Report for the Doctoral Dissertation Thesis

Title: Data mining in social network analysis
Author: Mgr. Peter Zvirinský
Supervisor: doc. RNDr. Iveta Mrázová, CSc.
Opponent: RNDr. Petra Vidnerová, Ph.D.

The thesis is focused on social network analysis (SNA) application to Czech insolvency proceedings. The topic is up-to-date, exceptionally interesting and offers a lot of space for possible innovations and novel research.

The text comprises 155 pages and is divided into 6 chapters accompanied by an introduction and conclusion. In addition, there is dataset scheme documentation and attachments containing dataset and source code information. The structure of the thesis is logical.

After the introduction, the first chapter defines the goals of the thesis. There are four goals:

Goal 1: Model the insolvency process utilizing a social network approach.
Goal 2: Model the insolvency process in time utilizing a dynamic social network approach.
Goal 3: Enrich the insolvency network by metadata extracted from the Insolvency Register.
Goal 4: Predict the future development of the insolvency network.

The goals are well-described and comprehensible, however, they resemble more the workflow of the thesis than the goals.

Chapter 2 is an overview of the SNA field. It is quite exhaustive and looks like a result of complex and extensive literature research. It defines and explains terms, methods and concepts that will be used later in the experimental part. The chapter definitely proves the student's solid knowledge of the field, on the other hand, it is written rather generally (without the focus on the goals), and social networks like Facebook are often mentioned, despite this kind of network not relating too much to the subject of the work.

Chapter 3 covers the insolvency system in the Czech Republic. Like Chapter 2, it is a summary and overview of existing information to provide background for the research topic. It brings documentation of the Insolvency Register (IR) and its value stems from the fact that it is the only documentation of IR in English.

Chapter 4 describes the Czech Insolvency Dataset (CID) and the process of its creation. The complete pipeline is proposed, which is a novel work of the student. The pipeline uses known algorithms and methods only, but is nontrivial, combining the processing of both structured and unstructured data (including the use of the OCR approach).

The dataset itself can be seen as a side-product of a thesis, however, it has a significant value itself. It is a unique original dataset, while there is a lack of publicly available datasets of this kind for research purposes.

Chapter 5 starts with definitions of terms needed further and its main purpose is to describe the GraphSlices software tool. Existing tools are reviewed, which again proves the student's broad knowledge of the field and available methods. The GraphSlices software tool is a novel piece of software, written in Scala, implementing a graph interface and existing algorithms needed for the analysis done further in the thesis.

Chapter 6 covers experiments and is the core chapter of the thesis. It consists of five experiments, each in an individual section with a unified structure - first defining the research questions, then defining the construction of the graph, and finally explaining the results. Though the chapter is perfectly structured, I miss a general introduction to the experimental part.

Individual experiments match the goals defined at the beginning. The first experiment demonstrates the usage of SNA and fulfils Goal 1, on the other hand is quite straightforward and its results are not surprising. The second experiment adds a time dimension to the problem and thus fulfils Goal 2. It also demonstrates the use of association rule mining being a preliminary step for future prediction of graph development. The following two experiments refer to Goal 3 since they analyse the data gained from unstructured documents, namely information about debt origin and the value of the claimed debt. The last experiment then fulfils Goal 4, being the most complex and interesting part of the experiments. Generally, the experimental part demonstrates the feasibility of the proposed methodology, takes advantage of the created CID dataset, and uses the concepts and approaches described in previous chapters. As a drawback, one can see that it also uses approaches not described before (such as the SOFM algorithm), but these are existing methods referred to in references.

The thesis as a whole is well-written, readable and comprehensible. Although it does not contain any novel theoretical results, it presents a complex solid piece of work. As the main contributions of the work, I reckon:

1. The novel unique dataset - the Czech Insolvency Dataset
2. The GraphSlices software tool
3. The complex methodology for both static and dynamic analysis of insolvency data

The work may serve a broader scientific community to build on and it has also a real-world application potential since it works with real data and problems.

The work comprises many SNA methods and a large variety of machine-learning approaches used both during dataset creation and experiments, which proves the student's broad knowledge of both fields. In addition, the creation of the CID dataset required work with big data, which was also perfectly managed.

The student demonstrated his ability to work independently on complex research problems. Therefore I recommend the thesis be accepted as a dissertation thesis.

Prague, 14. 3. 2024

RNDr. Petra Vidnerová, Ph.D.

Questions:

- When creating the CID dataset one has access to many personal information about the actors. Did you face the problem with GDPR and how it was solved?
- The GraphSlices software may be useful to a wide range of researchers, however, the last changes on the GitHub repository are quite old. Do you plan to maintain it in the future? Is it compatible only with SCALA or is it possible to make it accessible through the interface to other programming languages?
- During the experiments, you tackled various research questions. What are the limits of the proposed methodology? I.e. what kind of problems cannot be solved by this approach?
- Recently, graph neural networks emerged. Do you think they can be useful for your research and how?

Notes:

- Page 12: The term "degree of separation" is used without being defined
- Page 13: The terms "directed" and "undirected" should be defined.
- n in equation (2.8) is not defined
- Page 45: "contains amendments up to the end of 2013" - it describes Acts from 2017 and 2019
- Page 117: GraphSlices library is mentioned. Does it mean, it is used only in this last experiment and not in the first four experiments?