be influential factors to choice of medium. DCN was considered to be faster and cheaper than other forms of communication.

Work (32%) and social/entertainment (22%) were the two most frequently-cited purposes for the current DCN activity. Examining the individual users' characteristics, older users were more likely than younger ones to view personal media as alternatives to the current DCN activity. Higher income respondents were disproportionately likely to view in-person, alternative Internet access, and object delivery with travel as alternatives to the current activity .

Applicability to transportation problem solving:

This study looks at the interactions among different forms of communication, including travel as an alternative form. It has been hypothesized that electronic forms of communication may replace some travel. The results of this study show that this is not the dominant effect. On net, DCN appeared to: greatly increase the number of electronic communications; leave the number of in-person communications essentially unchanged (generating some communications, but

eliminating or substituting just as many); decrease the number of communications through physical objects (such as a book or diskette); and decrease the number of trips. Although the overall impact of DCN as far as travel is concerned appears to be one of substitution, analysis of additional data collected under this project (the subject of another report) indicates that when interactions among the various communication alternatives are more fully accounted for, the predominant impact appears to be complementarity or generation.

CONTENTS

	Title	**i**
	Acknowledgments	**ii**
	Abstract	**iii**
	Summary	**iv**
	Contents	**vii**
	List of Tables and Figures	**ix**
Chapter 1	**Introduction**	**1**
Chapter 2	**Literature Review**	**4**
2.1	Travel and Telecom Relationships - Concepts	4
2.2	Travel and Telecom Relationships - Empirical Studies	8
2.3	Computer-mediated Communication and Community Networks	11
Chapter 3	**Davis Community Network Project**	**15**
3.1	Project Overview	15
3.2	Description of Survey Instruments and Survey Sample	17
Chapter 4	**Investigation of Communication Patterns**	**23**
4.1	Conceptual Framework	23
4.2	Macroscopic Analysis	35
4.3	Microscopic Analysis	38
	4.3.1 Similarity Measures	45

4.3.2 Variable Weighting 49

4.3.3 Clustering Algorithms 53

4.4 Communication Patterns 58

4.5 Conclusions 69

Chapter 5 Investigation of Media-choice Behavior 72

5.1 Media Choice Clusters 72

5.2 Media Choice Behavior 79

5.2.1 Media Characteristics 80

5.2.2 Activity Characteristics 85

5.2.3 Individual Characteristics 89

5.4 Conclusions 94

Chapter 6 Conclusions and Suggestions for Future Research 96

6.1 Conclusions 96

6.2 Suggestions for Future Research 105

List of References 107

LIST OF TABLES AND FIGURES

TABLES

Table 3.1	Application Usage in the DCN Study	21
Table 3.2	Distribution of Activities by Purpose	22
Table 4.1a	Distribution of DCN Impacts by Medium (N=636)	36
Table 4.1b	Self-generated Activities	37
Table 4.2	Net Volume of Communication by Medium	37
Table 4.3	Distribution of Patterns Obtained by Sorting 12 Variable String	41
Table 4.4	Variable Weights Calculated Using Equations 4.6	58
Table 4.5a	Clusters obtained with Density Linkage Clustering with K=5 on 8-bit Effects Vector and Weighted Distances	60
Table 4.5b	Clusters obtained with Density Linkage Clustering with K=5 on 8-bit Effects Vector and Unweighted Distances	60
Table 4.6	Crosstabulation of Neutral Communication Activities	62
Table 4.7	Purposes Associated with Stand-alone Self-contained Activities	63
Table 4.8	Comparison of Communication Purposes for Neutral and Active Communications	64
Table 4.9	Comparison of Application Usage for Neutral and Active Communications	65
Table 5.1	Media Alternative Clusters Obtained by Density Linkage	

	Clustering with K=5 and the Weights Used	75
Table 5.2	Cluster Descriptions	78
Table 5.3	Reasons for Using DCN by Media Alternative Cluster	82

FIGURES

Figure 2.1	Conceptual Representation of Telecommunications-Travel Interactions	8
Figure 3.1	Distribution of Income Levels	18
Figure 4.1	The Activity Interaction Diagram	25
Figure 4.2	Self-generated Activities	31
Figure 4.4	Match Table for Binary Variables	45
Figure 4.5	Activity Purposes Associated with Top 19 Effects Clusters	66
Figure 4.6	Applications Associated with Top 19 Effects Clusters	66
Figure 5.1	Inverted Media Choice Model	74
Figure 5.3a	Distribution of Age in Different Media Alternative Clusters	92
Figure 5.3b	Distribution of Computer Experience in Different Media Alternative Clusters	92
Figure 5.3c	Distribution of Gender in Different Media Alternative Clusters	93
Figure 5.3d	Distribution of Income in Different Media Alternative Clusters	931

INTRODUCTION

It is being recognized today that interactions between communication systems and travel and those among communication systems themselves are rather complex. The belief that advanced communication systems will substitute for or reduce travel is gradually being replaced by the recognition that other effects such as generation and modification are also likely. Starting with the concept that travel is yet another communication medium, this thesis attempts to investigate two main empirical issues covering the broader topics of media-usage and interactions: First, "*How does the availability and usage of one medium of communication (computer-mediated in this case) affect communication activities conducted via other media?*" and second, *"What are the key factors that influence the choice of one medium over another in a certain activity context?*"

The data used for the study comes from the Davis Community Network (DCN) Evaluation Project. The community network provided internet and email access to its users and limited community information and services at the time of the evaluation. One of the primary goals of the Evaluation Project was to evaluate the effect of the community network on its users' travel behavior in an attempt to measure the effectiveness of such networks as a travel reduction strategy. As part of the study several survey instruments were designed and used to collect socio-demographic information, data on communication activities and system usage. This thesis primarily utilizes the data from the "Activity Diary" survey instrument, wherein the users of the community network provided information on communication activities carried out using the system. Apart from the details of the current activity, the respondents also described details of

any antecedents of the current activity and also any expected "effects" in terms of activities that might be generated, modified or eliminated by the current activity.

Following the model that influenced the design of the study in general and the "Activity Diary" survey in particular, an activity-based approach was taken to investigate the two questions identified earlier. To investigate the first of the two issues, "*How does the availability and usage of one medium of communication (computer-mediated in this case) affect communication activities conducted via other media?*", each current activity was first tagged not only by its own characteristics, but also by whether it was prompted by other activities, and by the media of the activities that it was anticipated to generate, modify or eliminate. The analysis was then conducted at two levels of resolution. At the lower resolution, the net volumes of communication carried out using the different media were calculated. To investigate the problem at a finer resolution, the data were classified into commonly occurring communication patterns and the specific tradeoffs between the different media was investigated in each of the clusters.

To investigate the second empirical question, "*What are the key variables that influence the choice of one medium over another in a certain activity context?"* the data were clustered into sets marked by common alternatives to using the DCN system. Then, an exploratory analysis was conducted to identify the key variables that distinguished between these media-alternative-clusters and thus aided in our understanding of why a certain medium was chosen in a certain context.

The primary conceptual challenge was to develop a comprehensive and coherent activity model which would form a basis for the analyses to follow. The methodological and analytical challenge was in the grouping of data into relevant and cohesive sets that would enable the more detailed analyses. Similarity and clustering techniques were used to accomplishing this latter task. Anecdotal information provided by the respondents was used to place the data in perspective. This thesis hopes to add to our understanding of CmC and help refine similar studies in the future.

The organization of this thesis is as follows: following this introduction the next chapter describes the state of the art and the current thinking with regards to topics related to this work. Chapter 3 gives a brief overview of the Davis Community Network project, the survey instruments used to collect data for this study and a brief description of the data itself. The next two chapters, Chapter 4 and 5, form the crux of the thesis. In Chapter 4 the topic of Communication Patterns is investigated. The chapter presents the conceptual framework used in the study and the results of the concomitant analysis. The theoretical background on similarity and cluster analysis is also presented in this chapter. Chapter 5 deals with the topic of Media Choice Behavior in a CmC setting. Following the conceptual model laid out in the previous chapter, the results of the analysis are presented. Finally, the last chapter summarizes the conclusions of the analyses from the previous two chapters and also discusses future research directions.

2

LITERATURE REVIEW

The topic of Computer-mediated Communication (CmC) and its effect on travel and communication patterns is fairly nascent and therefore there is not a significant body of directly relevant literature. However, several of the basic areas that this thesis touches upon are well researched and documented. In this chapter a summary of selected significant literature covering the following areas will be presented:

1. Travel and telecommunications relationships - Concepts
2. Travel and telecommunications relationships - Empirical studies
3. Computer-mediated Communication
4. Community Networks

A major issue not covered in the literature review presented here is that of the analytical methods employed in this thesis - similarity and cluster analysis. These topics are covered in great detail in Chapter 4 as it was deemed more logical to present the crux of the analytical theory immediately prior to the actual analysis itself.

2.1 TRAVEL AND TELECOMMUNICATIONS RELATIONSHIPS - CONCEPTS

The relationship between telecommunications and transportation has been discussed in the transportation literature for at least 25 years (Nilles *et al.,* 1976). In more recent years, an exponential increase in telecommunication technology and its accessibility; and the persistent

attractiveness of reducing travel, has led to a resurgence of interest among transportation planning academicians and practitioners. While simple substitution of travel has been the predominantly discussed relationship (Toffler, 1981), it is becoming clear that other interactions exist.

Substitution, the interaction most desired by transportation planners, assumes that as the availability of telecommunication technologies expands to allow more applications at reduced costs, the need for travel will diminish as telecommunications will be used instead (Salomon, 1986). Telecommuting from home using a communications link to the work place leading to the elimination of a work trip is an example of substitution. If it is assumed that this is the only kind of interaction that exists, the total volume of activities (whether travel or communication) will not be affected by the assignment of traffic to either medium. This statement in fact refers to a basic premise of this thesis that transportation is yet another mode of communication. It has been noted that most travel involves communication, either (most often) as its primary end (e.g., drive to a conference) or as an adjunct (e.g., making hotel and airline reservations for a vacation). Thus, it is evident that both physical transport and telecommunications are subsystems of the larger communications system.

Given the premise presented above, it becomes clear that interactions other than mere substitution are not only possible but likely. The second such interaction is that of *enhancement*. Salomon (1986) describes enhancement as the generation of additional travel between two nodes due to additional telecommunication. Travel to a rare philately exhibit after learning about it through a newsgroup on the Internet would be an example of enhancement, for had the individual

not had access to the newsgroup (and the relevant communication medium) she would not have engaged in the trip.

Operational efficiency is a third type of interaction and is described by Meyburg (1983) as any situation where one service is contributing to the efficiency of the other. Salomon (1986) points out that greater efficiency of transportation systems in terms of necessary capital, operating costs, energy and pollution costs, safety and other effects can be achieved through the employment of telecommunications. The use of Electronic Road Pricing (ERP) (being tested on the Golden Gate Bridge, San Francisco) for toll collection is a good example of the use of telecommunications to make the transportation system more efficient.

A fourth type of interaction is that of *indirect, long-term interactions* such as the effect of telecommunications on land-use and thus indirectly on transportation. The popular hypothesis (Kellerman, 1984) is that increased use of telecommunication will result in the geographical dispersion of activities and residences and though the number of trips might decrease, the sheer magnitude of the distances traveled to engage in face-to-face interactions will lead to a net increase in the vehicle-miles traveled.

An outcome of the complementarity of the transportation and telecommunication systems, is the hypothesis presented by some researchers (Mokhtarian, 1990) that "communication breeds communication". That is, while there is likely to be a migration of activities from one medium to

the other, there is also likely to be an increase in the volume of communication activities themselves.

To summarize the discussion presented so far, we recognize that:

- it is logical to consider travel to be yet another communication medium,
- there exist interactions, both substitutions and enhancements, between media, and
- there is likely to be an increase in the volume of communication in general as such interactions occur.

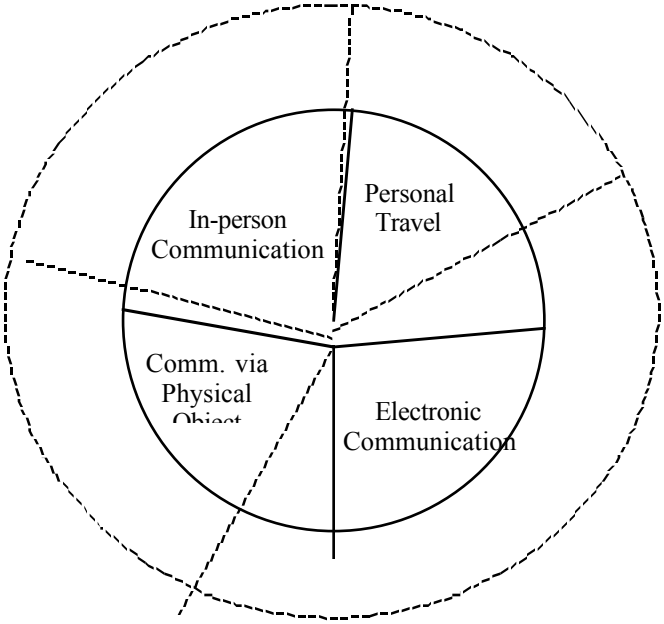
Figure 2.1 below embodies the concepts summarized above. The figure below uses the following different categories of communication media (these are described in considerable detail in Chapters 4 and 5):

1. In-person communication (face to face meetings)
2. Electronic communication (phone calls, electronic messages, TV)
3. Communication via physical object/surroundings (letters, books)
4. Personal travel

The figure only represents, *qualitatively*, the increasing envelope of communication activities and a change in the proportion of the activities using each of the media as this occurs. It should not be interpreted to mean that say, electronic communication gathers a large share of the activities using personal travel and communication via physical activity media. There is likely to be trans-migration of activities between various pairs of media. And in the case of a broad category such as electronic communication, it is very likely that activities shift from one specific technology to

another, as in electronic mail substituting for phone calls. It is precisely such specific migrations that this thesis attempts to study and document.

Figure 2.1 Conceptual Representation of Telecommunications-travel Interactions (Modified from Figure 2 in Mokhtarian, 1990)



2.2 TRAVEL AND TELECOM RELATIONSHIPS - EMPIRICAL STUDIES

The previous section detailed the various interactions that are likely between the different communication media. In this section some of the empirical studies in this area are mentioned. To date, disaggregate empirical studies of the relationships between telecommunications and travel have focused on the impacts of applications such as telecommuting and videoconferencing on travel. The studies of telecommuting [Henderson and Mokhtarian, 1996; Koenig *et al*., 1996; Henderson *et al*., 1996; Sampath *et al*., 1991] have unanimously found substantial net reductions in individual travel for the short term, although several of them caution that system-level impacts are negligible now and likely to remain so for quite some time. At least one study of

teleconferencing [Mokhtarian, 1988] documented a net *increase* in travel, due to more people traveling (albeit shorter distances per person) to the videoconference location than would have gone to the more distant central meeting location. Several other studies looking at specific communications technologies have also been carried out. One such study of the relationship between telephone use and travel carried out by Claisse and Rowe (1991) is particularly interesting because the concepts underlying the study are akin to those underlying the current effort.

The study was based on telephone activity and attempted to analyze telephone calls categorized under the following uses or "traffic":

1. Specific Market: call has no direct relationship to travel
2. Induction: call leads to a trip
3. Substitution: call is made instead of a trip
4. Management: call influences the particulars of a trip

The study sample was based on 663 people and 7252 phone calls. The key survey instrument was a one-week diary of all telephone calls. For each call the respondent described the "action" and "aim". Further, the respondents were asked whether the call had led to an unplanned trip (*induction*) and whether an alternate mode (personal travel, mail or telegram, waited/nothing) would have been used to carry out the action had the whole telephone network been down (*substitution*).

Claisse and Rowe used two methods to determine the volume of calls in each "traffic category". The first method used the answers to the latter two questions directly to measure volume. While calls in the induction and substitution categories were fairly clear, management traffic was defined as a call that did lead to an unexpected trip and for which the respondent would have made a trip (or sent someone) as an alternative had the network been unavailable. This deduction of management traffic is inherently weak and could have been better served by a direct question about whether the particulars of a pre-planned trip changed due to the telephone call. Using a second method, Claisse and Rowe categorized the actions into three general types: management, information, and discussion and the aims into: private life, social life, and working life. This leads to eight possible action/aim combinations (discussions were not deemed possible for working life). Additionally they distinguished calls based on the motivation for the call, classifying calls as relational (whose purpose is to keep in touch), functional (calls with a specific objective), functional plus incidentally relational and relational plus incidentally functional. Combinations of correspondent, aim and action were used to determine whether a call was relational, functional or hybrid.

The first approach yielded a net substitution of 22% (27% trip substitution, 5% induction) and the second found a net substitution of 17.1% (20.5% trip substitution, 3.4% induction). The net effect on urban mobility was defined as follows:

Change in Urban Mobility =

Net Substitution x (Calls/Day/Person) /(Trips/Day/Person)

and yielded a saving of 6% of all trips. Handy and Mokhtarian (1993) point out that the study is limited by its focus on residential phone habits, its exclusion of any incoming calls and the lack of an activity-based approach, emphasizing a single event rather than a chain of events.

2.3 COMPUTER-MEDIATED COMMUNICATION AND COMMUNITY NETWORKS

The studies discussed so far involve specific technologies or a narrow set of activities. The present study however focuses on a much broader and now, more ubiquitous set of technologies enabling what is called Computer-mediated Communication (CmC). Fisser *et. al.* define Computer-mediated Communication as "the process by which people create, exchange and perceive information using networked telecommunications systems that facilitate encoding, transmitting, and decoding messages. It typically includes computer conferencing, electronic mail, and electronic bulletin boards. CmC is a means to establish an electronic environment that is accessible to participants who might otherwise be separated by time zones and physical distance." Simply put, CmC is a form of communication conducted via computers and computer networks.

Computer-mediated Communication can be enabled through several technologies. The specific system studied here is a Community Network providing electronic mail and Internet access capabilities to its users. Handy and Mokhtarian (1993) provide an exhaustive survey of related technologies. Among them the Minitel system in France (late 1980s) appears to be the closest precursor of Community Network type technologies. Created as a replacement for printed

telephone directories and operating using dumb terminals, the system has evolved to include messaging services, chat facilities, games, banking services and general interest information on topics such as the weather.

The Davis Community Network evaluation project, which forms the basis for this study, attempted to study Community Networks as they are commonly understood. Beamish (http://alberti.mit.edu/arch/4.207/anneb/thesis/toc.html, 1995) in her thesis defines a Community Network as " ... a network of computers with modems that are interconnected via telephone lines to a central computer which provides:

- community information; and
- a means for the community to communicate electronically.

Unlike the similarly named 'on-line communities' or 'virtual communities', community networks are based in a physical place, what participants have in common are their cities and neighborhoods." While specific technologies mentioned in the definition of a Community Network might get outdated (ISDN lines might replace modems) the motivation and the basic architecture is likely to remain the same.

Existing non-profit community networks could be categorized into four groups (Beamish, 1995):

- Free-Nets,
- bulletin boards,
- government-sponsored networks, and

- wired cities.

Free-Nets are members of the National Public Telecomputing Network (NPTN) and follow their policies and procedures. The NPTN defines them to be loosely organized, community-based, volunteer-managed electronic network services, providing local and global information sharing and discussion at no charge to the Free-Net user or patron. The Cleveland Free-Net is an example. A *neighborhood bulletin board* is a small-scale bulletin board which usually focuses on a particular neighborhood rather than a city. These bulletin board systems (BBSs) are frequently scaled-down versions of city networks. MUSIC (Multi-User Sessions in Community) in Boston is an example of a neighborhood BBS. *Government-sponsored networks* are city-wide networks established with the primary purpose of making city records and municipal information available to residents. Santa Monica's Public Electronic Network (PEN) is an example. *Wired cities* are usually experimental projects providing information and services to households and businesses in a city. The DCN and the Blacksburg Electronic Village are examples of wired cities.

Thus, community networks not only provide the features incorporated in all other implementations of networked communication, such as electronic messaging facilities and Internet access, but go a step further by providing local, community related information and services. Unfortunately the DCN system at the time of evaluation lacked the elements that qualify it to be a Community Network. However, even just the provision of electronic mail and Internet access to a sizable segment of the community provided a rich test-bed for studying the influence of these capabilities on communication and travel activities. This thesis to our knowledge is the only one to date to explore at an activity level, the effects of Computer-mediated Communication

on communication and travel activities. While this thesis focuses on a not so prototypical Community Network, both the substantive insights gained from it and the evaluation methodology itself are likely to be applicable to the more typical ones.

Having introduced the reader to some key literature in the various areas related to this thesis, we can now move on to the actual study itself. The next chapter presents a very brief description of the Davis Community Network from which data was collected and analyzed for the purpose of this thesis. Chapters 4 and 5 form the crux of the thesis and present both the conceptual models and the analytical results that evolved from the study.

3
DAVIS COMMUNITY NETWORK PROJECT

The data used in this thesis were collected under the aegis of the Davis Community Network (DCN) Project. The project, funded by Caltrans, consisted of two pieces - the implementation piece and the evaluation piece. The implementation piece aimed at establishing a Computer Network in the City of Davis providing access to electronic mail, the World Wide Web and local community information and services. The evaluation piece aimed to evaluate the effect of this computer community network on users' travel and communication behavior, in an effort to measure the effectiveness of such systems as travel reduction measures. This chapter presents a brief overview of the DCN project and the data collection instruments used in the evaluation process.

3.1 PROJECT OVERVIEW

Davis, California is a town of about 45,000 people. Surrounded by agricultural land, it is nevertheless only 15 miles from the State Capitol of Sacramento, and a significant portion of its residents commute to state jobs in the capitol. The University of California, Davis is by far the major employer in Davis, with more than 20,000 employees. Hence, the sample for this study will be better educated and more affluent than average, and these characteristics should be kept in mind while evaluating the results.

Davis Community Network is a community-based computer network, established to provide a level of telecommunications capability to Davis residents and businesses well beyond what was previously available to the average home or office. Started on an experimental basis in January 1994, it had nearly 900 subscribers in August 1995 with membership expected to grow substantially in the coming months. The vision statement crafted by the DCN board declares that the mission of the Davis Community Network is to

" ... strengthen the community by helping its members understand and benefit from participation in the electronic information age",

with one of the priority goals being to

"lessen the need to drive vehicles, improve air quality and traffic." [DCN WWW Page]

In the time period that the system was evaluated, the primary utility provided by DCN was access to the Internet; specifically electronic mail facilities, access to the World Wide Web (WWW) and newsgroups. At that time it also offered, to a limited extent, a community web site hosting home pages for some City government offices, local businesses and an event calendar. The web site also served as a vehicle for community planning groups for exchange of documents (agendas, minutes, and so on) related to the process of preparing the General Plan. Today, in keeping with its vision, the Davis Community Network has matured into a full-fledged community network, with the web site featuring everything from the local transit routes and timetable to the ability to order pizza and groceries from local stores. The reader is encouraged to visit the Davis Community Network web page at http://www.dcn.davis.ca.us to get a feel for a prototypical community network.

The implementation and evaluation of the DCN was part of a 3-year research contract with the California State Department of Transportation (Caltrans), the evaluation part being fully funded by Caltrans. The evaluation was funded contemporaneously with the implementation, which meant that when the network was slower to develop than anticipated, the system being evaluated was less sophisticated than initially anticipated. Development of the network has continued, whereas the evaluation funding terminated June 30, 1995. Data collection was postponed as late as possible in the evaluation period to allow the system to develop as fully as possible within the scheduled time frame. Two of the data collection instruments used in this thesis - the Background Survey and the Activity Diary - will now be described.

3.2 DESCRIPTION OF SURVEY INSTRUMENTS AND SURVEY SAMPLE

A *Background Survey* was designed to obtain information on user characteristics. The survey is composed of four sections, collecting information on the following:

Part A: Use of Computers and Other Communication Technology

Part B: Job Characteristics

Part C: Commute Travel Characteristics

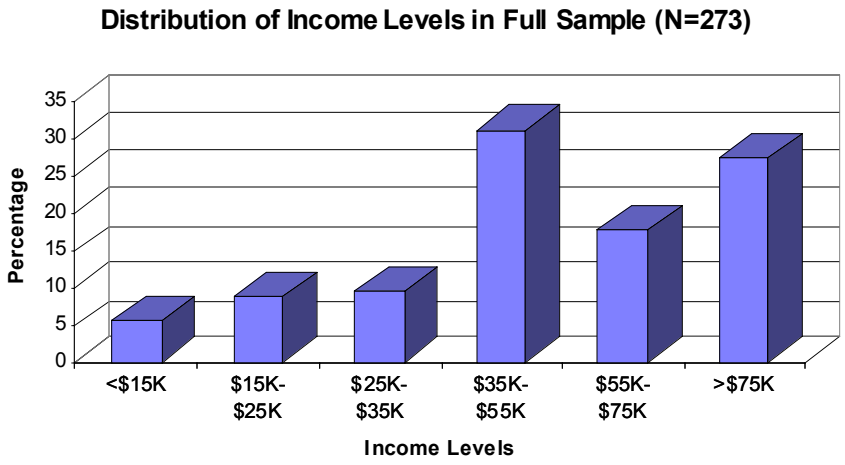
Part D: General Socio-Economic Information

Of these, information on Computer Experience (Part A), Age, Income and Gender (all from Part D) were used in this thesis, as these variables were hypothesized to have some affect on both communication patterns (Chapter 4) and media-choice behavior (Chapter 5). Some basic statistics related to each of these variables will now be presented.

A total of 273 people responded to the survey out of 505 people to whom the survey was sent out. Of these a majority of the sample (85.9%) had more than four years of experience with personal computers. However, a large percentage (50-54%) had seldom or never used on-line services (newsgroups, gopher searches or electronic bulletin boards). An overwhelming majority of the sample (96.7%) had computers at home. Also, 81% of the sample reported that they used computers *primarily* for work-related activities.

The gender distribution in the sample population was fairly even, with the composition being 42.1% females and 57.9% males. The age of the respondents was distributed mainly among the 25 to 34 (22.9%), 35 to 44 (29.2%) and 45 to 54 (29.2%) groups. The annual household income of the sample was fairly high. The distribution of the income levels is shown in Figure 3.1 below.

Figure 3.1 Distribution of Income Level



While the Background Survey provided information on respondent characteristics, the *Activity Diary* provided information on the various effects of using a computer-mediated system such as

DCN. Data collected from the Activity Diary forms the basis for the analyses to be presented in the chapters that follow. The Activity Diary is a collection of an individual's DCN Communication activities. Two forms of the diary were used to collect information: one group of randomly selected respondents were administered the survey through a mail-back paper version, while the other half received an electronic version of the diary. Though subtle differences between the two versions exist which might have influenced the responses to some of the questions in the survey, controlling for such effects was beyond the scope of this thesis and were disregarded. Seventy-one people responded to the paper version of the Activity Diary (out of a total of 125) and seventy-seven people responded to the computer version of the diary (out of 146 persons). Each diary requested information on up to five uses of the DCN ("communication descriptions"). A total of 636 descriptions were available and used in the current analyses, an average of 4.3 per person.

For each DCN communication description the respondent provided information on both the current activity s/he was engaged in and the effects of that activity. Activities were broadly divided into four categories depending on the medium used to carry them out. These categories, to be discussed in greater detail in the next chapter, are:

- In-person communication (example: interacting face-to-face)
- Electronic communication (example: phone calls and electronic mail)
- Communication via object delivery (example: reading a magazine)
- Travel

Questions regarding the current communication included the *initiator* of the communication (respondent or somebody else), the purpose of the communication (work, social/entertainment, school etc.), the nature of the communication (sent/received/requested information, conducted a transaction etc.) and the antecedents of the activity (whether any previous activities prompted the current communication and their details). Questions regarding the effects covered three distinct types of effects:

1. *Generation* of *additional* activities.
2. *Elimination* of *previously planned* activities.
3. *Modification* of *previously planned* activities.

The respondents were also asked to report if an alternate medium (in lieu of DCN) would have been used to carry out the current activity.

For each of these questions the respondent was first asked whether such an effect existed and if it did, details were solicited regarding the outcome of the effects. These details included information on the type of the activity (in-person, electronic etc.) and the initiator. The other activity specific data that was collected concerned such issues as to how exactly an object would be delivered or what mode of transportation would be used if travel was involved. For the purposes of this thesis and for the analysis presented in Chapter 4 (Investigation of Communication Patterns) however, only information regarding the presence or absence of each of the effects was used. Finally, respondents were asked to indicate the advantages and disadvantages of using DCN as the communication medium for the activity described. This along

with the information on the alternatives to DCN were used to study the question of media-choice behavior in Chapter 5 (Investigation of Media-choice Behavior).

Some key tabulations on the 636 communication descriptions are given below. The majority of the communication activities, more than 81%, were initiated by the respondent. Email was the predominant application, used for nearly 64% of the communications. The usage of the other applications is given in the table below.

Table 3.1 Application Usage in the DCN Study

Application Used	**Frequency**	**% Frequency**
Email	405	63.68
File Transfer Protocol (FTP)	38	5.97
Gopher (Search Engine)	23	3.62
Internet Relay Chat (IRC)	1	0.16
Lynx (text-based hyperlinked browser)	8	1.26
Newsgroups	73	11.48
WWW	86	13.52
Miscellaneous	2	0.31
Total	636	100

For the question "*What did you do during the current communication?"*, respondents answered *"sent/received/requested information"* for 477 (75%) of the 636 communication activities, *"conversed*" for 153 (24.01%) descriptions and *"conducted a transaction*" in only 5 (0.79%)

cases. While these results are fairly intuitive, it is interesting to see a significant number of "conversed" activities, which are most likely to be inter-personal email communication.

The distribution of the stated purpose of the reported communication activity reveals that the system was used primarily for work-related activities more than any other (more than 32% of the activities). However, it is also significantly used for social/entertainment purposes, browsing and hobby.

Table 3.2 Distribution of Activities by Purpose

Purpose	**Frequency**	**% Frequency**
Browsing	102	16.04
Work	206	32.39
Social/Entertainment	143	22.48
Shopping	12	1.89
School	12	1.89
Hobby	69	10.85
Volunteer Work	59	9.28
Traveling	14	2.20
Other	19	2.99
Grand Total	636	100

In this chapter the source and nature of the data used in the analyses were briefly described. We are now ready to discuss the conceptual framework developed, the methodologies used and the

results obtained in greater detail. In Chapter 4 the question of communication patterns is taken up and in Chapter 5 media-choice behavior is investigated.

4

INVESTIGATION OF COMMUNICATION PATTERNS

In this chapter we take a closer look at the various interactions that are likely to exist among communication media. Following this introductory section, the conceptual framework that guides the rest of the study is laid out. In section 4.2 the results of a high-level macroscopic analysis are presented, wherein the aggregate values of communications for the different media are calculated. These figures, while only being qualitative indicators of communication flow, do suggest a certain pattern of tradeoffs among the different media. To investigate these tradeoffs in greater detail an attempt was made to uncover the commonly occurring communication patterns in the data set. The primary methodological tools used are similarity and cluster analyses, which are discussed in Section 4.3. These techniques were then applied to the available data set and the results are presented in Section 4.4. Anecdotal information provided by the respondents is used to place the data in context. The data for the analysis presented in this chapter and the next comes from the Activity Diary and the Background Survey, discussed in some detail in Chapter 3.

4.1 CONCEPTUAL FRAMEWORK

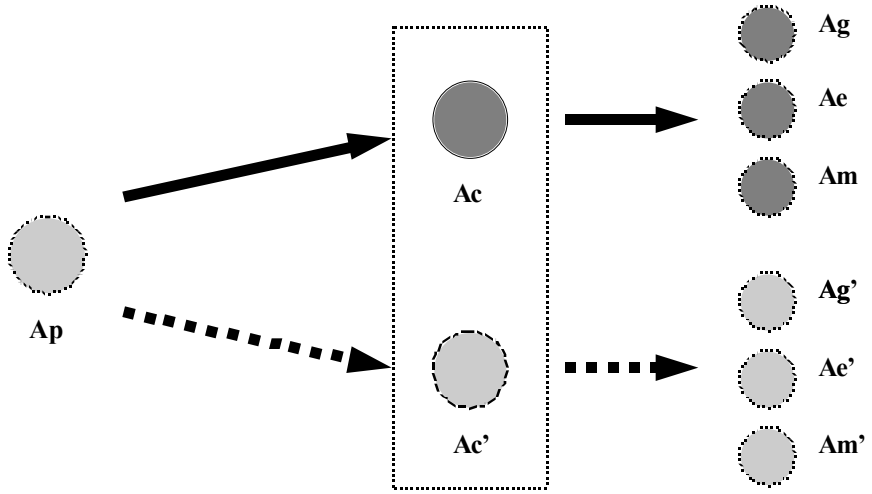
This section describes certain conceptual models and hypotheses that motivate and guide the exploratory study that follows. The study concentrates on Computer-mediated Communication (CmC) based Activities. An activity is described as a reaction to one or more stimuli, which as Salomon [1985] points out, is in most cases, the need for fulfillment at some psycho-physical

level [Maslow, 1954]. These motivations are usually first realized at a subconscious level and are later translated into a cognizable purpose, and could either be external (as when we return a friend's phone call) or internal (as when we decide to call a friend we haven't spoken to in a while). To accomplish this activity (in order to achieve the cognizable purpose) a person usually uses a certain medium of communication. Thus an activity A is characterized by a purpose P and a *medium* of communication m.

Having laid out a working definition for an activity, let us now describe the activity interaction model. The model, best discussed in conjunction with the figure below, has underlying it the hypothesis that activities affect other activities by either generating them, eliminating them or by modifying their details. Thus, activities form a sequence or more specifically a directed graph with each node linked to multiple parents and multiple offspring. Parent activities (nodes) are *causally* responsible for their offspring, or in other words the offspring activities are generated by the parent activities. To sidetrack a little, causality is emphasized here because a common mistake found in the survey responses was that of mistaking chronological precedence as a sufficient condition for qualification as a parent activity. Consider the following two cases. In the first situation, a person plans to visit a friend in person but calls to confirm she is there before making the trip. In the second case, the person spontaneously calls a friend, and she equally spontaneously invites him over for lunch. The phone call actually causes the trip only in the second instance. Attributing causality in the first instance is a classic case of the *post hoc ergo propter hoc* fallacy in logic, wherein the fact that one activity *follows* another leads to a faulty conclusion that the later activity was *caused* by the earlier one. Causality is important

because otherwise the wrong activities might be attributed with generating, eliminating or modifying certain activities. However, the notion of causality is fuzzy and primarily relates to "being responsible in some way" for an activity's effects. Finally, an activity can also have a complementary activity (or a set of them), where a complementary or alternate activity is one that accomplishes an identical objective (purpose) but through different means, in our case, a different medium.

Figure 4.1 The Activity Interaction Diagram



Media Choice between c and c'

The model has two main components, the interaction model shown above and the media choice model which will be described shortly. In the figure above, $\mathbf{A_c}$ corresponds to the current activity, and $\mathbf{A_g}$, $\mathbf{A_e}$ and $\mathbf{A_m}$ correspond to the set (possibly) of activities that are generated, eliminated and modified respectively, by the current activity. $\mathbf{A_p}$ corresponds to the set of parent activities that prompted the current communication. $\mathbf{A_c}$, represents the set of complementary or alternate activities, activities that would have substituted for the current

activity had the current mode of communication been unavailable or inaccessible in some way. And finally, $\mathbf{A_g}'$, $\mathbf{A_e}'$, and $\mathbf{A_m}'$, are the sets of activities that are likely to have been generated, eliminated and modified by the set of alternate activities $\mathbf{A_c}'$. The dashed lines around some of the activity sets represent that those activities need not necessarily exist. Thus, the only activity that must necessarily exist is the current activity.

A major conceptual hurdle faced in this study was that of the adequate time horizon of events that one is interested in. To paraphrase, how far back does one look for the event that sparked off the series of events that eventually led to the current activity and how far into the future is one to follow the chain of events resulting from the current activity. In this discussion we limit ourselves to a certain level (corresponding to the current activity) and one generation above and one generation below this level. It is to be noted however that by a generation we are merely referring to the set of activities reported by the respondent, which might actually include activities several generations deep temporally (if the respondent has thought out the ripple effects of the current activity beyond the effect that just immediately follows). Practical considerations of survey design and respondent ability motivate this restricted view, and though it does not fully take into account the long term effects of an activity, it should provide some broad insights into the effects on communication activities.

Having discussed the idea of activity interactions, we now turn our attention to the question of media choice. As is indicated in Figure 4.1, media choice is closely related to the concept of complementary activities. Media choice behavior dictates the medium of communication that is

finally chosen to carry out a specific activity. As was discussed earlier, in a relatively straightforward CmC activity scenario, a person is likely to go through a series of steps. First, s/he will transform a subconscious need for psycho-physical fulfillment into an activity with a cognizable purpose. Having become aware of the need to accomplish a certain activity, the person will then attempt to choose a medium to do so.

The problem now reduces to one of discrete choice among several communication media alternatives. By communication media we refer to combinations of individual media which could be broadly classified as instances of the following:

- In-person communication
- Communication through physical object/surrounding
- Electronic communication
- Travel

Let us consider the categories one at a time: *In-person communication,* as the name suggests, involves *face-to-face* communication between individuals or groups of individuals. Talking to one's spouse at home or a colleague at work falls under this category. In-person communication is primarily inter-personal (as opposed to primarily involving information acquisition) and is characterized as being very intimate, interactive, intensive and information rich. It is one of the few forms listed above that involves extensive use of non-verbal cues. *Communication through physical object* includes obvious formal means such as writing a letter to a friend or reading a

book, but also non-formal means involving sensory perception such as wine tasting, a walk through the park or receiving a gift. This medium has probably the lowest level of real-time interactivity with regards to inter-personal communication and in the traditional forms such as letter writing, is limited in the non-linguistic communication possible. The third category, *electronic communication*, covers the entire gamut of electronic media ranging from traditional telephones to advanced teleconferencing systems and computer-mediated systems and includes broadcast media such as radio and television. The fourth and final category is *travel* and its inclusion is a key feature of the conceptual model. Following the derived demand theory of travel which states that people travel for a purpose, it is hypothesized that travel occurs primarily in conjunction with the first two media, namely, in-person and physical object/surrounding. While occasional joy riding occurs, the majority of travel is carried out with a purpose in mind at the destination or during the process of travel itself.

Thus, when a person decides to choose a medium to carry out an activity s/he has available the following *primary* alternatives:

1. In-person communication (I)
2. Communication via physical object/surrounding (P)
3. Electronic communication (E)
4. Travel and in-person communication (TI)
5. Travel and communication via physical object/surrounding (TP)

Note that I and P nearly always involve travel of some kind. We distinguish alternatives 1 and 2 from alternatives 4 and 5 in that 4 & 5 refer to travel that wouldn't have occurred otherwise. For

example, if in-person communication is said to be generated by an activity, it would be classified as 1 if no additional travel were involved and as alternative 4 otherwise.

This list of primary alternatives does not mean that a person engaging in a certain activity is restricted to selecting only *one* of these modes. In fact, in many cases, more than one primary mode is used. But, any combination of communication media can be broken down into a set of primary alternatives. If more than one primary mode is used in the activity, they only augment each other and do not compete. If for example, one sends out an email (electronic communication) inviting friends to a party and then drops by the home of a close friend to extend a personal invitation (in-person communication) they are not competing but are acting separately to augment each other. The point being made here is that a person never uses two primary modes simultaneously for a given activity (being carried out with a certain purpose in mind). The use of multiple primary modes (in different points in time) occurs only under two scenarios, when the two augment each other as described above, or when the first primary mode that is chosen, fails to generate the desired result, and the person enters into an iterative process where s/he picks the medium that appears to be most appropriate, carries out the activity and compares the results with the desired purpose in mind, with the process terminating either when the desired result is achieved or when the person has exhausted all possible available modes of communication. But in any case, when a person first attempts an activity he or she makes a conscious decision to pick one of the five primary modes. However, it may be the case that some of these alternatives are not available to the person or are inapplicable to the activity context or situation. Thus, if a *utility* could be associated with each medium for a certain

communication context setting, the ones not in the choice set are those with zero utility and assuming that people are utility-maximizers, the medium with the maximum utility is picked to accomplish the activity.

This utility is believed to be a function of the following [Moore and Jovanis, 1988]:

1. *Individual characteristics* (I) - characteristics of the initiator and recipient(s) and the relationship between them.
2. *Activity characteristics* (A) - the purpose of the activity, the communication context and other intrinsic factors such as urgency, complexity and desired level of security of the activity.
3. *Media characteristics* (M) - characteristics of the communication medium such as speed, cost, ease of use, interactivity etc.
4. *Situational constraints* (S) - fundamental constraints such as availability or accessibility of a certain medium in space-time.

Hence, depending on these and probably other unknown factors, a person decides to use a certain medium as the primary mode of communication to carry out the activity s/he has in mind. Having discussed the two aspects of the model (interactions and media choice), we will now concentrate on the first and investigate the second in some detail in the next chapter.

Let us first attempt to estimate the "amount" of communication that occurs in each medium. This amount or volume would be estimated by measuring, over some unspecified time window,

the number of activities that would *eventually* occur in the medium under consideration. Let us first begin by looking at an individual activity. Theoretically one could estimate the volume V_c (across all media) for the current activity as:

$$\mathbf{V}_c = G_c - E_c - S_c + SG_c \qquad \ldots 4.1$$

where the first two terms are direct effects of the current activity, representing activities that were **g**enerated or **e**liminated by the current activity, irrespective of medium. The third term is included because alternate activities, which are **s**ubstituted by the current activity, are in effect being eliminated by it. The last term represents "**s**elf-**g**enerated" activities and is an important yet not so obvious component of the equation which needs some further explanation.

Referring to Figure 4.2, we see that the set of activities could be viewed in terms of two dichotomies, the first being whether a given activity was prompted or not and the second being whether it had an alternative (i.e. whether it could have been carried out by some other means).

Figure 4.2 Self-generated Activities

	Alternatives Exist (A)	No Alternative (A')
Prompted by prior activities (P)		
Not prompted (P')		

P represents the set of current activities that were prompted, its complement **P'** would represent the set of activities that were not. Similarly, considering the other dichotomy, if **A** represents the set of current activities that did have alternatives, its complement **A'** is the set of activities that took place due to some idiosyncrasy of the current medium (DCN: CmC). The set of self-generated activities is represented by the shaded cells of the matrix. Thus, the final term of equation 4.1 is the number of elements in the set $A' \cup P' = (A \cap P)'$.

Keeping in mind that the goal of this exercise is to qualify the effect of the *current* activity and *not* activities that preceded it, the second definition $(A \cap P)'$ becomes clear. The activities that are being excluded are actually the effect of activities that prompted them and not the activities themselves. In that case shouldn't we be excluding all activities that *were* prompted (**P**)? No. The reason becomes clear if we consider the set of activities that were prompted but did not have an alternative. These activities would not have taken place if the medium they used did not exist (because they have no alternative). In other words these activities, though prompted by prior activities, are in reality taking place due to some unique feature of the "prompter activity - current medium" combination. The current medium is a characteristic of the current activity, and hence in that sense the current activity is responsible for its own generation.

We now have an equation (4.1) that represents the net number of activities (irrespective of the medium) eventually resulting from any given *individual*. While this is far from answering our original question about the volume of communication in any given medium, it has served its

purpose of identifying the various components that make up this volume. Let us now refine our model for volume further by adding a medium dimension to it and looking at the effect of a given activity **c**, carried out using medium **M**, on any generic medium **m**. Following the tone set by Equation 4.1, we compute the volume as:

$$V_{c(M),\,m} = G_{c(M),\,m} - E_{c(M),\,m} - S_{c(M),\,m} + SG_{c(M),\,m} \quad \ldots 4.2$$

It is to be noted that the last term for self-generated activities would be 0 for $m \neq M$ because an activity can only "self-generate" in the same medium. And also, ideally, the term for substitution would be 0 for $m = M$, because by its definition, substitution would only occur between complementary or alternate media. However, with the broad descriptions of the different media in the current study, this condition had to be relaxed because there was significant intra-medium substitution (electronic mail messages replacing phone calls though both are defined as electronic media).

Thus, any arbitrary medium m could be affected by activities conducted in that same medium (i.e., for which $M = m$) and by activities carried out in other media (i.e., for which $M \neq m$) and the sum over all such effects would provide the total volume for that medium. This total volume is simply given by:

$$V_m = \Sigma_c \, V_{c(M),m} \quad \ldots 4.3$$

In this study the current activity is always a computer-mediated communication. While this is an integral element of the study design (its purpose being to study the effects of computer-mediated communication), it has the result that a CmC activity (potentially from among several in a chain of events) will always be reported as the current activity.

This constraint has two implications: First, we are basically analyzing only the impacts of CmCs in this study, not the total impacts of all communications media. Thus, the picture presented here is necessarily incomplete. For example, even if it is found that, say, CmCs have a net reduction impact on travel, one could not know whether travel overall was increasing or decreasing without examining the impacts of all other communication modes on travel. The second implication is best explained through a hypothetical scenario: the information acquired through reading an article in a magazine (Activity I, communication through physical object) prompts a person to send an email to the editor (Activity II, electronic communication), reading which a fellow subscriber calls on the person (Activity III, electronic communication). In the current study setup, if the person involved (the respondent) chooses to report this scenario, s/he will be reporting Activity II as the current activity (II being the only activity involving CmC), Activity I as its parent and Activity III as its offspring. However, one could just as easily visualize Activity I or III being the current activity, with the concomitant changes in the parent and offspring activities. Thus, as the data are analyzed and the generation, elimination and modification of activities studied, one needs to bear in mind that the entire discussion is centered around CmC activities. To obtain a broader perspective on how activities and media usage interact, a more generalized version of the same study wherein the respondent is not constrained

to report on any specific medium could be carried out. This topic will be revisited under suggestions for future research in Chapter 6.

4.2 MACROSCOPIC ANALYSIS

Having laid out the conceptual framework for the analysis to follow, we are now ready to carry out the actual computation. The first step in calculating the net communication volume is to calculate the number of times that activities in each of the different media categories are reported to be generated, eliminated and substituted by the current activity. It is important to note that the numerical figures presented in the discussion that follows are to be viewed merely as qualitative indicators and not quantitative values for the actual number of activities generated, eliminated or substituted. The survey was designed to capture the existence of a certain type of effect and not the extent of the same. For example, if a person responds that some form of electronic communication was eliminated by the current activity, it could mean that one, seven or twenty seven instances of the same were eliminated.

Calculation of the net volume of communication is done by tabulating responses to the generated, eliminated and alternate activities questions in the Activity Diary survey. Next, the number of times an activity is not prompted by prior activities is calculated using the response to the "what prompted the current communication?" question. Finally, the number of self-generated activities is calculated by calculating the number of current activities that were either not prompted by parent activities ($\mathbf{A_p} \neq$ NULL) or did not have any alternate activities ($\mathbf{A_c}$, = NULL). The

specific questions used from the survey are given below and Table 4.1 gives the response tabulations to these questions. Table 4.2 gives the net volume of communication by medium, calculated as per equation 4.3.

- Did any previous activities PROMPT the current communication?
- Do you expect the current communication to GENERATE any additional activities?
- Do you expect the current communication to ELIMINATE any future activities?
- Do you expect the current communication to CHANGE something about any previously scheduled activities?
- If DCN had not been available for the current communication, would something else have likely occurred to complete the activity instead?

Table 4.1 below gives the distribution of the various types of impacts of DCN activities by communication medium. The union column provides the number of cases in which any one (or more) of the preceding media are reported. Referring to the table, 66% of all current communication activities reported were prompted by other activities, 65% of CCs generated some later activity that would not have occurred otherwise, 22% eliminated activities that were planned and would have occurred otherwise, and nearly 18% modified the details of previously scheduled activities. More than 50% of the current communications had an alternative medium which would have been used if DCN had not been available

Table 4.1a Distribution of DCN Impacts by Medium (N=636)

	In-Person	Electronic	Object Del.	Travel	Union
Prompted by	113 (17.8%)	321 (50.5%)	55 (8.6%)	-	419 (65.9%)
Generated	140 (22.0%)	351 (55.2%)	81 (12.7%)	81 (12.7%)	413 (64.9%)
Eliminated	66 (10.4%)	69 (10.8%)	57 (9.0%)	68 (10.7%)	137 (21.5%)
Modified	13 (2.0%)	26 (4.1%)	6 (0.9%)	15 (2.4%)	112 (17.6%)
Substituted	82 (12.9%)	248 (39%)	94 (14.8%)	110 (17.3%)	354 (55.8%)

Table 4.1b Self-generated Activities

	Alternative Exists	No Alternative	Total
Prompted	255	164	419
Not Prompted	99	118	217
Total	354	282	636

Now, using the methodology outlined at the start of this section and Equation 4.3, the values for the net volumes could be calculated using the values provided in the previous tables. These values are presented in Table 4.2 below.

Table 4.2 Net Volume of Communication by Medium

	Net Volume	Direction
In-person	-8	Decreasing
Electronic	415	Increasing
Physical Object/Surrounding	-70	Decreasing
Travel	-97	Decreasing

Let us run through the calculation of the net volume for the electronic medium to clarify the computation process. From Table 4.1a we find that 351 electronic activities were generated, 69 eliminated, and 248 substituted. These are the numbers corresponding to G, E and S respectively in equations 4.2/4.3. Now, from Table 4.1b we calculate the number of self-generated activities by considering all activities that *either had no alternative or were not prompted* (A' U P') and this value turns out to be the sum of the three shaded cells (118+164+99) giving us 381. This value corresponds to the SG term in the equations for volume. Putting them together we obtain a high positive value of 415 (351-69-248+381).

As is evident from the table above, the general effect of the use of DCN seems to be towards generating additional electronic communication (evident from a positive or increasing net volume of communication) and elimination/substitution of the other three media illustrated by a negative or decreasing net volume of communication. While we study these quantities the caveats discussed above need to be borne in mind, and specifically that the numbers indicate (qualitatively more than quantitatively) only the effects of current communications carried out using CmC.

4.3 MICROSCOPIC ANALYSIS

The previous section looked at the net changes in the communication carried out through different media. The computations suggest that there is a net reduction in communication through the travel and non-electronic media, while there is a net increase in electronic communication. While these results give us an overall indication of medium exchange, the exact nature of the interplay

between the different media is not clear. To study these interactions, a disaggregate microscopic analysis is conducted at the individual communication level. This investigation is motivated by two questions:

- What are the commonly occurring patterns of generation, elimination and modification (for example, is the generation of communication in one type of medium typically associated with elimination in another medium) ?
- What, if any, are the factors associated with communications that give rise to such patterns?

For the investigation of the effects of a computer-mediated communication, 12 variables were created based on the survey questions related to the generation, elimination and modification of activities. The variable names, which will be referred to often, have the following format:

GEN_X - generation of an activity carried out using medium X, where X is either I (in-person), E (electronic), O (physical object) or T (travel).

ELIM_X - elimination of an activity carried out using medium X, where X takes on values as defined above.

MOD_X - modification of an activity carried out using medium X.

The variables are binary, with a 0 indicating negative and 1 indicating affirmative. For example, if GEN_T were 1, it implies that the corresponding communication would generate travel and so on. Information on substitution was gathered from a question in the modify category which asked whether the medium of communication had changed. As substitution can be viewed as generation of activity in one medium and elimination in another, the answers to the substitution question were recoded to reflect generation and elimination in the respective media involved.

As the problem involves collating cases into suitable groups, an obvious method would be to simply sort the cases based on a string of the 12 binary variables. This was done as a first cut procedure to identify primary patterns in the data set. The result of such a procedure is given in Table 4.3 below. The *pattern* in the table below is simply a concatenation of the 12 variables listed above. The indicators appear in three groups of four, where the groups correspond to generation, elimination, and modification, respectively and the bits within each group correspond to I, E, O & T respectively. Thus, a pattern such as 0110 1001 0000, indicates that the group (or case) is most likely to *generate* activities involving electronic communication and communication through physical object (bits 2 & 3), *eliminate* activities involving in-person communication and travel (bits 5 & 8) and not *modify* any other activities. Several results are apparent: (1) a large proportion (more than 30%) of the communications do not result in either generation, elimination or modification. Such communications will henceforth be referred to as *Neutral Communications*. (2) The second largest group (more than 23%) involves just the generation of electronic communication, a result that seems to explain the positive value of the CFI index seen during the macroscopic analysis. (3) Most patterns in the table are relatively simple, containing two or fewer "ones". Only 5 patterns exhibit three or more simultaneous effects. (4) Although there are 2^{12} (4096) possible patterns, the 18 most frequent ones shown in the table account for fully 80% of the cases. While this indicates a remarkable degree of regularity in the data, it is desirable to see if some of the remaining 20% may be similar enough to the 18 patterns shown in the table (or to each other) to be grouped together. To explore this question, we introduce the technique of cluster analysis.

Table 4.3 Distribution of Patterns Obtained by Sorting 12 Variable String

Rank	Pattern	Freq	%Freq	Description
1	0000 0000 0000	191	30.03	Neutral activity
2	0100 0000 0000	149	23.43	Generation of electronic comm.
3	1100 0000 0000	26	4.09	Generation of in-person and electronic comm.
4	1000 0000 0000	15	2.36	Generation of in-person comm.
5	0110 0000 0000	15	2.36	Generation of electronic comm. and object
6	0101 0000 0000	15	2.36	Generation of electronic comm. and travel
7	1001 0000 0000	14	2.2	Generation of in-person and travel
8	1101 0000 0000	11	1.73	Generation of in-person, elec. and travel
9	0010 0000 0000	11	1.73	Generation of comm. via physical object
10	1110 0000 0000	10	1.57	Gen. Of in-person, electronic comm. and comm. via physical object
11	0100 1001 0000	9	1.42	Gen. Of elec. comm. and elim. of in-person comm. and travel
12	1111 0000 0000	7	1.1	Gen. Of comm. via all four media
13	0100 0010 0000	7	1.1	Gen. Of electronic comm. and elimination of comm. via physical object
14	0000 1001 0000	6	0.94	Elimination of in-person comm. and travel
15	0000 0100 0000	6	0.94	Elimination of electronic comm.
16	0100 0100 0100	5	0.79	Generation, elimination and modification of electronic comm.
17	0100 0100 0000	5	0.79	Generation and elimination of electronic comm.
18	0100 0000 0100	5	0.79	Generation and modification of electronic comm.
	Total	507	79.72	

Cluster analysis, though known by many other names such as numerical taxonomy (in biology) and unsupervised pattern recognition (in artificial intelligence literature), essentially addresses the following problem [Everitt 1993, p. 4]:

> *Given a collection of n objects, each of which is described by a set of p characteristics or variables, derive a useful division into a number of classes. Both the number of classes and the properties of the classes are to be determined.*

The solution generally sought is a *partition* of the *n* objects, that is a set of clusters where an object belongs to one cluster only, and the complete set of clusters contains all the objects. However, the conclusion of a classification investigation may well be that any such summary of the data would be misleading. Almost all clustering procedures involve three important steps, each closely related to the other:

1. The choice of the clustering dimensions or the variables on which the data set is to be clustered.
2. The choice of a (dis)similarity measure.
3. The choice of the clustering algorithm.

Step 1 of the process was briefly discussed above, where it was mentioned that 12 binary categorical variables were taken as the clustering dimensions. Let us now revisit this selection process. To quote Everitt [1985, p. 37] on the selection of variables:

> *The initial choice of the particular set of measurements used to describe each case to be clustered constitutes a frame of reference within which to establish the clusters; this choice presumably reflects the investigator's judgment of relevance for the purpose of classification.*

Consequently the first question to ask about the chosen variables is whether they are relevant to the type of classification being sought. As mentioned earlier, the variables selected in this study arose naturally from the context.

The next question concerns the number of variables that should constitute the dimension on which the individuals are clustered. Here the primary concern is to select all those variables that are deemed necessary to the clustering process while excluding those that are likely to obscure the "true" cluster structure. While in the most generalized clustering problems, the methodology adopted is to use some dimensionality reduction technique such as Factor Analysis and then cluster the individuals on the most appropriate factors, for the problem at hand a cursory analysis indicated which variables were likely to be redundant in terms of clustering information and hence to be excluded. Referring to Table 4.1 it is readily seen that the four variables related to modification have significantly smaller frequencies than the generate or eliminate categories, and hence are not likely to contribute significantly to the clustering process. This and reasons related to the weighting procedure (which will be discussed shortly) motivated the exclusion of these four variables from the clustering process.

Having selected the variables which would be used to characterize the individual unit, the next step involved the standardization and/or weighting of these variables. Standardization is a technique by which variables which have varying ranges and units are reduced to a common scale in the hope of eliminating any effect of magnitude or order on the clustering process. This was a non-issue considering the data at hand, which were binary categorical variables.

Weighting of variables is a means of giving one variable more importance than the other, and has been a rather controversial issue in the area of cluster analysis. Several authors [Sneath and Snokal, 1973] question the validity of such a procedure, arguing that as most weighting techniques are based on intuitive judgments of what is important, they may simply reflect existing classifications of the data or a biased view that the investigator takes. However, it is to be noted that the variable selection process itself involves variable weighting, with the variables that were not chosen essentially being accorded zero weights.

Most of the weighting approaches in the literature appear to be ad hoc and data driven. In our context, on the other hand, it was considered desirable to weight dimensions based on some a priori *conceptual* considerations that similarity on some variables was more important than similarity on others, or that matching might be a matter of degree rather than a binary outcome. For example, two communication activities, both of which generated new activities but in different modes, would be considered more similar than two activities, one of which generated an activity in one mode and the other of which eliminated an activity in a different mode. Ultimately, however, we were unsuccessful in finding a weighting approach which satisfactorily reflected these considerations, and hence fell back on the conventional approaches found in the literature. The relatively small number of clusters ultimately obtained, and the relative homogeneity within clusters, suggests that the conventional approaches are satisfactory in this context. Before we examine some variable weighting techniques, let us discuss Step 2 of the clustering process, viz., the calculation of similarity measures. In the section that follows a brief

outline of the commonly used similarity measures is given and the rationale behind selecting the one used in the study discussed.

4.3.1 Similarity Measures

All clustering methods operate on a notion of similarity or more specifically dissimilarity, with objects that are more similar forming clusters. Similarity can be measured in several different ways depending on the type of data and there exist several different measures for any given type as well. As the data type being dealt with in this study is binary categorical (dichotomous variables), we shall limit our discussion to those measures pertinent to this data type. The most common of these measures are what are termed *matching coefficients*, which are best explained in relation to Figure 4.4 below. To derive a matching coefficient for two cases *i* and *j* described by *p* binary variables, a tally of the number of matches is made. Referring to the figure below, **a** refers to the number of variables on which both cases score a 1; **b** is the number of variables on which case *i* scores a 0 and case *j* scores a 1; **c** is the number of variables on which the converse happens and finally, **d** refers to the number of variables on which both cases score a 0.

Figure 4.4. Match Table of Binary Variables

	Individual i			Total
Individual j		1	0	
	1	**a**	**b**	a+b
	0	**c**	**d**	c+d
Total		a+c	b+d	p

Using this match table as a basis, several similarity measures have been proposed and discussed extensively in the literature [Anderberg, 1973; Everitt, 1993]. Two of the most commonly used measures in cluster analysis are the Simple Matching Coefficient (SMC) and Jaccard's Coefficient defined below:

$$S_{\text{simple matching}} = (a+d)/p \qquad \ldots 4.4$$

$$S_{\text{jaccard}} = a/(a+b+c)$$

Thus, the SMC is simply the number of variables that match (1-1 or 0-0) over the total number of variables and is the probability that the two data units have the same score (either 0 or 1) on a randomly chosen variable. Jaccard's coefficient neglects the so called *negative* matches (0-0) in both the numerator and the denominator and is the probability that both the data units score 1 on a randomly chosen variable given that all 0-0 matches are discarded. The rationale behind discarding these 0-0 matches comes from the biological taxonomy perspective, where a zero on a variable (characteristic) might imply that the individual lacks that attribute. The argument is that the absence of an attribute does not necessarily make individuals similar. In the context of the problem at hand though, both 0s and 1s are equally important to defining similarity of two communication patterns. That is, it is just as important that two communications both did not lead to a generation as it is that they both did.

The other alternative is to measure dissimilarity rather than similarity. A common way of doing this is to use the Minkowski metric given by:

$$d_{ij} = [\Sigma_{r=1,p} \mid x_{ir} - x_{jr} \mid^{m}]^{1/m}$$

which reduces to the familiar Euclidean representation of distance between two cases for m = 2. The use of Euclidean distance as a dissimilarity measure has the conceptual drawback that categorical variables are not intuitively representable in Euclidean space. Also, this representation assumes that the variable values are uncorrelated [Everitt, 1993] with one another, which might not be the case in this study. An alternative that is suggested, incorporating possible correlation, is to use the more generalized *Mahalanobis* $\boldsymbol{D^2}$ metric, which for two individuals *i* and *j* is given by:

$$d_{ij} = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j)$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the vectors of measurements on the two individuals and **S** is usually taken to be the pooled within groups covariance matrix. The same conceptual drawback regarding representation of categorical variables that holds for Euclidean measures applies for the *Mahalanobis* $\boldsymbol{D^2}$ too. With Euclidean distance and Jaccard's coefficient deemed inappropriate for the data type and problem at hand, the Simple Matching Coefficient was chosen as the basis for the similarity measure used in this study.

Clustering algorithms operate on dissimilarities, rather than on similarities directly. There are several transformations that can be used to convert similarities into dissimilarities, and depending on these transformations they possess certain desirable properties listed below:

$$d_{ij} = 0 \qquad \text{iff } i = j \qquad \ldots 4.5a$$

$$d_{ij} \geq 0 \qquad \forall\, i, j \qquad \ldots 4.5b$$

$$d_{ij} = d_{ji} \qquad \forall\, i, j \qquad \ldots 4.5c$$

$$d_{ik} + d_{jk} \geq d_{ij} \qquad \forall\, i, j, k \qquad \ldots 4.5d$$

$$\max(d_{ik}, d_{jk}) \geq d_{ij} \qquad \forall\, i, j, k \qquad \ldots 4.5e$$

Measures that satisfy the first three properties (4.5a through 4.5c) are termed *semimetrics* [Anderberg, 1973], those additionally satisfying the *triangle inequality* (4.5d) are termed *metrics* and those satisfying all five conditions are termed *ultrametrics*. Johnson [1967] showed that certain hierarchical clustering algorithms such as single linkage and complete linkage induce an ultrametric on the data and this is used in some weighting techniques to be discussed shortly. In addition to this classification, a metric that is representable in Euclidean space is simply called a Euclidean. While a measure needs to be at least a semimetric to be operable, being a metric, Euclidean or ultrametric strengthens its qualities progressively. However, not all similarity measures can be easily transformed to yield dissimilarities that are ultrametric, Euclidean or even metric. While it intuitively appears that the properties listed above are desirable in a dissimilarity measure, there are no studies that demonstrate that the absence of at least the stronger properties (4.5d and 4.5e) is detrimental to the clustering process itself. In fact, Milligan [1980] in his study of the effects of several different types of error perturbations on various clustering algorithms, reports that the use of non-Euclidean indices and unstandardized variables tends to produce only slight decrements in mean recovery of cluster structure.

Gower and Legendre [1986] discuss in great detail the sufficient conditions for certain transformations to be metric and/or Euclidean. They show that if s_{ij} is the simple matching coefficient,

$$d_{ij} = 1 - s_{ij}$$

is a metric (possessing properties 4.5a - 4.5d) and that the transformation,

$$d_{ij} = \sqrt{(1 - s_{ij})}$$

is Euclidean. Presenting a generalized approach for investigating the metric and Euclidean properties of (dis)similarity measures, they show that if the similarity matrix **S** can be demonstrated to be positive semi-definite, then the dissimilarity measure $(1-s_{ij})$ is metric and $(1-s_{ij})^{1/2}$ is Euclidean respectively. However, this is only a sufficient condition and the measure could still be a metric (and Euclidean) if the condition doesn't hold. The next section describes the specific weighting algorithm adopted in the study and investigates the specific properties that the algorithm does (or does not) possess.

4.3.2 Variable Weighting

Now we are ready to discuss variable weighting strategies. The most prevalent variable weighting techniques involve the prescription of weights that give greater importance to variables that are able to significantly discriminate between the clusters [DeSarbo *et al.,* 1984]. DeSoete *et al.* [1985] describe an interesting technique that prescribes weights such that the computed distances approximate ultrametric distances as closely as possible in a least square sense. Milligan [1989] reports that this technique is also effective in identifying variables that are irrelevant to the

clustering present in the data, a very desirable property in a weighting procedure. Let us now take a brief look at this procedure.

Consider a weighted Euclidean distance similarity measure wherein a weight w_k is assigned to each of the p variables to yield:

$$d_{ij} = [\Sigma w_k (x_{ik} - x_{jk})^2]^{1/2}$$

The problem then is to determine the weights w_k which minimize the following expression:

$$\Sigma(d_{ik} - d_{jk})^2 / (\Sigma\Sigma d^2_{ij})_{i<j}$$

subject to the constraints that all weights $w_k \geq 0$ and the sum of the weights is 1. The sum of the numerator in the expression above is over the set of ordered triples which *violate* the ultrametric inequality described above (4.6e):

$$\{(i, j, k) \mid \max(d_{ik}, d_{jk}) \geq d_{ij} \text{ and } d_{ik} \neq d_{jk}\}$$

While the methodology is unimpeachable and the results encouraging, the assumption of a weighted Euclidean distance measure is not appropriate to the current problem and hence this weighting procedure, unfortunately, cannot be used.

A different approach to the weighting problem is to prescribe larger weights to matches that occur only sparsely in the population. The rationale behind this approach is that cases that match on attributes that occur frequently in the population are likely to be less similar than those matching on uncommon attributes. First suggested by Smirnov [Sneath and Snokal, 1963] for use

in categorical data with multiple classes, a modified weighting procedure for binary variables is suggested in this study and is described below.

Consider a sample of N cases, with each case characterized by *p* binary variables. Also, let N_k represent the number of times the k*th* variable assumes a value of 1 for the *population* of variable strings. As this value is not known, the frequency of occurrence in the sample, denoted by N_k, is used as an estimate of this figure and thus the estimated probability that the k*th* variable is 1 is given by:

$$p_k = N_k / N .$$

Thus, assuming that cases are independent in terms of the values that the variables may assume, the probabilities that any two cases have a certain pairing of the k^{th} variable are:

$$p_{0\text{-}0} = (1\text{-}p_k)^2$$

$$p_{1\text{-}1} = p_k^2$$

$$p_{0\text{-}1} = p_{1\text{-}0} = p_k(1\text{-}p_k)$$

$$p_{mismatch} = 2p_k(1\text{-}p_k)$$

The weight associated with this variable should be some inverted function of the corresponding probability. Two obvious alternatives exist,

$$w = 1 - p \quad \text{and}$$

$$w = 1/p \quad .$$

Anderberg [1973] in his discussion of a related method for weighting classes comments that the complement is not particularly discriminating and recommends the second function for inverting the probability.

Now, using the Simple Matching Coefficient described in equation 4.4 of Section 4.3.1 as a base, the weighted measure would be computed as:

$$S_{ij} = A / B \quad \ldots 4.6$$

where, $A = \Sigma_k [w[k]_{0\text{-}0} (1\text{-}x_{ik})(1\text{-}x_{jk}) + w[k]_{1\text{-}1} x_{ik} x_{jk}]$ . . . 4.6a

$B = A + \Sigma_k [w[k]_{0\text{-}1|1\text{-}0} \{(1\text{-}x_{ik})x_{jk} + x_{ik}(1\text{-}x_{jk})\}]$ . . . 4.6b

and x_{ik} = value of the i*th* case on the k*th* variable

Let us see which of the properties listed in Section 4.3.1 are satisfied by this customized similarity measure. Considering the situation when the two cases are identical, we observe that the second term in the denominator is zero and hence the similarity measure reduces to 1. Recalling that the dissimilarity d_{ij} is simply $(1- S_{ij})$, we see that the dissimilarity is 0 when the two cases are identical. Also, since the weights are all greater than zero, the second term in B can never be zero but for the situation that the two cases are identical, thus satisfying requirement 4.5a. The same reasons also prove that 4.5b is satisfied. The measure is also obviously reflexive, satisfying 4.5c. While checking for the first three properties was quite trivial, the triangle inequality does not yield itself to as simple a check. The only recourse is to check by exclusion, that is if we can find at least one triplet in the data set that does not satisfy the inequality, the measure fails for that property and is hence not a metric. Instead of checking for all triplets ($^{N}C_3$), a method which is computationally less intensive was used. It is based on a parameter θ, introduced by Gower and Legendre [1986]:

$$S_\theta = (a + d)/\{a + d + \theta (b + c)\} \quad \ldots 4.7$$

which defines a family of well-known similarity measures for different non-negative values of θ. For θ = 1, the parameter reduces to the simple matching coefficient. They show that 1- S_θ is metric for θ ≥ 1 and $\sqrt{(1-S_\theta)}$ is Euclidean for θ ≥ 1/3. Thus, if one could show that the θ corresponding to the measure S defined by 4.6 is greater than 1, the corresponding dissimilarity measure is a metric. However, given the definition of S, every pair of cases is likely to have a different θ. If any of these values are less than 1 then the corresponding dissimilarity measure *may* be non-metric. The value of θ for a pair of cases (the operation is $^{N}C_2$, and hence less computationally intensive than the comparison of all triplets) can be computed from the expression below, obtained by equating 4.6 and 4.7:

$$\theta = \frac{\Sigma \text{ (weights for all mismatches)}}{\Sigma \text{ (weights for all matches)}} \frac{(a + d)}{(b + c)}$$

It was found that several pairs yielded values of θ less than 1. However, while this does not prove that the measure will yield a metric, it does not disprove it either. While the status of the similarity measure remains inconclusive and not fully resolved, the results obtained from the clustering process and the fact that the similarity measure was definitely semi-metric, encouraged the use of this measure.

4.3.3 Clustering Algorithms

In the previous two sections we discussed similarity measures and variable weighting techniques. Now we turn our attention to the actual clustering algorithms themselves. The discussion that

follows is only cursory, and the reader is referred to the many excellent books and publications on the topic for methodological details [Sneath and Snokal 1963, Everitt 1993, Anderberg 1973, Fukunaga 1994, Hartigan 1975]. Two main classes of algorithms exist: *hierarchical* and *non-hierarchical*. In the former class of algorithms, the data are not partitioned into a particular number of classes or clusters in a single step. Instead the classification consists of a series of partitions which may run from a single cluster containing all individuals, to *n* clusters each containing a single individual. Most hierarchical algorithms yield a *dendogram* or tree structure of clusters. The non-hierarchical techniques on the other hand yield only one partition though the actual algorithm itself may be iterative. While the former class of algorithms has the drawback that clusters once formed at a certain level can never be undone later, they were chosen over the non-hierarchical methods for the following reasons:

1. Non-hierarchical techniques such as K-means need the specification of number of clusters and initial seed values. While several runs could be made for different numbers of clusters, the seed values pose a serious problem. Milligan [1980] in his paper examining the effect of different types of error perturbations on different clustering algorithms, reports that the "starting partition must be close to the final solution if the K-means algorithms are to be expected to give good recovery." Determining a starting partition close to the final solution is in general a difficult task, and more so in an exploratory study such as this one.
2. A more mundane problem arose from the requirements for running non-hierarchical algorithms. All the major statistical software packages that run K-means algorithms required as an input the original data matrix, from which Euclidean distances would be

calculated by the software. As discussed in Section 4.3.1, this measure being unsuitable for the problem at hand, it is inappropriate to use the algorithms on the data set.

Now among the many hierarchical algorithms, one that satisfied the following requirements was to be selected:

1. An algorithm that could operate on a customized dissimilarity matrix was needed.
2. The chosen algorithm was not to have significant biases towards finding clusters with certain properties. This was an important requirement as the nature of the "true" clusters was not clear.

Most algorithms satisfy the first requirement. However, many common hierarchical clustering algorithms have known biases or otherwise suffer serious recovery degradation in the presence of error perturbations. Some of these are described here. Single Linkage Clustering, which involves the joining of clusters at each stage by the single shortest link between them, is notorious for its tendency to "chain" or yield long serpentine clusters [Anderberg, 1973]. Centroid Clustering which involves the joining of clusters at each stage based on the distance between the centroids of the clusters, is known to impose a "spherical" solution on the partitions [Everitt, 1993]. Milligan [1980] reports that the Complete Linkage method wherein the intercluster distance is calculated as the maximum of the distances between individual cases in the two groups is highly susceptible to outliers. The procedure is also biased towards producing clusters with roughly equal diameters [SAS/STAT User's Guide, 1990]. Average Linkage Clustering wherein the distance between two clusters is computed as the average of the distances between the individual

cases in the clusters is known to be biased toward producing clusters with the same variance [SAS/STAT User's Guide, 1990].

After a thorough literature survey on the advantages and disadvantages of various clustering algorithms, it was decided that a variant of the Density Linkage Clustering algorithm using k*th*-nearest neighbor estimates [Wong and Lane, 1983] be used. In the paper describing the procedure, Wong and Lane [1983] prove that the kth-nearest neighbor clustering procedure is *strongly set-consistent* for high-density clusters, meaning that the procedure does not impose any geometrical structure on the clusters it produces. To quote from the SAS/STAT User's Guide [1990]:

> *Density linkage applies no constraints to the shapes of the clusters and, unlike most other hierarchical clustering methods, is capable of recovering clusters with elongated or irregular shapes. . . . If, like many people doing exploratory cluster analysis, you have no idea what kinds of clusters to expect, you should include at least one of the relatively unbiased methods, such as density linkage . . .*

Density Linkage Clustering is based on the intuitive concept of finding high density regions in a data set [Hartigan, 1975]. If objects are points distributed in an *N*-dimensional space (one for each variable), clusters may be thought of as regions of high density separated from other such regions by regions of low density. Consider a distribution of an infinite number of points characterized by a density $f(\mathbf{x})$ at each point $\mathbf{x}$. The number $f(\mathbf{x})$ is proportional to the number of objects per unit volume at the point $\mathbf{x}$. Hartigan shows that such clusters form a tree and hence

satisfy the informal requirement that the sample space consists of a hierarchy of high-density regions surrounded by low-density regions.

The k*th*-nearest neighbor method uses the k*th*-nearest neighbor density estimates of $f(\mathbf{x})$ and essentially involves two steps. Consider a closed sphere centered around **x** enclosing k neighbors with radius $r_k(\mathbf{x})$. The density estimate is first calculated as $f_N(\mathbf{x}) = k/(NV_k(\mathbf{x}))$, where $V_k(\mathbf{x})$ is the volume of the sphere and N the finite number of points. Using these estimates, a new dissimilarity measure is then computed as:

$$d^*(\mathbf{x_i}, \mathbf{x_j}) = (1/2)(1/f(\mathbf{x_i}) + 1/f(\mathbf{x_j})) \quad \text{if } \max(r_k(\mathbf{x_i}), r_k(\mathbf{x_j})) \geq d(\mathbf{x_i}, \mathbf{x_j}) \qquad \ldots 4.8a$$

$$= \infty, \quad \text{otherwise.} \qquad \ldots 4.8b$$

Hence, finite distances are defined only for pairs of observations which are in the same *neighborhood* (satisfying the condition associated with equation 4.8a). After this first step, the single linkage algorithm is applied to the new dissimilarity matrix to obtain a hierarchical partition. For a more detailed discussion of the methodology the reader is referred to Hartigan [1975] and Wong and Lane [1983].

The kth-nearest neighbor algorithm takes as an input the value of the *smoothing parameter* k. As k increases from 1 to N, the density estimate becomes smoother or less bumpy; that is, the number of modes of the density estimate increases. From an applications viewpoint, increasing k tends to pool clusters together with the opposite effect being seen for decreasing values of k.

Wong and Lane report that there is empirical evidence that the rule of thumb $k = 2\log_2 N$ is effective.

Having discussed all the three steps involved in a clustering problem, we now proceed to discuss the application of these methods to the data set at hand and the results obtained therefrom.

4.4. COMMUNICATION PATTERNS

The clustering procedure was carried out with Density Linkage Clustering as the primary algorithm. Clustering was carried out on both weighted and unweighted data. The weights used in the clustering process are given in Table 4.4 (N = 636).

Table 4.4 Variable Weights Calculated Using Equations 4.6

	GEN_I	**GEN_E**	**GEN_O**	**GEN_T**
N_k	140	351	81	81
p_k	0.22	0.55	0.13	0.13
$w[0\text{-}0] = 1/(1-p_k)^2$	1.64	4.98	1.31	1.31
$w[1\text{-}1] = 1/p_k^2$	20.64	3.28	61.65	61.65
$w[1\text{-}0\|0\text{-}1] = 1/[2p_k(1-p_k)]$	2.91	2.02	4.50	4.50

	ELIM_I	ELIM_E	ELIM_O	ELIM_T
N_k	66	69	57	68
p_k	0.10	0.11	0.09	0.11
$w[0\text{-}0] = 1/(1-p_k)^2$	1.24	1.26	1.21	1.25
$w[1\text{-}1] = 1/p_k^2$	92.86	84.96	124.50	87.48
$w[1\text{-}0\|0\text{-}1] = 1/[2p_k(1-p_k)]$	5.38	5.17	6.13	5.24

	MOD_I	MOD_E	MOD_O	MOD_T
N_k	13	26	6	15
p_k	0.02	0.04	0.01	0.02
$w[0\text{-}0] = 1/(1-p_k)^2$	1.04	1.09	1.02	1.05
$w[1\text{-}1] = 1/p_k^2$	2393.47	598.37	11236.00	1797.76
$w[1\text{-}0\|0\text{-}1] = 1/[2p_k(1-p_k)]$	24.97	12.75	53.50	21.71

As was discussed in Section 4.3.1, the very sparse occurrence of the four "modify" variables results in variable weights that are disproportionate. In order not to bias the clustering procedure, it was decided that the clustering would be carried out on the 8 variables corresponding to generation and elimination only. Tables 4.5a and 4.5b show the results of using the kth-nearest neighbor procedure on both weighted and unweighted data for the 8 variable case.

As is evident from the latter table, the procedure is unable to differentiate between the cases when they are not weighted. This is an interesting observation and is probably arising because of the fact that the Simple Matching Coefficient yields the same (dis)similarity for all pairs of cases having the same number of matches irrespective of the actual variables that match. On the other hand, the weights being unique for each variable, the weighting procedure yields different values for pairs that match on different variables even if the *number* of matches is the same.

The results shown are for a value of k=5 for the smoothing parameter. While the thumb rule estimate yielded k=6 ($2\log_2 N$, N=636), literature [Wong and Schaak, 1982] suggests that several values of the parameter in the vicinity of this estimate be tried. The results shown here were deemed to be the best considering the explanatory power of the clusters, the number of cases captured by the clusters in relation to the number of clusters formed.

A cursory comparison of the clusters obtained with the groups given in Table 4.3 indicates that the clustering procedure has been successful in correctly recovering all the "pure" clusters in the data set. Cluster #1 corresponds to the neutral cluster from Table 4.3 and cluster #2 corresponds to the second group, cluster #3 corresponds to the third group from Table 4.3, cluster #8 matches up with the fifth group from the table, cluster #9 is identical to the sixth, cluster #10 to the fourth, cluster #11 to the seventh group, cluster #12 corresponds to the eighth group from Table 4.3, cluster #14 is a superset of groups 16 and 17, cluster #15 is identical to group nine from the table, while the other groups "marry" some of the existing groups while including several more from the ones that were not grouped in Table 4.3.

Table 4.5a: Clusters obtained with Density Linkage Clustering with K=5 from 8-bit Effects Vector and Weighted Distances

CLUSTER	FREQ	%FREQ	GEN_I	GEN_E	GEN_O	GEN_T	ELIM_I	ELIM_E	ELIM_O	ELIM_T	MOD_I	MOD_E	MOD_O	MOD_T
1	197	30.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.02
2	158	24.84	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.01	0.01
3	31	4.87	0.26	0.90	0.45	0.10	0.65	0.55	1.00	0.71	0.06	0.06	0.06	0.00
4	27	4.25	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
5	19	2.99	0.16	0.58	0.00	0.00	0.95	0.00	0.00	1.00	0.05	0.00	0.00	0.00
6	17	2.67	0.12	0.82	0.00	0.18	0.00	0.00	1.00	0.41	0.00	0.06	0.00	0.00
7	17	2.67	1.00	0.71	0.71	1.00	0.41	0.00	0.00	0.35	0.12	0.06	0.00	0.24
8	15	2.36	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	15	2.36	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	15	2.36	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	14	2.20	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	13	2.04	1.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15
13	13	2.04	1.00	0.77	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00
14	12	1.89	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.17	0.50	0.00	0.00
15	11	1.73	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	10	1.57	0.00	0.00	0.30	0.00	0.00	1.00	0.00	0.00	0.00	0.30	0.10	0.00
17	9	1.42	1.00	0.67	0.00	0.00	0.00	1.00	0.00	0.00	0.22	0.11	0.00	0.00
18	5	0.79	0.20	0.60	0.00	0.00	1.00	1.00	0.00	1.00	0.20	0.20	0.00	0.00
19	5	0.79	0.00	0.40	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20	3	0.47	1.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00
Top 20	606	95.28												

Table 4.5b: Clusters obtained with Density Linkage Clustering with K=5 on 8-bit Effects Vector and Unweighted Distances

CLUSTER	FREQ	%FREQ	GEN_I	GEN_E	GEN_O	GEN_T	ELIM_I	ELIM_E	ELIM_O	ELIM_T	MOD_I	MOD_E	MOD_O	MOD_T
1	587	92.30	0.19	0.54	0.11	0.11	0.06	0.08	0.04	0.05	0.02	0.04	0.01	0.02

The algorithm yielded 25 modal clusters, of which the top 20 are displayed. These 20 clusters capture more than 95% of the cases. The value in each cell is the average for the corresponding variable and thus represents the proportion of cases in the cluster that scored a 1 on that variable. Thus, a cluster with a variable average of 1 implies that all members of the cluster possess that variable attribute and conversely, a variable average of 0 indicates that no member of the cluster possesses that variable attribute. Cells representing a frequency of 50% or greater are shaded, with frequencies of 50-66% having the lightest shading and those with frequencies between 84-100% having the heaviest.

The largest cluster (~31%) is that of communication activities that neither generate nor eliminate activities. This is a significant result in itself. In terms of the activity interaction model (Figure 4.1) for these activities, $\mathbf{A_g}$, $\mathbf{A_e}$ and $\mathbf{A_m}$ are essentially null sets. Let us now consider these neutral activities in some detail. Table 4.6 gives a cross tabulation of this cluster on the number of activities that were prompted and/or had alternative means of carrying them out.

The table below follows the format of Figure 4.2. We observe that the largest group in the neutral activities cluster is that of activities that are neither prompted by other activities nor do they have other alternatives. These could be termed *stand-alone self-contained* activities. One would expect that these activities are characterized primarily by relatively discretionary purposes such as browsing and social/entertainment related activities. This is precisely the case as is evident from Table 4.7.

Table 4.6 Crosstabulation of Neutral Communication Activities

		Alternative		
		1	**0**	Total
Prompt	**1**	36	49	85
	0	37	75	112
	Total	73	124	197

Table 4.7 Purposes Associated with Stand-alone Self-contained Activities

Purpose	**Freq**	**%Freq**
Browse	29	38.7
Work	11	14.7
Soc/Ent	14	18.7
Hobby	8	10.7
VolWork	4	5.3
Other	9	12
Total	75	100

It is to be expected that the kinds of scenarios that give rise to neutral activities are likely to be significantly different from those that give rise to the non-neutral or "active" communications. We hypothesize that neutral communications are likely to arise from communications that involve minimal interaction and those involving a one-time one-way transfer of information. A typical example of such an activity would be one involving browsing. To informally test these hypotheses, the distribution of purposes in the two classes of activities were compared. A χ^2 test ($p=0.0001$) revealed that the distribution of the top 5 purposes was significantly different.

Referring to Table 4.8, we observe that while the proportion of work related activities is smaller in the neutral cluster, the proportions of browsing and social/entertainment related activities increase. A similar comparison (Table 4.9) based on the distribution of applications used reveals that while the proportion of WWW use remains the same, email use decreases and the use of newsgroups increases for neutral communications compared to the active ones. The decrease in email is expected because most email based activities are very likely to either elicit responses (generation of activities) or to eliminate the need for the same. Also, keeping in mind that for the data set being examined the largest proportion of browsing activities is associated with newsgroups the other results also become clear. A χ^2 test ($p=0.004$) revealed that the distributions were significantly different.

Table 4.8 Comparison of Communication Purposes for Neutral and Active Communications

	Active		**Neutral**	
	Freq	**%Freq**	**Freq**	**%Freq**
Browse	52	11.85	50	25.38
Work	163	37.13	43	21.83
Soc/Ent	90	20.50	53	26.90
Hobby	47	10.71	22	11.17
VolWork	45	10.25	14	7.11
Shop	9	2.05	3	1.52
School	12	2.73	0	0.00
Travel	13	2.96	1	0.51
Other	8	1.82	11	5.58
Total	439	100.00	197	100.00

Table 4.9 Comparison of Application Usage for Neutral and Active Communications

	Active		Neutral	
	Freq	%Freq	Freq	%Freq
Email	296	67.43	109	55.33
WWW	58	13.21	28	14.21
Newsgroups	37	8.43	36	18.27
FTP	24	5.47	14	7.11
Other	24	5.47	10	5.08
Total	439	100.00	197	100.00

Figures 4.5 and 4.6 document the distribution of activity purposes and application usage in all of the 19 clusters.

Next, we move onto cluster #2 which constitutes nearly 25% of the cases in the data set and is characterized by the generation of electronic communication. In fact a glance through the clusters reveals that the generation of electronic communication is a feature that is common to several clusters. This supports the hypothesis made earlier that communication, particularly in a medium as facile as electronic communication, is likely to generate more communication. It also clarifies the positive value of net volume obtained for the electronic medium of communication, which indicated that there was a net increase in the use of electronic communication.

Figure 4.5 Activity Purpose Associated with Top 19 Effects Clusters

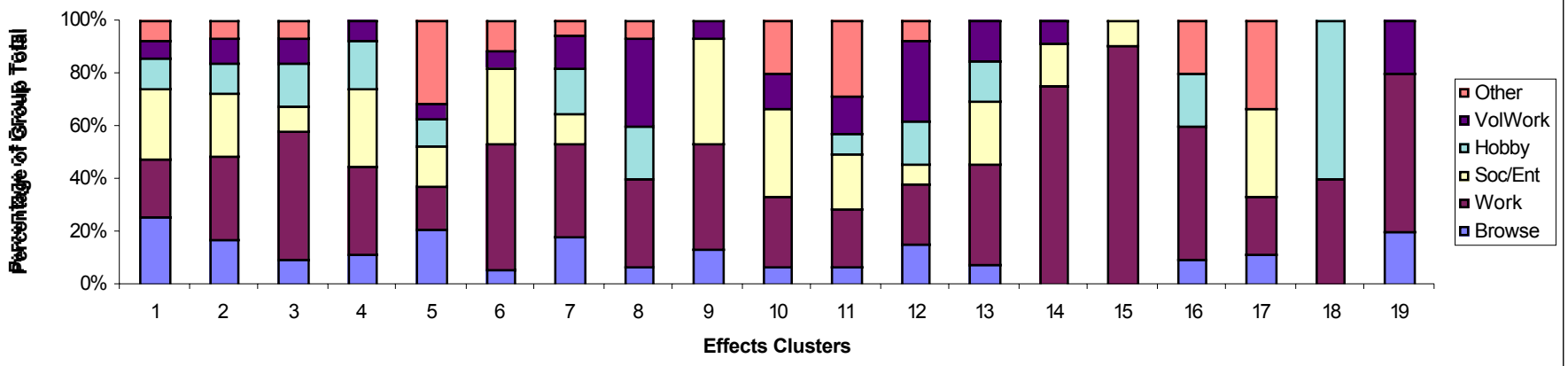
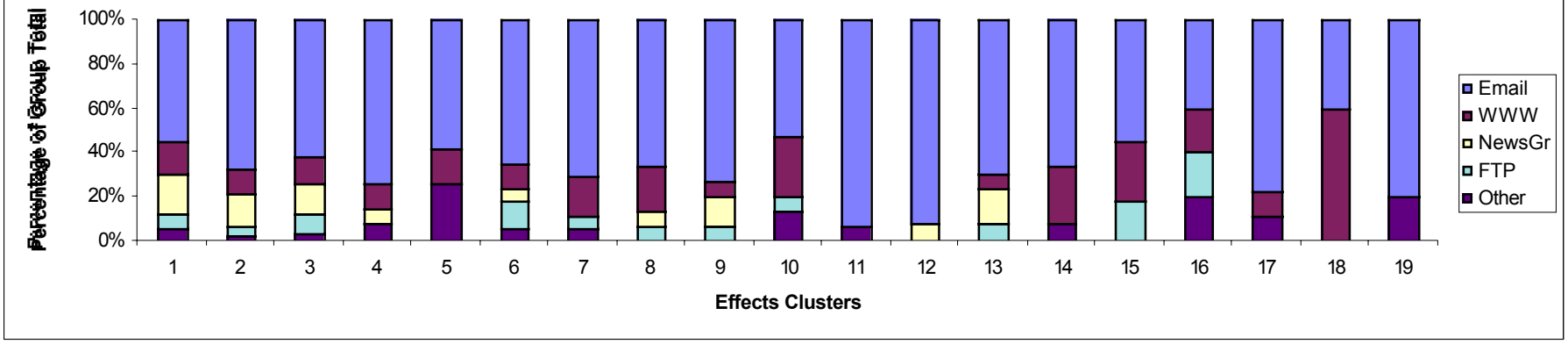


Figure 4.6 Applications Associated with Top 19 Effects Clusters



After the top two clusters, the cluster sizes decrease markedly. However, the cluster sizes are still big enough to command attention. Let us look at some of the more interesting clusters in some detail. Considering the third biggest cluster (comprising nearly 5% of the sample), we see that it is characterized by the near universal generation of electronic communication and the universal elimination of communication through physical object. The cluster members also exhibit occasional elimination of travel and the infrequent elimination of in-person and electronic communication. What we are seeing here is a classic example of medium substitution, with the current activity (carried out using the electronic mode) either eliminating the need for a future activity carried out using the OD/ODT media or substituting for the same. A significant cluster of such communications reassures us that people do in fact engage in activities resulting in medium substitution. A simple cross tabulation of the cluster members also reveals that more than half of the communications in this cluster are likely to result in the elimination of communication through in-person and/or travel. The cluster characteristic is embodied in this comment by one respondent:

". . . Email communication eliminates the necessity for an in-person meeting and formal transfer of review document hardcopy."

A cluster very similar to #3 is cluster #6 which consists of 17 (3%) cases. This cluster too is characterized by universal elimination of object delivery and frequent generation of electronic communication. However, no other medium is eliminated to a significant degree.

Cluster #4, characterized by the generation of in-person and electronic communications, captures more than 4% of the cases. This is typical of several other clusters where we see the generation

of communication activities using multiple media. Clusters 7, 8, 9 and 11 belong to this category. Anecdotal information provided by the respondents brings to light a typical communication scenario that characterizes these clusters:

". . . I downloaded software via Netscape (after reading) about a new version of netscape on the netcom homepage. . . . I'll tell friends and family about the new release."

In this example the current communication involves the downloading of the software. And the additional activities that are likely to be generated involve the dissemination of information regarding the new software via multiple modes (*electronic* and *in-person* in this case). However, in some cases the media that have been reported are alternatives and not used in conjunction as is evident from this scenario:

". . . a friend sent me a copy of a letter she sent to Pres. Clinton about Nuclear Dumping. I adapted it & sent it to Senator Boxer. . . . I expect a response to my letter of concern, probably by email, possibly by snail mail."

The generation of communication could occur via either the electronic medium (email) or via object delivery (snail mail). Though this issue will not significantly affect the results, it needs to be borne in mind as we analyze the data set.

Cluster #5 is yet another cluster exhibiting medium substitution and is characterized by the elimination of in-person communication and travel, accompanied by infrequent generation of electronic communication. Comprising nearly 3% of the cases, this cluster is described by the following communication scenario:

". . . Documentation for meetings and documentation that normally is delivered in-person will be substituted at times through e-mail, fax, and phone, which will eliminate possible traveling to City Hall and/or the Post Office for mailing products/labels and such."

While clusters such as #3, #5 and #6 support the fact that there is a significant portion of the communication universe where electronic communication can substitute for other traditional forms such as in-person, object delivery and travel, the existence of clusters such as #7, #11 and #19 supports the argument that the generation of activities using these traditional media is also possible. This latter group of clusters constitute essentially mirror images of the former, wherein traditional forms of communication are generated by the current activities. One communication scenario makes this point succinctly:

". . . I usually drive many times per week to the Bay Area. With e-mail, some travel is not necessary. Of course, I am also issueed (sic) invites by mail, so sometimes it leads to travel. The phone bill is less, though."

4.5 CONCLUSIONS

As was hypothesized earlier, there clearly are two main kinds of interactions taking place between the various communication media. In one type of interaction, electronic communication through DCN generates significant amounts of electronic communication which appears to be generated as an artifact of the very nature of the medium. This generation is accompanied in some cases by the elimination of the more traditional forms of communication, specifically the in-person, travel and communication through physical object modes, and in many cases by the

elimination of other instances of electronic communication such as the telephone. In the largest set of cases, however, the generation of electronic communication is unaccompanied by any other impacts.

The second main interaction involves the generation of communication conducted through the more traditional media. This latter set of interactions is smaller than the former and unlike the first, is likely to be influenced by factors such as the context of communication and the individuals involved, rather than by the medium itself. Thus, on the whole we are seeing a decrease in communication through media such as in-person, travel and communication through physical object and an increase in electronic communication.

It is interesting to note the evolution of the more "advanced" computer based systems: from teletype systems which were purely text (character) based, we have now reached a stage where multimedia messaging is quite common. Web browsers that allow the transmission and display of text and in-line graphics are almost ubiquitously available, if not yet widely used. Having conquered the static visual domain, systems that are truly multimedia, featuring real-time animation and audio are becoming increasingly common and in the coming years technology such as VRML (Virtual Reality Modeling Language) will make 3D animation and graphics just as common. While several questions remain as yet unanswered, it could be hypothesized that substitution of the more conventional media by advanced systems is likely to occur under the following scenarios: 1) where the communication process and any concomitant travel is tedious (grocery shopping for example) and 2) where primarily non-personal communication needs long-

distance travel as a precursor activity (computer-conferencing for example, in lieu of physical business travel). The subject of the exact scenarios under which substitution (and hence elimination) occurs is one of importance and must be taken up in future studies.

In this chapter we carried out analyses at both the macroscopic and microscopic levels to investigate the effects of communicating in a computer-mediated scenario in terms of the activities that are generated or eliminated. While the high-level analysis showed that there was a net reduction in non-electronic modes of communication and a net increase in communication using electronic modes, a more detailed analysis revealed that both generation and elimination of activities were taking place as initially hypothesized. In the next chapter we take an informal look at the factors influencing choice of various media of communication.

5
INVESTIGATION OF MEDIA CHOICE BEHAVIOR

In Chapter 4 we discussed the commonly occurring communication patterns in the data set and their distinguishing attributes. In this chapter, we turn our attention to the other major study question, that of media choice behavior. As was pointed out in Section 4.1 it is assumed that people engaging in communication activities are making conscious media choices to carry out a specific activity. Also, these media choices are influenced by several factors including the characteristics of the individual and the medium, the initiator-recipient relationship and the communication context [Moore and Jovanis, 1986]. In this section an effort will be made to explore the relationship between media choices (or "not choices", as will be explained shortly) and the factors affecting the same. In a future study, it would be ideal to attempt to model media choice behavior using traditional discrete choice models. Despite the use of non-traditional techniques several interesting exploratory analyses can be conducted with the data in hand and these are presented below.

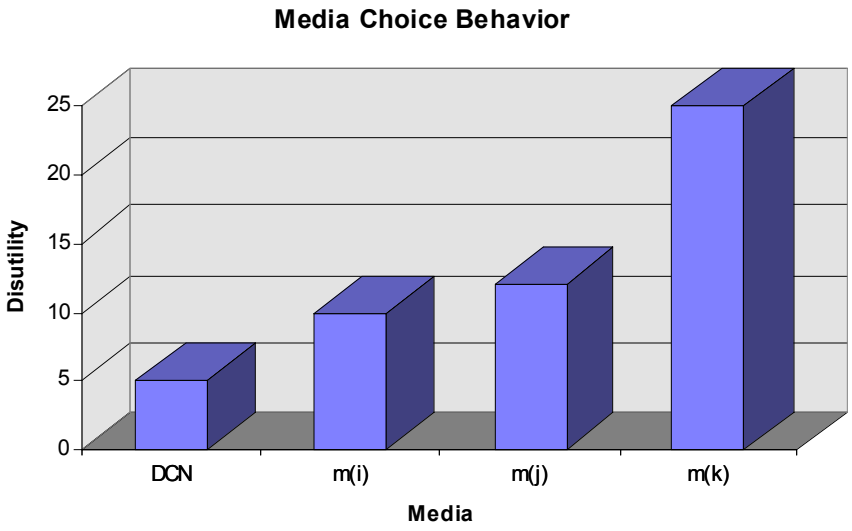
5.1 MEDIA CHOICE CLUSTERS

The topic of media choice has commanded considerable attention from researchers in various fields [Moore and Jovanis, 1986, 1988; Murray, 1991] and most studies on this issue have primarily attempted to answer the question, "How do individuals choose one form of communication over another?" Though the DCN study was not designed to answer such

questions, certain sections of the questionnaire revealed enough information to make for some interesting inquiry. The main data for this investigation comes from the question: *"If DCN had not been available, would something else have likely occurred to complete the current activity instead?"* The question is then followed up by asking about the specific media which might have been used and other relevant details. As must be obvious by now, the experimental setup in this study corresponds to the choice of one medium (namely, DCN which is an instance of a computer-mediated communication medium) over a set of others applicable for a given activity. And it is this set of media (that were not chosen in deference to DCN) that is revealed by the response to this question. To paraphrase, one might equivalently ask of the respondent: *"What other alternate media were you considering, and why were they* ***not*** *chosen?"* Thus, if a respondent reveals that media m_1 and m_2 were also in the choice set, one would like to know what factors contributed to their *non-choice* in favor of DCN. This essentially amounts to inverting the focus of conventional discrete-choice models, taking DCN as the base alternative, and looking at the factors contributing to the *disutilities* of the various media.

Consider the hypothetical scenario presented by Figure 5.1, wherein four media, namely DCN, m(i), m(j), and m(k) are mapped to their respective disutilities. The medium with the minimum disutility is chosen to carry out a given communication activity and in this study that medium is always DCN. However, we could look at the media that were included in the choice set and the reasons contributing to their disutility. While this simplistic scenario describes four unique media, it could well be that media occur in pairs or even triplets as was described in Section 4.1.

Figure 5.1 Inverted Media Choice Model



Thus, the first step in this investigation would be to identify those alternative media combinations that occur commonly in the data set. To do this a clustering procedure incorporating all the steps described in Section 4.3 was used. The specific variables used in the process were obtained from the responses to the question cited earlier: "*If DCN had not been available, would something else have likely occurred to complete the current activity instead ?"*. From the information obtained, nine binary variables were created to include each of the four major media categories: in-person (ALT_I), communication via physical object (ALT_O), travel (ALT_T) and electronic (ALT_E). To get more specific information on the type of electronic communication used, variables for *phone, fax, radio, TV, alternate Internet access* and *other* were created and used in lieu of ALT_E. Table 5.1 displays the characteristics for the top 17 clusters capturing nearly 95% of the 636 cases.

Table 5.1 Media Alternative Clusters Obtained by Density Linkage Clustering with K=5 and the Weights Used

Cluster	Freq	%Freq	ALT_I	PHONE	FAX	RADIO	TV	INTACC	OTHER	ALT_O	ALT_T
1	282	44.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	106	16.67	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	45	7.08	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.02	0.00
4	27	4.25	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
5	26	4.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
6	25	3.93	0.00	0.08	0.00	0.00	0.00	0.00	0.00	1.00	1.00
7	18	2.83	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	12	1.89	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
9	11	1.73	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
10	10	1.57	0.00	0.50	1.00	0.00	0.00	0.10	0.00	0.10	0.10
11	9	1.42	0.22	1.00	0.11	1.00	0.00	0.11	0.00	0.00	0.33
12	8	1.26	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00
13	6	0.94	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
14	5	0.79	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
15	5	0.79	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00
16	5	0.79	0.00	0.00	0.00	0.00	1.00	0.20	0.00	0.20	0.00
17	3	0.47	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.67
other	33	5.19	0.48	0.58	0.33	0.03	0.15	0.52	0.06	0.45	0.67
Top 17	603	94.81									
Total	636	100									

Variable Weights

	ALT_I	PHONE	FAX	RADIO	TV	INTACC	OTHER	ALT_O	ALT_T
N(k)	82	180	22	10	5	70	5	94	110
p(k) = N(k)/636	0.13	0.28	0.03	0.02	0.01	0.11	0.01	0.15	0.17
w[0-0] = 1/(1-p(k))^2	1.32	1.95	1.07	1.03	1.02	1.26	1.02	1.38	1.46
w[1-1] = 1/p(k)^2	60.16	12.48	835.74	4044.96	16179.84	82.55	16179.84	45.78	33.43
w[1-0] = 1/[2p(k)(1-p(k))]	4.45	2.46	14.97	32.31	64.10	5.10	64.10	3.97	3.50

Note: The procedure generated 18 modal clusters

The value in the cell under each of the alternatives is the average figure for that cluster and so is indicative of the proportion of cluster members choosing that media alternative. Thus, if the value is unity, then *all* cluster members chose that media alternative, while if a cell has a value of zero, then *none* of the cluster members chose that media alternative, with all other values falling in between these two extremes. Unlike the clusters in the previous chapter, we observe that the media alternative clusters are significantly more homogeneous and hence a simple shading scheme was used. Only cells with a value of one were shaded to indicate the universal presence of certain attribute in the cluster.

Table 5.2 gives brief descriptions of each of the 17 clusters. The first important observation is that the largest cluster (comprising more than 44% of the cases) represents those respondents who felt that DCN was their only alternative for the communication in question. This group is referred to as the *captive group* for the remainder of the chapter. The fact that no alternative media were checked implies that these respondents had a single medium in their choice set, namely, DCN, and that the respondents perceive DCN to be the *only* medium capable of performing the current activity. While it might be reasonable to assume that there exist several activities that can only be carried out via a computer-mediated system, the same argument cannot be extended to explain the non-inclusion of alternate Internet access. The fact that this medium is also not perceived as an alternative implies that either the respondent does not have access to an alternate Internet provider (analogous to not having a car, and hence having to take transit) or that DCN provides certain functionalities needed for the activity, which are unavailable in the alternate means of Internet access as far as the respondent is aware (analogous to a person

choosing a car with an automatic transmission instead of one with manual transmission, because the individual is incapable of driving the latter). Thus, as we analyze the factors influencing these media choices (or not choices), one needs to bear in mind that the captive group unfortunately represents both respondents who made a conscious decision (influenced by individual, media, and activity characteristics) to exclude alternate media from the choice set and those who had to exclude them due to situational constraints such as inaccessibility.

The next six clusters represent the *primary* medium alternatives described in Section 4.1, wherein phone and alternate Internet access (which could be either through personal subscription to a commercial service or through a service available at the workplace) are specific types of the broader electronic medium. We observe that travel rarely occurs apart from in-person or physical object communication (cluster #9 and cluster #17) and it is the conjecture that where it does, it is probably because respondents overlooked those forms of communication. The difference between the *pure* in-person cluster and the IPT cluster is obviously the fact that in the latter the respondent would have had to travel to some location other than from where the DCN activity was conducted to engage in in-person communication. Similarly, comparing the pure object delivery and the ODT cluster, in the latter the respondent would have had to travel to a different location in order to engage in a communication through physical object/surrounding. Thus, while sending a letter by snail mail would be an object delivery alternative, going to the library to pick up a book would be an ODT alternative.

Table 5.2 Cluster Descriptions

Cluster	Freq	%Freq	Media Alternative(s)
1	282	44.34	No alternative
2	106	16.67	Phone
3	45	7.08	Alternate Internet access
4	27	4.25	In-person and travel
5	26	4.09	Object delivery
6	25	3.93	Object delivery and travel
7	18	2.83	In-person
8	12	1.89	Phone and object delivery
9	11	1.73	Phone and travel
10	10	1.57	Fax and sometimes phone; occasionally other media
11	9	1.42	Phone and radio; occasionally other media
12	8	1.26	In-person, object and travel
13	6	0.94	In-person, phone and travel
14	5	0.79	Phone and Internet access
15	5	0.79	In-person, phone, object and travel
16	5	0.79	TV
17	3	0.47	Other electronic communication; frequent travel
other	33	5.19	Frequent travel, phone, Internet access, in-person or object delivery; occasional fax; infrequent TV and other electronic media

The next tier of 10 clusters indicates the use of multiple primary media. However, not all members in these clusters suggest the joint use of the multiple media. There are cases when the respondent feels that any of the media listed would have served as an alternative in the absence of DCN. For example, in cluster #8 while many respondents indicated the use of phone and object delivery occurring jointly, there were some who indicated that either one could serve as an

alternative medium. One solution to this problem might have been to identify all cases falling in the latter category and creating multiple occurrences, one for each of the media indicated as possible alternatives. The first major drawback in doing this was that several respondents did not provide enough explanations to clearly distinguish the activity. And secondly, pooling respondents who would have used *either* medium 1 or medium 2 in the absence of DCN, with those who considered medium 1 alone or medium 2 alone was not theoretically justified. This is because the former set of respondents viewed *both* medium 1 *and* medium 2 as being the next best alternative to DCN, while the latter group viewed only one of them as being the next most suitable alternative. And so for the group that chose phone as the medium of choice in the absence of DCN, communication through physical object was either not being considered as an alternative or was not ranked as having the same utility as phone. Because of these restrictions it was decided that the clustering would be conducted with the data as such and the results would be studied keeping in mind the caveats mentioned above. Also, the fact that the top seven clusters are primary mode clusters greatly reduces the severity of the "and/or" problem.

5.2 MEDIA CHOICE BEHAVIOR

Having thus obtained clusters representing various media alternatives, the next step was to correlate these clusters with various factors that are likely to influence media choice behavior. Recalling the discussion in Section 4.1 regarding these factors, it was pointed out that they could be summarized as belonging to one of the following four groups: Media Characteristics, Activity

Characteristics, Individual Characteristics or Situational Constraints. Of these, the current study provides, albeit in an indirect manner, information about the first three groups. Media characteristics were obtained from the question pertaining to the advantages of DCN (and hence, the disadvantages of the other alternative media) with regard to the current communication. Individual characteristics were obtained from the Background Survey, and some broad socio-economic variables were studied to identify possible correlations. And finally, the main source of contextual activity information was the question related to the purpose of the activity from the Activity Diary. Only minimal information on situational constraints could be gleaned from the respondents' activity descriptions and hence these could not be formally studied.

5.2.1 Media Characteristics

Information on media characteristics was obtained from the following question in the Activity Diary: *"Why was DCN chosen for the current activity?"* The respondents could check multiple reasons from a list of 11 advantages, so to speak, of DCN. These advantages were that the system was cheap (CHEAP), easy to use (EASY), was fast (FAST), provided high quality of information (HIQUAL), was less disruptive to the household (LDH), was less disruptive to the recipient (LDR), was used because there was no need for interaction (NOINT), was spontaneous (SPONT), the respondent wanted to try it out (TRYOUT), or DCN was the only option available (ONLYOPT). An *other* category was provided for responses that did not fit any of the above categories. Table 5.3 gives the proportion of respondents in each primary alternative

media cluster (i.e., the first seven clusters of Table 5.2) who checked the corresponding advantage.

Informally these values could be interpreted as being indicative of a quality *opposed* to the variable indicated for the alternative media. For example, when object delivery and travel is the alternative, the perceived advantages of DCN are that it is fast and easy to use - the implication being, faster and easier than traveling and delivering an object. When another means of Internet access is the alternative, the main perceived advantage of DCN is that it is cheap - i.e. cheaper than other Internet access alternatives. Comparing across alternatives, the degree to which a particular advantage of DCN is cited is indicative of the degree to which that quality is a *disadvantage* of the alternative in question. Thus, the fact that ease of use is cited as an advantage of DCN more often when the alternative is object delivery and travel than for any other alternative, suggests that the object delivery and travel alternative is least easy to use overall. To take another example, the fact that respondents in the cluster whose alternative was phone were more likely than any other group to cite "less disruptive to the recipient" as the comparative advantage of DCN, suggests that phone is the alternative perceived as the most disruptive to the recipient.

Using this argument and taking the highest number in each row of Table 5.3 (shaded), we see that compared to DCN: object delivery is the slowest, least spontaneous and (when travel is involved) most difficult; phone offers the lowest quality of information and is most disruptive;

and other forms of Internet access are more expensive for the communication activities in the sample.

Table 5.3 Reasons for Using DCN by Media Alternative Cluster[1]

	None	Phone	Alt. Int. Acc.	In-person +Travel	Obj. Del.	Obj. Del. +Travel	In-person
CHEAP	0.23	0.39	0.51	0.15	0.42	0.36	0.22
EASY	0.18	0.44	0.29	0.37	0.38	0.56	0.28
FAST	0.19	0.54	0.20	0.41	0.69	0.56	0.28
HIQUAL	0.07	0.14	0.04	0.04	0.08	0.12	0.06
LDH	0.04	0.20	0.00	0.19	0.12	0.08	0.06
LDR	0.10	0.41	0.00	0.19	0.12	0.08	0.22
NOINT	0.11	0.20	0.09	0.11	0.04	0.00	0.06
SPONT	0.11	0.16	0.11	0.22	0.27	0.04	0.06
TRYOUT	0.12	0.10	0.16	0.11	0.08	0.20	0.22
OTHER	0.04	0.08	0.04	0.04	0.08	0.08	0.00
ONLYOPT	0.59	0.08	0.16	0.41	0.12	0.08	0.44

A multiple comparison procedure was used to test the significance of the difference in responses of the respondents in different clusters. Unlike t-tests, where the confidence level is applicable to each of the comparisons individually, a multiple comparison procedure like Tukey's or Scheffe's method offers a confidence level for multiple comparisons collectively. Tukey's method was used in the current study as it applies to the family of all possible pairwise comparisons of group means.

[1] The highest number in each row is shaded.

Tukey's multiple comparison method at a confidence level of 95% revealed that a significantly higher proportion of respondents in the alternative Internet access (AIA) cluster checked *cost* as an advantage of DCN compared to those in the in-person and travel (IPT) alternative cluster and

the captive group. Also, a higher proportion of respondents with phone as the alternative checked cost as an advantage over the captive group. One could infer that the cost of the system was perceived as an advantage of DCN when compared to both phone and the AIA alternative. Looking at it from a media choice perspective, one could argue that cost is likely to be an important factor adding to the disutility of both. An explanation could be that DCN during the period of this study was a free service. Also, an Internet service lets the user access information and contact people virtually anywhere in the world for the cost of a phone call. And in the case of DCN and this study group, the phone call would have been a local phone call, making it considerably more attractive than a commercial Internet service which would involve a service fee or at least a long distance call (there being no local Internet access providers in Davis at that time).

For the ease of use variable, both the phone and the object delivery and travel (ODT) alternatives had a significantly higher proportion of respondents than the captive group. While it is understandable that the ODT alternative is seen as quite odious, it is interesting to note that the phone alternative also ranks rather low in the ease of use scale. One explanation could be that when respondents check ease of use as an advantage over phone, they are considering not only ease of "use" but also ease of "accessing" the recipient. Several respondents have raised this

issue while quoting ease of use as an advantage as was evident in a study of their activity explanations.

A significantly higher proportion of respondents in the object delivery (OD), phone and ODT clusters checked speed as an advantage when compared to the captive group and the AIA cluster. Also, the OD cluster had a significantly higher proportion of cases than the in-person alternative. It is hypothesized that "speed" here refers to the speed with which communication can be conducted, i.e. information received or conveyed, which would involve several factors such as how facile the system is, how fast a "connection" can be made with the recipient and so on. While OD and ODT being viewed as significantly slower media than a computer-mediated one comes as no surprise, it is interesting to note that phone is also perceived to be a slower medium. This again could be because of the issue that people find "getting hold of the recipient" to be rather difficult and hence time consuming.

Regarding the LDH variable, the phone cluster had a significantly higher proportion of cases than both the captive group and the AIA cluster. Similarly, regarding the LDR variable, the phone alternative cluster had a significantly higher proportion of cases than all the other groups except in-person communication. This is quite obvious because very few of the other alternatives involve direct interpersonal communication. The asynchronous nature of CmCs gives them a further advantage in this area.

Finally, regarding the "only option" alternative, it is not surprising that a significantly higher of cases in the captive alternative have checked this option in comparison to the other groups. It is however perhaps surprising that a significantly higher proportion of cases in the in-person and

IPT alternative clusters selected this option in comparison to the phone alternative cluster. Superficially there appears to be a contradiction here: these respondents have earlier said that if DCN weren't available they would have communicated in person, but here they indicate that DCN is the only option available. The resolution is likely to be found in the fact that performing the communication in person would have been so onerous that DCN is considered the only *practical* alternative. But the imperfect correlation between responses to these questions is illustrative of the complexity of responses to this survey.

No significant differences were seen across clusters for the variables dealing with quality of information, spontaneity, need for interaction or experimentation.

5.2.2 Activity Characteristics

Researchers have concluded that the activity context plays a very important part in the decision process [Moore and Jovanis, 1982]. In the current study the only contextual information that was available was from the question regarding the purpose of the activity. The questionnaire gave the respondent a choice of nine purposes. Table 5.4 crosstabulates those purposes against

the seven primary clusters identifying media alternatives. A χ^2 test ($p \sim 0.00$) indicates that the distribution of purposes differs significantly across cluster.

Stacked bars representing the relative proportions of different media alternatives for each communication purpose are shown in Figure 5.2 and are quite illuminating. This particular representation was chosen because, though the clusters are the primary variables under study, the fundamental issue that we are attempting to investigate is that of media *choice*, specifically, when given a certain activity with a purpose, how does an individual select a specific medium to carry out the activity. And this question was best answered by studying the distribution of clusters used for different purposes rather than the converse.

Considering browsing, we observe that the predominant medium is DCN itself. Browsing, as was discussed earlier, is primarily carried out through the World Wide Web or via Newsgroups. Thus, while it is not surprising that very few of the other traditional alternatives appear as media alternatives to DCN, it is noteworthy that the AIA cluster does not stand out either. One explanation for this may be found in the fact that most alternate Internet access (for the data set under consideration) is through the work place, where access to the Internet may be limited to facilities such as email and to a more limited extent newsgroups. It is only recently that many organizations and companies have started providing full access featuring the WWW accessible through browsers. DCN on the other hand provides complete Internet access making it the ideal candidate for browsing activity.

Work, on the other hand, can be accomplished through several media. In fact it is the activity purpose for which we see the most diverse distribution of media alternatives, with phone (~20%)

being the primary alternative to using DCN. Communications conducted for social and entertainment purposes also had a fairly broad distribution of media alternatives with phone again being the primary alternative. Phone and object delivery being the predominant alternatives suggests that most of the activities which are social/entertainment in purpose may be primarily long distance interpersonal communications. In-person and even IPT constituting a relatively small proportion of the work-purpose activities indicates that little substitution is likely in these categories. Stated another way, people are likely to continue to see each other (whenever geographically possible) despite the availability of email and other related CmCs.

The communication purpose of extended shopping which primarily included information acquisition prior to the actual purchase and software downloads, logically had IPT and ODT as the primary alternatives to DCN. One respondent who was using DCN to acquire information about cars prior to purchase (purpose: shopping) and would have visited several dealers instead (media alternative: IPT), writes:

> *"eliminated options (choices of car), helped to make the decision to purchase ... will not go to various dealers to talk about cars that I have now [after reading newsgroups] eliminated from my options."*

Another respondent who used DCN to download a graphics application writes:

"I shall not have to continue fruitlessly touring computer places in Sacramento looking for software."

Thus, the two alternatives relate to the information acquisition and actual purchase of the software. As DCN did not support any teleshopping facilities at the time of this study, only a limited number of shopping related communications were observed. But, as these anecdotes indicate, there is some scope for substitution of travel if such applications were to become widely available.

For the school purpose, which mainly involved the electronic transaction of assignments and student-teacher interactions, IPT was the primary alternative which constituted more than 30% of the cases in this group (with another 40% being captive to DCN). This result is quite natural as other modes are rather unsuitable for the activities that such a purpose entails. Volunteer work had phone making up more than 40% of the cases (with another 45% being captive to DCN). As most of the volunteer work activities involved scheduling meetings and other related organizational activities, this result is also quite natural. Travel and other purposes constitute a relatively small proportion of the total and did not exhibit any interesting characteristics.

The hobby category, which primarily involved information acquisition about topics relating to one's personal interests and other areas of self-development, is very similar to the browse category in terms of media alternative distribution. Hobbies pursued via the Internet usually

encompass the entire gamut of topics available via traditional means, including specialized and even "exotic" subjects.

5.2.3 Individual Characteristics

The next set of factors that were evaluated reflected variables that were expected to have some correlation with choice behavior. The individual characteristics that were chosen were *age, computer experience, gender* and *income*. Cross-tabulations were performed to see if any interesting patterns emerged and the results are given in the stacked bar charts in Figures 5.3a-d.

As we saw from the results in Section 5.2.1, the choice of a medium is likely to be related to the issue of how a user perceives a medium to fare on time, cost and convenience. While these are very much dependent on the context of the activity, the value that a user attributes to time, cost or convenience would be dependent on the individual's characteristics. To assess the value that people in different age groups or income levels attach to media characteristics, we now look at the distribution of these individual characteristics in the different alternative media clusters.

Starting with age, different age groups are likely to value the three characteristics differently. Time may be more of a concern for younger users than for those who are older (especially retirees). Thus, one may expect a significantly higher proportion of the former group to engage in communication activities that save time, and these would include the use of electronic media as opposed to travel or communication via physical object. Also a feeling of loneliness in older

people might motivate them to choose media that are perceived to be more intimate such as in-person communication or telephones. Referring to Figure 5.3a, this might be reflected in a disproportionately large number of older respondents (55-64) considering in-person communication as an alternative than their younger counterparts (25-34). Unfamiliarity with new technology may be yet another issue, where older respondents are likely to be less familiar with computer-mediated systems than their "Internet-age" counterparts. We also see that the middle aged group (35-44) contributes only minimally to the clusters involving media combinations that are rather time consuming, namely, ODT, IPT or OD, whereas they form the largest single group in the clusters corresponding to the phone, IP and AIA alternatives. This however may be more of a reflection of their income status than their age. We see this in Figure 5.3d, wherein the respondents in the higher income groups constitute a significant majority in the AIA and IP clusters.

A χ^2 test performed after collapsing the age groups into four categories 15-34, 35-44, 45-54, and 55+ yields a p value of 0.032 suggesting that the effect of age on media choice might not be as strong as was hypothesized earlier. But, bearing in mind that significant interaction effects are likely to exist between the different individual factors the arguments might not be entirely invalid. A similar test performed on household income after collapsing the cases into two categories (<55K, >55K) yielded a very small p value (~ 0.00) suggesting that respondent's in the two income groups had significantly different media choice behavior. One significant difference, as

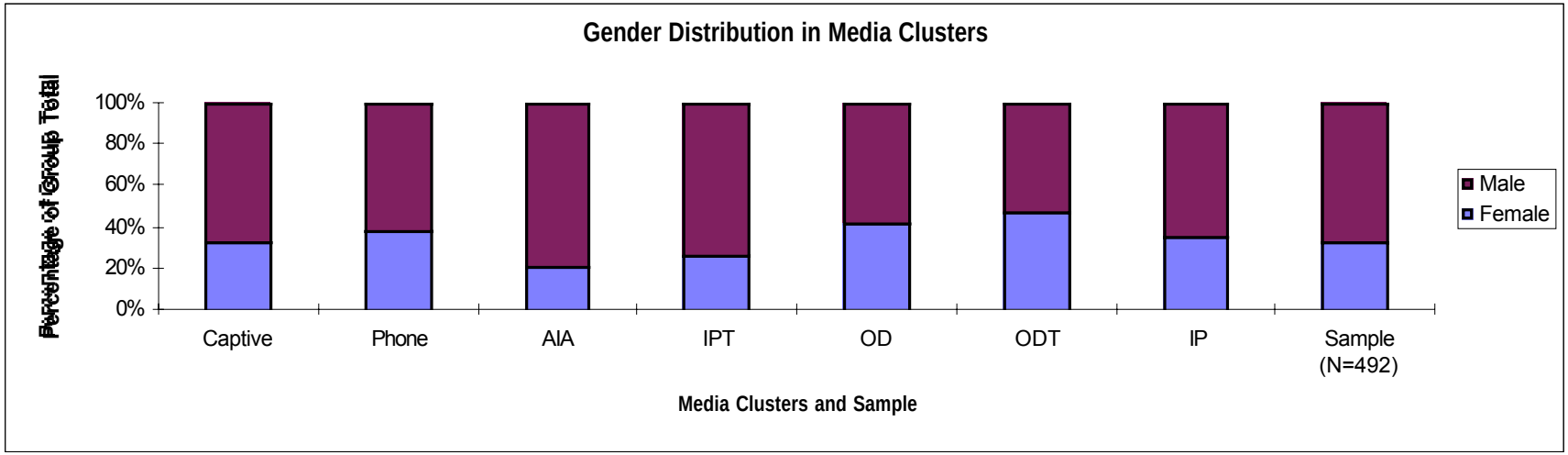
can be observed from Figure 5.3d, is that respondent's with higher household incomes are in some sense less captive and have more options.

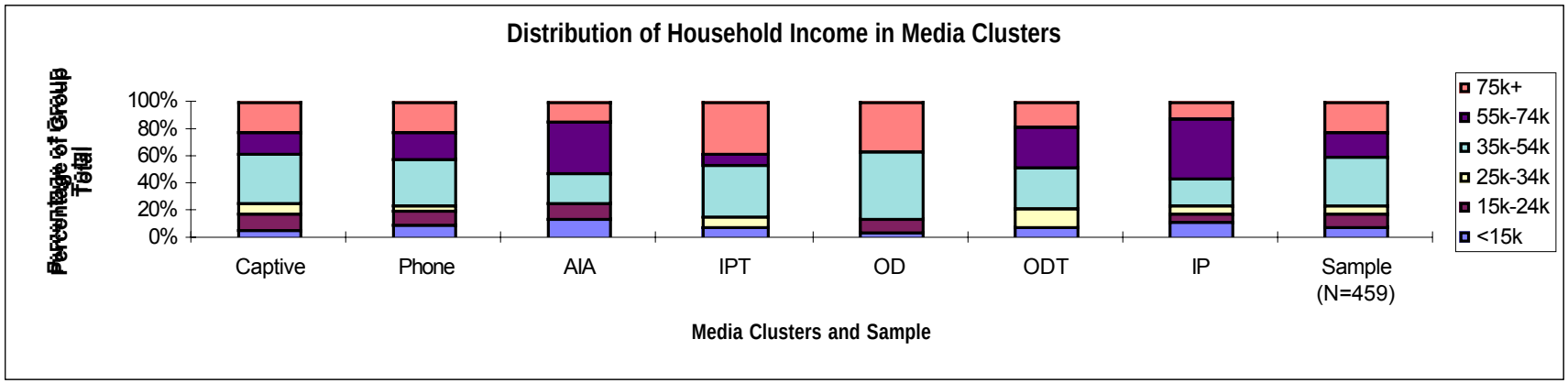
Considering computer experience, the sample not surprisingly has a rather skewed distribution, with nearly 57% of the respondents having considerable computer experience (more than 8 years). One hypothesis that can be formulated is that respondents with considerable computer experience are likely to be more aware of and hence comfortable with new communication technology and so are more likely to consider related media as alternatives. This might be demonstrated by a complete absence of novice users from the AIA cluster while they are present in some number, albeit small, in the ODT, OD and phone clusters, clusters representing the more traditional media. A χ^2 test revealed a p value of close to 0 on aggregating the cases into two categories of respondent's with less than 8 years of computer experience and those with more than 8 years, suggesting that computer experience has a strong influence on media choice behavior.

Regarding gender, the data set is composed primarily of males (nearly 67%). The fact that the AIA cluster is composed of a significantly smaller proportion of females might be indicative of a gender bias with respect to computer usage and familiarity, particularly Internet-related applications. However, this is likely to be correlated with variables such as age and occupation, with a young female professional being just as familiar and comfortable with CmC as her male

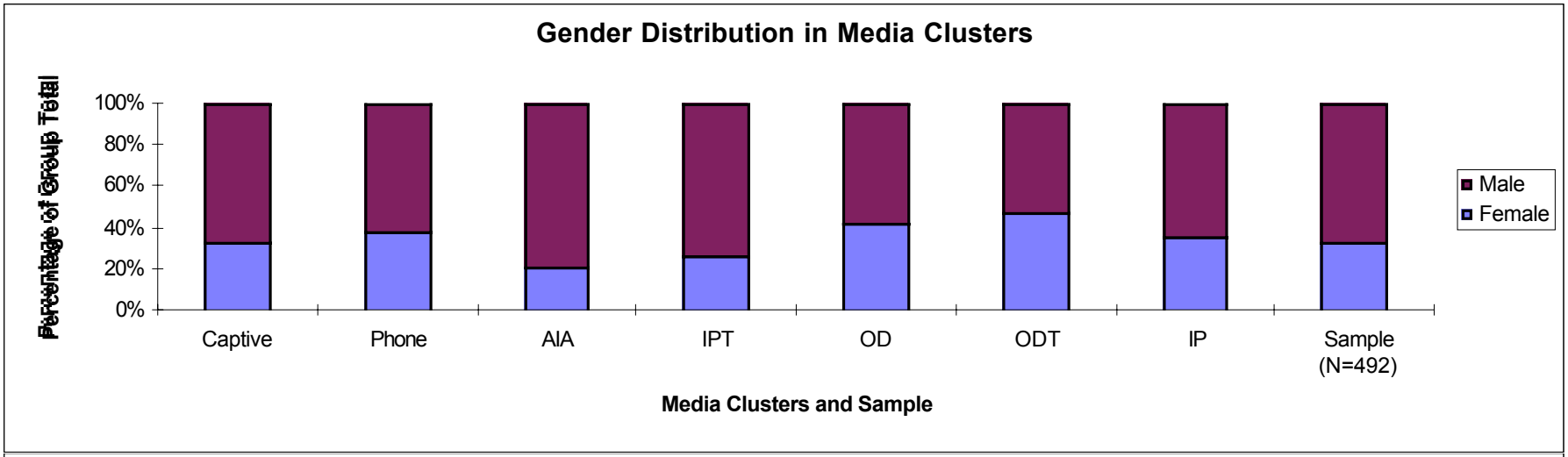
colleagues. Value of time might be another issue influenced by gender. To the extent that female respondents work less than males, they may be more likely to engage in discretionary travel, and this might be reflected in the higher proportion of female respondents in the ODT and OD clusters. As before, a chi-squared test was performed and the small p value suggests that significant differences exist between males and females in terms of their media choice behavior.

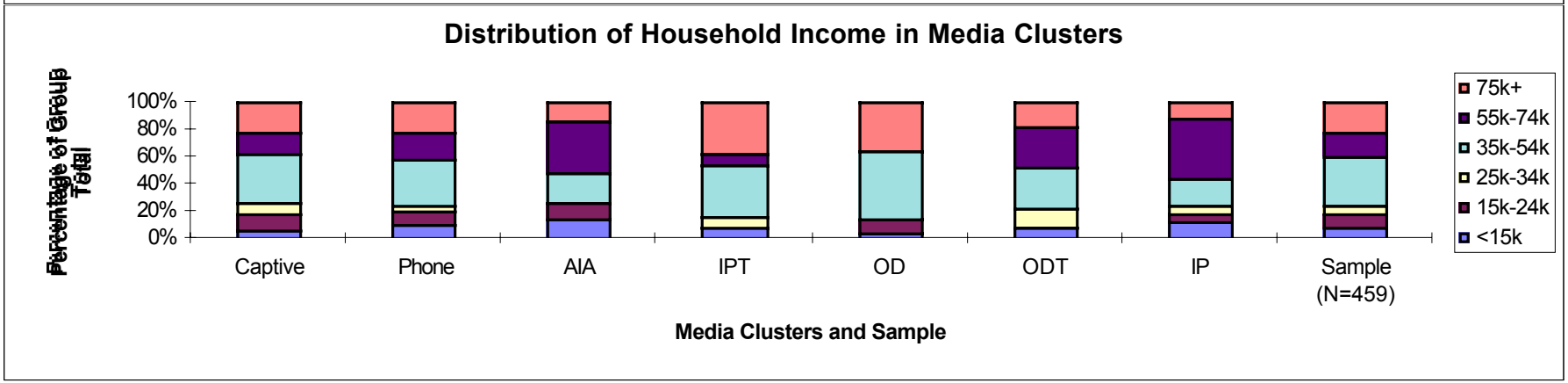
Figures 5.3a-b. Distribution of Age and Computer Experience in Different Media Alternative Clusters





Figures 5.3c-d. Distribution of Gender and Income in Different Media Alternative Clusters





5.4 CONCLUSIONS

To summarize, the seven largest clusters represented one captive group and six clusters corresponding to the set of primary communication media. Media, activity and individual characteristics were studied. Cost was cited as an advantage of DCN over phone and alternate Internet access. Both phone and the travel & object delivery alternatives were perceived to be more tedious than DCN. Speed (or rather lack thereof) was perceived to be a disadvantage of the OD, phone and ODT alternatives. With regard to disruption to the household and the recipient, both phone and in-person alternatives were perceived to be significantly disruptive when compared to computer-mediated systems. Thus economy, convenience, speed and non-disruptiveness appear to be important media characteristics influencing media choice behavior.

The purpose of the activity was found to be significantly related to the media an individual considers and evaluates. Work related activities did not seem to favor any particular media cluster. However, phone and object delivery were the preferred alternatives for the social/entertainment activities, which seems to suggest that most of the social/entertainment activities that are carried out using CmCs might be primarily long-distance interpersonal relationships. For geographically close interactions in-person communication still remains the preferred medium. Activities involving extended shopping (information acquisition prior to purchase and software downloads) show great promise in terms of substitutability. If explored further these appear to be the activities that are most likely to use CmCs in lieu of more traditional media.

Though individual characteristics did not show any significantly striking patterns, it is hypothesized that attributes such as age, gender, income and computer experience are likely to influence media choice behavior. The influence of the first three attributes, however, may become weaker in the future as new technology and CmCs in particular become ubiquitous.

Chapters 4 and 5 covered the two major topics of investigation for this study. Chapter 4 revealed that while on the whole there is likely to be a decrease in communication using traditional media and an increase in communication using more advanced systems, at a finer level both generation and elimination of activities were taking place. In Chapter 5 we discussed some of the factors that are likely to influence media choice behavior, and while the analysis was rather informal it is hoped that it will pave the way for more rigorous studies in this area. In the next and final chapter, we lay out in more detail some of the important conclusions and suggestions for future research.

6

CONCLUSIONS & SUGGESTIONS FOR FUTURE RESEARCH

In the preceding chapters, a conceptual model for the interaction of communication activities was laid out. Guided by this model, analyses were conducted to investigate, at both a macro and micro-level, the interplay between communication activities in a Computer-mediated setting. As a sub-model, the process of media-choice behavior was also investigated and primary factors influencing choice of communication media identified. In this concluding chapter, we try to reconcile and summarize the key results and insights gained from this exploratory study. While this study focused on a narrow set of activities and interactions, it is a small but important step towards understanding the broader ramifications of living in a wired world.

6.1 CONCLUSIONS

Let us first discuss the question "*How does the availability and usage of one medium of communication (computer-mediated in this case) affect communication activities conducted via other media?*" raised at the beginning of this thesis. To set the stage for the discussion let us begin by recapitulating the methodology used in this study. The data used in the analyses was obtained from 636 communication descriptions obtained from the DCN Evaluation project. In each communication description DCN system users described a communication activity carried out using the DCN system, detailing its purpose, the specific application used and the initiator(s) and recipient(s) involved. The users also described any previous activity that might have prompted the current activity and activities that might be generated, eliminated or modified due

to the current activity. Finally, the users described alternative media that they might have used in order to carry out the current activity and also gave the perceived advantages of using DCN over these media.

To give structure to the analyses that were to follow, a conceptual model was laid out to describe the interaction of communication activities. Borrowing heavily from ideas used to develop the survey described above, the Activity Interaction Model forms the basis for the two stage (macro and micro) analysis conducted to answer the question raised earlier: "*How does the availability and usage of one medium of communication (computer-mediated in this case) affect communication activities conducted via other media?*", and will now be briefly summarized. An activity is defined as a reaction to stimuli, where the stimuli themselves are outcomes of the need for fulfillment at some psycho-physical level. An activity is further characterized by a *purpose* and *medium*. The interaction model described in Figure 4.1 proposes that activities form a sequence or more specifically a directed graph (directed in time) with each node possibly associated with multiple parents and multiple offspring activities. A parent activity is defined as an activity causally responsible for its offspring. The specifics of the analyses center around a "current" activity which is used as a reference for the graph. For every such current activity there could hypothetically be a set of complementary or alternate activities, activities which *might* have been carried out to serve the same purpose as the current activity if the current medium had been unavailable or inaccessible in some way. Finally, the offspring of the current activity are classified into three broad categories - activities that were *generated*, *eliminated* and *modified* by the current communication.

Having defined the various categories of activities involved, the interaction model then describes a sub-model involving a media-choice process leading to the selection of the medium of the current activity. Five primary media categories are identified - In-person communication (I), Communication via physical object/surrounding (P), Electronic communication (E), Travel and in-person communication (TI) and Travel and communication via physical object/surrounding (TP). It is also argued based on the utility-maximizer theory that if a utility could be associated with every medium in a given context, an individual would choose the medium with the maximum utility. This utility itself is believed to be a function of the characteristics of the Individual, Activity, Medium and the Situation or context of the communication.

With these concepts in place, the study first attempted to understand the effect of the usage of a fairly new type of communication medium specifically the DCN system (a form of Computer-mediated Communication) on the "activity space" of individuals. Will it change the size of the activity space, i.e., will there be more or fewer activities? What kinds of trade-off will occur between different communication media, will the new medium substitute for more onerous traditional media or will it generate activities carried out in the same medium or other media?

To answer the first of these questions a macroscopic analysis was conducted to measure at an aggregated level the volume (in a qualitative sense) of communication conducted in any particular medium in the context of Computer-mediated communication. We again stress that this is a qualitative measure as no formal means were used to quantify and correlate communication in

different media (talking for 3 minutes on the phone is not equal to reading 72 pages with an Arial typeset!). This volume was essentially measured as all the activities that were generated in that medium less the activities that were eliminated either directly or indirectly via substitution. Activities that were considered to be self-generated were also added to this tally. A self-generated activity was defined to be an activity that did not have a parent or prompter activity or if it did, it did not have any alternatives.

This simple computation yielded some very significant results. Almost all media showed new activities being generated due to the use of the DCN system. This immediately affirms the hypothesis that communication activity interactions are not a zero sum game and that the use of a particular medium is capable of giving rise to new activities not only in that medium but in other media as well, contributing to a possible increase in Activity Space. Also, all media showed either a net decrease in volume (in-person, object delivery and travel) or a small increase (electronic communication) if one only considered the number generated and subtracted the numbers eliminated or substituted. However, the large number of activities that were self-generated in the electronic communication (the category under which CmC or DCN activities fell) medium led to a large net increase in volume for this medium. These self-generated activities are rather interesting, because they are in some sense a result of the availability of this new medium, once again, contributing to the increase in Activity Space. And finally, we do see substitution of one medium for another and more - specifically the DCN system (a form of Computer-mediated Communication) - for travel, indicating that it is very conceivable that in certain contexts, advanced communication media could be used as an alternative for travel.

While the macroscopic analysis revealed gross trends, a microscopic analysis was called for to unearth the actual interplay between the various media. Using similarity and cluster analysis, communication activities were grouped based on their *communication patterns*. A pattern was coded as an 8-bit vector, with the first four bits indicating the presence or absence of generation in each of the four basic media types and the next four indicating the same for their elimination. A Density Linkage Clustering technique (K=5) with a Simple Matching coefficient similarity measure and a customized weighting procedure was used. The algorithm partitioned the dataset into 25 modal clusters, the top 20 capturing more than 95% of the data. The clusters obtained and their attributes once again make for some interesting observations.

The largest cluster comprised Neutral activities (approximately 31% of all descriptions), activities which neither generate nor eliminate other activities. This is a very interesting and significant finding in itself, showing that for this sample, the largest chunk of computer-mediated communication is one-time, stand-alone and self-contained dominated by browsing activities. Whether directed information acquisition or undirected passive-information intake, it is not yet clear whether these activities result primarily because of the characteristic of the activity (is the individual interested only in gathering some information as opposed to interacting with other individuals) or whether they result due to the inherent limitations of the medium (with the exception of electronic mail and chat facilities the Internet is still rather limited in interactivity). It would be interesting to see if the share of neutral activities declines as the Internet and related CmC media become more interactive.

The second largest cluster involved the generation of Electronic Communication, a feature common to several other clusters too. While corroborating the positive value obtained for the medium's volume earlier, given the medium's ease of use, the result comes as no surprise.

After the top two clusters the cluster sizes decrease markedly, but can generally be differentiated into two broad categories: clusters exhibiting substitution of traditional media by electronic communication and clusters exhibiting generation of communication via traditional media themselves. This dichotomy underscores the hypothesis presented at the outset of this thesis that communication and information in general will only increase communication through all media, accompanied by media-substitutions at the margin.

The final step in this study was to make an informal attempt at understanding what prompts these media-shifts by exploring the underlying mechanisms of media-choice behavior. Following the basic ideas described earlier in the media-choice sub-process, characteristics of the individual, medium and activity were investigated to identify influential factors. However, as the survey was not designed with this express purpose in mind, a conceptual transformation had to be performed on the data to be able to view it in a media-choice context. Information on what the users would have done had DCN not been available and the quoted advantages of DCN over the other prospective media were used to carry out the analyses. As in the previous case each communication description was coded as a 9-bit vector which formed the basic clustering element. These nine bits corresponded to the presence or absence of the following media in the

choice set - in-person, phone, FAX, radio, TV, alternate internet access, physical object, travel and other miscellaneous. An identical clustering technique was used and 17 major clusters were identified. The largest cluster was of those activities for which the respondent believed there were no alternatives to DCN. The next six clusters included the five primary media mentioned earlier and alternate internet access.

Having identified these clusters the next step was to identify the media, activity and individual characteristics affecting the non-choice of one as opposed to another. The media characteristics were identified from responses to the question describing the advantages of using DCN, the activity characteristics were identified by the purpose of the activity and the individual characteristics were identified by the respondents' socio-economic characteristics.

Among media characteristics, *economy, ease of use, speed* and *non-disruptiveness* were found to be the most significantly influential factors. DCN was perceived to be significantly cheaper than either telephone usage or alternate internet access, probably due to the fact that Internet access via DCN enabled users to communicate with far-flung elements for the cost of a local phone call (DCN was a free service at that point in time). DCN was also perceived to be easier to use than either phones or the ODT medium. The activity descriptions reveal that in their measure of "ease of use" users include the ease with which the recipient can be accessed. Also, DCN was perceived to be faster than phones, object delivery and the object delivery + travel medium. Phones and in-person communication were perceived to be very disruptive media from the recipient's perspective, while computer-mediated systems (DCN and other internet access) were

perceived to be less disruptive to the household in general. In general computer-mediated systems seem to offer the advantages of being cheaper, faster, easier to use and less disruptive than more traditional media such as phones or object delivery.

Purpose was the key indicator of activity characteristic evaluated in this study. The most notable result is the acknowledgment of a passive-information intake activity called "browsing" or "web surfing". The Merriam Webster's dictionary defines the word browsing as:

"to look over or through an aggregate of things casually especially in search of something of interest"

A site on the Internet [http://www.sandybay.com/pc-web/surf.htm] describes "web surfing" as:

"To move from place to place on the Internet searching for topics of interest. ... The term surfing is generally used to describe a rather undirected type of Web browsing in which the user jumps from page to page rather whimsically, as opposed to specifically searching for specific information."

Characteristic of CmC media, surfing or browsing is a passive-information intake activity because the actor does not engage in it with a conscious effort to acquire *specific* knowledge or information. Not unlike browsing books in a bookstore, this activity is even more undirected and even more dependent on finding something "interesting". This is a fairly unique phenomenon with users being known to spend several hours each day surfing the web, with many users claiming to be "addicted" to the activity. However, it is usually restricted to novice users and as the fascination wears off the person is likely to engage in more directed or active-information acquisition activities.

The other interesting observation arose from shopping activities. Shopping in this context was defined to include both the pre-transaction information acquisition process and the actual business transaction itself. The Internet with its abundance of information on all kinds of products and services is the ideal source for engaging in the first half of the shopping process involving information acquisition. However, the types of products and services for which it will provide a viable alternative to the IPT medium (in particular) are those that provide minimal sensory gratification in the information acquisition (or window shopping) phase. It is unlikely that teleshopping will substitute for a trip to the mall, the pleasure of trying out different clothes or even eating at the food court! Fairly mundane activities, with minimal interpersonal interaction or sensory gratification and products and services which can be chosen based on well defined characteristics are the best choices for teleshopping and hence substitution of CmC media for travel.

Finally, Individual Characteristics such as age, income level, gender and computer experience were studied. It was hypothesized that individual characteristics would most likely influence the value attached to each of the media characteristics (time, cost, ease of use) discussed earlier. Some evidence was found to corroborate the fact that older users (more than the youngest groups) valued time much less and preferred more intimate and personal media such as phones and in-person communication. Unfamiliarity with technology might have been another factor contributing to this result. Income levels (greater than $55K and less than $55K) were found to be significantly different in their choice of media, simply the availability of additional choices for

the higher income levels might be an important factor. Computer experience (less than 8 years and greater than 8 years) was also found to be an important distinguishing factor and so was gender.

In conclusion, strong evidence was found to support the hypothesis that the availability and usage of a communication medium, specifically Computer-mediated Communication in the DCN context, is likely to result in an increase in communication activities in general. In this particular study, significant substitutions were seen between medium, wherein electronic media substituted for more traditional media such as in-person communication, object delivery and travel. However, the use of a new medium (DCN) also generated new activities carried out in both this and other media. A set of activities fairly unique to the specific medium of communication were also found. The single largest cluster of activities carried out via the DCN system were found to be neutral, neither eliminating nor generating other activities. Cost, speed, ease of use and non-disruptiveness were found to be the most influential factors affecting media-choice behavior. Income, gender and computer experience were found to be influential individual factors affecting media choice. However, these differences might not sustain with computer systems and other communication technology becoming more ubiquitous by the day.

6.2 SUGGESTIONS FOR FUTURE RESEARCH

There is scope for significant work in a wide array of related areas. Some of the key issues that merit further discussion are:

- A more generalized study of the inter-relationships between activities conducted via different communication media. This would involve a survey designed to assess the effects of communication in one medium on activities in other media and would not be restricted to computer-based communication alone. The challenge will be in designing a study that can be manageable and clear, while collecting relevant and useful information.
- A study aimed specifically at studying media-choice behavior using discrete-choice models should prove insightful. Incorporating qualitative contextual and situational information will be critical and can provide significant understanding into the mechanisms governing media-choice behavior.
- System statistics on application usage, temporal patterns and activity purpose can be a goldmine of information. Such statistics are routinely collected and maintained by network administrators and their study can yield considerable information regarding community network users and their behavior.

LIST OF REFERENCES

Anderberg, M. (1973) *Cluster Analysis and Applications*. Academic Press, London.

Beamish, A. (1995) *Communities On-line: Computer-based Community Networks*. Submitted to the Department of Urban Studies and Planning in Partial Fulfillment of the Requirements for the Degree of Master in City Planning at the Massachusetts Institute of Technology. On-line WWW address: http://alberti.mit.edu/arch/4.207/anneb/thesis/toc.html

Claisse, Gerard and Rowe, F. (1993) Domestic Telephone Habits and Daily Mobility. *Transportation Research A*, Vol. 27A, No. 4, pp. 277-290.

DeSarbo, W., Caroll, J. and Green, P. (1984) Synthesized Clustering: A Method for Amalgamating Alternative Clustering Bases with Different Weighting of Variables. *Psychometrika*, Vol. 49, pp. 57-78.

DeSoete, G., DeSarbo, W. and Carroll, J. (1985) Optimal Variable Weighting for Hierarchical Clustering: An Alternative Least-Squares Algorithm. *Journal of Classification*, Vol. 2, pp. 173-192.

Everitt, B. (1993) *Cluster Analysis*. Edward Arnold, London.

Fisser, P., Weghuis, J. and Slotman, K. Chapter 4: Computer-Mediated Communication. *On-line and Distance Learning Textbook '95.* Faculty of Educational Science and Technology, University of Twente, Netherlands. On-line WWW address: http://www.to.utwente.nl/ISM/ONLINE95/CAMPUS/LIBRARY/ONLINE95/CHAP4/chap4.HTM

Fukanaga, K. (1994) *Introduction to Statistical Pattern Recognition*. Academic Press, London.

Gower, J. and Legendre, P. (1986) Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification*, Vol. 3, pp. 5-48.

Handy, S. and Mokhtarian, P. (1993) *Evaluating the Travel and Communication Impacts of Advanced Residential Telecommunications Services: Lessons from the Literature*. Prepared for the State Department of Transportation under Research Technical Agreement No. 65T449. UCD-ITS-RR-93-23, Institute of Transportation Studies, University of California, Davis, CA 95616.

Hartigan, J. (1975) *Clustering Algorithms*. John Wiley & Sons, New York.

Henderson, D. and Mokhtarian, P. (1996) Impacts of Center-Based Telecommuting on Travel and Emissions: Analysis of the Puget Sound Demonstration Project. *Transportation Research D*, Vol. 1, No. 1, pp. 29-45.

Henderson, D., Koenig, B. and Mokhtarian, P. (1996) Using Travel Diary Data to Estimate the Emissions Impacts of Transportation Strategies: The Puget Sound Telecommuting Demonstration Project. *Journal of the Air and Waste Management Association*, Vol. 46, No. 1, pp. 45-57.

Johnson, S. (1967) Hierarchical Clustering Schemes. *Psychometrika*, Vol. 32, pp. 241-254.

Kellerman, A. (1984) Telecommunications and the Geography of Metropolitan Areas. *Progress in Human Geography*, pp. 222-246.

Koenig, B., Henderson, D. and Mokhtarian, P. (1996) The Travel and Emissions Impacts of Telecommuting for the State of California Telecommuting Pilot Project. *Transportion Research C*, Vol. 4, No. 1, pp. 13-32.

Maslow, A. (1954) *Motivation and Personality*. Harper, New York.

Meyburg, A. (1983) The Substitutability of Communication for Transportation. Paper presented at the *Third World Conference on Transport Research*, Hamburg, F. R. Germany.

Milligan, G. (1980) An Examination of the Effect of Six Types of Error Perturbations on Fifteen Clustering Algorithms. *Psychometrika*, Vol. 45, No. 3, pp. 325-342.

Milligan, G. (1989) A Validation Study of a Variable Weighting Algorithm for Cluster Analysis. *Journal of Classification*, Vol. 6, pp. 53-71.

Mokhtarian, P. (1988) An Empirical Evaluation of the Travel Impacts of Teleconferencing. *Transportation Research A*, Vol. 22A, No. 4, pp. 283-289.

Mokhtarian, P. (1990) A Typology of Relationships between Telecommunications and Transportation. *Transportation Research A*, Vol. 24A, No. 3, pp. 231-242.

Moore, A. and Jovanis, P. (1986) Analyzing the Choice of Telecommunications Media and Transportation Modes in Business Organizations. *Research For Tomorrow's Transport Requirements: Proceedings of the Fourth World Conference on Transport Research*, Vol. 1, pp. 24-39. Vancouver, British Columbia, Canada.

Moore, A. and Jovanis, P. (1988) Modeling Media Choices in Business Organizations: Implications for Analyzing Telecommunications-Transportation Interactions. *Transportation Research A*, Vol. 22A, No. 4, pp. 257-273.

Murray, D. (1991) *Conversation for Action: The Computer Terminal As Medium of Communication*. John Benjamin's Publishing Company, New York

Nilles, J., Carlson, F., Gray, P. and Hanneman, G. (1976) *The Telecommunications-Transportation Tradeoff: Options for Tomorrow*. Wiley, New York.

Salomon, I. (1985) Telecommunications and Travel: Substitution or Modified Mobility? *Journal of Transport Economics and Policy,* Vol. 19, No. 3, pp. 219-235.

Salomon, I. (1986) Telecommunications and Travel Relationships: A Review. *Transportation Research A,* Vol. 20A, No. 3, pp. 223-238.

Sampath, S., Saxena, S. and Mokhtarian, P. (1991) The Effectiveness of Telecommuting as a Transportation Control Measure. *Proceedings of the ASCE Urban Transportation Division National Conference on Transportation Planning and Air Quality*, Santa Barbara, CA, July 28-31, pp. 347-362.

SAS/STAT User's Guide. (1990). SAS Institute, Raleigh, NC.

Sneath, R. and Snokal P. (1973) *Principles of Numerical Taxonomy*. W. H. Freeman and Company.

Wong, A. and Schaak, C. (1982) Using the Kth Nearest Neighbor Clustering Procedure to Determine the Number of Subpopulations. *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 40-48.

Wong, A. and Lane, T. (1983) A kth Nearest Neighbour Clustering Procedure. *Journal of the Royal Statistical Society B*, Vol. 45, No. 3, pp. 362-368.