Determining Team Hierarchy from Broadcast Communications

Anup K. Kalia[1], Norbou Buchler[2], Diane Ungvarsky[2], Ramesh Govindan[3], and Munindar P. Singh[1]

[1] North Carolina State University, Raleigh, NC 27695, USA
[2] US Army Research Lab, Aberdeen Proving Ground, MD 21005, USA
[3] University of Southern California, Los Angeles, CA 90089, USA

Abstract. Broadcast chat messages among team members in an organization can be used to evaluate team coordination and performance. Intuitively, a well-coordinated team should reflect the team hierarchy, which would indicate that team members assigned with particular roles are performing their jobs effectively. Existing approaches to identify hierarchy are limited to data from where graphs can be extracted easily. We contribute a novel approach that takes as input broadcast messages, extracts communication patterns—as well as semantic, communication, and social features—and outputs an organizational hierarchy. We evaluate our approach using a dataset of broadcast chat communications from a large-scale Army exercise for which ground truth is available. We further validate our approach on the Enron corpus of corporate email.

1 Introduction

In an organization, a team is a purposeful social system created to get work done. Therefore, it is important to understand and characterize the degree to which team members coordinate with each other. In most organizations, a team hierarchy exists among the team members wherein a higher ranking team member sets high-level goals, and guides or motivates lower ranking team members, who are expected to carry out such commands. Although team members have clearly delineated roles, it is important to evaluate whether they are performing their jobs well or whether the team needs restructuring. One important factor for evaluating team performance is communication between team members. Eaton [4] provides insight that communication is essential for team members to build their inter-personal relationships which indirectly enhance team performance. Leonard and Frankel [13] describe that for effective teamwork communication is important because it creates predictability and agreement between team members. Resick et al. [17] suggest that information elaboration is important in evolving teams to maintain team performance. Our premise is that we can determine such indicators of organizational effectiveness and team member performance from members' communications, such as chats and emails, which provide an account of actual behavior while being unobtrusive.

Several works have identified team hierarchies from graphs extracted from online social networks such as Twitter, Flickr, Prison, and Wikivote [8, 14, 15] and text such as emails and short message service (SMS) communications [7, 18, 21]. Gupte et al. [8] and Enys et al. [14, 15] provide hierarchical measures called social *agony* and *global reaching centrality* (GRC), respectively, to extract hierarchies from online social networks. Rowe et al. [18] extract an undirected graph from Enron emails [5, 10] based on the number of emails exchanged between Enron employees whereas Wang et al. [21] compute hierarchy from Enron emails as well as from call and SMS data. Gilbert [7] emphasized analyzing text content to extract phrases that indicate hierarchy. The above works apply when social graphs can be extracted, such as from online social networks and directed messages (emails and SMS). However, these approaches do not apply for broadcast messages, where the receiver is not clear.

Our approach takes in broadcast messages recorded from a multiparty event and produces a team hierarchy among the participants. The basis of our approach is to identify communication patterns from messages that indicate a possible team hierarchy. Broadly, we identify three patterns: *directive*, *question*, and *informative*. We select these patterns based on the existing literature [7, 16] and the fact that they occur frequently in broadcast messages. The overall approach approximates Gilbert [7]. Whereas his approach identifies communication content that indicates power and hierarchy, we additionally compute the ranks and validate our approach versus ground truth. Also, Gilbert's approach is domain-dependent, whereas our approach is domain-independent and applies to broadcast as well as directed communications.

We analyze *semantic*, *communication*, and *social* features that can be extracted from messages to compute hierarchy. Semantic features include *responses* to communication patterns and *emotions* expressed in responses features extracted from text content. Communication features include the *average response time delay* and *messages* sent features. Social features include the *degree centrality* and *betweenness centrality* features. We hypothesize that semantic features, which capture the meaning of interactions, are better indicators of hierarchy than social features, which merely capture network statistics.

To identify the patterns, we select two chat rooms from a military exercise dataset. We use one chat room to refine our methods to identify patterns and test our method on the second chat room, obtaining an F-measure of 83% for identifying the patterns. From the patterns identified, we collect the features described above. Using these features we determine participants' ranks computed via hierarchical clustering. We evaluate our results against actual known ranks. In addition, we evaluate the generalizability of our approach to directed communications, as in Enron email corpus. In directed communications, emails exchanged between senders and receivers provide good indicators of hierarchy.

We find that for the chat corpus the accuracy in identifying ranks using the *informative* pattern is significantly higher than for the *directive* and *question* pattern. Additionally, we find that semantic features along with communication features are better indicators of hierarchy than social features. For Enron, we

obtain similar results regarding the identification of patterns though we find that social features are better indicators of hierarchy than semantic features, possibly because compared to the military dataset, the Enron corpus is much larger with more participants and messages. And it may be that in such a large corporate organization, the roles, responsibilities, and influence need to be ascertained socially. Also, compared to participants in Enron, participants in military communication networks have well-defined functional roles and prescribed work flows that lead to more structured communication and hence, semantic features may perform better than social features.

2 Communication Patterns in Broadcast Messages

Broadcast messages are sent by participants in a group and hence, everyone in a group can see and respond to messages. Before we infer a hierarchy from broadcast messages it is important to understand what each message means. For example, a message can indicate different illocutions [1] such as directives and commissives. Based on the literature [7, 16] and our preliminary analysis, i.e., manually finding the distributions of meanings of the messages in the military dataset, we hypothesize that hierarchical information can be extracted from messages via three communication patterns: directive, question, and informative. A *directive* is an order or request; a *question* is an inquiry; an *informative* is a report. *Directives* and *questions* correlate with the sender having a higher rank than the receiver; *informatives* the reverse.

An important challenge in dealing with broadcast messages is that the recipient of a message is not clear. To tackle the challenge, we define a *window* $\mathcal{W}$ consisting of two consecutive messages where we assume that the second message $\mathcal{W}^{next}$ is a response to the first message $\mathcal{W}^{curr}$. The two messages must occur in the same chat room and have different senders. A window $\mathcal{W}$ is instantiated as a *directive*, *question*, or *informative* pattern if, respectively, $\mathcal{W}^{curr}$ is a *directive*, *question*, or *informative* and correspondingly $\mathcal{W}^{next}$ is an acknowledgment, response, or acknowledgment. Table 1 provides examples of these patterns from military data.

Table 1. Examples of communication patterns from military chat data.

Window	Sender	Messages	Pattern
$\mathcal{W}_a^{curr}$	8_6i_256_s3	Cos, send all reports up to BN over this net	*Directive*
$\mathcal{W}_a^{next}$	8_6i_256_b_cdr	rgr	
$\mathcal{W}_b^{curr}$	8_6i_256_s3	B, whats your status on personnel?	*Question*
$\mathcal{W}_b^{next}$	8_6i_256_b_cdr	no casualties	
$\mathcal{W}_b^{curr}$	8_6i_256_b_cdr	have been engaging with SAF and MTRs with no effect	*Informative*
$\mathcal{W}_b^{next}$	8_6i_256_cdr	ack, keep me posted	

3 Process

Figure 1 shows the process we follow. In the process, we separately consider the *directive*, *question*, *informative* patterns as well as the combination of *directive* and *question* patterns to compute ranks. Next, we evaluate the accuracy of the ranks computed based on different patterns.

As an illustration, consider computing ranks using *directive* patterns. For each participant P in chat messages we extract the following features. First, we extract *directive* patterns $\mathcal{W}$ where $\mathcal{W}^{curr}$ indicates a directive message and P is the sender of $\mathcal{W}^{next}$. From the patterns, we assume that P responds to $\mathcal{W}^{curr}$ and hence, we calculate the total number of such responses to directives for P.

Second, we determine whether $\mathcal{W}^{next}$ indicates a positive, neutral, or negative emotion. We extract emotions because we hypothesize that they can be indicators of hierarchy. For example, P may be a team leader and may display positive emotions to motivate subordinates or P may be a subordinate and may express emotions with respect to outcome of his or her actions. We include responses to patterns and emotions within semantic features.

Third, based on the patterns $\mathcal{W}$ we find the *average response time delay*, i.e., the average of the time lags between $\mathcal{W}^{curr}$ and $\mathcal{W}^{next}$ extracted for P. Fourth, we find the *number of messages* that P broadcasts. We include the average response time delay and number of messages broadcast as communication features.

From the patterns $\mathcal{W}$ we create a graph that contains directed edges from responders (P) to respondees. Using the graph, we compute social features for P, i.e., P's *degree centrality* and *betweenness centrality* [2, 6]. We aggregate all features—semantic, communication, and social—for P. We repeat the feature extraction for all participants P^*. Finally, based on P^*s' features we compute hierarchical ranks for each P. We evaluate computed ranks against the ground truth of actual ranks. We carry out the above process for the informative and question patterns.

Prior works [8, 15, 18] focus primarily on social and communication features to compute ranks whereas we include semantic features based on the intuition that semantic features, being based on the message content, can reveal important hierarchical information. Below, we discuss the extraction of features in detail.

3.1 Extracting Semantic Features

To extract semantic features for each participant, first, we identify patterns $\mathcal{W}$. To identify patterns, we create a rule-based approach using training data and evaluate it on a test data. Both training and test data consist of broadcast messages labeled *directive*, *question*, or *informative*. To support our rules, for each dataset, we build a domain-specific lexicon of action verbs that includes words occurring frequently in the data.

3.1.1 Extracting Responses to Directives To extract a response to a *directive*, we determine if a message $\mathcal{W}^{curr}$ in $\mathcal{W}$ indicates a *directive*. To do so,

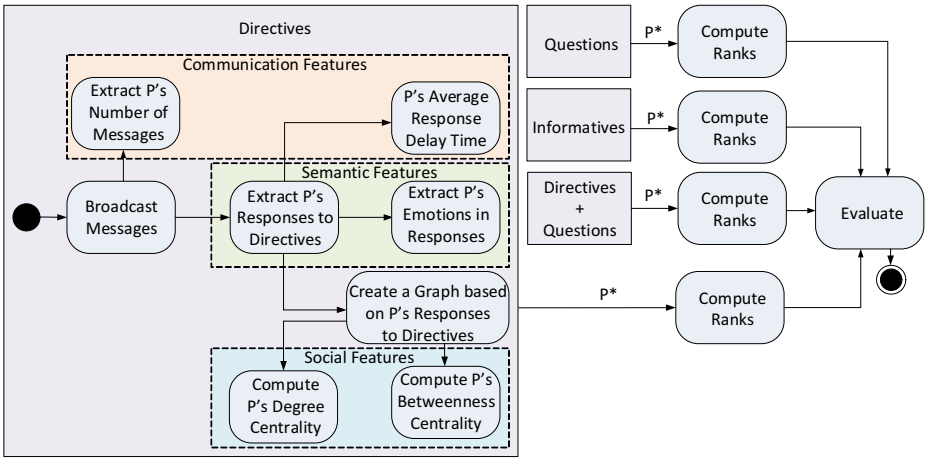


Fig. 1. Process followed to compute ranks with respect to directive, question, and informative patterns, and directives and questions combined for participants (P*) from broadcast messages.

we parse a message $\mathcal{W}^{curr}$ using the Stanford Natural Language Parser [9] and extract a parse tree. Figure 2 represents a parse tree for a sample message "Cos, send all reports up to BN over this net." In the parse tree, first, we look for a verb phrase (VP) indicated by the shading in Figure 2. Then, in the VP we look for an action verb (VB). If the action verb matches a verb in our domain-specific lexicon, we extract the rest, i.e., noun (NP) and prepositional phrase (PP), as shown in Figure 2. Hence, the words extracted from the example message are "send all reports up to BN over this net" which we identify as a *directive*. We assume the next message $\mathcal{W}^{next}$ is a response to the *directive* message.

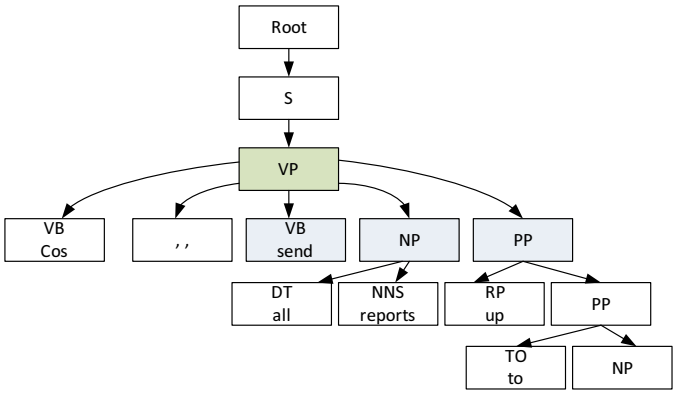


Fig. 2. Parse tree derived from "Cos, send all reports up to BN over this net" where Cos is the Chief of Staff position and BN is the Battalion.

3.1.2 Extracting Responses to Questions To extract a response to a *question*, we determine if a message $\mathcal{W}^{curr}$ in $\mathcal{W}$ indicates a *question*. If a message starts with a word such as *what*, *when*, *why*, *has*, *how*, *have*, and so on and ends with a *question mark* or if a message starts with a modal verb (MD) such as *will*, *shall*, *could*, *would*, *should*, and *can* followed by the word *you*, we mark the message as a *question*. If a message is identified as a *question*, we assume the next message $\mathcal{W}^{next}$ is a response to the *question* regardless of its grammar or content.

3.1.3 Extracting Responses to Informative To extract a response to an *informative*, we determine if a message $\mathcal{W}^{curr}$ in $\mathcal{W}$ indicates an *informative*. If a message begins with the following *rgr*, *Roger*, *ack*, *yes*, *yup*, *yep*, *okay*, *ok*, *thanks*, and so on we tag the message as the informative. Although some of the words (e.g., *Roger* and *ack*) are domain-specific, other words (*thanks*, *yes*, and *okay*) are domain independent. Such generic words make this pattern domain-independent. The next message $\mathcal{W}^{next}$ we assume is a response to the *informative* message.

For each participant, we calculate the count of all $\mathcal{W}^{next}$ or responses extracted for each pattern.

3.1.4 Extracting Emotions in Responses For each communication pattern $\mathcal{W}$, we determine if the response message $\mathcal{W}^{next}$ indicates an emotion, which could be positive, neutral, or negative. We use the Stanford Sentiment Parser [20], which computes the emotion corresponding to a message. For each participant, we compute the sums of the emotion polarities identified from response messages.

3.2 Extracting Communication Features

For each participant we extract two communication features. One, the number of messages sent by the participant and second, the *average response time delay* for a participant based on the messages that indicate *responses* to a pattern. The number of messages is a network statistic calculated independently of *responses* to patterns.

3.3 Extracting Social Features

To extract social features, we create a graph represented as an adjacency matrix $\mathcal{A}_{ij}$. In the matrix i and j represent the participants. An edge ij in $\mathcal{A}$ exists from the sender (responder) of $\mathcal{W}^{next}$ toward the sender (respondee) of $\mathcal{W}^{curr}$, if $\mathcal{W}^{next}$ indicates a response to a pattern, i.e., directive, question, or informative. If an edge ij exists, we mark $\mathcal{A}_{i,j} = 1$ else we mark $\mathcal{A}_{i,j} = 0$. We also mark $\mathcal{A}_{i,j} = 0$ if i equals j because we assume a sender does not respond to itself. We mark $\mathcal{A}_{i,j} = 1$ irrespective of one or more responses between i and j. From $\mathcal{A}_{i,j}$ we can

construct a directed graph $G(V, E)$ where V represents the participants and E represents the directed edge between the participants.

Using the directed graph $G(V, E)$ extracted from a pattern, we compute the social features of *degree centrality* and *betweenness centrality*. We consider these social features for two reasons. One, they have been used in the literature to interpret Rowe et al.'s [18] hierarchy. Two, we consider chatrooms that contain more intrateam messages than interteam messages, possibly, because we assume graphs derived from intrateam messages may be strongly connected than graphs derived from interteam messages. Our assumption is based on the notion that a chatroom mapping is not one-to-one direct and in general, people subscribe to chatrooms. In that sense the degree distribution is shared widely (observed) by all.

- **Degree centrality** is defined as the degree of a node or the number of edges directed to a node. The *degree centrality* $\mathrm{dc}(v_j)$ of a node v_j equals the number of edges ij directed to v_j, i.e., $\sum_i a_{ij}$ [2].
- **Betweenness centrality**, defined as the number of shortest paths passing through a node, is a measure of how important a node is. The *betweenness centrality* of a node v_j is calculated as $\sum_i \sum_k \frac{\delta_{ijk}}{\delta_{ik}}$ where δ_{ijk} is the number of shortest paths between i and k that include j and δ_{ik} is number of shortest paths between i and k [2, 6].

3.4 Computing Ranks

We compute ranks based on features extracted for participants. We adopt hierarchical clustering for two reasons. First, it being an unsupervised technique can be applied to datasets of any size. This is useful because we don't need to create a model from a large dataset and then use the model to produce predictions for a new dataset. Second, we want to infer a hierarchy among team members. The method helps cluster employees with similar rankings.

To compute ranks, we normalize all features extracted for each participant to the interval [0,100]. We construct a feature vector for each participant and use the *Euclidean distance* between them as a basis for hierarchical clustering. We plan to evaluate other distance metrics in future. We adopt the *single link* algorithm [19], which is a simple and popular technique. Figure 3 shows an example of a hierarchical cluster as a *single link dendrogram*. In Figure 3, d_1, ..., d_5 represent distances between the clusters. We assume that participants in the same cluster have the same rank. Next we provide rules to estimate rank orders between participants in clusters. We derive these rules by checking the consistency in rank outputs by applying the rules on multiple datasets.

Rank Rule 1 *For the directive and question patterns, increasing distance between clusters from bottom to top indicates decreasing rank.*

Rank Rule 2 *For the informative pattern, increasing distance between clusters from bottom to top indicates increasing rank.*

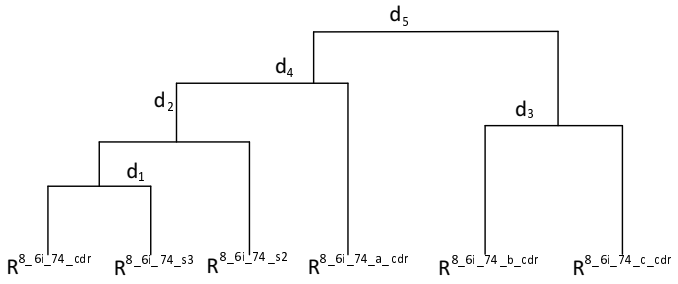


Fig. 3. An example of a single link dendrogram with distance d between clusters, applied to estimate rank R (bottom row).

4 Evaluation

We evaluate our approach primarily on our military broadcast chat dataset and secondarily on the Enron (directed) email dataset. The evaluation has two steps. First, we evaluate our methods to extract communication patterns, as described in Section 3.1. Second, we evaluate our estimation of ranks based on the patterns, as described in Section 3.4.

To evaluate the extraction of patterns we use the following metrics: precision, recall, and F-measure. Precision is given by $\frac{true\_positive}{true\_positive+false\_positive}$, recall by $\frac{true\_positive}{true\_positive+false\_negative}$, and F-measure by $\frac{2\times precision\times recall}{precision+recall}$. The mean absolute error (MAE) of a rank prediction is $\frac{\sum_i^N |predicted\_rank_i - actual\_rank_i|}{N}$. The accuracy of a rank prediction is $\frac{N-MAE}{N}$, where N is the highest rank.

4.1 Data Description

4.1.1 Military The military dataset was provided by the Mission Command Battle Lab at Fort Leavenworth, Kansas, and the US Army Research Laboratory, Maryland, from an Army simulation experiment (SIMEX). The dataset contains 20 chat rooms, on average, with 42 participants each and 6,998 messages. From the dataset, we consider the following chat rooms: Infantry Brigade Combat Team (IBCT), USMC Maneuver Brigade (MEB), Cavalry (CAV), and Commander (CDR) to evaluate our results. MEB has 546 messages and 50 participants, CAV has 481 messages and 48 participants, CDR has 409 messages and 37 messages, and IBCT Intel has 1027 messages and 64 participants. We consider these chat rooms because, first, they have more messages than the mean number of messages and, second, they have more intrateam messages than interteam messages.

The dataset includes the participants' actual ranks. (Rank 1 is the highest.) Table 2 shows the ranks of a few participants who sent more than one broadcast message and belong to a particular military team.

Some participant IDs in the dataset have OCR errors. For example, the ID 8_6i_256_s3 has spurious variants 8_61_256_s3 and 8_6i_256_53 in which i is substituted by 1 and s by 5, respectively. Such errors make it difficult to identify the IDs automatically. To handle such spurious IDs, we select participant IDs with the highest number of messages. For example, if 8_6i_256_s3, 8_61_256_s3, and 8_6i_256_53 have sent 25, 34, and 10 messages respectively, then for our evaluation we consider 8_61_256_s3 with 34 messages.

Table 2. Ranks of participants selected from the chat rooms.

Rank	MEB	CAV	CDR	IBCT Intel
1.	2meb_cdr	8_6i_74_cdr	8_6i_256_cdr	8_6i_s2
2.	2meb_s2	8_6i_74_s3	8_6i_256_s3	8_6i_156_s2
3.	2meb_s3	8_6i_74_s2	8_6i_256_s6	8_6i_256_s2
4.	2meb_fso	8_6i_74_fso	8_6i_256_fso	8_6i_256_s3
5.	2meb_mech_bn_cdr	8_6i_74_a_cdr	8_6i_256_alo	8_6i_35_s2
6.	2meb_mech_bn_s3	8_6i_74_b_cdr	8_6i_256_a_cdr	–
7.	2meb_mech2_bn_cdr	8_6i_74_c_cdr	8_6i_256_b_cdr	–
8.	2meb_helo_sqdn_cdr	8_6i_74_jtac	8_6i_256_c_cdr	–
9.	–	–	8_6i_256_wpn_cdr	–

4.1.2 Enron In the Enron email dataset [5, 10], we arbitrarily consider 62 employees who have sent 38,863 emails with a total of 360,708 email sentences. Prior to the evaluation, we obtain the actual ranks of these 62 employees [7]. The distribution of ranks from 0 to 6 is as follows: 8%, 2%, 29%, 11%, 6%, 36%, and 8%.

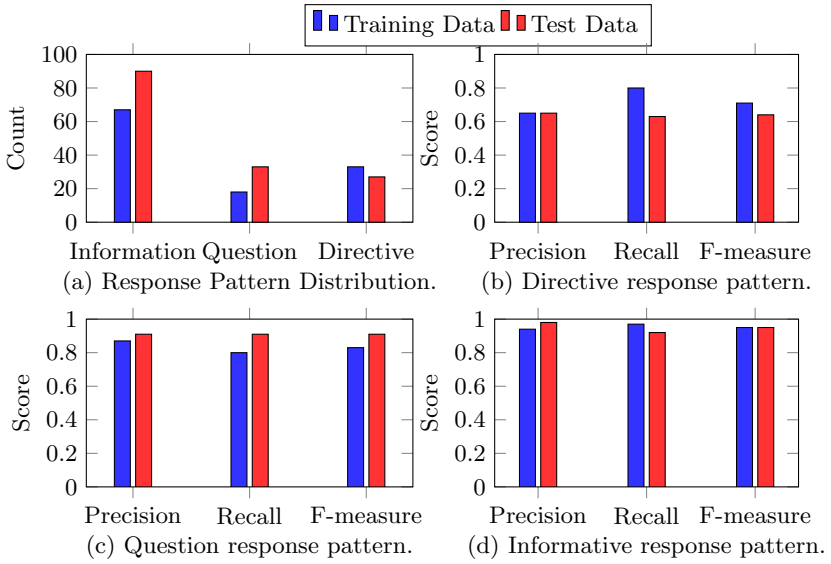
4.2 Results

We describe the results of our evaluation for extracting patterns and computing ranks on both the military chat dataset and the Enron dataset.

4.2.1 Extracting Communication Patterns We created the rule-based approach given in Section 3.1 using CDR (training data) and evaluated it on CAV (test data). Figure 4(a) shows distributions of the communication patterns in these datasets. Notice the high frequency of the *informative* pattern. Two raters (both graduate students in Computer Science) labeled the data with the various patterns. Their inter-rater agreement (*kappa* score [3]) was 0.76, which is fairly high. We arbitrarily selected one of the rater's assigned labels as the ground truth, because we cannot take the average. There are advanced approaches that

use Bayesian techniques to estimate a ground truth probability for each classification [12], but this is beyond the current scope and means of the paper.

Based on the training data, we constructed our rules, as described in Section 3.1, and evaluated them on the test data. For the training and test data, we found that the F-measures are respectively 0.71 and 0.64 (for the *directive* pattern), 0.83 and 0.91 (the *question* pattern), 0.95 for each (for the *informative* pattern), and 0.84 and 0.83 (overall). Considering the F-measure to identify different *patterns* as 0.83, we predicted the patterns for the dataset MEB and IBCT Intel.



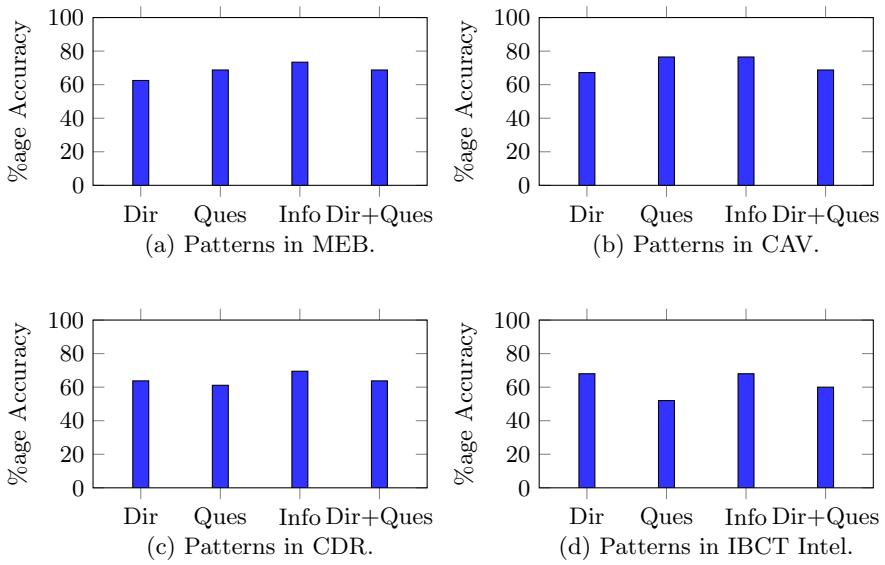
(a) Response Pattern Distribution. (b) Directive response pattern.

(c) Question response pattern. (d) Informative response pattern.

Fig. 4. Panel A: Distribution of response patterns. Panels B, C, D: F-measure scores for the response patterns (highest for the informative response pattern in Panel D).

4.2.1 Computing Ranks via Different Patterns We used the hierarchical clustering approach described in Section 3.4 to compute ranks. Specifically, we considered eight features F_1 to F_8 extracted for each pattern. F_1 represents the counts of *responses* to patterns, i.e., either *directive*, *question*, or *informative*; F_2, F_3, and F_4 represent the number of negative, neutral, and positive *emotions*, respectively; F_5 represents the *average response time delay*; F_6 represents the number of *messages* sent; F_7 represents the *degree centrality*; and F_8 represents the *betweenness centrality*. Since the *directive* and *question* patterns have the same relationship, we combined them into the *directive+question* pattern with the assumption that it would yield improved results over treating them separately.

Using the clustering method, we calculated the percentage accuracies for the four datasets MEB, CAV, CDR, and IBCT Intel for the four patterns respectively. From the mean absolute errors (MAE) we computed the percentage accuracy based on the highest rank N considered for the evaluation. Figure 5 describes the overall result. In each panel, the x-axis shows the patterns, i.e., *directive*, *question*, *informative* and *directive+question* and the y-axis shows the percentage accuracy. From the result, we observed that the percentage accuracy for *informative* is the highest for all the datasets (73.4%, 76.5%, 69.5%, 68%), which suggests that the *informative* pattern is a better indicator of hierarchy than other patterns. In addition, we performed one-tailed t-test to check if the accuracy for *informative* is significantly higher than for *directive*, *question*, and *directive+question* at the significant level of 5%. We find that the accuracy for *informative* is indeed significantly higher than *directive* (p=0.03) and *directive+question* (p=0.002), but not significantly so for question (p=0.06).



(a) Patterns in MEB. (b) Patterns in CAV.

(c) Patterns in CDR. (d) Patterns in IBCT Intel.

Fig. 5. Percentage accuracy of computing hierarchy using different response patterns directive (Dir), question (Ques), informative (Info), and directive+informative (Dir+Info) via different datasets.

4.2.2 Evaluating Features We compared MAEs obtained using only the semantic features with those obtained using only the social features. For the comparison, we performed one-tailed t-tests on the MAEs obtained from the four chat rooms for all patterns.

Table 3 summarizes these hypotheses and the results obtained.

In the table, we have stated hypotheses that compare the mean (μ) of the MAEs obtained using features F_1 to F_8 for the patterns. When the null hypothesis is rejected we accepted the alternative hypothesis, i.e., one mean is significantly less than the other. Among the features, F_1 to F_4 represent the semantic features, F_5 and F_6 represent the communication features, and F_7 and F_8 represent the social features. We found that the MAEs obtained based on features F_1 to F_4 were not significantly lower than the MAEs obtained based on features F_7 and F_8. Recall that F_5 is *average response time delay* and F_6 is *number of messages*. We found that the MAEs obtained based on features F_1 to F_5 or obtained based on features F_1 to F_6 were significantly lower than those obtained based on F_7 and F_8. When we added F_5 and F_6 to features F_7 and F_8, the MAEs obtained were not significantly lower than the MAEs obtained considering features F_1 through F_4. Similarly, when we added F_6 to features F_7 and F_8, the MAEs obtained were not significantly lower than the MAEs obtained using features F_1 through F_5. The foregoing suggests that the semantic features are better indicators of hierarchy than the social features.

Table 3. Statistically comparing semantic features with social features (sem-semantic, comm-communication, soc-social, avg-average, resp-response, del-delay, msg-messages, hyp-hypotheses, rej-rejected).

#	Alt. Hypotheses	Null hyp. p-val	Null hyp. rej. at 5%?
1.	**sem.** ($\mu_{F_1 toF_4}$) < **soc.** ($\mu_{F_7 toF_8}$)	0.23	no
2.	**sem. & avg. resp. time del.** ($\mu_{F_1 toF_5}$) < **soc.** ($\mu_{F_7 toF_8}$)	0.04	yes
3.	**sem. & comm.** ($\mu_{F_1 toF_6}$) < **soc.** ($\mu_{F_7 toF_8}$)	0.00	yes
4.	**soc. & comm.** ($\mu_{F_5 toF_8}$) < **sem.** ($\mu_{F_1 toF_4}$)	0.08	no
5.	**soc. & no. of msg** ($\mu_{F_6 toF_8}$) < **sem.** ($\mu_{F_1 toF_4}$)	0.13	no

4.2.3 The Enron Dataset We evaluated our approach on the Enron email dataset [5, 10] as well. A major challenge we faced is to create conversation threads based on a subject or a topic. Whereas in the military dataset we considered the counts of the response messages to *directive*, *question*, and *informative* messages, for the Enron dataset, we considered the counts of *directive* and *question* messages sent by an employee. We considered a message whose subject begins with "RE:" as an *informative* because it indicates that the message responds to a prior message. To identify patterns we used the rules described in Section 3.1. Once the messages were identified, we computed ranks using the

rules provided in Section 3.4. The features we considered to compute ranks were F_1, F_6, F_7, and F_8. We did not consider features F_2 to F_5 (*emotions* and *average response time delay*) because we could not create conversation threads. We constructed F_7 and F_8 based on the *number of messages* exchanged between employees.

We found that ranks computed using the *informative* pattern have higher accuracy (75%) than the directive (74.4%) and question (70.1%) patterns. This results coheres with our finding over the military data. However, unlike the military data, the accuracy from social features (72%) was slightly higher than for the semantic and communication features (71%). We also found that adding semantic features to social and communication features (75%) slightly improved the accuracy over considering only social and communication features (74%). Therefore, along with social and communication features semantic features were important in predicting hierarchy.

5 Discussion and Future Work

We provide a novel approach to computing team hierarchy from broadcast messages. To compute the hierarchy, first, we identify three patterns via text mining obtaining F-measures of 80%, 95%, and 60% respectively for *question*, *informative*, and *directive* patterns, and 83% overall. Second, once we identify the patterns, we extract disparate features: semantic, communication, and social. Third, using the features we compute ranks using the hierarchical clustering method. We find that the *informative* pattern is a better indicator of hierarchy than the other patterns, thus validating our approach. We find that semantic features added with communication features (i.e., the network statistics) are better indicators of ranks than using social features alone. We obtain similar results regarding the usage of patterns to infer hierarchy on the Enron dataset. We also find that semantic features added to social and communication features improve accuracy in predicting hierarchy. However, social features in Enron are better indicators of hierarchy than semantic features. This could be because the Enron dataset is much larger than the military dataset: on average, Enron participants sent more messages than military participants.

Although we consider only two datasets, our study provides some hints as to the differences in how people use chat communications versus email, at least in work-related settings. Email communications would tend to respect predefined organizational relationships (who writes to whom) and thus social features are predictive of hierarchy. In contrast, broadcast communications at the level of connectivity do not respect any predefined relationships. Thus their semantic features are better predictive of hierarchy. In the military setting, the ranks of the participants are well defined. We conjecture that, in settings where ranks are not predefined, such as in collaborations between peers as in open source software development or nascent political movements, broadcast communications would be a way for true hierarchies to emerge.

This work estimates intrateam hierarchy. In future work, we will consider interteam hierarchy. Also, we hope to extend our work on the estimation of a hierarchy to the estimation of team cohesion, trust, and performance. We plan to improve our domain-specific military lexicon to further improve performance. We expect that our results would be stronger on larger datasets where participants communicate more frequently with each other.

6 Related Work

There has been a small amount of research on inferring hierarchy from communications. Nishihara and Sunayama [16] compute hierarchy by two measures: based on request actions communicated by a speaker and the number of sentences sent by a speaker. In contrast, instead of identifying requests, we identify patterns such as *directive*, *question*, and *informative*. Moreover, Nishihara and Sunayama do not incorporate features such as *emotions*, *average response time delay*, or centrality features that can provide important clues to identify hierarchy. Also, they evaluate their work on directed messages but not on broadcast messages.

Gilbert [7] identifies words and phrases from Enron emails [10, 5] that indicate team hierarchy. This work is limited to finding such words and phrases rather than computing a hierarchy. Also, Gilbert's approach is domain-dependent because it requires words and phrases related to hierarchy. Preparing such lexicons for new datasets can be cumbersome. In contrast, we provide ways to identify patterns that generalizes to different datasets. Also the lexicon we prepare is easy to extract as the verbs are extracted based on their frequencies.

Rowe et al. [18] compute team hierarchy by extracting an undirected graph based on emails exchanged between senders and receivers. They consider centrality measures to compute hierarchy and do not focus on analyzing the content of emails. Hence, Rowe et al.'s contribution does not handle broadcast messages. In contrast, we emphasize understanding the content of the messages to identify the patterns and consider broadcast messages. In addition, we find that patterns and emotions extracted from messages are better indicators of hierarchy than are centrality measures.

Krafft et al. [11] propose a probabilistic model to visualize topic-specific subnetworks in email datasets. In specific, they associate an author-recipient edge (or an email) with different subtopics using K-dimensional topic-specific communication patterns. In our work we take a similar approach where we extract different communication patterns and features from emails and broadcast messages for participants to infer their hierarchy.

7 Acknowledgment

This work was primarily supported by the Army Research Laboratory in its Network Sciences Collaborative Technology Alliance (NS-CTA) under Cooperative Agreement Number W911NF-09-2-0053. Kalia and Singh were partially supported by the NCSU Laboratory for Analytic Sciences.

References

1. Austin, J.L.: How to Do Things with Words. Clarendon Press, Oxford (1962)
2. Borgatti, S.P., Everett, M.G.: A graph-theoretic perspective on centrality. Social Networks 28(4), 466–484 (2005)
3. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1), 37–46 (1960)
4. Eaton, J.W.: Social processes of professional teamwork. American Sociological Review 16(5), 707–713 (1951)
5. Fiore, A., Heer, J.: UC Berkeley Enron email analysis (2004), `http://bailando.sims.berkeley.edu/enron_email.html`
6. Freeman, L.C.: Centrality in social networks conceptual clarification. Social Networks 1(3), 215–239 (1979)
7. Gilbert, E.: Phrases that signal workplace hierarchy. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work. pp. 1037–1046. Seattle (2012)
8. Gupte, M., Shankar, P., Li, J., Muthukrishnan, S., Iftode, L.: Finding hierarchy in directed online social networks. In: Proceedings of the 20th International Conference on World Wide Web. pp. 557–566. Hyderabad (2011)
9. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics. pp. 423–430. Stroudsburg (2003)
10. Klimt, B., Yang, Y.: The Enron corpus: A new dataset for email classification research. In: Proceedings of the 15th European Conference on Machine Learning. LNCS, vol. 3201, pp. 217–226. Pisa (2004)
11. Krafft, P., Moore, J., Desmarais, B.A., Wallach, H.M.: Topic-partitioned multinetwork embeddings. In: Proceedings of 25th Annual Conference on Neural Information Processing Systems 2012. pp. 2807–2815. Curran Associates, Inc., Lake Tahoe (2012)
12. Lehner, P.: Testing the accuracy of automated classification systems using only expert ratings that are less accurate than the system. The MITRE Corporation pp. 1–28 (2014)
13. Leonard, M.W., Frankel, A.S.: Role of effective teamwork and communication in delivering safe, high-quality care. Mount Sinai Journal of Medicine 78(6), 820–826 (2011)
14. Mones, E.: Hierarchy in directed random networks. Physical Review E 87(2), 022817 (2013)
15. Mones, E., Vicsek, L., Vicsek, T.: Hierarchy measure for complex networks. PLoS ONE 7(3), e33799 (2012)
16. Nishihara, Y., Sunayama, W.: Estimation of friendship and hierarchy from conversation records. Information Sciences 179(11), 1592–1598 (2009)
17. Resick, C.J., Murase, T., Randall, K.R., DeChurch, L.A.: Information elaboration and team performance: Examining the psychological origins and environmental contingencies. Organizational Behavior and Human Decision Processes 124(2), 165–176 (2014)
18. Rowe, R., Creamer, G., Hershkop, S., Stolfo, S.J.: Automated social hierarchy detection through email network analysis. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. pp. 109–117. San Jose (2007)

19. Sibson, R.: SLINK: An optimally efficient algorithm for the single-link cluster method. The Computer Journal 16(1), 30–34 (1972)
20. Socher, R., Perelygin, A., Jean Y. Wu, J.C., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Seattle (2013)
21. Wang, Y., Iliofotou, M., Faloutsos, M., Wu, B.: Analyzing communication interaction networks (cins) in enterprises and inferring hierarchies. Computer Networks 57(10), 2147–2158 (2013)