Inferring Colocation and Conversation Networks from Privacy-Sensitive Audio with Implications for Computational Social Science

DANNY WYATT
University of Washington
TANZEEM CHOUDHURY
Dartmouth College
JEFF BILMES
University of Washington
and
JAMES A. KITTS
Columbia University

New technologies have made it possible to collect information about social networks as they are acted and observed *in the wild*, instead of as they are reported in retrospective surveys. These technologies offer opportunities to address many new research questions: How can meaningful information about social interaction be extracted from automatically recorded raw data on human behavior? What can we learn about social networks from such fine-grained behavioral data? And how can all of this be done while protecting privacy? With the goal of addressing these questions, this article presents new methods for inferring colocation and conversation networks from privacy-sensitive audio. These methods are applied in a study of face-to-face interactions among 24 students in a graduate school cohort during an academic year. The resulting analysis shows that networks derived from colocation and conversation inferences are quite different. This distinction can inform future research in computational social science, especially work that only measures colocation or employs colocation data as a proxy for conversation networks.

This work was supported by NSF grants IIS-0433637 and IIS-0845683.
Authors' addresses: D. Wyatt, University of Washington, Department of Computer Science and Engineering, Box 352350, Seattle, WA 98195-2350; email: danny@cs.washington.edu; T. Choudhury, Dartmouth College, 6211 Sudikoff Lab, Hanover, NH 03755; email: tanzeem.choudhury@dartmouth.edu; J. A. Bilmes, University of Washington, Seallte, Department of Electrical Engineering, Box 352500, Seattle, WA 98195-2500; email: bilmes@u.washington.edu; James A. Kitts, Columbia University, Graduate School of Business, 704 Uris Hall, 3022 Broadway, New York, NY 10027; email: jak2190@columbia.edu.

DOI 10.1145/1889681.1889688 http://doi.acm.org/10.1145/1889681.1889688

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]: Probabilistic Algorithms; H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing

General Terms: Algorithms, Experimentation, Human Factors, Security

Additional Key Words and Phrases: Social networks, mobile sensing

ACM Reference Format:

Wyatt, D., Choudhury, T., Bilmes, J., and Kitts, J. A. 2011. Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science. ACM Trans. Intell. Syst. Technol. 2, 1, Article 7 (January 2011), 41 pages.
DOI = 10.1145/1889681.1889688 http://doi.acm.org/10.1145/1889681.1889688

1. INTRODUCTION

Much social network research has relied on data collected via surveys that ask subjects to report their social ties (e.g., Goodreau et al. [2009]) or recall their previous social interactions (e.g., Lazega and van Duijn [1997]). When self-reports of recalled interactions have been compared to independent observations, however, the reliability of subjects' answers has been shockingly poor [Killworth and Bernard 1976; Bernard and Killworth 1977, 1979; Bernard et al. 1980, 1982]. An early study came to the dire conclusion that "people do not know, with any accuracy, those with whom they communicate" [Bernard and Killworth 1977]. Later studies found that durable, long-term patterns of communication are reliably reported, but moment-to-moment social interactions are not [Freeman et al. 1987]. More troubling for research into network structure, individuals tend to "fill in" non-existent interactions if they would increase the transitivity of the network [Freeman 1992]. Faced with these challenges, some researchers lamented that "unfortunately, most naturally occurring interactive behavior (the stuff of which networks are built) is neither observable nor conveniently recorded in some automated fashion" [Killworth and Bernard 1979].

That statement is no longer true. New technologies have made it possible to collect information about social behavior as it is enacted, instead of as it is recalled after-the-fact. Phone calls, text messages, emails, instant messages, on-line chat sessions, social media posts, and any other kind of electronically mediated communication can be automatically recorded for large groups of people, over long periods of time. Portable audio recording devices have grown in capacity while becoming smaller, cheaper, and more powerful, making it easier to record face-to-face conversations. In fact, wearable sensors now allow us to automatically record natural and spontaneous speech for an entire group of people over a long period of time.

The automated recording of real-world speech is crucial because, despite the rise in on-line interactions, face-to-face communication is still people's primary mode of social interaction [Baym et al. 2004]. Unlike methods previously employed for speech data derived from laboratory contexts, our proposed method would capture truly spontaneous speech that arises *in situ* as people enact their actual, lived relationships. For that reason, we refer to such data as *situated speech data*—data gathered *in the wild*—to contrast it with other speech data recorded in constrained or contrived settings.

Of course, obstacles to gathering situated spontaneous speech still remain, especially concerns about privacy. To capture truly natural interactions while providing a full picture of a social network, we must record people as they freely go about their lives. This requirement 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. Ideally, a privacy-sensitive recording technique will process incoming audio in order to discard any information deemed too invasive while still preserving data useful for sociological inquiry.

This dilemma illuminates what is perhaps a fundamental trade-off between privacy and quality when automatically recording social behavior. Subjects are unlikely to consent to large-scale, unrestricted recordings of their behavior, so some sociologically useful information must almost surely be destroyed. Therefore, set of features that allow us to balance this trade-off between privacy and quality is needed.

In this article, we present exactly such a set of privacy-sensitive features, together with a method for using this feature set to find colocation and conversation events in separately recorded streams of audio data. In evaluation using non-privacy-preserving test data—where access to ground truth is possible—our method performs better than earlier methods.

We then use our proposed method to derive networks of colocation and face-to-face conversation among 24 graduate students over the course of an academic year. Networks created from colocations and conversations appear to be quite different—a result that can impact and inform future research in computational social science.

The remainder of this article is divided into two broad sections. Section 2 presents the methods for discovering physically colocated and conversing people from privacy-sensitive audio data, then assesses the performance of these methods. Section 3 covers the Spoken Networks project, a data collection effort that employed the proposed methods to study a real-world network over an extended period of time, demonstrating new insights that may be available through these lenses.

2. PRIVACY-SENSITIVE CONVERSATION MODELING

When collecting situated conversation data it is necessary to protect the privacy of not just people who willingly consent to wear a recording device, but also of those who may come within range of the microphones. For this purpose, destructive processing of the audio should yield a feature set that prevents us from reconstructing intelligible speech or inferring the identities of anyone not wearing a device. A further constraint on the feature set is that all features must be computed in real-time within the limited computational resources of a wearable device—no raw audio should ever be stored, even temporarily.

At the same time, the features must still contain enough information to allow conversations to be found and meaningful inferences made about those

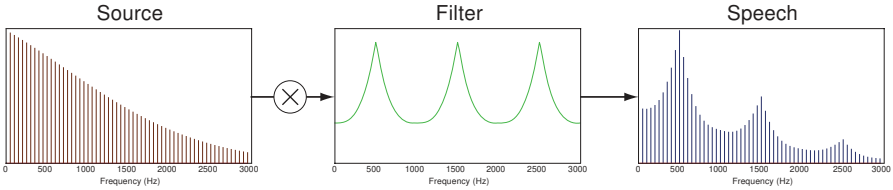


Fig. 1. Conceptual schematic of the source-filter model.

conversations. Fortunately, the nonlinguistic aspects of a conversation—who speaks when and for how long, how loud, and at what pitch—still allow for many useful analyses. Interruptions and speaking time reveal information about status and dominance [Hawkins 1991]. Speaking rate reveals information about a speaker's level of mental activity [Hurlburt et al. 2002]. Energy (loudness) can reveal a person or group's interest in the conversation [Gatica-Perez et al. 2005]. Pitch may be used for inferring emotion [Dellaert et al. 1996], and energy and duration of voiced and unvoiced regions are also informative emotional features [Schuller et al. 2004].

Here we present a set of privacy-sensitive features that can be extracted from an audio stream in real-time (Section 2.1), along with methods for using those features to automatically determine who is in conversation with whom (Section 2.2) and *how* people are speaking (Section 2.3).

Related Work. To the best of our knowledge, prior to this research, there were only two existing methods for finding conversations in separately recorded streams of audio. The method proposed by Corman and Scott [1994] computes normalized cross-correlation between raw audio signals and concludes that two people are in a conversation if their correlation coefficients are above a threshold estimated from labeled data. Obviously, using raw audio does not protect privacy, but a privacy-sensitive variant of their method is considered below. Similarly, the method proposed by Basu [2002] computes the mutual information between binary signals that represent voiced/unvoiced speech and places two people in a conversation if their mutual information is above a pre-specified threshold. Our work extends Basu's method in three important ways: (i) to detect multiperson conversations (not just dyadic), (ii) to operate at a finer time granularity while still producing a "smooth" inference over time, and (iii) to learn its threshold in an unsupervised manner.

2.1 Privacy-Sensitive Features

Following Basu [2002], our approach to extracting non-linguistic speech information builds on methods for detecting voiced human speech. A basic model for the production of human speech is the standard *source-filter model* [Quatieri 2001] shown in Figure 1. As its name suggests, the source-filter model posits two semi-independent components of speech production: (1) a source sound that is generated by the glottis and passed through (2) the filter, realized by the vocal tract, that shapes the spectrum of the source.

The source can be *voiced* or *unvoiced*. If it is voiced, the vocal cords are vibrating at what is called the *fundamental frequency*, or F0, which constitutes the pitch at which the person is speaking. A sequence of speech will alternate rapidly between voiced and unvoiced segments. Prosodic features of speech—intonation, stress, duration—are described by how the fundamental frequency and energy (volume) change during speech.

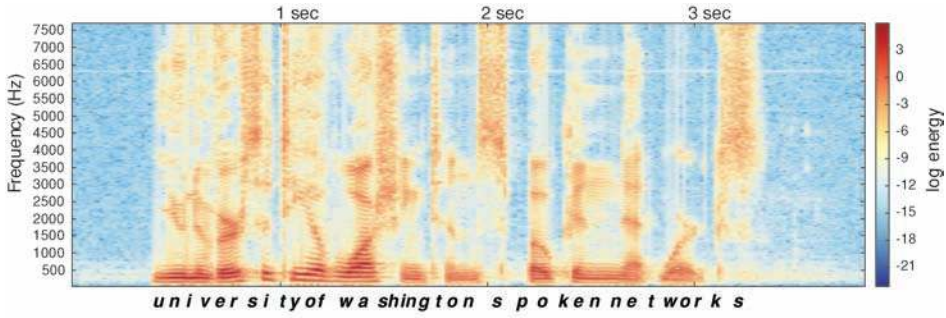
The source sound is shaped into words by changing the shape of the vocal tract. It is the frequency response of the vocal tract, particularly the resonant peaks known as *formants*, that contains information about the phonemes that are constituent parts of spoken words. Any processing of the audio that removes information about the formants will ensure that intelligible speech cannot be synthesized from the signal that remains.

Thus, to find conversations and retain information about how people are speaking, we save information about the source while discarding (almost) all information about the filter. We argue below that this preserves sociologically useful information.

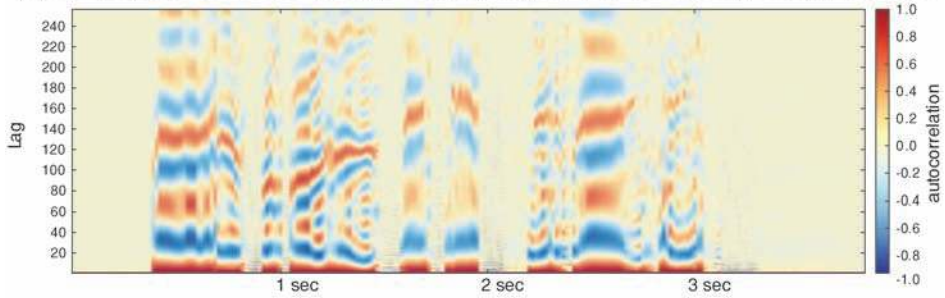
The first step in that process is finding voiced speech. Figure 2(a) shows the spectrogram for a male voice saying the phrase "University of Washington Spoken Networks." In a spectrogram, time runs along the x-axis and frequencies increase along the y-axis; color indicates energy at a given frequency. In this example, all of the phonemes are voiced except those for "s," "t," "sh," "p," and "k." The strong harmonics are indicators of voiced speech and we take advantage of that harmonicity to find segments of voiced speech.

Three features that have been shown to be useful for robustly detecting voiced speech under varying noise conditions are: (i) noninitial maximum autocorrelation peak, (ii) the total number of autocorrelation peaks, and (iii) relative spectral entropy [Basu 2002]. To provide an intuition for the first two features, Figure 2(b) shows the autocorrelogram for the example phrase. As in the spectrogram, time runs along the x-axis. The y-axis shows increasing lags at which the autocorrelation is computed, and colors show the value of the autocorrelation. The voiced segments show fewer, stronger peaks. All three features are shown in Figure 2(c). During voiced segments, the number of autocorrelation peaks drops, while the maximum autocorrelation value and relative spectral entropy rise.

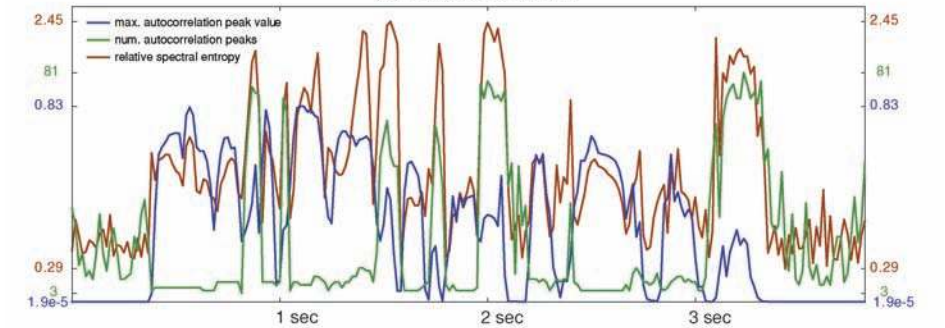
The harmonicity in the spectrogram shows that voiced speech has a low spectral entropy, compared to non-voiced regions. However, in many environments there can be noise centered strongly at a specific frequency. Figure 2(a) shows two possible examples of such noise: a low frequency hum (from 300 to 500 Hz) that may be an air conditioner, and a sharp high frequency noise (around 6400 Hz) that is probably a computer fan or hard drive. Such narrow spectrum noise will lower the general environmental spectral entropy. The relative spectral entropy is the relative entropy (also known as Kullback-Leibler divergence, see Eq. (2)) between an instantaneous normalized spectrum and the mean normalized spectrum for a much longer window of time. Relative spectral entropy captures the quick change in entropy caused by short segments of voiced speech while smoothing away any environmental reductions in entropy. Narrow spectrum noise can also create strong autocorrelation peaks.



(a) Spectrogram. The letters are approximately aligned with their corresponding phonemes.



(b) Autocorrelogram



(c) Features used for detecting voiced speech. Values are scaled to fit, so y axis labels indicate minimum and maximum values for each feature.

Fig. 2. Audio of a male voice saying "University of Washington Spoken Networks."

Fortunately, in settings where conversations can comfortably occur, such noise is usually low energy (compared to voiced speech) and its autocorrelation can be disrupted by adding low energy white noise to the signal.

The precise procedure for computing features is as follows: The 15360-Hz raw audio signal is split into frames of 512 samples (1/30th of a second) with overlaps of 256 samples (1/60th of a second). Let the vector $\mathbf{x}$ denote a single frame of raw audio data. Each frame has its mean subtracted and is multiplied with a Hamming window before applying a discrete Fourier transform resulting in a 256 point spectrum, denoted with the complex vector $\mathbf{s}$. The real-valued power spectrum $\mathbf{v}$ is derived from $\mathbf{s}$ so that $v_i = |s_i|^2$. The *energy* $r = \sum_i v_i$ of $\mathbf{v}$ is computed and saved. To disrupt low-energy, narrow spectrum noise, we

uniformly whiten the power spectrum with additional energy equal to 1% of the maximum energy possible per frame. The inverse Fourier transform of the whitened power spectrum is then taken to find the autocorrelation vector **a** [Gray and Davisson 2004] for the frame, where

$$a_j = \sum_{i=0}^{512} x_i x_{i+j} \tag{1}$$

is the autocorrelation of **x** at lag j. The number of autocorrelation peaks (defined as positive regions between zero-crossings in **a**) is counted and the value and lag of the highest peak is saved (which naturally excludes the initial maximum at lag 0). The normalized power spectrum **p**, with $p_i = v_i / \sum_j v_j$, is computed and a running mean $\overline{\mathbf{p}}$ of the normalized spectra for the last 500 frames (≈8.33 seconds) is kept. The relative spectral entropy is computed between the current normalized spectrum and that running mean as

$$D(\mathbf{p}||\overline{\mathbf{p}}) = \sum_i p_i \log(p_i/\overline{p_i}) \tag{2}$$

Altogether, we save six acoustic features: (i) value and (ii) lag of the non-initial maximum autocorrelation peak, (iii) the total number of autocorrelation peaks, (iv) instantaneous and (v) relative spectral entropy, and (vi) energy.

On the specific device we used (described in Section 3.2.1), all computations are carried out in the frequency domain using fixed point arithmetic. The logarithm required to compute entropy is not practical given the device's limited processing power. However, the device's comparatively large amount of RAM allows us to instead use a look-up table pre-populated with logarithms for all 16 bit values.

The energy is used later to determine who is speaking. The lag of the maximum autocorrelation peak is not needed for detecting voiced speech, but it is useful for determining a speaker's F0 [Rabiner 1977]. The peak will sometimes correspond not to the exact F0 but instead to one of its harmonics. Formants are expressed through the attenuation of many of the harmonics present while letting only those near the resonant peaks of the vocal tract pass through. This means that at least one harmonic (often more) will correspond to single formant. To reproduce intelligible speech, information on at least three formants is required [Donovan 1996]. Since we save at most one harmonic, and that harmonic is most often F0, we believe that our features are privacy-sensitive.

2.2 Extracting Conversation Data

To gather data about face-to-face conversations, we ask multiple people to wear recording devices each of which saves separate streams of the privacy-sensitive features described above. We then combine the streams and find conversations using four steps, each of which is described in a following subsection. First, we must find voiced speech in each person's recording (Section 2.2.1). Second, people must be partitioned into colocated groups where all the members of a group are considered "together" with each other and not together with any

person in any other group (Section 2.2.2). Third, we must infer who is speaking within each colocated group (Section 2.2.4). Finally, once colocated groups and speakers have been identified, we can conclude that people who are colocated and speaking are in conversation together and then extract further features of their conversation (Section 2.3). Figure 3 provides an overview of the entire process.

Evaluation Data. All of the techniques presented here were evaluated using a small set of labeled data collected using the same wearable devices as the large Spoken Networks corpus. To record this smaller dataset, five people wore devices for just over 50 minutes while moving around a building and entering and leaving different conversations with one another. The participants were told where to go and with whom to speak, but were not told what to talk about. The two primary locations were a quiet meeting room and a loud and noisy public space (where most of the background noise was other speech), but conversations also occurred while the participants walked together and rode elevators between locations. In order to label the data, raw audio was saved for this small set.

2.2.1 *Finding Voiced Speech.* Our method relies on first inferring whether a recorded stream contains voiced speech. We use a hidden Markov model (HMM) with one time step per 60 Hz frame (16.67 milliseconds) of audio features. The HMM's observation variable is a 3-dimensional vector containing the three features previously described as useful for voicing detection: the value of the non-initial autocorrelation peak, the number of autocorrelation peaks, and the relative spectral entropy. Let $\mathbf{x}_a$ denote the vector of observations for person a with $\mathbf{x}_a^t$ being the three observed variables at time t. Similarly, let $\mathcal{V}_a$ be the vector of hidden states for person a. The HMM defines the probability of a sequence of voicing states and observations as

$$p(\mathbf{v}_a, \mathbf{x}_a) = \prod_{t=1}^{T} p\left(\mathbf{v}_a^t | \mathbf{v}_a^{t-1}\right) p\left(\mathbf{x}_a^t | \mathbf{v}_a^t\right), \tag{3}$$

for a length T sequence, where $p(\mathbf{v}_a^1 | \mathbf{v}_a^0) = p(\mathbf{v}_a^1)$.

The observation probability is modeled with a full covariance three dimensional Gaussian

$$p\left(\mathbf{x}_a^t | \mathcal{V}_a^t = v\right) \frac{1}{(2\pi)^{3/2} |\mathbf{\Sigma}_v|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}_a^t - \boldsymbol{\mu}_v)\mathbf{\Sigma}_v^{-1}(\mathbf{x}_a^t - \boldsymbol{\mu}_v)}, \tag{4}$$

and the state transition probabilities are modeled with the usual transition matrix $\mathbf{A}$ with $A_{ij} = p(\mathcal{V}_a^t = i | \mathcal{V}_a^{t-1} = j)$. The means $\boldsymbol{\mu}$, covariances $\mathbf{\Sigma}$, and transition matrix $\mathbf{A}$ of the voicing HMM are learned from data that does not contain any speakers in our evaluation data (or in our larger corpus). This voicing HMM has been shown to be speaker-independent and robust across environmental conditions [Basu 2003].

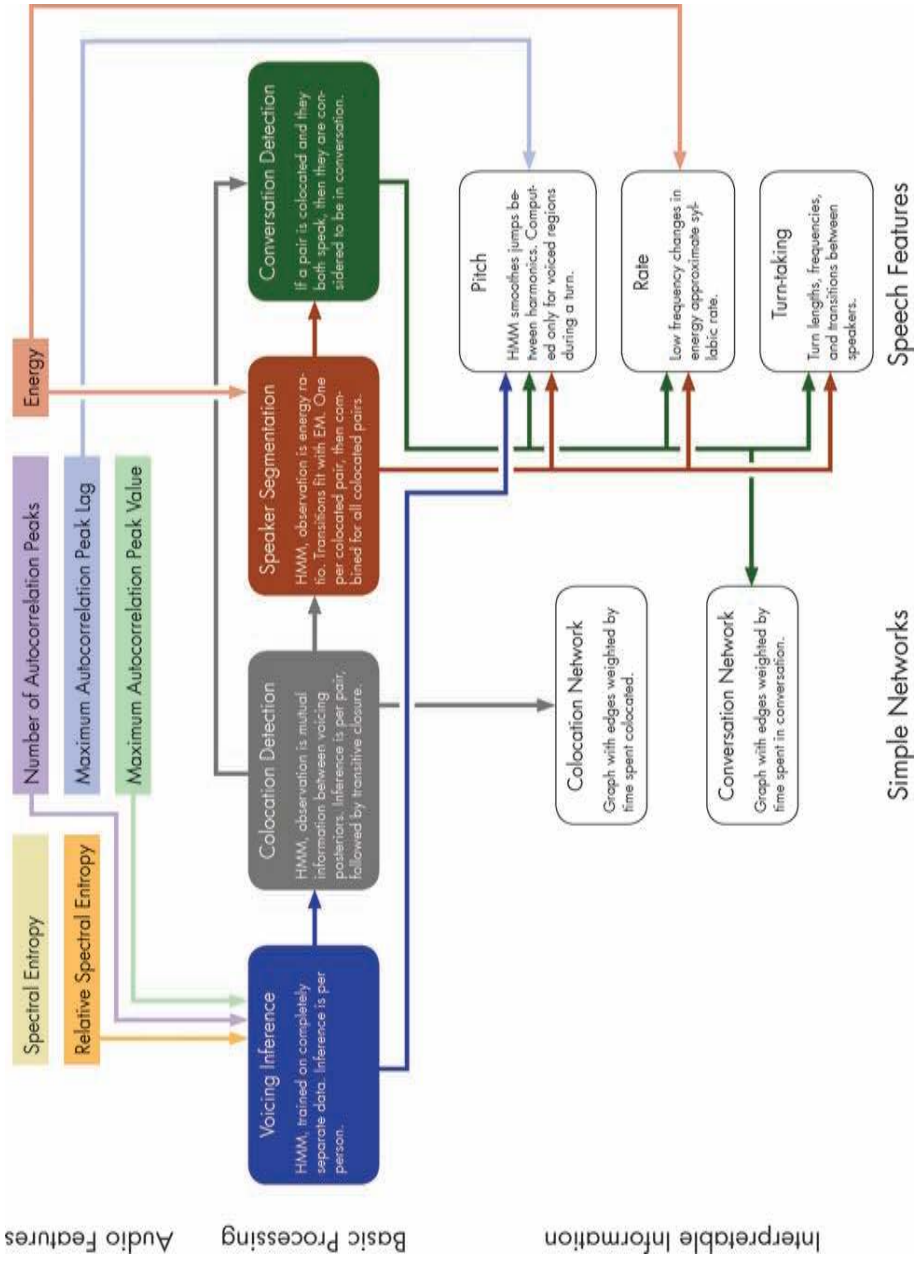


Fig. 3. A schematic illustrating the process required to go from recorded audio features to meaningful conversation data.

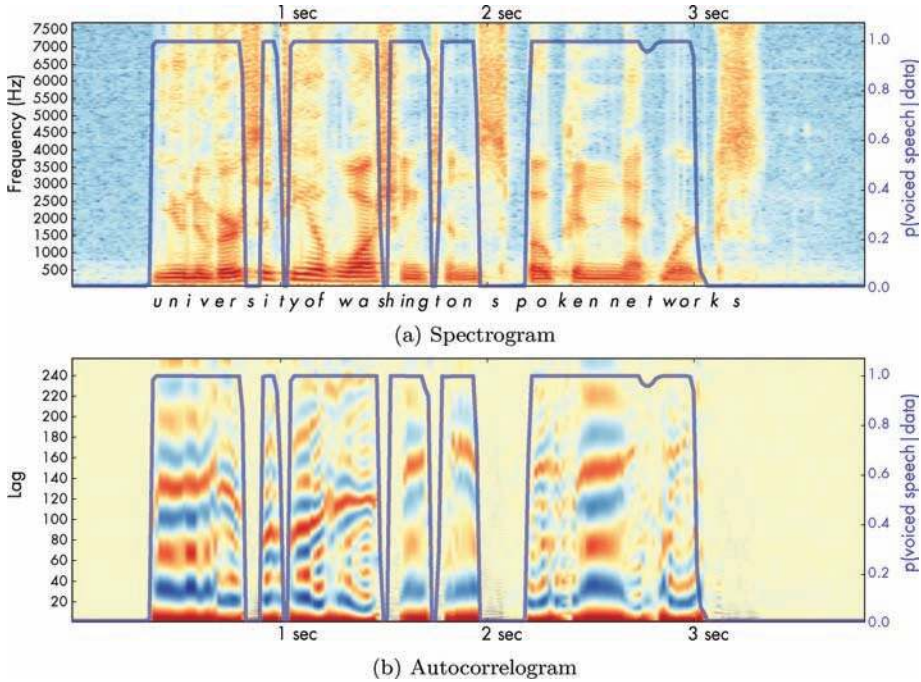


Fig. 4. Inferred voicing posterior (blue line, right y axis) overlayed on examples from Figure 2.

For each recorded stream, we use the forward-backward algorithm [Rabiner 1989] to infer $p(\mathcal{V}_a^t|\mathbf{x}_a)$: the posterior probability of voiced speech in each frame, given the entire recorded stream. Figure 4 shows the spectrogram and autocorrelogram from Figure 2 with the inferred voicing posterior for the example recording overlaid.

2.2.2 *Finding Colocated People.* We treat finding colocated groups within the multiple streams of data as a clustering problem. Successful colocation detection requires clustering together segments of data from miked individuals when they are in a conversation. Once the voicing posteriors are computed, the voicing frames are aggregated into *colocation windows* of size $W = 1200$ voicing frames (20 seconds), with no overlap between windows. To determine whether two people are colocated, we examine the mutual information between simultaneous colocation windows from each of their streams. The mutual information between persons a and b during colocation window w is

$$I\left(\mathcal{V}_a^w, \mathcal{V}_b^w\right) = \sum_{(v,v')\in\{0,1\}^2} p\left(\mathcal{V}_a^w = v, \mathcal{V}_b^w = v'\right) \log \frac{p\left(\mathcal{V}_a^w = v, \mathcal{V}_b^w = v'\right)}{p\left(\mathcal{V}_a^w = v\right) p\left(\mathcal{V}_b^w = v'\right)}, \qquad (5)$$

where $p(\mathcal{V}_a^w = 1)$ is the probability that any of the 1200 frames from person a is voiced, and $p(\mathcal{V}_a^w, \mathcal{V}_b^w)$ is the joint distribution over the 4 possible combinations of voiced states for a simultaneous frame for both a and b. Since the voicing states

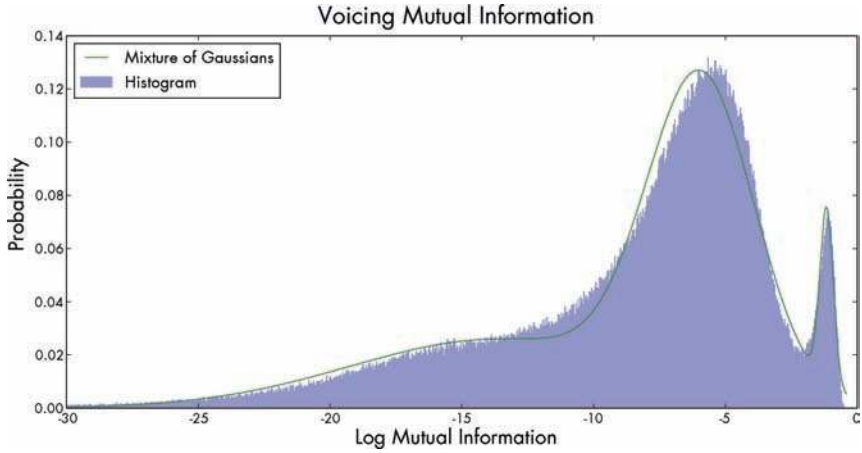


Fig. 5. Histogram of voicing mutual information values for one week of data with fitted mixture model.

are not directly observed, we estimate these aggregate voicing probabilities as

$$p(\mathcal{V}_a^w = v, \mathcal{V}_b^w = v') = \frac{1}{W}\sum_{t=\tau}^{\tau+W} p(\mathcal{V}_a^t = v)p(\mathcal{V}_b^t = v') \tag{6}$$

$$p(\mathcal{V}_a^w = v) = \frac{1}{W}\sum_{t=\tau}^{\tau+W} p(\mathcal{V}_a^t = v) \tag{7}$$

where τ is the first time index in window w. That is, we estimate the aggregate voicing distributions using their expected sufficient statistics according to the posterior distribution $p(\mathcal{V}_a^t|\mathbf{x}_a)$ computed by the voicing HMM. This allows uncertainty in the voicing inference to carry through to the conversation inference. The earlier method of Basu [2002] estimated the same probabilities using the *maximum a posteriori* (MAP) sufficient statistics (calculated from a Viterbi decoding of the voicing HMM). We gain slightly in accuracy (Tables III and IV) by using this "soft" mutual information computed from expected sufficient statistics instead of a "hard" one computed from a MAP estimate.

While there are many methods for computing a similarity between two signals, mutual information between voicing inferences seems uniquely suited to finding conversations between people wearing microphones. At the expected physical distances for a face-to-face conversation, all microphones worn by participants in the conversation will pick up the speech of any speaker in the conversation. It is extremely unlikely that two microphones that are not close enough to be in a conversation will observe the same speech signal, as we empirically demonstrate in Section 2.2.3. Other metrics (e.g., correlation between energy, considered below) do not have this property.

The voicing mutual information of Eq. (5) is computed for all windows and all pairs. The empirical distribution of the logs of the resulting values, shown for one week in Figure 5, makes the division between colocated and separated pairs clear. There is a sharp peak of high mutual information values

Table I. Colocation Inference Compared to True Room-Level Colocation

	Accuracy		Precision		Recall		Specificity		Partial Precision	
Mics	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
5	0.980	-	0.809	-	0.864	-	0.987	-	1.000	-
4	0.959	0.002	0.812	0.019	0.833	0.007	0.975	0.002	1.000	0.000
3	0.920	0.005	0.849	0.009	0.804	0.011	0.955	0.003	0.999	0.001
2	0.868	0.021	0.994	0.004	0.770	0.037	0.997	0.002	0.994	0.004
Overall	0.910	0.011	0.896	0.016	0.799	0.016	0.976	0.004	0.997	0.002

When fewer than 5 microphones are included, means and standard errors (SE) are computed across all permutations of the given number of microphones.

corresponding to colocated pairs, and two broader, overlapping peaks of lower values for separated pairs. That distinctness makes it easy to learn, in a completely unsupervised manner, different conditional distributions over log mutual information for colocated and noncolocated pairs. For that, a mixture of three Gaussians is first fit to all of the observed values (also shown in Figure 5). The component with the highest mean is taken to be the conditional distribution of the log mutual information for a colocated pair. A mixture containing the other two components (with their mixture probabilities renormalized) is taken to be the conditional distribution for a noncolocated pair.

Since the colocation windows do not overlap, temporal smoothness in the colocation inference is enforced by using another HMM to infer colocation for a pair. The hidden state of the colocation HMM is a binary variable indicating whether the pair is colocated, and its observation variable is the log of the mutual information between their voicing posteriors. The observation probabilities are set to be those of the mixtures of Gaussians and the transition probabilities are fixed so that the expected duration in either state is one minute. In an earlier technique [Wyatt et al. 2007], we did not use an HMM for colocation but instead averaged together mutual information values from neighboring time steps using a normalized triangular window. One minute was found to be the optimal window length, hence the expected duration for the HMM. The HMM-based method does not perform any differently on labeled data than the simple window-smoothed method, but on the Spoken Networks corpus it produces much more plausible colocation inferences.

To ultimately partition people into colocated groups, the MAP sequence of colocation states for each pair is computed using the Viterbi algorithm. The transitive closure of the separate pairwise inferences is then calculated within each colocation window to ensure a consistent grouping.

2.2.3 *Evaluation.* As presented so far, there is not a single, well defined ground truth for the concept of colocation. Are two people colocated if they are in the same room? What if the room is a large hall and they are on opposite sides? The evaluation data includes labels for location at the room level as well as who is in conversation with whom. Each of those could provide ground truth for the colocation inference. Table I shows our technique's performance when

Table II. Colocation Inference Compared to True Conversation Grouping

	Accuracy		Precision		Recall		Specificity		Partial Precision	
Mics	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
5	0.987	-	0.928	-	0.876	-	0.995	-	0.953	-
4	0.977	0.001	0.933	0.003	0.879	0.009	0.991	0.001	0.952	0.003
3	0.960	0.003	0.933	0.005	0.893	0.006	0.980	0.001	0.945	0.006
2	0.943	0.007	0.928	0.022	0.938	0.011	0.947	0.012	0.928	0.022
Overall	0.958	0.004	0.931	0.009	0.907	0.007	0.970	0.006	0.940	0.009

When fewer than five microphones are included, means and standard errors (SE) are computed across all permutations of the given number of microphones.

compared to "in the same room with" ground truth. Table II shows performance when compared to "in conversation with" ground truth.

There are five performance metrics presented in the tables, all derived from counts of true and false positives and negatives. To compute these metrics, we consider the set of all possible groupings of two or more people for each colocation window. If a grouping occurs in both the labeled data and the inference, then it is a true positive. If the grouping occurs in the inference but not in the labeled data, it is a false positive. A true negative is a grouping that is in neither the labeled data nor the inference and a false negative is a grouping that is in the labeled data but not in the inference. Additionally, we define the *contained false positives* to be the false positives that are nevertheless valid subgroups of a true grouping—that is, inferred groups that are missing one or more true group members but contain no erroneous members. The derived metrics are then defined as

$$\text{accuracy} = \frac{\text{tp} + \text{fp}}{\text{tp} + \text{fp} + \text{tn} + \text{fn}} \tag{8}$$

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \qquad \text{(positive predictive value)} \tag{9}$$

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \qquad \text{(sensitivity, true positive rate)} \tag{10}$$

$$\text{specificity} = \frac{\text{tn}}{\text{fp} + \text{tn}} \qquad \text{(1—false positive rate)} \tag{11}$$

$$\text{partial precision} = \frac{\text{tp} + \text{contained fp}}{\text{tp} + \text{fp}} \tag{12}$$

To test the performance of our methods in the presence of unmiked speakers, we selectively removed streams from the dataset and performed inference using only the remaining streams. Results reported for fewer than five microphones are averaged over all permutations of that number of microphones with standard errors also reported. For $k < 5$ microphones, results are computed for all $\binom{5}{k}$ combinations of excluded microphones, and the means and standard errors across these "folds" are reported. The overall result at the bottom of each table is the mean over all folds for all numbers of excluded microphones.

Table III. Other Colocation Techniques Compared to True Room-Level Colocation

	Accuracy		Precision		Recall		Specificity		Partial Precision	
Method	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
HMM										
Soft MI	0.910	0.011	0.896	0.016	0.799	0.016	0.976	0.004	0.997	0.002
Hard MI	0.900	0.012	0.893	0.018	0.769	0.018	0.977	0.004	0.999	0.001
Energy	0.944	0.008	0.937	0.010	0.872	0.015	0.981	0.004	0.991	0.002
Threshold										
Soft MI	0.909	0.012	0.898	0.015	0.795	0.016	0.977	0.004	0.991	0.003
Hard MI	0.896	0.012	0.888	0.017	0.758	0.017	0.975	0.004	0.993	0.002
Energy	0.940	0.007	0.929	0.010	0.863	0.014	0.978	0.004	0.986	0.003

The conversation comparison is slightly more favorable, suggesting that the definition of colocation implicit in our voicing-based method is "close enough to converse." That is exactly what is needed to automatically collect data about face-to-face conversations.

There are two instances in the evaluation data when the inferences disagree with one labeling or the other. First, there is one case where the five people are in two groups (of three and two people) sitting at adjacent tables in the large public space. Their room-level location label is the same (the "large public space"), but the voicing-based colocation inference separates them according to table and the energy-based colocation puts them together. By contrast, there is another case where the 5 are again in two groups but at opposite ends of a conference table in a quiet meeting room. In this case, both the voicing-based and energy-based colocation inferences place them all in one group—matching the room-level labeling but not the conversation labeling.

Comparing to Other Methods. The two previous methods for acoustic colocation detection [Corman and Scott 1994; Basu 2002] differ from ours in two ways: (i) the choice of a similarity metric, and (ii) the method of using that metric to classify pairs as either colocated or separated. Neither previous approach proposes using any method to temporally smooth the colocation classification (as the HMM does for our method). Instead, both suggest classifying windows independently of all others using a threshold learned in a supervised way from labeled data. Unfortunately, neither proposes a specific learning algorithm or loss function. As such, it is difficult to make a direct comparison between our method and the others. We can, however, use their different similarity metrics with both the simple threshold learned through our mixture of Gaussians approach as well as with our HMM.

As mentioned above, Basu's similarity metric is the "hard" mutual information between voicing inferences computed from a MAP inference of voiced states. Corman and Scott's [1994] similarity metric is cross-correlation between raw audio signals. We can approximate that metric in a privacy-sensitive way by using the energy computed for each frame of features in place of the raw audio signal.

Table III shows the results for these alternate similarity metrics when compared to "in the same room with" ground truth. The soft MI row in the HMM

Table IV. Other Colocation Techniques Compared to True Conversation Grouping

	Accuracy		Precision		Recall		Specificity		Partial Precision	
Method	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
HMM										
Soft MI	0.958	0.004	0.931	0.009	0.907	0.007	0.970	0.006	0.940	0.009
Hard MI	0.953	0.005	0.936	0.007	0.881	0.012	0.975	0.005	0.949	0.008
Energy	0.932	0.008	0.861	0.013	0.877	0.016	0.929	0.019	0.866	0.013
Threshold										
Soft MI	0.951	0.004	0.920	0.009	0.890	0.009	0.967	0.006	0.929	0.009
Hard MI	0.943	0.005	0.920	0.008	0.859	0.011	0.969	0.005	0.937	0.008
Energy	0.929	0.008	0.855	0.013	0.870	0.016	0.926	0.019	0.860	0.013

section includes the Overall line from Table I. The other rows report the same overall evaluations for different similarity metrics and classification methods. The HMM generally performs slightly better than the simple threshold, and soft mutual information generally performs slightly better than hard, but neither of those improvements is significant. More interestingly, the energy cross-correlation metric generally outperforms both voicing mutual information metrics. However, Table IV shows the results when compared to "in conversation with" ground truth. As before, the soft MI row in the HMM section includes the Overall line from Table II and the other rows also report overall evaluations. Our method (the HMM using soft mutual information) outperforms all others, significantly so for some metrics.

This suggests that voicing mutual information is a better metric for finding people who are actually in conversation, while energy cross-correlation is better for finding people who are simply physically colocated. A plausible explanation for that is that when people are colocated but in separate conversations, they are not taking turns with one another and will talk at overlapping times. The lower level voicing inference may only make inferences about the louder signal—that of the wearer—and thus the two signals will not be similar. When people are colocated and in conversation, they take turns, allowing each persons' speech to be clearly recorded on each microphone and the voicing inferences to be similar. So the voicing inference may be filtering out some "noise" that corresponds to speech that is not part of the microphone wearer's conversation.

2.2.4 *Segmenting Speaker Turns.* Once colocated groups have been found, we want to infer, in each grouping, who was speaking when. This is a task known as speaker diarization and there are a number of existing methods for it [Ajmera et al. 2004; Reynolds and Torres-Carrasquillo 2005; Anguera 2006]. However, all of the existing methods use features (primarily mel-frequency cepstral coefficients) from which the verbal content of the signal can easily be inferred, violating our privacy requirements. Our method relies on the output of our voicing classifier combined with the saved energy feature. Like our approach to colocation detection, our speaker segmentation method begins with separate inferences for each pair of people that are later combined into a global inference.

Pairwise Speaker Segmentation. First, for a given person a, the 60-Hz voicing frames are aggregated into longer *speaker frames*. We use a speaker frame size of 0.26 second (16 voicing frames) with an overlap of 0.13 second (8 voicing frames). The longer speaker frames reduce the sensitivity of the speaker segmentation algorithm to small errors in the voicing inference. The specific frame size was chosen because the NIST standard for evaluating speaker segmentation [NIST 2009] allows for 0.25 second of forgiveness around speaker turn transitions, so we are operating at the maximum conventional granularity.

Two quantities are computed for each speaker frame s for person a: (i) g_a^s, the mean energy of its constituent voicing frames, and (ii) v_a^s, the log of the sum of the constituent voicing posteriors.

For these speaker frames, we instantiate a new HMM whose hidden state $\mathcal{S}$ has four values:

(1) n: no one is speaking
(2) a: person A is speaking
(3) b: person B is speaking
(4) u: someone other than A or B is speaking.

The observations for this speaker HMM are the log ratios of the speaker frame energies: $r^s = \log g_a^s - \log g_b^s$. The speaker HMM observation probabilities, $p(r^s|\mathcal{S}^s)$, are modeled as a one-dimensional Gaussian distribution. The mean of the Gaussian for states n and u is set to 0. The mean for states a and b is learned from 3 minutes of data collected in a location and from a set of speakers that are different from those in our evaluation data. A single mean $\hat{g}$ is estimated for all pairs of speakers, and states a and b have their means set to $\hat{g}$ and $-\hat{g}$. The variances of the Gaussians for all four states (identical for a and b) are also estimated from this training data.

Generally, the log ratio r^s is greater than zero when $\mathcal{S} = a$ is speaking, less than zero when $\mathcal{S} = b$ is speaking, and $r^s \approx 0$ when $\mathcal{S} = n$ or $\mathcal{S} = u$. To disambiguate between states n and u, the probability that any person is speaking during speaker frame s is computed as $p(w^s|v_a^s) = (1+e^{\alpha-\beta k_a^s})^{-1}$ where $k_a^s = \sum_{t\in s} v_a^t$ is the sum of voicing probabilities in speaker frame s for person a. In other words, $p(w^s|v_a^s)$ is computed with a logistic regression. The parameters α and β of that logistic regression are estimated from the same training data used to learn the HMM's observation probabilities.

The speech probability $p(w^s|v_a^s)$ is incorporated into the speaker segmentation HMM as soft, or virtual, evidence [Bilmes 2004]. Virtual evidence introduces a pseudo-observation vector $\mathcal{X}$ whose value is always observed to be 1, that is, $\forall s\, \mathcal{X}^s = 1$. The observation probability for that pseudo-observation is then defined to be

$$p(\mathcal{X}^s = 1|\mathcal{S}^s = a) \triangleq p\big(w^s|v_a^s\big) \tag{13}$$

$$p(\mathcal{X}^s = 1|\mathcal{S}^s = b) \triangleq p\big(w^s|v_b^s\big) \tag{14}$$

$$p(\mathcal{X}^s = 1|\mathcal{S}^s = u) \triangleq \frac{1}{2}\big(p\big(w^s|v_a^s\big) + p\big(w^s|v_b^s\big)\big) \tag{15}$$

$$p(\mathcal{X}^s = 1|\mathcal{S}^s = n) \triangleq 1 - p(\mathcal{X}^s = 1|\mathcal{S}^s = n). \tag{16}$$

Note that the information about the voicing posterior is not incorporated through any variable's value, but instead through the inhomogeneous parameterization of $p(\mathcal{X}^s|\mathcal{S}^s)$, which varies with s.

For each conversation, the transition probabilities are set to intuitive initial values and refined using expectation-maximization (EM). We tried using the entire dataset of all conversations to learn the transition probabilities, but that degraded performance. Learning the observation probabilities, $p(r^s|\mathcal{S}^s)$, using EM also reduced overall accuracy. This suggests that speaker transitions vary for different pairs of people in different conversations, and that energy ratios are difficult to separate in an unsupervised manner. Once the EM procedure converges, we infer the posterior distribution for each speaker frame using the forward-backward algorithm.

Combining Pairwise Segmentations. Once posterior distributions over speaker states have been inferred for all pairs, those posteriors are combined into a single, global distribution for the entire group of colocated people. This is done by expanding each pairwise distribution into a larger distribution that has more than four states. Specifically, the expanded distribution has one state for each speaker who has been grouped together with the pair in the colocation step; one state for no speaker; and one state for any unmiked speakers. If there are m speakers in a conversation the probability that was assigned to state u (for a given pair a and b) is divided evenly among the remaining $m-2$ speakers' states and the unmiked speaker state. The probabilities for the other states, a, b, and n, remain unchanged.

The expanded distributions from each pair are then combined to form the global distribution. We evaluated two simple methods of combining distributions: summing $p(\mathcal{S}^s = y) = \frac{1}{Z}\sum_{a,b} p_{ab}(\mathcal{S}^s = y)$ and multiplying $p(\mathcal{S}^s = y) = \frac{1}{Z}\prod_{a,b} p_{ab}(\mathcal{S}^s = y)$, where $p_{ab}(\mathcal{S}^s = y)$ is the posterior probability computed by pair (a, b) and Z is a re-normalizing term. The summing approach achieved better empirical results so we used it to construct the final global distribution.

From this global speaker distribution, it is then easy to construct a MAP speaker segmentation vector $\mathbf{s}$ with $s_i = \text{argmax}_y\, p(\mathcal{S}^i = y)$. Note that for a conversation with m participants the values of $\mathbf{s}$ will range from 1 to $m + 2$, where the two "extra" values denote silence (no one speaking) and some unmiked other speaking.

2.2.5 *Evaluation.* To evaluate speaker segmentation, for each speaker frame we choose the most likely state from the combined speaker distributions and compare it to the ground truth in our evaluation dataset. We perform this evaluation on two versions of our evaluation data: a raw version and a smoothed version. The raw evaluation considers all frames in the data. The smoothed evaluation, in accordance with the NIST standard [NIST 2009], merges any pause shorter than 0.3 s in a single speaker's turn and ignores 0.25 s of data around a change in speaker.

Since the segmentation problem has more than two states, simple metrics (like (8) to (11)) do not readily apply. However, a full confusion matrix for each conversation is also uninformative since it is not very interesting to see how

Table V. Raw Speaker Pseudo-Confusion Matrix

		Inferred Class: None		Miked, Correct		Miked, Incorrect		Un-miked Other	
		Mean	SE	Mean	SE	Mean	SE	Mean	SE
True Class	None	**0.067**	0.003	-	-	0.072	0.003	0.017	0.002
	Miked Speaker	0.013	0.001	**0.569**	0.017	0.011	0.001	0.028	0.002
	Un-miked Other	0.018	0.002	-	-	0.091	0.007	**0.115**	0.011

Table VI. Raw Speaker Segmentation Performance

	Accuracy		Precision		Recall	
Mics	Mean	SE	Mean	SE	Mean	SE
5	0.817	-	0.825	-	0.967	-
4	0.781	0.006	0.788	0.006	0.961	0.003
3	0.750	0.010	0.756	0.010	0.956	0.004
2	0.730	0.015	0.736	0.016	0.955	0.007
Overall	0.751	0.008	0.757	0.009	0.957	0.003

often a specific person a is confused with any other specific person. We can examine pseudo-confusion matrices that show three ground truth states: no one, miked speaker, unmiked other; and four meaningfully collapsed inferred states: no one, the correct miked speaker, an incorrect miked speaker, and an unmiked other.

From these pseudo-confusion matrices, we compute three summary evaluation metrics:

(1) *Accuracy.* The fraction of frames in which the inferred state matches the ground truth state.
(2) *Precision.* The fraction of inferred-spoken frames for which the correct speaker is inferred.
(3) *Recall.* The fraction of truly-spoken frames for which the correct speaker is inferred.

Table V shows the pseudo-confusion matrix for the raw evaluation, and Table VI shows the corresponding summary metrics. Overall, the results are promising. The correct state is inferred most of the time. Importantly, miked speakers are rarely confused with one another. The most common mistake is when an unmiked other is incorrectly inferred to be one of the miked participants.

Table VII shows the pseudo-confusion matrix for the smoothed evaluation, with the corresponding summary metrics in Table VIII. Both accuracy and precision improve significantly when the ambiguous boundaries at the starts and ends of speaker turns are excluded. The confusion between unmiked others and miked speakers has also been reduced.

Table VII. Smoothed Speaker Pseudo-Confusion Matrix

		Inferred Class							
		None		Miked, Correct		Miked, Incorrect		Un-miked Other	
		Mean	SE	Mean	SE	Mean	SE	Mean	SE
True Class	None	**0.047**	0.003	-	-	0.051	0.003	0.011	0.002
	Miked Speaker	0.006	0.001	**0.645**	0.020	0.007	0.001	0.018	0.002
	Un-miked Other	0.013	0.002	-	-	0.088	0.008	**0.114**	0.013

Table VIII. Smoothed Speaker Segmentation Performance

	Accuracy		Precision		Recall	
Mics	Mean	SE	Mean	SE	Mean	SE
5	0.876	-	0.883	-	0.983	-
4	0.838	0.006	0.846	0.005	0.978	0.003
3	0.806	0.012	0.813	0.012	0.973	0.003
2	0.782	0.017	0.789	0.017	0.972	0.006
Overall	0.806	0.009	0.813	0.010	0.974	0.003

2.3 Conversation Data

The steps described so far provide ways of determining who is physically colocated with whom and who is speaking when, but they do not provide a method for determining who is in conversation with whom. Such a method is difficult to define because the ground truth for the relation "in a conversation with" is more ambiguous than physical location or speaking state.

For example, imagine two officemates a and b who work mostly in silence for two hours while occasionally talking. a makes a comment, b responds, and a short exchange ensues before they fall back into silence. When does the conversation start and when does it end? If a makes a comment later but b does not explicitly respond, is that a conversation? If a third person c enters the room and speaks to b but only a responds, who is in conversation with whom?

To define conversations for our subsequent analyses we make the following three assumptions: (i) to converse, two people must be colocated according to the voicing-based method ("close enough to converse"); (ii) all people considered to be in a conversation together must speak at least once; and (iii) "enough" intervening silence ends a person's participation in a conversation.

Making those assumptions concrete, we say that a person is *active* during a 20 second colocation frame if he speaks for at least half a second during that frame. We also say that he is active for 20 seconds before the first frame in which he first speaks (to account for people beginning to join an ongoing conversation) and that he is active for 40 seconds after the last frame in which he speaks (an ad hoc threshold for "enough silence"). If two people are colocated in "acoustic proximity" and active, then they are considered to be in conversation with each other.

Note that this is a pairwise relation. a and b can be in conversation for a long period of time if they continue speaking, but c may only occasionally be put into a conversation with them if she speaks infrequently. This may seem to exclude more silent people, but the short threshold required to be active should capture even the slightest back channel communication required for a conversation to proceed smoothly. Additionally, the previous enforcing of transitivity for the colocation relation will ensure that the conversation relation is properly transitive.

Evaluation. These heuristics happen to match our evaluation data perfectly. Consequently the accuracy of the conversation detection step depends on how well the colocation detection works. If conversing individuals are detected as being colocated then all of the true conversations are accurately classified. Thus, an evaluation comparing the resulting inferred conversations to the "in conversation with" ground truth label yields exactly the same results as comparing our voicing-based colocation inference to the "in conversation with" ground truth. The table of those results is identical to Table II, and thus we omit it for space. We plan to test the generalizability of these heuristics on future datasets.

3. THE SPOKEN NETWORKS CORPUS

Using the conversation detection methods from the previous section, we collected a corpus of real-world face-to-face conversations among 24 research subjects over an extended period of time. This section first contrasts our effort with earlier data collection projects (Section 3.1), it then explains the procedure used to gather the data (Section 3.2), provides summary statistics about the data (Section 3.3), and shows novel measures of social behavior that can be easily extracted form the data (Section 3.4).

3.1 Related Work

Our project integrates two distinct streams of research. First, it collects situated speech data for an entire subject population, building on earlier efforts in both spontaneous speech data collection and real-world social interaction measurement. Second, it monitors these interactions over an entire academic year, building on previous work in collecting temporal social network data.

Spontaneous Speech Data. Existing efforts at collecting real-world speech data have considered settings—meetings, phone conversations, interviews [Ang 2002; McCowan et al. 2003; Dielmann and Renals 2004; NIST 2009; Stupakov et al. 2009; Lian and Hsu 2009]—where the content of the speech is unpredictable, but the decision to have a conversation is made in advance. In these scenarios the dialogue is spontaneous, but the existence of the conversation is not. As such, the datasets do not capture information about their subjects' social networks.

Beyond that, most of the existing research on speech and emotion has either used acted speech data [Douglas-Cowie et al. 2003]—which is known to poorly reflect natural emotion [Batliner et al. 2000]—or small datasets limited to a

handful of observations of each subject, which cannot be used to compare one person's speech across different situations or over time (e.g., Greasley et al. [1995], Douglas-Cowie et al. [2000], and Ang [2002]). Most are also recorded in relatively unnatural settings (television shows, interviews) that are not representative of ordinary human communication. We have found only one other attempt at collecting similar data in settings as natural and spontaneous as ours [Campbell 2002], but it only recorded single participants in isolation (i.e., only one side of a conversation).

Social Behavior and Temporal Network Data. Several studies have used cell phone call data to measure real-world social interactions. For example, Onnela et al. [2007] construct an undirected network of reciprocated cell phone calls with ties weighted according to time spent in conversation. Another temporal study comes from Kossinets and Watts [2006], who analyze email sent between all students, faculty, and staff at a university over one academic year.

Of course, new data collection methods are not limited to only virtual communication. Borovoy [2002] developed a wearable badge capable of detecting physically proximate people. The badge used infrared sensors and thus could only detect people facing each other with a clear line of sight, but has been used to analyze face-to-face encounters. For example, Ingram and Morris [2007] use such infrared badges to study patterns of interaction among executive MBA students at a mixer party. Similarly, Connolly et al. [2008] use data collected from motion sensors [Wren et al. 2007] to infer social events like walking together, attending the same meeting, or coincidentally meeting in a break room. Eagle and Pentland [2006] present a system for inferring physical proximity from the short-range Bluetooth radios in cell phones, and for inferring coarse absolute location using cell tower IDs. Using this system they collected interaction data for graduate students from two different departments at one university.

The most relevant and groundbreaking real-world social behavior data collection (and the immediate ancestor of this work) has used the *sociometer*: a wearable platform combining infrared, motion, and—most importantly—audio sensors [Choudhury and Pentland 2003]. Choudhury [2004] recruited 23 members of the MIT Media Lab—including graduate students, faculty and staff—to wear the sociometer for two weeks. She was able to automatically extract conversations from the data with accuracies ranging from 64% to 88%. (That study saved raw audio, so the conversation detection could be compared to labeled data.)

Methods employing audio data have many advantages. Audio-based inference methods are not restricted to line of sight like infrared or motion sensors and will not infer colocation through walls like Bluetooth. This approach aims to capture actual interactions, not just physical proximity. And it also allows for a much finer-grained observation of the behavior during an interaction, not just the fact of whether or not an interaction occurred.

3.2 Data Collection Method

The data collection effort presented in this work descends from the original sociometer study, but differs in the research context and design. The dataset is

also richer compared to the original sociometer work as a result of the improved conversation and colocation detection techniques that we have developed.

In this study, we recruited new students as they entered a graduate program at a large research university. We collected data on two student cohorts in the period 2004-2007 and analyze one of these cohorts here, including 24 of the 27 students who enrolled in the department in that year. This research site allows us to study the initial formation of a network among a group of peers as well as the dynamics of this network over the academic year. This design builds on Choudhury's [2004] study, which included professors as well as students of varying seniorities, began measuring their interactions after they were already acquainted, and monitored them for only two weeks.

Our subjects recorded data by wearing a personal digital assistant (PDA) with an attached sensing device (described in more detail below). Subjects recorded data during whatever period they considered their "working hours." They recorded daily for one work week (five days) each month over the nine-month course of an academic year. The first week had only three working days and the last only four, for a total of 42 collection days. Aside from the days and hours, no other restrictions were placed on data collection. The subjects recorded data everywhere they went, indoors and outdoors: class, lunch, study groups, meetings, spontaneous social gatherings, etc.

Data was saved to a 2-GB Secure Digital (SD) flash memory card on the PDA. Subjects were asked to upload their collected data at the end of each collection day, but because their memory cards could hold an entire week of data most waited until the end of the week. The subjects were paid for each day of data that they submitted. They were also allowed to use the PDA during noncollection weeks and were given the PDA at the end of the study.

Research subjects completed questionnaires before beginning the school year, at the end of each data collection week, and following the end of the school year. These surveys used conventional retrospective questions to measure subjects' substantive relations with one another (e.g., collaborations on homework or research, social visits outside of school), their research interests, and other basic information (e.g., race, gender, age, languages spoken).

3.2.1 *Hardware and Software for Data Collection.* All conversation data discussed here was collected using the same platform: an HP iPAQ hx4700 PDA with an attached multi-sensor board (MSB, Figure 6) containing eight different sensors.

The PDA was carried in a small over-the-shoulder bag and the MSB was connected to the PDA via a USB cable that ran discreetly down the bag's strap (Figures 7(a) and 7(b)). The MSB was clipped to the strap, like a lapel microphone. Recording could be started and stopped with the press of a single hardware button on the side of the PDA and the screen indicated whether the device was recording, how much data had been recorded, how much battery power remained, and an estimate of recording time left with the available battery power (Figure 7(c)). The PDA has an Intel XScale PXA270 624-MHz processor, with no floating-point unit, and 64 MB of RAM. As mentioned above, all data was saved to an SD card, with files rotated every half hour. The file

Fig. 6. The MSB. Microphone is at top.

(a) Front: MSB is on right shoulder (b) Back: PDA is in bag. (c) PDA and data collection program.

Fig. 7. The data collection kit worn by each subject.

rotation was implemented to prevent any accidental corruption from spoiling an entire data collection session, but in practice corrupted files were very rare.

The most important sensor for conversation detection is clearly the microphone. The MSB's microphone is an inexpensive electret condenser microphone that records 16-bit audio at a rate of 15,360 Hz. Though not addressed in this paper, the MSB also contains seven other sensors that sample at varying rates: triaxial accelerometer (550 Hz), visible light (550 Hz), digital compass (30 Hz), temperature and barometric pressure (15 Hz), infrared light (5 Hz), and humidity (2 Hz). These sensors can be used to infer the wearer's physical activity (e.g., walking, sitting, standing, etc.) and whether she is indoors or outdoors [Lester et al. 2005]. In addition to the data gathered via the MSB, the PDA records (at 0.5 Hz) the MAC addresses and signal strengths of the 32 strongest

visible WiFi access points. We had hoped that the WiFi data could be used to determine the wearer's absolute physical location [Ferris et al. 2006], but repeated attempts to infer locations from the recorded data were unsuccessful. Unlike audio, the raw data from the additional sensors and the WiFi readings are saved in their entirety with no initial feature processing. We also intended to collect GPS data using a separate unit which communicated to the PDA via bluetooth but the subjects often forgot to recharge the separate unit. In addition, most of the data was collected indoors so the GPS data did not add enough information to justify the additional bluetooth communication cost.

For the 24 subject population, raw audio was not saved—even temporarily—on the device. The privacy-sensitive features described in Section 2.1 were computed in real-time on the PDA and only those features were saved. For the five subject group that generated the evaluation data described in Section 2.2, raw audio was saved in addition to the privacy-sensitive features. That group contains no subjects from the larger study population and all members of the evaluation group consented to have raw audio recorded during their 50 minutes of observed interactions.

3.2.2 *Problems Encountered.* We encountered four significant technical problems during data collection. First, batteries died faster than anticipated. We discovered that the PDA's operating system was attempting to connect to known WiFi networks in weak signal conditions that we had not previously tested. We alleviated this problem by reconfiguring the OS to never attempt to connect to any network while the data collection application was running. Second, although subjects found it easy to recharge their PDAs at the end of each day, they would often forget to charge them between collection weeks. Because all of the PDA's software and settings are stored in volatile RAM and are completely lost if the battery fully discharges, this led to many Monday mornings of lost recording time while PDAs were reconfigured. Third, the PDAs' clocks are shockingly unreliable. We found them to drift up to 5 minutes between collection weeks, requiring resynchronization with a time server. The fourth significant problem was that the cable connecting the MSB to the PDA's USB card was not durable enough. Over time, the PDA would intermittently lose its connection to the MSB, requiring replacement of the cable.

Each of these problems ultimately arose from our stretching the iPAQ PDA well beyond its intended use. It was meant to be turned on only sporadically for short tasks, not to run continuously as its user goes about her day. The PDA was also intended to be attached to a computer regularly, providing it with the opportunity to charge its battery and synchronize its clock. While these PDAs were handy portable platforms for short data collection efforts, they were not suited to long term collection efforts such as ours. Fortunately for subsequent efforts, newer platforms—particularly smart phones—are much better suited to running long-lived, independent data collection tasks.

3.3 Collected Data

Our subjects gathered a total of 4,401.51 hours of data, an average of 183.40 hours per subject. The amount of data collected for each participant

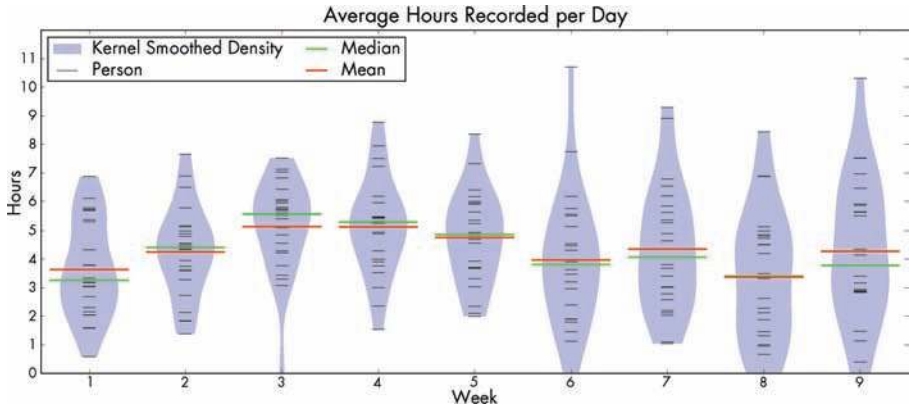


Fig. 8. Average hours recorded per day for each subject in each week. Black lines are data points: the average for one person for that week. Blue "beans" are kernel density estimates. Green lines are medians and red lines are means.

varied greatly, from a maximum of 321.53 hours to a minimum of 88.41 hours. Each subject recorded 4.27 hours per study day on average, with a minimum of zero and a maximum of 10.71 hours recorded in a single day.

Figure 8 shows beanplots [Kampstra 2008] of the average number of hours collected per day for each collection week. (Beanplots are an alternative to box plots that allow for comparison across weeks while also showing more information about the specific distribution of data within each week.) The first three weeks (i.e., representing the first three months of the academic year) show an increase in the amount of data collected as the subjects initially learned how to use the devices and we resolved battery and software problems as previously described. Recording hours diminished slightly in the later weeks, also due partly to technical problems with the cables and perhaps because the participants became fatigued or the study became less novel to them.

Since colocated people and their conversations can only be found when participants are simultaneously recording, the number of overlapping recordings is more important than the raw amount of data collected. Figure 9 shows histograms of the number of people simultaneously recording any 20 second window in the data (a window is only in the data if at least one person recorded it). While there is no moment when all subjects are recording (the maximum number of simultaneous recordings is 21), there is enough overlap in the data for it to contain many interactions among our subjects. The average number of simultaneous recordings per window is 8.10 for the entire corpus, and 88.53% of all recorded windows are covered by at least two recordings. Simultaneous recording time varies across pairs from a minimum of 16.13 hours to a maximum of 215.18 hours.

3.4 Basic Behavioral Inferences

Data processing follows the three steps described in Section 2: colocation detection, speaker segmentation, and conversation extraction. Recall from

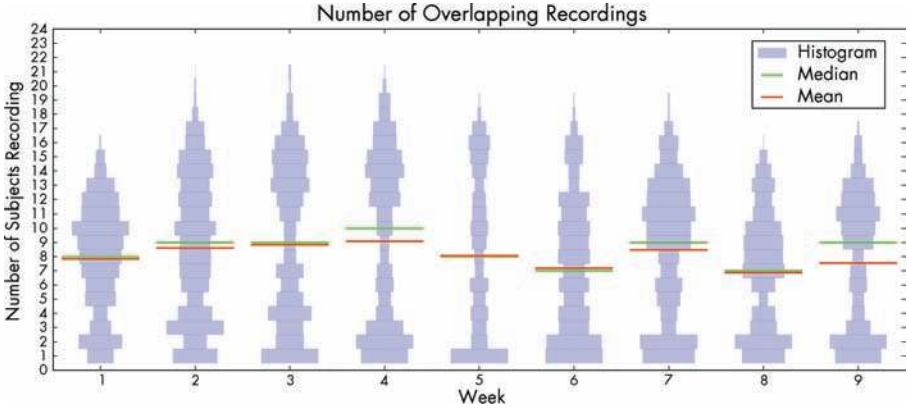


Fig. 9. Number of people simultaneously recording each 20 second window with at least one person recording. Stacked blue boxes are histograms with one bin for each possible number of simultaneous recordings. The width of the box reflects the number of windows simultaneously recorded by the corresponding number of subjects. Green lines are medians and red lines are means.

Section 2.2.2 that colocation inference based on energy is more accurate when compared to ground truth for physical location, but colocation inference based on voicing mutual information is more accurate compared to ground truth for conversations. Since each method presents distinct strengths for sociological analysis, both were used to create separate colocation inferences for each week. The heuristics in Section 2.3 are used to group subjects into actual, interacting conversations, and pairs are only considered for conversation grouping if they are first determined to be colocated using the voicing-based method.

3.4.1 *Inspecting Daily Patterns.* The times of day that recording devices are active provides information about our subjects' daily schedules. Figure 10 shows the number of subjects recording over the course of each day during week 4. Most subjects begin recording between 9 am and 11 am and gradually stop between 5 pm and 7 pm. The long slopes at both ends of the day show that students keep different hours but most are present and recording during the middle of the day.

There is a noticeable increase in the number of subjects who begin recording shortly before 10:30 am on Tuesday and Thursday. During this academic term, most subjects attended a class that met from 10:30 am to 12:00 pm on these days, so many students arrived at school and began recording before that class.

The colocation inferences in Figure 11 show the class much more clearly. Figure 11 shows the inferences for colocation using both energy and voicing mutual information, as well as the conversation grouping. At each point in time, the number of pairs inferred to be together or in conversation is normalized by the number of pairs simultaneously recording at that moment. Thus, each line is interpretable as the proportion of currently recording pairs grouped together according to each method. Because of that, when few people are recording (see Figure 10) even a small group of interacting subjects will appear as a larger proportion in the plot. This is most apparent at the end of the day.

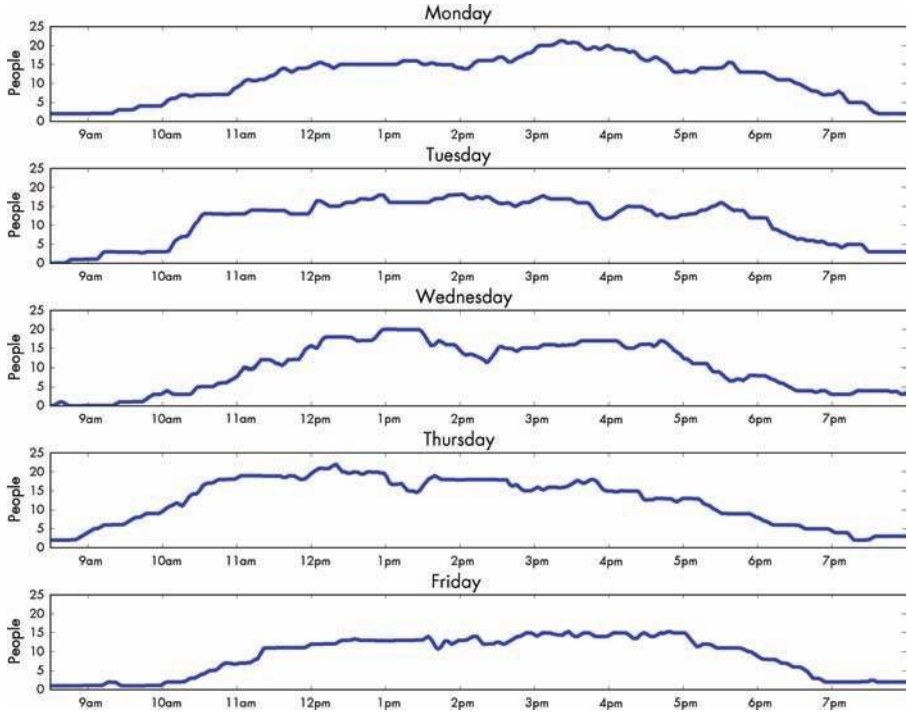


Fig. 10. Number of people simultaneously recording over the course of each day during week 4.

During class on Tuesday and Thursday morning, the two colocation methods largely agree with one another since the quiet of the class and the common signal of the instructor's voice will match in both energy and voicing inference. There are also classes on Monday and Wednesday from 12:00 pm to 1:30 pm, and on Thursday from 3:00 pm to 4:30 pm. All of these classes appear similarly in the colocation and conversation inferences.

There is a department-wide social gathering on Friday afternoon. The energy-based colocation puts many pairs together, but the voicing-based colocation does not. This corroborates the earlier observation that during periods where the background noise represents other conversations, the voicing colocation groups people into smaller (conversation-sized) groups while the energy colocation groups them by broader physical location. This informs our interpretation of the two colocation measures.

3.5 Basic Network Analyses

Constructing networks from survey data is usually simple: they are often just the union of self-reported ties for each actor in the network. Deriving networks from social behavior data is not so straightforward. Many short interaction records need to somehow be aggregated into a single network. This process of aggregation generally involves two steps: (i) aggregating observations across time into *temporal windows*–periods during which all observations

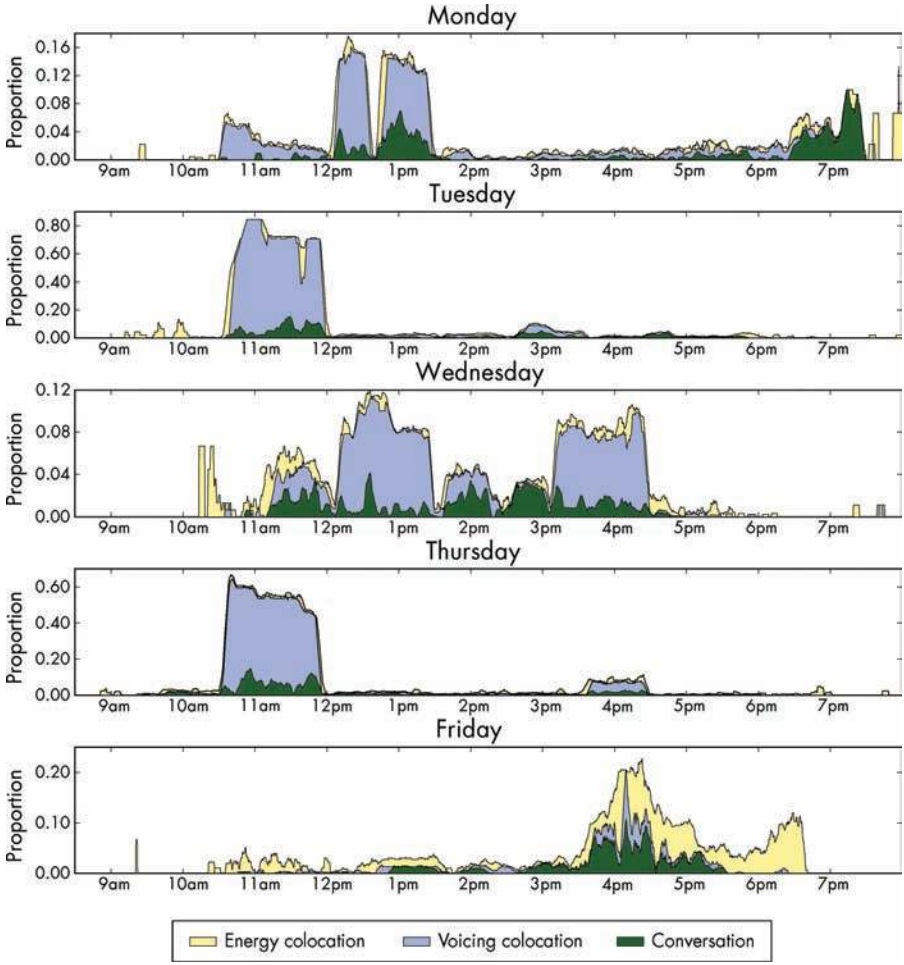


Fig. 11. Proportion of recording pairs that are physically colocated in accordance with energy correlation, voicing mutual information, and in conversation.

are assumed to correspond to a single, static network, and (ii) deriving some measure of an edge from the raw interaction data within a window.

Aggregation across time is often necessary because short, nearly instantaneous observations rarely contain enough network structure to be interesting. For example, in the Spoken Networks data it is theoretically possible to observe interaction at a granularity of 20 seconds, but it is unlikely such small snapshots would contain more than a handful of ties and the structures observed at any point in time would be ephemeral. By contrast, windows that are too long risk "blurring together" separate stages of the network's evolution, obscuring changes in the network over time and producing structures that do not correspond to any observed network. A balance between the two—one that offers a meaningful view of robust structural patterns while also allowing analysis of how those patterns change over time—must be found.

Each pair may be connected by a number of interaction events within a time window. If simple networks are desired, some method is needed for deriving a single value for each edge from this rich information. Since most network analysis techniques have been developed for binary networks, many studies of social behavior data have resorted to defining simple thresholds that distinguish binary ties from non-ties (e.g., Kossinets and Watts [2006], Palla et al. [2006], and Leskovec et al. [2008]). An alternative is to use weighted edges, but this requires using less conventional network analysis methodology.

For the simple analyses presented in this section, we define our temporal window size to be one work week. We aggregate data to the entire study-week because we are not interested here in the variation of interaction patterns from day to day (e.g., induced by class schedules, as suggested by Figure 11) and we are interested in changes in these robust interaction patterns over the course of the year. Note that one week is the longest window of contiguous observation, because we sample only one week per study month.

For each week, we construct two networks: the colocation network and the conversation network. In the colocation network edges between pairs indicate time spent in the same physical location (using the energy-correlation colocation detection, not the voicing-based method). Similarly, in the conversation network edges reflect time spent in conversation. Instead of selecting an arbitrary threshold, we consider weighted networks. However, since we can only observe an interaction between two people if both are simultaneously recording, we normalize the observed interaction times by the amount of data available. Specifically, let o_{ij}^t be the amount of overlapping time in i's and j's recordings during week t. Let l_{ij}^t be the time the pair is inferred to be physically colocated, and c_{ij}^t the time they are inferred to be in conversation. We define two networks: (i) the colocation network $\mathbf{L}^t$ where $L_{ij}^t = l_{ij}^t / o_{ij}^t$: the proportion of observed time that i and j spend colocated; and (ii) the conversation network $\mathbf{C}^t$ with $C_{ij}^t = c_{ij}^t / o_{ij}^t$, the proportion of observed time that i and j spend in conversation. Defining edge weights to be proportions has the added benefit of ensuring that they are between zero and one. Many metrics developed for binary networks can then be applied without much modification, since a binary network is a special case of such a normalized weighted network where all ties (and non-ties) take on only the most extreme values.

Figure 12 shows the conversation networks constructed for each week. Obviously, a visual comparison of the networks provides limited insight. The rest of this section considers four simple network properties that can be more easily compared: network density, degree distributions, two measures of transitivity, and path lengths. We examine both how these properties change over time, and how they contrast between colocation and conversation networks.

3.5.1 *Density.* The density of a network is its mean edge value:

$$d(\mathbf{Y}) = \frac{1}{\binom{N}{2}} \sum_{i,j} Y_{ij} \tag{17}$$

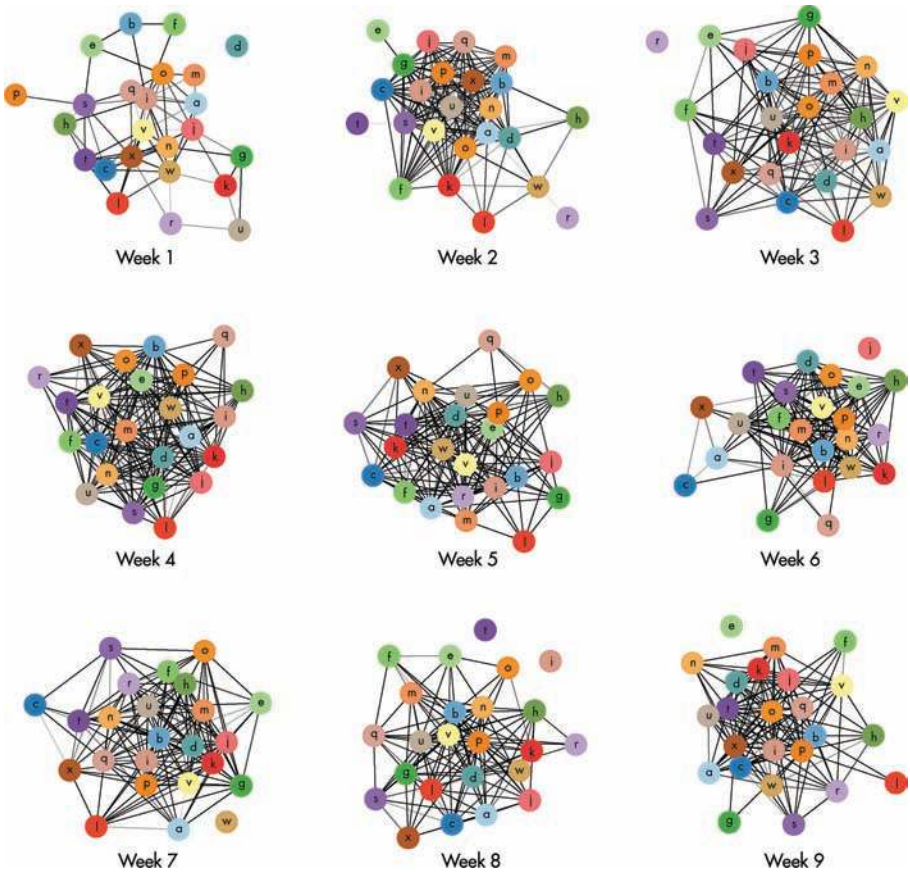
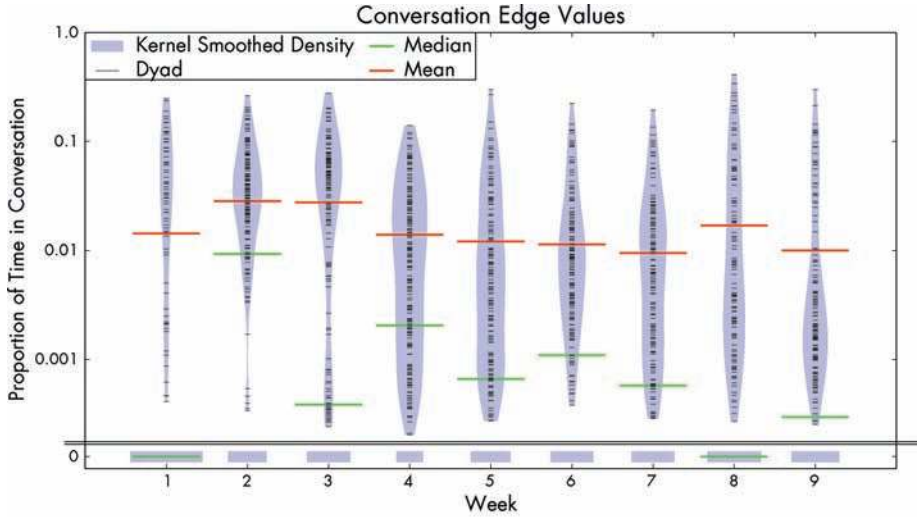


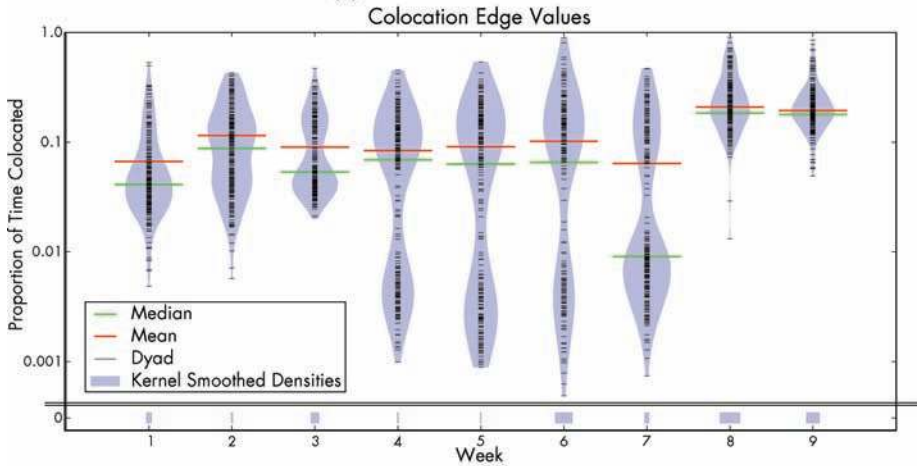
Fig. 12. Conversation networks for each week. Edge shades correspond to proportion of time spent in conversation.

For weighted networks, this has all the ambiguities inherent in summarizing a dataset with its mean. For example, a weighted network with a few very strong edges may have the same density as one with many weak edges, despite the fact that they are very different networks from other perspectives. Density shows how much interaction exists in the network, but it does not reflect how that interaction is distributed. It is more illuminating, then, to consider the full distribution of edge values together with its mean.

Figure 13 shows those distributions as beanplots for the conversation and colocation networks across all weeks. The red line on each bean is the value of (17) for the week. Since most edges have very small values, it is helpful to show them on a logarithmic scale in order to see all of the variation in the data. Of course, zero values cannot be shown on a log scale. Figure 13 thus shows a separate box or bin whose width corresponds to the number of zero-valued edges in the network. The blue beans are kernel-smoothed densities for the log transformed data. Thus, the width of the bean at some point y on the y axis corresponds to $p(\mathcal{Y}_{ij} = y|\mathcal{Y}_{ij} > 0)$. The width of the box corresponds



(a) Conversation networks



(b) Colocation networks

Fig. 13. Edge value distributions. The data has been split into zero and non-zero valued edges. The width of the blue box at the bottom corresponds to the number of zero-valued edges for that week. The blue beans are kernel smoothed densities of log-transformed non-zero edge values. The width of the zero boxes and the beans can be compared: a wide zero box shows that there are many zero-valued edges and results in a thinner bean for the non-zero edges.

to $p(\mathcal{Y}_{ij} = 0)$. The width of a box and that of the corresponding bean can be compared: a wide box means there are many zero valued edges, and the bean will be thinner. Note that the means and medians are computed from *all* values, both zero and non-zero.

This distinction is necessary for the log scale display, but it also corresponds to a very natural intuition about weighted networks. There is a difference of kind, one beyond the simple difference in value, between zero valued edges and non-zero edges. Adding a new edge, even one with a minuscule value,

can have drastic effects on the path lengths, reachability, and connectivity of the network. The box/bean split in Figure 13 can quickly provide a picture of the ratio of zero-valued edges to non-zero-valued edges. The median lines also provide information about the ratio: for weeks 1 and 8 the median proportion of time spent in conversation is zero and thus more than half of all pairs are not connected by an edge in the conversation network.

When comparing across weeks, the conversation edge values in Figure 13(a) show very different distributions. The early weeks seem almost bimodal, while the later weeks have elongated densities with gradual, almost linear decreases. Since the plot is on a log scale, this linearity corresponds to a roughly exponential decrease in probability for higher valued edges, a fact also reflected in the distance between the means and medians. There are certainly differences between weeks, but no pattern is immediately obvious.

A more useful comparison is that between the conversation and colocation networks. Figure 13(b) shows the same edge value distributions as in Figure 13(a), only derived from the colocation networks. The differences between the networks are discussed further in Section 3.5.5.

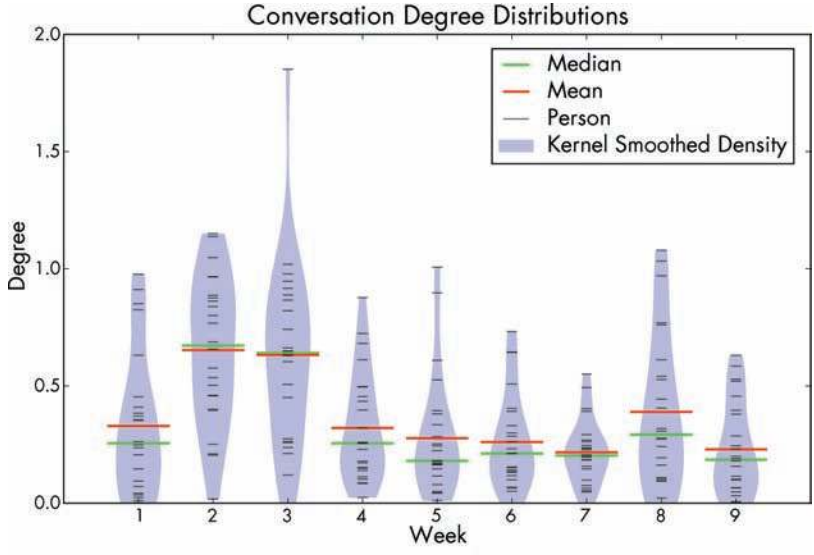
3.5.2 *Degree.* The degree of a node is the sum of the values of the edges incident to it: $d_i(\mathbf{Y}) = \sum_j Y_{ij}$. Different people may have different levels of interaction, and patterns in those differences can be seen in the network's degree distribution.

Figure 14 shows beanplots of the degrees of each person for the conversation and colocation networks over all weeks. As with the edge value distributions, the values for the colocation degrees are much higher than those for conversation degrees and the two kinds networks seem to be very different with regard to degree. This difference is also discussed below in Section 3.5.5.

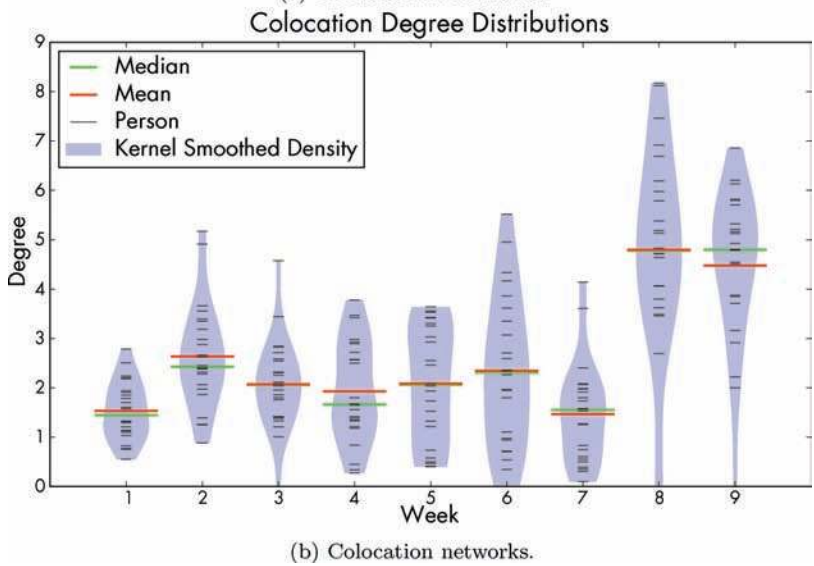
3.5.3 *Transitivity.* An important property of social networks is their tendency to be transitive: people who are tied to one another tend to both have ties to the same people. More colloquially, people who share a mutual friend tend to be friends. Transitivity expresses itself through a greater number of triangles in the network, and thus metrics for quantifying transitivity are usually based on counts of triangles. In this section, we will consider two such metrics: the clustering coefficient and the global triangle count.

The clustering coefficient for a person is defined as the fraction of pairs to whom she is tied who also have ties to each other [Watts and Strogatz 1998]. Equivalently, it is the number of triangles that involve her divided by the total number of triangles that could involve her given her observed set of ties. Since the metric relies on the discrete existence or nonexistence of ties, it does not generalize to weighted networks as easily as density and degree do. There are, however, several proposed variants of the clustering coefficient that can be used with weighted networks. The one we use is the weighted clustering coefficient defined by Saramäki et al. [2007]:

$$C_i(\mathbf{Y}) = \frac{1}{k_i(k_i - 1)} \sum_{j,k} (\hat{Y}_{ij}\hat{Y}_{ik}\hat{Y}_{jk})^{1/3} \tag{18}$$



(a) Conversation networks.



(b) Colocation networks.

Fig. 14. Degree distributions.

$\mathbf{Y}$ is a weighted adjacency matrix and $\hat{\mathbf{Y}} = \mathbf{Y}/\max(\mathbf{Y})$ is the normalized adjacency matrix where the maximum edge value is one. The weighted clustering coefficient defines the "intensity" of a triangle to be the geometric mean of the edges involved and thus is equivalent to the traditional clustering coefficient if edges take only zero or one values. $k_i = \sum_j \mathbf{1}_{[Y_{ij}>0]}$ is the "structural" degree of person i, and thus (18) captures the amount of triangle intensity that exists, divided by the total possible intensity (e.g., if i belonged to a clique where all edges have value one).

Another more global measure of transitivity is the simple count of all triangles in the network [Davis 1970; Holland and Leinhardt 1975]. As with the clustering coefficient, the triangle count does not generalize as easily to weighted networks as degree and density, but, following Saramäki et al. [2007], we can define a weighted triangle value as

$$T_{ijk} = (Y_{ij}Y_{ik}Y_{jk})^{1/3} \tag{19}$$

As with (18), this value is equivalent to the ordinary triangle indicator if $\mathbf{Y}$ contains only binary values. As with edge values, looking at the distribution of the weighted triangle values will provide more information about transitivity in the network than the mean (or sum) alone would.

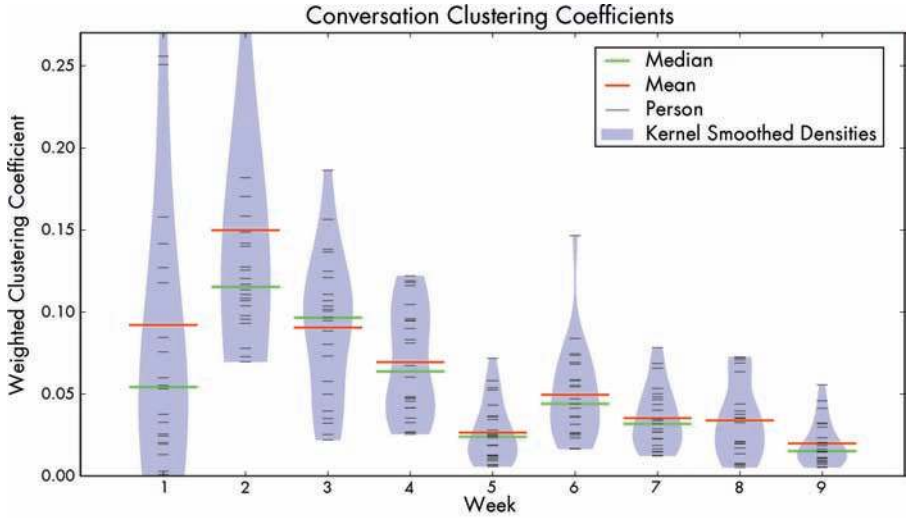
Figure 15 shows beanplots of the weighted clustering coefficients, and Figure 16 shows beanplots of the log scaled weighted triangle values. In Figure 16, the ratio of zero to non-zero values is shown as it was in Figure 13.

For both metrics, there are extreme differences between the conversation and colocation networks. The clustering coefficient values are much higher in the colocation networks. The median triangle count for the conversation networks is always zero: of $\binom{N}{3}$ potential triangles, the majority do not exist. For the colocation networks, the median is never zero. These differences for the clustering coefficients and triangle counts may reflect the greater density for the colocation networks, so it is also important to note the changes over time for the colocation networks are different from those in the conversation networks.

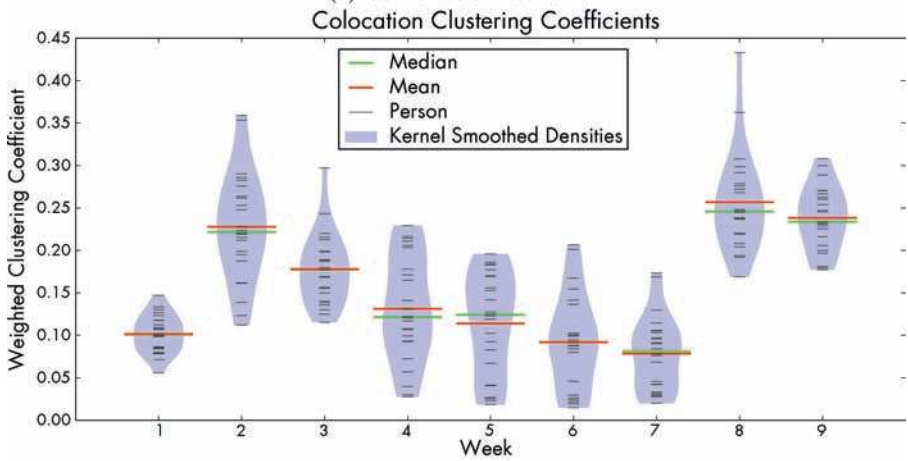
3.5.4 *Path Lengths.* A final property of the networks to consider is the distribution of path lengths. To compute path lengths, we define the length of edge (i, j) to be $1 - Y_{ij}$ if $Y_{ij} > 0$. In other words, the more time a pair spends interacting, the shorter the edge is. (If $Y_{ij} = 0$, then there is no edge between i and j and the length is undefined). The shortest path is found for all pairs and the distribution of path lengths are shown in Figure 17.

The conversation path lengths display a pronounced bimodality that corresponds to how many edges are involved in the path: values around 1 involve a single edge, values around 2 involve two edges, etc. This is unsurprising given the fact that most conversation edge values are small, as seen in Figure 13(a), and thus most edge lengths are approximately one. The maximum point at each time step is the diameter of the network. We can see that paths are generally short, usually involving at most one intermediary. Indeed, in all but the first week, a majority of the shortest paths involve only a single edge. This is unsurprising given the strong connectivity of the network seen in Figure 12.

Paths in the colocation networks are also short, but much more so than paths in the conversation networks. The shorter paths partly reflect the greater density of colocation networks, but again we see further qualitative differences between the two networks: The clustering of lengths around one does not reflect the same semi-discrete path lengths as in the conversation networks. There is much more variation in edge values in the colocation network (Figure 13(b)) so paths that traverse two edges can be as short as those that traverse only one, even as most paths are short.



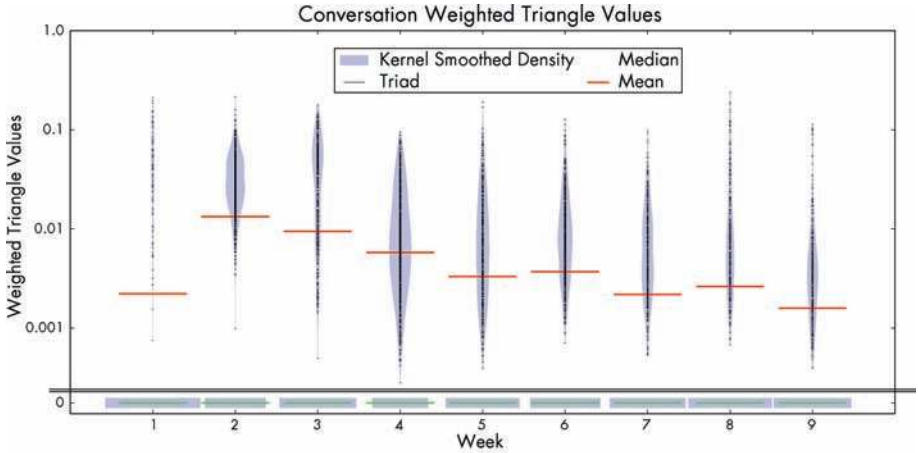
(a) Conversation networks.



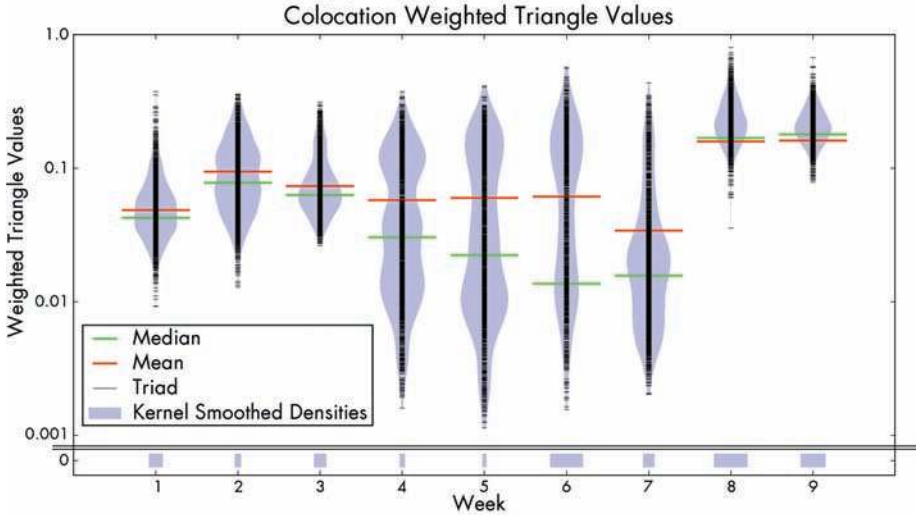
(b) Colocation networks.

Fig. 15. Weighted clustering coefficient distributions.

3.5.5 *Discussion.* All of the metrics above reveal that the colocation and conversation networks are very different. The colocation networks are denser, and show correspondingly higher transitivity and shorter path lengths. Those differences may be largely due to the simple phenomenon of shared classes. When many subjects attend the same class they are all colocated for a long period of time. This provides the opportunity for a single interaction event—the shared class—to create a large clique with heavily weighted edges in the network. Such large, strong cliques will naturally increase their members' degrees and clustering coefficients as well as the weighted triangle count of the entire network. Indeed, those three metrics are much higher for the colocation networks than the conversation networks.



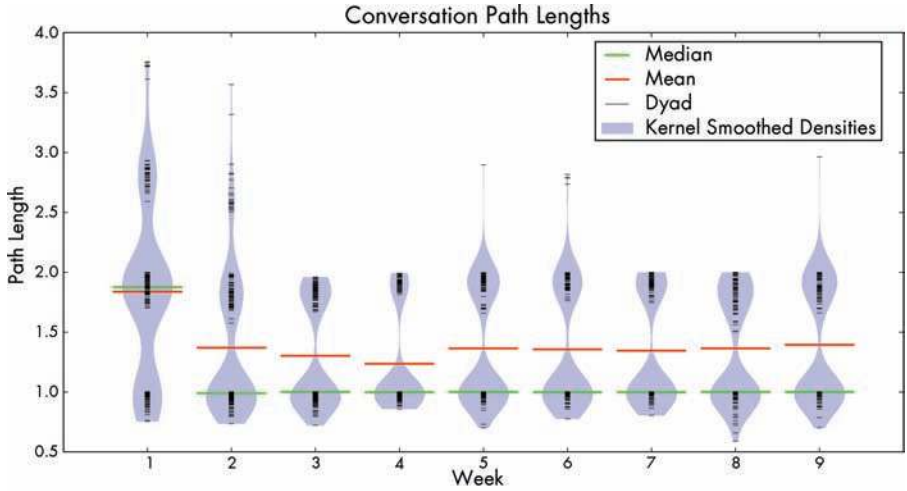
(a) Conversation networks.



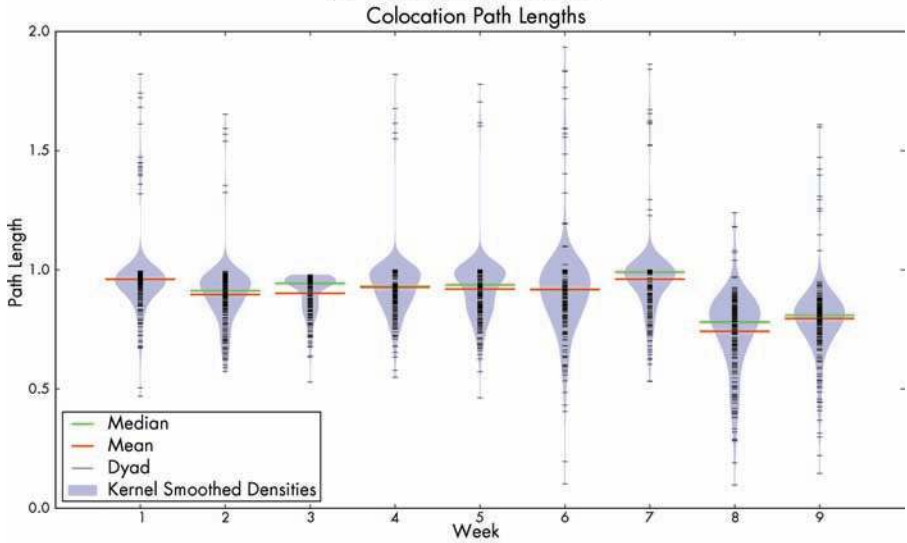
(b) Colocation networks.

Fig. 16. Weighted triangle value distributions.

Additionally, since all members of the cohort have offices in the same building, they have many opportunities to be physically proximate. That is reflected in the very different distributions of edge values for the conversation and colocation networks. The conversation networks have far more zero-valued edges, and lower non-zero values. The colocation networks have comparatively few zero-valued edges, suggesting that almost any subject is physically near most other subjects at least briefly during the week. Of course, subjects that share a class will have decidedly non-brief periods of time spent colocated. That difference may explain the bimodal colocation degree distributions of weeks 4 through 7 (Figure 13(b)), where there seems to be a distinction between pairs who spend much time together and pairs that only come together in passing.



(a) Conversation networks.



(b) Colocation networks.

Fig. 17. Path length distributions.

When examining the changes in degree distributions as time progresses, the conversation distributions seem to become more stable, while the colocation distributions continue changing. That is perhaps because some durable social network begins to form. The influence of that durable network on time spent in conversation may gradually become greater than the influence of external factors, such as time spent together in class. Time in class would certainly have a larger effect on the colocation distribution. The simple summary statistics presented here are not capable of distinguishing the relative importance of different factors on the network's evolution, but this remains a promising direction for future work within this research program.

Whether the colocation networks or the conversation networks are to be preferred also depends on the substantive research question considered. For example, colocation networks may be more relevant for the spread of the flu, while conversation networks are obviously more relevant for the spread of information through face-to-face speech.

Regardless of which is more important, the measures above reveal that the two networks are very different for this population, and that distinction that should inform future studies of real-world social networks, especially those that use colocation as a proxy for social interaction.

4. CONCLUSION

In this article, we have outlined a set of privacy-sensitive features that can be computed from incoming audio data in real-time. We have shown how to use those features to determine who was physically colocated with whom, both at the granularity of a room in a building and at the more elastic "acoustic proximity" needed to have a face-to-face conversation. We have used the privacy sensitive features to infer who was speaking when, and combined those inferences with colocation inference to determine who was in conversation with whom. This conversation detection can handle conversations with any number of participants, extending beyond previous methods that were limited to dyadic conversations only. We also demonstrated the performance of these methods using labeled conversation recordings in a variety of contexts.

We recounted an extensive project that collected privacy-sensitive situated speech data from a subject population of 24 graduate students, and applied our colocation and conversation detection methods to extract records of face-to-face conversations within the study cohort during the academic year. We constructed weighted networks of social behavior and examined basic descriptive statistics in order to compare social networks defined by colocation events to networks defined by conversation events. We found colocation and conversation networks to be quite different, providing new insight into earlier studies that had access to only colocation data, or that interpreted colocation records as an approximation of interpersonal interaction.

ACKNOWLEDGMENTS

We thank the editor and three reviewers for their helpful feedback and suggestions.

REFERENCES

AJMERA, J., LATHOUD, G., AND MCCOWAN, I. 2004. Clustering and segmenting speakers and their locations in meetings. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

ANG, J. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

ANGUERA, X. 2006. Robust speaker diarization for meetings. Ph.D. dissertation, Universitat Politècnica de Catalunya.

Basu, S. 2002. Conversational scene analysis. Ph.D. dissertation, Massachusetts Institute of Technology.

Basu, S. 2003. A linked-HMM model for robust voicing and speech detection. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Batliner, A., Fisher, K., Huber, R., Spilker, J., and Nöth, E. 2000. Desperately seeking emotions or: actors, wizards and human beings. In *Proceeding of the ISCA Tutorial and Research Workshop on Speech and Emotion*.

Baym, N., Zhang, Y. B., and Lin, M. C. 2004. Social interactions across media: Interpersonal communication on the internet, telephone and face-to-face. *New Media Society 6*, 299–318.

Bernard, H. R. and Killworth, P. D. 1977. Informant accuracy in social networks II. *Hum. Comm. Resear. 4,* 1, 3–18.

Bernard, H. R., Killworth, P. D., and Sailer, L. 1980. Informant accuracy in social network data IV: a comparison of clique-level structure in behavioral and cognitive network data. *Social Netw. 2,* 3, 191–218.

Bernard, H. R., Killworth, P. D., and Sailer, L. 1982. Informant accuracy in social network data V: An experimental attempt to predict actual communication from recall data. *Social Sci. Resear. 11*, 30–66.

Bilmes, J. 2004. On soft evidence in bayesian networks. Tech. rep. 16, Department of Electrical Engineering, University of Washingon.

Borovoy, R. 2002. Folk computing: Designing technology to support face-to-face community building. Ph.D. dissertation, MIT MediaLab.

Campbell, N. 2002. The recording of emotional speech: JST/CREST database research. In *Proceedings of the Annual Conference on Language Resources and Evaluation (LREC)*.

Choudhury, T. 2004. Sensing and modeling human networks. Ph.D. dissertation, MIT Media Lab.

Choudhury, T. and Pentland, A. S. 2003. Sensing and modeling human networks using the sociometer. In *Proceedings of the International Conference on Wearable Computing*.

Connolly, C. I., Burns, J. B., and Bui, H. H. 2008. Recovering social networks from massive track datasets. In *Proceedings of the IEEE Workshop on Applications of Computer Vision*.

Corman, S. R. and Scott, C. R. 1994. A synchronous digital signal processing method for detecting face-to-face organizational communication behavior. *Social Netw. 16,* 2, 163–179.

Davis, J. A. 1970. Clustering and hierarchy in interpersonal relations: Testing two graph theoretical models on 742 sociomatrices. *Amer. Socio. Rev. 35,* 5, 843–851.

Dellaert, F., Polzin, T., and Waibel, A. 1996. Recognizing emotion in speech. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

Dielmann, A. and Renals, S. 2004. Multi-stream segmentation of meetings. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing*.

Donovan, R. 1996. Trainable speech synthesis. Ph.D. dissertation, Cambridge University.

Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. 2003. Emotional speech: Towards a new generation of databases. *Speech Comm. 40*, 33–60.

Douglas-Cowie, E., Cowie, R., and Schroeder, M. 2000. A new emotion database: considerations, sources and scope. In *Proceedings of the ISCA Tutorial and Research Workshop on Speech and Emotion*.

Eagle, N. and Pentland, A. S. 2006. Reality mining: Sensing complex social systems. *Person. Ubiq. Comput. 10,* 4, 255–268.

Ferris, B., Haehnel, D., and Fox, D. 2006. Gaussian processes for signal strength-based location estimation. In *Proceedings of Robotics: Science and Systems*.

Freeman, L. 1992. Filling in the blanks: A theory of cognitive categories and the structure of social affiliation. *Social Psych. Quart. 55,* 2, 118–127.

Freeman, L., Romney, A. K., and Freeman, S. C. 1987. Cognitive structure and informant accuracy. *Amer. Anthropol. 89*, 311–325.

Gatica-Perez, D., McCowan, I., Zhang, D., and Bengio, S. 2005. Detecting group interest-level in meetings. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Goodreau, S. M., Kitts, J. A., and Morris, M. 2009. Birds of a feather or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography 46*, 103–125.

GRAY, R. M. AND DAVISSON, L. D. 2004. *An Introduction to Statistical Signal Processing*. Cambridge University Press.

GREASLEY, P., SETTER, J., WATERMAN, M., SHERRARD, C., ROACH, P., ARNFIELD, S., AND HORTON, D. 1995. Representation of prosodic and emotional features in a spoken language database. In *Proceedings of the International Congress of Phonetic Sciences*.

HAWKINS, K. 1991. Some consequences of deep interruption in task-oriented communication. *J. Lang. Social Psych. 10*, 185–203.

HOLLAND, P. W. AND LEINHARDT, S. 1975. The statistical analysis of local structure in social networks. In *Sociological Methodology*, Jossey-Bass, 1–45.

HURLBURT, R., KOCH, M., AND HEAVEY, C. 2002. Descriptive experience sampling demonstrates the connection of thinking to externally observable behavior. *Cogn. Therapy Res. 26,* 1, 117–134.

INGRAM, P. AND MORRIS, M. 2007. Do people mix at mixers? Structure, homophily, and the "life of the party". *Adminis. Sci. Quarter. 52,* 4, 558–585.

KAMPSTRA, P. 2008. Beanplot: A boxplot alternative for visual comparison of distributions. *J. Stat. Softw. 28,* 1, 1–9.

KILLWORTH, P. D. AND BERNARD, H. R. 1976. Informant accuracy in social network data. *Human Org. 35,* 3, 269–286.

KILLWORTH, P. D. AND BERNARD, H. R. 1979. Informant accuracy in social netw. data: III a comparison of triadic structure in behavioral and cognitive datasets. *Social Netw. 2*, 10–46.

KOSSINETS, G. AND WATTS, D. J. 2006. Empirical analysis of an evolving social network. *Science 311*, 88–90.

LAZEGA, E. AND VAN DUIJN, M. 1997. Position in formal structure, personal characteristics and choices of advisors in a law firm: A logistic regression model for dyadic network data. *Social Netw. 19*, 375–397.

LESKOVEC, J., LANG, K. J., DASGUPTA, A., AND MAHONEY, M. W. 2008. Statistical properties of community structure in large social and information networks. In *Proceedings of the International World Wide Web Conference (WWW)*.

LESTER, J., CHOUDHURY, T., KERN, N., BORRIELLO, G., AND HANNAFORD, B. 2005. A hybrid discriminative-generative approach for modeling human activities. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

LIAN, C. AND HSU, J. 2009. Probabilistic models for concurrent chatting activity recognition. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

MCCOWAN, I., BENGIO, S., GATICA-PEREZ, D., LATHOUD, G., MONAY, F., MOORE, D., WELLNER, P., AND BOURLARD, H. 2003. Modeling human interaction in meetings. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

NIST. 2009. NIST rich transcription evaluations. http://www.itl.nist.gov/iad/mig/tests/rt/2009/index.html.

ONNELA, J.-P., SARAMÄKI, J., HYVÖNEN, J., SZABÓ, G., DE MENEZES, M. A., KASKI, K., BARABÁSI, A.-L., AND KERTÉSZ, J. 2007. Analysis of a large-scale weighted network of one-to-one human communication. *New J. Physics 9*, 179.

PALLA, G., BARABÁSI, A.-L., AND VICSEK, T. 2006. Quantifying social group evolution. *Nature 446*, 664–667.

QUATIERI, T. 2001. *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall.

RABINER, L. 1977. On the use of autocorrelation analysis for pitch detection. *IEEE Trans. Acoustics, Speech, Sig. Process 25,* 1, 24–33.

RABINER, L. R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE 77,* 2, 257–286.

REYNOLDS, D. A. AND TORRES-CARRASQUILLO, P. 2005. Approaches and applications of audio diarization. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

SARAMÄKI, J., KIVELÄ, M., ONNELA, J.-P., KASKI, K., AND KERTÉSZ, J. 2007. Generalizations of the clustering coefficient to weighted complex networks. *Phys. Rev. E 75,* 027105, 1–4.

SCHULLER, B., RIGOLL, G., AND LANG, M. 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network

architecture. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

STUPAKOV, A., HANUSA, E., BILMES, J., AND FOX, D. 2009. COSINE—A corpus of multi-party conversational speech in noisy environments. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

WATTS, D. J. AND STROGATZ, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature 393*, 440–442.

WREN, C. R., IVANOV, Y. A., LEIGH, D., AND WESTHUES, J. 2007. The MERL motion detector dataset. Tech. Rep. 2007-069, MERL.

WYATT, D., CHOUDHURY, T., AND BILMES, J. 2007. Conversation detection and speaker segmentation in privacy-sensitive situated speech data. In *Proceedings of Interspeech*.

Received August 2010; revised October 2010; accepted October 2010