The Expected Value Test: A New Statistical Warrant for Theoretical Saturation

Zachari Swiecki$^{1[0000-0002-7414-5507]}$

Monash University, Clayton Victoria 3800 AUS
zach.swiecki@monash.edu

Abstract. The basic goal of quantitative ethnography (QE) is to use statistics to warrant theoretical saturation. In QE studies that use epistemic network analysis (ENA), statistical comparisons between two or more samples are often used as warrants. However, no standard quantitative techniques have been developed to provide warrants absent of differences between samples—e.g., for the data overall or for particular individuals in the data. In this paper, I introduce the expected value test (EVT), a technique for finding statistical warrants for single samples in the context of ENA-based QE analyses. Building on the concepts of agreement due to chance and randomization, the EVT generates a distribution of networks whose edge weights approximate the expected value of the edge weights due to chance. Using these distributions affords tests of statistical significance been observed networks and chance-based networks. These tests provide warrants that connections observed qualitatively and measured by ENA are theoretically saturated.

Keywords: quantitative ethnographic methods · epistemic network analysis · one-sample tests · Cohen's kappa · randomization tests

1 Introduction

Quantitative ethnography (QE) is a system of reasoning that uses statistics to warrant claims about *theoretical saturation*—when qualitative observations constitute a meaningful pattern in some discourse [8]. These warrants are typically in reference to qualitative and quantitative differences between two or more samples, for example differences between the discourse patterns of control and treatment groups.

However, in many cases, researchers may not be able to, or do not wish to, compare samples. Instead, they can or want to say something about the data overall, one group, or one particular individual. For example, perhaps they only have qualitative data on one class of students with no meaningful subgroups. Or, they are more interested in *describing* the discourse patterns of individuals rather than using those patterns to *predict* whether individuals behave more like one group or another. Unfortunately, there are no standard quantitative techniques for warranting theoretical saturation in the context of QE for situations like these.

In this paper, I introduce a novel technique—the expected value test (EVT)—that addresses this limitation of QE. The technique works by generating data that simulates random connections in the discourse. Models produced from this simulated discourse can then be compared to the model produced from the observed discourse. If the observed model is found to be statistically significantly different from the random models, there is evidence that the connections present in the discourse are not produced at random, but instead are a systematic property of the discourse. In turn, there is evidence that these connections are theoretically saturated. Here, I describe the EVT and use it to warrant that observed connections in the discourse of military teams are theoretically saturated.

2 Background

2.1 Saturation

The grounded theory perspective on qualitative analysis argues that theoretical saturation is the point at which researchers repeatedly find similar instances of the same phenomena of interest; it is the point at which further analysis would reveal nothing new [4]. As Shaffer describes, "the basic idea of Quantitative Ethnography is simple: Use statistics to warrant theoretical saturation in qualitative analyses," [8] (pg. 393). Statistics—specifically, significance tests—warrant theoretical saturation because they give evidence that a particular observation is not simply a random property of the sample we have, but a pattern in the larger population of data we could sample from—in other words, if we were to continue to sample data, we would see the same things over and over again. In QE, statistics typically come in two forms: significant interrater reliability measures, such as Cohen's kappa and Shaffer's rho, and significant parameters from epistemic network analysis (ENA), such as differences between ENA scores [8]. This paper is concerned with the later kind of statistics.

In a standard QE analysis using ENA, researchers develop qualitative interpretations about the connections that individuals or groups make in their discourse. They then analyze the discourse using ENA, which quantifies connections in a particular way. ENA produces parameters that can be tested using well-known statistical methods.[1] For example, this was the process that my colleagues and I used to analyze the discourse of military teams in training [11].

In that study, which compared individuals in treatment and control conditions, our qualitative analysis found that treatment individuals tended to respond to information with decisions, and control individuals tended to respond to information by seeking information—that is, asking questions. To statistically warrant these claims, we analyzed the data using ENA. The ENA dimension that accounted for the biggest difference between treatment and control individuals distinguished them in terms of connections to decisions versus connections to

[1] Crucially, though less important to the arguments in this paper, researchers also *close the interpretive loop*—that is, they judge whether their qualitative interpretations align with the quantitative results.

seeking information. Using the average ENA scores for individuals in each condition on this dimension, we conducted Mann-Whitney tests and found that treatment individuals made significantly more connections to decisions and control individuals made significantly more connections to seeking information. This result provided a statistical warrant for our original qualitative claims.

2.2 Samples

As the example above highlights, our statistical warrant came in the form of a significant difference between two groups, or *samples*, of individuals. Such comparisons between samples are common statistical warrants in ENA-based QE analyses [1, 5, 10, 12]. However, in cases where researchers wish to say something about the connections that an individual or group made absent of the differences from others, the appropriate statistical test to use is less clear. Given some ENA-based parameter about one individual or one group of individuals or the ENA model overall, what statistical warrant should be used?

This question is not a new one in the field of statistics. Essentially, it is the difference between a one-sample and a two-sample test. Like the analysis of military teams, many ENA-based approaches to QE use two-sample tests to warrant qualitative claims. In their most basic form, the parameter difference between the samples (say the difference between mean ENA scores on a particular dimension) is calculated and that differences is compared to zero. One-sample tests work similarly, however, differences between samples cannot be calculated, so the difference between a parameter on the sample is compared to some other meaningful value. In many cases, this value is zero—if it is significantly different from zero, then some difference is actually there—but the meaning of zero may differ from test to test. Cohen's kappa provides one such example.

2.3 Chance

Cohen's kappa is a widely applied statistic for measuring inter-rater reliability—the extent to which two raters make similar categorizations of data [2]. In the learning sciences and other fields, kappa is used to measure the agreement between two process of assigning qualitative codes to data. Often, these processes are done by human raters, but in other cases they may be automated. Regardless of who or what codes the data, kappa measures how often the two raters agree.

Importantly, however, the kappa statistic differs from simply calculating the number of cases in which the raters agree divided by the total number of cases—percent agreement. It is possible that the two raters may agree simply due to chance—for example, if both raters randomly assigned codes to the data, we would likely find that their assignments agreed some of the time. To account for this, the kappa calculation incorporates the expected level of agreement due to chance. The outcome of this correction is a statistic whose zero value corresponds to the level of chance agreement. One way in which Cohen's kappa is meaningful, then, is the extent to which it measures a pattern that is different from what we would expect due to chance.[2]

The expected value test (EVT) draws inspiration from Cohen's kappa in the sense that it suggests a meaningful value to which we can compare an ENA-based parameter for a single sample. Just as two-raters might agree on their ratings by chance—that is, their rating values may co-occur frequently—individuals may make connections between concepts at random. As a thought experiment, say we took a transcribed and coded conversation between two friends and shuffled the codes and order of talk—that is, we simulated a scenario where *when* a person talks and *what* they talk about is not based on their interaction with others. If we found that the connections we observed in the real conversation and the connections we observed in the simulated conversation were the same (or very similar) it would cast doubt on the claim that the two friends were meaningfully responding to one another. In statistical terms, our observations would be common under the null hypothesis of randomization. In other words, our claims about this conversation would be statistically unwarranted.

This example illustrates the basic idea of the EVT: compare an ENA parameter value for a sample to the value we would expect that parameter to take if the connections in the discourse were occurring due to chance alone.

2.4 Related Work

The randomization process of the EVT builds upon work done by Csandi and colleagues [3]. In their study, they sought to test whether the temporal ordering of the discourse was meaningful. To do so, they compared an ENA model produced from an observed dataset to an ENA model produced from a dataset in which they randomized the order of the lines. The EVT extends this randomization process in two important ways.

First, Csandi and colleagues only randomized the order of the lines. This is appropriate for testing whether the temporal nature of the data is important, but it is not sufficient for testing whether the connections in the discourse are theoretically saturated. In co-occurrence-based models like ENA, it matters where the discourse occurs because connections are identified within particular segments of discourse and may form between an individual's lines and the lines of others. However, connections may also form within single lines. Simply reordering the lines ignores the impact of within line connections.[3] So, even if the order of the lines are randomized, systematic patterns of connections may still be present. To achieve the effect of chance-based co-occurrences, we therefore need to randomize the presence or absence of the codes in the discourse as well as the order of the lines.

[2] Another way in which Cohen's kappa is meaningful is the extent to which it generalizes, which is addressed by Shaffer's rho.

[3] In the coded data examined by Csandi and colleagues, each line could be coded for only one code, therefore this issue did not apply to their data. However, in other contexts, multiple codes may be assigned to the same line.

Second, to produce a model with which to investigate the importance of line order, Csandi and colleagues randomized the dataset *once* and then compared the original ENA model to the model created from the randomized data. They did so by testing for statistical differences between the original and random networks in the low-dimensional space produced by ENA. While this method is adequate for demonstrative purposes, it has two important limitations. First, by randomizing only once, it ignores the variation that may be present between different random datasets. Second, by comparing the networks in the low-dimensional space, the method tests for differences between summaries of the networks and not their overall structure. This means that other potentially important differences (or similarities) between the observed and random networks might be missed. The EVT addresses both of these issues by creating a *distribution* of chance-based networks and making statistical comparisons between the *complete networks*, rather than their low-dimensional projections.

2.5 Research Questions

To demonstrate the EVT, I used it to address the research question: RQ1—Does the EVT suggest that the connections observed in the discourse of military teams are theoretically saturated? Furthermore, because any reliable test should identify situations where the test passes and where it fails, I also address the research question: RQ2—Can the EVT suggest when observed connections are theoretically saturated and when they are not? In other words, to provide statistical warrants for claims about theoretical saturation regarding single samples, I used the EVT to test whether the connections observed overall in these data were significantly different from chance. To examine the discriminatory power of the technique, I used it to test whether the connections observed for specific individuals were significantly different from chance.

The data used here have been analyzed multiple times from a QE perspective (e.g., see [11]). Each of these prior analyses suggested that the observed connections in the discourse were theoretically saturated. The warrants for these claims were in-depth qualitative analyses of the data, quantitative results that suggest that different groups of individuals made different patterns of connections, and alignment between the qualitative and quantitative results. However, no quantitative warrants have been provided for the connections observed in the data overall, or for particular individuals in these data. The EVT was designed to provide such warrants.

3 Methods

3.1 Data

Data from air defense warfare (ADW) teams was collected as part of the Tactical Decision Making Under Stress project [6]. 94 individuals in 16 teams participated in simulated scenarios to test the impact of a decision-support system (DSS) and teamwork training on team performance. During the scenarios, teams performed the *detect-to-engage sequence*, which entailed using computer-based tools to detect and identify ships and aircraft in the vicinity—referred to as *tracks*—assess whether they were threats and decide whether to respond with warnings or combat orders.

Each team consisted of six officers assigned to particular roles. Two officers held command roles and four held supporting roles. The command roles were the Commanding Officer (CO) and the Tactical Action Officer (TAO). They were responsible for making tactical decisions, such as when to warn or engage tracks. Support roles were the Identification Supervisor (IDS), the Air Warfare Coordinator (ADWC), the Tactical Information Coordinator (TIC), and the Electronic Warfare Supervisor (EWS). They were responsible for reporting critical information to commanders, such as the detection of threats, their identification—for example, whether the threat was a jet, ship, or helicopter—and their behavior—for example, their speed and location relative to the warship.

During the training scenarios, team members communicated via an open-channel radio system. The dataset analyzed here consists of transcripts of team talk that were segmented for analysis by turn of talk for a total of 12,027 turns.

3.2 Coding

The transcripts were labelled for the presence or absence of the codes described in Table 1. These codes were developed, validated, and applied to the data as part of a prior analysis [11]. The coding scheme was validated using Cohen's kappa and Shaffer's rho as part of a process that compared the ratings of two human raters with automated classifiers. The automated classifiers were validated using the standard threshold for kappa (>.65) and a rho threshold of < 0.05. All pairwise combinations of raters (humans and automated classifier) achieved kappa > 0.83 and rho $(0.65) < 0.05$.

3.3 Analysis

The EVT includes the following main steps, which are described in more detail below:

1. Generate a model from the observed data.
2. Generate a distribution of models from data in which the codes and lines have been repeatedly randomized—that is, a distribution of chance-based models.
3. Calculate the similarity of the observed model to the average, or *centroid*, of the chance-based models. Calculate the distribution of similarities of the chance-based models to the centroid. Compare the observed similarity to the distribution.

Table 1: Codes, definitions, and examples.

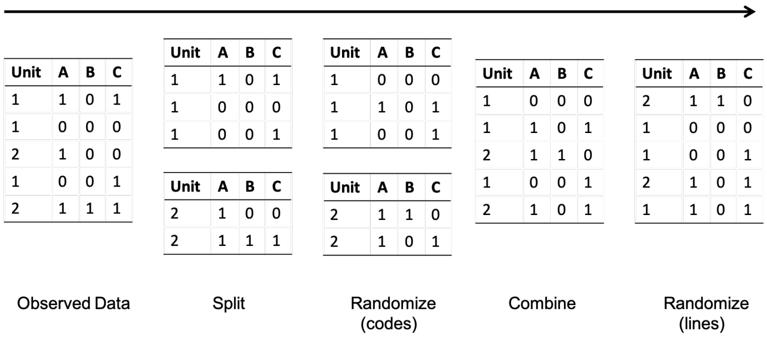
Code	Definition	Example
Detection	Talk about radar detection of a track or the identification of a track, (e.g., vessel type).	IR/EW NEW BEARING, BEARING 078 APQ120 CORRELATES TRACK 7036 POSSIBLE F-4
Track Behavior	Talk about kinematic data about a track or a track's location.	AIR/IDS TRACK NUMBER 7021 DROP IN ALTITUDE TO 18 THOUSAND FEET
Assessment	Talk about whether a track is friendly or hostile, the threat level of a track, or indicating tracks of interest.	TRACKS OF INTEREST 7013 LEVEL 5 7037 LEVEL 5 7007 LEVEL 4 TRACK 7020 LEVEL 5 AND 7036 LEVEL 5
Status Updates	Talk about procedural information, e.g., track responses, or talk about tactical actions taken by the team.	TAO ID, STILL NO RESPONSE FROM TRACK 37, POSSIBLE PUMA HELO
Seeking Information	Asking questions regarding track behavior, identification, or status.	TAO CO, WE'VE UPGRADED THEM TO LEVEL 7 RIGHT?
Recommendation	Recommending or requesting tactical actions.	AIR/TIC RECOMMEND LEVEL THREE ON TRACK 7016 7022
Deterrent Orders	Giving orders meant to warn or deter tracks.	TIC AIR, CONDUCT LEVEL 2 WARNING ON 7037
Defensive Orders	Giving orders to prepare defenses or engage hostile tracks.	TAO/CO COVER 7016 WITH BIRDS

Observed Model. I modeled the observed data using the R implementation of ENA [7]. ENA uses a moving window to construct a network for each line in the data, here defined as each turn of talk. Edges in the network are defined as the co-occurrence between codes in the current line and codes within the window, operationalized as a specific number of prior lines. To create networks for individuals, ENA aggregates the networks associated with their lines and normalizes this aggregated network. ENA represents each individual's network visually as a weighted network graph in which the nodes correspond to codes, and the edges are the normalized rate of co-occurrence between the codes.[4] These networks can be plotted and subtracted from other networks to show the connections that are stronger in one network relative to the other.

Distribution of Chance-based Models. To create a distribution of chance-based models, the EVT constructs multiple simulated datasets based on the frequency of talk for each individual in the data and the frequency of their code occurrences. Figure 1 provides an overview of this process. Given the observed data, for each conversation, the EVT algorithm splits the data by each unique unit of analysis. For each unit, the algorithm then shuffles their code values by code. Next, the data is recombined in the original order in which the lines of data occurred. Finally, the line order is randomized. The effect of this process is to simulate situations in which these units randomly produced coded discourse—for example, if they randomly chose when to speak and what to speak about given their overall talk and code frequencies in a given conversation. Due to the

[4] Typically, ENA also performs a dimensional reduction on the collection of normalized and centered networks via singular value decomposition. However, the analysis presented here focused on the complete networks rather than their projections in a low dimensional space.

stochastic nature of the process, the simulation is repeated a large number of times—here, 1000 times. For each of the randomized datasets, an ENA model is created using the same specifications used for the observed model. This produces a distribution of chance-based ENA models.



Observed Data

Unit	A	B	C
1	1	0	1
1	0	0	0
2	1	0	0
1	0	0	1
2	1	1	1

Split

Unit	A	B	C
1	1	0	1
1	0	0	0
1	0	0	1

Unit	A	B	C
2	1	0	0
2	1	1	1

Randomize (codes)

Unit	A	B	C
1	0	0	0
1	1	0	1
1	0	0	1

Unit	A	B	C
2	1	1	0
2	1	0	1

Combine

Unit	A	B	C
1	0	0	0
1	1	0	1
2	1	1	0
1	0	0	1
2	1	0	1

Randomize (lines)

Unit	A	B	C
2	1	1	0
1	0	0	0
1	0	0	1
2	1	0	1
1	1	0	1

Fig. 1: Overview of the EVT randomization process.

Significance Testing. Constructing an ENA model from the observed data and a distribution of models from the randomized data affords a statistical comparison between the observed model and the chance-based models. Broadly, this comparison involves calculating the centroid (average) of the chance-based models. Next, the similarity of the observed model to the centroid is calculated. Additionally, the similarity of each chance-based model to the centroid is calculated. This last step yields a distribution of similarity metrics under the null hypothesis that the connections identified in the models were due to chance. In other words, if the connections were due to chance alone, this is the distribution of similarities between models we would see. If the similarity of the observed model to the centroid is greater than 95% of the values in this distribution, we reject the null. That is, with a type I error rate of 0.05, we conclude that the observed model is significantly different than chance.

In general, then, the EVT is simply an empirical hypothesis test. It empirically generates the null hypothesis distribution rather than assuming it has a particular form (as is done in parametric tests) and the p value is calculated as the number of observations in the null hypothesis distribution that are more extreme than the test statistic—here, the similarity between the observed model and the chance-based centroid.

More formally, let the normalized network for unit i using the observed data be $\mathbf{v}^i$ for $i = 1, \ldots, m$, where m is the number of units of analysis, and let the normalized network for unit i using randomized data j be $\mathbf{v}_j^{*i}$ for $j = 1, \ldots, n$, where n is the total number of randomized datasets. $\mathbf{v}^i$ and $\mathbf{v}_j^{*i}$ are both vectors where each element of the vector is the normalized co-occurrence rate for a given connection—that is, the edge weight of a given line typically represented in ENA visualizations. $\mathbf{v}^i$ is the observed model for the unit; $\mathbf{v}_j^{*i}$ is the chance-based model for the unit, given some randomized data.

Next, construct the centroid (i.e., average of the co-occurrence vectors) of the n chance-based models for a given unit: $\bar{\mathbf{v}}^{*i}$. The similarity of the observed model to this centroid is defined as the Euclidean distance between the two:

$$d^i = \left|\mathbf{v}^i - \bar{\mathbf{v}}^{*i}\right| \tag{1}$$

Likewise, the similarities of the chance-based models to the centroid for each randomized dataset j are:

$$d_j^{*i} = \left|\mathbf{v}_j^{*i} - \bar{\mathbf{v}}^{*i}\right| \tag{2}$$

Next, construct the distribution $D(d_1^{*i}, d_n^{*i})$—that is, the distribution of distances between the chance-based models and their centroid for the ith unit. We can think of d^i, then, as the test statistic and $D(d_1^{*i}, d_n^{*i})$ as the null hypothesis distribution. We can then calculate an empirical p value by counting the number of observations in $D(d_1^{*i}, d_n^{*i})$ that are more extreme (greater than) d^i. If this value is less than α (here, 0.05), we conclude that the observed model is significantly different from chance.[5]

Equations (1) and (2) test whether a given *individual's* model is significantly different from their chance model. We can modify these equations to test whether the overall model (that is, the model on the entire dataset of all individuals) is significantly different from chance. To do so, we simply compute the average distance from the observed models to their centroids:

$$\bar{d} = \frac{\sum_{i=1}^{m} \left|\mathbf{v}^i - \bar{\mathbf{v}}^{*i}\right|}{m} \tag{3}$$

and the average distance between the chance-based models and their centroids:

$$\bar{d_j^*} = \frac{\sum_{i=1}^{m} \left|\mathbf{v}_j^{*i} - \bar{\mathbf{v}}^{*i}\right|}{m} \tag{4}$$

Finally, construct the distribution of average distances from the chance-based models to their centroids, $\bar{D}(\bar{d_1^*}, \bar{d_n^*})$, and use this distribution and (3) to compute the empirical p value.

To address RQ1, I conducted a qualitative analysis of the overall ADW discourse informed by findings from my prior work [11]. I then sought a quantitative warrant for my observations using an EVT that compared the average distance of the observed models to the centroids of the chance-based models (equation 3) to the distribution of average distances of the chance-based models to their centroids (distribution of equation 4). To address RQ2, I conducted qualitative

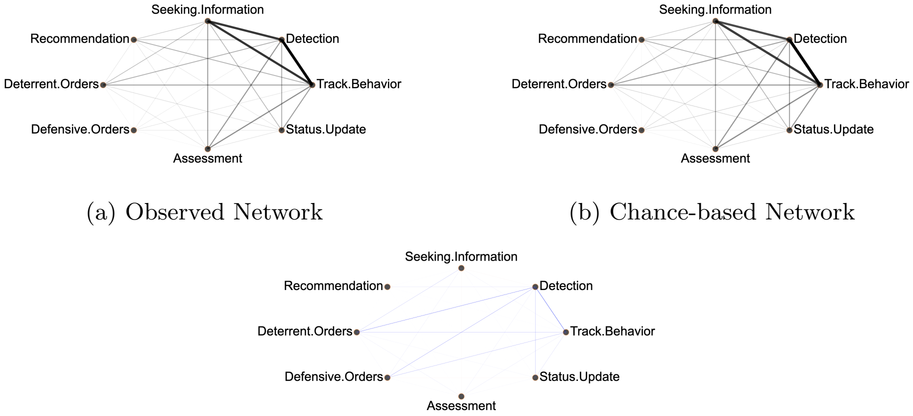
[5] In this sense, the EVT is essentially an empirical version of Hoetelling's T^2 test. However, for that test, the squared Mahalanobis distance is used in place of the Euclidean distance and a normal distribution is assumed.

analyses of the discourse of two specific individuals in the data—one whose observed model was highly *dissimilar* to chance and one whose observed model was highly *similar* to chance, as measured by equation 1. I conducted two separate EVTs for these individuals to determine whether their observed models differed significantly from chance. For both research questions, I compared observed and chance-based networks using network subtractions and coordinated these findings with qualitative interpretations of the data.

4 Results

4.1 RQ1

The results of the overall qualitative analysis suggested that many individuals had patterns of connections between Detection and Track Behavior and between Deterrent Orders and Detection. In the data, it was common to see individuals report information about the detection and behavior of tracks in the same line. For example, when an EWS announced to their team: "NEW TRACK AT BEARING 068 APQ120 CORRELATES TO 7036 POSSIBLE F-4," they are reporting the detection of a new track whose radar (APQ120) suggests it is a F-4 jet. Behavior information is passed as the track's location at bearing 068. Similarly, it was common in the data to see commanding officers make connections between Detection and Deterrent Orders. For example, when a TAO said: "TRACK 7036 WAS ID AS F-4. ISSUE LEVEL 2 WARNINGS TO TRACK 7036.", they passed information about the detection and identification of the track as an F-4 and immediately gave an order to issue warnings to the track. Here, warnings are deterrents that announce the presence of the warship and request that the track identify themselves and their intentions.



(a) Observed Network

(b) Chance-based Network

(c) Difference Graph

Fig. 2: Network comparison for overall data.

Figure 2 shows the shows the network models for these data.[6] On the left is the observed mean network. On the right is the mean chance-based network—that is, the mean network representation of the chance-based centroids. Below is their difference graph. Here, connections in red occurred more frequently than chance and connections in blue occurred less frequently than chance. Both the observed and the chance-based networks suggest strong connections between Detection and Track Behavior and Detection and Deterrent Orders. However, the difference graph suggests that these connections (along with most others) occurred less frequently than chance.

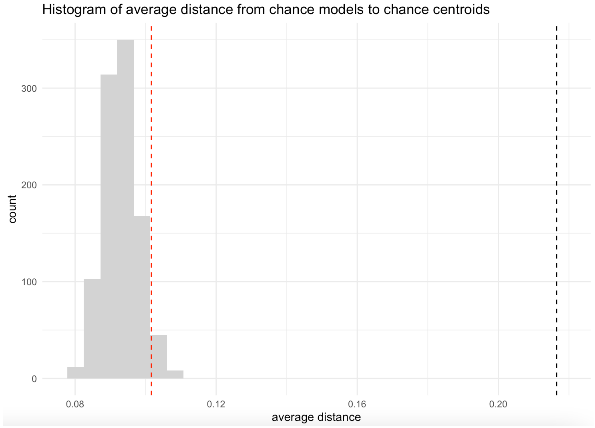


Fig. 3: Histogram of average distance from chance models to chance centroids.

The results of the EVT for the overall data are summarized in Figure 3. The histogram shows the distribution of the average distances between the chance-based models and their centroids—that is, the null hypothesis distribution. The red dashed line marks the 95th percentile of the distribution. The black dashed line marks the value of the average distance of the observed models from the chance-based centroids (our test statistic). Because this distance is greater than 95% of the null hypothesis distribution, the EVT indicates that the observed models are significantly different from chance with an empirical p value of 0. The effect size of this difference, as measured by Cohen's d, is massive ($d = 25$). The results suggest that there is a strong pattern of connections in this discourse and the qualitative observations are theoretically saturated.

4.2 RQ2

The results of the qualitative analysis of the individual whose observed network was highly dissimilar to chance, a CO, suggests that they made systematic connections between information about potential threats and deterrent orders (see

[6] Because a dimensional reduction was not used to project the networks in a low-dimensional space, I have positioned the nodes of these networks in a circle to simplify their presentation.

Table 2). Recall that in ADW discourse, potentially hostile radar contacts are referred to as tracks and they are given identification numbers by the team. Also note that in these data, speakers often began their turns of talk by addressing the member(s) of the team for whom the message is intended followed by who was speaking. Thus, "CO/TAO" should be read as "CO, this is TAO."

Table 2: CO

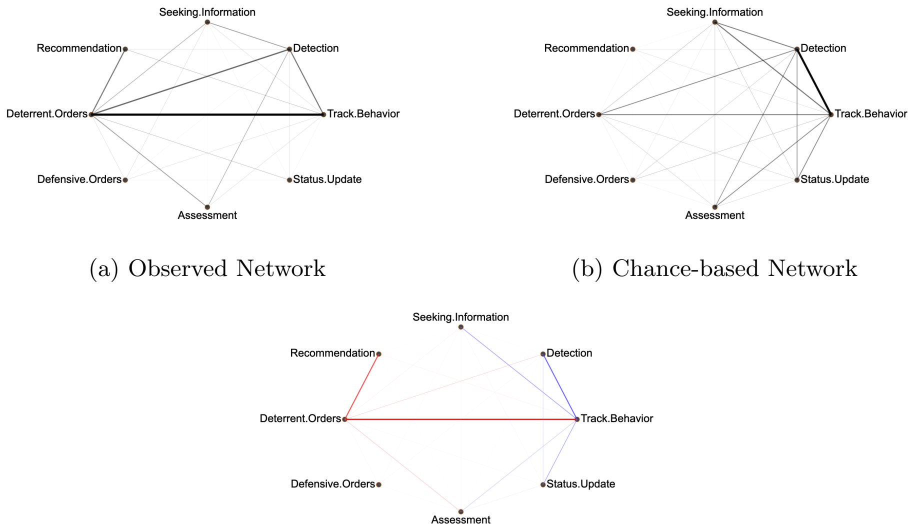
Line	Speaker	Utterance
1	TAO	CO/TAO FOR TRACK 7002 THE LA COMBANTE DSS IS GIVING ORIGION AS EVIDINCE OF A THREAT HOWEVER IT IS COMING FROM ORMANI UAE TERRITORIAL WATERS
2	CO	I DON'T FEEL TO COMFORTABLE WITH THAT LET ME SEE WHAT HIS RANGE IS TO THE SOUTH
3	TAO	14
4	CO	ONCE HE CLOSES INTO 10 LETS GO AHEAD AND GIVE HIM A LEVEL ONE

The excerpt begins as the TAO passes information about a new track (7002) identified as a La Combattante ship. The decision-support system (DSS) suggests classifying this track as threat based on its origin, but the TAO is questioning this classification. The CO responds (line 2) saying that they are not "comfortable" with the ship and proceeds to check the distance from their warship to the potential threat. The TAO provides the distance in the next line (3) as 14 nautical miles. Then, the CO gives a deterrent order, telling the TAO to send a "LEVEL ONE" warning to the track once the ship is within 10 nautical miles (line 4).

The excerpt illustrates that this CO made meaningful connections between their own discourse and the discourse of the team. In particular, it shows that they made tactical decisions ("GO AHEAD AND GIVE HIM A LEVEL ONE") based on information from another team member.

Figure 4 shows the network models for this individual. On the left is their observed network. On the right is their average chance-based network—that is, the network representation of their chance-based centroid. Below is their difference graph. Their observed network suggests that the CO made strong connections between Deterrent Orders and information such as Track Behavior, Assessment, and Detection. The difference graph suggests that these connections occurred more frequently than expected due to chance.

Figure 5 summarizes the results of the EVT for this individual. The black-dashed line indicates that the distance between the observed model and the chance-based centroid is greater than 95% of the distances between the chance-based models and the centroid (red dashed line). Some values in the distribution do exceed the test statistic, hence the empirical p value for this test is 0.001,



(a) Observed Network

(b) Chance-based Network

(c) Difference Graph

Fig. 4: Network Comparison for CO

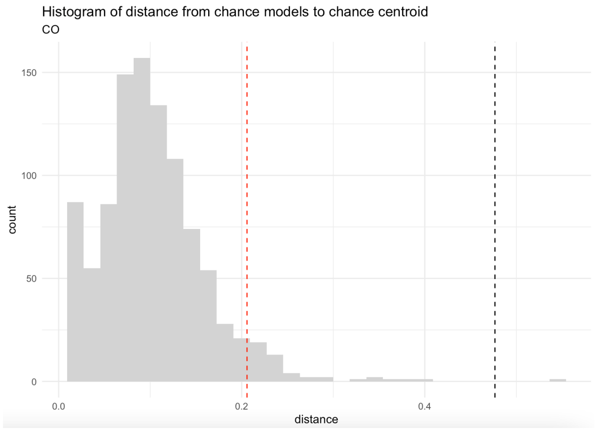


Fig. 5: Histogram of distances for CO.

indicating that the observed model is significantly different from chance. The effect size of this difference is very large, $d = 6.5$. The results suggest that there is a strong pattern of connections in this individual's discourse and the qualitative observations are theoretically saturated.

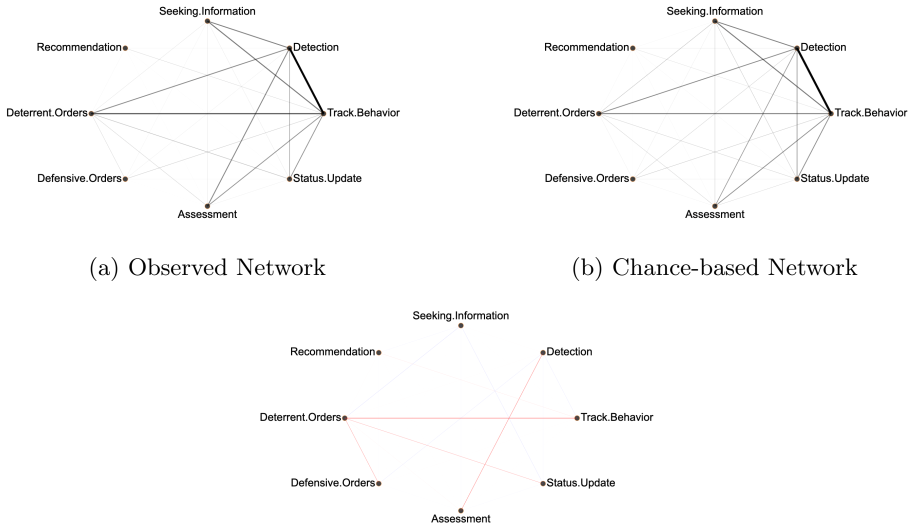
The results of the qualitative analysis of the individual whose observed network was highly similar to chance, a TAO, suggests that they had densely connected talk (see Table 3). The excerpt begins as the TAO provides a large amount of information about their highest priority tracks: They call out a La Combattante ship that has the warship within weapons range and helicopter ("SUPER PUMA HELO") approaching from the south that may be carrying anti-ship missiles ("POSSIBLE THREAT EXOCET"). Immediately after, the TAO seeks information from the warship's bridge to ask if they have any "VISUALS" (line 2). BRIDGE responds that they may see a merchant craft at bearing 270 but visibility is low (line 3). In the next line, the TAO asks if it is a big merchant ship (line 4) and BRIDGE tells them it is medium-sized (line 5). In the last line, the TAO provides identification information that the merchant ship corresponds to track 7007 on his DSS (line 7).

Table 3: TAO

Line	Speaker	Utterance
1	TAO	MY NUMBER ONE PRIORITY IS THE LA COMBANDANTE WE'RE WITHIN WEAPONS RELEASE RANGE ALSO TRACK 7037 COMING UP FROM THE SOUTH SHINING A PRIMUS 40 I HOLD THAT TO BE A SUPER PUMA HELO AIR UNKNOWN ASSUMED HOSTILE POSSIBLE THREAT EXOSET
2	TAO	BRIDGE/TAO DO YOU HAVE ANY VISUALS
3	BRIDGE	TAO THIS IS BRIDGE I KIND OF MAKE OUT CONTACT BEARING 270 REAL HAZY HARD TO MAKE OUT LOOKS LIKE A POSSIBLE MERCHANT CRAFT
4	TAO	BIG MERCHANT?
5	BRIDGE	MEDIUM-SIZED
6	TAO	MEDIUM. AYE
7	TAO	THAT CORRELATES TO TRACK 7007

The excerpt illustrates that this TAO's talk is densely connected, including connections among Detection ("THAT CORRELATES TO TRACK 7007"), Track Behavior ("COMING UP FROM THE SOUTH"), Assessment ("MY NUMBER ONE PRIORITY"), and Seeking Information ("DO YOU HAVE ANY VISUALS").

Figure 6 shows the observed (left) and chance-based (right) networks for this individual. Both networks reflect the connections discussed above. The difference



(a) Observed Network

(b) Chance-based Network

(c) Difference Graph

Fig. 6: Network Comparison for TAO

graph (bottom) includes only faint thin edges, suggesting that the connections in the two models occurred at similar rates. In other words, the difference graph suggests that the observed network is not very distinguishable from the chance-based network.

Figure 7 summarizes the results of the EVT for this TAO. The dashed-lines indicate that the distance between the observed model and its chance-based centroid (black dashed line) is less than the 95% of the distances between the chance-based models and the centroid (red dashed line). In particular, the empirical p value for this test is 0.64. The EVT indicates that this individual's model is *not* significantly different from chance and suggests that the qualitative observations are not theoretically saturated.

5 Discussion and Conclusion

In this paper, I introduced the EVT, a novel technique for providing statistical warrants of theoretical saturation in the context of ENA-based QE analyses. Traditionally, such warrants are derived from differences between two or more samples; however, the EVT is designed to provide warrants in situations where researchers examine single samples, such as the data overall or a given individual. To demonstrate the EVT, I conducted an overall qualitative analysis of the discourse of ADW teams and qualitative analyses of two particular individuals. Separate EVTs suggested that, overall, the qualitatively observed connections were theoretically saturated. For the individuals, the tests suggested that the ob-

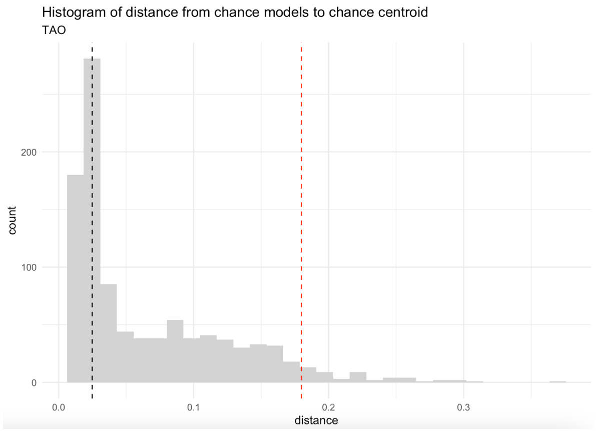


Fig. 7: Histogram of distances for TAO.

served connections for one individual were saturated while connections observed for the other were not. Thus, the technique has discriminatory power.

The case in which the EVT suggested that the observed connections were not theoretically saturated points to potential features of discourse that may impact whether theoretical saturation is achievable. The TAO's excerpt in Table 3 illustrates two important features of their discourse. First, the TAO dominates the discussion. Of the seven lines shown, five are from TAO, two are from the BRIDGE, and none are from the remainder of the team. Second, the TAO's talk is densely connected, including connections between Detection, Track Behavior, Assessment, and Seeking Information. To put these points in perspective, on average, the CO shown in Table 2 accounted for only 11% of the team's coded talk and 2% of the talk coded for two or more codes. In contrast, this TAO accounted for 36% of the team's coded talk and 20% of the talk coded for two or more codes—the highest percentages of any member of the team.

These measures suggest that highly active and densely connected individuals may be less likely to produce connections for which we can warrant theoretical saturation. In other words, individuals like this TAO may make connections that are statistically indistinguishable from random noise. This seems plausible—if an individual talks about everything all of the time, it is unlikely that they will make systematic patterns of connections. Future work will explore this hypothesis in more detail by varying features of the simulated data (e.g., code and talk frequencies) and relating those features to EVT outcomes.

The results I presented here have at least two important limitations. First, this study was based upon data collected from a single context. Future work will investigate the utility of the EVT in a variety of contexts. Second, the demonstration of the EVT relied on parameters derived from a specific modeling technique—ENA. However, the only constraint that the EVT places on models is that they produce vectors of co-occurrences. In turn, the EVT can accommodate other co-occurrence-based modeling techniques such as social network analysis, sequential pattern mining, and lag sequential analysis. Future work will explore the EVT with such techniques.

Despite these limitations, the results suggest that the EVT provides plausible warrants for theoretical saturation in QE. The EVT adds to the QE toolkit by providing a technique for warranting theoretical saturation in the absence of differences between samples. As such, it can strengthen quantitative ethnographic claims in situations where researchers are unable to or do not wish to compare samples.

References

1. Bressler, D. M., Bodzin, A. M., Eagan, B., Tabatabai, S.: Using epistemic network analysis to examine discourse and scientific practice during a collaborative game. Journal of Science Education and Technology, 28(5), 553-566 (2019).
2. Cohen, J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20, 37–46 (1960).
3. Csanadi, A., Eagan, B., Kollar, I., Shaffer, D. W., Fischer, F.: When coding-and-counting is not enough: Using epistemic network analysis (ENA) to analyze verbal data in CSCL research. International Journal of Computer-Supported Collaborative Learning, 13 (4), 419–438 (2018).
4. Glaser, B., Strauss, A.: The Discovery of Grounded Theory. Chicago: Aldine (1967).
5. Hod, Y., Katz, S., Eagan, B.: Refining qualitative ethnographies using epistemic network analysis: A study of socioemotional learning dimensions in a humanistic knowledge building community. Computers & Education, 156 (2020).
6. Johnston, J. H., Poirier, J., Smith-Jentsch, K. A.: Decision making under stress: Creating a research methodology. In J. A. Cannon-Bowers and E. Salas (Eds.), Making decisions under stress: Implications for individual and team training, pp. 39–59. American Psychological Association, Washington D.C (1998).
7. Marquart, C., Swiecki, Z., Collier, W., Eagan, B., Woodward, R., Shaffer, D. W. (2018). rENA: Epistemic Network Analysis (Version 0.2.1.2). https://cran.rstudio.com/web/packages/rENA/index.html
8. Shaffer, D. W.: Quantitative ethnography. Madison, WI: Cathcart Press (2017).
9. Shaffer, D. W., Collier, W., Ruis, A. R.: A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. Journal of Learning Analytics, 3 (3), 9–45 (2016).
10. Ruis, A. R., Rosser, A. A., Quandt-Walle, C., Nathwani, J. N., Shaffer, D. W., Pugh, C. M.: The hands and head of a surgeon: modeling operative competency with multimodal epistemic network analysis. American Journal of Surgery, 216 (5), 835-840 (2018).
11. Swiecki, Z., Ruis, A. R., Farrell, C. Shaffer, D. W.: Assessing individual contributions to collaborative problem solving: A network analysis approach. Computers in Human Behavior. Volume 104, 2020, 105876 (2020).
12. Wu, B., Hu Y., Ruis, A. R., Wang, M.: Analysing computational thinking in collaborative programming: A quantitative ethnography approach. Journal of Computer Assisted Learning, 35, 421-434 (2019).