Network-Based Tools for the Visualization and Analysis of Domain Models

Paper presented as the annual meeting of the American Educational Research Association, Philadelphia, PA

Hua Wei

April 2014

ALWAYS LEARNING PEARSON

Abstract

A domain model depicts relationships among the important knowledge and skills that students are expected to learn in the subject domain. A graphic representation of a domain model facilitates understanding of the complex relationships in the domain and informs subsequent efforts of assessment development. The most important type of relationships in domain models is the prerequisite relationship, which spells out the idea that a knowledge element must be acquired before learning another. How to visualize the network of prerequisite relationships and represent it in ways that help assessment designers gain insights into the domain is essential in domain modeling. This paper illustrates the applicability of NodeXL, a network visualization and analysis tool, in visualizing prerequisite relationships and discusses how key aspects of network structure can be identified. This paper also demonstrates how network metrics are used to diagnose and explore domain model structure and reveal details about individual elements of the structure.

Keywords: domain model, network visualization, prerequisite relationship, NodeXL

Network-Based Tools for the Visualization and Analysis of Domain Models

Domain modeling is an important stage in educational assessment design (Mislevy & Riconscente, 2006). At this stage, information collected from domain analyses, which are typically conducted by subject matter experts (SMEs), is used to structure relationships among the important knowledge and skills that students are expected to learn in the subject domain. A graphic representation of a domain model, either a high-level sketch or a more detailed relationship map, facilitates understanding of the complex relationships in the domain and informs subsequent efforts of assessment development.

A visual display of the domain model, such as graphs or maps, has the great advantage of forcing people into "recognizing patterns and interpreting spatial relationships" (Mislevy et al., 2010, p. 7). As educational assessment systems become more complex and the assumptions upon which they are based become more hidden, there is an increased need for tools to explore, analyze and communicate these foundational structures. In these more complex arrangements, communication of networks of dependencies requires the use of network-based graphics as well as corresponding algebraic expressions of network structure.

Following the logic of exploratory data analysis and data visualization (Behrens et al., 2012), insights into the structure of a complex set of dependencies can be communicated both to analysts and end-users using network visualization tools. This paper illustrates the use of a network visualization tool in domain modeling and discusses how key aspects of network structure (and logical dependence) can be identified. A second goal of the paper is to illustrate the application of network metrics in the quantification of network structure. In the network literature there are a number of common node-level metrics including in-degree and out-degree as well as betweenness and closeness centrality. While originally devised to represent social

importance, these metrics have corresponding dependency interpretations as well. In this paper we demonstrate how these metrics can be used to diagnose and explore domain model structure for both design and communicative purposes.

Network of Prerequisite Relationships

The most important type of relationships in domain models is the prerequisite relationship, which spells out the idea that a knowledge element must be acquired before learning another. Prerequisite relationships are usually determined for curriculum design. The typical question to ask when defining a prerequisite relationship between instructional units is that given an instructional unit, which prior units significantly influence students' success in this unit. Prerequisite relationships can also be determined between aspects of content, such as learning objectives. Definitions of prerequisite relationships are usually based on expert opinions or theoretical models (Chi & Koeske 1983).

In a well-defined domain such as mathematics where the content is typically structured as a hierarchy of strands, topics, and objectives, relationships among the elements of the hierarchy can be one-to-one, but in many cases, one-to-many and possibly many-to-many. How to visualize the network of prerequisite relationships and represent it in ways that help assessment designers gain insights into the domain is essential in domain modeling.

NodeXL: A Network Visualization and Analysis Tool

NodeXL is an open-source network visualization and analysis software package for Microsoft Excel 2007/2010. It can be freely downloaded at http://nodexl.codeplex.com/. It is a powerful and easy-to-use interactive tool that leverages Microsoft Excel as the platform for representing generic graph data, and performing advanced network analysis and visual

exploration of network data. NodeXL is widely used for exploring and analyzing social media network data, such as data imported from sources like Facebook and Twitter.

As an add-in for Excel, NodeXL extends the existing features of Excel spreadsheet with added data analysis and charting capabilities. NodeXL opens as an Excel workbook with identified worksheets containing elements of graph structures such as edges and vertices, and graph metrics. Figure 1 shows the interface of NodeXL. The added graph display pane provides instant graphical representation of relationships of complex network data.

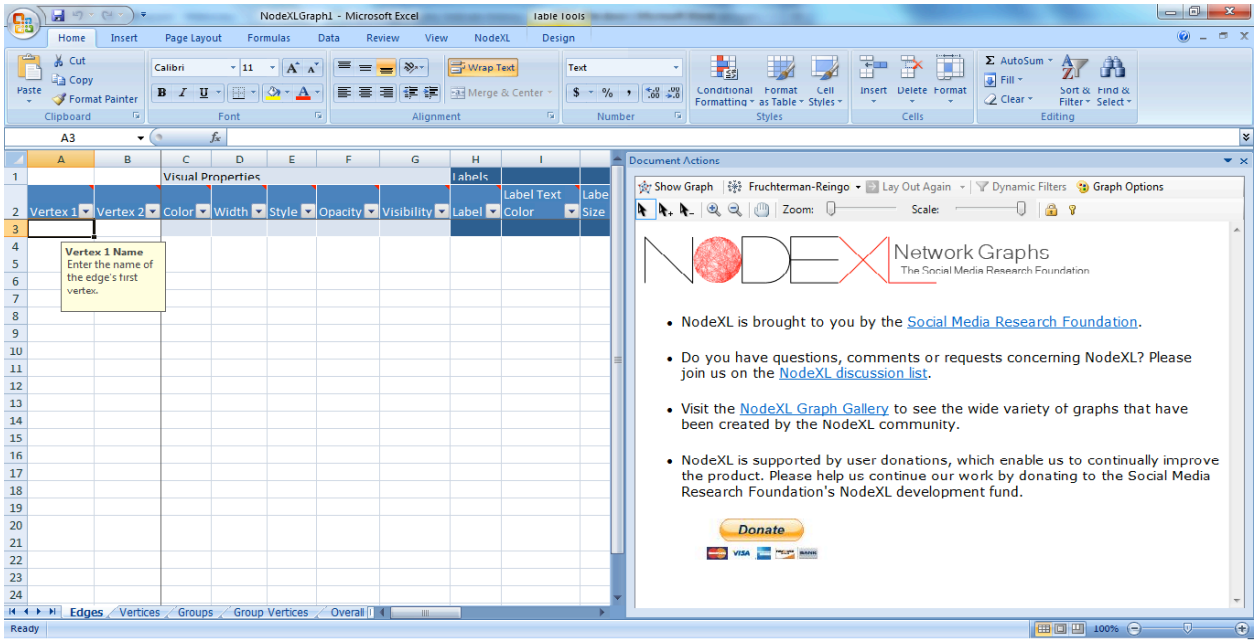

Figure 1. NodeXL Interface

Network data are entered or imported into NodeXL in the form of edge lists, i.e., pairs of vertices/nodes. Each vertex represents a unique entity in the network. Each edge connecting two vertices represents a relationship that exists between them. The relationship can be directed or undirected. Each relationship can be annotated with additional columns that contain information about the relationship, such as strength or date range.

NodeXL calculates statistics for individual vertices, and these statistics are added to the spreadsheet as additional columns that can be further used during analysis and visualization. NodeXL also calculates metrics for the overall graph, and these metrics are displayed in the workbook. Table 1 is a summary of the (often used) metrics generated by NodeXL and their descriptions.

Table 1

Network Analysis Metrics Generated by NodeXL

Classification	Type	Description
Overall Network Metrics	Density	• The number of relationships observed to be present in a network divided by the total number of possible relationships that could be present. • Used to describe the level of interconnectedness of the vertices
	Geodesic Distance	• Maximum geodesic distance (also known as diameter) is the distance between the two most distant vertices. • Average geodesic distance is the average distance between any two vertices in the network.
Vertex-Specific Metrics	Degree	• Simple counts of the total number of connections linked to a vertex • In-degree is the number of edges that lead into the vertex. Out-degree is the number of edges that lead out of the vertex.
	Betweenness Centrality	• A measure of how often a given vertex lies on the shortest path between any two other vertices • Can be considered as a "bridge" score, which indicates how much removing a vertex would disrupt the connections between other vertices in the network
	Closeness Centrality	• Average shortest distance from a vertex to every other vertex in the network • Can be considered as a "distance" score
	Eigenvector Centrality	• Measures the degrees of the vertices that a vertex is connected to • Useful in determining which vertex is connected to the most connected vertices • A vertex with high eigenvector centrality is connected to other vertices with high eigenvector centrality.
	Clustering Coefficient	• The number of edges connecting a vertex's neighbors divided by the total number of possible edges between the vertex's neighbors • A measure of how connected a vertex's neighbors are to one another

An Example of NodeXL Usage

The data analyzed in this study are a list of prerequisite relationships that accompanies a widely used textbook intended for college-level developmental mathematics classes. The prerequisite relationships are defined between learning objectives, which are grouped into sub-topics and then topics in the textbook. Each relationship is indicated as either strong or weak. There are 517 learning objectives and 966 prerequisite relationships in the data, as shown in Table 2.

Table 2

Network Data Statistics

Metrics	Value
Vertices	517
Unique Edges	966
Graph Density	0.0036

Figure 2 displays a graph of all the prerequisite relationships in the data. A prerequisite relationship is directed in the sense that a relationship from A to B means A is a prerequisite to B.

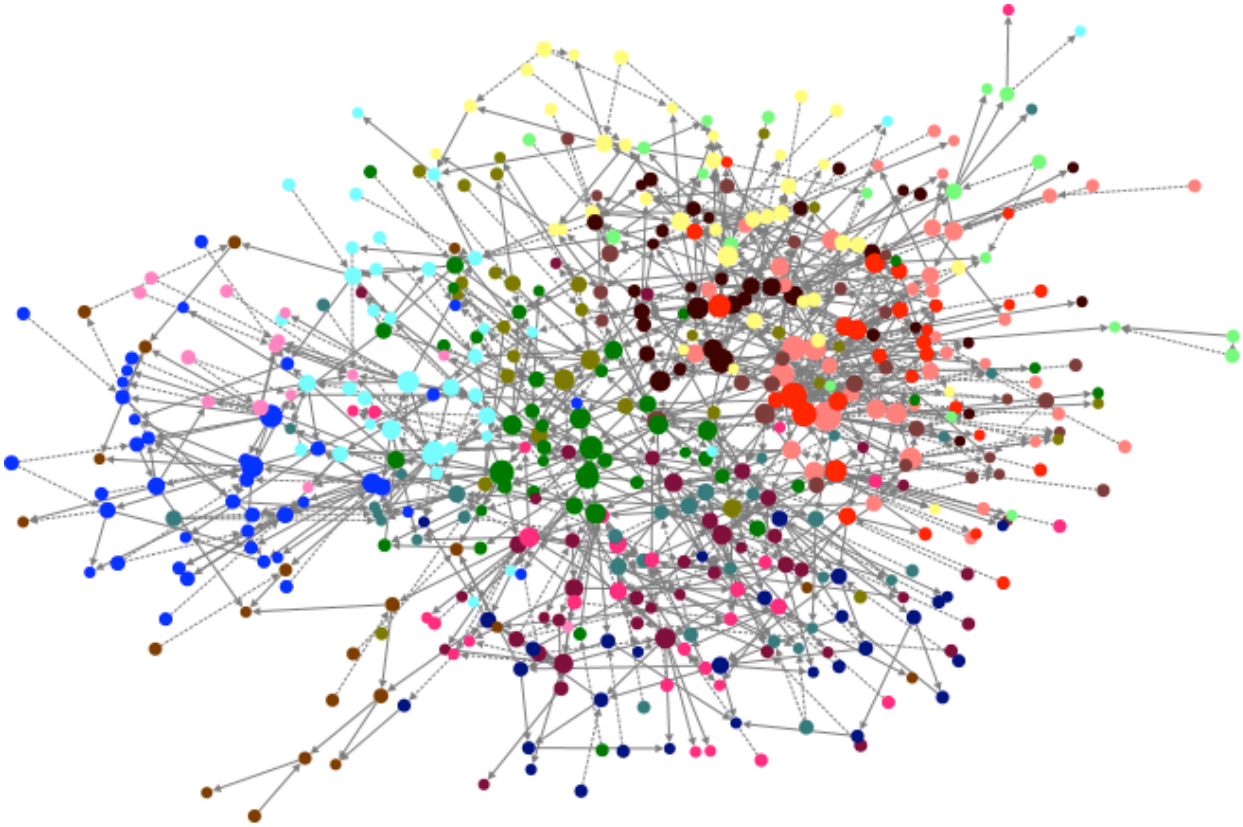

Figure 2. Graph Visualization of All Network Data

NodeXL provides a variety of display options to specify the appearance of individual edges and nodes as well as the overall layout of the network. The lines between nodes that represent edges can have different widths, colors, styles, and levels of opacity depending on the values of the selected attributes of the data. Similarly, each node can be set to have a different location, size, color, opacity, and shape. In the visualization as shown in Figure 2, nodes with different colors indicate objectives from different content topics. For example, red nodes are objectives under the topic of multiplying and dividing fractions. The size of each node is an indication of its out-degree. Nodes with greater out-degrees appear bigger than nodes with lower out-degrees. In other words, the bigger the node is, the more frequent it is a prerequisite to other objectives. In addition, the strength of a relationship is indicated with different line styles. A strong relationship is represented by a solid line and a weak relationship by a dotted line. The

graph is laid out using the Harel-Koren Fast Multiscale algorithm, which minimizes line crossings, and tends to produce a graph more aesthetically pleasing and readable.

Although different display options are employed in the visualization, the whole graph looks chaotic and is hard to read and interpret. This is so because a network with a few hundred nodes and a thousand edges can be very challenging for visualization and interpretation. Networks with this many elements need to be reduced to a limited set either by aggregation or by selectively focusing on a small region of the larger network.

NodeXL's strength is the ease with which you can filter and customize the network visualization. To refine the graph visualization, only links between nodes from different content topics were retained and analyzed. In other words, prerequisite relationships between objectives from the same content topic were discarded and we were interested in visualizing only relationships between objectives from different topics. Figure 3 shows the refined network graph. In this visualization, edges representing weak relationships were not shown.

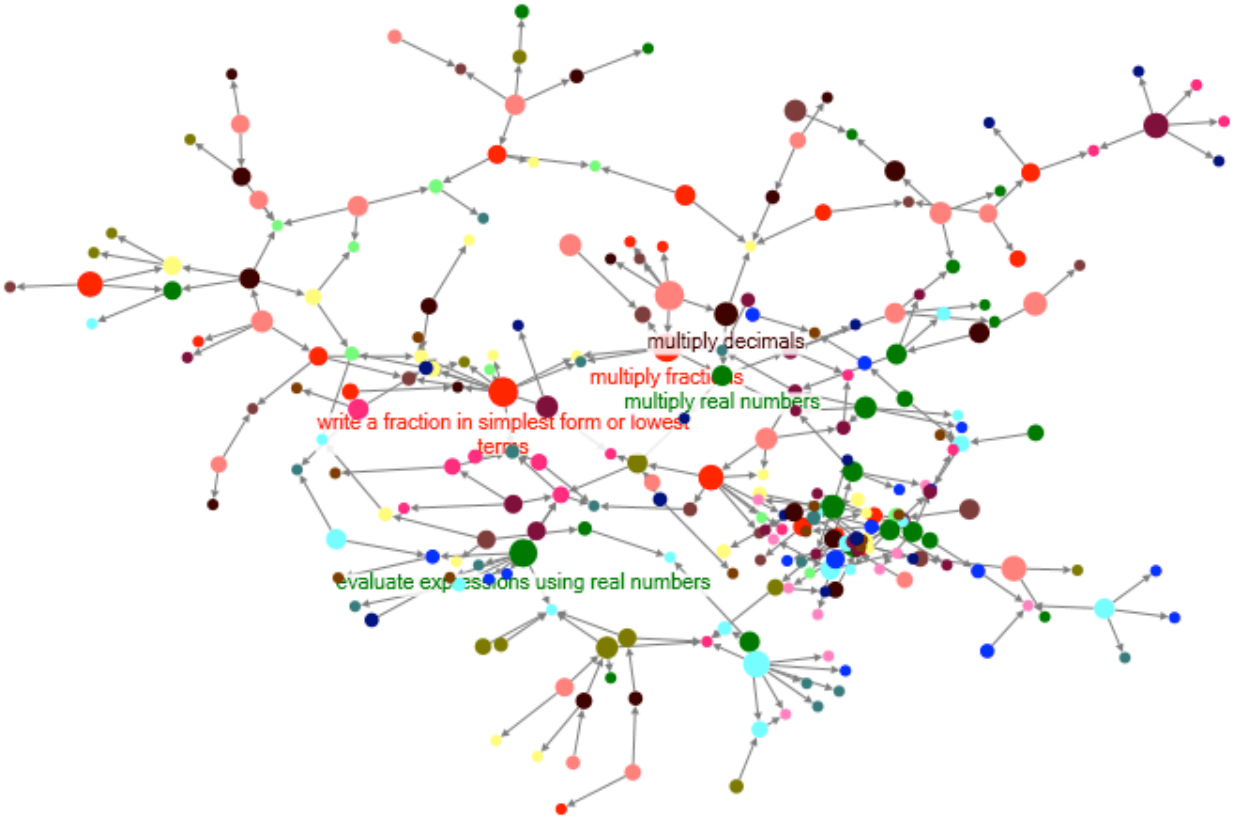

Figure 3. Graph Visualization of Filtered Network Data

As shown in the graph, the labeled nodes are visible "bridging points" in the network. They represent skills that must be acquired before advancing to learn other skills. These nodes have large betweenness centrality values, and many of them have greater out-degrees than other nodes. The "bridging" position of these objectives is easily observed in this graph though may be easily overlooked in a long list of object-to-objective dependency tables.

In order to facilitate understanding of the prerequisite relationships between different content topics in the book, I created a new network dataset which was based on the original network of relationships between objectives. The new dataset consists of connections between pairs of topics, with strength of each connection indicated by the weighted (i.e., the weighting assigned to a weak relationship was half of the weighting assigned to a strong relationship) total

number of prerequisite relationships that exist between objectives from the two topics. Figure 4 shows the network graph of relationships for the new dataset.

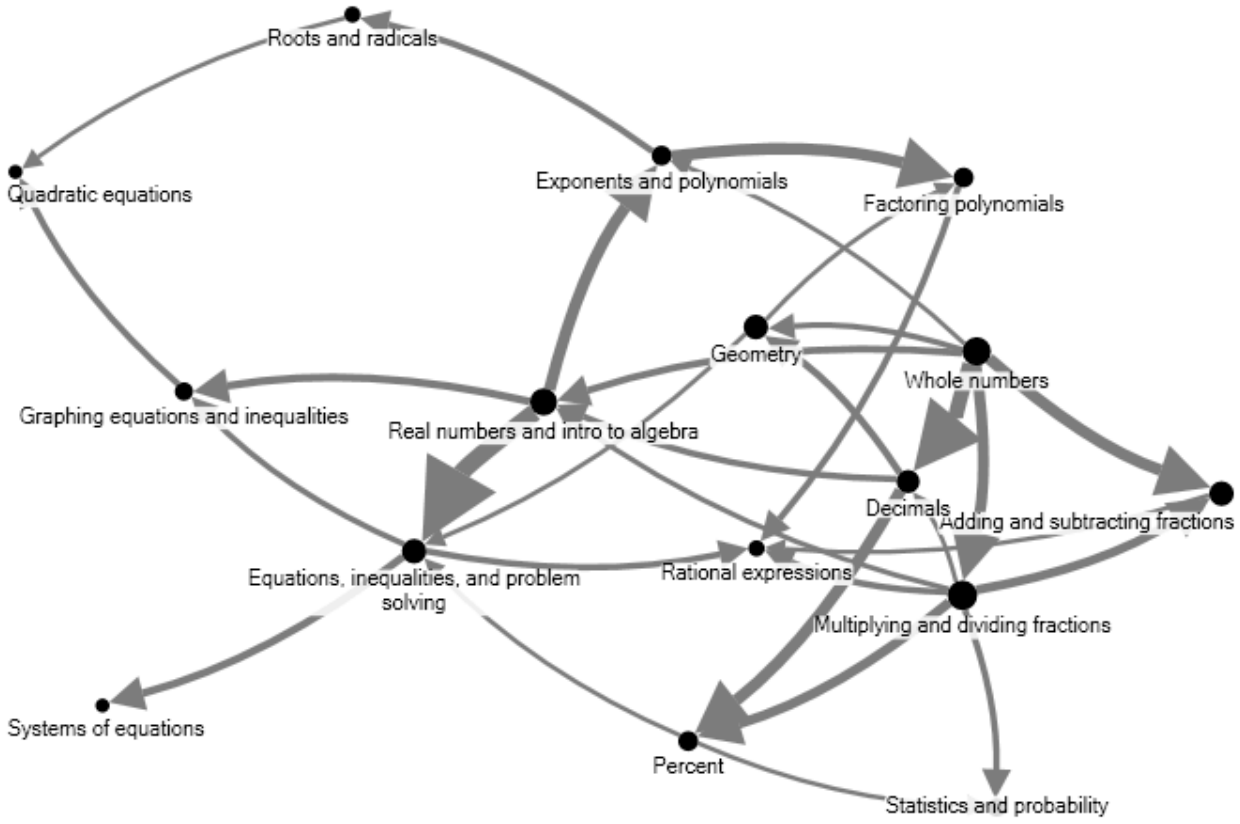

Figure 4. Graph Visualization of Customized Network Data

In this visualization, the size of each node again indicates its out-degree value, and the width of each line indicates the strength of the connection. Obviously, the topics that are introduced earlier in the curriculum, such as whole numbers, and multiplying and dividing fractions, are prerequisites to many other topics that are taught later. The topic of real numbers and introduction to algebra is easily discerned as the choke point, which means proficiency with this topic is essential to the learning of subsequent topics in the textbook.

Conclusions

In this paper, we illustrated the applicability of social network analysis tools in representing prerequisite relationships for the purpose of domain modeling. We introduced NodeXL, a social network visualization and analysis software, for visualizing the structure in relational data and quantifying the network structures. We analyzed the prerequisite relationship data for a mathematics textbook and showcased how the combination of visual representation and statistical analysis provide additional non-trivial observations about the dependency structures of different content elements in the book. We believe that the patterns discovered in the visualization and information revealed in the analysis are valuable to content experts as well as to assessment designers.

One advantage of NodeXL is that it does not involve the use of a programming language, and thus enables domain experts to explore network data in their field without the need to acquire technical skills or experience. Other network analysis tools like Pajek and UCINet that provide graphical interfaces and richer sets of graph metrics require complex data manipulation, graphing, and visualization, and thus are only intended for expert practitioners. However, NodeXL is not without limitations. For example, there is a limit on the number of vertices and edges that NodeXL can handle. Besides, some of the interactivity features are lost once the graph is exported and viewed outside NodeXL. Therefore, with massive network data sets or more advanced analysis and visualization needs, programming languages such as R or Java may be more appropriate. For example, heatmaps (Vuong, Nixon, & Towle, 2013) and hive plots (see http://bost.ocks.org/mike/hive/ for an example and a brief explanation) are created to represent dependency structures. With the wide variety of network visualization and analysis tools

available, the goal is to produce visual representations that facilitate qualitative interpretation of dependency data and quantitative evaluation of the properties of the dependency structures.

References

Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2012). Evidence centered design for learning and assessment in the digital world. In M. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.). *Technology-based assessments for 21st Century skills: Theoretical and practical implications from modern research* (pp. 13-54). Charlotte, NC: Information Age Publishing.

Chi, M., & Koeske, R. (1983). Network representation of a child's dinosaur knowledge. *Developmental Psychology, 19*(1), 29–39.

Hansen, D., Shneiderman, B., & Smith, M. A. (2010). *Analyzing social media networks with NodeXL: Insights from a connected world*. Burlington, MA: Morgan Kaufmann.

Mislevy, R.J., Behrens, J.T., Bennett, R.E., Demark, S.F., Frezzo, D.C., Levy, R., Robinson, D.H., Rutstein, D.W., Shute, V.J., Stanley, K., & Winters, F.I. (2010). On the roles of external knowledge representations in assessment design. *Journal of Technology, Learning, and Assessment, 8*(2).

Mislevy, R.J., & 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.

Vuong, A., Nixon, T., & Towle, B. (2011). A method for finding prerequisites within a curriculum. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, & J. Stamper (Eds.), *Proceedings of the Fourth International Conference on Educational Data Mining* (pp. 211-216). Eindhoven, Netherlands.