Applying Graph-based Data Mining Concepts to the Educational Sphere

András London, Áron Pelyhe, Csaba Holló, Tamás Németh

Abstract: *In this study, we discuss the possible application of the ubiquitous complex network approach for information extraction from educational data. Since a huge amount of data (which is detailed as well) is produced by the complex administration systems of educational institutes, instead of the classical statistical methods, new types of data processing techniques are required to handle it. We define several suitable network representations of students, teachers and subjects in public education and present some possible ways of how graph mining techniques can be used to get detailed information about them. Depending on the construction of the underlying graph, we examine several network models and discuss which are the most appropriate graph mining tools (like community detection and ranking and centrality measures) that can be applied on them. Lastly, we attempt to highlight the many advantages of using graph-based data mining in educational data against the classical evaluation techniques.*

Key words: *Data mining, Graph mining, Educational evaluation, Complex networks, Mathematical modelling*

INTRODUCTION

In recent years, handling large data sets to extract information by system modelling and applying complex networks and graph mining techniques has become more popular. The large amount of data available allows us to study large-scale systems that appear in a wide range of fields from biology to economics and the social sciences. Often these complex systems can be represented by graphs or networks, where the vertices or nodes, stand for individuals or entities, while the edges or links represent the interaction between pairs of these individuals (for an excellent review on complex networks, see, for instance [1]). Research on mining graph (and network) data has been steadily growing over the past few years, and it has become the most promising way forward for extracting knowledge from relational data [2]. The complex network approach is not only useful for simplifying and visualizing this huge amount of data, but it is also effective for picking out the key elements of the system and finding their most important interactions. Besides this, many effective tools have been developed to explore the deeper and refined topological characteristics of networks, like the community structure [3], core-periphery structure [4] or small-world property [5] and scale-free property [6]. Moreover, ranking individuals and finding out how important or "central" they are, based on their role and location in the network. have also become a useful direction of study, since it is now widely accepted that the network encodes more information about an entity than the simple descriptor(s) of this entity itself.

Educational Data Mining [7] is concerned with the development, research and application of computerized methods to find patterns and features in large collections of educational data, features that would be hard to analyse due to the huge amount of information available and the high-level complexity of such databases. Data of interest is not restricted to interactions of individuals in an educational system (e.g., navigation behaviour, input to quizzes and interactive exercises), but might also include data from collaborating students (e.g. text chat), administrative data (like school, school district, teacher), and demographic data (like gender, age, school grades). For some discussions on educational data mining, we refer the reader to [8, 9, 10]. Databases of educational institutes, where the data is produced by complex administration systems of the institutes,

CompSysTech '15, June 26-27, 2015, Dublin, Ireland
 ISBN 978-1-4503-3357-3/15/06…$15.00
DOI: http://dx.doi.org/10.1145/2812428.2812436

contain the administration of the daily work of teachers and students, like descriptions of the lessons including the equipment and educational methods that were used, the areas of competence that have been developed, the students who participated and their marks and level, among other things. Since a large amount of detailed data has become available via administration activities, there is an opportunity to get more information about the participants of the educational system, than e.g. using classical questionnaire methods. Such relevant issues, which have long been of interest, like measuring the improvements and achievements of the students, the efficiency of the teacher's work, level of difficulty, data visualization and the detection of incidental problems of the students (like drug or alcohol abuse, crisis in the family) may be investigated and addressed using new kinds of data processing techniques. In this study, we propose several suitable network representations of certain educational data and indicate which are the most appropriate graph mining tools for analysing it and what kind of additional information can be extracted by their usage. Depending on the construction of the underlying graphs, we present four families of network models. These families are a directed network of students (which is similar to that proposed earlier in [11]), an undirected network of students based on the similarity of their marks and two bipartite networks that represent students and teachers, and students and subjects, respectively.

The paper is organised as follows. In the next section, we briefly review the relevant notions from the theory of complex networks and basic graph mining tools. After, we describe the proposed network models and discuss how the network approach and graph mining techniques can be applied on them. Then, we talk about the advantages of using of complex network for examining data and mention some directions for future work.

GRAPH MINING

Basic definitions

Informally, a graph is a set of nodes and the pairs of nodes might be connected by edges. In many cases, data can be intuitively mapped into a graph format. For example, the road network of a country consists of cities (nodes) and roads (links) between each of them; social networks consist of individuals and their interconnections, which can be based on friendship or business relations, or they can be defined by a set of similarities among the individuals. The problems of generating synthetic but realistic graphs (of individuals), detecting outliers in and characteristic features of the graph have received much attention recently. Below, we attempt to describe these steps to construct a model and explain how it can be done and we will show in a concrete example, taken from the sphere of public education, how graph mining can be succesfully applied in information extraction.

Formally, the pair $G = (V, E)$ is a graph or network, where $V = (1,2, \dots, n)$ denotes the set of nodes and $E \subseteq V \times V$ stands for the set of edges. The graph is directed if the edges have a direction (i.e. if the elements of E are ordered pairs) and undirected otherwise. If a function $w: E \to \mathbb{R}$ is given that assigns a real number w_{ij} to each (i, j) edge, then the graph is weighted. The adjacency matrix of G is the $n \times n$ matrix $A = [a_{i,j}]_{i,j}$ with entries $a_{ij} = 1$, if (i, j) an edge ($i \to j$ directed edge from i to j), and $a_{ij} = 0$ otherwise. The *degree* of node i is $d_i = \sum_{j=1}^{n} a_{ij}$, which is the number of links that are connected to it. If the graph is directed, we can distinguish the in-degree and out-degree of a node, these being the number of incoming links to it and outgoing links from it, respectively.

Similarities, local properties and ranking

The problem of assigning scores to a set of individuals based on their bilateral relationships appears in many areas. For instance in sports, players or teams are ranked

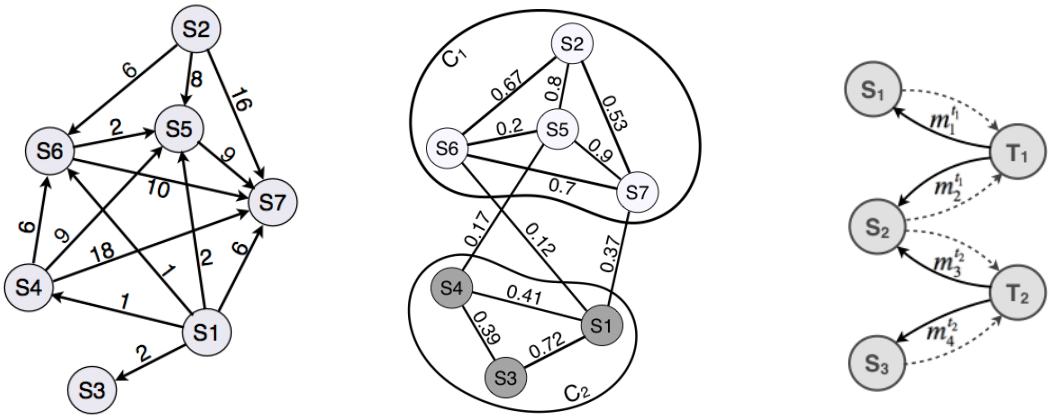

Figure 1. Toy examples for the network models. 1st: directed weighted graph of the students. 2nd: similarity-based weighted graph of the students with communities 3rd: bipartite graph of students and teachers.

according to the results of their games; the impact of scientific publications can be measured using their citation relations; Web search engines rank websites based on their (hyper-) linking structure, importance and centrality; or the special characteristics of individuals can also be evaluated according to their social relations. In real social networks, the nodes usually represent people and links might represent friendship or some other social relation between them. However, one can construct artificial social networks where the links between individuals provide some information about their similarity or some comparison based on information about them. As a concrete example, in educational institutes, pairwise comparisons can be made between students based on the lists of their marks. Depending on the marks that are considered (for example, marks of the common subjects, marks of the subjects taught by the same teachers, end-of-year reports or time series of the marks), several network representations of the students can be defined. Once we have the network model, we can apply graph mining techniques to analyse it with refined methods. First, one can examine the global characteristics of the network, such as degree distribution, topological properties (like community structure, existence of a core and periphery) and network density (i.e. the fraction of links and nodes). Measuring these properties usually provides useful information about the modelled system as a whole. Second, other information about the individuals in the system can be determined by investigating the structure, and also the dynamics of the network. This information offers us a relative picture of the actors in the system and hence provides more detailed statistics about them and comparisons between them. As a concrete example, in [11] the authors examined a network of students in a secondary school, which was constructed based on the students' end-of-year reports and proposed a PageRank-based [12] evaluation and ranking of them. They defined a directed and weighted network of the students, where a link between two students represented how much better one student was than the other (a toy example of the construction of the network can be seen in Fig. 1 first pic.). They mentioned that the network approach provides the possibility of finding talented students and filtering out the relatively "easy" subjects.

Communities in networks

Among the features of complex networks, a common one is the community structure [3,13]. In practice, community detection in a graph is a partition of the nodes into sets, such that nodes in the same community are more densely connected to each other than to the rest of the graph. Generally speaking, the communities in a network reflect the similarity and common features of the nodes that they contain. For instance, in social networks communities refer to the common location, interest or job of the actors in them. In the past

decade, several algorithms have been designed and proved to be very efficient for finding an acceptable partition of the nodes (for a comprehensive review on community detection in graphs, see [13]). One common algorithm used for community detection is the *modularity maximization method* [14]. Roughly speaking, modularity measures how strong a community structure is in a graph, compared to a random graph with the same degree distribution. The Newman modularity M of a graph [15] can be computed as

$$M = \frac{1}{2m}\sum_{i<j}(a_{ij} - \frac{d_i d_j}{2m})\delta(C_i, C_j), \quad (1)$$

where m is the number of links, C_i denotes the community where node i belongs and $\delta(C_i, C_j) = 1$ if i and j belong to the same community; and 0 otherwise. The value of M lies in the range $[-1/2,1]$, and if it is positive, then the number of edges within the communities is higher than expected if the links were randomly rewired. The concept can be readily extended to weighted networks using the w_{ij} weight value and the weighted degree of each node instead of the a_{ij} value and simple node degree.

PROPOSED GRAPH MODELS OF EDUCATIONAL DATA

Directed graphs based on the marks of the students

The first network model of the students is a generalization of the one defined in [11]. In this model, each node represents a student and a link between two students is defined in the following way. We assume that two student can be compared directly if they received an end-of-year mark in at least one common subject. If the end-of-year mark of the students i and j are $(m_1^1, \dots, m_t^1)$ and $(m_1^2, \dots, m_t^2)$, respectively, then we can calculate the weight $w_{ij} = \sum_{i=1}^{t} c_i(\, m_i^1 - \, m_i^2)$, and add a directed edge with weight w_{ij} between nodes i and j. The link goes from j to i, if $w_{ij} > 0$, and it goes in the opposite direction if $w_{ij} < 0$. The constant term c_i refers to the level of difficulty of a subject, which can also be measured by a network-based approach (see below) or applying statistical methods. In a short concrete example, suppose Anne and Bob received the end-year marks $(4,5,5,5,5)$ and $(5,3,3,3,4)$ for Mathematics, Literature, History, English and Art, respectively. Then $w_{AB} = 6$ with $c_i \equiv 1$ means that Anne is 6 points better than Bob, if the subject difficulty is thought to be the same (see Fig 1. *first pic.*). One possible way of determining the subject difficulty values is to use the average of the end-of-year marks of each subject and assume that the higher the average, the less difficult the subject is. By using the cumulative distribution of the marks, one can define an alternative way for calculating the c_i values by comparing these distributions. It is also possible to find out how difficult it is to a get a certain mark from a teacher and incorporate this parameter into the formula that calculates the edge weights.

Undirected graphs based on similarities of the marks of the students

The second network model is a family of undirected and weighted networks. As before, the nodes represent students, while a weighted edge between two students is defined by a similarity measure S of the lists containing the end-of-year marks of their common subjects (that were not necessarily taught by the same teachers). For example, the Jaccard similarity measure [16] is defined as the fraction of the marks that are the same as all the marks in common for two students (a toy example can be seen in Fig. 2, *second pic.*). One may use several similarity functions to define the weight of similarity of two students, such as the Cosine [17], Hamming [18] or Adamic-Adar [19] similarity measures. Figure 2 shows the community structure of the network of 255 students in their tenth year in a Hungarian secondary school. The weights were defined by the Jaccard similarity measure. We observed in our preliminary studies that the network contained two main communities of

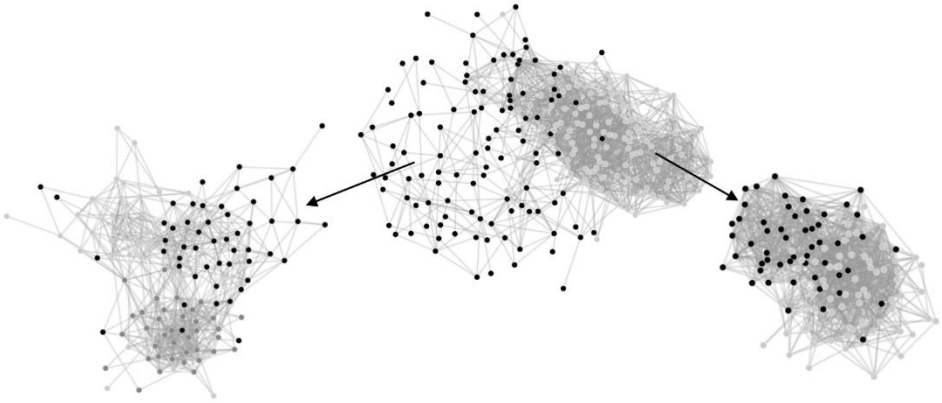

Figure 2. Community structure of a network of students (middle). The two subnetworks (left and right) induced by the two communities were re-clustered to get a more detailed structure of the network.

students who performed well in the school (Fig 2, middle, grey community) and students with a weaker academic performance school (Fig 2, middle, black community), respectively. We also found that the network had a more refined structure by re-clustering the two main communities, and we identified clusters of students who were better in the natural sciences and students who were better in the arts, respectively. We should add that these studies were not too detailed scientifically, but such investigations could be the subject of a future study.

Bipartite graphs of students and teachers

In order to evaluate how difficult it is to get a good mark from a certain teacher, we propose a family of bipartite graphs as network models based on the earlier results of [20] and [21]. We consider a *bipartite graph,* $G = (X \cup Y, E)$, whose vertices can be divided into two disjoint sets X and Y such that each edge in E connects a vertex in X to one in Y; or equivalently, there is no edge between two vertices in the same set. In the model, the elements of X are students from the same school, while the set Y stands for their teachers. We define a directed edge from a node $y \in Y$ to a node $x \in X$ with weight m, if the teacher who is represented by y gave an end-of-year mark m to student who is represented by x. However, we also define a directed edge from x to y, based on the assumption that it is more difficult to get a good mark from this teacher if the mark he or she gave is lower than the average of the student's marks (a toy example can be seen in Fig. 2, *third pic*). Next, we can easily construct a weighted directed graph of the teachers using the same technique as that described in [21] (since a discussion of the technical details is beyond the scope of this communication, we refer the reader to [20, 21]). With this "projection", a network of the teachers can be constructed, where a directed and weighted link from a teacher y_i to another teacher y_j shows how much more "consistent" a teacher is than the other, where consistency is measured via the average difference of the marks that the teacher gave to each of his or her students and the average of the student mark. Once this network is given, we can apply the PageRank method on it in order to assign scores to the teachers. These scores may provide a realistic evaluation of the consistency of their marking habits; moreover, these scores can be used to compare students by normalizing their marks using this evaluation of the teachers.

Bipartite graphs of students and subjects

Similar to the evaluation of the teachers, we can also evaluate how difficult it is to a get a good mark in a certain subject. For this purpose, we consider a bipartite graph of students and subjects, i.e. we simply substitute the set of teachers Y (defined in the previous section) by the set of subjects Z. A directed and weighted link from a subject (say Maths) to a node

$x \in X$ (which represents the student x) is defined with the weight m if the student x got the end-of-year mark m. Then, from the student a weighted link to the subject Maths is defined, where the weight represents the difference between mark m and the average of the student's marks. Without going into the technical details (which can be found in [21]), a network of the subjects can be defined and by using some evaluation technique (e.g. PageRank), a ranking of the subjects according to their level of difficulty can be obtained. These scores can be used as weights for the calculation of the students' performance and also for the evaluation of the teachers.

DISCUSSION AND FUTURE WORK

In the past decade, graph-based algorithms and data mining have been applied efficiently in a variety of areas. Following the usual methodology, in this study we proposed four different suitable network representations of students, teachers and subjects in public education and presented some possible ways of how graph mining techniques could be used to get detailed information about them. First, we defined a directed and weighted network of students, and pointed out that using graph mining methods, a more detailed picture of the achievements and ranking of the students may be obtained, instead of performing a simple statistical analysis. Then, we defined an undirected weighted graph of the students by supposing that two students are more "close" to each other if their marks are similar. Using community detection algorithms on this network, we can divide the students into groups where the groups may encode important information about them. Lastly, we defined two bipartite networks of students and teachers, and students and subjects, respectively, which could be used to measure how difficult it is to get a good mark from a certain teacher or in a certain subject.

One of the most important questions that arises is how we can efficiently use the network concepts described above. Based on graph-mining techniques, the achievements and ranking of the students can not only be analysed and determined by simple statistics, but the pairwise comparisons and the complex network representation of them also provide a more detailed and quantified picture about the real relations among them. As a by-product, these methods are able to find those students who are outstandingly better or weaker than their schoolmates. Furthermore, the common drawback of the standard statistical methods (i.e. they are not sensitive to the evaluation habits of the teachers and the real level of difficulty of the subjects) can be eliminated by using refined data mining techniques.

Several ideas and potential direction appears by using pairwise comparison methods. One of them is to categorize the subjects (e.g. as natural sciences, arts and languages) and perform separately a network-based analysis on them. By evaluating the students according to these networks, one can develop the type of learning groups (or classes) that contains at least one outstandingly better student who can help his or her classmates in the corresponding topic or subject. These students can help their weaker classmates in a cooperative learning environment. Most probably, the impact of this kind of classification can be improved if each student in a group is talented (or relatively good) in at least one subject category. Here, it should be added, that an evaluation of the learning groups should be made according to the weaker students' achievements, not by using different competence assessments. Categorizing the subject in an appropriate way can be performed based on the communities determined by a similarity-based network of the students. A crucial property of many community detection algorithms is that the number of communities can be tuned before looking for the communities. Generally speaking, using the mark-based similarity measures, the constructed network will have a few large communities with students that are generally good at science subjects and humanities, and students who are generally weaker than the others. However, creating too many groups is usually not possible, hence fine

tuning the number of communities and examining some other viewpoints (described above) might be appropriate.

Besides the possibilities of the future, technically it is already possible to include and test data-mining tools in the administration software systems of educational institutes. Using new methods to visualize and evaluate the complex system of students, the teachers can continuously monitor the achievements and relative performance of each of them. The time series of the students' marks, missing, test points, etc. are all computerized now, and plenty of information is contained in the databases; and most of the attributes and parameters are quantifiable. We think the data mining and the network approach could provide a better understanding of the educational systems and it could be a common tool for evaluation, decision making and planning in educational institutes.

Acknowledgement. This work is supported by the European Union and cofunded by the European Social Fund. Project title: The denouement of talent in favour of the excellence of the University of Szeged. Project number: TAMOP-4.2.2.B-15/1/KONV-2015-0006.

REFERENCES

[1] M. E. J. Newman, "The structure and function of complex networks", *SIAM Review* vol. 45, no.2, pp. 167-256, 2003

[2] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery in database", *AI magazine* vol. 17 no. 3 pp. 37, 1996

[3] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks, *Physical review E*, vol. 69, no. 2, 026113, 2004

[4] P. Csermely, A. London, L-Y. Wu and B. Uzzi, "Structure and dynamics of core/periphery networks", *Journal of Complex Networks*, vol. 1, no. 2, pp. 93–123, 2013

[5] D.J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks", *Nature*, vol. 393, no. 6684, pp. 440–442, 1998

[6] A.L. Barabási, "Scale-free networks: a decade and beyond", *Science*, vol. 25 no. 5939, pp. 412, 2009

[7] C. Romero, S. Ventura, M. Pechenizkiy and R. Baker, Ryan, *Handbook of educational data mining*, CRC Press, 2011.

[8] C. Heiner, N. Heffernan, and T. Barnes, "Educational data mining", In *Supplementary of Proceedings of the 12th International Conference of Artificial Intelligence in Education,* 2007

[9] C. Romero and S. Ventura, "Educational data mining: A survey from1995 to 2005", *Expert systems with applications*, vol. 33, no.1, pp. 135–146, 2007

[10] O. Scheuer and B.M. McLaren, "Educational data mining", In *Encyclopedia of the Sciences of Learning*, pp 1075–1079, Springer, 2012

[11] A. London and T. Németh, "Student evaluation by graph based data mining of administrational systems of education", In *Proceedings of the 15th International Conference on Computer Systems and Technologies*, pp. 363-369, 2014

[12] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", *Computer networks and ISDN systems*, vol. 30, no. 1, pp. 107-117, 1998

[13] S. Fortunato, "Community detection in graphs" *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010

[14] M.E.J. Newman, "Fast algorithm for detecting community structure in networks", *Physical review E*, vol. 69, no. 6, 066133, 2004

[15] M.E.J Newman, "Modularity and community structure in networks", *Proceedings of the National Academy of* Sciences, vol. 103, no. 23, pp. 8577–8582, 2006

[16] P. Jaccard, "Etude comparative de la distribution florale dans une portion des Alpes et du Jura", *Bulletin de la Societé Vaudoise des Sciences Naturelles,* vol. 37, pp. 547, 1901

[17] G. Salton, A. Singhal, M. Mitra and C. Buckley, "Automatic text structuring and summarization", *Information Processing & Management,* vol. 33, no. 2, pp. 193-207, 1997

[18] R. W. Hamming, "Error detecting and error correcting codes", *Bell System technical journal*, vol. 29, no. 2, pp. 147-160, 1950

[19] L. A. Adamic and E. Adar, "Friends and neighbors on the Web" *Social networks*, vol. 25, no. 3, pp. 211-230, 2003

[20] H. Deng, M.R. Lyu and I. King, "A generalized Co-HITS algorithm and its application to bipartite graphs. In *Proceedings of the 15th ACM SIGKDD International conference on Knowledge discovery and data mining, pp. 239-248*, 2009

[21] A. London and T. Csendes, "HITS based network algorithm for evaluating the professional skills of wine tasters", In *Proceedings of the International Symposium on Applied Computational Intelligence and Informatics*, pp. 197-200, 2013

ABOUT THE AUTHORS *

András London, PhD student, Institute of Informatics, University of Szeged,
phone: +36 62 343444, email: london@inf.u-szeged.hu
web: http://www.inf.u-szeged.hu/~london/

Áron Pelyhe, Research assistant, Institute of Informatics, University of Szeged,
phone: +36 62 343444, email: pelyhe@inf.u-szeged.hu

Csaba Holló, Assistant professor, Institute of Informatics, University of Szeged,
phone: +36 62 544130, email: chollo@inf.u-szeged.hu
web: http://www.inf.u-szeged.hu/~chollo/

Tamás Németh, Assistant professor, Institute of Informatics, University of Szeged,
phone: +36 62 343435, email: tnemeth@inf.u-szeged.hu
web: http://www.inf.u-szeged.hu/~tnemeth/