Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

INSTRUCTIONS

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

Predicting the global impact of authors from the learning analytics community – A case study grounded in CNA

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Ionita, Remus	University Politehnica of Bucharest	
Dascalu, Mihai	University Politehnica of Bucharest	
Corlatescu, Dragos-Georgian	University Politehnica of Bucharest	
McNamara, Danielle	Arizona State University	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) May 2021

Check type of content being submitted and complete one of the following in the box below:

- ☐ If article: Name of journal, volume, and issue number if available
- ☑ If paper: Name of conference, date of conference, and place of conference
- ☐ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

The 23rd International Conference on Control Systems and Computer Science, May 26-28 2021, Bucharest, Ro

DOI or URL to published work (if available) 10.1109/CSCS52396.2021.00078

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** Institute of Education Scien through **[Grant number]** R305A18014 to **Institution]** Arizona State U. The opinions expressed are those of the authors and do not represent views of the **[Office name]** Institute of Education Scie or the U.S. Department of Education.

Predicting the Global Impact of Authors from the Learning Analytics Community – A Case Study grounded in CNA

Remus Florentin Ionita
Computer Science Department
University Politehnica of Bucharest
Bucharest, Romania
ionitaremusflorentin@gmail.com

Dragos-Georgian Corlatescu
Computer Science Department
University Politehnica of Bucharest
Bucharest, Romania
dragos.corlatescu@upb.ro

Mihai Dascalu
Computer Science Department
University Politehnica of Bucharest
Bucharest, Romania
mihai.dascalu@upb.ro

Danielle S. McNamara
Psychology Department
Arizona State University
Tempe, Arizona, USA
dsmcnama@asu.edu

***Abstract*—Exploring new or emerging research domains or subdomains can become overwhelming due to the magnitude of available resources and the high speed at which articles are published. As such, a tool that curates the information and underlines central entities, both authors and articles from a given research context, is highly desirable. Starting from the articles of the International Conference of Learning Analytics & Knowledge (LAK) in its first decade, this paper proposes a novel method grounded in Cohesion Network Analysis (CNA) to analyze subcommunities of authors based on the semantic similarities between authors and papers, and estimate their global impact. Paper abstracts are represented as embeddings using a fine-tuned SciBERT language model, alongside a custom trained LSA model. The extrapolation between the local LAK community to a worldwide importance was also underlined by the comparison between the rankings obtained from our method and statistics from ResearchGate. The accuracies for binary classifications in terms of high/low impact predictions were around 70% for authors, and around 80% for articles. Our method can guide researchers by providing valuable information on the interactions between the members of a knowledge community and by highlighting central local authors who may potentially have a high global impact.**

Keywords—Cohesion Network Analysis, Semantic and Co-authorship Links, Global Author and Paper Impact

I. INTRODUCTION

Researching a new domain is a difficult task due to the high volume of articles that are being published. In addition, a general ranking based on surface indicators (e.g., citation counts) is not sufficient for identifying the most relevant articles or the most influential authors. Thus, a semantic modeling of the interactions between authors and papers can provide insight in terms of informing recommendations. As such, a semantic query can provide a more personalized response focused on the user's interests, rather than an average response that may or may not fit the intent of the search.

A drawback of semantic search is the time required to retrieve results. Given a large corpus, it can take a considerably larger amount of time to compute text embeddings and similarities between texts, in contrast to search engines relying on only keywords. A potential solution to this problem is to compute the semantic relations on a smaller corpus; however, the following research question arises: to what extent features extracted from a local perspective match the global framing, when accounting for the impact of authors and of papers? We propose an approach based on state-of-the-art language models, coupled with statistical features derived from the generated graph representation of interactions between papers and authors, to create local-to-global extrapolations of their importance.

This paper introduces a semantic analysis grounded in Cohesion Network Analysis (CNA) [1] of the Learning Analytics community (LAK) in its first decade, coupled with a novel approach of computing the similarities between two articles or authors. CNA extends Social Network Analysis and builds a graph representation of both authors and articles connected using semantic links [2]. Similar to Social Network Analysis, a multitude of metrics can be computed to better understand the relationships between the entities of the graph. The method presented in the following sections uses a fine-tuned SciBERT [3], alongside a custom Latent Semantic Analysis model [LSA; 4], to compute the cosine similarity between articles. This approach was evaluated on the GLUE benchmark [5] with good results before applying it to the LAK dataset. Furthermore, the rankings of articles and authors based on the CNA graphs obtained from this method were good predictors for the worldwide impact of each entity (i.e., article or author).

Following the introduction, the paper continues by presenting state-of-the-art approaches for community analysis based on graphs representations. The third section introduces the corpus, alongside the neural architecture, where the fine-tuning processes are detailed, followed by the description of the architecture used to generate the embeddings of the articles. The results of the experiments suggest that the local metrics computed on the LAK dataset successfully predict global metrics extracted from ResearchGate. The last section provides conclusions, together with future improvements.

II. STATE OF THE ART

Understanding the ideas behind community modeling are a key aspect of this research. We believe that having the correct (or as accurate as possible) relations between the entities in a community graph is the steppingstone for further

in-depth analyses. The following articles describe different ideas or applications of community modeling and/or their real-life applications.

Cruz et al. [6] explored a method to integrate both structural and semantic information in community detection. A common practice to find a community consists of applying a custom clustering algorithm; however, the authors argue that this approach by itself can deteriorate the quality of the clusters. Thus, they first obtain a partial grouping of the nodes starting from the semantic relations between them; afterwards, a common clustering algorithm is applied. The networks were then analyzed from perspectives pertaining to different types of nodes.

Community analysis can be applied to other domains than scientist networks and can help extract meaningful insights. Heller et al. [7] performed an analysis on the open-source software community of Github starting from a dataset of 500,000 follow links, over 1 million commits, and 50,000 users. The method does not provide any conclusive answers, but rather offers a starting point for further analysis on productivity and communication in open-source projects that have contributors from different parts of the world. For visualization purposes, they used linked geo-scatter maps, small multiple displays, and matrix diagrams.

Graph theory and methods can also improve the representation of the communities and the relations between them. For example, Reda et al. [8] approached the social network analysis from a different perspective. In general, the analysis of a community is static (in a precise moment in time), but they argue that a dynamic analysis that evolves across time can provide a more insightful view of the relationships between the members of a community. As the main study material, they used the community structure (based on votes and opinions) of the House of Representatives between March 3 and 18, 2010. They showcased using visual graphs how multiple subcommunities were formed and how the dynamics between them evolved in time, depending on popular subjects from that period.

Vehlow et al. [9] approached the problem of community visualization with the same idea as the previous study, namely dynamic graphs. The authors introduce an interface where a community evolution layer can be used to observe community transitions, alongside the possibility to order communities and vertices using various criteria. The UI could be customized using colors that were carefully studied to provide a clear separation between communities and their stages. The graphs could also include interactive highlighting for better pattern recognition processes.

Machine learning can be applied to obtain a better perspective of the relations between nodes in a graph. This is known as node embeddings, and Tang et al. [10] applied this notion to build a model that supports large scale graphs, as well as small scale graphs. The first-order proximity and second-order proximity are used to ensure relevant representations between vertices in a graph. The model – Line– is trained to maximize the likelihood of predicting the neighbor vertices of a node. They argued for the efficiency of the embeddings in various tasks such as language networks, social and citation networks.

DeepWalk [11] is another approach in the area of node embeddings consisting of a model capable to recognize related vertices based on random walks – i.e., the main idea is to estimate the likelihood of sequences of vertices in a corpus. The random walks can have different lengths and the overarching goal is to explore the graph. Based on this exploration, the next step involves updating the representation of vertices. An important advantage of this approach in comparison to other methods is the ease of parallelization: the random walks can be executed separately and the model can be updated using the asynchronous version of stochastic gradient descent.

Taking it one step forward, Cavallari et al. [12] researched community embeddings starting from the node embeddings. This study introduces a mathematical approach to select the "members" of a community alongside steps required to train the model. Their experiments analyzed multiple networks, but the most relevant for this paper is DBLP – an academic paper citation network built upon the DBLP repository. Each article had one of five available labels based on the conference theme: Natural Language Processing, Computer Vision, Data Mining, Database, and Networking. Community detection was performed with promising results, concluding with a visualization of clusters in the graph.

Tu et al. [13] explored another method of representing node and community embeddings. While LINE and DeepWalk use only contextual information about vertices, this study proposes an additional feature, namely the community information to which a node belongs. The model learns both vertex and community embeddings by maximizing the log probability of predicting context vertices using both the vertex and the community to which it is assigned.

Embedding methods were studied for dynamic graphs, as well. Goyal et al. [14] describe the DynGEM model trained in a semi-supervised manner, which provides promising results in link prediction and anomaly detection. The main challenge, as exposed in the paper, was the scalability of the model. Their solution involved heuristics and the idea of reusing previously generated graphs.

III. METHOD

A. Corpora

Our dataset consists of papers published in the first decade of the International Conference on Learning Analytics & Knowledge (LAK), starting from 2011 up until 2020. The abstracts of the articles, alongside additional metadata, were freely available from the ACM Digital Library and a crawler was implemented to extract this information. The final dataset contained 557 articles written by 1194 distinct authors from 312 different institutions. A cleanup was required to remove articles belonging to "Workshop" or "Poster" sections that were not consistently present across all years. As such, 143 articles were removed from the initial dataset which contained 700 articles.

B. Building the 2-Mode CNA graph

The local impact of entities is modelled by the SNA centrality scores computed using the 2-mode CNA graph [15] that is now obtained from the embeddings given by SciBERT and LSA. The first step in building these embeddings involved fine-tuning SciBERT [3] on the Stanford Natural Language Inference - SNLI [16], Multi-Genre Natural Language Inference - MultiNLI [17] and Semantic Textual Similarity - STSb-train [18] datasets using the Siamese Networks proposed by Reimers and Gurevych [19]. Our aim was to learn semantically meaningful paragraph embeddings that can be

further used to measure the semantic similarly between them. The model was evaluated using the STSb-test, one of the General Language Understanding Evaluation (GLUE) benchmarks.

Fine-tuning was achieved by training one epoch on each of these three datasets, using the guidelines provided by Devlin et al. [20]. All layers were updated, the batch size was set at 8, the Adam optimizer had a learning rate of 2e-5, and the linear learning rate warm-up was used for 10% of the fine-tuning data. The entire fine-tuning procedure lasted about two hours on a GeForce GTX TITAN X, with CUDA 10.0 and cuDNN 7.

Figure 1 presents performance on the STSb-dev during fine-tuning. The evaluation on the STSb-dev dataset tracked the model performance during fine-tuning at each 1000 batches (8000 samples). The final major upper-jump was generated after fine-tuning on the STSb-train which, despite its quite small size (5749 samples), helped in improving considerably the overall performance.

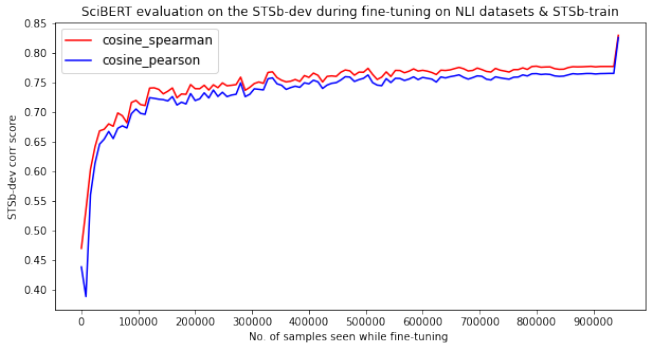


Fig. 1. SciBERT evaluation on the STSb-dev during fine-tuning.

Note that both SciBERT and BERT models fine-tuned using the same method proposed by Reimers and Gurevych [19] are available online (https://huggingface.co/gsarti/sciBERT-nli, https://github.com/UKPLab/sentence-transformers); however, the SciBERT is fine-tuned only on the NLI datasets, and both previous models have a maximum sequence length of 128 word pieces. Thus, they were not a viable option because our pipeline would ignore a considerable part of the input, as abstracts tend to have more than 128 words (M = 170.333, SD = 56.038, Min = 28, Q1 = 134, Q2 = 168, Q3 = 199, Max = 689). Moreover, fine-tuning also on the STSb-train improves model performance significantly, despite its reduced size, as reflected in the final major upper-jump from Figure 1 and previously suggested by Reimers and Gurevych [19].

The Siamese Network received as input two articles and encoded them using two BERT models with tied weights. The final document embeddings were obtained using mean pooling - the average of output vectors for BERT tokens. The maximum sequence length was chosen to be 384 after encountering multiple memory errors when using the max sequence length normally accepted by Transformers (512). Longer sequences were considerably more computationally expensive as attention mechanisms are quadratic in terms of sequence length.

Moreover, a BERT-based model was also fine-tuned on these datasets using the same architecture and configurations (including the same maximum sequence length); however, we found it more appropriate to use the fine-tuned SciBERT for encoding the meaning of the LAK articles because they are highly related to the computer science domain and SciBERT was pre-trained, as its name suggests, on scientific text, including text from the computer science domain.

Moving forward, a custom LSA model was trained using the titles and abstracts of the LAK 2011-2020 articles to obtain domain-specialized dense representations; our intuition was also to compensate for the extra words ignored by the Transformer model due to the maximum sequence length descried above (384-word pieces). The custom LSA model used Tf-Idf over 1-gram, 2-gram, and 3-gram tokens, ignoring the least and most frequent 0.01% of the terms (min_df=0.01 & max_df=0.99), followed by an SVD to reduce the Tf-Idf document embeddings to the most significant dimensions. In the end, 100 dimensions corresponding to the highest singular values were used.

The final embedding of a document represented the dense vector obtained by concatenating the document's embeddings given by the fine-tuned SciBERT model (size 768) and the domain specialized LSA model (size 100). Thus, domain-specialized knowledge was "injected" into the SciBERT dense vectors by appending the LSA embeddings. Fewer dimensions might have been selected for the projection of the SVD from the LSA model; however, the contribution of the LSA embeddings to the final article's dense representations would have been insignificant because the BERT embeddings were larger.

The cosine distance was applied using the SciBERT and LSA embeddings to quantify the semantic similarity between articles. Other similarity functions were considered (i.e., Euclidean distance, Manhattan distance, and dot product), but the evaluations on the STSb-dev showed that the cosine distance performed slightly better, as previously stated by Reimers and Gurevych [19]. Nevertheless, the similarity scores when using the LSA embeddings of size 100 alone were positively correlated with the similarity scores when using the fine-tuned SciBERT embeddings alone (Spearman correlation of .4). The scores of the final embeddings (after concatenation) were strongly correlated with the scores using the SciBERT embeddings alone (Spearman correlation is .99). Thus, the contribution of LSA was not significant, but the effects were visible and powerful for certain pairs of articles. For example, Figure 2 presents semantic similarity scores between 7 sample articles. The title of the articles is sufficiently expressive for the high similarity scores, whereas the values are further sustained by the content of the article's abstracts.

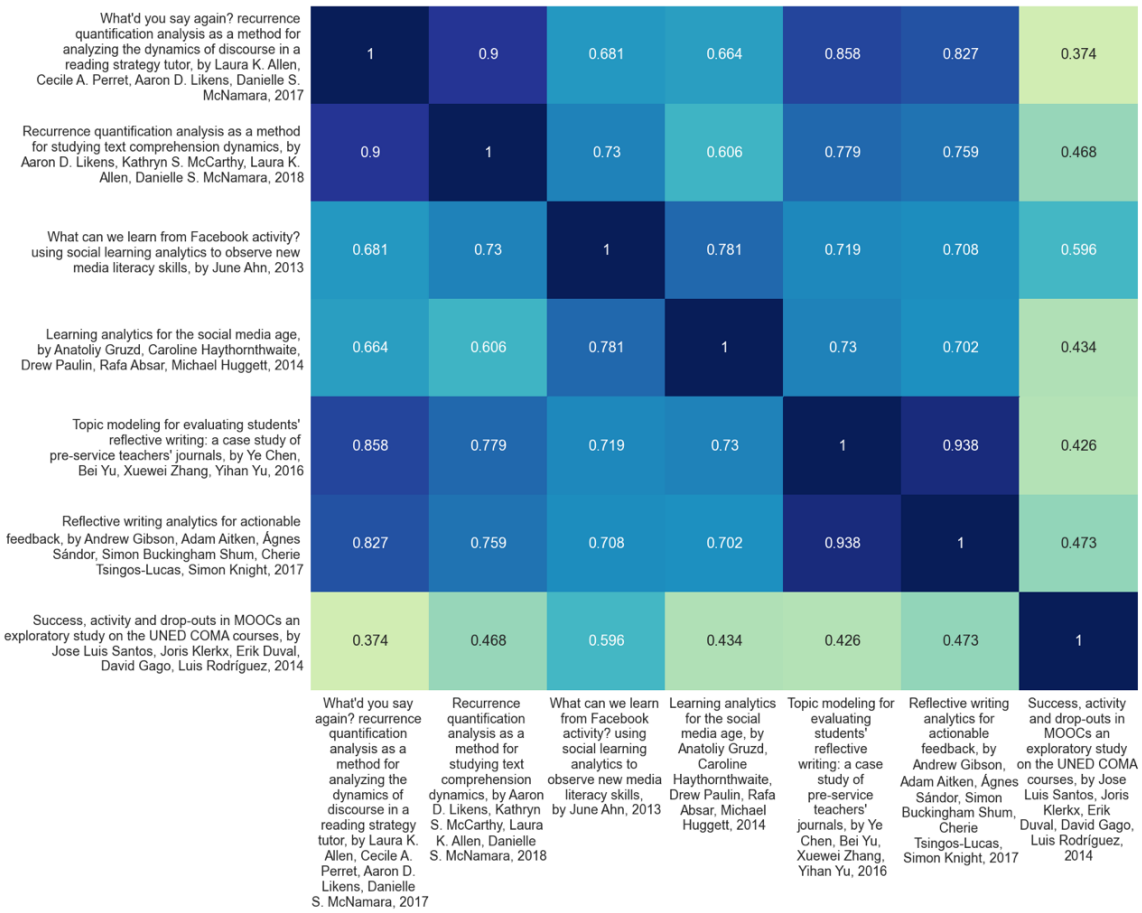


Fig. 2. Heatmap of semantic similarity scores between selected articles.

While considering the previous concatenated embeddings, a 2-mode CNA graph including both authors and papers is created. The values of the edges in the 2-mode graph are computed as follows:

- The weight of the connection between two articles is their semantic similarity;
- The score between an article and an author is represented by the mean of the similarities between the article and all of the author's articles;
- The semantic similarity between two authors is represented by the maximum coupling of the subgraph, where only the articles of the two authors are present.

C. *Predicting the global impact of authors and articles based on their local impact*

Starting from the previous 2-mode CNA graph, our objective is to argue the generalizability of SNA metrics computed locally to reflect the worldwide impact of an author or an article. SNA centrality scores consider degree, closeness, and betweenness centralities.

The authors' global scores are retrieved from Research Gate (www.researchgate.net), namely the Research Gate score ('rgScore'), the percentile of the Research Gate score ('rgScorePercentile'), h-index ('hIndex') and the h-index excluding self-citations ('hIndexExcludingSelfCitations'). The metrics retrieved from Research Gate are expressive for measuring the worldwide impact of an author.

The global impact of articles is assessed in terms of research interest, the total number of citations, the total number of reads and the number of reads in the last week, all extracted from Research Gate. Research interest quantifies the scientists' interest for the certain article (weighted average taking into account the number of reads, full text reads, citations, and recommendations). In addition, two new scores were added, namely the normalized number of reads and normalized number of citations. The normalization was a linear on the total number of reads (and citations in order to attenuate the influence of time on these metrics (see equation 1).

$$scoreN(art_i) = score(art_i)\,\frac{1}{(2020 - year(art_i) + 1)} \qquad (1)$$

While considering authors, diverse regression models were trained to predict the author's worldwide scores (metrics retrieved from Research Gate), given their local context (SNA scores computed on the co-authorship graph and on the 2-mode CNA graph, as well as the overall number of articles published in the community). Additionally, the problem was simplified and a Random Forest model for binary classification was trained to learn a global 'importance' relation between two distinct authors, starting from their local SNA metrics on both co-authorship and semantic graphs. The model receives as input the SNA metrics (on both co-authorship and semantic graphs) corresponding to two distinct authors, and predicts which author has a greater impact or is more 'important' at worldwide level. The task is a binary classification problem with the following labels: 0, if the first author has a higher Research Gate score, and 1 otherwise. All the possible pairs of two LAK authors were considered, dropping the pairs having the same target value.

Similarly, regression models were trained to predict the global scores of articles (metrics retrieved from Research Gate), given their local context (SNA metrics for articles from the 2-mode CNA graph). In addition, the binary classification experiment was conducted also for articles in order to reflect their relative importance. The same labels were applied: 0 if the first article has a higher Research Gate score, and 1 otherwise. All possible pairs of two LAK articles were considered, dropping the pairs with the same target value.

IV. Results

The evaluation of the fine-tuned SciBERT and BERT models was performed on the STSb-test GLUE benchmark. Moreover, diverse regression and classification models were trained to predict the global impact of an author or an article, based on the entity's local impact within the LAK community.

A. Evaluate the SciBERT model on the STSb dataset

The fine-tuned SciBERT and BERT versions were evaluated on the STSb-test - one of the GLUE benchmarks for NLU problems (sentence similarity). A sample represents a pair of two sentences. First, the embeddings of the two input sentences were computed, and then their semantic similarity was measured using the cosine distance. Then, the evaluation was conducted using the Spearman correlation between the computed distances and the human annotation scores included in the STSb-test dataset (scores range from zero - no relation to five - sentences are semantically equivalent). The model was fine-tuned using the method proposed by Reimers and Gurevych [19]. The uncased variants of the models were used - uncased BERT model, uncased SciBERT model.

Figure 3 presents the evaluation on the STSb for different models using different pooling strategies. If the task name (NLI, STSb) appears on the model name axis, then the corresponding model is fine-tuned on the task's specific datasets. The fine-tuned BERT model performs better than SciBERT, but note that STSb does not contain data related to the computer science or learning domains, only news, captions, and forum data. Moreover, using out of the box versions of BERT or SciBERT for computing embeddings of sentences or paragraphs seems unsuitable for tasks that further involve measuring their similarity score.

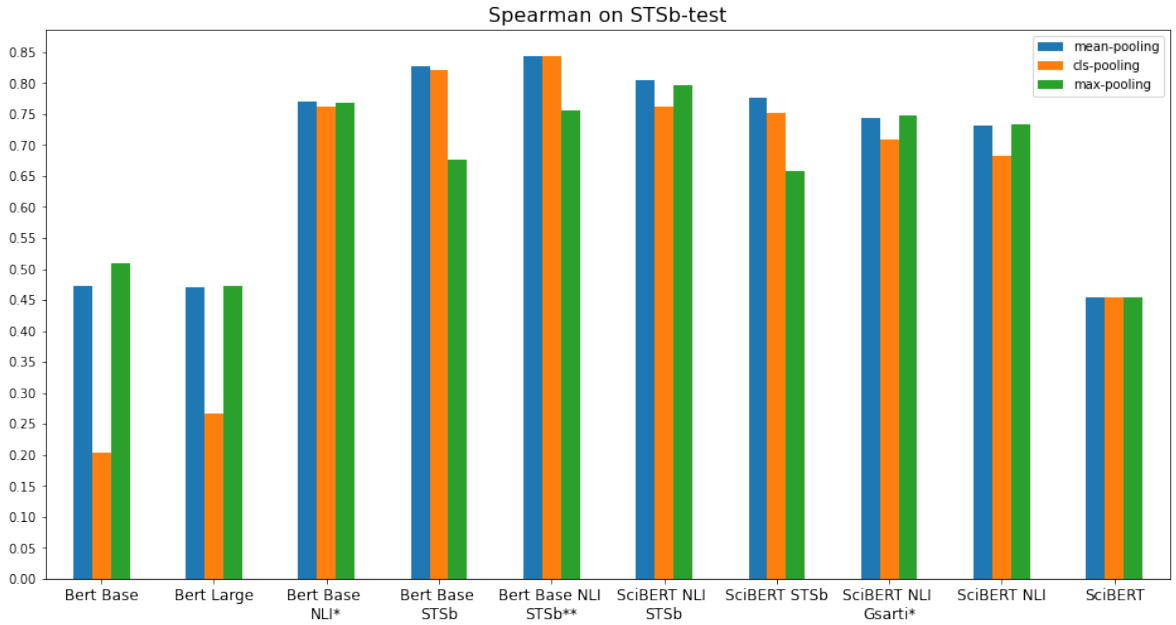


Fig. 3. Evaluation on the STS benchmark test set for different models and pooling strategies.

* public model already fine-tuned on the NLI datasets;
** public model already fine-tuned on the NLI datasets, but further fine-tuned by us on the STSb-train (https://huggingface.co/gsarti/sciBERT-nli, https://github.com/UKPLab/sentence-transformers).

B. 2-mode CNA Author Rankings

Figure 4 illustrates the Spearman correlations between local features and global author metrics from Research Gate. The correlations take into account only the authors who have published at least three articles within the LAK conference, as the SNA metrics are more expressive for these authors whose centrality ranking scores were computed starting from multiple articles. Note the positive correlations between the computed SNA scores and the metrics retrieved from Research Gate, especially the betweenness centrality scores computed on the co-authorship graph and 2-mode CNA Graph. Both measures are also positive correlated with the number of articles published by the authors in this conference. Overall, the computed semantic scores are more positively correlated with the Research Gate scores than simple co-authorship local metrics, indicating that extra knowledge or structure is inferred from the semantic links.

The best regressor for predicting global author impact after hyperparameter tuning was a Support Vector Regressor (RBF kernel, $C = 10.0$ - regularization parameter, and gamma = 1.0). The outliers were removed using the classical boxplot model, and only the samples with the target score between [Q1 - 1.5 * IQR, Q1 + 1.5 * IQR] were retained, where IQR = Q3 – Q1 (interquartile range). The model performance on the test set (MAE = 5.694, MSE = 59.378, RMSE = 7.677) is slightly better than the performance of a baseline regressor that always predicts the median value (MAE = 5.971, MSE = 65.800, RMSE = 8.112), but improvements are small. The boxplot model removed about 5% of the authors.

Figure 5.a shows the density plot relative to the target column, namely the h-index scores of the authors. A density plot is a continuous and smoothed version of the histogram that is approximated by summing Gaussian kernels at each data point. A note from this chart is that the model mostly over-estimates the lower h-index scores, but also under-estimates the higher h-index scores. Figure 5.b shows the distribution of residuals (prediction minus true score) for the same target. Ideally, it would be normally distributed, meaning that the model is wrong in the same amount in both lower and higher directions. Nevertheless, the conclusion is the same as in the previous chart (Figure 5.a), namely that the trained regressor over-evaluates the lower h-index scores (right tail), but there are a high number of under-estimation errors (left tail). Similar results are obtained using other worldwide metrics as targets.

The binary classification model of relative author importance uses the default Random Forest configuration, except of the number of trees that was set to 10. Figure 6 shows the evaluation scores on the test set using different Research Gate metrics as target. Most of the metrics are about 70%. This result argues that the local SNA scores developed for authors may be used to infer the worldwide impact or the 'importance' of an author.

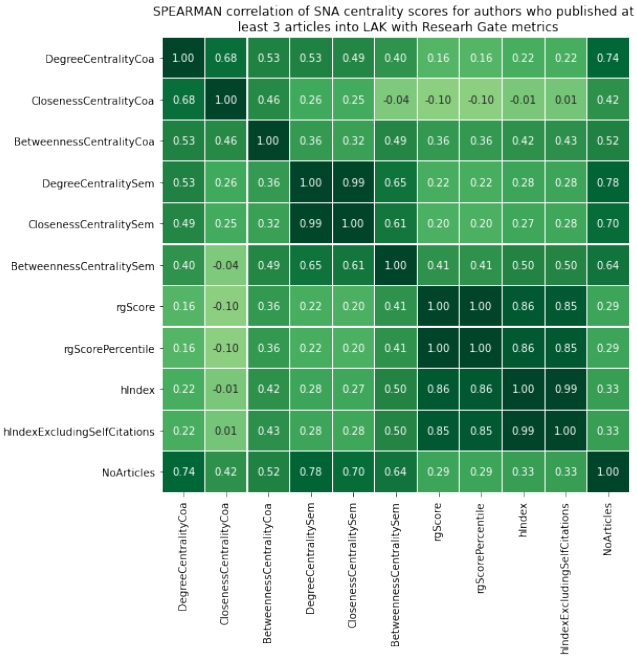
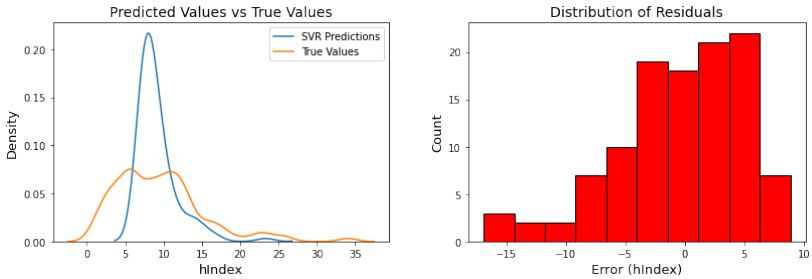


Fig. 4. Sperman correlations between SNA centrality scores and Research Gate metrics for authors (Coa – coauthorship; Sem – semantic).



(a) (b)

Fig. 5. Support Vector Regressor results for authors - (a) Density plot, (b) Distribution of Residuals.

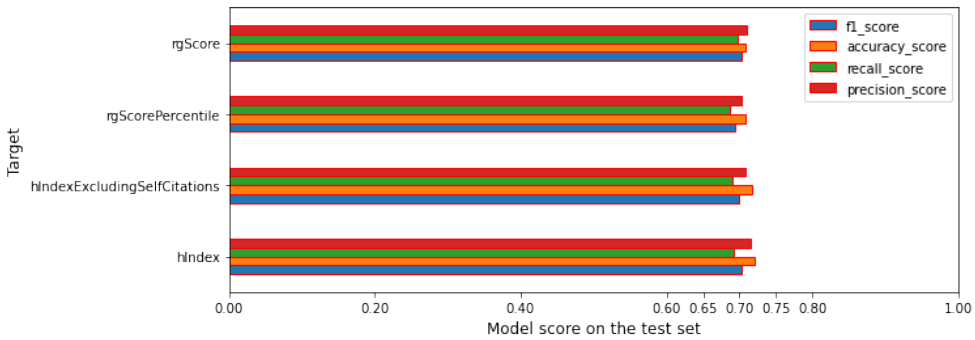


Fig. 6. Evaluation on test set using as target different Research Gate metrics for authors.

C. 2-mode CNA Articles Rankings

Figure 7 introduces the Spearman correlations between local importance scores and global article scores retrieved from Research Gate. Correlations are now lower as absolute values, and only betweenness centrality reaches correlations of .25 or above with global metrics.

When building the regressor for predicting the article's impact, the best model after hyperparameter tuning was a Support Vector Regressor (sigmoid kernel, $C = 1.0$ - regularization parameter and gamma = 100). Again, the outliers whose target values are lower than the first quartile minus one and a half interquartile range or greater than the third quartile plus one and a half interquartile range were removed. However, the regressor performances are not great (MAE=1.599, MSE=4.782, RMSE=2.155), being almost equal to the performances of a naïve model and the baseline that always predicts the target median value (MAE=1.608, MSE=4.708, RMSE= 2.1697). These scores are computed using the normalize citations values as targets, but the performance is similar when using other article's worldwide metrics. The SNA centrality scores (degree and betweenness and centrality scores) are used as features and the boxplot model removed less than 10% of the article. Inherently, the

objective to predict an article's impact is more difficult than estimating an author's impact.

Figure 8.a shows the density plot relative to the target column, namely the normalized number of citations of the articles. Note from this chart that the model still under-estimates higher citation scores. Figure 8.b shows the distribution of residuals (prediction minus true score) for this target. The previous observations for authors are also present: the trained regressor for articles under-evaluates the higher citations scores (left tail), but it has a high number of tiny over-estimation errors (right tail).

Finally, a Random Forest model was used to learn a global 'importance' relation between two distinct articles, starting from their local SNA metrics. The model uses the default configuration, except of the number of trees that was set to 10. Figure 9 shows the evaluation scores on the test set using different Research Gate metrics as target. Most of the metrics are about 80%. This result suggests that the local SNA scores for articles provide valuable insights in terms of inferring the global relative 'importance' of an article – i.e., if an article is more central and important from a semantic point of view within a community, it will most likely have a higher global score compared to less important local articles.

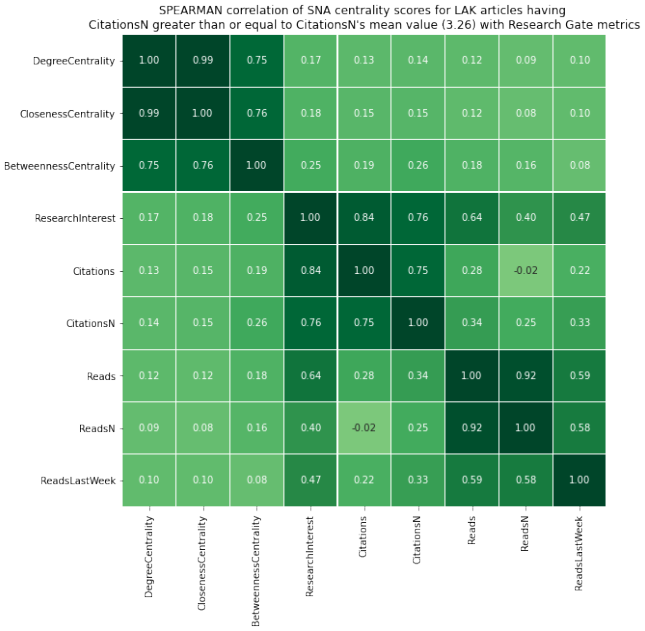
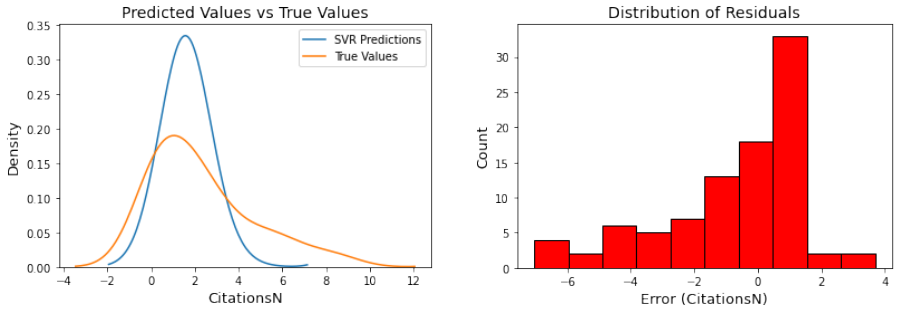


Fig. 7. Correlation of SNA centrality scores and Research Gate metrics for articles.



(a) (b)

Fig. 8. Support Vector Regressor results for articles - (a) Density plot, (b) Distribution of Residuals.

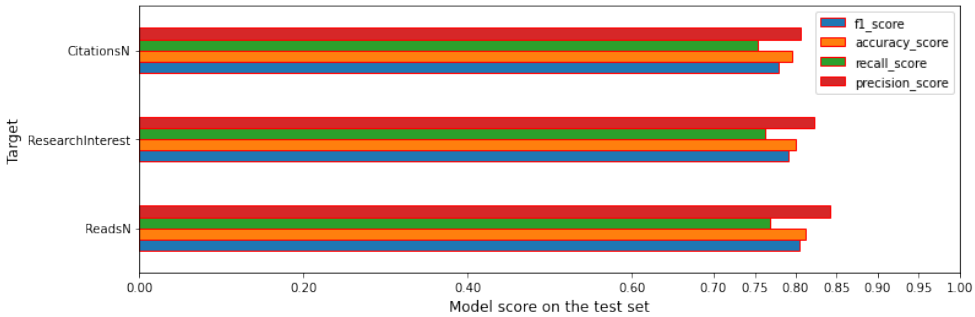


Fig. 9. Evaluation on test set using as target different Research Gate metrics for articles.

V. Conclusions and Feature Work

The current study introduces a novel method to compute the similarity between articles or authors using state-of-the-art models such as SciBERT. These measurements grounded in CNA showed promising modeling of the global impact of both authors and articles, transcending the local community. As such, semantic features extracted from the local LAK community provided reliable predictors for global statistics of both authors and papers.

In terms of future work, a principal improvement would be to explore and enhance the current method used for measuring the semantic similarity between articles, respectively between authors and between authors and articles. The current state-of-the-art techniques are represented by Transformers using their cross-encoder architectures to encode the semantic similarity between two text (sentences or paragraphs). However, this solution is not feasible for larger corpora due to the huge number of possible combinations. Our approach is a trade-off between performance and scalability, while also considering processing speed and resource consumption. The proposed solution involved training Transformer models to map each article to a vector space, where semantically similar articles

are closer, instead of using the traditional Transformer cross-encoder architecture. Then, the semantic similarity distance between each combination of two articles was easier and less computationally intensive to measure (using cosine distance, for example). In conjunction, LSA provided more contextualized information for the considered domain. Using the proposed finetuning method for Transformer architectures, combined with simpler methods (such as Tf-Idf, LSA or word2vec [21]) for filtering and ranking, might be a viable solution on the long run and may be applied to similar tasks. Furthermore, a scale up process is intended by using the full text of the articles, including all their paragraphs, instead of just considering their titles and abstracts when computing the semantic similarities.

Another direction worth exploring is how Transformer models perform when fine-tuned on Paraphrase Tasks datasets, such as the Microsoft Research Paraphrase Corpus [MRPC; 22] or the Quora Question Pairs (QQP) (https://www.kaggle.com/c/quora-question-pairs), that are considerably larger. Further evaluations of performance and computational efficiency might be performed. Additional evaluations could also be conducted on other sentence/paragraph embeddings benchmarks, such as SentEval [23] and Argument Facet Similarity (AFS) [24], as previous performed by Reimers and Gurevych [19].

Moreover, further steps may be taken to improve our method by analyzing the evolution of a research community across time. CNA and SNA techniques enable these types of analyses, but other new research leads involve exploring and extending the method to GNNs (Graph Neural Networks) for capturing and aggregating information contained in the 2-mode CNA graph structure.

ACKNOWLEDGMENT

This research was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number TE 70 *PN-III-P1-1.1-TE-2019-2209,* ATES – "Automated Text Evaluation and Simplification", and by the Office of Naval Research (Grants: N00014-17-1-2300 and N00014-19-1-2424) and the Institute of Education Sciences (R305A180144). The opinions expressed are those of the authors and do not represent the views of these granting agencies.

REFERENCES

[1] M. Dascalu, D. S. McNamara, S. Trausan-Matu, and L. K. Allen, "Cohesion Network Analysis of CSCL Participation," Behavior Research Methods, vol. 50, pp. 604–619, 2018.

[2] I. C. Paraschiv, M. Dascalu, D. S. McNamara, S. Trausan-Matu, and C. K. Banica, "Exploring the LAK Dataset Using Cohesion Network Analysis," in 3rd Workshop on Social Media and the Web of Linked Data (RUMOUR 2017), in conjunction with the Joint Conference on Digital Libraries (JCLD 2017), Toronto, Canada, 2017, pp. 17–21.

[3] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 2019, pp. 3613–3618.

[4] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge," Psychological Review, vol. 104, pp. 211–240, 1997.

[5] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in BlackboxNLP at EMNLP 2018, Brussels, Belgium, 2018, pp. 353-355.

[6] J. D. Cruz, C. Bothorel, and F. Poulet, "Community detection and visualization in social networks: Integrating structural and semantic information," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 5, pp. 11:1–11:26, 2014.

[7] B. Heller, E. Marschner, E. Rosenfeld, and J. Heer, "Visualizing collaboration and influence in the open-source software community," in ICSE11: International Conference on Software Engineering, Waikiki, Honolulu HI, USA, 2011, pp. 223–226.

[8] K. Reda, C. Tantipathananandh, A. Johnson, J. Leigh, and T. Berger‐Wolf, "Visualizing the evolution of community structures in dynamic social networks," Computer Graphics Forum, vol. 30, pp. 1061–1070, 2011.

[9] C. Vehlow, F. Beck, P. Auwärter, and D. Weiskopf, "Visualizing the evolution of communities in dynamic graphs," Computer Graphics Forum, vol. 34, pp. 277–288, 2015.

[10] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in WWW '15: 24th International World Wide Web Conference, Florence, Italy, 2015, pp. 1067–1077.

[11] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in KDD '14: The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2014, pp. 701–710.

[12] S. Cavallari, V. W. Zheng, H. Cai, K. C.-C. Chang, and E. Cambria, "Learning community embedding with community detection and node embedding on graphs," in CIKM '17: ACM Conference on Information and Knowledge Management, Singapore, Singapore, 2017, pp. 377–386.

[13] C. Tu, H. Wang, X. Zeng, Z. Liu, and M. Sun, "Community-enhanced network representation learning for network analysis," Computing Research Repository, vol. abs/1611.06645, 2016.

[14] P. Goyal, N. Kamra, X. He, and Y. Liu, "Dyngem: Deep embedding method for dynamic graphs," CoRR, vol. abs/1805.11273, 2018.

[15] I. C. Paraschiv, M. Dascalu, D. S. McNamara, and S. Trausan-Matu, "Finding the Needle in a Haystack: Who are the most Central Authors within a Domain?," in 11th European Conference on Technology Enhanced Learning (EC-TEL 2016), Lyon, France, 2016, pp. 632–635.

[16] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015, pp. 632–642.

[17] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, USA, 2018, pp. 1112–1122.

[18] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation," CoRR, vol. abs/1708.00055, 2017.

[19] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 2019, pp. 3980–3990.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representation in Vector Space," in Workshop at ICLR, Scottsdale, AZ, 2013.

[22] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in Third International Workshop on Paraphrasing at the 9th International Joint Conference on Natural Language Processing, Jeju Island, Korea, 2005.

[23] A. Conneau and D. Kiela, "Senteval: An evaluation toolkit for universal sentence representations," in Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan, 2018.

[24] A. Misra, B. Ecker, and M. A. Walker, "Measuring the similarity of sentential arguments in dialog," in The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Los Angeles, CA, USA, 2017, pp. 276–287.