Skoltech
Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Sergei Ivanov

PhD Program: Computational and Data Science and Engineering

Title of Thesis: Combinatorial and Neural Graph Vector Representations

Supervisor: Prof. Evgeny Burnaev

Name of the Reviewer: Maxim Panov, Assistant Professor, CDISE, Skoltech

I confirm the absence of any conflict of interest.	**Signature:**  **Date: 15-11-2019**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications

The summary of issues to be addressed before/during the thesis defense

The considered thesis targets the important area of constructing vector representations (or alternatively embeddings) for the graph data. The problem is very relevant as allows to proceed inherently discrete graph structures by means of standard data analysis algorithms working with data represented as vectors in Euclidean space. In this area, the particularly complex part is to construct embeddings which reflect the more complex structure of a graph than just the degree distribution in it. Ideally, such structural embedding of a graph should be able to test on isomorphism or near-isomorphism of graphs. That's exactly the area which is targeted by the thesis, which is very relevant for the modern network science as only few methods attacking the problem exist, none of them showing consistently high quality in practice.

The thesis focuses on the so-called anonymous walk embeddings (AWE), which were introduced in the earlier paper [Micali, 2016] and were proved to contain the information sufficient to reconstruct the sub-graph of any size given the sufficient length of walks. However, the proof uses very long walks, which are impossible to be realized in practical algorithms. The thesis argues that we can use much shorter walks and us their frequencies as a structural embeddings of a graph. Unfortunately, the authors both of [Micali, 2016] and the thesis don't explicitly give the sufficient length of walk to reconstruct the ball of given radius, which would be very useful to compare with proposed solution. However, the experimental results indeed show that the proposed embeddings work well. The neural model processing AWEs and inspired by modern NLP algorithms is also proposed which tries to improve over the feature based approach. The experimental results show the significant improvement in some cases. Overall, anonymous walk embeddings are proved to be useful for applications surpassing the state-of-the art competitors in terms of the quality of downstream tasks' solution.

I have several questions about this part of the thesis:

1. It is not clear for me, in what sense anonymous walks are "non-linear graph objects".
2. The sampling procedure for AWE is called "efficient" but it is not clear what does it mean exactly.
3. The neural model makes the direct analogy with NLP algorithms using, for example, notion of window size. However, anonymous walks are naturally not ordered and it is not clear how the emphasized analogy with NLP helps.
4. Considering very large standard deviation, may one confidently say that AWE is better than other methods in the experiment summarized in Table 2.6?
5. What can be the reason for relatively low performance of AWE on biological datasets (Table 2.5)?

The last chapter of the thesis focuses on the graph embeddings for combinatorial problems such as product recommendation in graphs and influencer recommendation in social networks. The overall type of the research here is similar to previous chapters: the formulated problems are NP-hard and the author proposed approximate solution, which works very well in practice. Despite that applications here are less general than in the case of AWE, the results are still relevant for practice. I also have a feeling that this part should be better linked to the other parts of the thesis as even the style of writing is different here. For example, the formulas are not properly integrated into the text due to issues with punctuation. I mostly liked very detailed and well-designed experiments which strongly support the significance of results.

Finally, the results of the thesis research were published in the proceedings of well-reputed conferences including two conferences having A* rating according to CORE conference rating. Thus, the quality of the publications well supports the overall good scientific quality of Sergei's thesis research.

While I have overall positive opinion about the research contents of the thesis I think that the text deserves serious improvement. The list of suggestions which should be incorporated in the final version of the manuscript (in order of appearance in the text):

1. The Section 1.1 contains only very few references while touching many important concepts. I think that that the references should be provided to all the major concepts discussed. The same applies (in the greater extent) to Section 2.1.
2. The formulation of some of research goals in Section 1.1.1 is very vague (like “development and analysis of new graph embeddings for the graph classification problem” or “formulation of new graph-based problem of medical diagnostics”). In my opinion the goals should be focused on solving certain problems, not just developing something new.
3. The publications at the end of section 1.1.1. seem to be listed without following any particular scientific references style. The publications should at least have full citation data including, for example, pages.
4. The transition to descriptors for molecules on the page 33 is rapid and completely unexplained.
5. Everywhere references to formulas should contain () around. This is done by \eqref in Latex.
6. There are 2 definitions of anonymous walks (2.8 and 2.9). The necessity of having 2 definitions and the difference is not clear.
7. Some phrases in the text are less formal than they should be. For example, after Theorem 2.13 is said that “By Theorem 2.13 it’s very unlikely to have a polynomial-time algorithm…”. However, Theorem 2.13 is strict and not-probabilistic which makes this sentence not correct.
8. The object in Formula 2.21 is called the likelihood, while it is not likelihood but evidence low bound for it.

There are also multiple misprints and minor language issues appearing throughout the manuscript. The list of such minor issues was communicated to the candidate directly.

To sum up, I think that the issues found do not decrease the scientific quality of the thesis and Sergei Ivanov deserves to be awarded with Skoltech PhD degree.

Provisional Recommendation

☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

☒ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate’s thesis according to the recommendations of the present report*

☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*