The Virtual Speech Community:
Social Network and Language Variation on IRC

John C. Paolillo <john@ling.uta.edu>
Program in Linguistics, University of Texas at Arlington

Abstract
Many scholars anticipate that online interaction will have a long-term effect on the evolution of language, but little linguistic research yet addresses this question directly. In sociolinguistics, social network relations are recognized as the principal vehicle of language change. In this paper, I develop a social network approach to online language variation and change through qualitative and quantitative analysis of logfiles of Internet Relay Chat interaction. The analysis reveals a highly structured relationship between participants’ social positions on a channel and the linguistic variants they use. The emerging sociolinguistic relationship is more complex than what is predicted by current sociolinguistic theory for offline interaction, suggesting that sociolinguistic investigation of online interaction, where more detailed and fine-grained information about social contacts can be obtained, may offer unique contributions to the study of language variation and change.

1. Introduction
Both popular wisdom and professional scholarship in a range of academic disciplines make a diversity of predictions about how the Internet will shape our lives and our language. If we are to understand truly how the Internet might shape our language, then it is essential that we seek to understand how different varieties of language are used on the Internet. In sociolinguistics, social network relations — the variety and frequency of contacts among people in a society — are recognized as the principal vehicle of language change. People in regular contact with one another tend to share more linguistic features, and tend to borrow more features of each others’ language varieties, even in situations where those varieties are different languages. Likewise, people who have less contact with one another tend to share fewer linguistic features with one another. Thus, to answer questions about how the Internet might affect the language that we use, we need to ask how it affects social contact among individuals, and what kinds of linguistic features that contact transmits to users.

There is also an urgent need to study language contact on the Internet; the rapid expansion of the Internet, and changes in the technology and the ways it is used make the Internet a very dynamic social force. Moreover, where sociolinguistics is concerned, theories of language contact and change developed through offline observations can be tested and refined through exploiting the rich data made available by persistent, digital text. Prior to the existence of the Internet, sociolinguists had to rely on crude approximations of frequency of contact to construct social networks. With the potential to log digital texts, a researcher can now compile a comprehensive corpus of interactions taking place among a selected group of people on any mode of computer-mediated communication, and from that record be able to identify recurrent producers and recipients of messages; in this way, a very detailed understanding of the frequency and nature of contact among members of a group can be constructed. Taken together with the fact that certain groups of people interact almost solely online in “virtual communities” [17], these developments mean that the relation of social network to language change can be studied more closely than ever before.

1.1. Internet Relay Chat
One example which illustrates well the importance of the Internet for the study of language contact is the recent explosion in the popularity of Internet Relay Chat (IRC) [20]. Like other computer-mediated communication (CMC), IRC is hosted by networks of servers that are globally distributed, so participants on IRC come from many different national backgrounds. What sets IRC apart from other modes of CMC is that interaction is conducted almost in real-time: all participants in an interaction must be electronically present at the same time, and messages are immediately transmitted through the intermediate servers to all participants, wherever they may be. Thus, IRC is characterized by much shorter propagation delay than Usenet news and Listserv messages. In addition, IRC is multi-participant, and message length is very short (typically one or two lines) so that IRC interaction is similar to multi-participant face-to-face conversation. Using IRC, people who are located in geographically distant locales, who are of different national and linguistic backgrounds, and who might otherwise never come into contact, can engage in real-time interactions that resemble the immediacy of in-person face-to-face encounters.

Unlike most face-to-face interaction, IRC interaction is constant. An IRC network such as EFNet, the original IRC network, typically hosts thousands of topically
themed “channels” where interaction takes place; many are occupied 24 hours a day. Since people must necessarily attend to other aspects of life, such as work or school, the participants of an IRC channel are constantly in flux. Further complicating matters, the “real life” identities of the interlocutors on chat are often uncertain or unknown; prior to “joining” a channel, participants select a “nick” (nickname) to be known by, and a person using a given nickname on one day may be quite different from one who has used it on another day. Nonetheless, IRC channels appear to develop a readily identifiable character. Interaction on a channel comes to center on topics that are related to the channel theme. Certain participants regularly return to the same IRC channels, to resume and maintain relationships they have initiated there. And recurring patterns of language use develop. On the channel #india, for example, I have noted several characteristic patterns of such language use, including the following five:

Use of Hindi

1. <Lone_Wolf> sajniji aap kahan se hain? Where are you from, Sajni?
2. <Raju> Gabbar hahaha........ sab ladke all the boys
tujhse pyaar karte hain adore you
dunno abt gals though:PP

Use of “r” for “are”

3. <sunaina> raju where r u?

Use of “u” for “you”

(See examples 3 and 4.)

Use of “-z” for “-s”

4. <Dr-pepper> nagin..it seemz u have already
lot of admirers

Use of obscenity

5. <D-zire> d-pac fuq?? fuq?? no babe, i dont
fuq. i make luv ;)

Such linguistic developments often arise entirely through interaction on the IRC channel. For instance, the “r” and “u” variables can be found on many IRC channels [20], and participants in other CMC types, including synchronous ones such as MUDs and MOOs tend to regard their use as coming from IRC, rather than being native to their own virtual cultures [3]. Given that participants on IRC may be separated by linguistic and international boundaries, and given that the body of participants on a channel is constantly changing and that many have never met face-to-face, how are such norms established and propagated?

2. Social networks, language and IRC

To investigate how the linguistic practices of IRC channels are established and propagated requires that the researcher study both the network of participants’ social interactions and the relationship of those social patterns to the distribution of linguistic variables such as those described in examples 1-5. This approach is that of social network studies in sociolinguistics. In prior social network studies, the linguistic variables tend to be more thoroughly studied than the social network relations. This is because it is relatively easy to collect representative examples of a participant’s speech in an interview, whereas directly studying social contacts requires knowledge of all the participants’ interactions with one another. What sociolinguists tend to do is to analyze network relationships using other information, such as the mutual naming of friends [10]. These techniques provide an approximation of frequency of contact, without requiring that the researcher observe such contact directly. In some cases, notably the work of Milroy [11] and Milroy and Milroy [12], researchers use participant observation to study social network relationships more directly, but this requires an extensive research commitment on the part of the researcher, and still does not permit the researcher to directly count the frequency of linguistic interaction between any pair of participants. With online interaction, the social network analysis can be based on a record of interaction contained in a log of online conversation. In this way, only the linguistically relevant social contacts can be studied. At the same time, the textual log of interaction allows the frequency of interaction to be used directly in studying the spread of linguistic variables.

2.1. Strong and weak ties

Sociolinguistic theory generally recognizes that social network ties vary in quality. On the one hand, there are “strong ties”, typical of relationships among family and close friends, characterized by frequent interaction, association in more than one social capacity (e.g. two people being both siblings and business partners), and territorially based groups (e.g. neighborhood gangs). On the other hand there are “weak ties”, typical of relationships among casual acquaintances, characterized by less frequent and more transient contact, not anchored to any territory [1, 12]. These different types of network ties are associated with different norms for the use of linguistic variables. Individuals at the center of networks with predominantly strong ties tend to use more non-standard, vernacular linguistic variants. People at the periphery of the same networks, with fewer of the same strong ties tend to use fewer vernacular variants. Strong ties thus tend to enforce non-standard, vernacular linguistic norms.

---

1 The punctuation “:PP” at the end of example 2 is an ASCII icon or “emoticon” representing someone with their tongue sticking out. Its function on IRC is to mark a taunt, often a friendly one.
Conversely, people who have predominantly weak ties tend to have higher incidence of variants associated with the recognized standard variety. These patterns have led Milroy and Milroy to speculate that linguistic changes in the direction of the standard variety are propagated through weak network ties, while changes diverging from the standard variety in the direction of vernacular, non-standard varieties are propagated through strong network ties [11].

2.2. Social network and virtual community

Strong and weak network ties can potentially describe any sort of social interaction; if we use these notions to describe the social contacts made through IRC, then we can make predictions about the influence of IRC on language change. Given that IRC participation is transient and constantly changing, and given that the participants on an IRC channel tend not to be territorially localized, we might expect the social network relationships expressed on IRC to be dominated by weak ties, and therefore to promote changes in the direction of the prestige variety. But IRC does permit the development of strong ties in a different way, leading some to describe IRC channels as “virtual communities” [17]. Regular IRC participants often become veritable “addicts”, spending several hours per day on IRC, frequenting a small number of channels and having sustained interactions with other, similar, online addicts. This activity is much like the “hanging out” that urban youth engage in that leads to territorial strong tie networks. While the IRC social networks formed this way are not territorial in the literal sense, they can have territorial interpretations, where the territory is a particular channel or set of channels. This territoriality is expressed through built-in structural asymmetries among users inherent in the IRC medium.

On IRC, ordinary users and new users (“newbies”) lack the powers of the channel operators (“ops”) who are, among other things, empowered to exclude (“kick”) people from the channel. The first person to join a channel on IRC creates the channel and becomes its first operator. An operator may grant operator privileges (also called “ops”) to any others they choose. Thus, operators collectively and cooperatively define the boundaries of the social interaction on the channel. Sometimes operators accomplish this with the aid of computer programs known as “(ro)bots”; bots connect to IRC channels much as normal users do, and are typically granted ops by their creators. When a bot is opped, it may in turn automatically grant ops to channel participants (especially the bot’s owner) whose names and passwords are stored in a database, or automatically execute actions that the owner and other operators may desire, such as kicking and banning other participants, or de-opping certain operators (usually in preparation for kicking and banning them) [20]. In this way, a two-tiered social system is maintained such that members of the upper tier (operators and especially bot owners) have more or less guaranteed privileges on the channel. Members of the lower tier (the non-operators, ordinary users and newbies) are always at the mercy of the operators, whose actions range from benevolent to capricious [5]. At times there is social conflict among different operators, and “op-wars” can result, where different factions of operators struggle to impose control over the channel. All these patterns of social differentiation have concrete implications for the tie strength experienced by different participants on a channel.

Operators will tend to be regular participants with strong ties to one another and perhaps others. Other participants’ rights and privileges are acquired through social contact with operators, so their positions on the channel will vary according to the strength of their ties with different operators.

2.3. Linguistic characteristics of IRC

Since IRC messages are typed at a keyboard, there is a tendency to use conventions of written English, particularly spelling. Yet, as indicated in examples 1-5, a number of distinctive IRC spelling practices have emerged some of which can be found on many channels. The practices in examples 2, 3 and 4, namely substituting the letters $u$ and $r$ for the English words $you$ and $are$, and substituting $z$ for $s$, especially in word-final position, are three such IRC spellings. All three spellings diverge from standard written English, and so would be considered “vernacularizing” changes, and we would expect to find them used especially among members of strong-tie networks on a given channel. Likewise, obscene language is often found to be a marker of strong-tie networks [4, 10], where it is associated with values of toughness and masculinity. Finally, when a channel such as #india has a regional or cultural theme, it is common to find a language like Hindi being used to mark in-group identity, much as is found offline (see [13] and references therein).

Since in-group identity characterizes strong-tie social networks, we might expect such languages to be most common among the stronger ties on a channel. But, are patterns of language variation on IRC — in the use of languages like Hindi, in IRC spellings, and in the use of obscenity — in fact correlated with the density of social interaction among participants of a channel? Do social networks with dense, high frequencies of interaction on IRC favor these vernacular features, as do strong tie

---

2 The use of “z” for “s” is also common offline in stylized spellings emanating through popular American culture, e.g., “boyz in da hood”, suggesting a strong connection with vernacular culture and values.
networks in “real life”? Does the preponderance of weak ties on IRC influence language use in any way? And what can these patterns tell us about the nature of social interaction on IRC?

3. Data and methods

To address these questions I conducted an exploratory investigation of social networks and language variation on the channel #india on EFNet IRC. Participants on #india are mostly Indian nationals living abroad, ethnic Indians and children of nationals living in other countries (as of the date of the log studied here, taken in Fall 1997, I still had seen no participants connected from India itself). The largest number of participants connect from the US, the UK and Canada, although some also connect from other countries such as Indonesia and Thailand. The community of #india is thus a virtual community [17], in that its participants are widely distributed geographically and interact principally on-line. At the same time #india is situated at the intersection of two real-world cultures, with their own social and linguistic norms [14, 15]. On the one hand there is the bilingual culture of expatriate Indian nationals. Many Indian expatriates reside in English-speaking countries and are fluent speakers of English. Indian expatriates nonetheless place a high value on linguistic and cultural maintenance; their home/community language is typically a regional or local Indian language, and they retain ties with their Indian home communities, especially via marriage. Among the languages of India, Hindi holds a dominant position, both numerically and in social status, as it is the language of the national capital and of the largest linguistic group in India. On the other hand, younger members of the Indian expatriate community are immersed in the cultures of the local youth. In Australia, the US and the UK, this means that they attend public schools with English-speaking youth, where they are exposed to popular culture, and are expected to integrate into the dominant culture.

3.1. Data collection and coding

Interaction on #india was recorded for a complete 24-hour period by connecting with an IRC client program and capturing the entire session to a log file. The resulting 794K file was then imported into a relational database to enable coding of linguistic and interactional features. A typical portion of the log appears as in Example 6, showing the different types of messages that appear on a user’s screen when connected to the channel (email addresses have been changed to avoid identifying individual participants).

6 A representative log from #india on EFNet IRC.

```
<Pavitra> Gujju: hiya:
*** amesha (~xxxxxxxx@ppp92.xxxxxxxx.xxx.sg) has joined channel #india
*** Signoff: Sheraz (Connection reset by peer)
* Gujju walks slowly... towards.... Lilly.... and screams.... HIIIIIlili ya Lilly... to kisko milli?
<Gujju> Hi pavitra..hows it going?
* flamenco wonders if gujju is ajit the villain
<Pavitra> Gujju: its going great..and urself?
*** pyckle (xxxxxxxx@xxxxxxxx.xxxxxxxx.ca) has joined channel #india
* Gujju is no villan
* Gujju is a gujju
* Gujju laughs
* flamenco wonders if ajit the villain is gujju
*** Gujju has left channel #india
*** Gujju (-xxxxxxxx@xxxxxxxx.xxxxxxxx.edu) has joined channel #india
<Gujju> tnt2 OPEE
<lilly> no i don’t dr-k9
<Dr_K9> lilly- why not ?
*** Mode change “—— Gujju” on channel #india by KhAm0sHi
*** Signoff: DESIRABU (Leaving)
<flamenco> scatman:)) hows u?
*** Akshay (xxxxxxxx@xxxxx.xxxxxxxx.net) has joined channel #india
*** pyckle has left channel #india
*** k00LaS|Ce is now known as SEXPLOS|V
*** umdhall1 (xxxxxxxx@xx.xxxxxxxx.ca) has joined channel #india
<lilly> cause i don’t wanna
* Gujju tells Dr_K9 there is no one talking through the bot hahaha...u idiot...Khamoshi is programmed to do that..once in a while..randomly..
*** umdhall1 is now known as Panchaud
```

Each line of the log file was coded as one of four message types: system messages, commands to bots, participant turns and actions. Lines that begin with a user’s nick between angle brackets are participant turns. These lines are typed directly by users, except for the bracketed nick, which the system adds. Lines beginning with a single asterisk represent “actions”, which appear when a user types “/me” followed by some text. The system simply substitutes the user’s name for “/me”, and adds an asterisk, so that for the line “* Gujju laughts”, the user “Gujju” simply typed “/me laughts” [20, cf. 3]. Actions are usually employed to present a message in a third-person narrative voice. Many are transitive and

---

3 EFNet originally stood for “Eris Free Network” after its original (now decommissioned) hub, eris.berkeley.edu.
4 With some notable exceptions, such as a romantically involved couple, and a small group of students who attend the same university in Malaysia, as reported through participants’ interactions on the channel.
5 A difficulty with this method of data collection is that it leaves the client idle, and thus vulnerable to takeover by “IP spoofing”, i.e. connection by a hacker who falsifies the client’s IP address.
involve other participants on the channel, and so can be
taken to be directed toward other users. Lines beginning
with three asterisks are system messages representing
changes in the state of the channel (what participants are
on, who the operators are, etc.); they do not represent
communication, but contain information about the
identities of users and changes in their status.

Since only participant turns and actions are the focus of
the linguistic interaction, these lines were separated for
subsequent analysis. System messages and commands were
collected in a separate database for tracking participant
identities. Comings, goings and nick changes of users
were all compared, to arrive at a reduced set of “true”
identities (based on the email addresses recorded there). In
the analysis of participant turns and actions, initially the
“speaker” of a participant turn or action was identified as
the first word, i.e. the nick in angle brackets, or the nick
following the initial asterisk. These nick were then
located in the database of system messages to be matched
with a participant identity. A similar analysis of addressees
was also conducted. Often a speaker identifies the addressee
directly in a turn or action [20]; if they do not, then often
the addressee of a turn or action can be identified by reading
the log. When an addressee was identified in either of these
ways, the system message database was again consulted to
establish the identity of the addressee by matching the
appropriate nick with an email identity. Lines not
addressed to a specific addressee were coded as being
addressed to “all”. All told, 8199 lines of participant turns
and actions were identified. Turns addressed to “all” and
bots were excluded from the analysis, leaving 6317 lines
for further analysis.

3.2. Social network and tie strength

Since the aim of this study is to relate network tie
strength to the frequency of use of the different linguistic
features, I undertook what is known as a positional
analysis of the participants’ interactional patterns [19]. In a
positional analysis, participants are grouped in (more or
less) equivalence classes, according to their patterns of
interaction with other participants. Frequencies of
interaction among all the pairs of participants were
tabulated by sorting the database records by speaker and
addressee, and arranging the table so that speakers were
listed in the rows of the table, and addressees were listed
in the columns, resulting in a 350 speaker by 288 addressee
directional matrix [18, 19]. Such a large table, of course is
very sparse, containing mostly zeros in its cells, since
most people will never converse with one another. I thus
further chose to narrow the focus to only the 92 most
frequent participants, while investigating participants’
positions as both speakers and addressees. This necessitated
that two smaller matrices be constructed from the first; one
a 92 by 288 matrix where the 92 most frequent
participants are speakers, the other being a 92 by 350
matrix where they are addressees. The two tables thus
represent non-identical but overlapping sets of data, the
combination of which represents the total interactional
behavior of the 92 most frequent participants on the
channel. Structurally equivalent participants in each table
were then identified using factor analysis [7, 8, 16, 18];
from the first table, the 92 participants would be grouped
according to their patterns of shared behavior as speakers,
and from the second table they would be grouped according
to their shared patterns of behavior as addressees. Then
the results of the two factor analyses were collated, so that
participants’ complete interactional patterns could be
compared.

3.3. Linguistic variables

For each of the five linguistic features of examples 1-5
(Hindi and Indian languages, “r”, “u”, “z”, and obscenity),
each of the turns and actions was given a code indicating if
that feature was present or absent. For each turn, if a given
feature appeared only once or if it appeared several times, it
was merely coded as having that feature present.
Subsequently, the database was sorted by participants, first
as speakers and next as addressees, and the frequency of
each participant’s use of each feature was counted. The
database was re-sorted according to the participant addressed
in each turn, and the frequency of the linguistic features
received by each participant were also counted. The two
sets of frequencies, by use and by receipt, were then
compared with the factor coefficients obtained from the
social network analysis, as measures of participants’ social
position. All five features are predicted to be most frequent
among participants with the strongest network ties. These
predictions were tested by correlating the factor coefficients
of the 92 most frequent participants with the frequencies
of each linguistic feature, and by plotting the distribution of
social interaction and linguistic features in a reduced
sociogram.

4. Results

The results obtained from this investigation and
described below suggest a highly-structured pattern of
social interaction, wherein a particular group of
participants with a large proportion of operators is
disproportionately sought out for interaction by other
participants. This group’s language use is characterized by
greater use of Hindi and avoidance of most of the other
features, which are localized elsewhere in the network.
With respect to tie strength, only Hindi use is localized.

6 Alternatively one could interpret both the factor
coefficients and the factor scores in the factor analysis of a
single table, in order to get at both aspects of interaction.
The present approach was adopted since it allows one to
use all available information about the most frequent
participants in assigning their social positions.
among the network’s strongest ties, while the other features are localized in areas of weaker tie strength.

4.1. Social network analysis

The two factor analyses produced four-factor solutions, as determined by a scree-test [7, 16]. Participants were assigned to the factor group for which they had the highest factor loading, labeled s1 through s4 for the speaker groups and a1 through a4 for the addressee groups. Participants whose factor loadings were lower than 0.3 on all factors were considered not to load on any factor; these participants were placed in the groups labeled s0 and a0 in the speaker and addressee analyses respectively. Collating the two factor analyses resulted in identifying 13 distinct speaker/addressee types, labeled A through M for convenient reference. Additionally, participants were classified according to their operator/non-operator status, and gender, insofar as it could be determined from the information on participants’ identities. Table 1 presents a summary of these relationships.

Male and female participants are distributed over the speaker and addressee types at roughly chance levels. At the same time, there is a preponderance of operators among the a1 and a2 addressee factors, suggesting that these participants are grouped together because they are consistently sought out for interaction, and possibly administrative favors (e.g. opping or kicking other people). None of the speaker factors shows a similar preponderance of ops.

4.2. Tie strength and the participant groups

In order to reveal the nature of the tie strength among the thirteen groups, a 13 × 13 table was constructed with A-M in both the rows and the columns. Each cell of the table contains the total number of times any member of the group represented by the row addressed anyone in the group represented by the column. A grand total of 4169 turns were counted this way. This table was then compared to another similar table representing a hypothetical sample of the same size with a uniform level of interaction among all groups, adjusted to take into account the number of participants in each group. Deviances for each cell were calculated, and the magnitude and direction of the deviances were compared.

These relationships were then represented in a reduced sociogram, given as Figure 1. In Figure 1, the weakest relationships represented begin at approximately the level expected in the model of uniform interaction (See [2] for similar measures for constructing reduced sociograms). Stronger ties are represented by broader lines, and since not all interactions are symmetrical, the relative size of the arrowheads (and their absence in some cases) is used to represent such asymmetries. Finally, since members of a group often interact with members of the same group, such “self-ties” are represented by loops directed back at the

Table 1. Classification of the 92 most frequent participants on #india, by speaker (s0-s4) and addressee factors (a0-a4)

<table>
<thead>
<tr>
<th></th>
<th>s0 (48)</th>
<th>s1 (13)</th>
<th>s2 (7)</th>
<th>s3 (15)</th>
<th>s4 (9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a0 (40)</td>
<td>A</td>
<td>F</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 m+op</td>
<td>2 m+op</td>
<td>3 m-op</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17 m-op</td>
<td>2 m-op</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11 f-op</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a1 (21)</td>
<td>B</td>
<td>G</td>
<td>J</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 m-op</td>
<td>4 m+op</td>
<td>5 m+op</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 f+op</td>
<td>2 m+op</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 m-op</td>
<td>2 f+op</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 f-op</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a2 (10)</td>
<td>C</td>
<td>H</td>
<td>L</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 m+op</td>
<td>2 m+op</td>
<td>1 m-op</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 f-op</td>
<td>1 f+op</td>
<td>3 f-op</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a3 (7)</td>
<td>D</td>
<td></td>
<td>K</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 m-op</td>
<td></td>
<td>1 m+op</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a4 (14)</td>
<td>E</td>
<td></td>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 m-op</td>
<td></td>
<td>3 m-op</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 f+op</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 m-op</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Interaction among the thirteen participant groups.
originating group. The resulting diagram iconically represents the nature and strength of social ties among the thirteen groups of participants, in a way that takes group size into account. Of the 4169 turns exchanged among the members of the different groups, 2991 (72%) are represented in the reduced sociogram.

It can be seen from Figure 1 that the majority of interaction among the different groups involves members of J. All groups except A, M, L and D have a substantial number of turns directed toward J. Moreover, when a particular group addresses J some number of times, J tends to return proportionately fewer turns. The one exception to this pattern is I, which receives more turns from J than it addresses to J. Finally, J has the highest rate of self-address of any of the groups (the deviation for J’s self address is more than 7 times greater than the nearest deviance). For these reasons, J is clearly the central group in the social network of the channel #india. Since J is also among the groups with the greatest proportion operators (the other being G), we can infer that part of J’s social position comes from its preponderance of operators, and the need for other participants to interact with operators in order to obtain favors. The second operator-heavy group, group G, does not enjoy this attention, however. Members of G interact reasonably heavily with members of J and also with other members of G, suggesting cooperation among the operators of the two groups, but others in the network (except for B) do not appear to seek G out nearly so much as J. G’s social position then, although close to J’s, is decidedly less central to the channel. Members of F, another group closely linked to J, also engage in self-address. Another group that has some degree of self-address is A. Although group A is the largest of the groups identified, it is also the most heterogeneous, consisting of all those participants whose patterns of interaction did not correlate with others on the channel, whether as speakers or as addressees. Consequently A’s patterns of interaction do not cohere in the manner of those of the other groups.

At the very periphery of the network are the groups L and D. These two groups can be considered to have only weak ties connecting them to any other members of the network. Next most peripheral is M, a full three links from J, followed by A (the only group connected to M) and E, at two links away from J. E’s position is perhaps more central than A because the ties connecting it to J (through F) are stronger than those of A (through C or H). At one link away are the groups C, F, G, H, I and K, although the ties of F, G and I to J are stronger than those of C, H and K. Although the position of H is superficially similar to J, in that there are a fair number of turns addressed to J, at the same time H has no self-address, and the number of incoming ties does not outbalance the number of outgoing ties as it does for J.

Thus we can propose the following hierarchy of tie strength among the groups on #india. J exhibits by far the strongest ties of any group on the channel. Second are groups G, F and I, with B, H, K, and possibly E in a third rank. H and C are peripheral with predominantly weaker ties, followed by A, M and finally L and D. The theory of language variation as it relates to tie strength predicts that we should find all five of the linguistic features of examples 1 through 5 to be concentrated principally around J, becoming less and less frequent moving down the hierarchy, until we reach groups L and D, where they should be least frequent. It is this prediction that is tested in the next subsection.

4.3. Correlations with Linguistic Variables

In order to ascertain the relationships between the five linguistic features under study and the 13 groups of participants, the two sets of factor coefficients from the two factor analyses were correlated with the frequencies of the four linguistic variables. Since s0 and a0 represent people whose patterns of participation doesn’t correlate with anyone else’s (all of their factor coefficients are between -0.3 and 0.3), their data were excluded from the linguistic feature correlations with the speaker factor coefficients and the addressee factor coefficients, respectively. Two sets of correlations were run for each set of coefficients, one comparing the coefficients to the frequency of the features used by a given participant, the other comparing the coefficient to the frequency of the features used in address to a given participant. These correlations are presented in Table 2 (for the speaker factor coefficients) and Table 3 (for the addressee factor coefficients). The significant correlations (at p ≤ 0.05) between a factor coefficient and a linguistic feature are indicated by a box around the corresponding cells.

From Table 2A, we can see that speaker factor s1 correlates positively with obscenity, but negatively with the “r” and “u” variables, or, in other words, the higher a participant’s s1 score is, the higher his/her use of obscenity is likely to be, and the lower his/her use of the “r” and “u” variables is likely to be. Likewise, members of s2 use more “r” and “u”,7 members of s3 use more Hindi, and members of s4 use more “z” and more “u”. Notably, receipt of Hindi is the only feature that correlated with a particular speaker group; s3 is the group that is more likely to receive Hindi from other participants. As seen in table 3, members of addressee factor a1 use less Hindi, “r” and “u”; members of a2 use more “r” and “u” and members of a3 use more “z”. Members of a4 do not have any characteristic usage of any of the five variables.

The five linguistic features can be used to identify the linguistic behaviors of each of the groups of Table 1, as indicated in Table 4. The linguistic feature correlations of the speaker and addressee factors are represented at the beginning of the column or row associated with the factor in Table 3. The linguistic features associated with a particular group A-M can be read by locating the row and

---

7 The “r” and “u” variables turn out to be significantly correlated as well, r = 0.542, p ≤ 0.01.
column of the group and reading the features associated with its row and column.

Groups G and H appear to be roughly complementary in their use of “r” and “u”, in that G, being at the intersection of both s1 and a1, disfavors the “r” and “u” linguistic variables while H, being at the intersection of a2 and s2, favors both variables. G also appears to favor the use of obscenity, since G has the majority of members of the s1 factor, and while no addressee factor favors obscenity, this could be because members of G comprise less than half of the members of a1, the rest of whom (mostly from J) “swamp” any pattern of correlation between a1 and the obscenity feature. Group J (s3, a1) initially appears ambiguous regarding the use of Hindi, but the inverse correlation of factor a1 and Hindi could be at least partly due to swamping by members of G having lower use of Hindi, and scattered participants throughout having higher use of Hindi. Inspection of the database for the use of Hindi by members of J does show a high use of Hindi, while members of G use very little. Moreover, members of s3, which includes members of J, do appear to receive more turns using Hindi, and this is the only correlation for receipt of a feature that is significant at p ≤ 0.05. Group L (s4, a3) appears to be characterized by use of “z”. Other groups are possibly too small (B, C, D, F, I, K, M) or simply do not show a correlation with the features examined (E).

4.4. Linguistic features compared to tie strength

In order that the distribution of the linguistic features be more readily compared to tie strength, a second figure was prepared in which the correlations indicated in Tables 2 and 3 were superimposed on the sociogram in Figure 1. This is displayed in Figure 2. The first thing that can be noted about Figure 2 is that contrary to prediction, the features do not cluster around J. J is involved in the avoidance of two of the features, “r” and “u”. Moreover, there seems to be a central/peripheral distinction that is marked by the distribution of the linguistic features, especially “r” and “u” (which predominate among the peripheral groups), and their avoidance (Especially by the most central groups J, G, and F). The distribution of “u” includes that of “r”, but extends to other peripheral groups that do not especially use “r”. The outermost periphery (L, M and D) is characterized by the use of “z”. Hindi is represented somewhat ambiguously: groups J, K and I are involved in the use of Hindi, but groups J, F, G, and B are involved in the avoidance of Hindi — group J appears in both correlations. In fact, when the turns of J are isolated for comparison, it is evident that members of J use a substantial amount of Hindi. The contradiction can be resolved upon inspection of the turns of G, which belongs to the same addressee factor as J: members of G use almost no Hindi, effectively swamping what might have been a positive correlation of Hindi with the addressee factor a0. Since I, J and K are also identified as receiving more Hindi, it makes sense to suppose that a greater proportion

### Table 2: Correlations of linguistic variables with factor coefficients

<table>
<thead>
<tr>
<th>A. Speaker factors by use of feature:</th>
<th>Hindi</th>
<th>r = are</th>
<th>u = you</th>
<th>z=s</th>
<th>obscene</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s1</td>
<td>0.138</td>
<td>-0.362</td>
<td>-0.519</td>
<td>0.005</td>
<td>0.319</td>
</tr>
<tr>
<td>s2</td>
<td>0.291</td>
<td>0.450</td>
<td>0.342</td>
<td>-0.010</td>
<td>0.135</td>
</tr>
<tr>
<td>s3</td>
<td>0.385</td>
<td>0.129</td>
<td>-0.032</td>
<td>0.037</td>
<td>0.220</td>
</tr>
<tr>
<td>s4</td>
<td>0.132</td>
<td>0.174</td>
<td>0.366</td>
<td>0.477</td>
<td>-0.168</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B. Speaker factors by receipt of feature:</th>
<th>Hindi</th>
<th>r = are</th>
<th>u = you</th>
<th>z=s</th>
<th>obscene</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s1</td>
<td>0.205</td>
<td>-0.137</td>
<td>-0.055</td>
<td>-0.149</td>
<td>0.010</td>
</tr>
<tr>
<td>s2</td>
<td>0.088</td>
<td>-0.090</td>
<td>-0.017</td>
<td>-0.220</td>
<td>-0.050</td>
</tr>
<tr>
<td>s3</td>
<td>0.335</td>
<td>0.090</td>
<td>-0.011</td>
<td>-0.090</td>
<td>-0.213</td>
</tr>
<tr>
<td>s4</td>
<td>-0.080</td>
<td>0.292</td>
<td>-0.034</td>
<td>0.141</td>
<td>-0.213</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C. Addressee factors by use of feature:</th>
<th>Hindi</th>
<th>r = are</th>
<th>u = you</th>
<th>z=s</th>
<th>obscene</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a1</td>
<td>-0.384</td>
<td>-0.508</td>
<td>-0.434</td>
<td>-0.065</td>
<td>0.187</td>
</tr>
<tr>
<td>a2</td>
<td>0.224</td>
<td>0.498</td>
<td>0.450</td>
<td>0.080</td>
<td>0.216</td>
</tr>
<tr>
<td>a3</td>
<td>0.096</td>
<td>0.213</td>
<td>0.175</td>
<td>0.350</td>
<td>-0.049</td>
</tr>
<tr>
<td>a4</td>
<td>-0.151</td>
<td>-0.039</td>
<td>-0.202</td>
<td>0.058</td>
<td>0.110</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D. Addressee factors by receipt of feature:</th>
<th>Hindi</th>
<th>r = are</th>
<th>u = you</th>
<th>z=s</th>
<th>obscene</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a1</td>
<td>-0.100</td>
<td>-0.246</td>
<td>-0.105</td>
<td>0.027</td>
<td>0.237</td>
</tr>
<tr>
<td>a2</td>
<td>0.031</td>
<td>0.015</td>
<td>0.015</td>
<td>-0.188</td>
<td>-0.060</td>
</tr>
<tr>
<td>a3</td>
<td>0.058</td>
<td>0.248</td>
<td>-0.019</td>
<td>0.033</td>
<td>-0.231</td>
</tr>
<tr>
<td>a4</td>
<td>-0.047</td>
<td>0.061</td>
<td>0.032</td>
<td>-0.024</td>
<td>-0.051</td>
</tr>
</tbody>
</table>

### Table 3: Association of linguistic features with participant groups.

<table>
<thead>
<tr>
<th>participant groups</th>
<th>s0</th>
<th>s1</th>
<th>s2</th>
<th>s3</th>
<th>s4</th>
</tr>
</thead>
<tbody>
<tr>
<td>s0</td>
<td></td>
<td></td>
<td>+obs</td>
<td>+r</td>
<td>+z</td>
</tr>
<tr>
<td>s1</td>
<td>+r</td>
<td>+u</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s2</td>
<td></td>
<td></td>
<td>+Hindi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s3</td>
<td></td>
<td></td>
<td></td>
<td>+z</td>
<td></td>
</tr>
<tr>
<td>s4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+u</td>
</tr>
</tbody>
</table>

| a0 | A | F | I |
| a1 | B | G | J |
| a2 | C | H |   |
| a3 | D |   | L |
| a4 | E | K | M |
of Hindi is in use in interactions among members of I, J, and K. When members of these groups interact with those outside the “Hindi belt”, Hindi tends to be avoided. This is especially the case for J, whose members have considerable interaction with members of other groups. Finally, obscenity, which is a classic marker of the vernacular values associated with strong ties, is characteristic of the central non-core groups G and F, but not of the central core (J) with the strongest network ties.

In short, the relationship of the linguistic variables to tie strength reveals a far more complex arrangement than predicted. There is a central/peripheral distinction in the distribution of the features, with use of “r” and “u” functioning as a clear marker of network peripherality. Use of “z” further marks the outer periphery. The central region of the network also has a rather complex distribution of features. Hindi is preferred among the most central participants, but avoided by one prominent non-core central group. Obscenity is also found principally among the non-core central groups, and not in the core-most group. The characteristic most clearly shared by the central groups is the avoidance of “r” and “u”.

5. Discussion and Conclusions

These results appear to disconfirm the notion that within a single network “vernacularizing” linguistic variables — those that diverge from the standard — will necessarily correlate with one another and with social network ties in any simple way. Different vernacularizing linguistic variables may instead be localized in different areas of a social network. Network tie strength, while it does not predict the distribution of the variables, does provide clues to their interpretation. Hindi does appear to characterize the core-most group of participants on the channel #india, and so should be construed as a marker of core-ness to the social network of the channel associated with strong ties. Obscenity, as used by members of F and G, may express something more like the exercise of power than a commitment to the value of in-group identification.

A participant’s use of obscenity is often an excuse for operators to kick that participant; operators, who are partly immune from such actions on account of their special privileges, can get away with using more obscenity [5]. Members of F and G might therefore represent participants who are nominally Indian in their identification, and hence are drawn to #india, where they may obtain operator privileges by reason of national background. From their avoidance of Hindi, members of G may be members of the Indian community whose native language is not Hindi or second-generation immigrants with lower proficiency in Hindi. Either circumstance would limit the opportunities of members of G to interact with the central core group J, and would relegate them to a less central social position. Finally, members of C, D, H, L, and M, whose participation on the channel is more peripheral, use the “r”, “u” and “z” variables that are widespread throughout IRC, marking, in effect, that they are not “newbies”, but regular users of IRC. These variants may represent a “standard” usage particular to IRC (not necessarily standard for English speakers in general). If so, they would be naturally spread through weaker social network ties, and one would not necessarily find them among the members of the central groups on a channel such as #india. Such a hypothesis would need to be examined in greater detail by studying other IRC channels to see if similar patterns of use of these variables can be found.

The results of this study indicate that standardizing and non-standardizing linguistic changes do not map onto tie strength in any simple way. Rather, careful consideration of the relation between network tie strength and the linguistic variables offers a rich and detailed view of the function of linguistic variables as markers of social position which cannot be readily obtained through other means. The findings of this study raise new questions about the propagation of linguistic variables through social networks that could be profitably investigated through further studies using the approach presented here.
An essential characteristic of the present research was its use of persistent, textual logs of social interaction as the primary object of investigation. This characteristic is important in two ways. First, it enables the researcher to utilize quantitative analytical methods that would be unthinkable if the research were conducted on face-to-face interaction. Investigating computer-mediated interaction enables the sociolinguist to rapidly collect far more data with far less labor than would be possible if transcribed audio tapes were the medium of choice. Moreover, the textual logs contain two forms of information that are of interest to the sociolinguist: information on social interaction and information about linguistic usage. Prior methodologies would need to obtain these different forms of data from entirely different sources. The second importance of the use of persistent textual logs is the range of possibilities available for future studies. The techniques used in the present analysis can readily be applied to any other mode of CMC in which people send and receive messages. Email, Listserv messages, Usenet newsgroups, bulletin boards, and many other systems have these properties, and the relation of their linguistic (and other) characteristics to their social structures could readily be investigated, much as the IRC data used here. These systems bring with them a broad range of social environments, so a similarly broad range of questions about the relationship between linguistic variation and the social (as well as the technological) environment could be investigated. Further investigations that undertake the direct observation of language change can be envisioned, as in the diachronic study of email by Herring [6]; log samples taken at regular intervals could be used to investigate the change in patterns of tie strength and linguistic variable usage. Such studies would not be feasible without the ready availability of logable, computer-mediated social interaction.

This study thus demonstrates the feasibility of applying social network analysis, a mainstay of variationist sociolinguistic research, to the study of social interaction and linguistic variation on new media such as IRC. At the same time, the present study points out the inadequacies of current sociolinguistic theories. Further study of CMC data, representing readily recorded, detailed social and linguistic interaction, offers new empirical evidence. This new evidence brings both challenges to the old theories of language change, while promising new perspectives that will facilitate the development of new theoretical models of language variation and change.

References