JOURNAL OF LEARNING ANALYTICS

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

A Tutorial on Epistemic Network Analysis: Analyzing the Structure of Connections in Cognitive, Social, and Interaction Data

David Williamson Shaffer
Wisconsin Center for Education Research
University of Wisconsin–Madison, United States
dws@education.wisc.edu

Wesley Collier
Wisconsin Center for Education Research
University of Wisconsin–Madison, United States

A. R. Ruis
Wisconsin Center for Education Research
University of Wisconsin–Madison, United States

ABSTRACT: This paper provides a tutorial on epistemic network analysis (ENA), a novel method for identifying and quantifying connections among elements in coded data and representing them in dynamic network models. Such models illustrate the structure of connections and measure the strength of association among elements in a network, and they quantify changes in the composition and strength of connections over time. Importantly, ENA enables comparison of networks both directly and via summary statistics, so the method can be used to explore a wide range of qualitative and quantitative research questions in situations where patterns of association in data are hypothesized to be meaningful. While ENA was originally developed to model cognitive networks — the patterns of association between knowledge, skills, values, habits of mind, and other elements that characterize complex thinking — ENA is a robust method that can be used to model patterns of association in any system characterized by a complex network of dynamic relationships among a relatively small, fixed set of elements.

Keywords: Epistemic network analysis, network models, learning assessment

1 INTRODUCTION

The purpose of this paper is to provide a tutorial on *epistemic network analysis* (ENA), a novel method for quantifying, visualizing, and interpreting network data (Shaffer et al., 2009; Rupp et al., 2009; Rupp, Gustha, Mislevy, & Shaffer, 2010; Rupp, Sweet, & Choi, 2010; Svarovsky, 2011; D'Angelo, Clark, & Shaffer, 2012; Orrill, Shaffer, & Burke, 2013; Nash & Shaffer, 2013; Knight, Arastoopour, Shaffer, Buckingham Shum, & Littleton, 2014; Chesler et al., 2015; Andrist, Collier, Gleicher, Mutlu, & Shaffer, 2015; Hatfield, 2015; Arastoopour, Shaffer, Swiecki, Ruis, & Chesler, 2016; Collier, Ruis, & Shaffer, 2016; Siebert-Evenstone et al., 2016; Quardokus Fisher, Hirshfield, Siebert-Evenstone, Arastoopour, & Koretsky, 2016). ENA is a set of techniques that identifies and measures connections among elements in

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

coded data and represents them in dynamic network models. Such models illustrate the structure of connections and measure the strength of association among elements in a network, and they quantify changes in the composition and strength of connections over time. Importantly, ENA enables comparison of networks both directly and via summary statistics, so it can be used to explore a wide range of qualitative and quantitative research questions in situations where patterns of association in data are hypothesized to be meaningful.

ENA was originally developed to model *cognitive networks*: the patterns of association between knowledge, skills, values, habits of mind, and other elements that characterize complex thinking. However, ENA is a versatile method that can be used to model patterns of association in any system characterized by a complex network of dynamic relationships among a relatively small, fixed set of elements. That is, it is a method optimized for the analysis of networks too large for multivariate parametric techniques, such as latent class models, but not so large that they require analysis exclusively via summary statistics, as with many traditional network analysis techniques (Batagelj & Mrvar, 1998).

In what follows, we will 1) describe the theory behind ENA, 2) detail the process by which ENA creates network models, and 3) explore example analyses using the ENA WebKit Version 4.[1] The ENA WebKit is an online tool that enables users to analyze and visualize network data using ENA. For rhetorical clarity, we will combine the discussion of methods, examples, and tool, but we stress that the *method* with which ENA models network data is not dependent on the interface or features of the ENA WebKit tool.

2 STRUCTURES OF CONNECTIONS AND NETWORK THEORY

In this section, we provide some background on the development of ENA and discuss the theoretical foundations of the method. It is important to note, however, that while ENA was *originally designed* to address certain challenges in learning analytics, the method is *not limited* to analyses of learning data. For example, ENA has been used to model social gaze coordination (Andrist, Collier, Gleicher, Mutlu, & Shaffer, 2015), surgeons' communication during operative procedures (D'Angelo, 2015), neuronal coactiviation in functional MRI data (Collier, 2015), and conceptual semantics in historical records (Ruis, 2016). In what follows, we first describe the theory of ENA in the context of cognitive networks, and then we extend this discussion to describe the general properties of networks that can be modelled using the ENA approach.

ENA was initially developed to model cognitive networks based on a fundamental assumption in some theories of learning analytics: that *the structure of connections among cognitive elements is more important than the mere presence or absence of those elements in isolation*. DiSessa (1988), for example, characterizes learning as a process whereby *phenomenological primitives* — isolated elements of experiential knowledge — are connected through theoretical frameworks to develop not just new

[1] To access the ENA WebKit and other ENA resources, visit http://www.epistemicnetwork.org/live/login.html

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

knowledge but deep, systematic understanding. Shaffer (2006, 2007, 2012) similarly characterizes learning as the development of an *epistemic frame*: a pattern of associations among knowledge, skills, habits of mind, and other cognitive elements that characterizes *communities of practice* (Wenger, 1999; Hutchins, 1995; Shaffer, 2004), or groups of people who share similar ways of framing, investigating, and solving complex problems. The important point that both DiSessa and Shaffer make is that learning is defined not by the *possession* of isolated bits of knowledge and other competencies but by the *structure of connections among them*.

Based on these theoretical foundations, ENA was designed to model the structure of connections among cognitive elements. A "classical" approach to modelling data about cognitive processes is to use multivariate analyses to look at correlations among knowledge, skills, and attitudes (Ruppet et al., 2009; Lattin, Carroll, & Green, 2003), including the presence or absence of different elements of cognition and the ways in which different elements in the model interact. However, traditional statistical analyses suffer from two weaknesses in addressing structures of connections of the kind hypothesized by DiSessa and Shaffer. First, the number of interactions rises exponentially in relation to the number of elements, so even models with a moderate number of interacting elements require very large amounts of data. Second, the *structure* of connections is fundamentally a question about *networks*. In a network analysis, the objects of interest are the interactions between elements of the model. Those interactions are typically (and in some ways necessarily) secondary to the elements themselves in multivariate models (Collier, Ruis, & Shaffer, 2016; Jaccard & Turrisi, 2003).

Network analysis is thus a good approach to modelling cognitive interactions, but most network analysis methods are not designed for networks with a relatively small, fixed set of nodes characterized by highly dynamic and often dense patterns of weighted connections. Rather, many extant network analysis techniques are optimized to analyze very large networks, with thousands, millions, or even billions of nodes. Such networks can only be analyzed via summary statistics, such as structural cohesion, clustering coefficients, or density. This is problematic, however, because such analyses can easily obscure important differences in the structure of the underlying networks. Consider a simple example. Figure 1 shows two weighted networks. Each network has the same weighted density, based on the number and strength of connections relative to the number of nodes, but there are significant differences between their patterns of connections. In other words, if we looked only at network density, the two networks would appear equivalent, despite the fact that their structures are not at all similar. Of course, there are other summary statistics that would show differences between the two networks; for example, counting the number of unconnected nodes. But it can be difficult to interpret the differences in a large set of networks in terms summary statistics, and many summary statistics, including degree, betweenness, and closeness centrality metrics, as well as transitivity, show little variance in the dense patterns of connection that often characterize cognitive networks.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

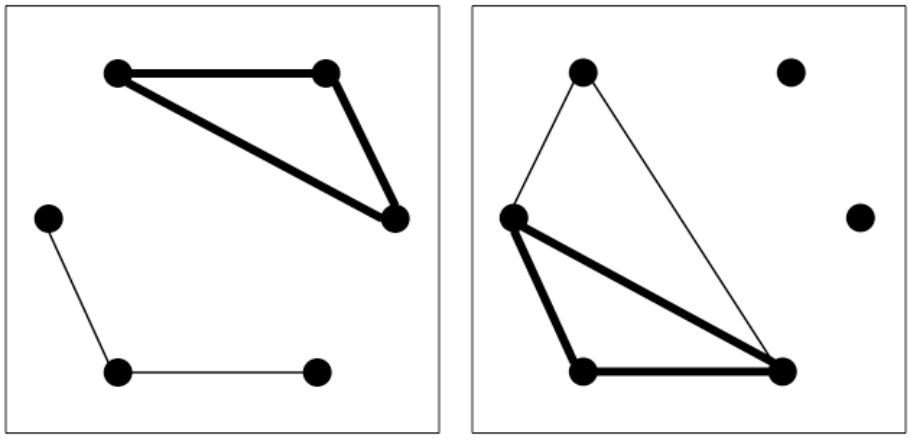


Figure 1. Two weighted networks with the same weighted density but significantly different patterns of connections.

ENA is thus optimized for the analysis of networks too large to be analyzed using multivariate parametric techniques but not so large that they can be analyzed only via summary statistics.

Whether ENA is used to model cognitive networks or any other kind of network, a key assumption of the method is that the structure of connections in the data is most important in the analysis, whether that is the structure of cognitive connections that students make while engaged in complex problem solving, the structure of connections among regions of the brain while participants perform simple tasks, the structure of connections in eye-gaze behaviour when demonstrating a new procedure, or any other context where the structure of connections is meaningful.

3 EXPLORING AND INTERPRETING NETWORKS USING ENA

To understand how ENA works, it may be helpful to begin with an example, which will provide context for a more detailed discussion of ENA methodology in subsequent sections. Consider Figure 2, which shows the cumulative cognitive network of a first-year undergraduate student (Student A) who participated in an online engineering design simulation. The student's network models the structure of connections among knowledge, skills, and other aspects of engineering practice: the elements of the student's engineering epistemic frame. This model of Student A's epistemic frame during the simulation shows a number of connections among knowledge elements, such as *knowledge of actuators*, *knowledge of range of motion* (ROM), and *knowledge of data*; skill elements, such as the *skill of collaboration* and the *skill of design*; and the *identity of an engineer*.[2] The network is also *weighted*: darker, thicker lines indicate connections that the student made more often (stronger connections), whereas lighter, thinner lines indicate connections that the student made less often (weaker connections).

[2] For a complete list of codes with definitions and examples, see Shaffer and Arastoopour (2014). The elements of the epistemic frame of a particular community of practice can be identified *a priori* from a theoretical or empirical analysis, or from an ethnographic study of the community in action.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

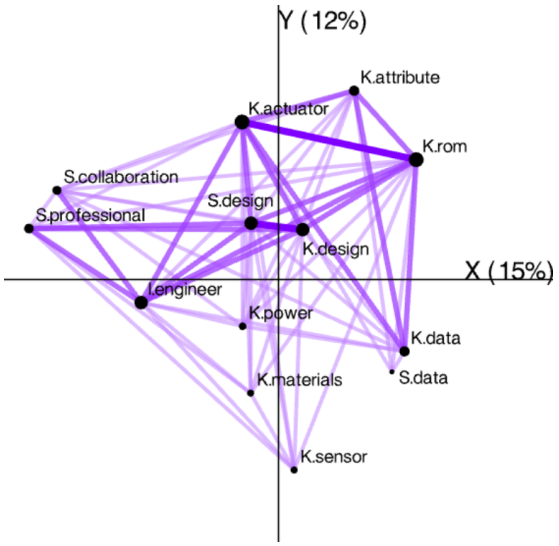


Figure 2. Cognitive network of a first-year undergraduate engineering student (Student A) representing the connections the student made while solving a simulated engineering design problem.

While we can draw some conclusions from the structure of connections in a single network, in many cases, the salient features become more apparent in comparison with other networks. Figure 3 shows the structure of connections among engineering epistemic frame elements of a second student (Student B) during the same engineering design simulation. Like Student A, Student B made a number of cognitive connections among knowledge and skill elements, but Student B's network exhibits more connections overall as well as connections to additional elements, most notably the *epistemology of design*, the *epistemology of data*, the *epistemology of the client*, and the *value of the client*. These additional elements reflect the extent to which Student B was explicitly epistemic — meaning that she explained and justified her thinking — and also the fact that Student B considered the role of the client in the design process. In other words, Student B exhibits a denser, more diverse, and in some important ways more sophisticated structure of cognitive connections than Student A.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

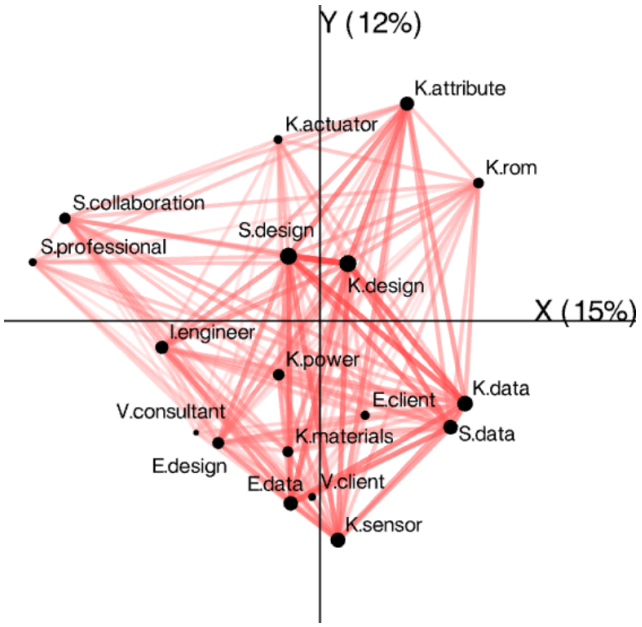


Figure 3. Cognitive network of a first-year undergraduate engineering student (Student B) representing the connections the student made while solving a simulated engineering design problem.

To understand what makes this comparison possible in an ENA analysis, it is helpful to superimpose the two networks (see Figure 4a). Note that the nodes in the networks of both students appear in exactly the same places in the network projection space — and not only for these two students, but for all the students in the dataset.[3] The placement of nodes in fixed positions allows for meaningful comparison of the patterns of connections in two or more networks, but it also allows for interpretation of the projection space itself. In this example, we can characterize the space as follows: toward the left are basic interpersonal skills, such as collaboration and professionalism; toward the bottom are values and elements related to epistemology, or the justification of decisions; toward the right are elements related to data; and toward the top are elements related to the knowledge and skill of design and to specific design attributes.

[3] The theory and mathematics behind node placement are discussed in a subsequent section.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

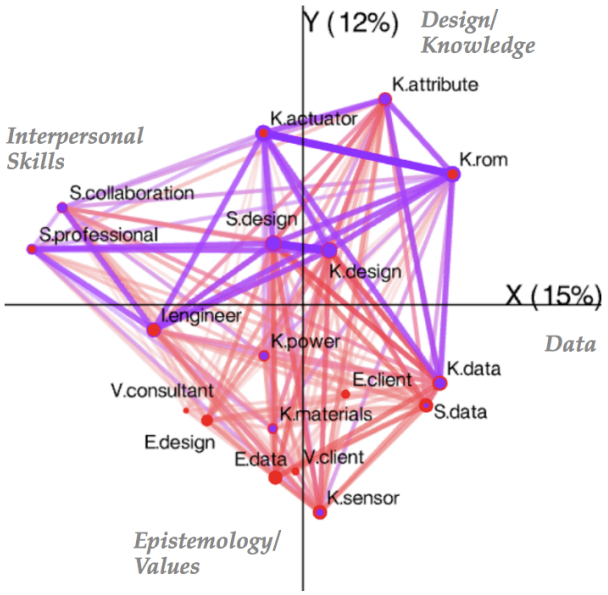


Figure 4a. The superimposed cognitive networks of two first-year undergraduate engineering students (Student A in purple, Student B in red) representing the connections each student made while solving a simulated engineering design problem. The space can be interpreted as shown (gray italics) based on the locations of the nodes.

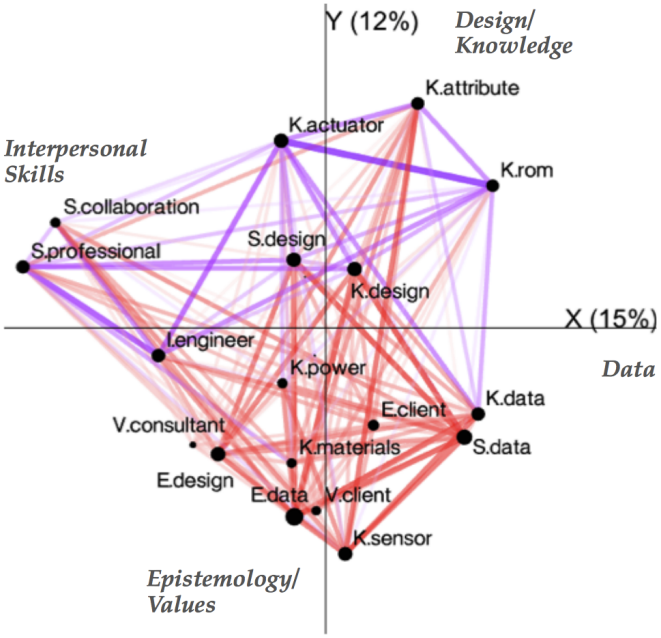


Figure 4b. The subtracted cognitive network of two first-year undergraduate engineering students (Student A in purple, Student B in red), in which Student B's connections are subtracted from Student A's connections. Connections coloured purple are stronger in Student A's cognitive network, and connections coloured red are stronger in Student B's cognitive network.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

Because of the fixed node positions, ENA can also construct a subtracted network, which enables identification of the most salient differences between two networks. To do this, ENA subtracts the weight of each connection in network one from the corresponding weighted connection in network two, then visualizes the difference. Darker, thicker lines indicate larger differences in connection strength, and lighter, thinner lines indicate smaller differences in connection strength; each line is coloured to indicate which of the two networks contains the stronger connection. Figure 4b shows the network resulting from the subtraction of Student B's cognitive network from Student A's. The purple lines indicate connections that are stronger in Student A's network, and the red lines indicate connections that are stronger in Student B's network. For example, the connection between *knowledge of actuators* and *identity of an engineer* was stronger in Student A's network, while the connection between *knowledge of design* and *epistemology of data* was stronger in Student B's network. Overall, the subtracted network shows that Student A has the strongest connections (relative to Student B) in the upper part of the space, and Student B has the strongest connections (relative to Student A) in the lower part of the space.

Comparing network projections directly, as in the examples above, is useful if we want to compare only two or perhaps three networks at a time. But what if we want to compare dozens or even hundreds of networks simultaneously? To do that requires summarizing pertinent network features but, as discussed above, *it is critical for that summary to accurately reflect the salient features of the structure of connections*. ENA does this by representing each network as a single point in the space, where each point is the *centroid* of the corresponding network.

The centroid of a network is much like the centre of mass of an object. Specifically, the centroid of a network graph is the arithmetic mean of the edge weights of the network model distributed according to the network projection in space. In Figure 5, for example, two networks are shown with the same connectivity but differently weighted connections, as indicated by the widths of the lines connecting the nodes. In both networks, the weights of connections AB and AC are the same, but the network on the right has more heavily weighted connections BC, BD, and CD than the network on the left. As a result, the centroid of the network on the right is higher and farther and to the right than the centroid of the network on the left. The critical point here is that *the centroid of an ENA network summarizes the network as a single point in the projection space that accounts for the structure of connections in the specific arrangement of the network model*.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

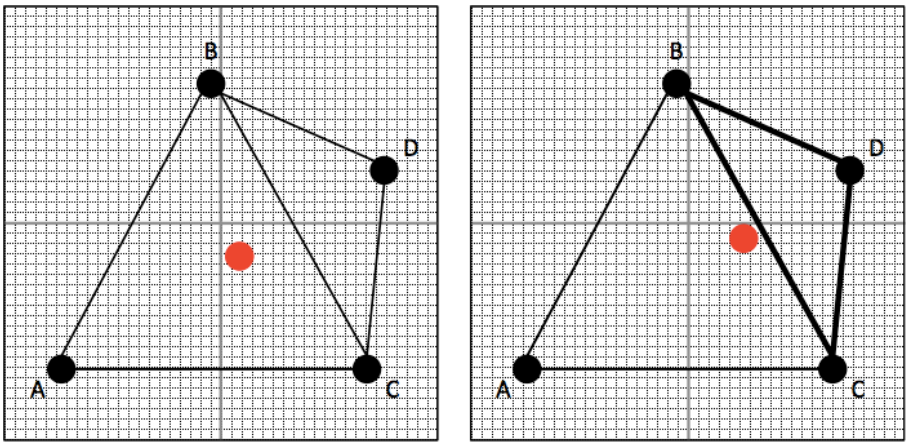


Figure 5. Two networks with the same nodes connected but with differently weighted connections and their corresponding centroids in red. Note that the centroid in the network on the right is higher and farther right due to the heavier weights of connections BC, BD, and CD.

If we return to the example of the two engineering students, we can now view each student's network as a single point in the projection space (see Figure 6). Note that because the centroid of a network is the arithmetic mean of the edge weights in the projection, the position of each centroid reflects the structure of connections in each network. Student A made connections mostly among elements in the upper part of the projection space, while Student B also made connections to elements in the lower part of the projection space. As a result, the centroid of Student A's network appears higher than the centroid of Student B's network.

This makes it possible to compare a number of different networks simultaneously because centroids located close together represent networks with similar patterns of connections, while centroids located far apart represent networks with different patterns of connections. Furthermore, once we have characterized the projection space by looking at the node positions, we can identify the nature of the most prominent connections in a network based on the position of its centroid. In other words, *we can interpret changes in the centroid of a network relative to the weights of the lines in the network, which allows us to interpret the dimensions of the space*. To illustrate this point, consider the centroid of a third student (Student C) in Figure 6.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

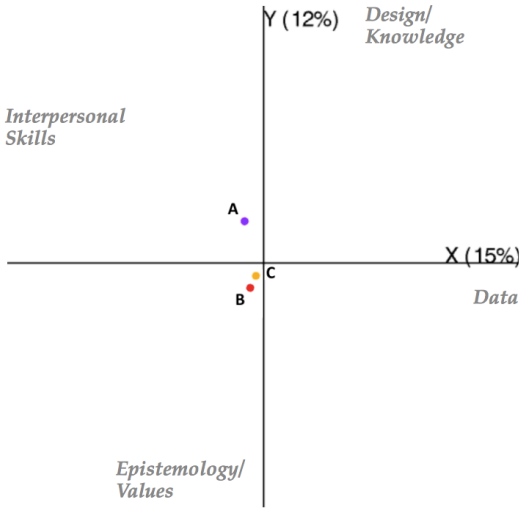


Figure 6. Centroids of the cognitive networks of three first-year undergraduate engineering students (Student A in purple, Student B in red, and Student C in yellow) representing the mean locations of their weighted networks in the projection space.

As Figure 6 shows, Students B and C have centroids located very close together, so we should expect that their networks have very similar structures. When we look at the two students' networks (Figure 7), they are in fact highly similar.[4]

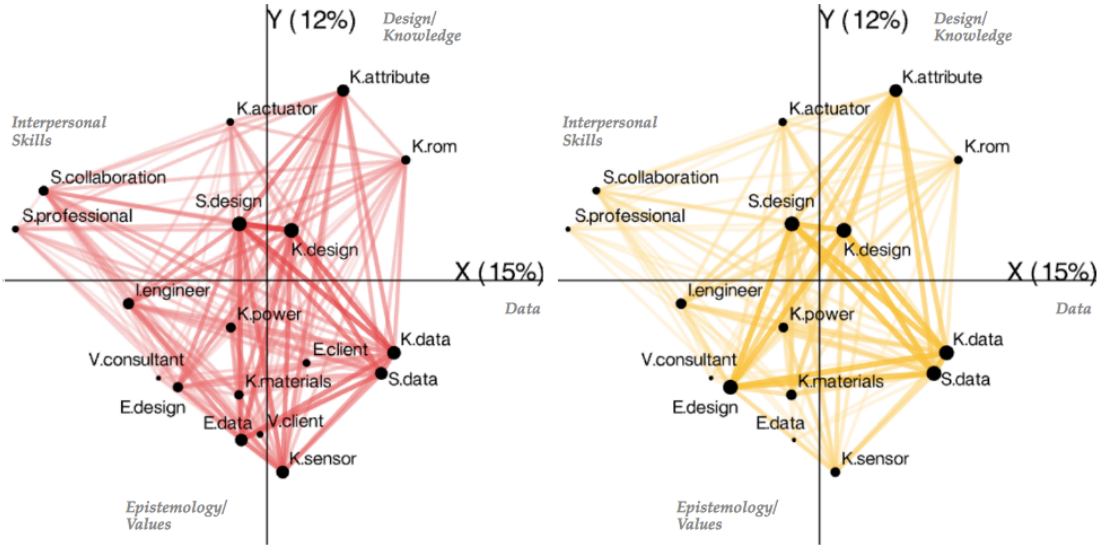


Figure 7. Cognitive networks of two first-year undergraduate engineering students (Student B in red, left, and Student C in yellow, right). Note that their patterns of connection are highly similar.

[4] It is, of course, possible to construct networks with very different structures of connections that share similar centroids. For example, a network with many connections might have a centroid near the origin; but the same would be true of a network that had only a few connections at the far right and a few at the far left of the network space. For obvious reasons, no summary statistic in a dimensional reduction can preserve all of the information of the original network.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

One of the advantages of the centroid is that it enables easy comparison of a large number of networks via summary statistics. For example, one question we might ask is whether students' cognitive networks are different in the first half of the engineering design simulation than in the second half. In other words, do they think about the engineering design problem differently in different parts in the design process? When we look at the centroids of students' networks from the first and second halves of the simulation (see Figure 8), we can see that there does appear to be some difference on the second (*y*) dimension. To determine if the difference is statistically significant, we can calculate the mean positions of the two groups with their confidence intervals on both dimensions and perform an independent samples *t* test. This indicates that the two groups are significantly different on the second dimension ($mean_A = 0.079$, $mean_B = -0.079$; $p < 0.001$; $t = 4.231$). In Figure 8, the mean positions of the two groups are shown as coloured squares; the boxes around those squares indicate the confidence intervals, where the width of the box is the confidence interval on dimension one (*x*), and the height of the box is the confidence interval on dimension two (*y*).

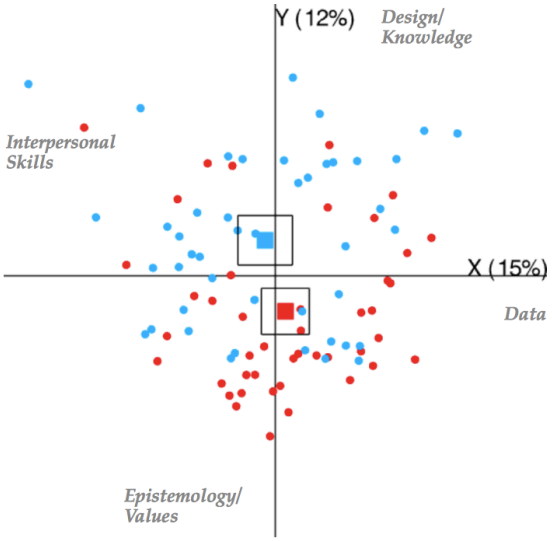


Figure 8. Students' centroids from the first (blue) and second (red) halves of an engineering design simulation with means and confidence intervals. Note that the two groups are significantly different on the second dimension.

To determine what factors account for this difference between the two groups, we can compare their mean networks (see Figure 9). In the first half of the simulation, students made connections mostly among basic knowledge and skills, while in the second half, students made more connections to epistemological elements and considerations of design.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

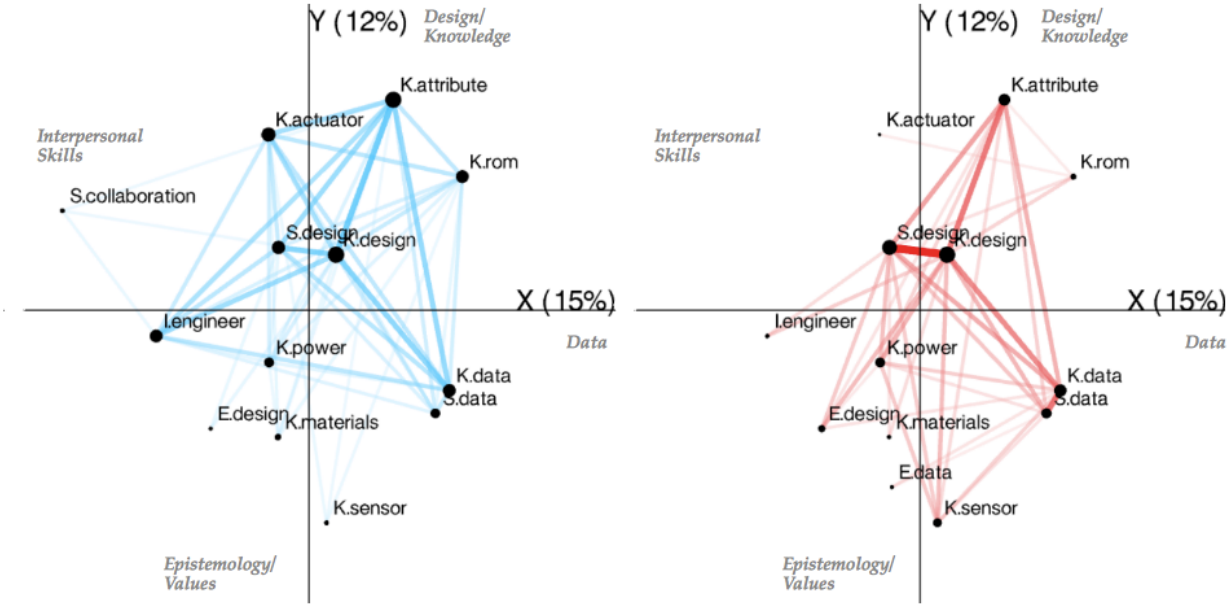


Figure 9. Mean networks of students' from the first (blue, left) and second (red, right) halves of an engineering design simulation.

The important point here is that the correspondence of the centroids with the structures of connections in the network projections allows comparison of multiple networks via common summary statistics *without the loss of important information about the structures of connections*. As with the previous example, we can use the network projections to identify the differences in the structures of connections that account for statistically significant differences between units.

Now that we have explored some of the basic elements of how ENA works, in the next section we will examine in greater detail how the technique identifies connections in the data and models them as a network.

4 DATA FORMATTING AND SEGMENTATION FOR ENA ANALYSIS

Before we describe how ENA models network data, it will help to provide more information about the context of the previous examples and the nature of the dataset from which they are drawn. The students whose cognitive networks are modelled in the previous section participated in an engineering *virtual internship* (Chesler et al., 2015; Arastoopour, Shaffer, Swiecki, Ruis, & Chesler 2016; Shaffer, 2007; Chesler, Arastoopour, D'Angelo, Bagley, & Shaffer, 2013), an online learning simulation in which students solve a realistic engineering design problem in a simulated work environment. In this case, the students participated in the virtual internship *RescuShell* (Chesler et al., 2015). In *RescuShell*, students work as interns at a fictitious company to design the robotic legs for a mechanical exoskeleton. Over the course of the internship, students review internal technical documents from the company, conduct background research, and examine research reports based on actual experimental data. Then, they develop hypotheses, test those hypotheses in the provided design space, and analyze the results, both

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

individually and in small teams. Students also become knowledgeable about consultants within the company who have a stake in their design choices. These consultants value different performance metrics for the exoskeleton, and no one design can satisfy all the consultants' requests. During the final days of the internship, students present their final prototypes and justify their design decisions.

Throughout the internship, students communicate via online instant message (chat) with other members of their team. These chats are recorded in a *log file*, which provides raw data for analysis. In order to model the cognitive connections that students make over the course of the internship using ENA, these data must be *formatted* appropriately and *coded* for the cognitive elements of interest.

To see how this works, it will help to look at an example excerpted from an actual coded log file of student work in *RescuShell* (see Figure 10). The source data (column labelled "text") consist of student chats among members of a project team (Group 5).

METADATA COLUMNS						CODE COLUMNS				
	STANZA COLUMN	UNIT COLUMN	RAW DATA COLUMN	RAW DATA COLUMN	RAW DATA COLUMN					
Line	activity	Group	created	text	username	E.data	E.design	I.engineer	K.actuator	K.power
1	Design Batch	5	10/17/13 9:38	@Kevin: did you find it fairly simple to meet at least the minimum requirements of the consultants with your prototypes?	josephk	0	0	1	0	0
2	Design Batch	5	10/17/13 9:39	Nope. The only requirement series elastic was good for was safety.	kevin	0	0	0	1	0
3	Design Batch	5	10/17/13 9:39	@Nassim Tehrani: is there any way that we can share our final batch results with eachother?	josephk	0	0	1	0	0
4	Design Batch	5	10/17/13 9:40	Alright so it's fair to say that series elastic is pretty much out of the question unless we were able to use a bunch of ROM that would make the battery life unrealistic?	josephk	0	0	0	1	1
5	Design Batch	5	10/17/13 9:40	I don't mean to interrupt, but I would suggest the electric actuator.	christian	0	0	0	1	0
6	Batch Analysis	5	10/22/13 10:03	I thought these results were very interesting	josephk	0	0	0	0	0
7	Batch Analysis	5	10/22/13 10:04	I was really surprised by the payloads	kevin	0	0	0	0	0
8	Batch Analysis	5	10/22/13 10:04	Yeah. I think we should go with Final Final 5 because it has the highest payload and is the most well rounded	luis	1	1	0	0	0
9	Batch Analysis	5	10/22/13 10:04	yes I mentioned that in my notebook the electric and pneumatic were very similar	josephk	0	0	1	1	0
10	Batch Analysis	5	10/22/13 10:05	What made this prototype better than all the other options?	nassim	0	0	0	0	0
11	Batch Analysis	5	10/22/13 10:05	I think it's too bad that we don't get to make a final one to test if we could I'd rather test more with the composite material	josephk	0	0	0	0	0

Figure 10. Excerpt of a coded log file containing evidence of student work from the engineering virtual internship *RescuShell*.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

One research question of interest in this case may be: *How are these students making connections between different aspects of engineering in the* RescuShell *design space?* The engineering epistemic frame elements (i.e., the nodes of the network model) can be seen in the code columns (the five rightmost columns): knowledge (K) of actuators and power sources in mechanical exoskeletons, the identity (I) of an engineer, and the epistemology (E), or decision-making and justification processes, of engineering data and design. Of course, for the purposes of this example, these are just a subset of the engineering issues relevant to this particular problem and its solution. The full dataset used to compute the networks above has both more lines of talk (nearly 4,000) and more codes (20).[5]

There are two important points to consider in this dataset. First, sometimes we can see connections among the elements of the design space (the codes) in a single line of data. In Line 4, for example, JosephK references knowledge of both actuators and power sources:

> *JosephK:* Alright so it's fair to say that series elastic is pretty much out of the question unless we were able to use a bunch of ROM that would make the battery life unrealistic?

While each row is coded for the elements of engineering design referenced in that row, the lines of talk are also related to one another. In Line 9, for example, JosephK talks about recording actuator properties in his own notebook:

> *JosephK:* yes I mentioned that in my notebook the electric and pneumatic were very similar.

But he is actually making a connection between those elements of the design space, and the data and design considerations Luis used in his decision making process in the preceding chat (Line 8):

> *Luis:* Yeah. I think we should go with [the design labelled] Final Final 5 because it has the highest payload and is the most well rounded.

We thus want to look at relations among elements not just within specific turns of talk but across turns of talk.

The second important point is that we don't necessarily want to look at relations among elements across *all* turns of talk. For example, an examination of the columns labelled "Activity" and "Created" in Figure 10 shows that there are actually two separate conversations here. Both involved the same group of students (Group 5), but the conversations took place on two different days while the students were working on two different activities. On the first day, they designed prototypes and were deciding which ones to send to a fabrication lab for testing. On the second day, they were reviewing the results of those tests. So in this case, although JosephK's comment in Line 9 should be connected to Luis's comment in Line 8, it does not make as much sense for it to be linked to the comments in Lines 1–5, which were made in a conversation from the previous week.

[5] For more information on this dataset, see Shaffer and Arastoopour (2014). For a detailed discussion of formatting data for ENA, see Shaffer, 2014a.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

Of course, precisely which rows should be linked for the purpose of analysis will depend on the nature of the data and the research question being investigated. From the perspective of data formatting, the key issue is that there has to be some column or columns in the data table that indicate how to segment the data for analysis. Those segments are referred to as *stanzas*.

Stanza is a term from discourse analysis (Gee, 1999), but it is similar to the stanzas of a poem. The idea of a stanza is that all the lines within a stanza are related to one another, while lines in different stanzas are not related to one another. Consider the poem "The Walrus and the Carpenter" from Lewis Carroll's *Through the Looking-Glass* (1871). The poem begins:

> *The sun was shining on the sea,*
> *Shining with all his might:*
> *He did his very best to make*
> *The billows smooth and bright—*
> *And this was odd, because it was*
> *The middle of the night.*

The first stanza is about the sun, and all the lines in the stanza refer to the sun. This is distinct from the next stanza, which is about the moon, and all the lines in that stanza refer to the moon. Similarly, the third stanza is about the sea, and so forth. The lines within a stanza are thus closely related, while the lines in different stanzas are not.

In the context of an ENA analysis, relations between elements can be explicitly specified in the model, but the idea of a stanza is that elements present in the same stanza are conceptually connected, while elements in different stanzas are not. In other words, elements that *co-occur* in a stanza are conceptually linked. Researchers (Chesler et al., 2015; Landauer, McNamara, Dennis, & Kintsch, 2007; Dorogovtsev & Mendes, 2013; Lund & Burgess, 1996; i Cancho & Solé, 2001) have shown that co-occurrences of concepts in a given segment of discourse data are a good indicator of cognitive connections, particularly when the co-occurrences are frequent (Newman, 2004). In Figure 10, for example, the stanzas are defined by the activity the students were engaged in (second column from the left). Based on this, rows 1–5 are grouped into one stanza, and rows 6–11 are grouped into a second stanza. Because interpreting the structure of connections is the ultimate goal of an ENA analysis, determining which connections in the data are meaningful is critical. The stanza determines which connections are meaningful for the purpose of a given analysis.

Abstracting from this specific example of modelling a cognitive network, *in general* ENA models networks using *stanza-based interaction data.* Stanza-based interaction data are data that contain information about 1) a set of objects, 2) the way they relate to one another, and 3) a series of stanzas that 4) reveal evidence about the relations between the objects. In the example network analyses above, the objects were elements of engineering thinking, the relations were the way in which these elements of engineering thinking were used together in discourse, and the stanzas were lines of chat

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

during discrete design activities, and the evidence of connection between elements of engineering thinking is their co-occurrence within a stanza of chats.

In general, however:

1. Objects can refer to people, concepts, or anything whose structure of connections is being modelled.
2. Relations between objects can refer to associations, such as strength of social ties, conceptual similarity, or any connection, interaction, or association that links one object to another.
3. Stanzas can be units of time, steps in a process, or any way of identifying a unit in the data for quantifying relations between objects.
4. Evidence refers to any specific elements of the data that can be used to identify the relations being modelled.

Segmenting data into stanzas thus makes it possible to accurately model the relations between objects based on evidence in the data.

5 MODELLING NETWORKS WITH ENA

In this section, we explore in more detail how ENA creates a network model from stanza-based interaction data.

5.1 Creation of adjacency matrices representing the co-occurrence of codes in each stanza

To identify the relations among objects in the data, ENA creates a series of adjacency matrices, where each matrix represents the co-occurrence of codes in a single stanza. If the stanzas consist of multiple rows in the data table, all the rows that comprise each stanza are accumulated into a single row in a new data table, which contains all the codes present in the rows of that stanza. Each code is summed, and the summation can be either binary or weighted. A binary summation assigns a one to each code that appears at least once in the rows of the stanza, and a zero to each code that doesn't appear in any row of the stanza; a weighted summation accumulates each code (see Figure 11).

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

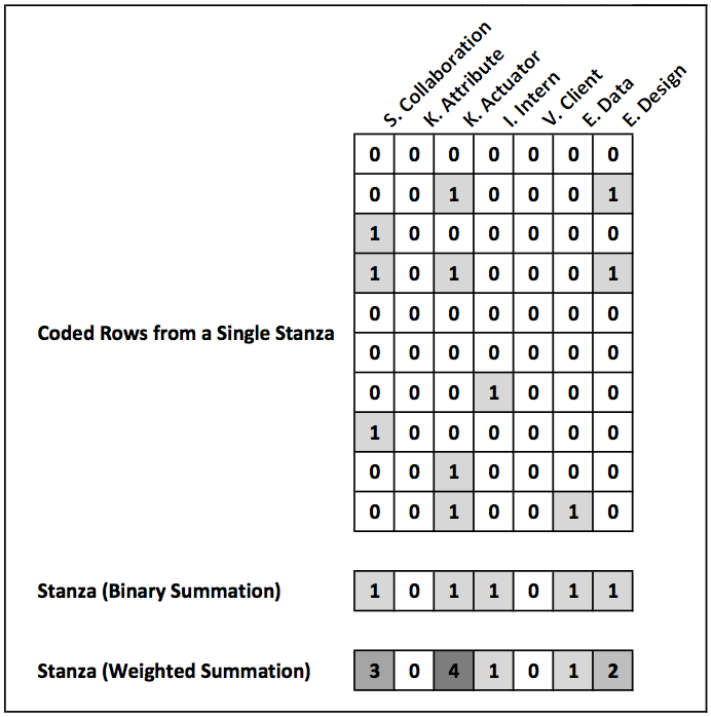


Figure 11. The accumulation of coded rows of data into a single coded row for each stanza.

In the sample excerpt from the previous section (Figure 10), there were two stanzas based on activity: Design Batch and Batch Analysis. Figure 12 illustrates how the rows in those stanzas are collapsed into single lines using binary summation. In this case, binary summation makes sense because there is no reason to assume, for example, that if a student says something twice as often, then it means that the student understands it twice as well.

Activity	Group	E.data	E.design	I.engineer	K.actuator	K.power
Design Batch	5	0	0	1	1	1
Batch Analysis	5	1	1	1	1	0

Figure 12. The accumulation of coded rows of data into a single coded row for each stanza based on the dataset in Figure 10.

Because we are ultimately interested in the relations among objects, ENA uses the new data table, where each stanza is represented as a single row, to produce an adjacency matrix for each stanza. If two codes both occur in the same stanza, a one is placed in the cell in the adjacency matrix for that stanza that corresponds to the intersection of the two codes; cells for codes that do not co-occur in the stanza receive a zero. (If stanzas were accumulated via weighted summation, a weighted adjacency matrix is produced. In that case, the adjacency matrix for each stanza then represents not the presence or absence of the co-occurrence of each pair of codes *ij*, but the square root of the product of the value of

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

code i and code j in the stanza.) ENA also zeroes out the diagonal, because the intersection or co-occurrence of codes with themselves does not represent a connection between two different elements of the model. The stanzas shown in Figure 12 would thus produce the adjacency matrices shown in Figure 13. Note that the matrices are symmetrical because the interaction data is symmetrical. ENA can also model asymmetrical (i.e., directional) interaction data, discussed in greater detail in a subsequent section (see also Shaffer, 2014a, 2014b).

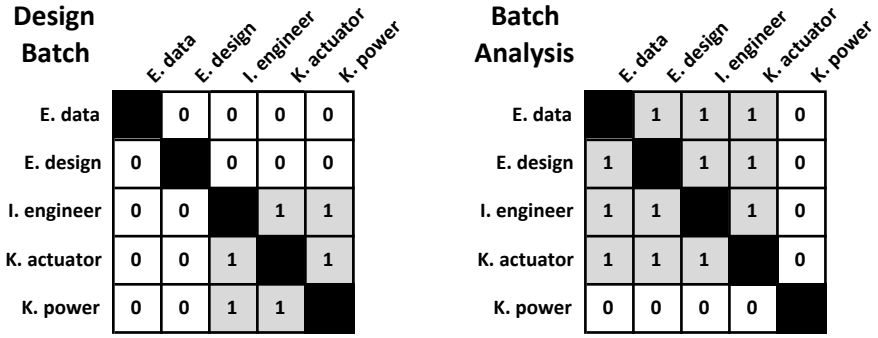


Figure 13. Adjacency matrices, indicating the co-occurrence of codes, for the two stanzas shown in Figure 12.

Each adjacency matrix thus represents the relations among the different objects in one stanza. There is one adjacency matrix per unit per stanza in the dataset, which means that each unit in the dataset is represented by a series of adjacency matrices.

5.2 Accumulation of adjacency matrices for each unit of analysis

To identify the structure of connections in the data, ENA sums the adjacency matrices for each unit of analysis *u* into a cumulative adjacency matrix, C^u, where each cell C^u_{ij} represents the number of stanzas in which codes i and j were both present. In the excerpted dataset shown in Figure 10, the unit of analysis is "group" because we are interested in modelling the structure of cognitive connections in the conversations that each group had while solving the *RescuShell* design problem. Because both of the stanzas in this dataset were from Group 5, they are accumulated as shown in Figure 14.

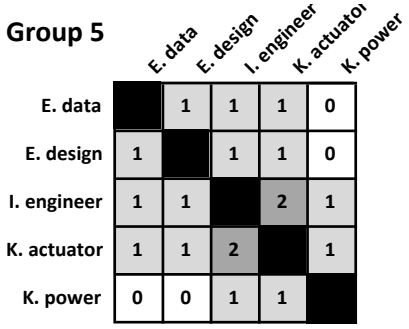


Figure 14. Cumulative adjacency matrix of Group 5, which sums the adjacency matrices shown in Figure 13.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

Of course, this is a simplified example. The full dataset from which Figure 10 was excerpted contains 17 activities and nearly 4,000 rows of data coded for 20 engineering epistemic frame elements. Thus, the cumulative adjacency matrix for a given group would be the summation of up to 17 adjacency matrices, each 20×20 with 190 (20 choose 2) possible unique co-occurrences of codes.

Once ENA has created the set of cumulative adjacency matrices for all units in the data, each matrix is then converted into an adjacency vector, V^u, by copying the cells from the upper diagonal of the matrix row by row into a single vector. For example, the matrix in Figure 14 would be represented by the vector [1,1,1,0,1,1,0,2,1,1], where the third element of the vector (which is a 1) corresponds to the cell in row 1 column 4, and the eighth element of the vector (which is a 2) corresponds to the cell in row 3 column 4.

These vectors thus exist in a high-dimensional space, V, such that each dimension of V represents a unique pairing of two codes; the position of the vector V^u representing cumulative adjacency matrix C^u on the dimension corresponding to the unique pairing of codes i and j in V is given by C^u_{ij}.

In this way, the structure of connections among objects for each unit of analysis is represented by an adjacency vector in high-dimensional space that contains all the unique co-occurrences of codes summed across all the stanzas.

5.3 Spherical normalization of adjacency vectors to calculate relative frequencies of co-occurrence

In this high-dimensional ENA space, each adjacency vector represents the pattern of associations of a single unit, and the *length* of a vector is potentially affected by the number of stanzas contained in the unit of analysis. More stanzas are likely to produce more co-occurrences, which result in longer vectors. This is problematic because two vectors may represent the same *patterns* of association, and thus point in the same *direction*, but represent different *numbers* of stanzas, and thus have different *lengths*. As we noted above, saying something twice as often does not mean that it is understood twice as well; to accurately compare the networks — the patterns of association — of different units, then, we need to control for the fact that different units may have different amounts of talk. ENA solves this problem by spherically normalizing the vectors. Spherical normalization is accomplished by dividing each vector V^u by its length. (This is the equivalent of the cosine norm frequently used in natural language processing and automated content analysis.) The resulting normalized vector, N^u, thus quantifies for unit u the *relative* frequencies of co-occurrence of codes independent of the number of stanzas in the model for any given unit.[6]

[6] There are some contexts, of course, in which normalization of the adjacency vectors would obscure important information. The normalization step in an epistemic network analysis is thus optional.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

5.4 Dimensional reduction via singular value decomposition

In order to interpret and visualize the (normalized) adjacency vectors, ENA performs a dimensional reduction on the high-dimensional space using singular value decomposition. (A singular value decomposition is similar to a principal components analysis, but it does not rescale the data.) This provides a rotation of the original high-dimensional ENA space, such that the rotated space provides a reduced number of dimensions that capture the maximum variance in the data. For every unit u in the data, ENA creates a point, P^u, which is the location of the normalized vector N^u under the singular value decomposition.

6 POSITIONING THE NETWORK NODES

Normally, the basis vectors, or the loadings, would tell us how to interpret the positions of points in the space. The problem, though, is that these basis vectors represent connections between codes in the original data. That is, each point represents one of the cells in the cumulative adjacency matrix. That makes it hard to interpret, because if we have 20 codes, as in the *RescuShell* example, then we can have up to 190 basis vectors, each of which corresponds to a unique co-occurrence of codes. To interpret the dimensions of this rotated space, ENA takes the codes in the original data — which correspond to the nodes of the network — and uses an optimization routine to position them in ENA space such that for any unit u, the centroid of the network model corresponding to the cumulative adjacency matrix C^u is as close as possible to the location of the point P^u. ENA computes and reports the strength of the correlation between the centroids and the projected points in the model using both Pearson's and Spearman's *r*. Typically, they are extremely high ($r > 0.9$).[7]

The utility of this correspondence between the centroid of the network graph and the location of P^u is that we can compare the structures of two networks by comparing the locations of their projected points P^u. For example, if P^A is higher in the space (has a larger *y*-value) than P^B, the centroid of A will also be higher in the space, which means the network graph for A has more connections in the upper part of the ENA space than the network graph for B. (This is what we saw in Figures 2–7 above with Students A, B, and C.) Thus, the position of a point in ENA space summarizes the structure of connections in the networks being modelled, and as a consequence the layout of the graphs of the networks in ENA space shows *which parts of the structure of the networks account for the most variance between different networks.*

[7] Technically, for all pairs of units (*A*,*B*), ENA is computing the correlation between ($P^A - P^B$) and ($M^A - M^B$), where P^A is the ENA projection of *A* and M^A is the centroid of *A* under the optimized network layout. This is a useful measure of the interpretive validity of the network graph, in the sense that a high correlation means that if P^A is (for example) to the right of P^B, then M^A will also be to the right of M^B, meaning that the differences in the network representations of *A* and *B* correspond to the relative positions of P^A and P^B in the ENA space.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

The result is that:

1. The dimensional reduction models the variance among the different networks being analyzed;
2. The corresponding network graphs allow us to interpret the significance of the locations of the points in the ENA model; and
3. As a result, we can interpret what aspects of network structure explain the differences between units in the model.

7 CREATING AND EXPLORING NETWORK MODELS WITH THE ENA WEBKIT

Now that we have examined in some detail how ENA creates and visualizes network models, we will work through several example analyses using the ENA WebKit. In doing so, we will explore both the process of modelling and interpreting data using the online tool and some of the more advanced features of ENA. That being said, this is not intended as a comprehensive user guide for the ENA WebKit.[8] In the examples that follow, we use the dataset RSdata.csv, the same dataset from which the previous examples are drawn.[9]

At its most basic level, the ENA WebKit performs two main functions, as described in the previous section. First, it processes coded data: it takes the data table, divides the rows into stanzas, accumulates the codes by stanza, creates a set of adjacency matrices, creates a cumulative adjacency matrix representing the structure of connections among the coded objects for each unit of analysis, and produces a dimensional reduction to represent the dataset. Second, ENA uses the results of this analysis to create visualizations that facilitate exploration and interpretation of the data.

When you log in to the ENA WebKit, you will see the set creation interface (see Figure 15). On the left, you can select a dataset and upload it into the tool. If you have already uploaded a dataset previously, it will appear in the list at the left.

[8] For step-by-step instructions on using the tool and detailed descriptions of its features, visit http://www.epistemicnetwork.org, which contains a tutorial video, technical documentation (Shaffer, 2014a, 2014b), and links to sample datasets already formatted and coded for ENA analysis.

[9] The full dataset can be accessed at http://www.epistemicnetwork.org, along with a technical report (Shaffer & Arastoopour, 2014) containing a detailed description of the *RescuShell* virtual internship, a guide to the content and organization of the dataset, an explanation of the coding scheme, and some sample research questions. If you would like to work through the following examples in the ENA WebKit as you read, you can access the ENA WebKit by clicking "Launch" at http://www.epistemicnetwork.org.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

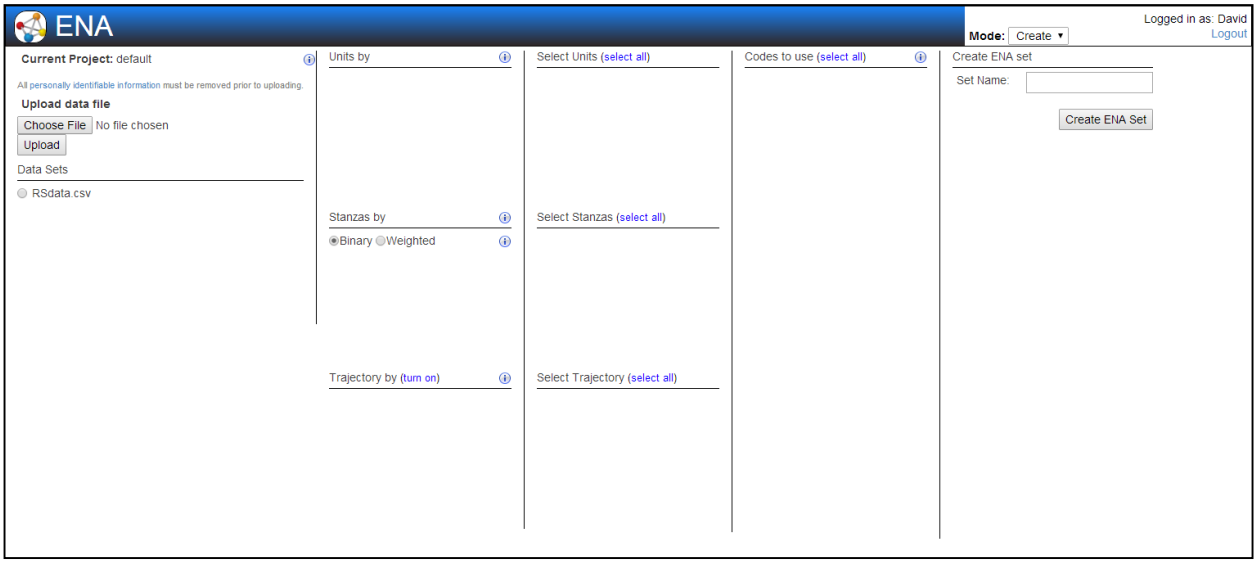


Figure 15. Set creation interface of the ENA WebKit version 4.

If you select a dataset, the set creation interface will populate with values based on the columns in the dataset (see Figure 16). At this point, you can set the parameters for your analysis.

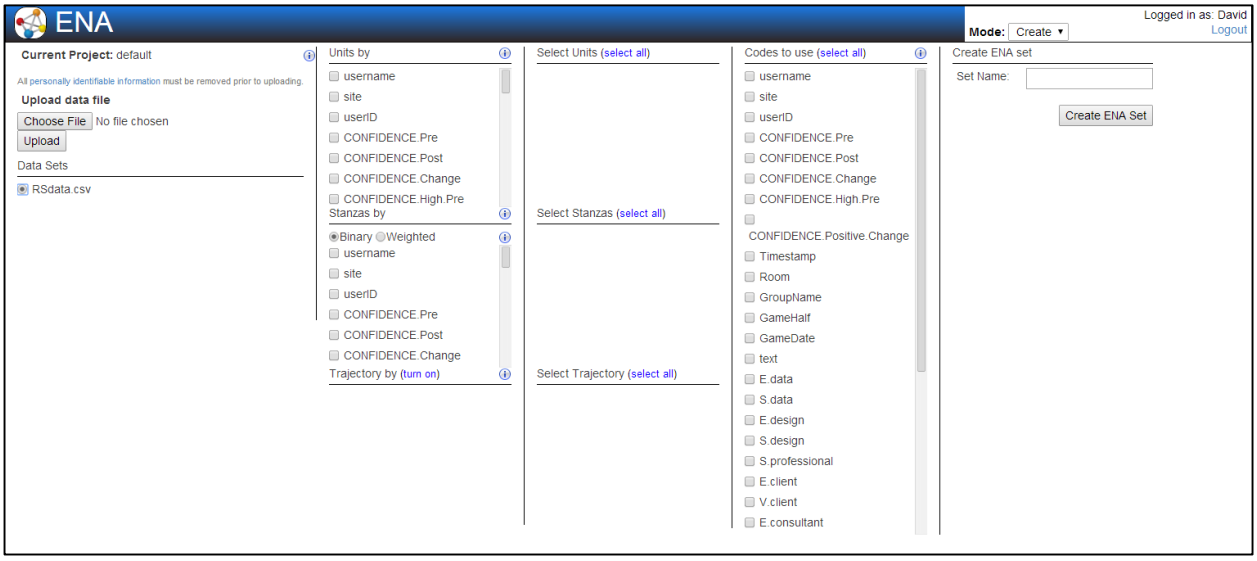


Figure 16. Set creation interface of the ENA WebKit version 4 populated with data from the dataset RSdata.csv.

For our first example, we will use the same dataset and parameters used in the analyses above. We will choose “UserName” as our unit of analysis in order to explore the networks of individual students. In this particular dataset, each day consisted of a discrete activity in the simulation, so we will define our stanzas by “GameDay.” And we will select all 20 engineering epistemic frame codes (see Figure 17). Once we have selected the unit, stanza, and code variables, we can name the set and create it (right side of the set creation interface). For example, we might call this ENA set “students.by.activity.allcodes” to

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

indicate the variables we chose. (Note as well that the workspace can be exported if you wish to perform additional analyses on the data that the ENA WebKit does not support.)

Unit Variable	UserName	*ENA will construct a network model for each student in the dataset*
Stanza Variable	GameDay	*ENA will construct a network model based on the co-occurrences of codes within each activity in the dataset*
Code Variables	E.data, S.data, E.design, S.design, S.professional, E.client, V.client, E.consultant, V.consultant, S.collaboration, I.engineer, I.intern, K.actuator, K.rom, K.materials, K.power, K.sensor, K.attribute, K.data, K.design	*ENA will construct a network model based on the co-occurrences of these variables*

Figure 17. Table showing the set creation parameters for exploring the cognitive networks of individual students in *RescuShell*.

At this point, the ENA WebKit will prepare the ENA set for visualization. Once that is done, the plot choices will appear on a new screen. When we select the ENA set (left side) and click "Open Plot" on the right, we will see the visualization interface (see Figure 18).

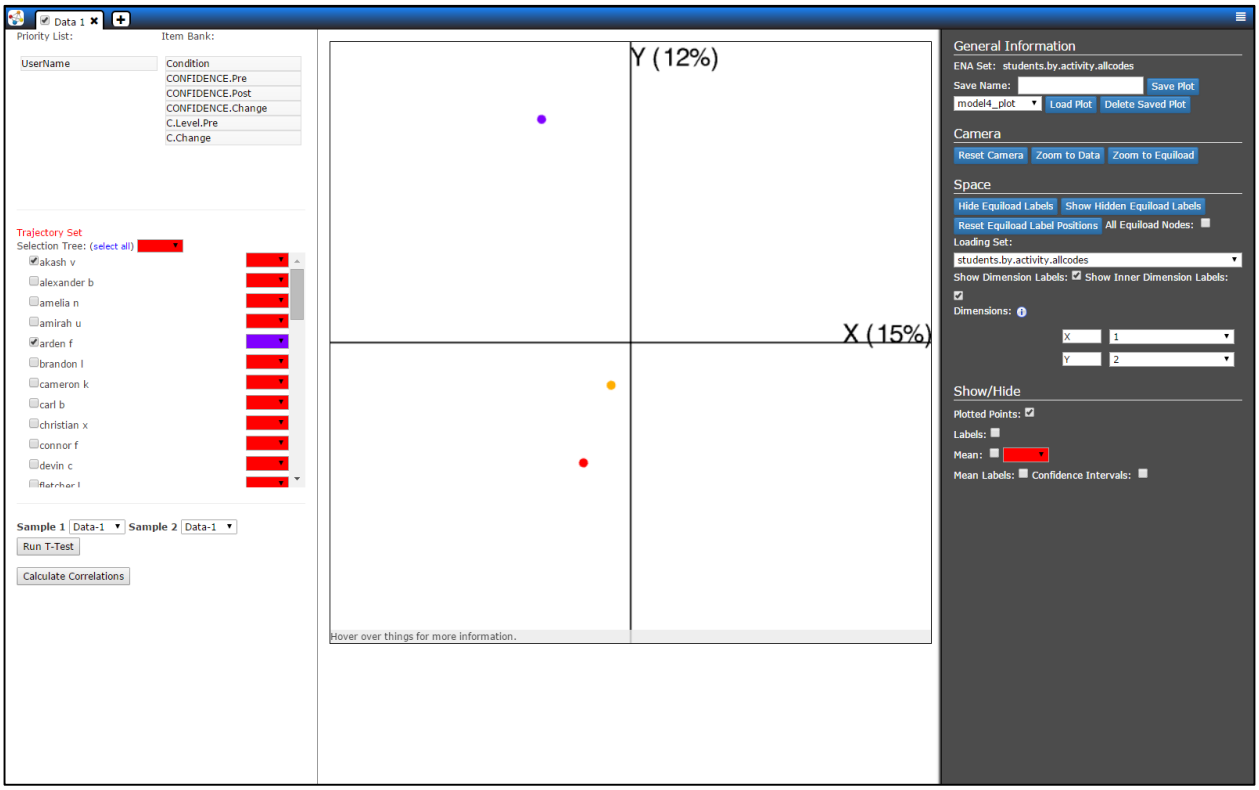


Figure 18. Visualization interface of the ENA WebKit version 4.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

The units we selected appear on the left. If we select a student, a point representing that student's network (which corresponds to the centroid for that students' network) will appear in the projection space. If we click "select all," all the students' centroids will appear, and hovering over a centroid with the cursor will identify to which student that centroid belongs. By default, the ENA space in which the centroids are visualized is defined by the first (*x*) and second (*y*) dimensions, the dimensions that account for the greatest variation in the data. (Recall that the singular value decomposition maximizes variance when performing the dimensional reduction.) If you would like to explore other dimensions, the drop-down menus on the right allow you to set the *x* and *y* axes as different dimensions. The numbers in parentheses next to the axis labels indicate the percentage of variance in the data for which those dimensions account. In this case, dimension one accounts for 15% of the variance in the data, and dimension two accounts for 12%.

As discussed above, the centroids and network graphs are visualized in the same space, and we can use the locations of the nodes in the equiload projection to interpret the dimensions of that space. Clicking on a centroid will bring up an equiload tab, which will display that student's network. For example, we can click on the centroid representing Akash's network. At first, it may be difficult to interpret the network. By default, the graph edges are weighted across the entire dataset. However, if we want to understand the structure of this particular student's connections, we can set the threshold values to make it easier to interpret the network graph. If we click the button "Over Plot" under the threshold options on the right side of the screen, the ENA WebKit sets the maximum line thickness and saturation equal to the largest weighted connection in this particular student's network, and it sets the thinnest and least saturated line to the lowest weighted connection in this particular student's network. We can also raise the lower threshold value to exclude the weaker connections, the connections that fall below that threshold, which makes it easier to see the overall structure (see Figure 19).

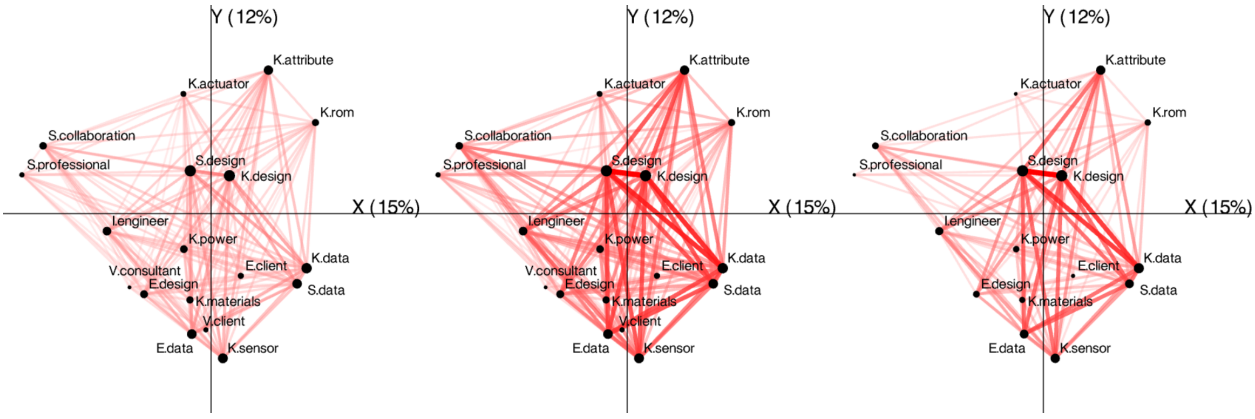


Figure 19. The same student's network with three different threshold settings. The network on the left is thresholded over the set and includes all connections. The network in the centre is thresholded over the plot and includes all connections. The network on the right is thresholded over the plot and excludes the weakest connections.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

Each connection in the equiload projection represents the total number of stanzas in which there is a connection between those two codes. To facilitate understanding the structure of an ENA network, the ENA WebKit makes it possible to see what portions of the dataset are represented by each part of the network. If we click the graph edge in Akash's network that connects *S.data* and *E.design*, toward the bottom of the network, a new window will open. If we select the column in the dataset that contains the chat utterances ("text"), we can see all of the utterances that contributed to this connection (see Figure 20).

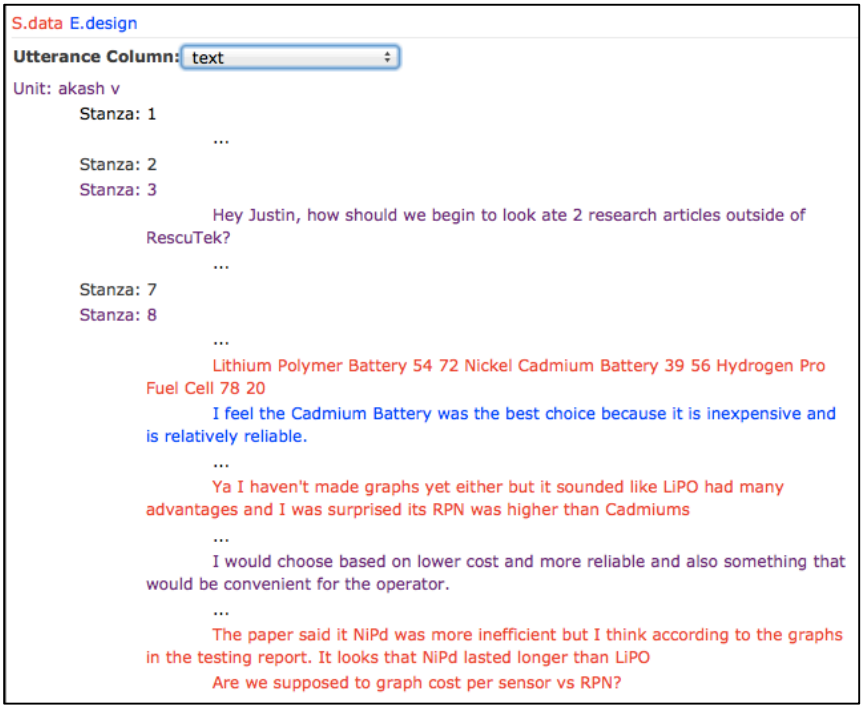


Figure 20. Excerpt of the chat utterances that contributed to the connection between *skill of data* and *epistemology of design* in Akash's network.

The codes are coloured such that excerpts in the data containing only the first code are shown in red, and excerpts containing only the second code are shown in blue. Excerpts containing both codes are shown in purple. (The intervening uncoded excerpts are hidden by default but can be shown by clicking the ellipses.) The utterances are grouped by stanza, and the colour of a stanza title indicates whether the two codes both occurred in that stanza, based on the same colour scheme. (Remember that the two codes don't need to co-occur in the same utterance; they only need to co-occur in the same stanza to contribute to a connection between the two nodes in the network graph.) This feature of the ENA WebKit allows us to close the interpretive loop: we started with a coded dataset; we used the coded data to create and visualize a network model; then, if we want to understand where any of the connections in the network model come from, we can go back to the raw data.

Thus far, we have simply been exploring individual students' networks. In many cases, though, we will want to begin with a specific research question and then use ENA to help answer it. In this dataset, for

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

example, half of the students had no prior experience with an engineering design simulation before they used *RescuShell*, and the other half of the students used *RescuShell* immediately after using a similar engineering design simulation on a different topic. Thus, one research question we might ask is:

> *Does prior experience with an engineering design simulation influence the structure of cognitive connections that students make in a second engineering design simulation?*

To create an ENA set with which to explore this question, we need to identify our unit of analysis, define the stanzas, and select which codes we want to model, just as we did in the previous example. We could ask this same question of *individual* students or of student *teams*. In this case, we will model the teams' networks. To do that, we will select "GroupName" as our unit of analysis. In this dataset, students worked with one team during the first half of the simulation and a different team during the second half of the simulation, so we need to keep this in mind when we do our analysis.[10] In addition, we will select "Condition," which indicates whether *RescuShell* was used first or second. As with our previous example, we will use "GameDay" to define the stanzas and select all of the engineering epistemic frame codes.

When we create the set and open the visualization interface, we can compare the two conditions. As before, we can click the check boxes next to each condition on the left to see the corresponding centroids, and we can change the colours using the drop-down menus to more easily distinguish one condition from the other. To determine whether there is a significant difference between the two conditions, we will select one of the conditions to display, then show the mean and confidence intervals by checking the appropriate boxes under "Show/Hide" on the right side of the visualization interface (see Figure 21). To include the second condition, we will open a new data tab (top of the screen) and repeat that process for the second condition, which will give us the projection shown in Figure 21.

[10] For more details on this dataset, see Shaffer and Arastoopour (2014).

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

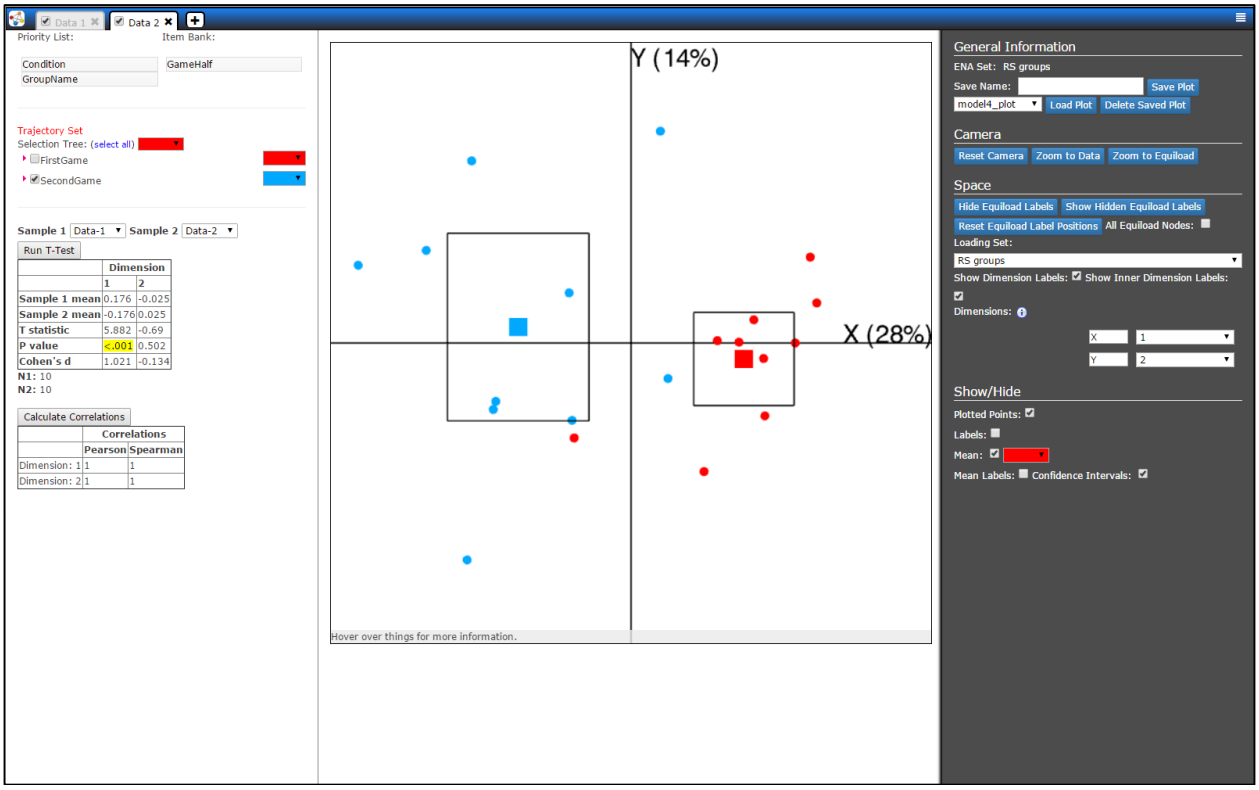
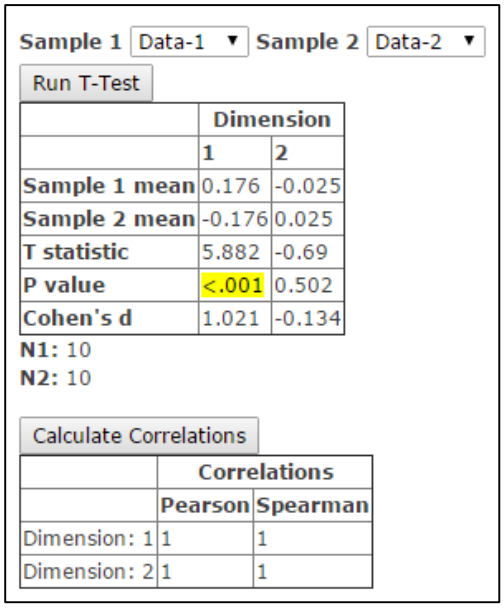


Figure 21. Groups who used *RescuShell* first (red) and second (blue). The two groups are significantly different on the first dimension.

As Figure 21 indicates, there is a significant difference between the two conditions on the first dimension. To determine the difference with more precision, we can run an independent samples *t* test. To do that, select the two samples you want to compare from the drop down menus in the lower left corner of the visualization interface (see Figures 21 & 22). When you do that, you will see the means for the two samples on both dimensions, along with the *t* score, *p* value, and Cohen's *d*, a measure of effect size. In this case, the difference on dimension one is significant: $mean_A = 0.176$, $mean_B = -0.176$; $t = 5.882$; $p < 0.001$; Cohen's $d = 1.021$. Furthermore, we can calculate the strength of the correlation between the centroids and the projected points in the model using both Pearson's and Spearman's *r*. In this case, both are equal to one for both dimensions because the number of units in the model is small compared to the number of dimensions (see Figure 23). The optimization is thus easy to resolve.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3



Sample 1 Data-1 Sample 2 Data-2

Run T-Test

	Dimension	
	1	2
Sample 1 mean	0.176	-0.025
Sample 2 mean	-0.176	0.025
T statistic	5.882	-0.69
P value	<.001	0.502
Cohen's d	1.021	-0.134

N1: 10
N2: 10

Calculate Correlations

	Correlations	
	Pearson	Spearman
Dimension: 1	1	1
Dimension: 2	1	1

Figure 22. Close-up of the *t*-test and correlations interface (left side of the visualization interface in Figure 21). The *t*-test shows that the two groups are statistically significantly different on the first dimension. The correlation between the centroids and the projected points is perfect: Pearson's *r* = 1 and Spearman's *r* = 1 for both dimensions.

To determine what accounts for this difference in the structures of connections, we can look at the equiload projections. As before, we can click on a group's centroid to see that group's network, but we can also click on a mean to see the mean network for all the groups in that condition. If we do that, we get the networks shown in Figure 23.

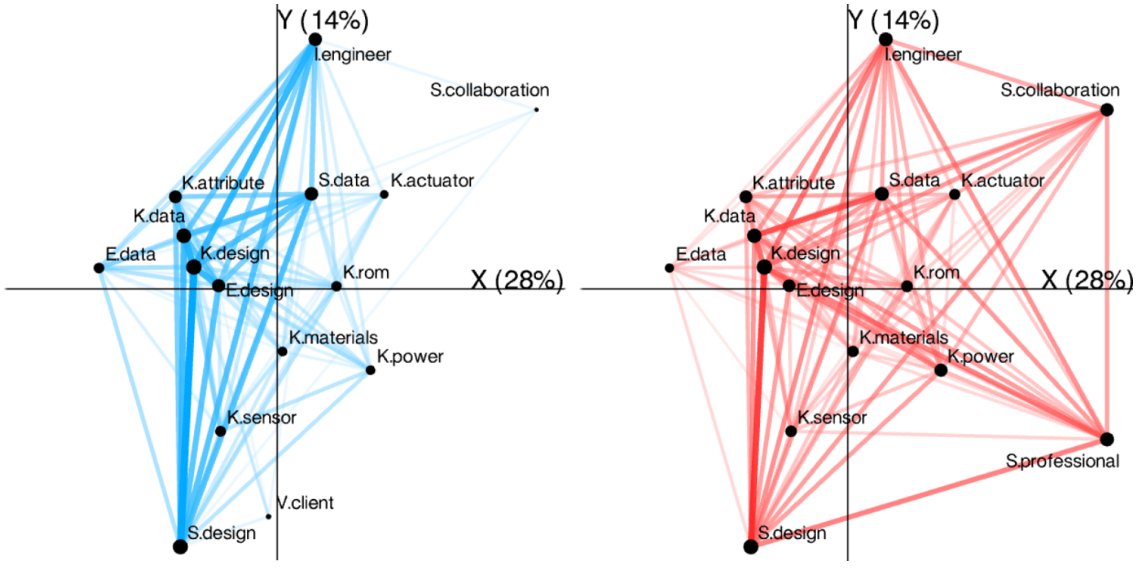


Figure 23. Mean networks of groups who used *RescuShell* first (red, right) and second (blue, left).

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

Notice that the dimensions are different from our previous examples. Remember that ENA creates the space in which the networks are visualized based on the parameters used to create the set. Because we are looking at groups rather than individuals, the space is different. Based on the mean networks for the two conditions, the most significant difference seems to be that groups with no prior experience (the red network on the right in Figure 23) made more connections to interpersonal skills, such as collaboration and professionalism, while groups with prior experience (the blue network on the left in Figure 23) made more or stronger connections to advanced aspects of engineering design, such as the *value of the client*, the *epistemology of data*, and the *epistemology of design*. A closer analysis of the qualitative data suggests that students who were less experienced spent a considerable amount of time figuring out how to manage their design teams, which accounts for the prevalence of connections to and among collaboration and professional tools (such as email).

Because we know which half of the simulation each group completed, we can make comparisons not only between groups in their first or second simulation, but also (a) between groups in the first and second halves of the simulation, and (b) between groups in the first half of their first simulation and groups in the first half of their second simulation. On the top left of the visualization interface (Figure 21) is a "Priority List," which shows the criteria we used to create the units: Condition and GroupName. There is also an "Item Bank," which shows all of the columns of metadata uniquely associated with each one of the units in the dataset. In this case, there is only one item: GameHalf. If we drag GameHalf from the Item Bank to the top of the Priority List, the data will be re-sorted based on the new category (see Figure 24).

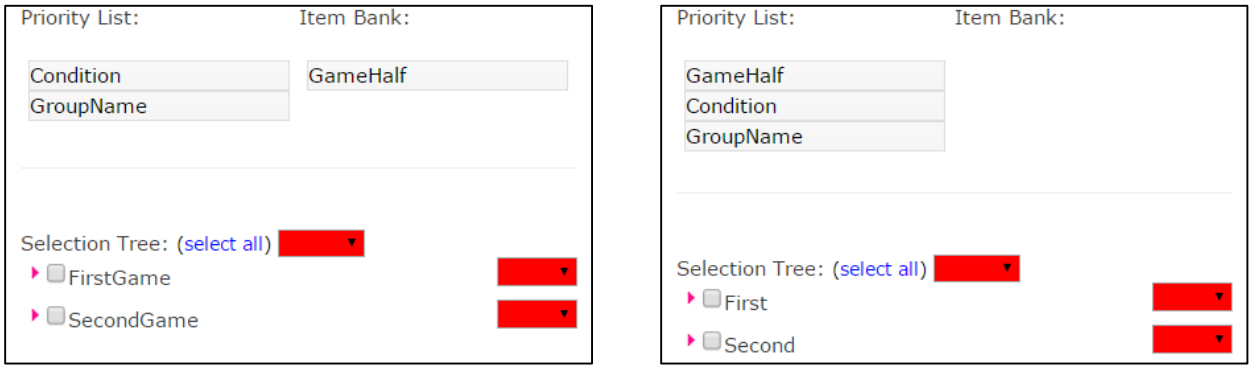


Figure 24. On the left is the Priority List for our initial analysis, which sorted the data by condition (first simulation or second simulation). When we drag GameHalf from the Item Bank to the Priority List, the data are resorted by half (first half or second half of a simulation).

After re-sorting the data by GameHalf, we can make the additional comparisons suggested above, using the same method we just described. This is left as an exercise for the reader, although we note that because of the small number of units in this model relative to the number of dimensions in the data this must be regarded as an exploratory ENA analysis, and the model created would not necessarily generalize beyond this particular dataset.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

8 MODELLING ASYMMETRIC WEIGHTED DATA AND NETWORK TRAJECTORIES USING ENA

So far, we have focused on *cumulative* network models of *symmetric* data. That is, we have looked at networks representing the structures of connections that accumulated over some particular time period (for example, over the course of an engineering design simulation) and based on symmetric codes: if, for instance, the skill of data is connected to the epistemology of design, then the epistemology of design is connected to the skill of data. For our last example, we will examine a case where the data are both *directional* and *weighted*, and show an example of using ENA to create a *trajectory* model, which indicates changes in structures of connections over time.

In this case, we will explore a classic experiment in social network analysis, Theodore Newcomb's 1961 study, *The Acquaintance Process*. Newcomb studied the patterns of social relationships among 17 men who lived in the same house for a semester after transferring to the University of Michigan. Each week for 15 weeks, the men ranked their housemates' relative popularity in the group. Newcomb found that as time went on, correlations between the men's reports increased. The experiment was run in two successive years, and the correlation coefficients between mutual pair attraction increased from 0.13 to 0.50 in Year 1 and from 0.16 to 0.56 in Year 2 (p. 75); the means of the individual correlations in rank-ordering for pairs of adjacent weeks increased from 0.51 to 0.91 in Year 1 and from 0.65 to 0.90 in Year 2 (p. 196).

To model these social relationships using ENA, we will use the data from Year 2. To account for the fact that rankings are not symmetrical (Person A's rank of Person B is not necessarily the same as Person B's rank of Person A), we refer to the person doing the ranking as the "sender" and the person being ranked as the "receiver." The dataset is structured such that each participant has a row and a column as a sender and a row and a column as a receiver. In the sender columns, there is always a one. In the receiver column is the rank that the person received from a given sender in a given week. However, we set up the data table so that the rankings are inverted. The person who is ranked as most popular (ranked #1) in a given week gets the highest score (17); the person who received the lowest rank (#17) in a given week receives the lowest rating (1). When ENA creates an adjacency matrix based on weighted values, it multiplies the receiver ranking by the sender value. Thus assigning all senders a value of one and the most popular ranks a high number correctly computes a strong association between a sender and his highest ranked receiver. In general, this approach — treating the "sending" and "receiving" functions of each individual in the model separately — makes it possible to analyze asymmetric data with ENA. One consequence of this approach, however, is that each participant will be represented by two nodes. One node represents the participant as a sender, and the other represents him as a receiver.

To re-examine Newcomb's results, we set "week" as the unit because we are interested in the structure of social relationships each week, and we use each row as a stanza because each row represents one person's rating of one other person during one week. When we select rows as the stanza variable, we

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

have a choice between binary and weighted. If we select binary stanzas, ENA converts all non-zero values to one before constructing the adjacency matrices. If we select weighted stanzas, ENA leaves the values as they are and multiplies them together when constructing adjacency matrices. So, in this case, because we want to preserve the ranking information, we use the weighted model. We also want to select all of the senders and all of the receivers for codes.

When we open the visualization interface and select all the weeks, we can see the centroids representing the ranking patterns of the senders for each week (see Figure 25). The ranking patterns seem to change from week 1 to week 15.

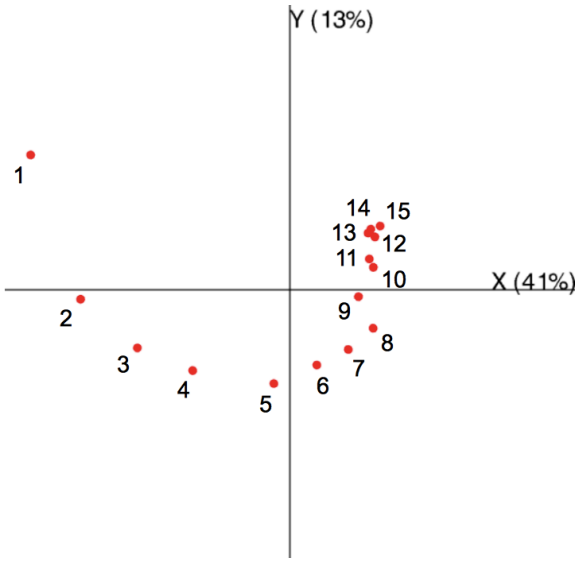


Figure 25. Centroids representing the ranking patterns of the senders for each week.

To understand whether the changes in ranking patterns from week to week are significant, we can plot the trajectory of the senders' ranking patterns from week 1 to week 15. To do this, we will return to the set creation interface. This time, we will select "sender" under units, use each row as a stanza, and select all of the senders and receivers for codes. We will also turn on the trajectory feature. That gives us the option to select either "accumulated" or "separate." In an accumulated trajectory, each time point is a summation of all the preceding units of time. In a separate trajectory, each time point is computed from zero, such that each point contains only the data within each unit of time. In this case, we will choose "separate" and select "week" in order to model each week independently from the preceding weeks.

When we open the visualization interface and select all the weeks, it appears that there were small groups of weeks where the ranking patterns didn't change substantially (see Figure 26). However, keep in mind that this model maximizes the variance along the dimensions shown. If we want to compare this with what we were looking at before, we have to visualize the data in the space created for the previous analysis.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

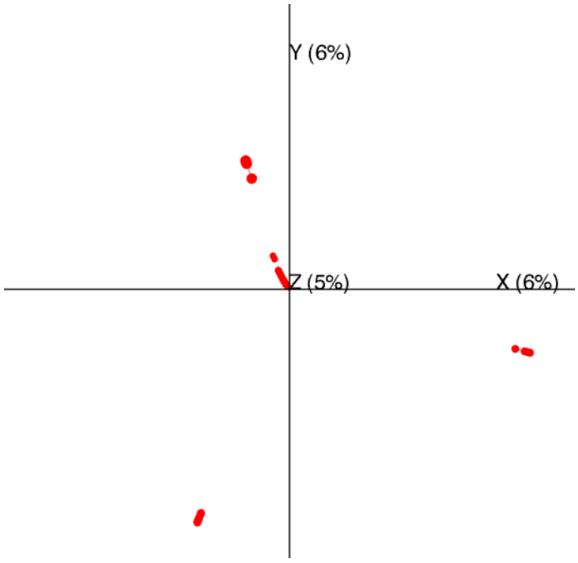


Figure 26. Trajectory model of the ranking patterns of the senders for each week.

To do that, we need to switch to the loadings that represented the group as a whole using the options on the right side of the visualization interface under "Loading Set" (see Figure 27). The ENA set from which the projection space comes is called the loading set because the projection space is determined by the rotation matrix from a singular value decomposition, and thus it functions like the loadings in a principal components analysis. In general, ENA set A can be used as a loading set for ENA set B as long as both sets were created using the same collection of codes.

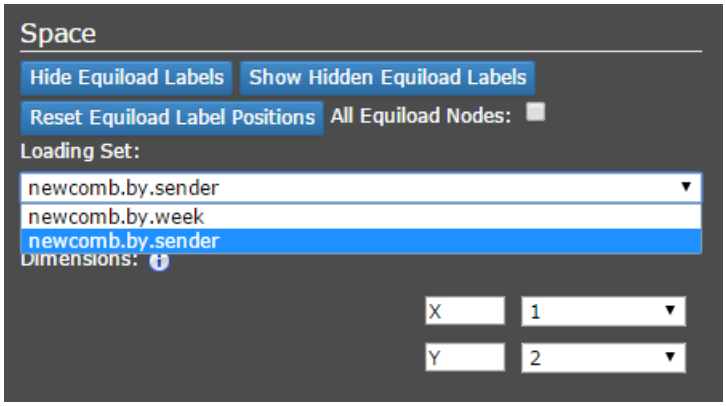


Figure 27. Close-up of the menu (right side of the visualization interface in Figure 21) that allows selection of a different loading set.

If we select as a loading set the set we created in the first part of this analysis, we will see the trajectories of all the senders in the same space as before. If we hide the individual trajectories and show only the mean trajectory (see Figure 28), we get basically the same model as shown in Figure 25. Now, however, we can compute the confidence intervals for each week.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

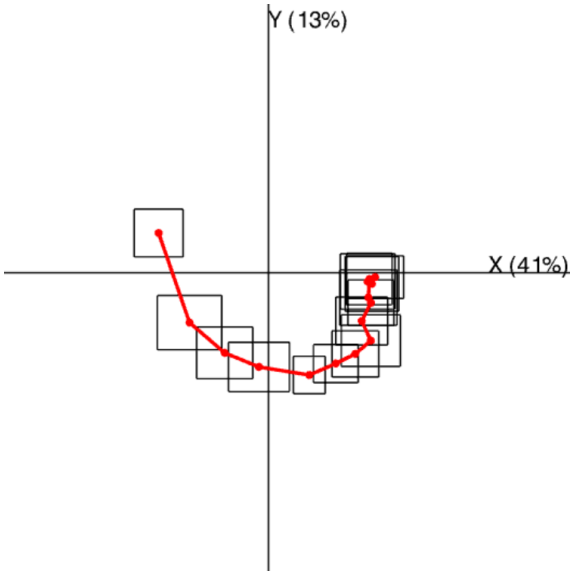


Figure 28. Mean trajectory model with confidence intervals of the ranking patterns of the senders for each week projected in the space defined by the initial analysis (see Figure 25).

Computing the confidence intervals reveals that there were only two places where a significant change in ranking patterns occurred from one week to the next. Between week 1 and week 2, there was a significant change in the rankings, and the same is true between weeks 4 and 5. Overall, however, there is a significant difference between the beginning of the semester, when all the men met for the first time, and the end of the semester, when they had been living with one another for several months. This suggests that the changes from week to week were mostly too small to seem significant, but overall, the accumulated small changes resulted in a significant shift in the participants' pattern of relationships. In addition, as Newcomb reported and this analysis confirms, the ranking patterns change more in the early weeks, becoming more similar toward the end of the semester.

The trajectory model also allows us to compare the trajectories of individual senders to the mean trajectory. Figure 29 shows the trajectories of two senders compared with the mean trajectory. The trajectory of the sender on the left (blue) is very similar to the mean trajectory (red), while the trajectory of the sender on the right (green) differs substantially from the mean trajectory.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

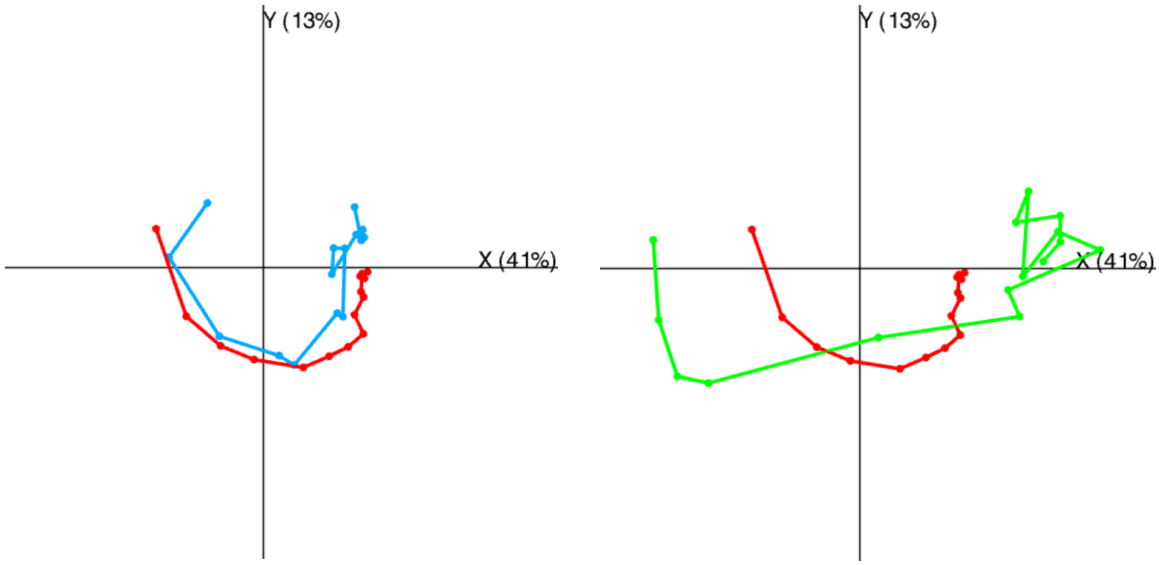


Figure 29. Trajectories of two individual senders (blue and green) compared with the mean trajectory (red). Note that trajectory of the sender on the left (blue) is similar to the mean trajectory, while the trajectory of the sender on the right (green) is different.

Trajectory models thus make it possible to explore how and to what extent structures of connections change over time.

9 CONCLUSION

ENA is a versatile method for quantifying, visualizing, and interpreting network data in a variety of contexts. If you would like to use the ENA WebKit, you can access the tool at http://www.epistemicnetwork.org. Numerous resources, including detailed user guides and coded datasets formatted for ENA analysis, are also available on the site. You can also upload your own datasets for analysis, but we recommend that you read the data formatting guide (Shaffer, 2014a) before doing so. If you use ENA in your research, please let us know by contacting us at ena.papers@gmail.com.

10 ACKNOWLEDGEMENTS

This work was funded in part by the National Science Foundation (DRL-0918409, DRL-0946372, DRL-1247262, DRL-1418288, DUE-0919347, DUE-1225885, EEC-1232656, EEC-1340402, REC-0347000), the MacArthur Foundation, the Spencer Foundation, the Wisconsin Alumni Research Foundation, and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison. The opinions, findings, and conclusions do not reflect the views of the funding agencies, co-operating institutions, or other individuals.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

REFERENCES

Andrist, S., Collier, W., Gleicher, M., Mutlu, B., & Shaffer, D. (2015). Look together: Analyzing gaze coordination with epistemic network analysis. *Frontiers in Psychology*, *6*(1016). http://dx.doi.org/10.3389/fpsyg.2015.01016

Arastoopour, G., Shaffer, D. W., Swiecki, Z., Ruis, A. R., & Chesler, N. C. (2016). Teaching and assessing engineering design thinking with virtual internships and epistemic network analysis. *International Journal of Engineering Education*, *32*(3B), 1492–1501.

Batagelj, V., & Mrvar, A. (1998). Pajek-program for large network analysis. *Connections*, *21*(2), 47–57.

Carroll, L. (1871). *Through the looking-glass, and what Alice found there*. London: Macmillan.

Chesler, N. C., Arastoopour, G., D'Angelo, C. M., Bagley, E. A., & Shaffer, D. W. (2013). Design of a professional practice simulator for educating and motivating first-year engineering students. *Advances in Engineering Education*, *3*(3), 1–29.

Chesler, N. C., Ruis, A. R., Collier, W., Swiecki, Z., Arastoopour, G., & Shaffer, D. W. (2015). A novel paradigm for engineering education: Virtual internships with individualized mentoring and assessment of engineering thinking. *Journal of Biomechanical Engineering*, *137*(2), 1–8. http://dx.doi.org/10.1115/1.4029235

Collier, W. (2015). Application of ENA-based network analyses to fMRI data on school-children's acquisition of number symbols (Poster Presentation). *Discovery Challenge Research Symposium, Wisconsin Institutes for Discovery*. Madison, WI: Wisconsin Alumni Research Foundation (WARF).

Collier, W., Ruis, A. R., & Shaffer, D. W. (2016). Local versus global connection making in discourse. In C.K. Looi, J.L. Polman, U. Cress, P. Reimann (Eds). *Transforming Learning, Empowering Learners: The International Conference of the Learning Sciences (ICLS) 2016, Volume I*, (pp. 426–433). Singapore: International Society of Learning Sciences.

D'Angelo, A.-L. (2015). *Evaluating operative performance through the lens of epistemic frame theory*. University of Wisconsin–Madison.

D'Angelo, C. M., Clark, D. B., & Shaffer, D. W. (2012, March). *Epistemic network analysis: An alternative analysis technique for complex STEM thinking*. Paper presented *at the National Association of Research on Science Teaching Conference* (NARST 2010), Indianapolis, IN, USA.

DiSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. Pufall (Eds.), *Constructivism in the computer age* (pp. 47–70). Hillsdale, NJ: Erlbaum.

Dorogovtsev, S. N., & Mendes, J. F. F. (2013). *Evolution of networks: From biological nets to the Internet and WWW*. Oxford, UK: Oxford University Press.

Gee, J. P. (1999). *An introduction to discourse analysis: Theory and method*. London, UK: Routledge.

Hatfield, D. (2015). The right kind of telling: An analysis of feedback and learning in a journalism epistemic game. *International Journal of Computer-Mediated Simulations*, *7*(2), 1–23. http://dx.doi.org/10.4018/IJGCMS.2015040101

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

i Cancho, R. F., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences*, *268*(1482), 2261–2265. http://dx.doi.org/10.1098/rspb.2001.1800

Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression*. Thousand Oaks, CA: Sage.

Knight, S., Arastoopour, G., Shaffer, D. W., Buckingham Shum, S., & Littleton, K. (2014). Epistemic networks for epistemic commitments. In Polman, J. L., Kyza, E. A., O'Neill, D. K., Tabak, I., Penuel, W. R., Jurow, A. S., O'Connor, K., Lee, T., and D'Amico, L. (Eds.), *Learning and Becoming in Practice: The International Conference of the Learning Sciences (ICLS) 2014, Volume 1*, (pp.150–158). Boulder, CO, USA: International Society for Learning Sciences.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Lattin, J. M., Carroll, J. D., & Green, P. E. (2003). *Analyzing multivariate data*. Pacific Grove, CA: Thomson Brooks/Cole.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. http://dx.doi.org/10.3758/BF03204766

Nash, P., & Shaffer, D. W. (2013). Epistemic trajectories: Mentoring in a game design practicum. *Instructional Science*, *41*(4), 745–771. http://dx.doi.org/10.1007/s11251-012-9255-0

Newcomb, T. M. (1961). *The acquaintance process*. New York: Holt, Rinehart & Winston.

Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review, E70*(5). https://doi.org/10.1103/PhysRevE.70.056131

Orrill, C., Shaffer, D. W., & Burke, J. (2013, April-May). *Exploring coherence in teacher knowledge using epistemic network analysis.* Paper presented at the American Educational Research Association Annual Conference (AERA 2013), 2013, San Francisco, CA, USA.

Quardokus Fisher, K., Hirshfield, L., Siebert-Evenstone, A. L., Arastoopour, G., & Koretsky, M. (2016, June). *Network analysis of interactions between students and an instructor during design meetings*. Paper presented at the Annual Conference of the American Society for Engineering Education (ASEE 2016), New Orleans, LA, USA. Retrieved from http://edgaps.org/gaps/wp-content/uploads/VirtualLabENA_FINAL.pdf

Ruis, A. R. (2016). "Trois Empreintes d'un Même Cachet": Toward a Historical Definition of Nutrition. *Annual Meeting of the American Association for the History of Medicine*, May 2016, Minneapolis, MN.

Rupp, A. A., Choi, Y. Y., Gushta, M., Mislevy, R. J., Theis, M. C., Nash, P., … Svarovsky, G. N. (2009, June). Modeling learning progressions in epistemic games with epistemic network analysis: Principles for data analysis and generation. Paper presented at the Learning Progressions in Science Conference (LeaPS 2009), 2009, Iowa City, IA, USA.

Rupp, A. A., Gustha, M., Mislevy, R., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning and Assessment*, *8*(4), 4–47.

(2016). 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. http://dx.doi.org/10.18608/jla.2016.33.3

Rupp, A. A., Sweet, S., & Choi, Y. (2010). Modeling learning trajectories with epistemic network analysis: A simulation-based investigation of a novel analytic method for epistemic games. In R. S. J. d. Baker, A. Merceron, P. I. Pavlik Jr. (Eds.), *Proceedings of the 3rd International Conference on Educational Data Mining* (EDM 2010), (pp. 319–321). Pittsburgh, PA, USA: International Educational Data Mining Society.

Shaffer, D. W. (2004). Pedagogical praxis: Using technology to build professional communities of practice. *Association for Computing Machinery (ACM) SigGROUP Bulletin*, *24*(3), 39–43. http://dx.doi.org/10.1145/1052829.1052838

Shaffer, D. W. (2006). Epistemic frames for epistemic games. *Computers and Education*, *46*(3), 223–234.

Shaffer, D. W. (2007). *How computer games help children learn*. New York: Palgrave Macmillan.

Shaffer, D. W. (2012). Models of situated action: Computer games and the problem of transfer. In C. Steinkuehler, K. D. Squire, & S. A. Barab (Eds.), *Games, learning, and society: Learning and meaning in the digital age* (pp. 403–431). Cambridge, UK: Cambridge University Press.

Shaffer, D. W. (2014a). *Formatting data for epistemic network analysis*. Madison, WI: Games and Professional Simulations Technical Report 2014–2.

Shaffer, D. W. (2014b). *User guide for epistemic network analysis web version 3.3*. Madison, WI: Games and Professional Simulations Technical Report 2014–1.

Shaffer, D. W., & Arastoopour, G. (2014). *Guide to RSdata.csv sample ENA data set*. Madison, WI: Games and Professional Simulations Technical Report 2014–3.

Shaffer, D. W., Hatfield, D. L., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E. A., … Frank, K. (2009). Epistemic Network Analysis: A prototype for 21st century assessment of learning. *International Journal of Learning and Media*, *1*(1), 1–21. http://dx.doi.org/10.1162/ijlm.2009.0013

Siebert-Evenstone, A. L., Arastoopour, G., Collier, W., Swiecki, Z., Ruis, A. R., & Shaffer, D. W. (2016). In search of conversational grain size: Modeling semantic structure using moving stanza windows. In C.K. Looi, J.L. Polman, U. Cress, P. Reimann (Eds). *Transforming Learning, Empowering Learners: The International Conference of the Learning Sciences (ICLS) 2016, Volume I*, (pp. 631–639). Singapore: International Society of Learning Sciences.

Svarovsky, G. N. (2011). Exploring complex engineering learning over time with epistemic network analysis. *Journal of Pre-College Engineering Education Research*, *1*(2), 19–30. http://dx.doi.org/10.5703/1288284314638

Wenger, E. (1999). *Communities of practice: Learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.