Towards the Automated Social Analysis of Situated Speech Data

Danny Wyatt
Dept. of Computer Science
University of Washington
danny@cs.washington.edu

Tanzeem Choudhury
Dept. of Computer Science
Dartmouth College
tanzeem.choudhury@dartmouth.edu

Jeff Bilmes
Dept. of Electrical Engineering
University of Washington
bilmes@ee.washington.edu

James A. Kitts
Graduate School of Business
Columbia University
jak2190@columbia.edu

ABSTRACT

We present an automated approach for studying fine-grained details of social interaction and relationships. Specifically, we analyze the conversational characteristics of a group of 24 individuals over a six-month period, explore the relationship between conversational dynamics and network position, and identify behavioral correlates of tie strengths within a network. The ability to study conversational dynamics and social networks over long time scales, and to investigate their interplay with rigor, objectivity, and transparency will complement the traditional methods for scientific inquiry into social dynamics. They may also enable socially aware ubiquitous computing systems that are cognizant of and responsive to the user's engagement with her social environment.

Author Keywords

Social network analysis, wearable computing, machine learning

ACM Classification Keywords

H.1.2 User/Machine Systems, J.5 Social and Behavioral Sciences, I.5 Pattern Recognition

INTRODUCTION

Most research on social networks is ultimately concerned with face-to-face contact, which remains the primary mode of social interaction [1]. However, direct observation of interaction is difficult and expensive, and so has been applied to small study populations over brief observation periods [6]. Researchers more often use indirect measures, such as survey self-reports [7], coattendance at events and comembership in organizations [4], or contact traces such as calendar appointments or e-mail headers [9]. In this study, we demonstrate a method for automatically and directly collecting data on face-to-face interaction over extended periods.

UbiComp'08, September 21-24, 2008, Seoul, Korea.

Access to longitudinal face-to-face interaction data could help address longstanding sociological questions such as: How do global structural properties of the social network relate to the local behavior that comprises the network ties? For example, how does our network position affect how we behave and how others behave toward us? Do we interact differently within our close relationships? Behavioral measures of social interaction may also be used to understand and improve traditional methods of data collection, such as surveys.

Recent advances in wearable and ubiquitous computing continue to make gathering such real-world interaction data easier. Portable recording devices have grown in capacity while becoming smaller, cheaper, and more powerful. But observing social behavior requires more than simple audio recording capability. To collect data that captures truly natural interactions, people must be recorded as they freely go about their lives. Requiring such unconstrained recording gives rise to two problems. First, uninvolved parties could be recorded without their consent—a scenario that, if raw audio is involved, is always unethical and often illegal. Second, people may change their behavior if they know they are being recorded. For both of those reasons, a level of privacy must be maintained. Privacy-sensitive recording techniques process incoming audio, preserving data useful for sociological inquiry while discarding information deemed too invasive. It is these recording techniques that necessitate ubiquitous *computing*, and not simply ubiquitous *recording*.

In this paper, we show that it is possible, using only privacy-sensitive techniques, to discover not only the link structure of a social network but also interesting characteristics of the social behavior within the network. We analyze the real world interaction between 24 subjects over a period of 6 months, and make the following contributions:

1. Demonstrate the feasibility of fine-grained modeling of social interaction and analysis of conversations throughout the day from situated speech data.
2. Identify a relationship between speaking style and the strength of social ties: a person changes his speaking style more when talking to someone he speaks to infrequently.

Figure 1. The data collection kit worn by each subject.

3. Establish a link between speaking style and network position by discovering correlations between a person's network centrality and how much others deviate from their normal speaking style in conversing with her.

We begin by describing the data collection and processing system, present the experimental analysis and results, and then conclude with discussion and description of future work.

THE SOCIAL NETWORK DATASET

We have collected a social network corpus containing sensor-derived measurements of conversations within a group of 24 subjects. All were members of the incoming graduate class (total size: 27) of a single department at a large research university. Each subject wore a Multi-Sensor Board (MSB) containing 8 different sensors useful for detecting conversations, activities, and environmental context: microphone, tri-axial accelerometer, infrared and visible light, digital compass, temperature, barometric pressure, and humidity. The MSB was connected to a PDA that performed some processing of the sensor streams and stored the resulting data on an SD card. As shown in Figure 1, the PDA was carried in a small over-the-shoulder bag. The MSB was clipped to the bag's strap at the position of a lapel mic. A complete description of the device can be found in [2].

Data were collected during working hours for one week per month over the 9 month course of an academic year. For the analysis in this paper, we consider only the 6 consecutive months with the most time recorded (3,021 hours). Subjects also submitted surveys at the end of each data collection episode, reporting on their interactions with other participants during the previous month. Full accounts of the data collection process, resulting corpus, and difficulties encountered along the way are in [15] and [8].

Privacy-sensitive speech processing

Most speech sounds can be modeled by two separate components: (i) the source of sound generated by the vocal chords and (ii) the filter (the vocal tract: mouth, nose, etc.) that shapes the sound spectrum [10]. Prosodic aspects of speech (intonation, stress, and duration) are described by how the pitch and energy (volume) from the source change during speech. The frequency response of the filter contains information about the phonemes that are the basis for words. To reproduce speech intelligibly, information on at least three resonant peaks from the filter is required [3]. Audio processing that removes information about these peaks practically ensures that intelligible speech cannot be reconstructed. Our privacy-sensitive recording technique extracts from the raw audio a set of features that contain information about the source and the prosodic content, but not about the formants. It thus attains an intermediate level of privacy that prevents the recording of the actual words being said, while preserving information about whether and how a person is speaking.

To evaluate our ability to detect conversations, we collected 50 minutes of data from 5 people who wore our recording devices while moving around a building and entering and leaving different conversations with one another. In that data, the raw audio is saved in order to assess performance. We find that we are able to detect conversations with accuracy ranging from 96.1% to 99.2%. A full recounting of that evaluation, along with a description of the features and method we use for finding multi-person conversations and inferring who was speaking in the conversations can be found in [13].

For this paper we have added one post-processing technique to our earlier method. Once conversations are identified, we mark each participant who speaks for more than 0.5 seconds as active for the preceding 20s and following minute. If a participant is not marked as active at a given time, he is removed from the conversation. This heuristic aims to remove false positives where subjects are in the same physical location but not actively conversing (e.g. working silently in the same office, or attending a lecture).

We only detect conversations between participants wearing a device. We cannot identify or even count any non-subjects who speak with our subjects, so we cannot describe the full set of interactions for any of our participants (such as those with strangers, professors, or friends outside of school). We focus only on analyzing interaction within a bounded network of 24 students, not describing the individual ego networks of those 24 students.

Daily interaction patterns

Once we have extracted conversations from the sensor data, we can begin building a picture of social interaction. Figure 2 shows histograms of the amount of speaking time grouped into 5-minute bins from 9am to 9pm. The figure includes one histogram per weekday, aggregated across all 6 monthly data collection episodes. These patterns reveal some rhythms of the week. For example, most participants were taking required graduate courses on Tuesday and Thursday and the department frequently held special events on those two afternoons. We notice more conversational activities during lunchtime and Friday afternoons when the department's weekly social event was scheduled.

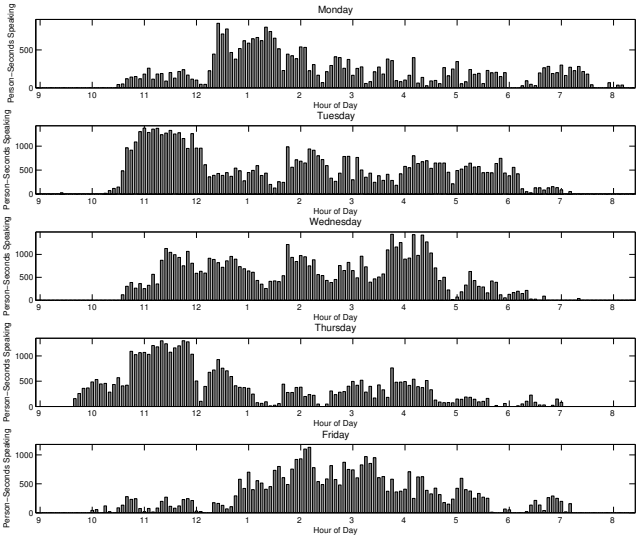

Figure 2. Histograms of person-seconds spent speaking.

SOCIAL NETWORK ANALYSIS: MICRO AND MACRO

We can also construct a network of face-to-face interaction from the extracted conversations. For the purposes of this paper, we estimate the network in the following manner: an edge exists for a pair if both subjects participate in at least 20 seconds of conversation with each other.

Although we independently assessed the accuracy of conversation detection using labeled data in controlled settings, there is no alternative source of streaming interaction data to validate the inferred networks in our subject pool. We then examine whether the inferred conversation networks correspond to meaningful social relations from the survey—subjects' self-reports of their collaboration on research or homework with peers. Indeed, the agreement between each month's conversation graph and the collaboration graph ranges from 61.6% to 82.2% and the aggregate agreement across all months is 71.3%. To put the sensor-to-survey comparisons in perspective, we also considered the expected agreement between random graphs (with same density as the survey graphs) and the survey graphs. For individual months, the difference in survey agreement between the conversation graph and random ranged from 2.1% ($p = .304$) to 13.9% ($p = 1.1 \times 10^{-4}$). When all months are aggregated, the conversation graph agrees with the survey graph 7.1% ($p = 6.0 \times 10^{-6}$) better than random. A full comparison for all months can be found in [14]. Adding a small threshold (requiring a pair to talk for a minimum amount of time) makes the difference in agreement between the random graph and the conversation graph much greater—so these comparisons are conservative.

Beyond aggregate network views

A key advantage of using ubiquitous sensing and inference tools to collect social network data is that they allow us to get not just a snapshot of the network, but also detailed records of how people interact with each other. Further, by continuously observing a group over an extended period of time, we can explore the relations between local behaviors and the global structure of the group. We present two such preliminary findings here.

Table 1. Correlation between change in speech features and tie strength

Rate		Pitch		Turn Length		Turn Frequency	
r	p	r	p	r	p	r	p
-.233	1.7E-9	-.138	.0001	-.118	.0008	-.068	.0525

Correlation between speaking style and strength of ties

The first hypothesis is that individuals change their speaking style less when interacting in their strong ties, i.e., with people they interact with often. This is based on the assumption that people's normal behavior is defined more by regular interaction partners than by rare partners.

To test this hypothesis, we estimate the proportion of mutually monitored time that persons i and j spend in conversation. We use two-party conversations only to ensure that any partner-dependent variation in speaking style is due to only one interlocutor.

We measure four features of subjects' speaking style: (i) rate, (ii) pitch, (iii) turn length, and (iv) turn frequency. To observe changes in person i's speaking style we estimate 3 quantities from the data. (1) $b_{i\setminus j}$: the mean of i's speech feature b when speaking with everyone except person j, (2) $b_{i\to j}$: the mean of i's speech feature b when speaking with j, and (3) s_i: the standard deviation of i's speech feature regardless of the interlocutor. We define $d_{ij} \triangleq |b_{i\setminus j} - b_{i\to j}|/s_i$ to be the amount that i's speech feature changes when in conversation with j. We then compute the correlation between c_{ij}, the proportion of time that i and j spend in conversation, and d_{ij} for the four speech features described above. Correlations are computed across all months by aggregating samples into one data set, and all correlations are conditional on $c_{ij} > 0$ since we can only estimate b_{ij} for pairs that spend some time in conversation.

We hypothesize a negative correlation: the more two people talk to each other, the less they change their speaking style. Table 1 shows that results support our hypothesis.

Correlation between speaking style and network centrality

The second hypothesis tested is whether individuals change their speaking style more when interacting with people who are more central to the network. For each person i, we compute his mean *incoming* change: the mean of d_{ki} for all k who speak to i. The higher this incoming mean, the more people change their speaking style when with person i.

We compute each person's centrality using a variant of closeness centrality [12], modified to take advantage of our continuous measure of the strength of social ties. We regard the 'length' of an edge i–j as $1 - c_{ij}$ (or the proportion of time that they do not spend in conversation) if $c_{ij} > 0$; if i and j do not converse at all, their edge is null and its length is undefined. We define the length of a path i–j as the sum

Table 2. Correlation between change in speech features and centrality

Rate		Pitch		Turn Length		Turn Frequency	
r	p	r	p	r	p	r	p
.307	.0003	.228	.0075	.164	.0558	-.0413	.6334

of all edge lengths along the shortest path between them. Given that each interaction graph includes one large connected component (along with, in some cases, isolated nodes with no incident edges), we compute centrality as the multiplicative inverse of the mean path length from a person to all others in the component. Individuals who are connected to others by short and strong paths are more central, and have higher closeness scores. We hypothesize that more central people will have higher incoming change.

We compute the correlation between a person's mean incoming change and her closeness centrality, aggregated across all months. Isolated nodes are excluded, since having no conversations also means having no behavioral data to consider.

We do observe a positive correlation between incoming change and closeness centrality for all speech features except turn frequency, which does not appear to be correlated with centrality (Table 2).

OTHER POTENTIAL APPLICATIONS

The two results presented in this paper demonstrate the use of ubiquitous computing technologies to make scientific queries about individual and group behavior. But the role of ubiquitous computing is not limited to data collection and basic research. The sociological insights gained through such inquiry can be used to enhance new ubiquitous computing applications.

One key piece of information needed by context-aware applications is knowledge of who the user is with [11]. We believe that additional information about a person's relationship to those whom she is with—beyond simply knowing their identities—will provide additional benefits. For example, whether an individual is engaged in a conversation is a moderately useful predictor of user interruptibility [5]. Knowledge of the tone of the conversation or the relationship between conversants may improve a system's ability to predict interruptibility.

Since the speech processing techniques we employ could theoretically run on almost any current cell phone, it is also possible for a cell phone to learn its user's conversational styles and use differences in speaking style to categorize the people to whom she speaks (assuming speakers are identifiable). Such categorization could subsequently be used to prioritize incoming calls or messages.

The two results presented in this paper suggest a relationship between the speaking styles of a conversation's participants and both the strength of the tie between the speakers and the position of the speakers within their larger social network. Ultimately, it may be possible to perform an automated social analysis of conversation data so that an application can infer the relationship between speakers. Clearly, further studies into conversational style and social relationships are needed beyond the limited one that we have presented here, but we have demonstrated a new direction for both the application and advancement of ubiquitous computing.

ACKNOWLEDGMENTS

This work was supported by NSF grant IIS-0433637.

REFERENCES

1. N. Baym, Y. B. Zhang, and M. C. Lin. Social interactions across media: Interpersonal communication on the internet, telephone and face-to-face. *New Media and Society*, 6:299–318, June 2004.
2. T. Choudhury, G. Borriello, S. Consolvo, D. Haehnel, B. Harrison, B. Hemingway, J. Hightower, P. Klasnja, K. Koscher, A. LaMarca, J. Landay, L. LeGrand, J. Lester, A. Rahimi, A. Rea, and D. Wyatt. The mobile sensing platform: An embedded system for activity recognition. *IEEE Pervasive Computing*, 7(2):32–41, April–June 2008.
3. R. Donovan. *Trainable Speech Synthesis*. PhD thesis, Cambridge University, 1996.
4. S. L. Feld. The focused organization of social ties. *Am J Sociol*, 86(5):1015–1035, March 1981.
5. J. Fogarty, S. E. Hudson, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang. Predicting human interruptibility with sensors. *ACM Trans. Comput.-Hum. Interact.*, 12(1):119–146, 2005.
6. D. R. Gibson. Taking turns and talking ties: Networks and conversational interaction. *Am J Sociol*, 110(6):1561–97, May 2005.
7. S. M. Goodreau, J. A. Kitts, and M. Morris. Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography*, 45(4), November 2008.
8. J. A. Kitts, H. I. Thielmann, E. Gleave, and R. Wang. *Survey of Social Relations and Interests Among Graduate Students*. Center for Statistics and the Social Sciences, Seattle, WA, 2006.
9. G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311:88–90, 2006.
10. T. Quatieri. *Discrete-time Speech and Signal Processing: Principles and Practice*. Prentice Hall, 2001.
11. B. Schilit, N. Adams, and R. Want. Context-aware computing applications. In *Proc. of Workshop on Mobile Computing Systems and Applications*, 1994.
12. S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
13. D. Wyatt, T. Choudhury, and J. Bilmes. Conversation detection and speaker segmentation in privacy sensitive situated speech data. In *Proceedings of Interspeech*, 2007.
14. D. Wyatt, T. Choudhury, and J. Bilmes. Learning hidden curved exponential random graph models to infer face-to-face interaction networks from situated speech data. In *Proc. of AAAI*, 2008.
15. D. Wyatt, T. Choudhury, and H. Kautz. Capturing spontaneous conversation and social dynamics: A privacy sensitive data collection effort. In *Proc. of ICASSP*, 2007.