Chapter 15: Epistemic Network Analysis: A Worked Example of Theory-Based Learning Analytics

David Williamson Shaffer and A. R. Ruis

Wisconsin Center for Education Research, University of Wisconsin–Madison, USA
DOI: 10.18608/hla17.015

ABSTRACT

In this article, we provide a worked example of a theory-based approach to learning analytics in the context of an educational game. We do this not to provide an ideal solution for others to emulate, but rather to explore the affordances of a theory-based - rather than data-driven - approach. We do so by presenting 1) epistemic frame theory as an approach to the conceptualization of learning; 2) data from an epistemic game, an approach to educational game design based on epistemic frame theory; and 3) epistemic network analysis (ENA), a technique for analyzing discourse and other data for evidence of complex thinking based on the same theory. We describe ENA through a specific analytic result, but our aim is to explore how this result exemplifies what we consider a key "best practice" in the field of learning analytics.

Keywords: Epistemic frame theory, epistemic game, epistemic network analysis (ENA), evidence centred design

In this chapter, we look at the role of theory in learning analytics. Researchers who study learning are blessed with unprecedented quantities of data, whether information about staggeringly large numbers of individuals or data showing the microscopic, moment-by-moment actions in the learning process. It is a brave new world. We can look at second-by-second changes in where students focus their attention, or examine what study skills are effective by looking at thousands of students in a MOOC.

As Wise and Shaffer (2016) argue in a special section of the *Journal of Learning Analytics*, however, it is dangerous to think that with enough information, the data can speak for themselves – that we can conduct analyses of learning without theories of learning. In fact, the opposite is true. With larger amounts of data, theory plays an even more critical role in analysis. Put in simple terms, most extant statistical tools were developed for datasets of a particular size and type: large enough so that random effects are normally distributed, but small enough to be obtained using traditional data collection techniques. Applying these techniques to datasets that are orders of magnitude larger in length and number of variables without a strong theoretical foundation is perilous at best.

In what follows, we look at this question not by analyzing the problems of applying statistics without a theoretical framework. What Wise and Shaffer suggest – and what the articles and commentaries in the special section of the *Journal of Learning Analytics* show – is that conducting theory-based learning analytics is challenging. As a result, our approach in what follows is to examine the role of theory in learning analytics through the use of a *worked example*: the presentation of a problem along with a step-by-step description of its solution (Atkinson, Derry, Renkl, & Wortham, 2000).

In doing so, our aim is not to provide an ideal solution for others to emulate, nor to suggest that our particular use of theory in learning analytics is better than others. Rather, our goal is to reflect on the importance of a theory-based approach – as opposed to an atheoretical or data-driven approach – to the analysis of large educational datasets. We do so by presenting *epistemic network analysis* (ENA; Andrist, Collier,

Gleicher, Mutlu, & Shaffer, 2015; Arastoopour, Shaffer, Swiecki, Ruis, & Chesler, 2016; Chesler et al., 2015; Nash & Shaffer, 2013; Rupp, Gustha, Mislevy, & Shaffer, 2010; Rupp, Sweet, & Choi, 2010; Shaffer et al., 2009; Shaffer, Collier, & Ruis, 2016; Svarovsky, 2011), a novel learning analytic technique. But importantly, we present ENA in the context of *epistemic frame theory* – the approach to learning on which ENA was based – and apply it to data from an *epistemic game*, an approach to educational game design based on epistemic frame theory. We thus describe ENA through a specific analytic result to examine how this result exemplifies the alignment of theory, data, and analysis as a "best practice" in the field.

DATA

The data we will use to explore this particular worked example come from an epistemic game (Shaffer, 2006, 2007), a simulation of authentic professional practice that helps students learn to think in the way that experts do. Specifically, the data come from the epistemic game *Land Science*, an online urban planning simulation in which students assume the role of interns at a fictitious firm competing for a redevelopment contract from the city of Lowell, Massachusetts. They work in small teams, communicating via chat and email, to develop a rezoning plan for the city that addresses the demands of different stakeholder groups. To do this, students review research briefs and other resources, conduct a survey of stakeholder preferences, and model the effects of land-use changes on pollution, revenue, housing, and other indicators using a GIS mapping tool. Because no rezoning plan can meet all stakeholder preferences, students must justify the decisions they make in their final proposals.

Land Science has been used with high school students and first-year college students more than 30 times. Our prior research (Bagley & Shaffer, 2009, 2015b; Nash, Bagley, & Shaffer, 2012; Nash & Shaffer, 2012; Shaffer, 2007) has shown that Land Science helps students learn content and practices in urban ecology, urban planning, and related fields, and it also helps them develop skills, interests, and motivation to improve performance in school.

As with many educational technologies, Land Science records all of the things that students do during the simulation, including their chats and emails, their notebooks and other work products, and every keystroke and mouse-click. This makes it possible to analyze not only students' final products but also the problem-solving processes they use.

In the worked example presented below, we examine the chat conversations from 311 students who used the same version of *Land Science*, including seven groups of college students (n = 155), eight groups of high school students (n = 110), and three groups of gifted and talented high school students (n = 46). In its entirety, this dataset contains 44,964 lines of chat.

THEORY

Our analysis of the chat data from *Land Science* is informed by *epistemic frame theory* (Shaffer, 2004, 2006, 2007, 2012). The theory of epistemic frames models the ways of thinking, acting, and being in the world of some *community of practice* (Lave & Wenger, 1991; Rohde & Shaffer, 2004). A community of practice, or a group of people with a common approach to framing, investigating, and solving problems, has a repertoire of knowledge and skills, a set of values that guides how skills and knowledge should be used, and a set of processes for making and justifying decisions. A community also has a common identity exhibited both through overt markers and through the enactment of skills, values, and decision-making processes characteristic of the community.

Becoming part of a community of practice, in other words, means acquiring a particular Discourse: a way of "talking, listening, writing, reading, acting, interacting, believing, valuing, and feeling (and using various objects, symbols, images, tools, and technologies)" (Gee, 1999, p. 719). A Discourse is the manifestation of a culture and, based on Goodwin's (1994) *professional vision*, an epistemic frame is the grammar of a Discourse: a formal description of the configuration of Discourse elements exhibited by members of a particular community of practice.

Importantly, however, it is not mere *possession* of relevant knowledge, skills, values, practices, and other attributes that characterizes the epistemic frame of a community, but *the particular set and configuration of them*. The concept of a *frame* comes from Goffman (1974) (see also Tannen, 1993). Activity is interpreted in terms of a frame: the rules and premises that shape perceptions and actions, or the set of norms and practices by which experiences are interpreted. An epistemic frame is thus revealed by the actions and interactions of an individual engaged in authentic tasks (or simulations of authentic tasks).

To identify analytically the connections among elements that make up an epistemic frame, we identify *co-occurrences* of them in student discourse – in this case, in the conversations they have in an online chat program. Researchers (Chesler et al., 2015; Dorogovtsev & Mendes, 2013; i Cancho & Solé, 2001; Landauer, McNamara, Dennis, & Kintsch, 2007; Lund & Burgess, 1996) 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). These concepts can be identified a priori from a theoretical or empirical analysis, or from an ethnographic study of the community in action.

ENA operationalizes epistemic frame theory by identifying co-occurrences in segments of discourse data and modelling the weighted structure of co-occurrences. ENA represents these patterns of co-occurrence in a dynamic network model that quantifies changes in the strength and composition of an epistemic frame over time – a process we describe in the next section.

ENA

ENA models the weighted structure of connections in discourse data, or in any kind of stanza-based interaction data. In what follows, we describe both the general principles of the ENA method and the specific process by which the current version of ENA software – www.epistemicnetwork.org – implements the ENA algorithms.

Stanza-Based Interaction Data

Before we describe how ENA operationalizes epistemic frame theory, it is important to understand how data is configured for analysis using ENA. Consider the simplified data in Table 15.1, which shows excerpts from two conversations held by one group of students in *Land Science*. In the five columns to the right are the concepts, or codes, whose pattern of association we want to model. In this case, the codes represent various aspects of professional urban planning practice – that is, various elements of an urban planning epistemic frame.

Note that sometimes we can see relations among the codes in a single utterance, as in In Line 3, where Jorge references knowledge of both social issues and environmental issues. In other cases, relations occur across utterances: in Line 10, Depesh talks about the trade-off involved in increasing open space, which responds to and builds on Natalie's more general comment about trade-offs in Line 8. However, we do not necessarily want to look at the relations among codes across *all* turns of talk. For example, two separate conversations are represented in Table 15.1. Both involve the same group of students (Group 3), but the conversations took place on two different days while the students were working on two different activities.

To create a network model of these data, we need to group the lines into stanzas. The key idea behind a stanza is that (a) codes in lines *anywhere within the*

Table 15.1. Edited Excerpt of Discourse Data Coded in ENA Format

Line	Activity	Group	Username	Created	Utterance	E.social.issues	S.zoning.codes	K.social.issues	K.zoning.codes	K.environment
1	VSV Meeting	3	Natalie	02/11/14 10:03	Okay, so what do the stakeholders want?	0	0	0	0	0
2	VSV Meeting	3	Depesh	02/11/14 10:03	talking w/ stakeholders, we learned that there are many issues within the city but there are some bariers that prevent these issues from being easily solved	0	0	1	0	0
3	VSV Meeting	3	Jorge	02/11/14 10:04	Yeah, the stakeholders care a lot about the environmental impact in the area as well as the need for low income housing	0	0	1	0	1
4	VSV Meeting	3	Depesh	02/11/14 10:04	they cared about different issues but they all wanted to create a healthy and livable community	0	0	1	0	0
5	VSV Meeting	3	Natalie	02/11/14 10:05	I agree. They are also worried about the quality of the water.	0	0	0	0	1
6	VSV Meeting	3	Jessie	02/11/14 10:06	and they want more housing opportunities for low-income residents	0	0	1	0	0
7	iPlan Meeting	3	Jorge	02/13/14 10:21	Quick question, what does the indicator P mean?	0	0	0	0	1
8	iPlan Meeting	3	Natalie	02/13/14 10:21	I found that certain indicators changed when altering the zoning designations of specific sites. Each change in zoning category came with its benefits and drawbacks. There was usually a tradeoff involved.	0	1	0	1	0
9	iPlan Meeting	3	Jessie	02/13/14 10:21	@Jorge: P = phosporous	0	0	0	0	1
10	iPlan Meeting	3	Depesh	02/13/14 10:22	yeah, if you add open space you can help run-off and nesting but hurt the job totals	1	0	1	0	1
11	iPlan Meeting	3	Jorge	02/13/14 10:25	Yeah, everything affects something.	0	0	0	0	0

same stanza are *related* to one another in the model, and (b) codes in lines that are *not in the same stanza* are *not related* to one another in the model. In this case, stanzas indicate which co-occurrences of concepts represent meaningful cognitive connections among the epistemic frame elements of urban planning.

ENA Models

To construct a network model from stanza-based interaction data, ENA *collapses* the stanzas. Usually this is done as a *binary* accumulation: if any line of data in the stanza contains code A, then the stanza contains code A. For example, the data shown in Table 1 would be collapsed as shown in Table 15.2 if we choose "Activity" to define the stanzas.

Table 15.2. Stanzas by Activity for Group 3

Activity	Group	E.social.issues	S.zoning.codes	K.social.issues	K.zoning.codes	K.environment
VSV Meeting	3	0	0	0	0	0
VSV Meeting	3	0	0	1	0	0

ENA then creates an *adjacency matrix* for each stanza, which summarizes the co-occurrence of codes (see Table 15.3). The diagonal of the matrix contains all zeros because codes in this model, and in general in ENA, do not co-occur with themselves. Each adjacency matrix, in this case, represents the connections that Group 3 made among urban planning epistemic frame elements during a particular activity. For example, in the VSV Meeting activity, K.social.issues, and K.environment both occurred in Group 3's discourse. The adjacency matrix representing that activity in Table 15.3 (left) thus contains a 1 in the cells that represent the co-occurrence of those two codes.

The adjacency matrices representing each stanza are then summed into a cumulative adjacency matrix for each unit of analysis in the dataset. The simple example shown in Table 15.3 would thus be represented by the cumulative adjacency matrix shown in Table 15.4. At the end of this process of accumulation, each unit in the dataset (in this case, each group) is associated with a cumulative adjacency matrix that represents the weighted pattern of co-occurrence (cognitive connections) among the codes (epistemic frame elements) for that unit.

To understand the structure of connections across different units – the relationships among their networks of connections, or the differences among their cumulative adjacency matrices – ENA represents each cumulative adjacency matrix as a vector in a high-dimensional space, where each vector is defined by the values in the upper diagonal half of the matrix. Note that the *dimensions* of this space correspond to the strength of association between every pair of codes.

Table 15.3. Stanzas by Activity for Group 3

Group 3 VSV Meeting	E.social.issues	S.zoning.codes	K.social.issues	K.zoning.codes	K.environment
E.social.issues	0	0	0	0	0
S.zoning.codes	0	0	0	0	0
K.social.issues	0	0	0	0	1
K.zoning.codes	0	0	0	0	0
K.environment	0	0	1	0	0

Group 3 iPlan Meeting	E.social.issues	S.zoning.codes	K.social.issues	K.zoning.codes	K.environment
E.social.issues	0	1	1	1	1
S.zoning.codes	1	0	1	1	1
K.social.issues	1	1	0	1	1
K.zoning.codes	1	1	1	0	1
K.environment	1	1	1	1	0

Table 15.4. Cumulative Adjacency Matrix for Group 3, Summing the Two Adjacency Matrices Shown in Table 3

Group 3	E.social.issues	S.zoning.codes	K.social.issues	K.zoning.codes	K.environment
E.social.issues	0	1	1	1	1
S.zoning.codes	1	0	1	1	1
K.social.issues	1	1	0	1	2
K.zoning.codes	1	1	1	0	1
K.environment	1	1	2	1	0

Before analyzing the data in ENA space, ENA divides each vector by its length to normalize the data. This is done because 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 *pattern* of association, and thus point in the same *direction*, but represent different *numbers* of stanzas, and thus have different *lengths*.

Once the data are normalized, ENA performs a *singular value decomposition* (SVD), a projection that centres the data but does not rescale it. This maximizes the variance accounted for in the data (similar to a principal components analysis). However, unlike a traditional PCA or factor analysis, (a) ENA is performed on the co-occurrences from the cumulative adjacency matrices, rather than on the counts or strengths of the codes themselves, and (b) ENA performs a sphere or cosine norm on the original data and centres it, but does not rescale the dimensions individually.

Interpretation of ENA Models

Once an ENA model is created, a suite of tools can be used to understand and create a meaningful interpretation. For example, in the *Land Science* dataset described above, the chat utterances of all students were coded for 24 urban planning epistemic frame elements (see Appendix I) using a previously developed and validated automated coding process (Bagley & Shaffer, 2015b; Nash & Shaffer, 2011). Codes relevant to authentic urban planning practice were developed based on an ethnographic study of how urban planners are trained (Bagley & Shaffer, 2015a).

ENA models are typically visualized using two-dimensions at a time, which facilitates interpretation. Figure 15.1, for example, shows the cumulative epistemic network of a high school student (Student A) who participated in *Land Science*. The network models the structure of connections among the elements of the student's urban planning epistemic frame. In this case, Student A's network shows a number of connections among knowledge elements, such as knowledge of social issues and knowledge of complex systems; epistemological elements, such as compromise; and the skill of using urban planning tools (such as a preference survey). The network is also weighted: thicker, more saturated lines represent stronger connections, whereas thinner, less saturated lines represent weaker connections. The thickness/saturation of a line is proportional to the number of stanzas in which the connection between the two epistemic frame elements occurred.

While we can draw some conclusions about this student's network – for example, Student A made cognitive connections mostly among basic knowledge and skills – in many cases, the salient features of a network are easier to identify in comparison with other networks. Figure 15.2 shows the urban planning epistemic network of a second high school student (Student B). Like Student A, Student B made a number of connections among basic knowledge elements, but Student B's network exhibits more and stronger connections overall as well as connections to additional elements, most notably to more advanced skills, such as scientific thinking, and to epistemological attributes.

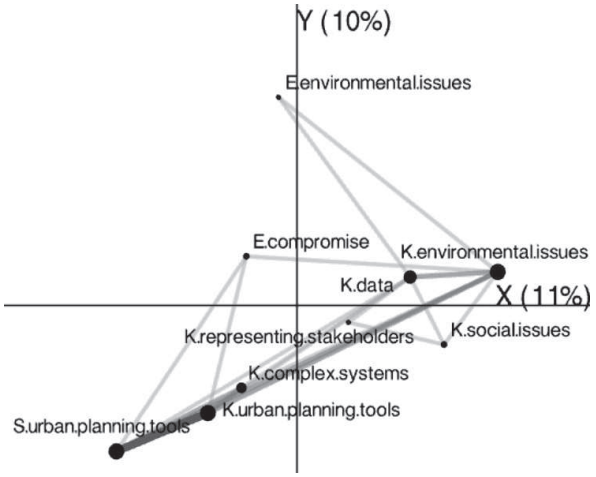

Figure 15.1. Epistemic network of a high school student (Student A) representing the structure of cognitive connections the student made while solving a simulated urban redevelopment problem. Percentages in parentheses indicate the total variance in the model accounted for by each dimension.the integration of multiple sources of data.

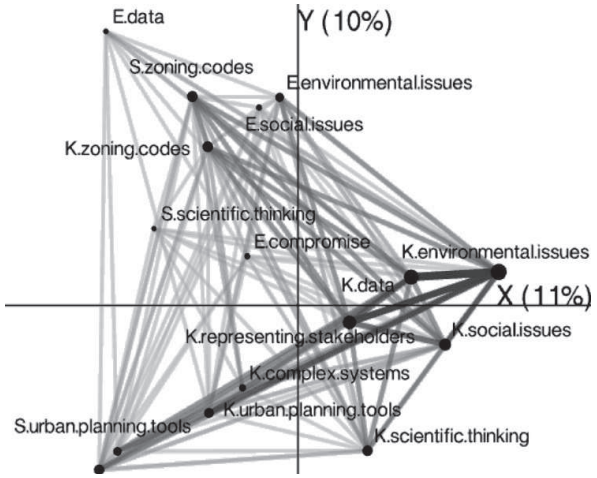

Figure 15.2. Epistemic network of a high school student (Student B) representing the cognitive connections the student made while solving a simulated urban redevelopment problem.

As discussed above, epistemic frame theory suggests that the epistemic frame of urban planning (or any community of practice) is defined by how and to what extent urban planning knowledge, skills, values, and other attributes are interconnected. In this example, ENA reveals that Student B's network is more overtly *epistemic*: she explained and justified her thinking in the way that urban planners do, and is thus learning to think like an urban planner.

What makes this comparison between Students A and B possible is that the nodes in both epistemic networks appear in exactly the same places in the network projection space – for these two students, and for all the students in the dataset. This invariance in node placement allows us to compare the network projections of different units directly, but this method of direct comparison only works for very small numbers of networks – what if we want to compare dozens or even hundreds of networks? For example, what if we want to compare all 110 high school students in this dataset, or compare the high school students with the college students? ENA makes this possible by representing each network as a single point in the projection space, such that each point is the *centroid* of the corresponding network.

The centroid of a network is similar to 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. The important 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 weighted structure of connections in the specific arrangement of the network model.*

The locations of the nodes in the network projection are determined by an optimization routine to minimize, for any given network, the distance between (a) the centroid of the network graph, and (b) the point that represents the network under the SVD rotation. Choosing fixed node positions to have the centroid of a network correspond to the position of the network in a projected space allows for characterization of the projection space – and thus of the salient differences among different networks in the ENA model. In this case, we can interpret the projection space in the following way: toward the lower left are basic professional skills, such as professional communication and use of urban planning tools; toward the right are knowledge elements related to the specific redevelopment problem and to knowledge of more general topics, such as data and scientific thinking; and toward the upper left are elements of more advanced urban planning thinking, especially epistemological elements – making and justifying decisions according to urban planning conventions – and the use of zoning codes.

We can thus compare a large number of different networks simultaneously because centroids located in the same part of the projection space represent networks with similar patterns of connections, while centroids located in different parts of the projection space represent networks with different patterns of connections[1]. This allows us to explore any number of research questions about students' urban planning epistemic frames. One question we might ask of the *Land Science* dataset is *How do the epistemic networks of the different student populations (college, high school, and gifted high school) differ*? For example, when we plot the centroids of the college students and the high school students (Figure 15.3), the two groups are distributed differently. To determine if the difference is statistically significant, we can perform an independent samples t test on the mean positions of the two populations in the projection space. The college students (dark) and high school students (light) are significantly different on both dimensions:

$\bar{x}_{College} = -0.083$, $\bar{x}_{HS} = 0.115$, $t = -7.025$, $p < 0.001$, *Cohen's* $d = -0.428$

$\bar{y}_{College} = 0.040$, $\bar{y}_{HS} = -0.045$, $t = 3.199$, $p = 0.002$, *Cohen's* $d = 0.186$

When the gifted and talented high school students are included in the analysis, in some respects they are more similar to the college students, and in others

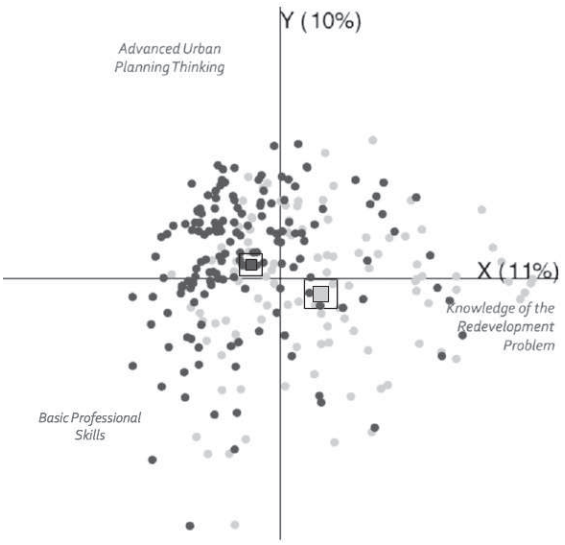

Figure 15.3. Centroids of college students (dark) and high school students (light) with the corresponding means (squares) and confidence intervals (boxes).

[1] It is possible, of course, that two networks with very different structures of connections will 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.

they are more similar to the high school students. The mean position of the gifted high school students in the projection space (Figure 15.4) is statistically significantly different from both the college students and the high school students only on the first (x) dimension:

$\bar{x}_{GiftedHS} = 0.007$, $\bar{x}_{College} = -0.083$, $t = 2.538$, $p = 0.013$, $Cohen's\ d = 0.202$

$\bar{x}_{GiftedHS} = 0.007$, $\bar{x}_{College} = 0.115$, $t = -2.736$, $p = 0.007$, $Cohen's\ d = -0.223$

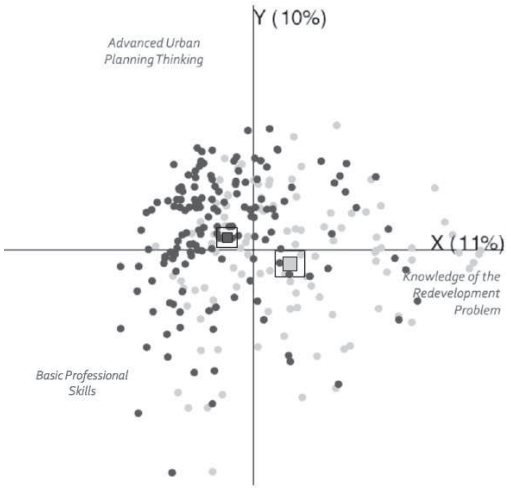

Figure 15.4. Mean network positions (squares) and confidence intervals (boxes) of the college students (left), high school students (right), and gifted high school students (center).

To determine what factors account for the differences among the three groups, we can compare their mean epistemic networks. As Figure 15.5 shows, the gifted high school students on average made more and stronger connections to elements of advanced urban planning thinking than the high school students, but not to the same extent as the college students. That is, they were somewhere between the high school and college students with respect to complex thinking in the domain. In contrast, the gifted high school students seem to be more similar to the high school students in that both populations made fewer connections than the college students between basic professional skills and advanced urban planning thinking. In other words, the gifted high school students are somewhere between the high school and college students intellectually, but they are more similar to the high school students in their level of basic professional and interpersonal skills.

Qualitative Triangulation of ENA Network Models

A key feature of ENA is the ability to trace connections in the model back to the original data – the chats, in this case – on which the connections are based. By clicking on the line connecting "epistemology of social issues" with "knowledge of data," we can access all the utterances that contributed to this connection in the network graph. Figure 15.6 shows an excerpt of the utterances that contributed to this connection in one college student's epistemic network.

The text is coloured such that stanzas or utterances containing only the first code are shown in red, those containing only the second code are shown in blue, those containing both codes are shown in purple, and those containing neither code are shown in black. The stanza (i.e., the activity) "Final Proposal Reflection," for example, is coloured purple because it contains utterances coded for both E.social.issues and K.data: the first (red) utterance justifies a land-use change based on a desire to improve the city (epistemology of social issues), while the second utterance references knowledge about the effects of zoning changes on atmospheric carbon dioxide levels (knowledge of data).

This feature of ENA allows us to close the interpretive loop (see Figure 15.7). We started with a dataset that was coded for urban planning epistemic frame elements; we used the coded data to create and visualize network models of students' urban planning thinking based on the co-occurrence of frame elements; then, if we want to understand the basis for any of the connections in the network models, we can return to the original utterances. ENA thus enables quantitative analysis of qualitative data in such a way that the quantitative results can be validated qualitatively.

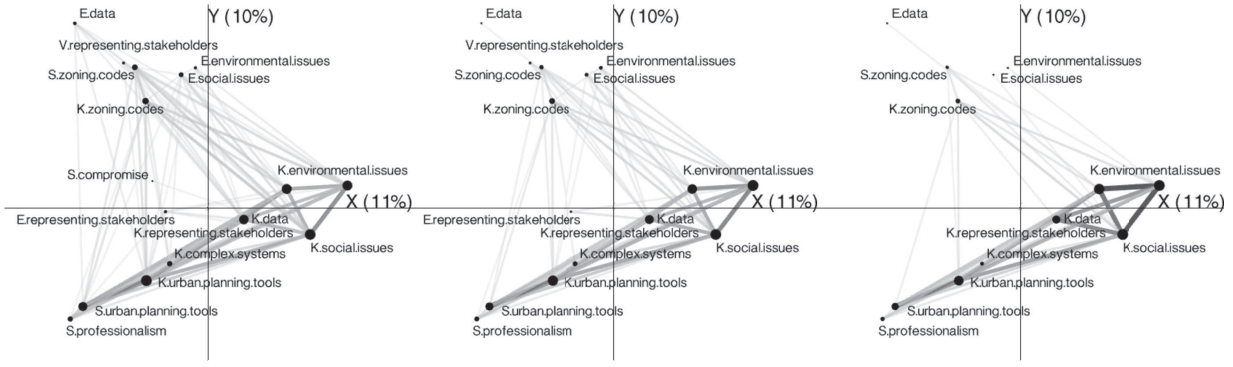

Figure 15.5. Mean epistemic networks of college students (red, left), gifted high school students (green, centre) and high school students (blue, right).

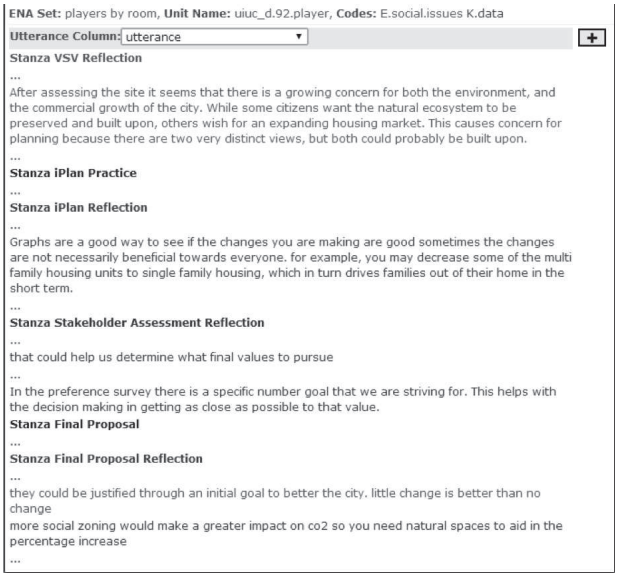

Figure 15.6. Excerpt of the chat utterances that contributed to the connection between epistemology of social issues and knowledge of data in one college student's epistemic network. In the ENA toolkit, the text of each utterance is coloured to indicate whether it contains code A (red), code B (blue), both (purple), or neither (black).

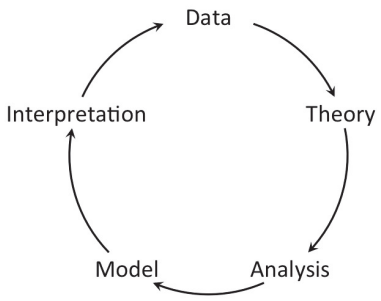

Figure 15.7. Good theory-based learning analytics "closes the interpretive loop" by making it possible to validate the interpretation of a model against the original data.

DISCUSSION

In working through this analysis, our aim was not to provide an *ideal* example for others to emulate, nor to suggest that epistemic frame theory has any particular analytic advantages over other learning theories, but to provide context for a more general discussion of methodology in learning analytics and educational data mining. As analyses of large educational datasets have become more common, a key application is obtaining empirical evidence to "refine and extend educational theories and well-known educational phenomena, towards gaining deeper understanding of the key factors impacting learning" (Baker & Yacef, 2009, p. 7). In other words, a theoretical framework guides the selection of variables and development of hypotheses, which can lead to an explanation for why observed phenomena are occurring.

In the worked example presented above, we used the theory of epistemic frames to guide our analysis of student chat data in an urban planning simulation. Epistemic frame theory suggests that learning can be characterized by the structure of connections that students make among elements of authentic practice. Our analytic approach, ENA, uses discourse data to construct models of student learning that are visualized as network graphs, mathematical representations of patterns of connections. The analysis is thus an operationalization of a particular theoretical approach to understanding learning.

One way to conceptualize the linkage between theory, data, and analysis is through *evidence centred design* (Mislevy & Riconscente, 2006; Rupp, Gustha et al., 2010; Shaffer et al., 2009). In evidence-centred design, an analytic framework is composed of three connected models: a student model, an evidence model, and a task model (see Figure 8; Mislevy, Steinberg, & Almond,

1999; Mislevy, 2006). The *student model* represents the characteristics of the student that we want to assess, or more generally the outcome we are trying to model or measure. The *task model* represents the activities and the data that will be used to measure the outcomes in the student model. The student (outcome) model and task (data) model are linked by an *evidence model*, which details the analytic tools and techniques that will be used to warrant conclusions about the outcomes based on the data.

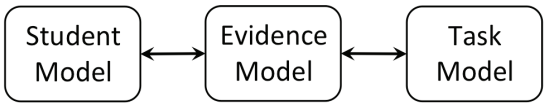

Figure 15.9. Models in an ECD Assessment (adapted from Mislevy, 2006).

Our worked example illustrates an approach to learning analytics in which each of the models (student, evidence, and task) are derived from the same theoretical framework – in this case, epistemic frame theory (see Figure 15.9).

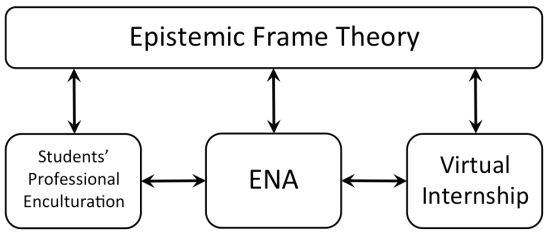

Figure 15.9. Mean network positions (squares) and confidence intervals (boxes) of the college students (left), high school students (right), and gifted high school students (center).

The result is an approach to analyzing expertise in the context of (simulated) complex problem solving that is guided by a particular theory of expertise and validated empirically. But critically, the empirical grounding of the results does not rely solely on statistical significance: because of the linkages between the different models or layers of the evidentiary argument, the interpretation of the statistics – the meaning of the model – can be verified in the original data.

Despite these advantages of a theory-based approach to data analysis, there has been a significant expansion in studies that take a radically atheoretical approach to discovery. Wired editor-in-chief Chris Anderson (2008) has even claimed that theory-based inquiry is unnecessary in the age of big data. "Petabytes [of data] allow us to say: 'Correlation is enough,'" Anderson suggests. "We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot." Despite the fact that most scientists would be deeply uncomfortable with the idea that causation is unimportant, Anderson's approach to the analysis of big data – "to view data mathematically first and establish a context for it later" – is a commonly applied method in data mining.

Of course, with a sufficiently large dataset and the ability to run it through dozens if not hundreds of mathematical models, statistically significant patterns will be found. But statistical significance does not imply conceptual or even practical significance. This does not imply that all theory-based approaches to analyzing large collections of data are ideal or even worthwhile. There is bad theory, just as there is bad empiricism – and even good theory badly operationalized or applied. Nor are we suggesting that the worked example above, or even more generally the theories and methods that we chose, are ideal in all circumstances.

Our argument, rather, is that there are distinct advantages to taking a theory-based approach to the analysis of large educational datasets. The worked example above illustrates how in theory-guided learning analytics, an explicit theoretical framework guides the search for understanding in a corpus of data and the selection of appropriate analytic methods. These linkages between data, theory, and analysis thus provide the ability to interpret the results sensibly and meaningfully.

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.

REFERENCES

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*, *16*(7). https://www.wired.com/2008/06/pb-theory/

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). journal.frontiersin.org/article/10.3389/fpsyg.2015.01016/pdf

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*(2), in press.

Atkinson, R., Derry, S., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, *70*(2), 181–214.

Bagley, E. A., & Shaffer, D. W. (2009). When people get in the way: Promoting civic thinking through epistemic game play. *International Journal of Gaming and Computer-Mediated Simulations*, *1*(1), 36–52.

Bagley, E. A., & Shaffer, D. W. (2015a). Learning in an urban and regional planning practicum: The view from educational ethnography. *Journal of Interactive Learning Research*, *26*(4), 369–393.

Bagley, E. A., & Shaffer, D. W. (2015b). Stop talking and type: Comparing virtual and face-to-face mentoring in an epistemic game. *Journal of Computer Assisted Learning*, *31*(6), 606–622.

Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17.

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), 024701:1–8.

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: Routledge.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Cambridge, MA: Harvard University Press.

Goodwin, C. (1994). Professional vision. *American Anthropologist*, *96*(3), 606–633.

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.

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

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, MA: Cambridge University Press.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

Mislevy, R. J. (2006). *Issues of structure and issues of scale in assessment from a situative/sociocultural perspective*. CSE Technical Report 668. Los Angeles, CA.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design*. Educational Testing Service. http://www.education.umd.edu/EDMS/mislevy/papers/ECD_overview.html

Mislevy, R., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 61–90). Mahwah, NJ: Erlbaum.

Nash, P., Bagley, E. A., & Shaffer, D. W. (2012). Playing for public interest: Epistemic games as civic engagement activities. *American Educational Research Association Annual Conference* (AERA). Vancouver, BC, Canada.

Nash, P., & Shaffer, D. W. (2011). Mentor modeling: The internalization of modeled professional thinking in an epistemic game. *Journal of Computer Assisted Learning*, 27(2), 173–189.

Nash, P., & Shaffer, D. W. (2012). Epistemic youth development: Educational games as youth development activities. *American Educational Research Association Annual Conference* (AERA). Vancouver, BC, Canada.

Nash, P., & Shaffer, D. W. (2013). Epistemic trajectories: Mentoring in a game design practicum. *Instructional Science*, 41(4), 745–771.

Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review E*, 70(5), 56131.

Rohde, M., & Shaffer, D. W. (2004). Us, ourselves and we: Thoughts about social (self-) categorization. *Association for Computing Machinery* (ACM) *SigGROUP Bulletin*, 24(3), 19–24.

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* (EDM2010), 11–13 June 2010, Pittsburgh, PA, USA (pp. 319–320). International Educational Data Mining Society.

Shaffer, D. W. (2004). Epistemic frames and islands of expertise: Learning from infusion experiences. *Proceedings of the 6th International Conference of the Learning Sciences* (ICLS 2004): *Embracing Diversity in the Learning Sciences*, 22–26 June 2004, Santa Monica, CA, USA (pp. 473–480). Mahwah, NJ: Lawrence Erlbaum.

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., Collier, W., & Ruis, A. R. (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.

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.

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.

Tannen, D. (1993). *Framing in discourse*. Oxford, UK: Oxford University Press.

Wise, A. F., & Shaffer, D. W. (2016). Why theory matters more than ever in the age of big data. Journal of Learning Analytics, 2(2), 5–13.

APPENDIX I

URBAN PLANNING EPISTEMIC FRAME CODE SET

Code	Code Description	Example
Epistemology of Social Issues	Using social issues to justify a decision or to ask for a justification of a decision (e.g., jobs, crime, housing)	Because it effects their [the stakeholders'] business
Epistemology of Environmental Issues	Using environmental issues to justify a decision or to ask for a justification of a decision (e.g., runoff, pollution, animal habitats)	except why wouldn't we think that they'd care about social and environmental issues?
Epistemology of Representing Stakeholders	Using the representation of stakeholders to justify a decision or to ask for a justification of a decision (e.g., referring to a specific stakeholders' needs by name, referring to the needs of the stakeholder group)	Try and understand what the stakeholders want and why. That might help you come up with a plan that they'll support.
Epistemology of Data	Using data to justify a decision or to ask for a justification of a decision (e.g., numbers, collecting information)	there are three different groups L-EDC, L-CAG, and LC-RWC and each group have different recommendations. all three of the numbers are different so it seems impossible to meet all three of those numbers at the same time. which group are we suppose to follow?
Epistemology of Compromise	Using compromise to justify a decision or to ask for a justification of a decision (e.g., balancing stakeholders' needs, referring explicitly to compromise)	You may have to make compromises, because the stakeholder groups sometimes disagree.
Value of Representing Stakeholders	Utterances indicating that players should, should not, must, must not, ought to care about representing stakeholders	First, let's make sure we agree on what the stakeholders want.
Value of Complex Systems	Utterances indicating that players should, should not, must, must not, ought to care about relationships between parts of a larger system	Flavian and Natalie both want to protect the enviornment in Lowell, whereas Lee and Nathanial both want to increase housing and economic groth
Value of Compromise	Utterances indicating that players should, should not, must, must not, ought to care about compromise	however, we may need to make compromises so everyone can live with the changes.
Skill of Professionalism	Utterance indicating that a skill related to professionalism was performed (e.g., sending an email)	OK. I finished the interview and I read the resources.
Skill of Data	Utterance indicating that a skill related to data was performed (e.g., entering values into the TIM, referring to values of TIM output and stakeholder assessment values)	we listened to the feedbacks and chose the best numbers according to what they wanted so i think 99.9999
Skill of Scientific Thinking	Utterance indicating that a skill related to scientific thinking was performed (e.g., making hypotheses, testing hypotheses, developing models)	so we put test numbers into the TIM to see how they react?
Skill of Compromise	Utterance indicating that a skill related to compromise was performed	however, we may need to make compromises so everyone can live with the changes.
Identity of Urban Planners	Utterance indicating that one or one's group identifies as an urban planner	Sure. iPlan is a model, so as a planner, when you make changes in iPlan, it shows us what might happen if you made those changes in the real world.
Identity of Interns	Utterance indicating that one or one's group identifies as interns	please remember we are professionals and all our chat and work that we hand in should reflect that.
Knowledge of Social Issues	Utterance referring to social issues (e.g., jobs, crime, housing)	I worked with a group that cared about nests, housing, phosphorous, and runoffs

Code	Code Description	Example
Knowledge of Environmental Issues	Utterance referring to environmental issues (e.g., runoff, pollution, animal habitats)	I worked with the Connecticut River Water council and they cared about the environment.
Knowledge of Representing Stakeholders	Utterance referring to representing stakeholders (e.g., referring to a specific stakeholders' needs by name, referring to the needs of the stakeholder group)	You may have to make compromises, because the stakeholder groups sometimes disagree
Knowledge of Complex Systems	Utterance referring to relationships between parts of a larger system	inorder to reduce co2 levels I had to increase the bird populations
Knowledge of Urban Planning Tools	Utterance referring to urban planning tools (e.g., iPlan, TIM, Preference Survey)	so wer making one iplan?
Knowledge of Zoning Codes	Utterance referring to zoning codes (e.g., open space, industrial space, housing, wetlands)	well if you changed a piece of land from open space to industrial, you create jobs, but the CO might increase as well
Knowledge of Data	Utterance referring to data (e.g., entering values into the TIM, referring to values of TIM output and stakeholder assessment values)	i got really positive feedback except for my runoff number. i need to reduce that a bit more
Knowledge of Scientific Thinking	Utterance referring to scientific thinking (e.g., making hypothesis, testing hypothesis, developing models)	I guess it would be a model that you can test and observe the effects that would happen in the real world.
Skill of Zoning Codes	Utterance indicating that a skill related to zoning codes was performed	Converting land to open space/wetlands increases the bird population. That's one of the stakeholder's concerns, so I should maybe mark more open space and less industrial or commercial
Skill of Urban Planning Tools	Utterance indicating that a skill related to urban planning tools was performed	To get the desired result you'd have to change a few different indicators.