Dynamic Multi-Valued Network Models for Predicting Face-to-Face Conversations

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

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

Jeff Bilmes
Dept. of Elec. Engineering
University of Washington
Seattle, WA
bilmes@ee.washington.edu

Abstract

We introduce a new probabilistic framework for collectively modeling people's social behavior from local sensor observations. Our approach extends curved exponential random graph models to (1) include features that account for multi-valued edges, and (2) model the change in edge values over time. We present empirical results on a real world dataset of face-to-face conversations collected from 24 individuals using wearable sensors over the course of 9 months. The results demonstrate that the model is capable of predicting not just whether but also for how long two people will converse and that the ordinality of discretized observations can be exploited to reduce the number of parameters.

1 Introduction

It is becoming increasingly easy to collect data that captures the simultaneous, real-world behavior of entire groups of people [4, 31, 29]. Such data sets often capture, either directly or indirectly, interactions between people. Despite that, much of the research on such data considers behavior only at the level of a single person (e.g. [17, 12, 15]). Models that do consider social behavior typically rise only to the level of the dyad [20] or small interacting group [8]. Conversely, an arsenal of techniques has been developed for social network analysis [21, 9, 28, 26, 3] but most of those methods consider only static, binary networks. Social networks derived from behavioral data will almost always be temporal and will often have finer grained observations about interactions than simple binary indicators. Work on multi-valued [22] or temporal [19] network models has been scarce since such data was previously hard to obtain.

The main contribution of this paper is a new modeling framework that simultaneously models the dynamics and structural properties (e.g., transitivity, network density) of automatically collected behavioral data. Our model extends curved exponential random graph models to learn the strengths of pairwise interactions and how those interactions evolve over time. To the best of our knowledge, this is the first implementation of dynamic, multi-valued CERGMs as well as the first application of them to behavioral data. We present experimental results on real-world data that demonstrate the strengths of our model.

2 The UW Spoken Networks Dataset

In order to build community-scale models of human behavior, we have collected a data set that captures the face-to-face conversations between a cohort of incoming graduate students. These students were from the same academic department at a large research university – 24 of 27 eligible subjects participated.

Each subject wore a sensing device containing 8 different sensors useful for detecting conversations, activities, and environmental context. Data was collected during working hours for one week each month over the 9 month course of an academic year. A complete description of the data is in [31]. In this paper, we only use information extracted from the audio.

This data set is novel in several respects. First, it directly captures real-world face-to-face conversations, which remain people's primary mode of social interaction [1]. While, there have been a few earlier efforts towards the direct recording of face-to-face interaction, those required human observers and manual coding [2, 7]—techniques that can only be applied to small study populations over brief observation periods. A second novel aspect of our data is that it is longitudinal. It is difficult to observe real-world interactions at even a single point in time; multiple observations at many different time points are clearly even more difficult.

Which brings up a third novel aspect of our data: it is automatically collected and processed. Automated recording and processing not only increases the scale—both in number of subjects and length of observation period—at which interactions can be studied, it also makes possible applications that have access to real-time information about a group's social network.

Inferring Conversational Behavior In earlier work, we developed techniques for determining who is in conversation with whom, and who speaks when in a conversation. These techniques involve a series of lower level probabilistic models whose outputs are fed into each other to produce high level inferences about conversations and speakers. They are capable of recovering who was in conversation with whom with an accuracy ranging from 96.1% to 99.2%. The details of those techniques are explained fully in [30]. The resulting high level inference produces a rich corpus of data about interactions which serves as observations for our network modeling step.

3 Modeling Network Structure

Traditionally, statistical analysis of social networks has focused on finding descriptive statistics—path lengths, degree distributions, clustering coefficients—that describe global features of the network [26]. In recent decades, a new class of models known as exponential random graph models (ERGMs, also sometimes called p-star models) has been developed [6, 27, 23]. ERGMs depart from traditional descriptive models by considering a social network as a realization of a set of random variables, one variable for each potential edge in the network. By considering a distribution over networks and network statistics (instead of considering just a single observed value), ERGMs can exploit and understand any underlying uncertainty in the data.

3.1 Curved Exponential Random Graph Models

Given an observed network, exponential random graph models (ERGMs) estimate the parameters of an exponential family model that describes the joint distribution of the edge variables. The probability distribution takes the form (typical for exponential families) of a log-linear combination of features and weights:

$$p(\mathbf{Y} = \mathbf{y}) = \frac{1}{Z_{\boldsymbol{\eta}}} e^{\boldsymbol{\eta}^\top \boldsymbol{\phi}(\mathbf{y})} \tag{1}$$

$\mathbf{Y}$ are the variables representing edges in the graph, $\boldsymbol{\phi}$ are feature functions defined on $\mathbf{y}$, $\boldsymbol{\eta}$ is a vector of weights to be learned, and $Z_{\boldsymbol{\eta}}$ is a normalizing constant. The features are deterministic functions (or statistics) of the network. Typical features are counts of subgraph occurrences, such as the number of triangles or even simply the number of edges. The strength of these models lies in their ability to capture the structural dependencies in a probabilistic manner. Properties of the network can then be interpreted in terms of how they affect the network's probability.

Despite the rich theory behind ERGMs, parameter learning has proven to be difficult due to model degeneracy. Models are considered degenerate if only a small set of parameter values lead to plausible networks. Slight changes in the parameter values of a degenerate model can cause it to put all of its probability on almost entire empty or entirely complete networks [10, 24].

Recently, [14] proposed using a curved exponential family model to avoid degeneracy but at the price of using more complicated features. A curved exponential family allows for non-linear constraints

to be placed on the values $\boldsymbol{\eta}$ is allowed to take. In that case, $\boldsymbol{\eta}$ is redefined as a non-linear function mapping a point $\boldsymbol{\theta}$ in q-dimensional space to a point $\boldsymbol{\eta}(\boldsymbol{\theta})$in p-dimensional space, where $q < p$. The points $\boldsymbol{\theta} \in \Theta$ then define a q-dimensional curved manifold in p-dimensional space and thus models defined in a such a way are called curved exponential families [5]. The likelihood for a curved exponential family is written as

$$p(\mathbf{Y} = \mathbf{y}) = \frac{1}{Z} e^{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{\phi}(\mathbf{y})} \tag{2}$$

This new formulation, known as curved ERGMs (CERGMs) has led to better model fits than linear ERGMs. CERGMs have the additional benefit of continuing to use intuitive features while also learning interesting aspects of those features.

3.2 Features for CERGMs

The simple subgraph counts used as features in ERGMs to model social ties can lead to model degeneracy [10], but they often also do not fully capture the intuitions that motivated the features. For example, one expects social networks to exhibit transitivity, but only up to a point. Networks do not eventually become their complete transitive closures.

[14] recasts ERGMs using a curved exponential family framework that allows entire histograms of statistics to be used as features while still requiring only a small number of parameters. For example, the traditional ERGM feature for capturing transitivity is the count of all triangles that appear in the network. [14] replaces that count with the edgewise shared partner histogram of the network: a vector where component i counts the numbers of edges whose endpoints have exactly i shared partners. For an n node network, there are $n - 2$ bins in that histogram. Each bin receives its own weight parameter but the weights are constrained so that the weight w_i for the i-th bin is

$$w_i = m \left[e^r (1 - (1 - e^{-r})^i) \right] \tag{3}$$

which is clearly a function of just two parameters: m, the usual multiplicative weight, and r, the rate at which the growth of w in i diminishes. Since that rate of diminishing increase is geometric, the above combination of features and constrained parameters is known as geometrically weighted edgewise shared partners (GWESP).

Other features that are functions of these statistics can be incorporated into the model by incorporating their weights into the function $\boldsymbol{\eta}$ without increasing the number of statistics (p). For example, network density can be computed from the degree distribution as $\sum_i \frac{1}{2} D_i(\mathbf{y})$. [13] provides a thorough history and derivation of these features.

[14] defined other similar features and we take three of them as a starting point for our model: (1) network density, (2) the geometrically weighted degree distribution (GWD) and (3) the GWESP. GWD is defined almost identically to GWESP but the edgewise shared partner histogram is replaces with the network's degree histogram.

4 Collective Modeling of Conversational Behavior

All specifications of CERGMs to date (along with many other kinds of social network analysis) have used only binary values for edges in the network and the features used to capture network properties are only defined for binary edge values. When using non-binary data researchers typically define simple thresholds or heuristics to discard observations that are believed *a priori* to not represent ties [18, 16]. To model social behavior, it is desirable to retain information about the intensity of a tie captured by that behavior and to consider a more nuanced representation of the network. For modeling our data we have extended the traditional CERGM feature set to handle multi-valued networks—networks whose edges can take more values than 0 or 1.

4.1 Multi-Valued CERGMs

Our model allows edges to take one of v discrete, ordinal values. These values represent the observed intensity of a social tie. Larger values indicate a stronger tie. To permit comparisons with binary-valued models, the values are scaled so that the smallest is 0 and the largest is 1. We redefine

traditional features in a straightforward manner. The density of a network is the sum of its edge values. A node's degree is the sum of the values of the edges incident to that node.

More complicated features that involve subgraphs require defining the intensity of a subgraph. We use the geometric mean of the edge values composing the subgraph as the intensity value for the subgraph. For example, a shared partner k for nodes i and j is defined to be a partner of intensity $(y_{ik}y_{jk})^{\frac{1}{2}}$, where y_{ij} represents the multi-valued edge between nodes i and j. The count of shared partners for a pair, SP_{ij} is the sum of these intensities:

$$SP_{ij} \triangleq \sum_k (y_{ik}y_{jk})^{\frac{1}{2}} \tag{4}$$

To model edgewise shared partners (i.e. mutual friends) we take the product of an edge's value with its shared partner sum: $ESP_{ij} \triangleq y_{ij}SP_{ij}$.

Note that if $v = 2$ and all values are either 0 or 1, then our features are equivalent to the traditional CERGM features.

4.2 Temporal Multi-Valued CERGMs

Beyond allowing edges to assume non-binary values, we further extend traditional CERGM models so that they may capture how edge intensities change over time. Call all of the features described so far static features. We model a dynamic social network as a discrete time Markov chain where each timestep is a complete network modeled by a set of static features. A set of dynamic features ties adjacent timesteps together and models how the network changes.

For the experiments in this work, we used a single dynamic feature: $\mathbf{T}$, a $v \times v$ matrix of transition counts. T_{rs} is the number of dyads that took value r at time t and moved to value s at time $t+1$, for all timesteps in the data. As with the static histogram features, the model has one weight per component of $\mathbf{T}$. However, given the flexibility of the function $\boldsymbol{\eta}(\boldsymbol{\theta})$ in (2) there can be fewer parameters than components of $\mathbf{T}$.

To smoothly set weights on the transition counts, we use a parameter constraint that assigns the weight w_{rs} for T_{rs} as

$$w_{rs} = c_r(m_r - s)^2 \tag{5}$$

There are only two parameters, m_r and c_r, and each row has its own pair of parameters. If c_r is negative, this weighting scheme is similar to a Gaussian, truncated to the range of allowable edge values, with mean m_r and concentration c_r. But c_r does not have to be negative. This constraint can also assign weights that *increase* with the distance from m_r. m_r is similarly unconstrained and can lie outside the allowable range of edge values. Altogether, that allows much flexibility in how weights may be assigned to rows of $\mathbf{T}$. When more accuracy in edge values is desired and the number of bins in the discretization is increased, the number of parameters will only grow linearly.

We compare the smoothed weights of (5) to a fully parameterized model with one parameter per component of $\mathbf{T}$, excluding one column. That is the same number of parameters that would be available to a typical model that estimates a complete matrix of transition probabilities between edge values. (The excluded column reflects the fact that only $v-1$ parameters are needed for each v-category multinomial of the transition matrix.) As the discretization grows finer, the number of parameters in the full parameter case obviously grows quadratically.

Altogether the models we test employ 3 static features (density, GWD, and GWESP) with 5 parameters (3 multiplicative weights and 2 geometric rates), and one dynamic feature with either $2v$ or $v^2 - v$ parameters.

Learning Maximum pseudolikelihood estimation is often used to learn parameter values in ERGMs [25]. The pseudo-loglikelihood for our model is defined as the sum of the conditional log-likelihood of each dyad, given the rest of the network. Given an observed network, the parameters that maximize the pseudo-loglikelihood can be found using quasi-Newton methods. We have used BFGS in our experiments and found that it performs acceptably, despite the fact the non-linear parameter constraints make both the likelihood and the pseudolikelihood non-convex. Figure 1 shows the progress during learning on synthetic data with a completely parameterized transition matrix. Note that $v = 5$ in that model so the distance is for 30 parameters.

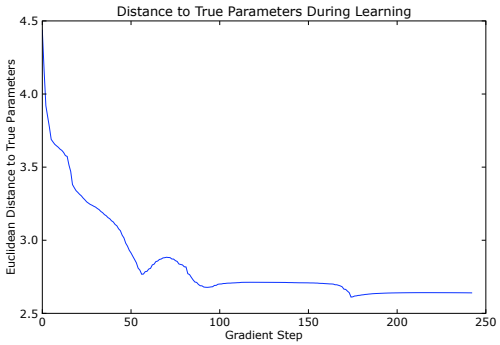


Figure 1: Progress during learning on synthetic data.

5 Experimental Results

For our experiments, we used the proportion of time that two people spend in conversation as the value of the edge between them. We varied v from 2 to 17 to test different discretization granularities. The lowest value (set to 0) is always used for pairs that spend no time in conversation. The remaining bins are spaced evenly from the minimum observed non-zero value to the maximum observed value. When $v = 2$ the model is equivalent to a traditional CERGM with binary edge variables.

To see how well our model can predict future observations, we learned parameters using leave-one-out cross validation, training on 9 splits with 8 weeks of data each. To predict edge values, we compute the conditional probability of each edge value for a tie given the other ties in the network and choose the value with the highest conditional probability as the prediction.

To evaluate the prediction, we compute the absolute error $|\hat{y}_{ij} - y_{ij}|$, where $\hat{y}_{ij}$ is the predicted value and y_{ij} is the true value. The mean absolute error for all predictions is defined as the total error. However, the total error does not provide a complete summary of the model's performance. In a social network there is a difference of kind between zero values and non-zero values beyond their simple absolute difference. Replacing a zero-valued edge with even a small valued edge can have large effects on network properties such as path lengths or reachability. To examine that source of error, we compute two additional evaluation metrics. The false positive error is the mean of all absolute errors where the true value is zero. When $v = 2$, the false positive error is equal to the false positive rate (one minus recall). The false negative error is the mean of all absolute errors when the predicted value is zero. When $v = 2$ the false negative error is equal to the false negative rate.

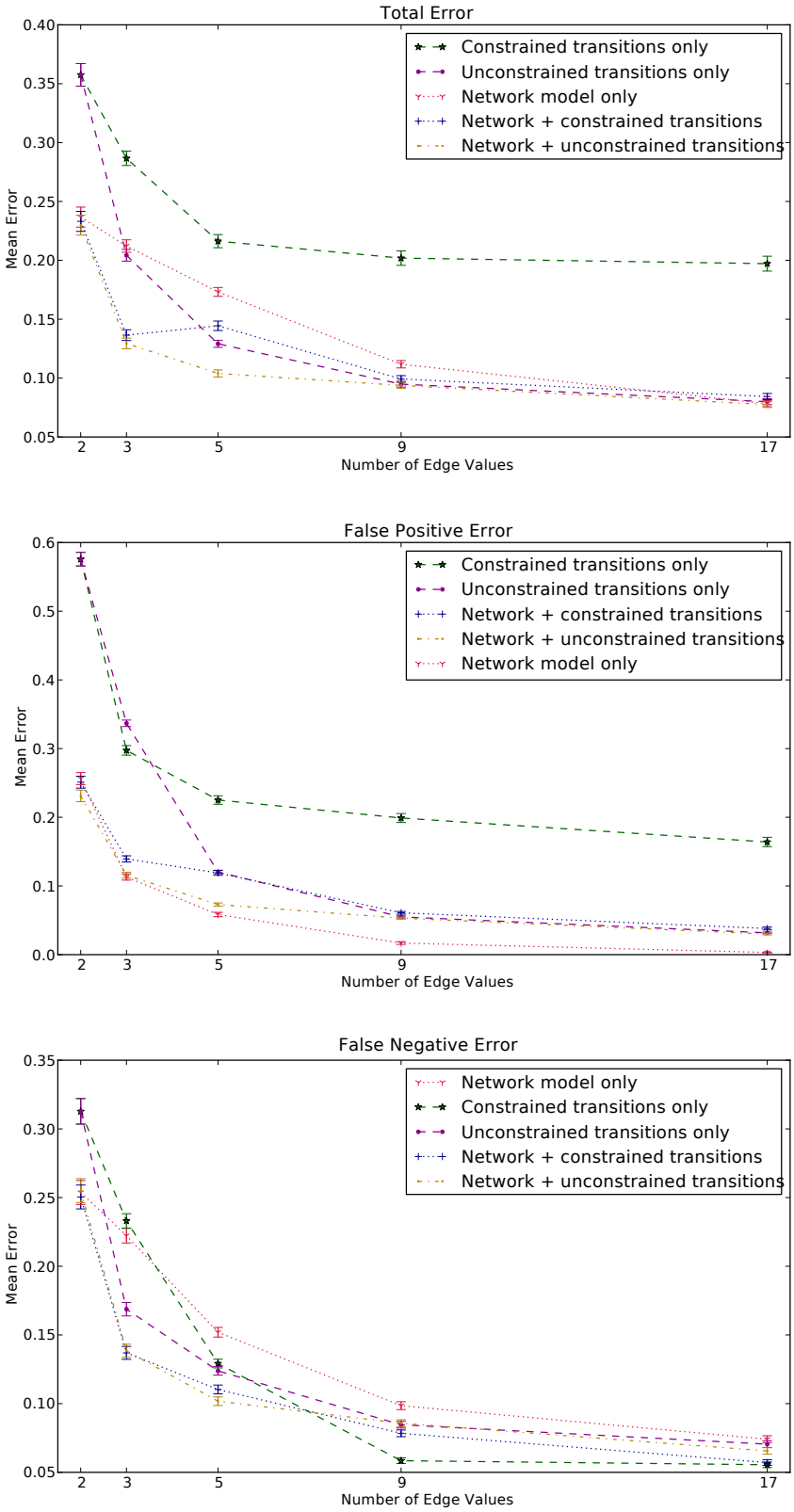
Figure 2 shows these three error metrics for 5 different models with increasing numbers of edge values. The models considered are: (i) a static network model with only network density, GWD, and GWESP as its features; (ii) the network model with the constrained transition parameters defined in (5); (iii) the constrained transition model alone, with no network structure model; (iv) the network model with the fully parameterized transition matrix; and (v) the fully parameterized transition matrix alone.

Unsurprisingly, the network plus full transitions model performs best. It has far more parameters than any other model. As the granularity of the data becomes finer, though, the network plus constrained transition model begins to perform nearly as well despite having e.g. only 23 parameters at 9 edge values compared to 77 for the fully parameterized model. The constrained network model has a similar false positive error, compared to the unconstrained models, but a lower false negative error. The static structural model also performs surprisingly well (and with a fixed number of parameters no matter how many edge values are used). Its predictions are also the most conservative with a low false positive error but high false negative error.

6 Conclusion and Future Work

We have presented two extensions to CERGMS—multi-valued edges and temporal features—that allow them to model networks of face-to-face conversations observed over time. We showed that the

Figure 2: Prediction results for 5 different models. Error bars show $\pm$ 2 standard errors.



models can be used to predict not just whether unseen edges exist, but also their specific intensity value; that prediction is improved by modeling the structural properties of the social network; and that constrained dynamic parameters perform nearly as well as unconstrained parameters.

Our transition model is very simple and ignores any temporal structural features—like closing a triangle—that may also improve prediction. We also have far more information about conversations than just the time spent. We can discover turn-taking patterns, or changes in pitch and rate, and use those to provide more information about the edges in the network. Features that connect conversation qualities to network structure could show whether a person's conversational behavior is related to her position in the network.

Finally, in addition to the leave-one-out experiment presented here, we also tried a more realistic prediction task using increasing amounts of data and always predicting future data. In that case more data seemed to have no effect on prediction accuracy which suggests that the usual assumption of time-inhomogeneity may not apply in our data. (Indeed, it may not apply to social networks in general: the only other temporal ERGM model that we are aware of found it necessary to discard early data to avoid time-dependent effect [11].) We are currently working on extending our parameter constraints to allow the modeling of time-inhomogeneous effects that still evolve smoothly and predictably.

References

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

[2] H. R. Bernard, P. D. Killworth, and L. Sailer. Informant accuracy in social network data IV: a comparison of clique-level structure in behavioral and cognitive network data. *Social Networks*, 2(3), 1980.

[3] P. J. Carrington, J. Scott, and S. Wasserman, editors. *Models and Methods in Social Network Analysis*. Cambridge UP, 2005.

[4] T. Choudhury and A. Pentland. Sensing and modeling human networks using the sociometer. In *Proc. of the Int'l Conference on Wearable Computing*, 2003.

[5] Bradley Efron. The geometry of exponential families. *The Annals of Statistics*, 6(2), 1978.

[6] O. Frank and D. Strauss. Markov graphs. *J. Am. Statistical Association*, 81, 1986.

[7] L. C. Freeman, S. C. Freeman, and A. G. Michaelson. On human social intelligence. *J. Soc. and Biol. Structures*, 11, 1988.

[8] D. R. Gibson. Taking turns and talking ties: Networks and conversational interaction. *Am. J. Sociol*, 110(6), 2005.

[9] M. Granovetter. The strength of weak ties. *Am. J. Sociol.*, 78(6), 1973.

[10] M. Handcock. Assessing degeneracy in statistical models of social networks. Technical Report 39, UW CSSS, 2003.

[11] S. Hanneke, W. Fu, and E. Xing. Discrete temporal models of social networks. arXiv, August 2009.

[12] M. R. Hodges and M. E. Pollack. An 'object-use fingerprint': The use of electronic sensors for human identification. In *Proc. of UbiComp*, 2007.

[13] D. Hunter. Curved exponential family models for social networks. *Social Networks*, 29(2), 2007.

[14] D. R. Hunter and M. Handcock. Inference in curved exponential family models for networks. *J. Computational and Graphical Statistics*, 15(3), 2006.

[15] T. Huynh, M. Fritz, and B. Schiele. Discovery of activity patterns using topic models. In *Proc. of UbiComp*, 2008.

[16] G. K. and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311, 2006.

[17] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition. In *Proc. of NIPS*, 2005.

[18] G. Palla, A.-L. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446, 2006.

[19] P. Pattison and G. Robbins. Random graph models for temporal processes in social networks. *J. Math. Sociol.*, 25, 2001.

[20] A. Pentland. Automatic mapping and modeling of human networks. *PhysicaA*, 2007.

[21] A. R. Radcliffe-Brown. On social structure. *J. Royal Anthropological Institute of GB and Irl.*, 70(1), 1940.

[22] G. Robins, P. Pattison, and S. Wasserman. Logit models and logistic regressions for social networks: Iii. valued relations. *Psychometrika*, 64(3):371–394, 1999.

[23] G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, 29(2), 2007.

[24] T. Snijders. Markov chain monte carlo estimation of exponential random graph models. *J. Social Structure*, 3(2), 2002.

[25] D. Strauss and M. Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85:204–212, 1990.

[26] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge UP, 1994.

[27] S. Wasserman and P. Pattison. Logit models and logistic regression for social networks: 1. an introduction to markov graphs and (p*). *Psychometrika*, 61, 1996.

[28] B. Wellman and S. D. Berkowitz, editors. *Social Structures: A Network Approach*. Cambridge UP, 1988.

[29] C. Wren, Y. Ivanov, D. Leigh, and J. Westhues. The merl motion detector dataset. In *MD '07: Proc. of the 2007 workshop on Massive datasets*, 2007.

[30] D. Wyatt, T. Choudhury, and J. Bilmes. Conversation detection and speaker segmentation in privacy sensitive situated speech data. In *Proc. of Interspeech*, 2007.

[31] D. Wyatt, T. Choudhury, and H. Kautz. Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort. In *Proc. of ICASSP*, 2007.