Internet-Based Social Contacts

Marijn Haverbeke
marijn@haverbeke.nl

Supervisor:
Wout Ultee

4th November 2004

Contents

Chapter 1

Background

The Internet has rapidly become prominent in the lives of almost everybody in the developed world. Internet-based means of communication are widely used, both as a replacement for more conventional ways of communication and in new ways. An example of such a new phenomenon is the practice approaching total strangers to ask for help, play a game, or flirt. This is uncommon offline, but customary online. The Internet contains a large variety of communication channels, such as newsgroups, chat boxes, electronic mail, games and bulletin boards, where one can communicate with whoever else happens to be using that channel.

Because the medium is so new and still developing quickly, it is not clear what effect these new forms of communication will have on society. As with practically every big new phenomenon in the past, some people are claiming this will radically change society as we know it—some say for the better, others say for the worse. The optimists claim the ability to communicate with people all over the world will increase understanding between people from different backgrounds, and strengthen democracy by providing a place where public discussion can take place (Rheingold 1993, Norris 2001). The pessimists, on the other hand, stress the difference between Internet communication (which is mostly text-only at this time) and 'real-life' communication, claiming that the former is not 'real' and only draws people away from real social contact (Rusciano 2001, Wellman & Gulia 1999). On both extreme sides of the spectrum outlandish claims are being made, but both sides also have reasonable theorists.

The 'realness' of online communication, however, is not something this thesis will concern itself with. Assuming they are at least somewhat real, these contacts with other people over the Internet form a social network between people. While these contacts are limited by the fact that they often consist only of text, they are often easier available than other contacts, which is a valuable property. Rheingold (1993) puts it this way:

> "There's always another mind there. It's like having the corner bar, complete with old buddies and delightful newcomers and new tools waiting to take home and fresh graffiti and letters, except instead of putting on my coat, shutting down the computer, and walking down to the corner, I just invoke my telecom program and there they are."

These contacts over the Internet, even those between people that never met before, can often be beneficial to the participants. The most common benefit is that of exchanging information (Kollock 1998, Matzat 2001), which is what the Internet was designed for in the first place. Beyond that many other forms of helping and cooperation take place online, for example people helping their online friends get a job or letting them stay in their house when they are in the

city (or country). These contacts can thus be treated as a form of social capital as Lin (2001) defines it: '*resources embedded in social networks accessed and used by actors for actions*'.

Social capital in general is an important part of the means people have for attaining their goals. The combined resources that someone's friends, family and even casual acquaintances have access to are usually much larger than the resources the person has direct access to. These resources can often be 'borrowed'[1] when the person needs them. Granovetter (1974) makes a distinction between weak and stong ties in social networks. Strong ties usually occur between people in a similar position who have a lot in common, it is usually easy to 'borrow' resources through such ties because they are relatively strong. On the other hand, both persons are likely to have access to mostly the same type of resources, which makes the exchange less profitable. The other kind of ties are the weak ties between people in different social positions or of different backgrounds, they are less 'close', making it less likely that an exchange will take place, but they have access to different types of resources, making these exchanges more profitable.

The Internet is doubtlessly used a lot for communicating with people that one has strong ties with. Most of these will be between people that did not meet on the Internet. Communication through weak ties also takes place on the Internet, things like occasional e-mails to an acquaintance or infrequent encounters in a chat box. An interesting subgroup of these contacts are the purely-online contacts, people who have met each other on the Internet. I will assume that most of these purely-online contacts are weak ties, because the text-only form of most online communication makes it hard to sustain a close relationship. Strong online contacts certainly occur, but they are a minority. This form of weak ties has the potential to connect very different people. Many things that would strongly hinder the emergence of ties offline, such as location and social background, are much less problematic online. Information sharing between such people could be a very valuable form of social capital.

This is actually one of the things that the Internet optimists see as a great benefit: The Internet connecting people all over the world, no matter where they are or what their background is. Increased contact between different groups creates a more healthy public debate and strengthens civic society. Pessimists counter that by claiming that the ability to easily find like-minded people on the net will lead to fragmentation and isolation of groups. As Rheingold (1993) puts it:

> "The present state of porosity between the boundaries of different online groups on the Net might be an artifact of the early stages of the medium—fragmentation, hierarchization, rigidifying social boundaries, and single-niche colonies of people who share intolerances could become prevalent in the future."

Thus it remains to be seen whether the Internet is a place of broad heterogenous social networks or just thousands of little homogenous clusters. In this thesis I will look into social contacts on the Internet and try to decide which of these images is the most accurate. The central question is:

> Do associations between people in different social positions and with different interests form more easily or less easily online?

To answer this I will conduct an online survey to gather data on the contacts people have, both online and offline. By comparing online and

[1] By 'borrowing' I mean any way of being granted (possibly temporary) benefit from another person's resources. This could be actually borrowing something like a tool, but also sharing information or having someone use their influence to get you something.

offline networks of a sample of Internet-users it should be possible to give an answer to the above question. Some attention is also given to the ways in which these people meet each other, the size of the networks and the amount of weak and strong ties in them.

Chapter 2

Theory

In this chapter several relevant theories are outlined. Based on these theories some hypotheses are formulated in section 2.5, which I will try to test in the next chapters.

2.1 Opportunity and Motivation

For contacts between people to emerge and persist two things are necessary. Firstly, the opportunity to interact has to be present and stay present. When two strangers meet an opportunity for establishing a contact is present, and as long as they have access to each other the opportunity to continue the contact is present.

The other required element for a tie is the motivation of both parties to initiate and continue it. As soon as one of the involved persons loses interest the contact will be lost or at least become very weak, and it becomes unlikely that resources will be exchanged through it. The motivation to continue a tie is related to the amount of social capital the tie represents. A person will only be motivated to maintain advantageous contacts. These advantages are things like goods, status, help, information or affection that are available through the contact.

This model of opportunity and motivation can be used to describe most interpersonal contacts in a more or less meaningful way. It is also possible to formulate expectations about the chance of contacts forming using this way. In the next paragraphs I'll use this model to compare various forms of online and offline contacts.

2.2 Finding other people

In the offline world it can be hard to find people who share one's interests or who can help with certain specific questions. There exist a variety of periodicals and organizations for things like this but most of them are local and cost money and effort to join. In the online world, interest or specialism-based communities are abound. For people with an Internet connection and some experience with the Internet these groups are easy to find. Especially for people with interests that are not common this is very valuable.

One would expect the existence this vast pool of easy to find and easy to reach people to stimulate people to make contacts with like-minded people. The most common occasion for contact is the asking and answering of specific questions. Some Internet groups permit only such functional discussion of issues related to the group's topic and sanction other use of the channel (Kollock & Smith 1996), others permit off-topic social talk as well. If one wants to enter an off-topic conversation with someone in the group there is almost always the possibility to send electronic mail to this person.

This ease of finding people represents a constant opportunity to meet people. The ability to send e-mails or use other forms of fast global

communication channels means the opportunity to stay in touch is present as long as the other person keeps using the Internet. Thus I expect opportunity to rarely be a serious problem when it comes to establishing and maintaining online contacts.

Being able to find like-minded people so easily could have the effect of reducing the incentive to associate with people who are different. Communicating with people who have a different background and different interests is supposed to take more effort and be less rewarding than communicating with people who are similar (Granovetter 1974, Lin 2001). Thus, the availability of like-minded people could lead to more narrow personal social networks online.

2.3 Bandwidth

The term bandwidth can be somewhat confusing in the context of Internet communication because it is used for two different concepts. Firstly, it is used to indicate computer-level bandwidth, the amount of bits that can be transferred over a network in a certain amount of time. Secondly, it is used at the human level, to refer to the number of social cues that can be exchanged between people through a communication channel. The human bandwidth of text-based communication is lower than that of in-person communication because things like facial expression and intonation are not conveyed (Wellman & Gulia 1999). This second meaning is the one I will be referring to when I use the term in this thesis.

How much does this lack of bandwidth set online communication apart from face-to-face communication? It has been argued (Rusciano 2001) that the lack of cues makes text-only communication unsuitable for any serious social contact. A less extreme view is that while contact is possible, the lack of cues causes people to misinterpret each other, making communication harder and often leading to polarization in debates. Others expect the lack of social cues to have a positive effect on social interaction by reducing the amount of prejudice based on irrelevant traits—most of those traits can not be seen online. Donath (1998) summarizes the issue as follows:

> "This dearth of social cues is both good and bad. One of the most widely hailed features of on-line communication is its democratic leveling: one's thoughts and ideas, rather than one's age, race, gender, etc., are the first things known about one. Yet social cues are not simply vehicles for prejudice; they play an essential role in the formation of community and in our comprehension of social interactions. In particular, cues that reveal who one has become, that show one's affiliations, beliefs and interests, (as opposed to those based on one's genetic traits) are an integral part of communication."

Many authors (Donath 1998, Rheingold 1993) have noted that people who use the Internet to communicate through text-only channels have developed various ways to convey social cues. A common method is the use of codes or 'emoticons'[1] to indicate simple cues that would be nonverbal in face-to-face communication (laughing, sarcasm, anger). Together with a piece of text, such a simple indicator can express a rather large range of meaning. Thus, online communication is not completely devoid of social cues. While it has been the experience of the author that polarization is more likely to occur online than in face-to-face interaction, it is not so that all

[1] Emoticons are combinations of easily typable characters that form some kind of little graphic together. The most common examples are the happy face, formed with a colon and a closing parenthesis - :) - and the winking face, with a semicolon instead of the colon - ;). The former indicated something like 'happy' or 'joking', and the second can also mean 'joking' or 'not serious'. By using an opening parenthesis instead the face can be made to look sad, etc.

online-interaction is destined to end in hostility. Many online relationships and communities interact in harmony (Kollock & Smith 1996), and manage to convey enough cues to get by.

Does this lack of bandwidth live up to the expectation that it reduces discrimination by hiding many characteristics of people? Because things like ethnicity, gender and social class of a person are often not obvious online, communication between people who might be uncomfortable with each other face-to-face becomes easier. On the other hand one has to ask how hidden these characteristics really are. Things like education and social background are often quite apparent in someone's writing style, and many online environments offer information about people that can be used in the same way as in-person cues to allow for a premature judgment. A good example is that many Internet message boards show the number of contributions the user has made to that board next to every message, thus distinguishing new people from regular visitors and making a whole new form of snobbery possible. Still, it is not very hard for someone to cover up certain parts of his or her identity online, or even to pretend he or she is someone else. And even if a lot of cues get through, these are probably less detrimental to communication than they are in an in-person situation, because they are less prominent.

2.4 Size and composition of online networks

Are online personal networks bigger than offline personal networks? Maybe for a few heavy Internet users, but in general I expect offline networks to be bigger. Internet use is increasing rapidly, but many users are only familiar with a small subset of the possibilities.

What about the amount of strong and weak ties in online personal networks? The issue of limited bandwidth (section 2.3) could be a serious problem for establishing and maintaining a strong tie through the Internet. Comparing that to the minimal effort needed for establishing and maintaining a weak tie online I expect the proportion of weak ties to be bigger in online personal networks compared to offline personal networks. It is also thinkable, however, that the lack of face-to-face contact makes people less restrained and more open in their communication, which would facilitate the establishing of strong ties. Maybe the lower risk involved in establishing such ties could make them less durable though.

Do the size and heterogeneity of online personal networks differ for people with different backgrounds? One thing that will certainly be important is the amount of experience someone has with computers and Internet. The opportunity for contact only emerges when one is able to find these other people. Furthermore, almost all online communities have their own set of rules, customs and folklore which can be difficult to grasp for the newcomer. Thus, experienced Internet users are bound to have bigger online networks than novices. Related to that, spending a lot of time on the Internet is also expected to lead to a larger online personal network. The establishing and maintaining of relationships takes quite some time.

2.5 Hypotheses

So what does all this mean for the heterogeneity of online contacts? When it comes to opportunities I argued in section 2.2 that they are very much present on the Internet. Opportunities to meet with similar people and different people are both available, but people might be more motivated to interact with similar people.

Because of the ways in which Internet communication differs from in-person communication certain differences between people, such as location, ethnicity and age, are less of a hindrance for forming contacts. It is possible that Internet contacts are a lot more heterogenous on these

points than offline contacts. With regard other characteristics, which stay relevant even in text-only communication, online contacts might be more homogenous than offline contacts because it is so easy to pick and choose your contacts online. This would apply to education, religion and fields of interest. Only education and interests appear in the data, so this is not tested with regard to religion.

H1: When it comes to location, ethnicity and age, online contacts are more heterogenous than offline contacts.

H2: When it comes to education, religion and interests, online contacts are more homogenous than offline contacts.

About the size and composition of online personal networks the following expectations were formulated in section 2.4:

H3: The average online personal network is smaller than the average offline personal network.

H4: Online personal networks contain a larger proportion weak ties than offline personal networks.

H5: Experienced Internet users have more online contacts than less experienced users.

H6: People who spend a lot of time using the Internet have a larger online personal network than less frequent users.

Chapter 3

Data

The data used in this thesis were collected by conducting an online survey. There are a number of issues associated with online surveys, especially those without a well-defined sample group. These will be discussed in the course of the chapter.

3.1 Questions

The initial plan was to gather a dataset with information about online social networks that could be combined with existing datasets measuring properties of offline social networks such as the United States' 1985 General Social Survey. In the end this turned out to be an impractical idea. The number of useful datasets available for this purpose was very small. On top of that some variables, such as the amount of overlapping interests between people, did not appear in any of them.

Thus I set out to measure three things. Firstly a number of background variables for the respondent, both to compare them with those of his or her contacts and as predictors of things like network size. Secondly I wanted data about people's online contacts, how many they have and what they have in common with them. And lastly I wanted to gather the same kind of data about people's offline contacts, to compare online and offline networks. The complete list of questions can be found in the appendix. These questions were asked in the form of an online survey.

The first series of questions asked for background characteristics. Respondents were asked to indicate their gender and age. They were asked whether they are currently following full-time education, and if not, how many years they went to school. Furthermore they were asked in what county they live. People who were not part of the ethnic majority of their country were asked to specify their ethnicity. Finally respondents are asked how long (in years) they have been using the Internet and how much time they spend online in a week.

The next batch of questions concerns online social contacts, that is contacts with people whom the respondent initially met online. A contact is defined as someone that you communicate with at least every two months, and a distinction is made between loose contacts and close ones, those that you would discuss personal matters with. For every contact the respondent was asked in what way they met that person, and for what purpose they communicated when they met. The purpose is a multiple choice questions with the choices "to exchange information or help for free", "to do business", "just to talk" and "other reason". The respondent was also asked to estimate how much interest or hobbies he and this contact share, the approximate age and education of this person, his or her gender, the county in which this person lives and his or her ethnicity. On all of these questions there was the possibility to indicate "unknown". The respondent was asked to list a maximum of five close contacts and 10 non-close contacts.

Finally the same questions, except for the questions about how people met, were asked about up to 5 close and up to 10 non-close *offline* social contacts. In retrospect asking people to fill in the same questions so often may not have been a very good idea. The attention spans of Internet users are not awfully long. The survey ended with a text box where people could add comments, and 6 people put remarks there about how boring and repetitive the survey was. I suspect a lot of people just decided they'd had enough before finishing their quota of contacts. In fact, only a few people submitted anywhere near the 30 people they were allowed to submit, and a large portion submitted less than three contacts. Still, miraculously, even though the non-close offline contacts came at the very end of the survey, a reasonable number of them was submitted.

An Internet site was set up with a short introduction and all the questions. The questions were divided over a number of pages, and whenever a respondent went to the next page his or her answers were saved. That way some data is available even for those who did not finish the survey.

3.2 Sample

The relevant population for my hypotheses is the group of all Internet users in the world. Obviously, no lists or files with the e-mail addresses of this population exist. I approached people with the question to fill in the survey in several different ways. Firstly, and most importantly, by placing messages on message boards asking readers to participate. I only did this in groups where such a request would not appear excessively offtopic and shameless. I'd rather not contribute too much to the enormous amount of unsolicited nonsense that Internet users are being subjected to these days, and when you approach people in an obnoxious way the chance of having them contribute to your survey is not very big anyway. I also entered two MUDs[1] and left messages on in-game virtual notice boards. Lastly I collected a number of e-mail addresses from newsgroups and sent those people an email asking them, as polite as I could, to participate. This came dangerously close to spamming and apparently was perceived as such. The response rate for these groups was very low. I only got one hostile reply out of 135 messages sent, but I suspect a lot of the e-mails were discarded without being read - only 7.4 percent participated in the survey.

I also had plans of contacting other people, for example people who use popular instant-messaging programs. These probably contain more inexperienced Internet users. These programs and services all hide the e-mail addresses of their users though, and instant-messaging people with a request to participate in a survey seemed too obtrusive.

This leaves the group of people who only use the Internet occasionally and are not experienced users pretty much out of the data set. If anybody who uses the Internet at all is considered part of the population, a big portion is left out here. People in this group are unlikely to have much online contacts though, and therefore are of little interest to this thesis. An effect of their underrepresentation will be that the average online network will be estimated to be larger than it actually is.

Another issue is language problems. The questions are formulated in English. For a lot of Internet users this will not be a problem, English is more or less the lingua franca of the Internet. A large portion of users are from the United States, Canada and the United Kingdom, and in most western European countries the English language is being taught intensively in school. Still there is no doubt that a large group of Internet user will not be able to fill in the survey because they do not speak the language.

[1] MUD stands for multi-user dimension (or domain, or dungeon), they are text-based games in which many people move their avatars through a virtual world in which they can interact with each other and the environment.

What is comes down to is that the sample we have here is not a good representation of the group of all Internet users. This is a pity, it makes generalizing findings impossible, but it is not a disaster. Because the hypotheses mostly describe relations between variables they do not suffer too much from not having a perfectly random sample.

After weeding out the people who filled in obvious nonsense or did not fill in anything at all, there were 195 people left. It is hard to say anything about the response rate on this survey, but it can be assumed to be dramatically low. The people who were so good as to fill in the questions are probably only the more helpful elements. If they are, they might have a bigger online social network than the average Internet user.

The resulting data set contains two tables, one with respondents and their characteristics, and one with the contacts of those respondents. This second table contains the properties specific to contacts, and a field to identify which respondent the contact belonged to. The data can be aggregated by adding the properties of the relevant respondent to each contact, and this is done for some tests in the next chapter.

3.3 Results

195 people filled in enough information to be included in the data. These people gave information on 469 contacts.

All but 3 people filled in a gender, 144 (75.0%) of these were male. This is quite a big proportion, but not a surprising finding. Men are overrepresented almost everywhere on the Internet.

The education of the respondents is shown in table 3.1. A large portion is still following full-time education. If the education of those who are still in school is taken to be their age minus 12 (not valid for a lot of countries, but a reasonable approximation), capped at 20, everyone has a score on this variable. The results of this are shown in table 3.2.

Table 3.1: Years of education, not including primary school

Education	N	valid %
Still in school	81	42.9
0 to 4 years	26	13.8
5 to 9 years	40	21.2
10 to 13 years	31	16.4
14+ years	11	5.8
Unknown	6	

Table 3.2: Years of education, recoded for people who are still in school

Education	N	valid %
0 to 4 years	30	13.8
5 to 9 years	58	21.2
10 to 13 years	61	16.4
14+ years	39	5.8
Unknown	7	

Table 3.3 shows the age distribution. Most people seem to be under 30, but even in the 50+ range some people can be found.

Table 3.3: Age distribution

Age	N	valid %
10 - 19	48	24.7
20 - 29	82	42.3
30 - 39	32	16.5
40 - 49	19	9.8
50 - 59	10	5.2
60+	3	1.5
Unknown	1	

The country in which the respondents live is shown in table 3.4. Two third of the respondents live in the United States, Canada or the United Kingdom. This is partially a result of the fact that the survey was formulated in English and invitations were placed mostly on English-based communication channels.

Tables 3.5 and 3.6 show the number of years people have been using the Internet and the

Table 3.4: Countries of residence

Country	N	valid %
Argentina	1	.5
Australia	16	8.3
Austria	1	.5
Belgium	2	1.0
Canada	27	14.1
Denmark	2	1.0
Egypt	1	.5
Estonia	1	.5
Finland	5	2.6
France	1	.5
Germany	7	3.6
India	2	1.0
Israel	2	1.0
Italy	1	.5
The Netherlands	5	2.6
New Zealand	7	3.6
Norway	1	.5
Pakistan	1	.5
Poland	1	.5
South Africa	1	.5
Sweden	4	2.1
Switzerland	2	1.0
United Kingdom	31	16.1
United States	70	36.5
Unknown	3	

Table 3.5: Years of experience with the Internet

Years	N	valid %
0 - 3	8	4.1
4 - 7	89	45.9
8 - 11	83	42.8
12+	14	7.2
Unknown	1	

Table 3.6: Hours online per week

Hours	N	valid %
0 - 24	84	43.3
25 - 49	82	42.3
50 - 74	23	11.9
75+	5	2.6
Unknown	1	

Table 3.7: Closeness and online character of contacts

	Online	Offline
Close	57	144
Non-close	76	192

Table 3.8: Number of contacts entered per respondent

Contacts	N	%
0	47	24.1
1-4	117	60.0
5-8	23	11.8
9+	8	4.1

number of hours they spend online every week. Only a few people in the sample have started using the Internet less than four years ago, and over fifty percent spend more than 25 hours online every week. The sample probably contains an overrepresentation of heavy Internet users.

A total of 469 persons were entered as social contacts. 133 (28.4%) of these are close contact, 201 (42.9%) of them are online contacts. Table 3.7 roughly shows how these properties relate to each other. There are more contacts that are both online and close than I would have expected. The ordering of the categories is not unexpected, offline weak ties are the most common, followed by offline strong ties, etc.

Table 3.8 shows the number of contacts entered per respondent. For some reason, 24.1 percent of the respondents did not enter any contacts. Two of these indicated in the comments field that this was correct information—they did indeed not have any social contacts at all. For how many of these others that is the case I do not know. I suspect a lot of people got distracted before they even got to the second page of questions. One individual filled in 28 (of a possible 30) social contacts, other than that person the maximum is 15 contacts.

For 458 of the 469 contacts a gender was known, 162 (35.4%) of these were women. The percentage of women in the contacts lies markedly higher than the percentage in the respondents. This could have something to do with women having a bigger social network (they submit an average of 3.5 contacts, men only 2.0).

The education of the contacts is shown in table 3.9. A sizable portion (10.9%) of the education levels is not known. It is possible that most people choose to err on the side of indicating a high eduction for their social contacts, for these figures show a *very* big proportion of people whose education is 'high'. When online and offline contacts are split the distribution turns out to be more skewed for the offline contacts than for the online contacts.

Table 3.9: Approximate education of contacts

Education	N	valid %
Low	13	3.1
Average	173	41.4
High	232	55.5
Unknown	51	

The approximate age of the contacts can be seen in table 3.10. It seems that people have less problems guessing ages than they have with educations—only in 4.1% of the cases is the age not known. The distribution is not very surprising given the fact that most respondent are young, most of their contacts are also young. The offline contacts are on average older than the online contacts.

The countries in which contacts live are shown in table 3.11. This distribution strongly resembles the one in table 3.4.

Table 3.10: Approximate age of contacts

Years	N	valid %
below 20	105	23.3
21 - 40	276	61.3
41 - 60	62	13.8
61+	7	1.6
Unknown	19	

Table 3.11: Countries of residence for contacts

Country	N	valid %
Argentina	2	.5
Australia	30	7.1
Austria	1	.2
Belgium	6	1.4
Canada	55	13.0
Denmark	1	.2
Egypt	4	.9
Estonia	4	.9
Finland	10	2.4
France	2	.5
Germany	9	2.1
India	7	1.7
Ireland	1	.2
Israel	3	.7
The Netherlands	11	2.6
New Zealand	22	5.2
Norway	5	1.2
Peru	1	.2
Poland	5	1.2
Russia	1	.2
South Africa	1	.2
Sweden	4	.9
Switzerland	3	.7
United Kingdom	73	17.2
United States	163	38.4
Unknown	45	

Table 3.12 shows the amount of overlap in interests between a respondent and a contact. Social contacts between people who hardly share interests seem to be uncommon.

For online contacts two other questions were asked—where people met and for what purpose they met. The open question on where people met has been sorted into categories by hand. The results are shown in table 3.13. The categories are somewhat cryptic, but that is because most of them are quite broad. 'Chat' means any kind of real-time text exchange that is not a game,

Table 3.12: Interests the contact has in common with the respondent

	N	valid %
A lot	170	36.4
A few	252	54.0
Almost none	45	9.6
Unknown	2	

Table 3.13: How did the respondent meet an online social contact

	N	valid %
Chat	28	14.2
Dating	3	1.5
Forum	74	37.6
Game	66	33.5
Group	23	11.7
Unknown	4	

mostly IRC[2] and instant messaging programs[3]. 'Dating' refers to online dating services. 'Forum' means any kind of open Internet forum, with the exception of Usenet groups, which fall under 'group'. 'Game' refers to online games, people who met playing a game. The most common games named here were MUDs and Everquest[4]. Finally, 'group' refers to Usenet newsgroups and mailing lists. The distinction between 'group' and 'forum' is somewhat arbitrary. In general fora are more open, they require less involvement to join. The fact that the 'forum' and 'game' categories are the most common is not very surprising, those are the channels in which I invited people to fill in the survey.

[2]Internet Relay Chat, an old and still very popular protocol for chat programs.

[3]Programs like ICQ and AIM.

[4]A so-called MMORPG, massively multi-player online role-playing game, a game in which thousands of players play at the same time taking the role of classic fantasy characters in the tradition of role-playing games like Dungeons & Dragons.

Table 3.14: Why did the respondent communicate with an online social contact for the first time

Reason	N	%
Exchange information (for free)	54	26.9
Business	5	2.5
Just to talk	101	50.2
Other	41	20.4

The reasons for which people met for the first time are shown in table 3.14. Most of the initial contacts appear to have taken place purely for social interaction.

Chapter 4

Analysis

In this chapter I'll try to test the hypotheses formulated in section 2.5. The data makes this rather hard for a few, especially those about network size, but some interesting results can be shown nevertheless.

4.1 Online and offline networks

The first two hypotheses make predictions about differences in heterogeneity of online and offline networks. For location, ethnicity and age, online contacts are expected to be the most heterogenous, while for education, religion and interests it is the other way around. Religion does not appear in the data, but for the other five properties we can test this.

The most straightforward way of measuring heterogeneity of characteristics measured as numbers is to take the correlations between a characteristic of the respondent and that same characteristic in his or her contacts. For this purpose variables measured as ordinal categories are treated as numbers, for example 'low', 'average' and 'high' education becomes 0, 1 and 2. Table 4.1 shows these correlations for gender, education and age.

Table 4.1: Correlations between characteristics of respondents and their contacts for offline and online contacts

	Online	Offline
Gender	.093	.376
Education	.067	.248
Age	.598	.690

For all three variables shown in table 4.1 online networks are more heterogenous than offline networks. Gender did not appear in the hypotheses, but it is interesting to note the big difference anyway. People have more acquaintances of the opposite sex online than offline. For age these correlations support the hypothesis that age differences are less of a barrier for communication online than they are offline, but the difference is not very big. For education these numbers strongly contradict the hypothesis that online contacts will be homogenous when it comes to education—the correlation is much smaller online than offline.

These hypotheses can also be tested with regression models. For example for education, the education of the contact is the dependent variable and three predictors are used. The education of the respondent, the yes/no variable indicating whether the contact is online, and the product of the other two variables. The effect of this last predictor is the difference in the effect of the respondent's education on the contact's education between offline and online contacts. Heterogeneity of age can be tested in a similar way.

For education the results of such a regression model are shown in table 4.2. The model was also fitted with the strength of ties and the products of the strength of ties and the respondent's education as predictors, but those effects were

Table 4.2: Regression model predicting education of a contact ($R^2 = .443$, $N = 409$)

	b	β	*s.e.*	p	
(Intercept)	1.280		.078	.000	**
Resp. Education	.032	.259	.008	.000	**
Online	.123	.108	.120	.306	
Online * Resp. Education	-.024	-.230	.012	.046	*

*: $p < .05$, **: $p < .01$

Table 4.3: Regression model predicting age of a contact ($R^2 = .433$, $N = 450$)

	b	β	*s.e.*	p	
(Intercept)	-.149		.079	.058	
Resp. Age	.039	.713	.003	.000	**
Online	.144	.108	.124	.246	
Online * Resp. Age	-.008	-.198	.004	.046	*

*: $p < .05$, **: $p < .01$

too small to be significant so they were omitted in the model presented here. The fact that the main effect of the respondent's education is significant indicates that for offline contacts education is definitely a factor. The difference between this effect and that for online contacts is also significant, and it is negative. This means that the conclusion drawn from table 4.1 still stands. Online contacts are more heterogenous on education.

For age a similar model is made. Table 4.3 shows the results. Again strength of ties was initially included in the model, but dropped because it did not have a significant effect. This shows a situation similar to that of education. Online the effect of age is smaller than offline. The difference between the effects for offline and online contacts is not as big as it was for education, but it is significant.

The degree in which interests overlap was measured with a direct question (see table 3.12). Table 4.4 shows how the distribution on this variable differs for online and offline contacts. It seems that contacts that take place offline have more overlap of interests than contacts that take place online. This contradicts the hypothesis.

If this is tested with regression, the model in table 4.5 is obtained. The strength of the ties has a significant effect here, but the online character of the contact does not. Taking another look at table 4.4 you could suspect this is a result of the odd distributions this variable has. For online contacts the distributions is much 'sharper' than for offline contacts. If the dependent variable is made dichotomous, with 0 meaning 'almost none' or 'a few' common interests, and 1 meaning 'a lot' of common interests, and a logistic regression is used, the difference does become quite significant all of a sudden. This is shown in table 4.6. The strength of the tie remains a strong predictor in this model.

Table 4.4: Overlap in interests for offline and online contacts

	Online		Offline	
	N	valid %	N	valid %
A lot	60	29.9	110	41.4
A few	126	62.7	126	47.4
Almost none	15	7.5	30	11.3
Unknown	0		2	

To analyse differences in the location of people several different approaches can be taken. Firstly I look whether the country of the respon-

Table 4.5: Regression model predicting overlap of interests ($R^2 = .004$, $N = 467$)

	b	β	$s.e.$	p	
(Intercept)	1.227		.042	.000	**
Online	-.077	-.061	.057	.178	
Close	.261	.188	.063	.000	**

*: $p < .05$, **: $p < .01$

Table 4.6: Logistic regression model predicting overlap of interests ($N = 467$)

	b	$s.e.$	p	
(Intercept)	-.640	.143	.000	**
Online	-.532	.203	.009	**
Close	.993	.213	.000	**

*: $p < .05$, **: $p < .01$

dent is the same as that of his or her contacts. Table 4.7 shows the results for this. These figures give strong support for the hypothesis that contacts between people from different locations are more likely online than offline.

In table 4.8 a logistic regression model is shown for homogeneity of location. Again the difference is significant. Online contacts are more heterogenous when it comes to location. The strength of the ties did not have a significant effect in this model and was left out.

This analysis only takes into account whether people live in the same country. This may be a somewhat limited way to look at location differences. For example people in a small country or a country in which hardly anybody is online will be more likely to form ties with people from abroad, purely because people from their own country are less available. I'll repeat this test in another way—dividing countries into categories.

A common way to categorize countries is into 'western' and other countries. North America, Western Europe, Australia and New Zealand are taken to be as western countries here. In table 4.9 a logistic regression model is shown with this division. The familiar pattern shows again. Offline contacts are more homogenous than online contacts.

Another way to categorize countries is by continent. Because of the small number of respondents in Asia, Africa and South America I'll use only three categories: America, Europe and Africa, and Asia, Australia and New-Zealand. In eurocentric fashion, I'll call these respectively West, Center and East from now on. Table 4.10 shows the cross-tables for these categories. Because of all the empty cells in the offline table, no meaningful results could be gotten from regression models. But the tables leave little doubt that this is a significant effect, for online contacts the numbers are quite evenly distributed while for offline contacts nearly everybody is on the diagonal.

Differences in ethnicity can also be approached in a few ways. I'll start with difference scores again. In this analysis the concept 'ethnicity' is a combination of the country where people live and the ethnicity given for them. People who live in the same country and have the same ethnicity code are considered to be 'of the same ethnicity'. Table 4.11 shows the distributions. Again, offline contacts are more homogenous than online contacts. This supports my hypothesis.

Table 4.7: Do contacts live in the same country? Split for online and offline contacts

	Online		Offline	
Location	N	valid %	N	valid %
Same	93	53.4	233	93.6
Different	81	46.6	16	6.4
Unknown	27		19	

Table 4.8: Logistic regression model predicting contact living in the same country ($N = 423$)

	b	$s.e.$	p	
(Intercept)	2.678	.258	.000	**
Online	-2.540	.300	.000	**

*: $p < .05$, **: $p < .01$

Table 4.10: General part of the earth where people live, split for online and offline contacts

	Online			Offline		
	West	Center	East	West	Center	East
West	65	17	8	128	2	0
Center	15	36	6	0	84	0
East	3	10	14	0	1	34

(Rows are respondents, columns are contacts)

Table 4.9: Logistic regression model predicting contact living in a 'western' country ($N = 423$)

	b	*s.e.*	p	
(Intercept)	-1.735	.626	.000	**
Online	7.164	1.182	.000	**
Western Country	1.735	.945	.066	
Online * Western	-4.305	1.419	.002	**

*: $p < .05$, **: $p < .01$

Table 4.11: Do contacts have the same ethnicity? Split for online and offline contacts

	Online		Offline	
Ethnicity	N	valid %	N	valid %
Same	47	42.0	146	73.4
Different	65	58.0	53	26.6
Unknown	89		69	

Table 4.12: Logistic regression model predicting contact having the same ethnicity ($N = 311$)

	b	*s.e.*	p	
(Intercept)	1.013	.160	.000	**
Online	-1.338	.250	.000	**

*: $p < .05$, **: $p < .01$

In a logistic regression model this conclusion holds up. Table 4.12 shows that the difference is significant, online contacts are more heterogenous on ethnicity. Again, the effect of the strength of the ties was too small to be included.

Ethnicities can also be divided into categories, but because a huge majority of people in the sample have a Caucasian background only two categories were used—Caucasian and non-Caucasian. Table 4.13 shows a logistic model for this. The effect of the multiplicative term has the predicted sign, but is not significant. This may have something to do with the very skewed distributions though, only about 10% of the people in the sample had a non-Caucasian ethnicity.

Table 4.13: Logistic regression model predicting contact having a Caucasian ethnicity ($N = 422$)

	b	*s.e.*	p	
(Intercept)	.194	.361	.591	
Resp. Caucasian	3.786	.620	.000	**
Online	.594	.649	.360	
Online * Caucasian	-1.349	.921	.143	

*: $p < .05$, **: $p < .01$

Online contacts turned out to be more heterogenous than offline ones on the variables age, education, location and ethnicity. For shared interests the results are somewhat unclear, but point in the same direction. This means that hypothesis 1 ('*When it comes to location, ethnicity and age, online contacts are more heterogenous than offline contacts*) gets confirmed. Hypothesis 2 ('*When it comes to education, religion and interests, online contacts are more homogenous than offline contacts*') gets rejected.

4.2 Network size

The hypotheses made several predictions about network sizes. Firstly, online social networks are expected to be smaller than offline ones (H3). People who have been using the Internet for a long time or are online many hours per week are expected to have a larger online network to show

Table 4.14: Regression model for number of online contacts ($R^2 = .175$, $N = 184$)

	b	β	$s.e.$	p	
(Intercept)	-.300		.347	.394	
Years of Internet experience	.047	.087	.038	.213	
Hours per week online	.007	.088	.005	.203	
Gender (woman)	1.108	.339	.224	.000	**
Years of education	.058	.173	.023	.014	*

*: $p < .05$, **: $p < .01$

for it (H5 and H6).

As mentioned earlier, the way network size is measured in my data is dubious. People are asked to enter as many contacts as they want, and the total number they enter is taken to be the network size. People who are not motivated to fill in a lot of data will end up with a small network size this way, and because the questions about offline contacts were asked after those about online contacts this might have caused people to enter relatively few offline contacts.

The mean number of online contacts people have is 1.03, for offline contacts this is 1.37. This is quite a big difference. A t-test shows that it is significant ($N = 195$, $t = 3.161$, $p = .002$). So that would mean the hypothesis that online social networks are smaller than offline ones gets confirmed. Note that the bias mentioned in the previous paragraph would make this difference smaller. Thus this conclusion is rather solid.

To test the effect of Internet experience and frequency of use on the size of one's online network a regression model is used. Gender and education were also included in the model, their effects are significant and they might correlate with the other predictors. One outlier (the person who entered 28 contacts, 14 of which were online) was excluded from the model. Table 4.14 shows the results. These hypotheses get rejected by the data, experience and hours online both have insignificant effects. This is not a very reliable conclusion though, it might be related to the problems in the data. Also, the effects of gender and education here could very well be caused by the fact that women and educated people are more helpful when it comes to filling in long boring questionnaires.

So the hypothesis that online networks are smaller than offline networks gets confirmed by the data. Rather surprisingly, the expectation that people who spend more time online and have been using the Internet for a long time have larger online networks is not supported.

4.3 Weak online ties

The fourth hypothesis in section 2.5 predicts that online networks contain a smaller proportion strong ties than offline networks. In table 3.7 a cross-tabulation of strength and online character of contacts is given. To test this hypothesis the proportion close ties in all offline and online networks was computed. Networks of 0 people have a missing score on this variable. This reduces the number of cases rather drastically, because all people who have no online or no offline contacts get excluded.

The mean proportion of close contacts for online social networks is .122, for offline networks this is .166. This difference is barely significant with $\alpha = .05$ ($N = 84$, $t = -2.05$, $p = .044$). This confirms the hypothesis that online social networks contain less strong ties, but again I have to point out that some problems with the data make this conclusion less than solid.

Table 4.15: Regression model for overlap in interests ("Meet on forum" is reference group, $R^2 = .153$, $N = 198$)

	b	β	$s.e.$	p	
(Intercept)	1.469		.058	.000	**
Met in group	-.382	-.218	.124	.002	**
Met in chat	-.219	-.136	.115	.058	
Met in game	-.499	-.418	.087	.000	**

*: $p < .05$, **: $p < .01$

4.4 Communication channels

The two variables concerning how people first met online (see tables 3.13 and 3.14) do not appear in the hypotheses. It would be a pity if they were completely unused. I adjusted all the models concerning heterogeneity discussed in section 4.1 to include these variables. That means I removed the 'online' variable from the model (because these variables are measured only for online contacts), and introduced these variables.

The variable giving the reason why people met did not produce any significant results. The other variable, indicating the type of communication channel where people first met, did produce significant effects for overlapping interests, location and ethnicity. The categories "Mail" and "Dating" were left out, because they contained too few people (respectively 4 and 3 people).

Table 4.15 shows the model for overlapping interests. "Forum" is the reference category here. Contacts that met in groups or games are significantly more heterogenous on interests than those that met in online fora.

Tables 4.16 and 4.17 show similar models. The dependent variable is contacts living in the same country or having the same ethnicity. "Chat" is the reference category here. In both models contacts that met in a forum or game are more heterogenous than those that met in a chat channel.

Table 4.16: Logistic regression model for living in the same country ("Meet in chat" is reference group, $N = 172$)

	b	$s.e.$	p	
(Intercept)	1.482	.495	.003	**
Meet in group	-1.040	.654	.112	
Meet on forum	-1.705	.557	.002	**
Meet in game	-1.652	.560	.003	**

*: $p < .05$, **: $p < .01$

Table 4.17: Logistic regression model for having the same ethnicity ("Meet in chat" is reference group, $N = 110$)

	b	$s.e.$	p	
(Intercept)	.956	.526	.069	
Meet in group	-.956	.726	.188	
Meet on forum	-2.101	.682	.002	**
Meet in game	-1.432	.606	.018	*

*: $p < .05$, **: $p < .01$

I also tested whether the reason for which people first met and the way in which they met had any effect on the strength of the tie between them, but no significant effects were found there.

Chapter 5

Conclusions

Even though the data used has some flaws caused by mistakes made during data collection, most of the hypotheses could be tested properly. Some interesting results were found.

The most important conclusion is that online social networks are more heterogenous than offline social networks on *all* variables taken into account here. That is on age, education, gender, interests, location and ethnicity. Education appeared to be almost irrelevant for the chance of forming online ties. The hypotheses did not predict this. The idea that people with similar interest and education might flock together online because it is so easy to avoid those with different ideas, as formulated in hypothesis 2, is rejected.

The fact that online contacts are more heterogenous on things like location and is hardly a surprise of course. Offline, people are surrounded by people who live near them and usually by people of the same ethnicity. Online these things, especially location, lose a lot of their relevance. Still, the fact that online contacts exist between people in different places is a meaningful piece of information. It suggests that the optimist theories about Internet communication, those that predict everyone will be talking to everyone, might be right. Global understanding through glass fiber cables actually remains a possibility, etc.

Some important characteristics, most notably religion and political ideas, on which heterogenous contacts could form 'bridges' between social groups, were not included in the data. The reason for this is the can of methodological problems that comes with measuring these variables in an international population. The amount of religions and political movements in the world is enormous, and especially when it comes to political affiliations terminology varies a lot between places. Still, if someone managed to measure heterogeneity of online contacts on those variables the results are bound to be very interesting.

Other findings were that even in a sample of heavy Internet users online social networks are smaller than offline social networks, and contain a smaller proportion of strong ties. Electronic contacts probably miss some characteristics that make in-person contacts more attractive. Face-to-face communication still has a much larger bandwidth, more space for social cues, than any Internet communication channel offers.

The expectation that people who have been using the Internet for a long time and people who spend a lot of time online have larger online networks than other people was not supported by the data. This certainly seems counter-intuitive, and I'm having a hard time explaining this. One probable explanation for this finding is that the deficiencies in the data were most problematic on this point, and may have reduced these effects. The effects that were found did go in the expected direction, but they were too small to be considered significant.

Finally, it was found that contact that emerged in a newsgroup, mailing list or game are more heterogenous on interests than those that were formed on an Internet forum. Fora often have a specific topic so that can explain the difference between fora and games, but newsgroups are also often very specific in their topics. On location and ethnicity contacts formed on fora and in games were more heterogenous than those formed in chat channels. Chats are probably more often locally-based than fora and games.

Appendix: Survey questions

These are the questions in the survey that was used to gather my data. These first questions about the respondent him or herself appeared on a separate page, and then for every contact a separate page with questions was shown. Before the online contacts and before the offline contacts a page appeared where they could indicate whether they had any contacts at all. At the end of the survey a possibility to enter comments was given.

This survey will consist of about 30 questions about your Internet use and the people you talk to online. It will probably take less than 10 minutes to fill them in. The first few questions ask some things about your background.

Are you a man or a woman? *(Choices: "Man", "Woman")*

What is your age? *(Open)*

Are you following full-time education at the moment? *(Choices: "Yes", "No")*

If you answered no to the previous question, how many years did you go to school? (Not including primary school.) *(Open)*

In which country do you currently live? *(Open)*

If you live in the USA or Canada, please enter the name the state or province you live in. *(Open)*

If your ethnicity differs from that of the majority of the people in your country, could you name it here? *(Open)*

How many years have you been using the Internet? *(Open)*

How many hours a week do you use the Internet? (Approximately) *(Open)*

Then the respondent was asked to enter 0 to 5 close online contacts. 'Online' meaning they first met these people online, and 'close' meaning they would discuss personal matters with these people. After that they could enter 0 to 10 non-close online contacts. For each online contact the following questions were asked.

In what way did you first meet him or her? (Please mention the type of Internet communication used.) *(Open)*

Why did this first contact take place? *(Choices: "To exchange information or help for free", "To do business", "Just to talk", "Other reason")*

Does this person share many of your important interests or hobbies? *(Choices "A lot", "A few", "Almost none")*

Please give some details about this person.

The approximate age. *(Choices: "Below 20", "Between 20 and 40", "Between 40 and 60", "Older than 60", "Don't know")*

The sex of this person. *(Choices: "Male", "Female", "Don't know")*

Approximate education. *(Choices: "Low", "Average", "High", "Don't know")*

The country this person lives in. *(Open)*

The ethnicity of this person. *(Open)*

Then the same questions, except the two about how they met, were asked about 0 to 5 close offline contacts and 0 to 10 non-close offline contacts.

The data that was gathered through this survey is available at 'http://marijn.haverbeke.nl/thesis/'. This document can also be downloaded from there.

Bibliography

Donath, J. S. (1998). Body language without the body: Situating social cues in the virtual world. **URL:** *http://duplox.wz-berlin.de/docs/panel/judith.html*

Granovetter, M. (1974). *Getting a Job: A Study of Contacts and Carreers*, Cambridge, Massachusetts: Harvard University Press.

Kollock, P. (1998). The economies of online cooperation, gifts and public goods in cyberspace, *in* M. Smith & P. Kollock (eds), *Communities in Cyberspace*, London: Routledge. **URL:** *http://www.sscnet.ucla.edu/soc/faculty/kollock/papers/economies.htm*

Kollock, P. & Smith, M. (1996). Managing the virtual commons: Cooperation and conflict in computer communities, *in* S. C. Herring (ed.), *Computer Mediated Communication. Linguistic, social and cross-cultural perspectives.*, Amsterdam: Philadelphia.

Lin, N. (2001). *Social Capital: A Theory of Social Structure and Action*, New York: Cambridge University Press.

Matzat, U. (2001). *Social Networks and Cooperation in Electronic Communities*, ICS.

Norris, P. (2001). *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide*, Cambridge: Cambridge University Press. **URL:** *http://ksghome.harvard.edu/~pnorris/Books/Digital%20Divide.htm*

Rheingold, H. (1993). *The Virtual Community: Homesteading on the Electronic Frontier*, New York: Addison-Wesley. **URL:** *http://www.rheingold.com/vc/book/*

Rusciano, F. L. (2001). Surfing alone: The relationships amoung internet communities, public opinion, anomie, and civic participation. **URL:** *http://cct.georgetown.edu/apsa/papers/Rusciano.pdf*

Wellman, B. & Gulia, M. (1999). Net-surfers don't ride alone: Virtual communities as communities, *in* B. Wellman (ed.), *Networks in the Global Village*, Boulder, Colorado: Westview Press.