Network analysis methods for smart inspection in the transport domain

Bruin, G.J. de

7 Conclusions

In this final chapter, we first answer the five research questions in Section 7.1. Subsequently, our answer to the problem statement is formulated in Section 7.2. Lastly, five future research directions (in addition to Section 6.7) are proposed in Section 7.3.

7.1 Answers to the research questions

We reiterate the research questions formulated in Chapter 1. Each research question is answered separately, along with references to relevant sections in which details can be found.

> **Research question 1:** *What is the relation between network structure and model performance in link prediction?*

In Chapter 2, we considered a large set of temporal, structurally diverse, real-world networks. We investigated the relationship between the structure of these networks and the model performance in link prediction for this set of networks. We found several structural network properties related to model performance in link prediction. Most notably, a negative correlation was discovered between network degree assortativity and link prediction performance. This negative correlation was also observed for real-world networks that had their degree assortativity artificially altered by means of a degree rewiring process. Our research showed that link prediction performance is generally higher in degree disassortative networks. In degree disassortative networks, the numerous low-degree nodes connect more frequently with hubs than with other low-degree nodes. For these low-degree nodes, the preferential attachment feature will provide higher scores for high-degree candidate node pairs. Hence, the supervised model can use this information to perform better (Finding 1).

In addition, regarding the temporal structure of networks, we distinguished between two classes of temporal networks, being temporal networks (1) containing only *persistent*

relations and (2) also containing *discrete events* (see Section 2.1). We found that model performance in link prediction improved significantly when in networks with discrete events, all events were explicitly taken into account. We coin this method "past event aggregation". It essentially is a method in which *all* information contained in both *persistent relations* and all *discrete events* is used (Finding 2).

Together, these two findings provide an answer to Research question 1.

Research question 2: *How can we obtain accurate estimates of the performance of link prediction models by using adequate splits into the train, validation, and test set?*

In Chapter 3, we described two dominant methods from the literature used to split network data in a train, validation, and test set for link prediction. We applied these two methods, called: the (1) random split and (2) temporal split, to six different temporal networks that have a considerable number of nodes and edges. We learned that the *random split* method provides (too) *optimistic* results. Therefore, the *temporal split* method should be used because we confirmed that it gives a *more realistic indication* of performance.

Research question 3: *How do network structure and vehicle attributes relate to co-driving behavior?*

In Chapter 4, we applied the link prediction approach to the truck co-driving network in an attempt to better understand the behavior of trucks and their drivers. Our research on the importance of features indicates that the network structure is better explained by co-driving behavior than by vehicle (node) characteristics. In particular, the *neighborhood features* that capture relevant information about the ego networks explained the observed co-driving behavior well.

Research question 4: *How can node attribute information be exploited to automatically create a good partitioning of a co-driving network into communities?*

In Chapter 5, we investigated the task of detecting communities of the truck co-driving network. The communities were detected by a modularity maximization algorithm, which has a resolution parameter. This parameter determines whether a more fine-grained or coarse-grained partition into communities is preferred. We proposed a method that considers node attributes to determine the best partitioning of the network into communities. In this method, a metric that we call average maximal community assortativity quantifies how well, on average, each community can be understood in terms of its node attributes. This metric was maximized to find the best choice for the resolution parameter. When applied to the truck co-driving network, results indicated that a good partitioning into communities was obtained by considering geographical aspects of the trucks as node attributes.

Research question 5: *How can ship behavior be utilized to enable smart inspection of cargo ships?*

The smart inspection entails the accurate, automated, fair, and interpretable assessment of (in our case) cargo ships. In Chapter 6, we proposed a machine learning model capable of predicting cargo ship noncompliance. We make use of (fair) random forests, because they allow humans to understand (1) what procedures were followed to make the model, (2) the inner workings of the model, and (3) how the model arrives at its predictions. The model's fairness was obtained using fair pre-trained models. The model decorrelates a ship's flag from the noncompliance prediction to reduce present bias in historical data and thereby prevent confirmation bias. The cargo ship network is constructed from behavioral data, which is less sensitive to manipulation than administrative information. Features derived from this cargo ship network served as input for the machine learning model. In summary, the entire approach led us to demonstrate how smart inspection should take place in the future.

7.2 Answer to the problem statement

After addressing the research questions, we now turn to the problem statement.

Problem statement: *How can network science methods leverage behavioral data for smart inspection of vehicles?*

The short answer to the problem statement is to be seen by applying the results of all five research questions. We summarize them below.

In answering Research question 1, we have shown that network science methods can generate useful features for a downstream machine learning task. This is *directly applicable* to the more fundamental link prediction task in networks, as seen in Chapter 2, and also *useful* in applied settings, for example, in identifying noncompliant ships (Chapter 6).

In answering Research question 2, we have shown that in link prediction, careful consideration must be given to splitting instances into an appropriate train, validation, and test set (Chapter 3).

Moreover, in answering Research question 3, we have explored other network science methods to better understand vehicle data, with a special focus on the relation between network structure, vehicle characteristics.

In answering Research question 4, we address the community structure (Chapters 4 and 5). The obtained results in these two chapters demonstrated that a network perspective on truck driving activities helps to uncover patterns that may ultimately be useful for promoting co-driving and reducing traffic congestion and fuel usage.

Finally, in answering Research question 5, we used network science tools to consider behavior as features in a machine learning model. By application of fair pre-trained models in Chapter 2, we achieved the desired smart inspection of vehicles.

7.3 Future research directions

The following are five directions (seen as addition to the four directions mentioned in Section 6.7) fruitful for future research.

1. **Argument:** Many current link prediction approaches have limitations in handling large and dynamic networks [101]. Applying dimensionality reduction before link prediction may improve scalability but could negatively impact interpretability.
   Future research: One straightforward direction is to produce interpretable techniques that scale well to larger networks. Many real-world networks are highly sparse, meaning the number of positive instances (pairs of nodes that will link) is very few compared to negative ones (pairs of nodes that do not link). Therefore, positive instances can be considered outliers, and thus outlier detection techniques may do well in link prediction, especially on large and dynamic networks.
2. **Argument:** We encountered limited availability of temporal network datasets.
   Future research: To advance link prediction, a *more diverse set* of temporal networks must be accessible to the public and not locked in private "silos" where they are accessible only by some [179]. To start, in Chapter 2, we presented a collection of 26 temporal networks.
3. **Assumption:** Incorporating features obtained from more sophisticated transport network models into smart inspection techniques may benefit prediction performance.
   Future research: Higher-order networks [170, 203] and evolutionary hypergraphs [212] have been proposed as more effective representations for capturing vehicle trajectories.
4. **Argument:** A natural progression of this work is to consider a more holistic approach toward inspection in the transport domain. Whereas in this work, we analyzed the cargo trucks and cargo ships separately, they are not independent in the real world. The containerization of the transport system facilitates smooth transfers between different modalities.
   Future research: A further study could assess the risk associated with the entire cargo journey.
5. **Argument:** In our work, we did not extensively consider the uncertainty in the network inferred from the available raw data. However, our data is likely partially incomplete, raising possible questions about to what extent the dataset is representative.
   Future research: A greater focus on measuring errors could shed more light on the difference between the data (i.e., what is measured) and the abstract, underlying network representation [150].

General goal and general recommendation

Ultimately, our goal is to improve *cleanliness* and *safety* in the transport domain. The proposed approach to smart vehicle inspection is just one of the actions needed to arrive at transportation without any danger or unnecessary environmental pollution. A combined and continuous effort is needed from many professions (policymakers, scientists, inspectors, and of course, ultimately, the vehicle drivers themselves) to offset all negative transportation consequences. We expect that the work in this thesis will contribute to the ongoing shift toward the smart inspection of vehicles.