"You're Mr. Lebowski, I'm the Dude": Inducing Address Term Formality in Signed Social Networks

Vinodh Krishnan
College of Computing
Georgia Institute of Technology
Atlanta, GA 30308
krishnan.vinodh@gmail.com

Jacob Eisenstein
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA 30308
jacobe@gatech.edu

Abstract

We present an unsupervised model for inducing signed social networks from the content exchanged across network edges. Inference in this model solves three problems simultaneously: (1) identifying the sign of each edge; (2) characterizing the distribution over content for each edge type; (3) estimating weights for triadic features that map to theoretical models such as structural balance. We apply this model to the problem of inducing the social function of **address terms**, such as *Madame*, *comrade*, and *dude*. On a dataset of movie scripts, our system obtains a coherent clustering of address terms, while at the same time making intuitively plausible judgments of the formality of social relations in each film. As an additional contribution, we provide a bootstrapping technique for identifying and tagging address terms in dialogue.[1]

1 Introduction

One of the core communicative functions of language is to modulate and reproduce **social dynamics**, such as friendship, familiarity, formality, and power (Hymes, 1972). However, large-scale empirical work on understanding this communicative function has been stymied by a lack of labeled data: it is not clear what to annotate, let alone whether and how such annotations can be produced reliably. Computational linguistics has made great progress in modeling language's informational dimension, but — with a few notable exceptions — computation has had little to contribute to our understanding of language's social dimension.

Yet there is a rich theoretical literature on social structures and dynamics. In this paper, we focus on one such structure: signed social networks, in which edges between individuals are annotated with information about the nature of the relationship. For example, the individuals in a dyad may be friends or foes; they may be on formal or informal terms; or they may be in an asymmetric power relationship. Several theories characterize signed social networks: in structural balance theory, edge signs indicate friendship and enmity, with some triads of signed edges being stable, and others being unstable (Cartwright and Harary, 1956); conversely, in status theory (Leskovec et al., 2010b), edges indicate status differentials, and triads should obey transitivity. But these theoretical models can only be applied when the sign of each social network connection is known, and they do not answer the sociolinguistic question of how the sign of a social tie relates to the language that is exchanged across it.

We present a unified statistical model that incorporates both network structure and linguistic content. The model connects signed social networks with **address terms** (Brown and Ford, 1961), which include names, titles, and "placeholder names," such as *dude*. The choice of address terms is an indicator of the level of formality between the two parties: for example, in contemporary North American English, a formal relationship is signaled by the use of titles such as *Ms* and *Mr*, while an informal relationship is signaled by the use of first names and

[1] Code and data for this paper is available at https://github.com/vinodhkris/signed-social.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1616–1626,
Denver, Colorado, May 31 – June 5, 2015.

placeholder names. These tendencies can be captured with a multinomial distribution over address terms, conditioned on the nature of the relationship. However, the linguistic signal is not the only indicator of formality: network structural properties can also come into play. For example, if two individuals share a mutual friend, with which both are on informal terms, then they too are more likely to have an informal relationship. With a log-linear prior distribution over network structures, it is possible to incorporate such triadic features, which relate to structural balance and status theory.

Given a dataset of unlabeled network structures and linguistic content, inference in this model simultaneously induces three quantities of interest:

- a clustering of network edges into types;
- a probabilistic model of the address terms that are used across each edge type, thus revealing the social meaning of these address terms;
- weights for triadic features of signed networks, which can then be compared with the predictions of existing social theories.

Such inferences can be viewed as a form of **sociolinguistic structure induction**, permitting social meanings to be drawn from linguistic data. In addition to the model and the associated inference procedure, we also present an approach for inducing a lexicon of address terms, and for tagging them in dialogues. We apply this procedure to a dataset of movie scripts (Danescu-Niculescu-Mizil and Lee, 2011). Quantitative evaluation against human ratings shows that the induced clusters of address terms correspond to intuitive perceptions of formality, and that the network structural features improve predictive likelihood over a purely text-based model. Qualitative evaluation shows that the model makes reasonable predictions of the level of formality of social network ties in well-known movies.

We first describe our model for linking network structure and linguistic content in general terms, as it can be used for many types of linguistic content and edge labels. Next we describe a procedure which semi-automatically induces a lexicon of address terms, and then automatically labels them in text. We then describe the application of this procedure to a dataset of movie dialogues, including quantitative and qualitative evaluations.

2 Joint model of signed social networks and textual content

We now present a probabilistic model for linking network structure with content exchanged over the network. In this section, the model is presented in general terms, so that it can be applied to any type of event counts, with any form of discrete edge labels. The application of the model to forms of address is described in Sections 4 and 5.

We observe a dataset of undirected graphs $G^{(t)} = \{i, j\}$, with a total ordering on nodes such that $i < j$ in all edges. For each edge $\langle i, j \rangle$, we observe directed content vectors $\boldsymbol{x}_{i \rightarrow j}$ and $\boldsymbol{x}_{i \leftarrow j}$, which may represent counts of words or other discrete events, such as up-votes and down-votes for comments in a forum thread. We hypothesize a latent edge label $y_{ij} \in \mathcal{Y}$, so that $\boldsymbol{x}_{i \rightarrow j}$ and $\boldsymbol{x}_{i \leftarrow j}$ are conditioned on y_{ij}. In this paper we focus on binary labels (e.g., $\mathcal{Y} = \{+, -\}$), but the approach generalizes to larger finite discrete sets, such as directed binary labels (e.g., $\mathcal{Y} = \{++, +-, -+, --\}$) and comparative status labels (e.g., $\mathcal{Y} = \{<, >, \approx\}$).

We model the likelihood of the observations conditioned on the edge labels as multinomial,

$$\boldsymbol{x}_{i \rightarrow j} \mid y_{ij} \sim \text{Multinomial}(\boldsymbol{\theta}^{\rightarrow}_{y_{ij}}) \quad (1)$$

$$\boldsymbol{x}_{i \leftarrow j} \mid y_{ij} \sim \text{Multinomial}(\boldsymbol{\theta}^{\leftarrow}_{y_{ij}}). \quad (2)$$

Parameter tying can be employed to handle special cases. For example, if the edge labels are undirected, then we add the constraint $\boldsymbol{\theta}^{\rightarrow}_{y} = \boldsymbol{\theta}^{\leftarrow}_{y}, \forall y$. If the edge labels reflect relative status, then we would instead add the constraints $(\boldsymbol{\theta}^{\rightarrow}_{<} = \boldsymbol{\theta}^{\leftarrow}_{>})$, $(\boldsymbol{\theta}^{\rightarrow}_{>} = \boldsymbol{\theta}^{\leftarrow}_{<})$, and $(\boldsymbol{\theta}^{\rightarrow}_{\approx} = \boldsymbol{\theta}^{\leftarrow}_{\approx})$.

The distribution over edge labelings $P(\boldsymbol{y})$ is modeled in a log-linear framework, with features that can consider network structure and signed triads:

$$P(\boldsymbol{y}; G, \boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{Z(\boldsymbol{\eta}, \boldsymbol{\beta}; G)} \times \exp \sum_{\langle i,j \rangle \in G} \boldsymbol{\eta}^{\top} \boldsymbol{f}(y_{ij}, i, j, G) \times \exp \sum_{\langle i,j,k \rangle \in \mathcal{T}(G)} \beta_{y_{ij}, y_{jk}, y_{ik}}, \quad (3)$$

where $\mathcal{T}(G)$ is the set of triads in the graph G. The first term of Equation 3 represents a normalizing constant. The second term includes weights $\boldsymbol{\eta}$, which apply to network features $\boldsymbol{f}(y_{ij}, i, j, G)$. This can include features like the number of mutual friends between nodes i and j, or any number of more elaborate structural features (Liben-Nowell and Kleinberg, 2007). For example, the feature weights $\boldsymbol{\eta}$ could ensure that the edge label $Y_{ij} = +$ is especially likely when nodes i and j have many mutual friends in G. However, these features cannot consider any edge labels besides y_{ij}.

In the third line of Equation 3, each weight $\beta_{y_{ij},y_{jk},y_{ik}}$ corresponds to a signed triad type, invariant to rotation. In a binary signed network, structural balance theory would suggest positive weights for β_{+++} (all friends) and β_{+--} (two friends and a mutual enemy), and negative weights for β_{++-} (two enemies and a mutual friend) and β_{---} (all enemies). In contrast, a status-based network theory would penalize non-transitive triads such as $\beta_{>><}$. Thus, in an unsupervised model, we can examine the weights to learn about the semantics of the induced edge types, and to see which theory best describes the signed network configurations that follow from the linguistic signal. This is a natural next step from prior work that computes the frequency of triads in explicitly-labeled signed social networks (Leskovec et al., 2010b).

3 Inference and estimation

Our goal is to estimate the parameters θ, β, and $\boldsymbol{\eta}$, given observations of network structures $G^{(t)}$ and linguistic content $\boldsymbol{x}^{(t)}$, for $t \in \{1, \ldots, T\}$. Eliding the sum over instances t, we seek to maximize the variational lower bound on the expected likelihood,

$$\begin{aligned}\mathcal{L}_Q =& E_Q[\log P(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\theta}, G)] - E_Q[\log Q(\boldsymbol{y})] \\ =& E_Q[\log P(\boldsymbol{x} \mid \boldsymbol{y}; \boldsymbol{\theta})] + E_Q[\log P(\boldsymbol{y}; G, \boldsymbol{\beta}, \boldsymbol{\eta})] \\ & - E_Q[\log Q(\boldsymbol{y})]. \end{aligned} \tag{4}$$

The first and third terms factor across edges,

$$\begin{aligned}E_Q[\log P(\boldsymbol{x} \mid \boldsymbol{y}; \boldsymbol{\theta})] =& \sum_{\langle i,j\rangle \in G} \sum_{y' \in \mathcal{Y}} q_{ij}(y') \boldsymbol{x}_{i \rightarrow j}^\top \log \boldsymbol{\theta}_{y'}^{\rightarrow} \\ & + q_{ij}(y') \boldsymbol{x}_{i \leftarrow j}^\top \log \boldsymbol{\theta}_{y'}^{\leftarrow} \\ E_Q[\log Q(\boldsymbol{y})] =& \sum_{\langle i,j\rangle \in G} \sum_{y' \in \mathcal{Y}} q_{ij}(y') \log q(y').\end{aligned}$$

The expected log-prior $E_Q[\log P(\boldsymbol{y})]$ is computed from the prior distribution defined in Equation 3, and therefore involves triads of edge labels,

$$\begin{aligned}E_Q[\log P(\boldsymbol{y}; \boldsymbol{\eta}, \boldsymbol{\beta})] =& -\log Z(\boldsymbol{\eta}, \boldsymbol{\beta}; G) \\ & + \sum_{\langle i,j\rangle \in G} \sum_{y'} q_{ij}(y') \boldsymbol{\eta}^\top \boldsymbol{f}(y', i, j, G) \\ & + \sum_{\langle i,j,k\rangle \in \mathcal{T}(G)} \sum_{y,y',y''} q_{ij}(y) q_{jk}(y') q_{ik}(y'') \beta_{y,y',y''}.\end{aligned}$$

We can reach a local maximum of the variational bound by applying expectation-maximization (Dempster et al., 1977), iterating between updates to $Q(\boldsymbol{y})$, and updates to the parameters $\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta}$. This procedure is summarized in Table 1, and described in more detail below.

3.1 E-step

In the E-step, we sequentially update each q_{ij}, taking the derivative of Equation 4:

$$\begin{aligned}\frac{\partial \mathcal{L}_Q}{\partial q_{ij}(y)} =& \log P(\boldsymbol{x}_{i \rightarrow j} \mid Y_{ij} = y; \boldsymbol{\theta}^{\rightarrow}) \\ & + \log P(\boldsymbol{x}_{i \leftarrow j} \mid Y_{ij} = y; \boldsymbol{\theta}^{\leftarrow}) \\ & + E_{Q(\boldsymbol{y}_{-(ij)})}[\log P(\boldsymbol{y} \mid Y_{ij} = y; \boldsymbol{\beta}, \boldsymbol{\eta})] \\ & - \log q_{ij}(y) - 1. \end{aligned} \tag{5}$$

After adding a Lagrange multiplier to ensure that $\sum_y q_{ij}(y) = 1$, we obtain a closed-form solution for each $q_{ij}(y)$. These iterative updates to q_{ij} can be viewed as a form of mean field inference (Wainwright and Jordan, 2008).

3.2 M-step

In the general case, the maximum expected likelihood solution for the content parameter $\boldsymbol{\theta}$ is given by the expected counts,

$$\boldsymbol{\theta}_y^{\rightarrow} \propto \sum_{\langle i,j\rangle \in G} q_{ij}(y) \boldsymbol{x}_{i \rightarrow j} \tag{6}$$

$$\boldsymbol{\theta}_y^{\leftarrow} \propto \sum_{\langle i,j\rangle \in G} q_{ij}(y) \boldsymbol{x}_{i \leftarrow j}. \tag{7}$$

As noted above, we are often interested in special cases that require parameter tying, such as $\boldsymbol{\theta}_y^{\rightarrow} = \boldsymbol{\theta}_y^{\leftarrow}, \forall y$. This can be handled by simply computing expected counts across the tied parameters.

1. Initialize $Q(Y^{(t)})$ for each $t \in \{1 \dots T\}$
2. Iterate until convergence:

   E-step update each q_{ij} in closed form, based on Equation 5.

   M-step: content Update $\boldsymbol{\theta}$ in closed form from Equations 6 and 7.

   M-step: structure Update $\boldsymbol{\beta}, \boldsymbol{\eta}$, and c by applying L-BFGS to the noise-contrastive estimation objective in Equation 8.

Table 1: Expectation-maximization estimation procedure

Obtaining estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ is more challenging, as it would seem to involve computing the partition function $Z(\boldsymbol{\eta}, \boldsymbol{\beta}; G)$, which sums over all possible labeling of each network $G^{(t)}$. The number of such labelings is exponential in the number of edges in the network. West et al. (2014) show that for an objective function involving features on triads and dyads, it is NP-hard to find even the single optimal labeling.

We therefore apply noise-contrastive estimation (NCE; Gutmann and Hyvärinen, 2012), which transforms the problem of estimating the density $P(\boldsymbol{y})$ into a classification problem: distinguishing the observed graph labelings $\boldsymbol{y}^{(t)}$ from randomly-generated "noise" labelings $\tilde{\boldsymbol{y}}^{(t)} \sim P_n$, where P_n is a noise distribution. NCE introduces an additional parameter c for the partition function, so that $\log P(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{\eta}, c) = \log P^0(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{\eta}) + c$, with $P^0(\boldsymbol{y})$ representing the unnormalized probability of $\boldsymbol{y}$. We can then obtain the NCE objective by writing $D = 1$ for the case that $\boldsymbol{y}$ is drawn from the data distribution and $D = 0$ for the case that $\boldsymbol{y}$ is drawn from the noise distribution,

$$\begin{aligned} J_{NCE}(\boldsymbol{\eta}, \boldsymbol{\beta}, c) \\ = \sum_t \log P(D = 1 \mid \boldsymbol{y}^{(t)}; \boldsymbol{\eta}, \boldsymbol{\beta}, c) \\ - \log P(D = 0 \mid \tilde{\boldsymbol{y}}^{(t)}; \boldsymbol{\eta}, \boldsymbol{\beta}, c), \end{aligned} \quad (8)$$

where we draw exactly one noise instance $\tilde{\boldsymbol{y}}$ for each true labeling $\boldsymbol{y}^{(t)}$.

Because we are working in an unsupervised setting, we do not observe $\boldsymbol{y}^{(t)}$, so we cannot directly compute the log probability in Equation 8. Instead, we compute the expectations of the relevant log probabilities, under the distribution $Q(\boldsymbol{y})$,

$$\begin{aligned} E_Q[\log P^0(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{\eta})] = \\ \sum_{\langle i,j \rangle \in G} \sum_y q_{ij}(y) \boldsymbol{\eta}^\top \boldsymbol{f}(y, i, j, G) \\ + \sum_{k: \langle i,j,k \rangle \in \mathcal{T}(G)} \sum_{y,y',y''} q_{ij}(y) q_{jk}(y') q_{ik}(y'') \beta_{y,y',y''}. \end{aligned} \quad (9)$$

We define the noise distribution P_n by sampling edge labels y_{ij} from their empirical distribution under $Q(\boldsymbol{y})$. The expectation $E_q[\log P_n(\boldsymbol{y})]$ is therefore simply the negative entropy of this empirical distribution, multiplied by the number of edges in G. We then plug in these expected log-probabilities to the noise-contrastive estimation objective function, and take derivatives with respect to the parameters $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and c. In each iteration of the M-step, we optimize these parameters using L-BFGS (Liu and Nocedal, 1989).

4 Identifying address terms in dialogue

The model described in the previous sections is applied in a study of the social meaning of **address terms** — terms for addressing individual people — which include:

Names such as *Barack*, *Barack Hussein Obama*.

Titles such as *Ms., Dr., Private, Reverend*. Titles can be used for address either by preceding a name (e.g., *Colonel Kurtz*), or in isolation (e.g., *Yes, Colonel.*).

Placeholder names such as *dude* (Kiesling, 2004), *bro*, *brother*, *sweetie*, *cousin*, and *asshole*. These terms can be used for address only in isolation (for example, in the address *cousin Sue*, the term *cousin* would be considered a title).

Because address terms connote varying levels of formality and familiarity, they play a critical role in establishing and maintaining social relationships. However, we find no prior work on automatically identifying address terms in dialogue transcripts. There are several subtasks: (1) distinguishing addresses from mentions of other individuals, (2) identifying a lexicon of titles, which either precede name addresses or can be used in isolation, (3) identifying

Text:	I	'm	not	Mr.	Lebowski	;	you	're	Mr.	Lebowski	.
POS:	PRP	VBP	RB	NNP	NNP	:	PRP	VBP	NNP	NNP	.
Address:	O	O	O	B-ADDR	L-ADDR	O	O	O	B-ADDR	L-ADDR	O

Figure 1: Automatic re-annotation of dialogue data for address term sequences

Feature	Description
Lexical	The word to be tagged, and its two predecessors and successors, $w_{i-2:i+2}$.
POS	The part-of-speech of the token to be tagged, and the POS tags of its two predecessors and successors.
Case	The case (lower, upper, or title) of the word to be tagged, and its two predessors and successors.
Constituency parse	First non-NNP ancestor node of the word w_i in the constituent parse tree, and all leaf node siblings in the tree.
Dependency parse	All dependency relations involving w_i.
Location	Distance of w_i from the start and the end of the sentence or turn.
Punctuation	All punctuation symbols occurring before and after w_i.
Second person pronoun	All forms of the second person pronoun within the sentence.

Table 2: Features used to identify address spans

a lexicon of placeholder names, which can only be used in isolation. We now present a tagging-based approach for performing each of these subtasks.

We build an automatically-labeled dataset from the corpus of movie dialogues provided by Danescu-Niculescu-Mizil and Lee (2011); see Section 6 for more details. This dataset gives the identity of the speaker and addressee of each line of dialogue. These identities constitute a minimal form of manual annotation, but in many settings, such as social media dialogues, they could be obtained automatically. We augment this data by obtaining the first (given) and last (family) names of each character, which we mine from the website `rottentomatoes.com`. Next, we apply the CoreNLP part-of-speech tagger (Manning et al., 2014) to identify sequences of the NNP tag, which indicates a proper noun in the Penn Treebank Tagset (Marcus et al., 1993). For each NNP tag sequence that contains the name of the addressee, we label it as an address, using BILOU notation (Ratinov and Roth, 2009): **B**eginning, **I**nside, and **L**ast term of address segments; **O**utside and **U**nit-length sequences. An example of this tagging scheme is shown in Figure 1.

Next, we train a classifier (Support Vector Machine with a linear kernel) on this automatically labeled data, using the features shown in Table 2. For simplicity, we do not perform structured prediction, which might offer further improvements in accuracy. This classifier provides an initial, partial solution to the first problem, distinguishing second-person addresses from references to other individuals (for name references only). On heldout data, the classifier's macro-averaged F-measure is 83%, and its micro-averaged F-measure is 98.7%. Class-by-class breakdowns are shown in Table 3.

4.1 Address term lexicons

To our surprise, we were unable to find manually-labeled lexicons for either titles or placeholder names. We therefore employ a semi-automated approach to construct address term lexicons, bootstrapping from the address term tagger to build candidate lists, which we then manually filter.

Titles To induce a lexicon of titles, we consider terms that are frequently labeled with the tag B-ADDR across a variety of dialogues, performing a binomial test to obtain a list of terms whose frequency of being labeled as B-ADDR is significantly higher than chance. Of these 34 candidate terms, we manually filter out 17, which are mainly common first names, such as *John*; such names are frequently labeled as B-ADDR across movies. After this manual filtering, we obtain the following titles: *agent, aunt, captain, colonel, commander, cousin, deputy, detective, dr, herr, inspector, judge, lord, master, mayor, miss, mister, miz, monsieur, mr, mrs, ms, professor, queen, reverend, sergeant, uncle.*

Placeholder names To induce a lexicon of placeholder names, we remove the CURRENT-WORD feature from the model, and re-run the tagger on all dialogue data. We then focus on terms which are frequently labeled U-ADDR, indicating that they are the sole token in the address (e.g., *I'm*/O *perfectly*/O *calm*/O, *dude*/U-ADDR.) We again perform a binomial test to obtain a list of terms whose frequency of being labeled U-ADDR is significantly higher than chance. We manually filter out 41 terms from a list of 96 possible placeholder terms obtained in the previous step. Most terms eliminated were plural forms of placeholder names, such as *fellas* and *dudes*; these are indeed address terms, but because they are plural, they cannot refer to a single individual, as required by our model. Other false positives were fillers, such as *uh* and *um*, which were ocassionally labeled as I-ADDR by our tagger. After manual filtering, we obtain the following placeholder names: *asshole, babe, baby, boss, boy, bro, bud, buddy, cocksucker, convict, cousin, cowboy, cunt, dad, darling, dear, detective, doll, dude, dummy, father, fella, gal, ho, hon, honey, kid, lad, lady, lover, ma, madam, madame, man, mate, mister, mon, moron, motherfucker, pal, papa, partner, peanut, pet, pilgrim, pop, president, punk, shithead, sir, sire, son, sonny, sport, sucker, sugar, sweetheart, sweetie, tiger.*

4.2 Address term tokens

When constructing the content vectors $\boldsymbol{x}_{i \rightarrow j}$ and $\boldsymbol{x}_{i \leftarrow j}$, we run the address span tagger described above, and include counts for the following types of address spans:

- the bare first name, last name, and complete name of individual j;
- any element in the title lexicon if labeled as B-ADDR by the tagger;
- any element in the title or placeholder lexicon, if labeled as U-ADDR by the tagger.

5 Address terms in a model of formality

Address terms play a key role in setting the formality of a social interaction. However, understanding this role is challenging. While some address terms, like *Ms* and *Sir*, are frequent, there is a long tail of rare terms whose meaning is more difficult to ascertain from data, such as *admiral*, *dude*, and *player*. Moreover, the precise social meaning of address terms can be context-dependent: for example, the term *comrade* may be formal in some contexts, but jokingly informal in others.

Class	F-measure	Total Instances
I-ADDR	0.58	53
B-ADDR	0.800	483
U-ADDR	0.987	1864
L-ADDR	0.813	535
O-ADDR	0.993	35975

Table 3: Breakdown of f-measure and number of instances by class in the test set.

Both problems can be ameliorated by adding social network structure. We treat $Y = \text{V}$ as indicating formality and $Y = \text{T}$ as indicating informality. (The notation invokes the concept of T/V systems from politeness theory (Brown, 1987), where T refers to the informal Latin second-person pronoun *tu*, and V refers to the formal second-person pronoun *vos*.)

While formality relations are clearly asymmetric in many settings, for simplicity we assume symmetric relations: each pair of individuals is either on formal or informal terms with each other. We therefore add the constraints that $\boldsymbol{\theta}^{\leftarrow}_{\text{V}} = \boldsymbol{\theta}^{\rightarrow}_{\text{V}}$ and $\boldsymbol{\theta}^{\leftarrow}_{\text{T}} = \boldsymbol{\theta}^{\rightarrow}_{\text{T}}$. In this model, we have a soft expectation that triads will obey transitivity: for example, if i and j have an informal relationship, and j and k have an informal relationship, then i and k are more likely to have an informal relationship. After rotation, there are four possible triads, TTT, TTV, TVV, and VVV. The weights estimated for these triads will indicate whether our prior expectations are validated. We also consider a single pairwise feature template, a metric from Adamic and Adar (2003) that sums over the mutual friends of i and j, assigning more weight to mutual friends who themselves have a small number of friends:

$$AA(i,j) = \sum_{k \in \Gamma(i) \cap k \in \Gamma(j)} \frac{1}{\log \#|\Gamma(k)|}, \qquad (10)$$

where $\Gamma(i)$ is the set of friends of node i. (We also tried simply counting the number of mutual friends, but the Adamic-Adar metric performs

slightly better.) This feature appears in the vector $\boldsymbol{f}(y_{ij}, i, j, G)$, as defined in Equation 3.

6 Application to movie dialogues

We apply the ideas in this paper to a dataset of movie dialogues (Danescu-Niculescu-Mizil and Lee, 2011), including roughly 300,000 conversational turns between 10,000 pairs of characters in 617 movies. This dataset is chosen because it not only provides the script of each movie, but also indicates which characters are in dialogue in each line. We evaluate on quantitative measures of predictive likelihood (a token-level evaluation) and coherence of the induced address term clusters (a type-level evaluation). In addition, we describe in detail the inferred signed social networks on two films.

We evaluate the effects of three groups of features: address terms, mutual friends (using the Adamic-Adar metric), and triads. We include address terms in all evaluations, and test whether the network features improve performance. Ablating both network features is equivalent to clustering dyads by the counts of address terms, but all evaluations were performed by ablating components of the full model. We also tried ablating the text features, clustering edges using only the mutual friends and triad features, but we found that the resulting clusters were incoherent, with no discernible relationship to the address terms.

6.1 Predictive log-likelihood

To compute the predictive log-likelihood of the address terms, we hold out a randomly-selected 10% of films. On these films, we use the first 50% of address terms to estimate the dyad-label beliefs $q_{ij}(y)$. We then evaluate the expected log-likelihood of the second 50% of address terms, computed as $\sum_y q_{ij}(y) \sum_n \log P(x_n \mid \theta_y)$ for each dyad. This is comparable to standard techniques for computing the held-out log-likelihood of topic models (Wallach et al., 2009).

As shown in Table 4, the full model substantially outperforms the ablated alternatives. This indicates that the signed triad features contribute meaningful information towards the understanding of address terms in dialogue.

Address terms	Mutual friends	Signed triads	Log-likelihood
✓			-2133.28
✓		✓	-2018.21
✓	✓		-1884.02
✓	✓	✓	-1582.43

Table 4: Predictive log-likelihoods.

V-cluster	T-cluster
sir	FIRSTNAME
mr+LASTNAME	*man*
mr+FIRSTNAME	*baby*
mr	*honey*
miss+LASTNAME	*darling*
son	*sweetheart*
mister+FIRSTNAME	*buddy*
mrs	*sweetie*
mrs+LASTNAME	*hon*
FIRSTNAME+LASTNAME	*dude*

Table 5: The ten strongest address terms for each cluster, sorted by likelihood ratio.

6.2 Cluster coherence

Next, we consider the model inferences that result when applying the EM procedure to the entire dataset. Table 5 presents the top address terms for each cluster, according to likelihood ratio. The cluster shown on the left emphasizes full names, titles, and formal address, while the cluster on the right includes the given name and informal address terms such as *man*, *baby*, and *dude*. We therefore use the labels "V-cluster" and "T-cluster", referring to the formal and informal clusters, respectively.

We perform a quantitative evaluation of this clustering through an intrusion task (Chang et al., 2009). Specifically, we show individual raters three terms, selected so that two terms are from the same cluster, and the third term is from the other cluster; we then ask them to identify which term is least like the other two. Five raters were each given a list of forty triples, with the order randomized. Of the forty triples, twenty were from our full model, and twenty were from a text-only clustering model. The raters agreed with our full model in 73% percent of cases, and agreed with the text-only model in 52% percent

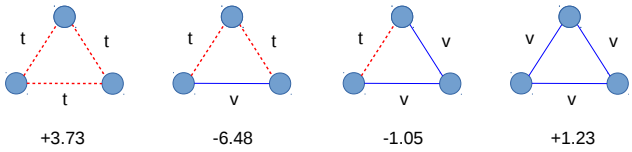

Figure 2: Estimated triad feature weights

of cases. By Fisher's exact test, this difference is statistically significant at $p < 0.01$. Both results are significantly greater than chance agreement (33%) by a binomial test, $p < 0.001$.

6.3 Network feature weights

Figure 2 shows the feature weights for each of the four possible triads. Triads with homogeneous signs are preferred, particularly TTT (all informal); heterogeneous triads are dispreferred, particularly TTV, which is when two individuals have a formal relationship despite having a mutual informal tie. Less dispreferred is TVV, when a pair of friends have an informal relationship despite both having a formal relationship with a third person; consider, for example, the situation of two students and their professor. In addition, the informal sign is preferred when the dyad has a high score on the Adamic-Adar metric, and dispreferred otherwise. This coheres with the intuition that highly-embedded edges are likely to be informal, with many shared friends.

6.4 Qualitative results

Analysis of individual movies suggests that the induced tie signs are meaningful and coherent. For example, the film "Star Wars" is a space opera, in which the protagonists Luke, Han, and Leia attempt to defeat an evil empire led by Darth Vader. The induced signed social network is shown in Figure 3. The V-edges seem reasonable: C-3PO is a robotic servant, and Blue Leader is Luke's military commander (BLUE LEADER: *Forget it, son.* LUKE: *Yes, sir, but I can get him...*). In contrast, the character pairs with T-edges all have informal relationships: the lesser-known character Biggs is Luke's more experienced friend (BIGGS: *That's no battle, kid*).

The animated film "South Park: Bigger, Longer & Uncut" centers on three children: Stan, Cartman, and Kyle; it also involves their parents, teachers, and friends, as well as a number of political and religious figures. The induced social network is shown in Figure 4. The children and their associates mostly have T-edges, except for the edge to Gregory, a British character with few speaking turns. This part of the network also has a higher clustering coefficient, as the main characters share friends such as Chef and The Mole. The left side of the diagram centers on Kyle's mother, who has more formal relationships with a variety of authority figures.

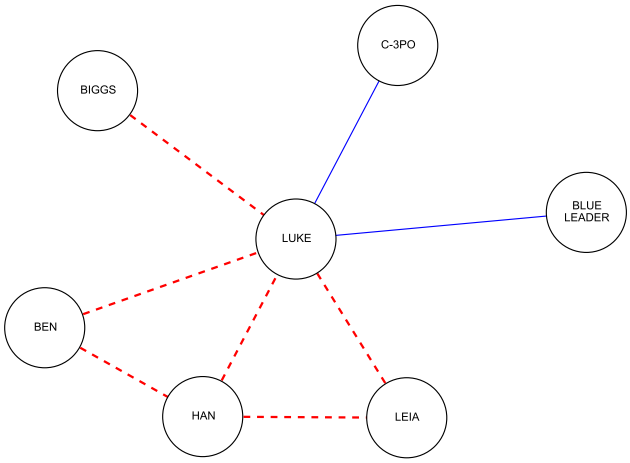

Figure 3: Induced signed social network from the film *Star Wars*. Blue solid edges are in the V-cluster, red dashed edges are in the T-cluster.

7 Related work

Recent work has explored the application of signed social network models to social media. Leskovec et al. (2010b) find three social media datasets from which they are able to identify edge polarity; this enables them to compare the frequency of signed triads against baseline expectations, and to build a classifier to predict edge labels (Leskovec et al., 2010a). However, in many of the most popular social media platforms, such as Twitter and Facebook, there is no metadata describing edge labels. We are also interested in new applications of signed social network analysis to datasets outside the realm of social media, such as literary texts (Moretti, 2005; Elson et al., 2010; Agarwal et al., 2013) and movie scripts, but in such corpora, edge labels are not easily available.

In many datasets, it is possible to obtain the textual content exchanged between members of the network, and this content can provide a signal for network structure. For example, Hassan et al. (2012) characterize the sign of each network edge in terms

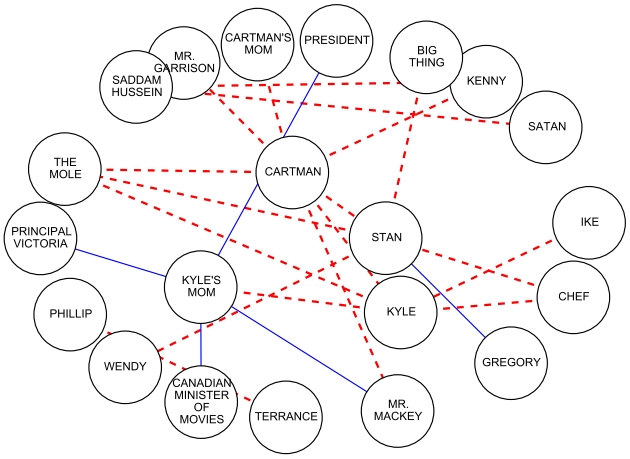

Figure 4: Induced signed social network from the film *South Park: Bigger, Longer & Uncut*. Blue solid edges are in the V-cluster, red dashed edges are in the T-cluster.

of the **sentiment** expressed across it, finding that the resulting networks cohere with the predictions of structural balance theory; similar results are obtained by West et al. (2014), who are thereby able to predict the signs of unlabeled ties. Both papers leverage the relatively mature technology of sentiment analysis, and are restricted to edge labels that reflect sentiment. The unsupervised approach presented here could in principle be applied to lexicons of sentiment terms, rather than address terms, but we leave this for future work.

The issue of address formality in English was considered by Faruqui and Padó (2011), who show that annotators can label the formality of the second person pronoun with agreement of 70%. They use these annotations to train a supervised classifier, obtaining comparable accuracy. If no labeled data is available, annotations can be projected from languages where the T/V distinction is marked in the morphology of the second person pronoun, such as German (Faruqui and Padó, 2012). Our work shows that it is possible to detect formality without labeled data or parallel text, by leveraging regularities across network structures; however, this requires the assumption that the level of formality for a pair of individuals is constant over time. The combination of our unsupervised approach with annotation projection might yield models that attain higher performance while capturing change in formality over time.

More broadly, a number of recent papers have proposed to detect various types of social relationships from linguistic content. Of particular interest are power relationships, which can be induced from n-gram features (Bramsen et al., 2011; Prabhakaran et al., 2012) and from **coordination**, where one participant's linguistic style is asymmetrically affected by the other (Danescu-Niculescu-Mizil et al., 2012). Danescu-Niculescu-Mizil et al. (2013) describe an approach to recognizing politeness in text, lexical and syntactic features motivated by politeness theory. Anand et al. (2011) detect "rebuttals" in argumentative dialogues, and Hasan and Ng (2013) employ extra-linguistic structural features to improve the detection of stances in such debates. In all of these cases, labeled data is used to train supervised model; our work shows that social structural regularities are powerful enough to support accurate induction of social relationships (and their linguistic correlates) without labeled data.

8 Conclusion

This paper represents a step towards unifying theoretical models of signed social network structures with linguistic accounts of the expression of social relationships in dialogue. By fusing these two phenomena into a joint probabilistic model, we can induce edge types with robust linguistic signatures and coherent structural properties. We demonstrate the effectiveness of this approach on movie dialogues, where it induces symmetric T/V networks and their linguistic signatures without supervision. Future work should evaluate the capability of this approach to induce asymmetric signed networks, the utility of partial or distant supervision, and applications to non-fictional dialogues.

Acknowledgments

We thank the reviewers for their detailed feedback. The paper benefitted from conversations with Cristian Danescu-Niculescu-Mizil, Chris Dyer, Johan Ugander, and Bob West. This research was supported by an award from the Air Force Office of Scientific Research, and by Google, through a Focused Research Award for Computational Journalism.

References

Lada A Adamic and Eytan Adar. 2003. Friends and neighbors on the web. *Social networks*, 25(3):211–230.

Apoorv Agarwal, Anup Kotalwar, and Owen Rambow. 2013. Automatic extraction of social networks from literary text: A case study on alice in wonderland. In *the Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 1–9, Portland, Oregon, June. Association for Computational Linguistics.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 773–782, Portland, OR.

Roger Brown and Marguerite Ford. 1961. Address in american english. *The Journal of Abnormal and Social Psychology*, 62(2):375.

Penelope Brown. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge University Press.

Dorwin Cartwright and Frank Harary. 1956. Structural balance: a generalization of heider's theory. *Psychological review*, 63(5):277.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems (NIPS)*, pages 288–296.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 699–708, Lyon, France.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 250–259, Sophia, Bulgaria.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

David K Elson, Nicholas Dames, and Kathleen R McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 138–147, Uppsala, Sweden.

Manaal Faruqui and Sebastian Padó. 2011. "I Thou Thee, Thou Traitor": Predicting Formal vs. Informal Address in English Literature. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 467–472, Portland, OR.

Manaal Faruqui and Sebastian Padó. 2012. Towards a model of formal and informal address in english. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 623–633.

Michael U Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The Journal of Machine Learning Research*, 13(1):307–361.

Kazi Saidul Hasan and Vincent Ng. 2013. Extra-linguistic constraints on stance recognition in ideological debates. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 816–821, Sophia, Bulgaria.

Ahmed Hassan, Amjad Abu-Jbara, and Dragomir Radev. 2012. Extracting signed social networks from text. In *Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing*, pages 6–14. Association for Computational Linguistics.

Dell Hymes. 1972. On communicative competence. *Sociolinguistics*, pages 269–293.

Scott F Kiesling. 2004. Dude. *American Speech*, 79(3):281–305.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010a. Predicting positive and negative links in online social networks. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 641–650.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010b. Signed networks in social media. In *Proceedings of Human Factors in Computing Systems (CHI)*, pages 1361–1370.

David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.

Dong C Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Franco Moretti. 2005. *Graphs, maps, trees: abstract models for a literary history*. Verso.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012. Predicting overt display of power in written dialogs. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 518–522.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Martin J Wainwright and Michael I Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1105–1112.

Robert West, Hristo Paskov, Jure Leskovec, and Christopher Potts. 2014. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics*, 2:297–310.