Socio-semantic Networks of Research Publications in the Learning Analytics Community

Soude Fazeli, Hendrik Drachsler, Peter Sloep

Open University of the Netherlands (OUNL)
Centre for Learning Sciences and Technologies (CELSTEC)
6401 DL Heerlen, The Netherlands
0031-(0)45-576-2218
{soude.fazeli,hendrik.drachsler,peter.sloep}@ou.nl

ABSTRACT

In this paper, we present network visualizations and an analysis of publications data from the LAK (Learning Analytics and Knowledge) in 2011 and 2012, and the special edition on Learning and Knowledge Analytics in Journal of Educational Technology and Society (JETS) in 2012.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Information filtering; K.3.m [computers and education]: Miscellaneous

General Terms

Algorithm, visualizations

Keywords

Network, recommender, visualization, dataset, learning analytics, degree

1. Introduction

The Society for Learning Analytics Research (SOLAR)[1] provided a dataset to solicit contributions to the LAK data challenge[2] sponsored by the FP7 European Project LinkedUp[3]. The dataset contains research publications in learning analytics and educational data mining for the years 2010, 2011, and 2012 (Taibi & Dietze, 2013). An overview of the dataset is shown in Figure 1. The dataset contains in total, 173 authors and 76 papers from the LAK (Learning Analytics and Knowledge) conference series in 2011 and 2012, and the special edition on learning and knowledge analytics in the Journal of Educational Technology and Society (JETS) in 2012. We found 24 authors who contributed to all three scientific proceedings.

Having access to a dataset always offers new opportunities, particularly in the educational domain, that lacks public datasets for running experimental studies (Verbert, Drachsler, Manouselis, Wolpers, Vuorikari, & Duval, 2011). Therefore, we used this dataset to present visualization of the authors and papers network, and to carry out a deeper analysis of the generated networks. Our overall aim is to use such a graph of authors and papers to recommend similar items to a target user. In the following sections, we evaluate the suitability of the LAK dataset for this purpose.

2. Motivation

It is often difficult for conference attendees to decide which workshops or sessions are suitable and relevant for them. Therefore, a list of recommended authors and papers based on shared interests could be supportive to plan the conference participation more efficiently and effectively. There already exist several papers published regarding awareness support for researchers (Reinhardt et al., 2012; Fisichella et al., 2010; Ochoa et al., 2009; Henry et al., 2009) and scientific recommender systems (Huang et al., 2002; Wang & Blei, 2010) but none of them has analyzed the Learning Analytics datasets for this purpose yet.

Our overall vision is to support the LAK attendees with a list of LAK authors and papers that are relevant for their own research interests. Such a recommendation could be created based on one or more of their own research papers but also on a short essay or even a tag cloud summarizing the research interest and objectives.

Such a priority list can support the awareness of the attendees and empower the network of like-minded authors in the attendees' particular research focus.

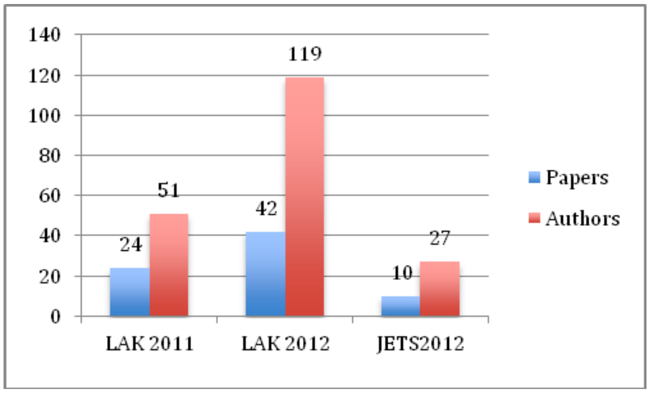


Figure 3. The used datasets

In this paper, then, we aim to explore and identify like-minded authors within the LAK dataset. Supposing that we have a network of all the LAK authors and papers, the main research questions are:

RQ1. How are the authors connected and which authors share more connections and are more central in terms of sharing commonalities with the others?

RQ2. How are the papers connected to each other in terms of similarity?

To answer these questions, we went through two main steps in our analysis: 1. Finding patterns of similarity between authors and

[1] http://www.solaresearch.org/

[2] http://www.solaresearch.org/events/lak/lak-data-challenge/

[3] http://linkedup-project.eu/

papers, 2. Visualizing networks of the LAK authors and papers. We will now describe each step in detail.

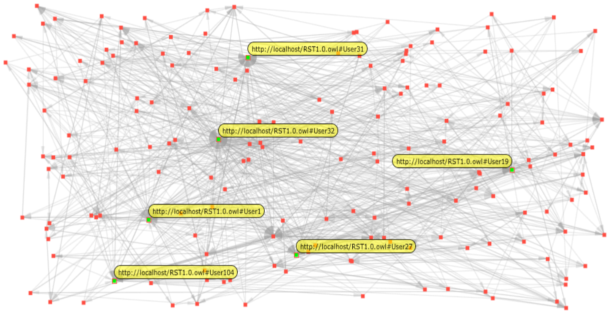

http://localhost/RST1.0.owl#User31
http://localhost/RST1.0.owl#User32
http://localhost/RST1.0.owl#User19
http://localhost/RST1.0.owl#User1
http://localhost/RST1.0.owl#User27
http://localhost/RST1.0.owl#User104

Figure 2. The LAK authors' network

(The Appendix shows a larger version)

3. Data processing

To find relationships between authors, we first computed the similarity of the papers with the TF-IDF[4] algorithm. TF-IDF can create a weighted list of the most commonly used terms in research articles. To generate the TF-IDF matrix for the LAK dataset, we first converted the LAK data from RDF to text files, which is an accepted format for the Mahout[5] system. Then, we ran the default TF-IDF algorithm provided by Mahout on the text files. We removed the stop words by setting the configuration variables within Mahout to 90%. Thus, if a word appears in 90% of the document, it is considered as a stop word (e.g. *and, or, the,* etc.) and is removed from the similarity matrix. As a final outcome we had:

- A so-called dictionary of all the terms in the LAK dataset
- A binary sequence file that includes the TF-IDF weighted vectors

For computing similarity between the LAK authors, we used the T-index algorithm (Fazeli, Zarghami, Dokoohaki, & Matskin, 2010) as a collaborative filtering recommender algorithm that generates a graph of users. In it the nodes are users and the edges show the relationship between users that originates from similarity of user profiles. The T-index algorithm originally makes recommendations based on the ratings data of users. We extended the T-index algorithm to be able to process tags and keywords extracted from the linked data e.g. RDF files. We used Jena[6] APIs to process RDF files and to handle Ontology Web Language (OWL) files that describe the generated graph of authors and papers. Jena helps to develop semantic Web application and tools.

4. Data visualization

We visualized the generated graphs of authors and papers with the Welkin[7] tool. Welkin takes an OWL file as input and provides visualization of the data as output. We present visualizations of the LAK authors and the LAK papers generated by Welkin in the following sub sections.

4.1. The LAK authors network

Figure 2 presents a network of the LAK authors in which red nodes represent the authors and the edges show the similarity between the publications of two authors. The result shows how the LAK authors are connected in terms of their publications' commonalities. Moreover, the network shows the users who share more commonalities than do other authors. We call them 'central authors'. In the next section, we show how they are connected with the other authors in the network.

4.2. The LAK authors' degree centrality

For some node in the network, the degree centrality shows the total number of incoming and outgoing edges. It is a metric commonly used for Social Network Analysis (SNA) (De Liddo, Buckingham Shum, Quinto, Bachler, & Cannavacciuolo, 2011; Gu´eret, Groth, Stadler, & Lehmann, 2012; Opsahl, Agneessens, & Skvoretz, 2010). In other words, the degree of a node describes how many other nodes are connected to the target node. In fact, it helps to measure how many hubs are in the network. We describe hubs as the nodes that have the most connections to the others in the network. The degree centrality metric may be used to strengthen a network by providing its nodes with more connections. In this data study, degree centrality is used to measure the relevance of an author's papers to the other authors in the network.

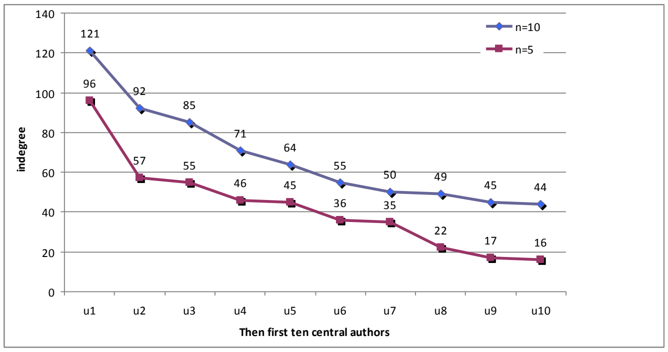


Figure 3. The degree centrality of the top ten central authors

Figure 3 shows the degree centrality for the first ten authors with the highest similarity degree with respect to the LAK publications. The horizontal axis (x) shows the top ten central users, e.g. *u1* is the author whose paper(s) has the highest degree. The vertical axis (y) shows the degree values that describe the number of relationships of a each user shown in the x-axis. Figure 3 also shows degree centrality for two different sizes of nearest neighborhoods (*n*). Such neighborhoods are commonly used in collaborative filtering recommender algorithms. By increasing the neighborhood size *n*, the degree of the authors increases accordingly. As a result, we will have a larger number of central authors when *n* is higher (e.g. *n*=10). As can be seen in Figure 3, degree for the first central author (*u1*) is equal to 121 if n=10 and 97 if n=5. These high scores show the high relevancy of *u1*'s publications to the authors. As a consequence, *u1* will appear in the top-n authors recommendations more often than the other authors.

[4] http://en.wikipedia.org/wiki/Tf–idf

[5] http://mahout.apache.org/

[6] http://jena.apache.org/

[7] http://simile.mit.edu/welkin/

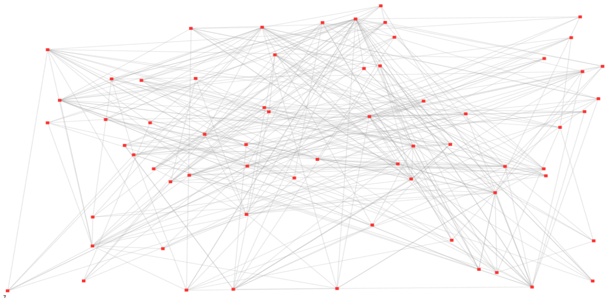


Figure 4. The LAK papers network

(The Appendix shows a larger version)

4.3. The LAK papers network

Figure 4 shows a network of the LAK papers. The red nodes are papers and the edges between them represent the similarity of the papers. By finding similar papers, we can recommend the most similar papers to specific authors. This increases the awareness of the authors about papers which are relevant to them and published in their communities.

Figure 4 shows that, some of the papers share more similarity with the others and own a higher degree number. As with the central authors, these papers *will appear more often in the top recommendation list than the other papers of the dataset*. One may interpret their degree as their popularity. Therefore, the papers with higher degree values are more popular and, presumably, they are more of interests to users. For the publication data, interests of users derives from the words and terms they have used more frequently in their papers.

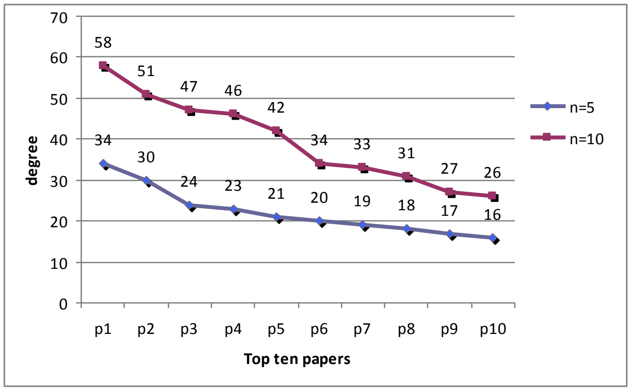


Figure 5. The degree centrality of Top ten papers

4.4. The LAK papers' degree centrality

Figure 5 shows the degree centrality for the first ten papers that are most similar to the other papers. We selected the first ten top papers with the highest degrees. The horizontal axis (x) shows the top ten papers e.g. *p1* is the paper with the highest similarity and thus, the highest degree value among the others shown by the vertical axis (y). Figure 5 shows degree centrality for two different sizes of nearest neighborhoods (*n*), 5 and 10. By increasing the *n*, the degree of the papers increases accordingly. As a result, we will have a larger number of top papers if *n* is higher (here, when *n*=10). In Figure 5, the degree for the first top paper (*p1*) is equal to 53 (n=10) and 29 (n=5). This shows how much *p1* shares similarity with other papers. As a consequence, *p1* can be considered as the most popular paper and it has the highest chance to appear in the top paper recommendations.

5. Discussion and conclusions

The results presented here, allow us to answer our research questions in the following way:

RQ1. How are the authors connected? Which authors share more connections and are more central in terms of sharing commonalities with the others?

We presented a visualization of the authors' network to provide an overview of how they are connected to each other. To justify the authors' connections and relationships, we evaluated the degree centrality for the first ten, most central authors. Table 1 presents the first ten central authors and their degree to show the authors with the highest relevancy of their publications with others in the network. Table 1 shows the degree of the authors for sizes of neighborhoods equal to 10.

Table 1. The first ten central authors

Author	Degree
Hendrik Drachsler	116
Kon Shing Kenneth Chung	87
Wolfgang Greller	80
Javier Melenchon	66
Brandon White	59
Vania Dimitrova	50
Erik Duval	45
Rebecca Ferguson	44
Anna Lea Dyckhoff	40
Simon Buckingham Shum	39

RQ2. How are the papers connected to each other in terms of similarity?

We presented degree centrality of the LAK papers to give insight in their relationships in the papers' visualized network. We selected the top ten papers that have the highest similarity with the other papers. To show which papers are placed in the top ten papers' list, we present the title and authors for each paper.

The top ten papers are not necessarily by the authors who are identified as the central authors. Although most of the central authors also appear in top ten papers' list (see Table 2), the order is not the same. As we investigated the LAK data, we found out that some of the central authors have more than one paper. For instance, Hendrik Drachsler has contributed to four papers. In this study, similarity is calculated based on all papers of an author. So, it is quite probable that not each and every one of the authors' papers individually has the highest similarity to the other papers. Although some of the central authors are common to the two tables, only one of the papers authored by those central authors appears in the top ten papers list shown by Table 2.

Table 2. The Top ten papers

Paper	**Authors**
Learning Dispositions and Transferable Competencies: Pedagogy, Modelling and Learning Analytics	Simon Buckingham-Shum, Ruth Deakin Crick
The Pulse of Learning Analytics Understandings and Expectations from the Stakeholders	Hendrik Drachsler, Wolfgang Greller
Social Learning Analytics: Five Approaches	Rebecca Ferguson, Simon Buckingham-Shum
Multi-mediated Community Structure in a Socio-Technical Network	Dan Suthers, Kar Hai Chu
Modelling Learning & Performance: A Social Networks Perspective	Walter Christian Paredes, Kon Shing Kenneth Chung
Teaching Analytics: A Clustering and Triangulation Study of Digital Library User Data	Beijie Xu, Mimi M Recker
Monitoring Student Progress Through Their Written "Point of Originality"	Johann Ari Larusson, Brandon White
Learning Designs and Learning Analytics	Lori Lockyer, Shane Dawson
A Multidimensional Analysis Tool for Visualizing Online Interactions	Eunchul Lee, M'hammed Abdous
Using computational methods to discover student science conceptions in interview data	Bruce Sherin

Overall, we found that the LAK dataset can help conference attendees to become more aware of their research network, which, in its turn, is useful for sharing knowledge and experiences. However, the current dataset contains no user feedback or evaluations to evaluate either an author or a paper recommender system in terms of common metrics such as prediction accuracy and coverage of the generated recommendations. For future analysis it would be helpful if the LAK dataset also contains references to the papers. The references could be used to identify the top cited authors and papers within the LAK dataset and beyond. As a further step, we are planning to try additional social network analysis measures besides degree, such as betweenness or closeness.

6. References

De Liddo, A., Buckingham Shum, S., Quinto, I., Bachler, M., & Cannavacciuolo, L. (2011). Discourse-centric learning analytics Conference Item. *LAK 2011: 1st International Conference on Learning Analytics & Knowledge*. Banff, Alberta.

Fazeli, S., Zarghami, A., Dokoohaki, N., & Matskin, M. (2010). Elevating Prediction Accuracy in Trust-aware Collaborative Filtering Recommenders through T-index Metric and TopTrustee lists. *JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE*, *2*(4), 300–309. doi:doi:10.4304/jetwi.2.4.300-309

Gu´eret, C., Groth, P., Stadler, C., & Lehmann, J. (2012). Assessing Linked Data Mappings using Network Measures. *Proceedings of the 9th international conference on The Semantic Web: research and applications* (pp. 87–102). Springer-Verlag Berlin, Heidelberg. doi:10.1007/978-3-642-30284-8_13

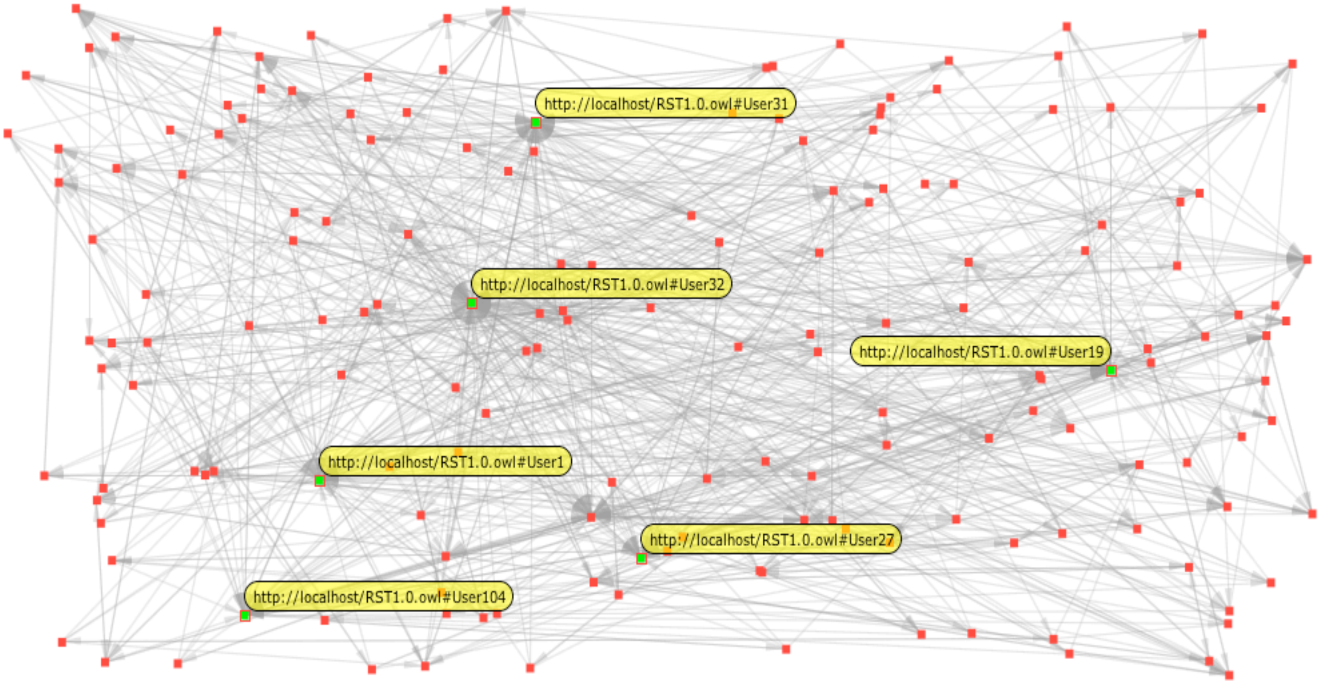
Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, *32*(3), 245–251. doi:10.1016/j.socnet.2010.03.006

Taibi, D., & Dietze, S. (2013). *Fostering analytics on learning analytics research: the LAK dataset.*

Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikari, R., & Duval, E. (2011). Dataset-driven research for improving recommender systems for learning. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (pp. 44–53). ACM, New York, NY, USA.

7. Appendix

7.1. The LAK authors' network


http://localhost/RST1.0.owl#User31
http://localhost/RST1.0.owl#User32
http://localhost/RST1.0.owl#User19
http://localhost/RST1.0.owl#User1
http://localhost/RST1.0.owl#User27
http://localhost/RST1.0.owl#User104

7.2. The LAK papers' network