Social Networks xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Detecting interaction links in a collaborating group using manually annotated data

Shobhit Mathur[a,*], Marshall Scott Poole[c], Feniosky Peña-Mora[b], Mark Hasegawa-Johnson[d], Noshir Contractor[e]

[a] 255 Warren St. Apt 807, Jersey City, NJ 07302, United States
[b] School of Engineering & Applied Science, 510 Mudd, Columbia University, New York, NY 10027, United States
[c] Department of Communication, 1207 W. Oregon, University of Illinois Urbana-Champaign, Urbana, IL 61801, United States
[d] Department of Electrical and Computer Engineering, Beckman Institute 2011, University of Illinois Urbana-Champaign, Urbana, IL 61801, United States
[e] Department of Industrial Engineering & Management Sciences, 2145 Sheridan Road, Northwestern University, Evanston, IL 60208, United States

ARTICLE INFO

Keywords:
Social networks
Dynamic Bayesian Network
Collaborative groups

ABSTRACT

Identification of network linkages through direct observation of human interaction has long been a staple of network analysis. It is, however, time consuming and labor intensive when undertaken by human observers. This paper describes the development and validation of a two-stage methodology for automating the identification of network links from direct observation of groups in which members are free to move around a space. The initial manual annotation stage utilizes a web-based interface to support manual coding of physical location, posture, and gaze direction of group members from snapshots taken from video recordings of groups. The second stage uses the manually annotated data as input for machine learning to automate the inference of links among group members. The manual codings were treated as observed variables and the theory of turn taking in conversation was used to model temporal dependencies among interaction links, forming a Dynamic Bayesian Network (DBN). The DBN was modeled using the Bayes Net Toolkit and parameters were learned using Expectation Maximization (EM) algorithm. The Viterbi algorithm was adapted to perform the inference in DBN. The result is a time series of linkages for arbitrarily long segments that utilizes statistical distributions to estimate linkages. The validity of the method was assessed through comparing the accuracy of automatically detected links to manually identified links. Results show adequate validity and suggest routes for improvement of the method.

Network data come from a variety of sources, including surveys, email depositories, analysis of documents and archives, and direct observation. Direct observation of networks has a long history, stretching back to the original sociometric studies (Moreno, 1951), the Bank Wiring Room studies conducted by Hawthorne researchers and analyzed by Homans (1951), and early anthropological work (e.g., Kapferer, 1969).

In recent years, gathering of network data through direct observation is less common than collecting data via surveys, reconstructing links from archival or media documents, and analysis of digital data on connections. Direct observation is extremely time and resource-intensive, and the expense becomes almost prohibitive if we want to study network dynamics over time.

However, there are compelling reasons for using direct observation to study networks. It offers a useful complement to self-reports of ties, especially if investigators can collect both types of data. If a permanent record of the observation can be made, for example by video or audio recording the observations, then additional facets of meaning can be adduced and used to supplement the network data. For example, semantic networks derived from transcriptions of interaction can be related to social or communication networks.

One important barrier to direct observation of networks is the time and effort it takes. For even small networks, coding links observationally requires multiple coders to watch different subsets of subjects in real time. For larger networks, such as networks of emergency responders who may number in the hundreds, the task becomes truly formidable and forbidding.

One way to reduce the time and effort required for direct observation of networks is to automate the process of link detection as much as possible. This manuscript reports on the development of an algorithm for detection of network links from video data that is amenable to automation. The algorithm utilizes visual cues which current automated computing systems can detect. It

* Corresponding author.
E-mail addresses: shomat@gmail.com (S. Mathur), mspoole@illinois.edu (M.S. Poole), feniosky@columbia.edu (F. Peña-Mora), jhasegaw@illinois.edu (M. Hasegawa-Johnson), nosh@northwestern.edu (N. Contractor).

0378-8733/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.socnet.2012.04.002

employs machine learning to train the system to recognize links based on a set of nonverbal cues. Machine learning enables the system to continuously improve its ability to recognize links. The algorithm promises to greatly improve the process of identifying network linkages through direct observation.

We will define a link to be formed any time one person in the network is communicating with another. The link exists so long as the two are communicating and is broken when they turn their attention elsewhere. Hence links in the network under observation are dynamic and change as members shift their attention to different members. The resulting network is dynamic, which is consistent with the nature of actual communication networks, and will generate data that can be modeled over time.

This report will first review previous work on automated link detection. Following this we will present a brief overview of the strategy we followed in developing and validating the algorithm. This will be followed by a step-by-step narrative of the development process. Then we discuss validation of the method. Finally, we consider the implications of the method and future steps.

1. Background

At least two methods for computational identification of interaction links have been developed. Choudhury and Pentland (2002) created a sensor device called Sociometer that could measure face-to-face interactions between people. Sociometer consisted of an IR Tranceiver, microphones, and accelerometers along with a power supply and storage. The device was capable of picking up face to face interactions within a range of approximately 6 ft. Subjects wore the device over a period of time during which their behavior such as proximity to other people, tonal variation and prosody and interaction time was monitored. Using the data collected and statistical pattern recognition techniques computational models of the group interactions were developed. These models detect conversation between individuals using the fact that speech pattern in conversations is highly synchronized.

Otsuka et al. (2006) were able to determine conversation structures such as whether the conversation was a monologue or dialogue and who was listening to whom using the gaze directions and utterance behavior of the participants. The study targeted small groups in closed environments, and direction of gaze of participants was determined from head direction tracked using a visual head tracker. Using gaze and utterance observations a Dynamic Bayesian Network (DBN) based algorithm was developed that determined the conversation structure of a meeting over time. The temporal dependencies of the DBN represented the dynamics of conversation such as turn taking.

The present effort differs from these two previous systems in that it does not rely on specialized devices such as Sociometer or head tracking hardware. Instead, it is based solely on analysis of video recordings. As such it should prove to be less intrusive and will require much less calibration of data gathering tools than will the specialized hardware employed in the Choudhury and Pentland and Otsuka et al. studies. Moreover the algorithm described in this work was successfully applied on data collected from a real world setting whereas most of the previous works have used laboratory setting as a testbed for their algorithms.

2. General strategy

The basic strategy for link detection was to employ multiple cues that reflect interaction as indicators of links. Utilization of multiple cues is more likely to yield valid and reliable values than single cues. The cues were also selected based on their ability to be automated using computer vision algorithms, since the ultimate goal of this effort is to provide a foundation for automated link identification. In this particular case we utilized three cues: location, gaze direction, and bodily posture.

Machine learning was then used to induce linkages from the three types of cues. A Dynamic Bayesian Network (DBN) approach associated the cues to a criterion variable indicating whether two individuals were linked.

The cues and the criterion variable were coded manually by 12 human coders from videotapes of an eight person group engaged in a planning task. These codes served as the input for the algorithm. Codes were assigned based on pooled aggregations of the codings of the 12 coders.

The derived DBN algorithm was then validated against a new set of coded cues and criteria.

3. Data collection and manual annotation

The basic data for this research were generated by groups of emergency responders planning a response to a simulated emergency situation. The Illinois Fire Service Institute (IFSI) regularly holds disaster simulation training exercises to introduce the firefighters to the National Incident Management System, which is a planning and decision making process designed to streamline the response to disasters. Generally, the groups involved consist of 8–10 incident command personnel, who are charged with planning their strategy to a disaster.

The IFSI simulations are held at different venues around the state of Illinois. The venue is usually a room with large rectangular center table and different charts and figures on the walls. Participants move around the room during the simulation and form different subgroups over time. The participants of the meeting were free to move about and enter and exit the room.

Video was captured using four digital video cameras placed at the corners of the room, providing a panoramic view of the session. Each participant in these events wore a high quality audio recorder, which yielded recordings of his/her deliberations and was later used to produce the transcripts of the verbal interaction. Participants wore color-coded vests so that each group member could be tracked throughout the recording. The computer vision tools likely to be used in automating cue collection key on different participants based on the color of outfit they wear (Ramanan et al., 2007).

The disaster scenarios provided to the emergency response team included leaking hazardous materials (HAZMAT), terrorist attacks and other events that require immediate response. Each emergency responder was given a specific role and they were required to follow specific procedures in coming up with an action plan. For this analysis we used a single IFSI session with 8 participants that lasted for about 2.5 h. The meeting room dimensions were approximately 490 cm × 460 cm. The central table occupied 183 cm × 153 cm of the floor space. Fig. 1 shows a snapshot of the room and numbering of participants used in this paper.

3.1. Manual data annotation

Since our prime motive for data extraction is to infer the interaction links among members, we focused on four types of verbal and nonverbal cues that have been shown to be good indicators of interaction or communication between two individuals:

- Direction of gaze: By observing what a person's eye is fixed on, one can guess the object of attention of the person. Previous studies have shown that eye gaze is one of the reliable indicators of what a person is "thinking about" (Henderson and Ferreira, 2004).

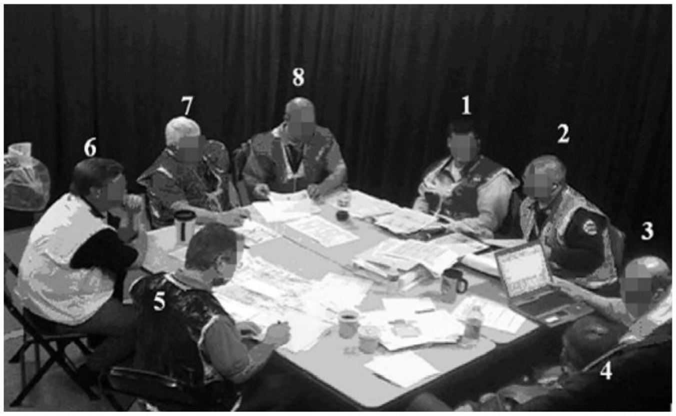

Fig. 1. Location of members.

Direction of gaze carries information about the focus of the user's attention (Prasov and Chai, 2008). It is estimated that 60% of conversations involve gaze and 30% mutual gaze (Sellen, 1992).

- Conversational distance: People tend to maintain the minimum normative interpersonal distance during conversations. Often the proximity between two members can be used as a cue in order to determine if the two are interacting. Interpersonal space maintained during conversation/interaction is often categorized into four regions:
  - (a) intimate (<1.5 ft),
  - (b) casual–personal (1.5–4 ft),
  - (c) social consultative (4–12 ft),
  - d) public (>12 ft).

  Generally the casual–personal region is correlated with conversation, but the social consultative region may be used for lecturing or addressing a group. However there are various factors that determine comfortable conversational distance including gender, age, cultural background and environment (Knapp and Hall, 2002).
- Body posture: Various parameters such as lean and orientation of the head indicate a person's focus of attention. Similarly during conversations various gestures are made by the speaker or listener indicating a communication channel between them. Touch, facial expressions, etc. are other important nonverbal cues.
- Vocalics and verbal cues: A member speaking indicates that he/she is communicating to someone. Often the tone, pitch or loudness of a persons' voice can give some indication of his/her linking behavior. Similarly verbal cues such as someone mentioning a person's name, someone answering question etc. can be used to determine who is interacting with whom.

3.2. Interface design

In order to extract relevant data from the videos a manual coding scheme was developed. A web-based interface was designed using Flash software with which volunteers were able to code the location, orientation, and bodily posture of individual group members, as well as the criterion variable, whether communication links formed between them. Snapshots of the videos were taken at 10 s intervals. The interval was chosen so as not to be so short that it would give redundant snapshots. Research has shown that in a typical conversation a speaker holds the floor for 6 s on average (Burgoon et al., 1996). Fig. 2 shows the interface.

The interface allowed users to code the following information:

- Location: The room was divided into a 6 × 6 grid (with four central rectangles occupied by the table) and coders were asked to represent the location in grid for each member. The location information was used to infer the conversation distances between the members.
- Head orientation: In order to record the direction of gaze of each member, the coders were asked to code the orientation of each members' head. Stiefelhagen and Zhu (2002) showed that head orientation estimates can be of great help in deducing focus of a person's gaze. The interface allowed the users to rotate an icon using mouse so as to match the head direction observable in the

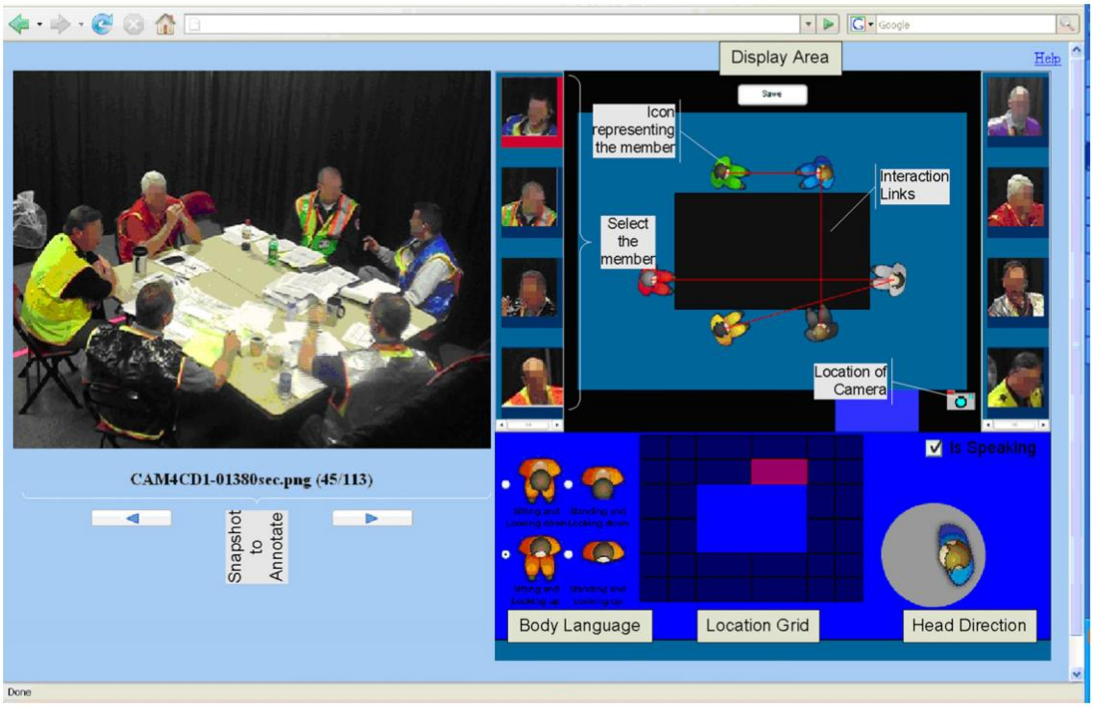

Fig. 2. Manual annotation interface.

photo. This resulted in a measure of the angle that the subject's head was oriented relative to the center of the room.

- Bodily posture: The interface was used to record two important components of posture: whether a person was sitting or standing, and whether a person was looking up or down. The first component was indicative of the height of gaze, while the latter was used as a supplemental cue in determining whether the member was gazing at another member or elsewhere. The two attributes were combined to form four icons and coders were asked to transcribe the most appropriate icon representing each member:
  – sitting and looking ahead,
  – sitting and looking down,
  – standing and looking ahead,
  – standing and looking down.
- Speaking (Y/N): In order to minimize coding time, we experimented with a system wherein coders labeled each member as "speaking" versus "not speaking" based only on still images without the aid of audio or transcripts. The coders could click on the checkbox if they believed that a member was speaking. The accuracy of this parameter was later determined by comparing with actual video to determine the effectiveness of just using photos in interpreting who was speaking. We consider this experiment useful, in part, because previously published work has not reported the detectability of locution from still images. As it turned out, intercoder agreement in this task was unacceptably low (see Section 3.4), and therefore automatic link detection algorithms were also tested using a speaker/nonspeaker label coded directly from the video (see Section 4.3).
- Links: To identify the criterion variable, whether two members are linked, an icon representing the member was displayed in the display area. The coders could then link two members who are interacting with each other by clicking on their icons.

The web based interface was designed to be easy to use. For each parameter the choices were kept at a minimum to speed the annotation task and achieve high agreement among coders. The coded members were displayed in the display area so that coders could compare their annotation with the photo. The snapshots were listed in random order (not chronological) so as to avoid repetition of similar snapshots and the bias this might introduce. A series of 113 photos were each coded by 12 student coders. We pooled the codes of twelve coders in order to maximize reliability and validity of the data, as discussed in the next section.

3.3. *Pooling of judgment*

Even though the annotation task is simple and straightforward, there still exists the possibility of error. To minimize error and generate the most valid codes possible, we pooled coder judgments into a composite code. In this section we describe how the results were pooled for different kinds of data and the effect of number of coders and individual accuracy on the composite result. Ideally we want to select the best coded value and use that as the composite judgment. However this is not possible without a priori knowledge. We will discuss pooled judgment for different types of data employed in this study:

- Continuous variables: The head orientation ratings were continuous (angle of orientation). Einhorn et al. (1977) described different models that can be used as yardsticks for evaluating the quality of a group judgment in determining a continuous quantitative measure. A group for example may randomly pick a single individual's judgment or use group mean as the group judgment. If ground truth were known, coders could identify the best individual judgment and use that value as the composite group judgment (referred to as the "best model"). For our study we assumed a homogenous group consisting of individuals whose judgments were distributed normally with mean x_n and variance σ_n^2, where x_n was also the true value to be predicted. Under this assumption, (Einhorn et al., 1977) showed that if for a group of size N, the mean of individual judgments is used as a measure of composite group judgment, the difference between true value to be predicted, x_n, and average of individual judgments, $\overline{X_N}$, is given by:

$$d' = \left| \frac{\overline{X}_N - x_n}{\sigma_n} \right| \tag{1}$$

$$E(d') = \frac{\sqrt{2}}{\sqrt{N\pi}} \tag{2}$$

$$\sigma^2(d') = \frac{1}{N} - (E(d'))^2 \tag{3}$$

where $E(d')$ is the expected value of d' and $\sigma^2(d')$ is the variance of d'. Using the above result we can evaluate the accuracy of an N-coder mean and accordingly control the number of coders required. Results in Einhorn et al. (1977) show that for 12 and more coders, the average of individual judgments gives a measure that is close to the accuracy of the most-accurate individual coder. For this reason we used 12 coders in our study.

- Binary variables: The speaking and linkage variables were binary. For binary variables the category coded by the majority can be used as the final result representing the composite judgment. If p is the average group competence in correctly coding a binary variable, Condorcet's Jury theorem states that for $p > 0.5$ the probability of getting correct judgment by majority vote for a group of size N, P_N, is monotonically increasing in N and $\lim_{N\to\infty} P_N \to 1$ (Grofman et al., 1984). More importantly the effect of increasing the number of coders can be given by the following recursion formula (assuming homogenous group with individuals having accuracy p)

$$\Delta P_{N+1} = P_{N+2} - P_N \cong \frac{2^{N+2}(p-(1/2))e^{-2N(p-(1/2))^2}}{\sqrt{\pi N}} \tag{4}$$

The above formula gives the effect of adding two coders to the overall accuracy of pooled values and is derived for odd N. Lam and Suen (1997) have extended the results to cases when N is even. This formula can then be utilized to find N for a required accuracy (of the composite measure) and known p. As apparent from the above formula, for $p > 0.5$, the benefit of an additional coder (the improvement in P_N per increment of N) decreases rapidly with N, whereas for $p < 0.5$ adding coders decreases the overall accuracy. For example, for $N \geq 11$ and $p \geq 0.8$ the improvement in P_N per additional coder is less than 0.01. The value of the parameter p can be found by benchmark tests; in the absence of any benchmark ground truth measure, the maximum likelihood estimate of parameter p is the proportion of all coded datum pairs on which two transcribers agree. We determined the number of coders per video by assuming that individual accuracy would be $p \geq 0.8$, which reflects generally accepted standards for coding reliability (Neuendorf, 2001), and thus found 12 annotators to be sufficient. Annotation results discussed in Section 3.4 suggest that the assumption $p \geq 0.8$ was met for link detection (for which the proportion of inter-transcriber agreements was 0.93 and transcription error rates were less than 20%), but not for identification of the speaker; this point is further discussed in Section 3.4.
- Nominal (polytomous) variables: The bodily posture and location variables were polytomous. Some polyotomous data, such as bodily language icons in our study, can be analyzed as a sequence of dichotomous choices. These icons were selected from a palette of four icons but the choice can be divided into whether a person is sitting/standing and whether a person is looking up/down. In

Please cite this article in press as: Mathur, S., et al., Detecting interaction links in a collaborating group using manually annotated data. Soc. Netw. (2012), http://dx.doi.org/10.1016/j.socnet.2012.04.002

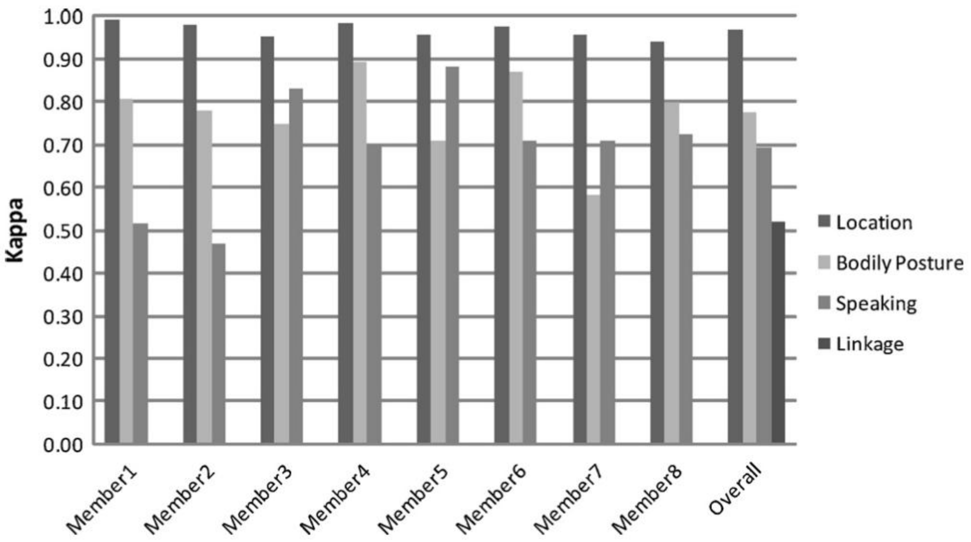

Fig. 3. Fleiss kappa for nominal data.

these cases results described for binary data can be used. For data that cannot be modeled using dichotomous choices, we can use plurality voting in order to pool the individual judgments. Plurality voting chooses the option that has received the maximum number of votes from among the coders. The number of votes need not be a majority. Lin et al. (2003) showed that under certain assumptions plurality voting is equivalent to minimum probability of error coding. No closed form solution exists to determine the accuracy of plurality voting as a function of the number of coders, however Lin et al. (2003) suggest using stochastic simulations in order to find the accuracy of pooled judgment.

To summarize, our analysis of judgmental accuracy suggested that 12 coders were sufficient to yield an accurate judgment. In the case of continuous variables we took the average of the twelve codes, for binary variables we took that category which had a majority of votes (discarding any data when there was a 6–6 split), and for polytomous variables we took that category that had a plurality of the votes.

3.4. Annotation results: reliability, errors and speed of coding

The reliability of coded data is described by its “dependability,” “reproducibility” and “consistency” (Neuendorf, 2001). In order to assess the reliability of the coded data we need statistics that measure the level of agreement or correspondence between different coder’s assessments. It is important that the annotated variables are reliable before they can be used for the link detection algorithm or any other analysis. We had four nominal variables (bodily posture, location, speaking, linked) and one continuous variable (head orientation, as indicated by the angle of the member’s gaze) annotated by the coders. This section discusses the reliability of the variables obtained using various statistical measures.

One popular measure for intercoder reliability of nominal variables is the Kappa family of statistics, that measures agreement beyond chance. It is calculated as

$$\kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}} \qquad (5)$$

where $\overline{P}$ is “proportion agreement” and $\overline{P_e}$ is “proportion agreement, expected by chance.” In the preceding formula the numerator is the amount by which actual agreement exceeds chance while the denominator is the maximum possible agreement over chance attainable. For our study we used Fleiss’ Kappa as it can handle multiple coders (Fleiss, 1971). Landis and Koch (1977) have defined kappa value between 0.61 and 0.80 as “substantial” and between 0.81 and 1.00 as “almost perfect.”

Since camera recordings and resulting snapshots are taken from a fixed camera angle, the degree of visibility of different members varies and as a result amount of agreement is different for different members. Fleiss’ Kappa was calculated for each nominal parameter for each member. Fig. 3 shows the Fleiss Kappa obtained for different parameters for members with the locations shown in Fig. 1. Although members move about in the room, Fig. 1 gives the locations where members are present in a majority of the snapshots.

The location coding was highly reliable (κ>0.9). Similarly a high overall reliability (κ = 0.77) was obtained for bodily posture. For the bodily posture measure, confusions were mostly between looking up or down. For the head orientation codings (continuous data) Cronbach’s Alpha was calculated using the reliability module of SPSS. The value of this measure was found to be 0.857 which is considered high (Garson, 2008). Thus, the location, head orientation and bodily posture measures showed substantial agreement.

Two measures were problematic. First the kappa for speaking was less than 0.7 and it was lower than 0.5 for members 1 and 2. These members were talking with each other in a majority of the snapshots. Coders annotated only one of them as speaking, but there was confusion and disagreement about which one of the two was speaking. Second the κ for links identified by coders was only 0.52. Although the overall proportion of agreement on links was high (0.93), since most of the members were coded as not linked with each other, agreement by chance was also high (0.85).

In order to make some determination of the validity of the data on speaking and links we compared the coders’ values, annotated using still images, to secondary annotations derived from a transcript of the video. These were regarded as criterion variables for determining the validity of the coders’ decisions. It proved difficult for coders to determine if a person is speaking just from the snapshot, but nonetheless their accuracy was still more than 60% compared to the videos and transcripts. Coders reported using cues such as hand gestures and facial expressions in order to determine if a person was speaking.

The coders were asked to identify links between two members who were verbally communicating with each other. To determine the validity of the linkages found by the coders, we identified the speakers from the codes and from the videos and transcripts. We calculated the proportion of cases in which coded links were not

Please cite this article in press as: Mathur, S., et al., Detecting interaction links in a collaborating group using manually annotated data. Soc. Netw. (2012), http://dx.doi.org/10.1016/j.socnet.2012.04.002

Table 1
Frequency of easily detectable types of transcription error. Error = 100 × (number of cases in error)/(total number of cases).

Statistics	Error (%)
Speakers coded as non-speakers	35.8
Non-speakers coded as speakers	23.7
Cases where coder showed a link between two speakers or two non-speakers[a]	12.6
Cases where coder showed a link between two speakers or two non-speakers[b]	15.2
Cases where coder missed a link for a speaker[a]	19.4
Cases where coder missed a link for a speaker[b]	4.3

[a] Speaker was identified from video/transcript.
[b] Speaker was identified from the manual coding using the interface.

consistent with the assumption that links are formed between a speaker and non-speaker (the links were definitely incorrect) and the proportion of cases in which a speaker was not linked to anyone at all (and thus the link was missed). Speakers were identified in two ways: from coders' annotations, or from videos and transcripts. There were possible inconsistencies for both types of speaker annotations. These error proportions are reported in Table 1. Note that the error statistics were calculated over the pooled data.

The results show that the coders were not particularly accurate in identification of speakers compared to the transcript and video. For this reason both coded speakers and speakers drawn from the video and transcript were used in the validation analysis.

Overall, using the web based coding interface resulted in two important gains. Firstly, it helped in accomplishing higher precision than what nonaided coders would be capable of. Secondly, for each picture, coding five attributes and linkage per member took an average of 73 s (standard deviation of 22 s), which is relatively fast for the amount of data generated and significantly sped up the process compared to what unstructured coding would permit.

4. Automatic annotation

Each snapshot taken from the video contains all or a subset of group members. For each member in each snapshot, we obtained data by coding the previously defined *attributes*: location, head orientation, bodily posture (sitting/standing crossed with looking up/looking down), and speaking. These data can then be used to infer direction of gaze, conversational distance, and whether two members are speaking or not, the higher-order variables that will be used to induce network linkages. The three higher-order variables will be referred to as *connectors*, because they can be used to infer network links. In the following sections we discuss how the annotated attributes were used to infer connectors and then how the connectors were combined in the link detection algorithm.

4.1. Gaze

It is challenging to infer the direction of gaze from unobtrusive observational data such as that obtained from video recordings. This section describes an algorithm to determine the subject of a member's gaze from his/her head orientation. The algorithm described here is based on one described by Stiefelhagen et al. (2001). They developed a method that utilized Gaussian mixtures to model the relationship between gaze and head orientation. The distribution of a member's head orientation given that he/she is gazing at a fixed subject, can be modeled as a Gaussian (normal) curve similar to the one shown in Fig. 4. Thus, in Stiefelhagen's method, a person's head orientation distribution over time can be modeled as a mixture of Gaussian curves where each curve corresponds to the person gazing at a particular subject.

In our group, however, members were moving about the room. Therefore we modeled head orientation using a mixture Gaussian, like Stiefelhagen, but instead of fixing the normal curves to a particular member, we fixed them to a particular location. Consider a matrix of Gaussian curves where the curve belonging to cell (i, j) corresponds to observable head orientation distribution over time for members at location i gazing at members at location j, then the head orientation distribution observed for members located at a particular location i is a mixture of Gaussian curves in the row i of this matrix. In equations below, μ_{ij} is the mean and σ_{ij} is the standard deviation of the normal curve in cell (i, j) of the Gaussian curve matrix, $P(\theta_i)$ is the distribution of head orientation angles (θ) observed at a location i and $gaze_i(j)$ refers to a binary variable indicating that the member at location i is gazing at another member who is at location j.

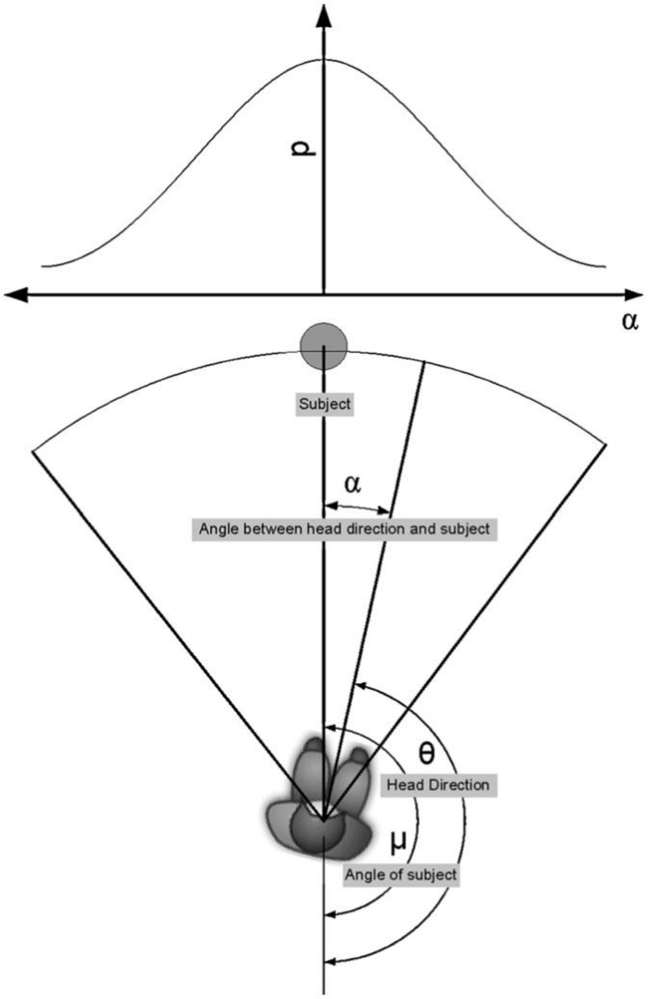

Fig. 4. Head direction distribution for a gaze subject.

$$P(\theta_i|gaze_i(j)) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-(\theta_i-\mu_{ij})^2/2\sigma_{ij}^2} \tag{6}$$

$$P(\theta_i) = \sum_{j=1}^{M} P(\theta_i|gaze_i(j))P(gaze_i(j)) \tag{7}$$

In addition to using a Gaussian mixture to model head orientation distribution we also used the concept of view volume as defined in Pan et al. (2007). A person can only see objects that fall within his/her view volume which is basically a cone of vision. The vision cone of a person is 30° either side of the direction of his head (Zmijewski, 2004), for a total cone of 60°. A person may be gazing at any object within this vision cone (Fig. 5). In order to account for

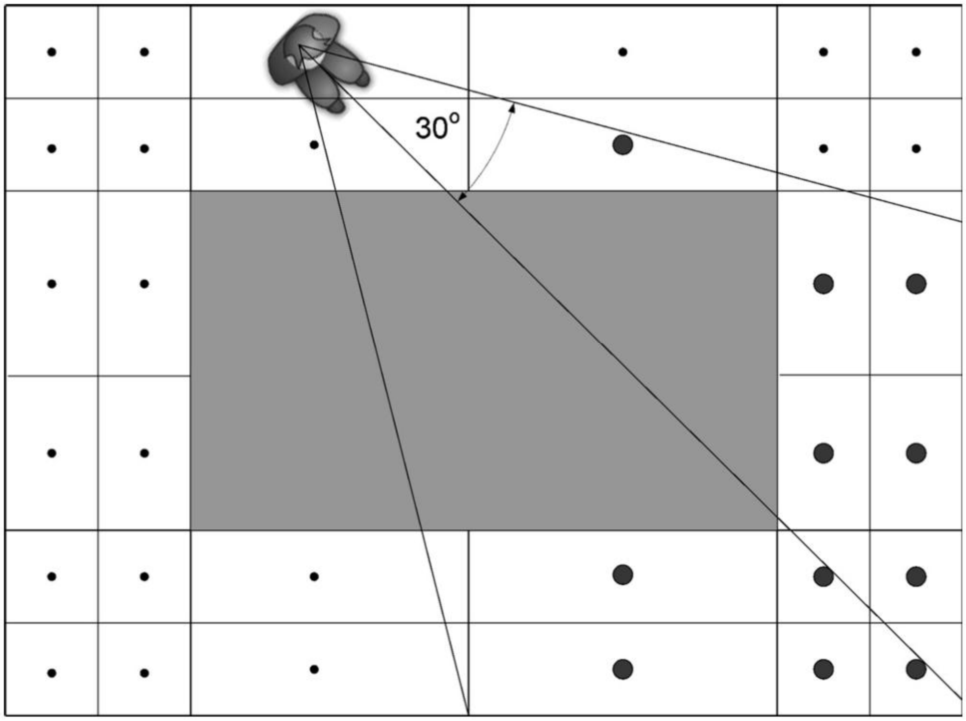

Fig. 5. Vision cone.

possible errors in coding head directions, the confidence interval of head direction angle was determined and vision cone adjusted to 30° on either side of confidence interval. Using the vision cone concept we can restrict the number of candidates that a member might be looking at. We also leverage the "looking up/down" attribute. If a member has been coded as looking down, we assume he/she is not gazing at any other member. If the member is looking up, we use Bayes' law in order to determine the most likely candidate of the member's gaze (out of those selected from her/his vision cone). The equation below gives Bayes' formula:

$$P(gaze_i(j)|\theta_i) = \frac{P(\theta_i|gaze_i(j))P(gaze_i(j))}{P(\theta_i)} \tag{8}$$

The member at location *j* with highest posterior probability found using the above formula is assumed to be the subject of member *i*'s gaze.

In order to determine the parameters of the Gaussian mixtures, the Expectation Maximization (EM) algorithm is used. The EM algorithm maximizes the likelihood of the head orientation distribution given the mixture model (Stiefelhagen and Zhu, 2002). The EM algorithm works iteratively, and the following update equations are used to readjust the mixture parameters.

$$\mu_{ij}^{new} = \frac{\sum_n P^{old}(gaze_i(j)|\theta_i^n)\theta_i^n}{\sum_n P^{old}(gaze_i(j)|\theta_i^n)} \tag{9}$$

$$(\sigma_{ij}^2)^{new} = \frac{\sum_n P^{old}(gaze_i(j)|\theta_i^n)(\theta_i^n - \mu_{ij}^{new})^2}{\sum_n P^{old}(gaze_i(j)|\theta_i^n)} \tag{10}$$

$$P^{new}(gaze_i(j)) = \frac{1}{N}\sum_n P^{old}(gaze_i(j)|\theta_i^n) \tag{11}$$

Further details of the implementation can be found in Mathur (2008). The algorithm used for inferring gaze direction is summarized as follows.

- Estimate the parameters of the mixture model using the EM algorithm.
- If the member is looking down he/she is not gazing at anyone.
- Else using vision cone obtain possible gaze candidates.
- If there are no gaze candidates within the vision cone, then member is not gazing at anyone, else find the most probable candidate using Bayes' law.

4.2. Conversational distance: constraints

The second important connector used in the link detection algorithm is conversation distance. As defined previously, we consider two individuals with 1.5–4 ft distance between them to be within each others' casual–personal space. We define two such individuals as neighbors. Our algorithm assumes that neighbors have a high probability of sharing a conversational link. Using the manually annotated data we can infer two members to be neighbors if they are located in adjacent rectangles of the location grid. Fig. 6 shows an example of how neighbors are inferred from the grid. All the adjacent grid squares are separated by roughly 4 ft each.

4.3. Speaking

Whether a member is speaking is the third cue used in the link detection algorithm. As described previously, coders were asked to annotate who was speaking from the photos. The reliability of this statistic was moderate ($\kappa < 0.7$), not sufficiently high to merit complete confidence. As noted previously, we also supplemented the speaker codes with speaking behavior directly annotated from videos and a transcript. The members who were speaking (as observable from the video) when the snapshot was taken, were coded as speaking. The algorithm was run on both sets of variables, annotated and manually derived from videos.

4.4. Link detection: constraints

The purpose of our algorithm is to detect interaction links formed between members. As noted at the outset, we define a link as occurring when two members interact via verbal

Please cite this article in press as: Mathur, S., et al., Detecting interaction links in a collaborating group using manually annotated data. Soc. Netw. (2012), http://dx.doi.org/10.1016/j.socnet.2012.04.002

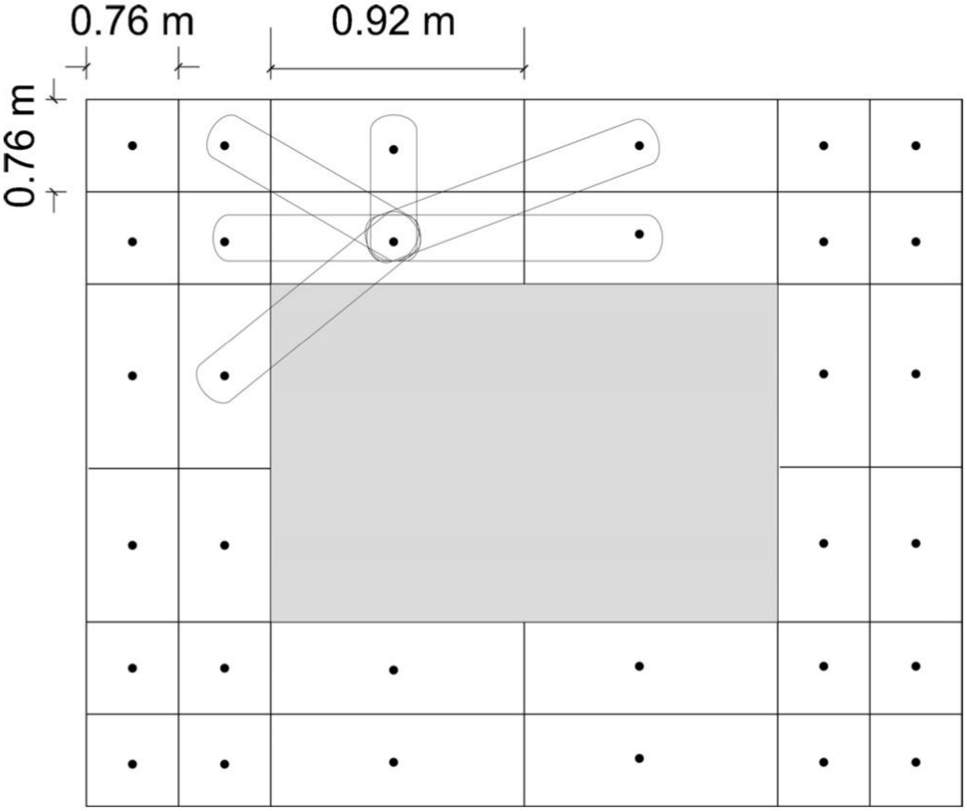

Fig. 6. Determining neighbors.

communication. The most typical mode of verbal communication is conversation. Burgoon et al. (1996) defined conversation as a series of opportunities to speak and listen. Conversation is composed of turns when a speaker has sole possession of the floor. Each turn is composed of one vocalization segment where the speaker holds the floor without any pauses longer than 200 ms (Feldstein and Welkowitz, 1978). Any conversation thus shows a process of turn taking in talking. Based on this model, the following assumptions are made for the link detection algorithm:

- By definition, conversation happens between a speaker and listener(s). A member who is speaking is always linked to at least one other member.
- A non speaker may or may not be linked to anyone at all.
- Since we are only interested in verbal communication, a non speaker can only be linked to a speaker.
- A non-speaker is linked to only one speaker (if any). Thus we assume a member can only be listening to one speaker at a time.
- If x and y are linked then at any given time, flow of information is unidirectional, i.e., either x or y is speaking

In addition to the conversation, other important forms of verbal communication include one-to-many (speech) and many to many communication. However, since these communication modes can be modeled as conversations, for our analysis we have not distinguished conversation from other types of verbal communication.

4.5. Dealing with simultaneous speech

Although it was assumed that conversations involve only one person speaking at a time, simultaneous speech does occur during a conversation. Simultaneous speech can occur in two forms (Burgoon et al., 1996).

- Noninterruptive simultaneous speech is a result of listener's responses (also called back-channeling). It starts and ends while the speaker continues to hold the floor. Backchannel verbal responses such as "uh-huh," "mm-hmm" and responses such as "yep," "right," "I see," etc. are examples of these (Burgoon et al., 1996).
- Interruptive simultaneous speech results in the speaker yielding the floor to the interrupter.

Since the snapshots were taken every 10 s, the probability of simultaneous speech occurring in any given snapshot was low, and hence it was disallowed by our algorithm. Noninterruptive simultaneous speech was avoided by ignoring short utterances when annotating who is speaking. During interruptive simultaneous speech, the interrupter's speech was ignored until the first speaker yielded the floor.

Similar to simultaneous speech, pauses and periods of silence exist between turns. These pauses are usually of short duration. Studies show that pauses longer than 0.3 s are noticeable in a conversation, and that a pause of 1.5 s is considered a very long pause in a conversation (Siegel and Wennerstrom, 2003). Since these pauses are a part of conversation they were incorporated into the speaker's speaking turn.

4.6. Link detection: a probabilistic formulation

Each snapshot consists of a set of links indicating the network at that particular point in time. This snapshot network will be referred to as its state. At any given point in time various states are possible, each of which constitutes a different configuration of links. Without any prior knowledge, the number of possible states for a snapshot having n members is $2^{\binom{n}{2}}$. However if we consider the constraints specified in Section 4.4, the number of possible states for a snapshot having s speakers and r non-speakers can be shown to be (Mathur, 2008):

$$\sum_{i=0}^{s}(-1)^i \binom{s}{i} (s+1-i)^r, \quad r \geq s,\ r+s=n \tag{12}$$

With eight members (the size of the group in this study), the number of valid states is always less than one thousand, which is quite tractable. For each snapshot we have deduced who is speaking, who is gazing at whom, and who is a neighbor of whom. Our task is to determine the best valid state for a snapshot given the observations.

As mentioned previously, verbal communications are modeled as conversations. Conversations show turn-taking behavior, and in order to model this behavior, we have assumed that a first order Markov property holds for links, namely that at any given time t, links are stochastically dependent on the state of the previous time $t-1$. Links are formed with high probability among members who belong to the same conversation group (subnetwork). We define two non-speaking members as indirectly linked to each other if they are in the same conversation group and thus listening to the same speaker. Thus whether two members are linked at time t is stochastically dependent on whether they were directly/indirectly linked in time $t-1$. The gaze and neighborhood observations are modeled as stochastically dependent on whether two members are linked. This model constitutes a Dynamic Bayesian Network. A Dynamic Bayesian Network is a graphical representation of stochastic dependency among time-indexed random variables (Murphy, 1998). In our DBN representation, links (L) are the hidden variables and the gaze (G), speaker (S), and neighborhood (N) are the observations.

Dynamic Bayesian Networks based on Hidden Markov Models (HMMs) have been utilized in a number of other studies of groups and meetings. McCowan et al. (2003) utilized several Hidden Markov Models (HMMs) to model and infer group actions in meetings. Brdiczka et al. (2005) represented synchronized turn taking in conversation using HMM. The authors used speech activity as the observed variable and interacting groups (the configuration of each subgroup in the meeting) were modeled as the state variable of an HMM. Using the Viterbi algorithm they were then able to detect group configurations in real time.

Dielmann and Renals (2007) used a DBN based algorithm to identify group actions such as discussion, monologues, presentation, note taking, etc. They tested the algorithm on the M4 Meeting Corpus, containing 69 short meetings recorded at IDIAP Research Institute in Switzerland (McCowan et al., 2003). The observations used for the DBN included video features, lexical features, prosodic features and speaker turns. Similar dynamic models have also been used for other meeting analysis research. Rienks et al. (2006) determined ranking and influence of individuals in small group meetings using DBN. Various features representing individual speech behavior, interaction behavior and semantics were extracted from the transcripts. Number of turns, turn duration and number of interruptions were used as the observations in a DBN model to automatically detect the influence of individuals.

Fig. 7 shows the graphical representation of the DBN. $L_{ij}(t)$ is a hidden binary random variable; $L_{ij}(t)=1$ if there is a link between members i and j at time t. $S_i(t)$ and $S_j(t)$ are observed binary random variables representing speaking by members i and j; $S_i(t)=1$ if member i is talking at time t. $N_{ij}(t)=1$ if members i and j occupy neighboring physical locations (as in Fig. 6); $G_{ij}(t)=1$ if member i is gazing at member j or vice versa. $I_{ij}(t)$ is a hidden binary random variable: $I_{ij}(t)=1$ if members i and j are indirectly linked, that is, they are both nonspeakers directly linked to the same speaker ($L_{ik}(t)=1$, $L_{jk}(t)=1$, and $S_k(t)=1$). While S_i, N_{ij} and G_{ij} are observed, I_{ij} is deduced from the values of L_{ij} once they have been inferred. An indirect link can only be formed between two non speakers. Direct links are formed between members i and j if i is speaking and j listening or j speaking and i listening. Thus the links are not treated as directional in this work.

4.7. Parameter estimation

Each node has a conditional probability table (CPT), representing stochastic dependency of a node on its parents. The first task prior to conducting the inference is learning these CPTs. The Bayes Net Toolkit was used to model the Dynamic Bayesian Network and estimate the parameters. The Bayes Net Toolkit (BNT) is an open-source Matlab package that can handle different kinds of static and dynamic graphical Bayesian models (Murphy, 2001). Using BNT we can model various static (intra-time point) and Markovian (inter-time points) dependencies. BNT allows use of numerous inference algorithms. For this work the junction tree algorithm was used for inference and applied to pairs of neighboring slices (timestamps) (Murphy, 2001). The parameters were estimated using the EM algorithm. Additional details of the implementation are given in Mathur (2008).

4.8. Inferring links with the Viterbi algorithm

After estimating the parameters, the next step is to perform inference on available data. Inference refers to finding the "most probable" state that fits a single time period. Inference can be performed by maximizing the likelihood of target variables and marginalizing others, or by choosing maximum likelihood values of all hidden variables; we used the latter method, often called the Viterbi algorithm (Forney, 1972).

In order to use the Viterbi algorithm for our DBN, a few variables need to be declared.

Let M_t: Set of all members in room at time t
Let Q_t: Set of all valid states for time t
$Q_t=\{q_1, \ldots\}$ is the set of all state vectors possible at time t. Each state vector is an array of mutually consistent settings of link variables. e.g., $q_n = [L_{12}^{q_n}, \ldots, L_{N-1,N}^{q_n}, I_{12}^{q_n}, \ldots, I_{N-1,N}^{q_n}]$ is the nth such vector.
Q_t is found by enumerating all the combinations of links valid with assumptions stated before.

In the following equations:

π_{q_m} is probability of state q_m at time $t=1$.
$c_{q_m \to q_n}$ is the probability of transition from state q_m to state q_n.
$b_{q_m}(O_t)$ is the probability of observation O_t given state q_m.

$$\pi_{q_m} = \prod_{(i,j)\in D_t} P(L_{ij}^{q_m} | S_i^t, S_j^t), \quad q_m \in Q_1 \tag{13}$$

$$c_{q_m \to q_n} = \prod_{(i,j)\in D_t} P(L_{ij}^{q_n} | L_{ij}^{q_m}, I_{ij}^{q_m}, S_i^t, S_j^t) \tag{14}$$

$$b_{q_m}(O_t) = \prod_{(i,j)\in D_t} P(G_{ij}^t | L_{ij}^{q_m}) P(N_{ij}^t | L_{ij}^{q_m}) \tag{15}$$

$$q_m \in Q_{t-1}, \quad q_n \in Q_t, \quad D_t = \{(i.j) | i \in M_t, j \in M_t, \quad i < j\}$$

The above equations assume the conditional independence of the different variables. Using the variables declared above, we can directly use the Viterbi algorithm as defined in Rabiner (1990). First, all valid states (valid configurations of valid links) for all the snapshots are enumerated. The single best sequence of states is then calculated using the Viterbi algorithm. Each observation sequence

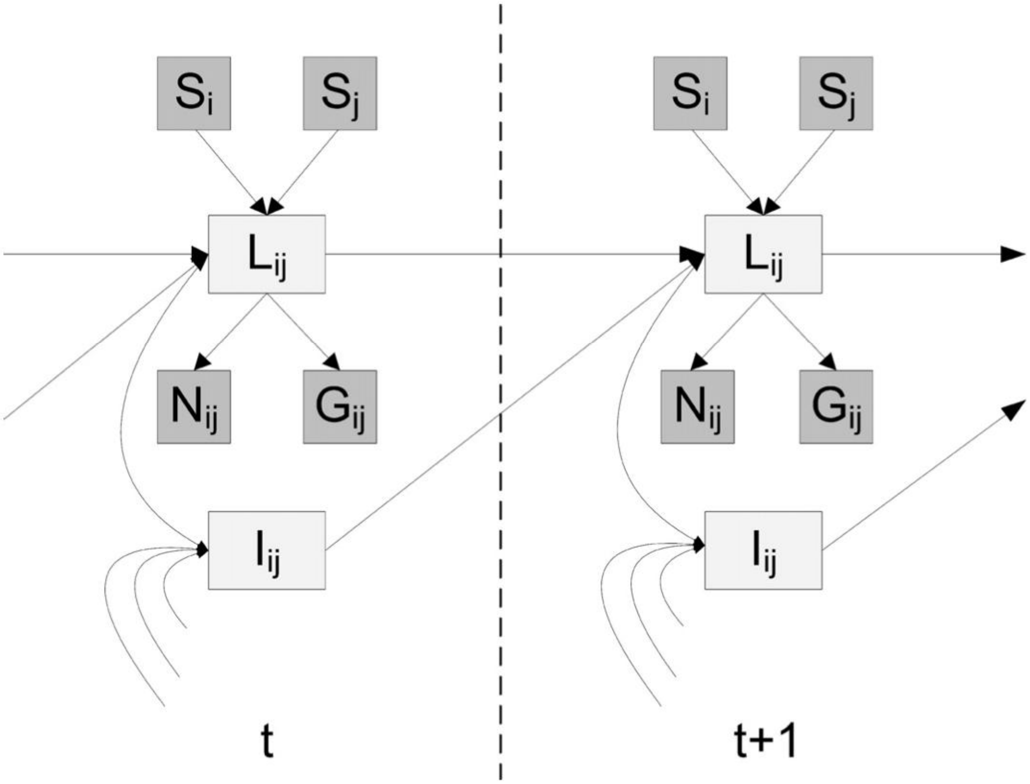

Fig. 7. Dynamic Bayesian Network (DBN) representing the group interaction process. Dark gray nodes are observed; light gray nodes are hidden. Variable names are defined in the text.

consists of a series of snapshots starting in silence, ending in silence, and spanning utterances by one or more members. The performance of Viterbi algorithm is determined by its time complexity which is $O(|Q|^2T)$ where Q is number of states and T is number of observations. As discussed in Section 4.6, the number of valid states at any given time is given by Eq. 12 and T is equal to the length of sequences for which algorithm is run.

5. Results

The accuracy of the algorithm was determined by comparing the detected variables, gaze direction and interaction links, with their annotated values.

5.1. Gaze behavior

In order to evaluate the results obtained by the algorithm described in Section 4.1, gaze directions were also manually extracted from the snapshots using the interface described in Section 3.2. The automatically detected gaze corresponded with manually extracted gaze in 68% of the cases. One major cause of the algorithm failure was "blank stares": when a member was looking up and not staring at some other member.

5.2. Link detection algorithm: accuracy

The results of the link detection algorithm were compared with the manually coded links. We used two measures in order to compare the result: precision and recall. Precision and recall are two widely used statistical indices used to assess the adequacy of the link classification algorithm. Precision is best understood as a measure of exactness or fidelity of classification, whereas recall is a measure of how completely the algorithm identified all cases in which there actually were links. For this case, precision is the number of true positives (the number of dyads correctly labeled as linked by the algorithm) divided by the total number of dyads labeled as positives by the algorithm (including false positives which are dyads incorrectly labeled as linked by the algorithm). Recall is defined as the number of true positives divided by the total number of actual linkages formed (i.e., the sum of true positives and false negatives, which are dyads not labeled as linked when they should have been). Due to coding errors some coded links were invalid as described in Table 1. Hence we also calculated recall if invalid links are removed from the ground truth (recall without error).

The manually coded links were found by pooling the links annotated by 12 coders. Besides running the algorithm on the automatically detected gaze (automatic gaze) and the speakers coded by the annotators using the interface (annotated speakers), we also determined the links using gaze information extracted manually (manual gaze) and speakers coded from the videos/transcripts (speakers from videos/transcripts). These latter two comparisons were intended to shed light on the sensitivity of the algorithm to inputs. The results obtained from the algorithm are shown in Table 2. As can be seen, in the cases where speaker and gaze information is more accurate (i.e., manual gaze and speakers from video/transcripts) we see an increase in the accuracy of the algorithm (as measured by precision and recall without error).

5.3. Link detection algorithm: estimated parameters

Table 3 shows the conditional probability estimated for various dependencies in the DBN. These values can be interpreted as the numerical representation of various qualitative results on conversation group. An interesting result for example was that probability of gaze given that members were linked was around 60% (on average) which is similar to what is found in previous research (Sellen, 1992). Similarly, as expected, the probability of link between members belonging to the same sub group

Table 2
Results for the link detection algorithm.

Type of input	Recall	Precision	Recall without error
Manual gaze/annotated speakers	74.4	69.5	85.1
Automatic gaze/annotated speakers	64.9	67.6	74.9
Manual gaze/speakers from videos and transcripts	72.2	56.9	84.7
Automatic gaze/speakers from videos and transcripts	69.8	57.9	82.6

($I_{ij}(t-1)=T$) is higher than the link probability of any other two unlinked members.

6. Discussion

This paper has described a partially automated algorithm to detect interaction links from video recorded in natural settings. It was specifically designed to use cues that are amenable to automated detection using computer vision and other analytics. To test the algorithm, the data were annotated from the recorded images using a web-based coding interface. This information was first used to infer gaze direction, conversational distance, and whether members were speaking to each other, and then these higher order inferences were combined to detect the interaction links.

The results for automated gaze inference (68%) are comparable to the 74% accuracy obtained by Stiefelhagen et al. (2001) who used a similar mixture of Gaussian approach. However our case was able to handle non-stationary members in a larger group setting in which they moved about.

For the link detection algorithm, comparison can be made with the work of Otsuka et al. (2006). The accuracy obtained using their method ranged from less than 80% to around 92% in small group settings with a controlled environment. Similarly Brdiczka et al. (2005) used an HMM with only audio observations to model turn taking in 4-person small group conversations, and their average recognition accuracy for finding group configurations was 84.8%. For our algorithm recall without error was between 74% and 85%. This is a somewhat lower level of accuracy, but in view of the fact that group members were moving about, the level of accuracy can be judged to be adequate. The annotated observations were substantially reliable. The gaze behavior detection algorithm produced results comparable to previous work. The link detection algorithm was able to recall valid subset of annotated links with an average recall of more than 80%. Considering the complexity of large groups and noisiness associated with manually annotated data, these results are encouraging and should motivate more researchers to extend their work on large group settings. The algorithms described here are generic enough to work independently of whether the observations are manually or automatically derived. Algorithms were specifically designed so that they can be fully automated when advances in video recognition enable automatic identification of the cues used to infer gaze direction, location, and speaking. With improvement in the algorithm and improvement in video recognition, it may someday be possible to have a fully automated link detection system. This would greatly enhance our ability to study large groups of interacting individuals, such as emergency responders, traders on the stock floor, and caregiving teams on hospital floors. The link detection software can be obtained from the first author.

Table 3
Conditional probability table for transitions and observations.

	$L_{ij}(t-1)=F$		$L_{ij}(t-1)=T$
	$I_{ij}(t-1)=F$	$I_{ij}(t-1)=T$	$I_{ij}(t-1)=F$
$L_{ij}(t)=F$	0.96	0.90	0.54
$L_{ij}(t)=T$	0.04	0.10	0.46

	$L_{ij}(t)=F$	$L_{ij}(t)=T$
$G_{ij}(t)=F$	0.98	0.42
$G_{ij}(t)=T$	0.02	0.58
$N_{ij}(t)=F$	0.94	0.63
$N_{ij}(t)=T$	0.06	0.37

Acknowledgments

This research was supported by the National Science Foundation CDI program, grant BCS 0941268 and the University of Illinois program for Critical Initiatives in Research and Scholarship. We would also like to thank coders Ahmed Ahmed, Joseph Blecha, Melissa Dobosh, Tamra Harkey, Erik Johnson, Sang Uk Han, Najung Kim, Avinash Kopparthi, Sejal Patel, Robert Schlehuber, Laura Snyder, Kwassy Sureyaho and Annie Wang. Melissa Dobosh, Diana Jimeno-Ingrum, Annie Wang assisted in data collection. The icons for denoting body language of the members were developed by Joyce Thomas. We would also like to thank Margaret Fleck for her useful advice on the coding interface.

References

Brdiczka, O., Maisonnasse, J., Reignier, P.,2005. Automatic detection of interaction groups. In: ICMI'05: Proceedings of the 7th International Conference on Multimodal Interfaces. ACM, New York, NY, USA, pp. 32–36.

Burgoon, J.K., Buller, D.B., Woodall, W.G., 1996. Nonverbal-Communication: The Unspoken Dialogue, 2nd edition. McGraw-Hill, New York.

Choudhury, T., Pentland, A., 2002. The sociometer: a wearable device for understanding human networks. In: Computer Supported Cooperative Work – Workshop on Ad Hoc Communications and Collaboration in Ubiquitous Computing Environments.

Dielmann, A., Renals, S., 2007. Automatic meeting segmentation using dynamic Bayesian network. IEEE Transactions on Multimedia 9, 25–36.

Einhorn, H., Hogarth, R.M., Klempner, E., 1977. Quality of group judgment. Psychological Bulletin 84 (1), 158–172.

Feldstein, S., Welkowitz, J., 1978. A chronography of conversations: in defense of an objective approach. In: Siegman, A., Feldstein, S. (Eds.), Nonverbal Behavior in Communication. Lawrence Erlbaum Associates, Hillsdale.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin 76, 378–382.

Forney, G.J., 1972. Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference. IEEE Transactions on Information Theory 18, 363.

Garson, G.D., 2008. Factor analysis: Statnotes, from North Carolina State University, Public Administration Program.

Grofman, B., Feld, S.L., Owen, G., 1984. Group size and the performance of a composite group majority: statistical truths and empirical results. Organizational Behavior and Human Performance 33, 350–359.

Henderson, J., Ferreira, F. (Eds.), 2004. The Interface of Language, Vision and Action: Eye Movement and Visual World. Psychology Press.

Homans, G.C., 1951. The Human Group. Harcourt Brace, New York, NY.

Kapferer, B.C., 1969. Norms and the manipulation of relationships in a work context. In: Social Networks in Urban Situations: Analyses of Personal Relationships in Central African Towns. Manchester University Press, Manchester, pp. 181–244.

Knapp, M.L., Hall, J.A., 2002. Nonverbal Communication in Human Interaction, 5th edition. Wadsworth Thomas Learning.

Lam, L., Suen, S.Y., 1997. Application of majority voting to pattern recognition: an analysis of its behavior and performance. IEEE Transactions on Systems, Man and Cybernetics A 27, 553–568.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Lin, X., Yacoub, S., Burns, J., Simske, S., 2003. Performance analysis of pattern classifier combination by plurality voting. Pattern Recognition Letters 24, 1959–1969.

Mathur, S., 2008. An IT framework to study large group interactions and dynamics. Master's Thesis. University of Illinois, Urbana-Champaign.

Please cite this article in press as: Mathur, S., et al., Detecting interaction links in a collaborating group using manually annotated data. Soc. Netw. (2012), http://dx.doi.org/10.1016/j.socnet.2012.04.002

McCowan, I., Bengio, S., Gatica-perez, D., Lathoud, G., Moore, F.M.D., 2003. Modeling human interactions in meetings. In: Proc. IEEE ICASSP, Hong Kong.

McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., 2003. Automatic analysis of multimodal group actions in meetings. Technical Report RR 03-27. IDIAP.

Moreno, J.L., 1951. Sociometry, Experimental Method and the Science of Society. Beacon House, Inc., Oxford, England.

Murphy, K., 1998. A Brief Introduction to Graphical Models and Bayesian Networks., http://www.cs.ubc.ca/murphyk/Bayes/bnintro.html.

Murphy, K., 2001. Applying the junction tree algorithm to variable-length DBNS. Tech Report. http://www.cs.berkeley.edu/murphyk/Papers/jtree_dbn.ps.gz.

Murphy, K., 2001. The Bayes Net Toolbox for Matlab. Computing Science and Statistics 33.

Neuendorf, K.A., 2001. The Content Analysis Guidebook. Sage Publications, Inc.

Otsuka, K., Yamato, J., Takemae, Y., Murase, H., 2006. Conversation scene analysis with dynamic Bayesian network based on visual head tracking. In: 2006 IEEE International Conference on Multimedia and Expo, pp. 949–952.

Pan, X., Han, C., Dauber, K., Law, K., 2007. A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations. AI & Society 22, 113–132.

Prasov, Z., Chai, J.,2008. What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In: IUI'08: Proceedings of the 13th International Conference on Intelligent User Interfaces. ACM, New York, NY, USA, pp. 20–29.

Rabiner, L.R., 1990. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 267–296.

Ramanan, D., Forsyth, D.A., Zisserman, A., 2007. Tracking people by learning their appearance. IEEE Transations on Pattern Analysis and Machine Intelligence 29, 65–81.

Rienks, R., Zhang, D., Gatica-Perez, D., Post, W., 2006. Detection and application of influence rankings in small group meetings. In: Proc. International Conference on Multimodal Interfaces (ICMI-06), pp. 257–264.

Sellen, A.J.,1992. Speech patterns in video-mediated conversations. In: CHI'92: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 49–59.

Siegel, A.F., Wennerstrom, A., 2003. Keeping the floor in multiparty conversations: intonation, syntax, and pause. Discourse Processes 36, 77–107.

Stiefelhagen, R., Yang, J., Waibel, A.,2001. Estimating focus of attention based on gaze and sound. In: PUI'01: Proceedings of the 2001 Workshop on Perceptive User Interfaces. ACM, New York, NY, USA, pp. 1–9.

Stiefelhagen, R., Zhu, J.,2002. Head orientation and gaze direction in meetings. In: CHI'02: CHI'02 Extended Abstracts on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 858–859.

Zmijewski, B., 2004. Me110a. http://www.zurb.net/me110a/2d.html (accessed 07.09.08).

Please cite this article in press as: Mathur, S., et al., Detecting interaction links in a collaborating group using manually annotated data. Soc. Netw. (2012), http://dx.doi.org/10.1016/j.socnet.2012.04.002