Investigation of the predictive power of the human capital of academic researchers on their financial performance using social network analysis and decision trees

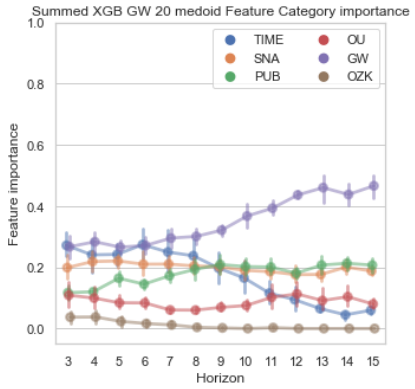
Wim Fyen[1 [0000−0001−9390−8816]⋆], Wannes Meert[1 [0000−0001−9560−3872]], Bart Thijs[1 [0000−0003−0446−8332]], and Koenraad Debackere[1 [0000−0002−3411−0668]]

KU Leuven, Leuven, Belgium

In this thesis, we investigated the relationship between human capital and (subsequent) financial performance of academics in a research-intense university [6] to see what insights could be learned for policy making [5]. In order to describe this relationship, we combined academic publication data with HR data and extracted various features as proxies representative for the 'human capital' for a set of academics. We then used a supervised machine learning approach in which financial performance (sum of subsidies and contract research revenue) was used as the target variable that we tried to predict for various lengths of time (from 3 to 15 years) between the source and target variables (the 'Horizon'). Important to note is that this approach gives information on the most important features but does not necessarily provide information on causality. The features describing the human capital of the academics were grouped into six categories: i) TIME: time-related features (such as tenure, ...), ii) SNA: social network analysis metrics (such as degree, betweenness, ...[7]), iii) PUB: publication related metrics (such as number of publications, average number of co-authors, ...), iv) OU: organisational unit of the academics, iv) GW: GraphWave network role metrics and vi) OZK[1]: describing access to people with a research managerial role in the research group. For the network role features, we built a weighted network from the publication data, with the academics as nodes and their co-publications as edge weights. The role of the academics was described using the GraphWave (GW) algorithm [2] that encodes the inter-node influences into a vector format [3]. These vectors are subsequently mapped into trajectories into the complex plane [4] (whereby academics playing a similar role in the network are mapped onto trajectories that are closer together). As features we used the distances between each academic and a set of prototype roles (using the medoids of a k-means clustering, i.e. not hand-crafted features, but automatically derived from the publication data). For the machine learning we used regression trees due to their high level of explainability. We compared the results of single trees, random forests and XGB[1] models. We found that the latter two were significantly better to relate human capital related features to financial performance than the first one. We also observed that the relationship was very dependent on the Horizon.

⋆ Ma thesis student 2020-2021

[1] from the Dutch 'Onderzoekskader': OZK



The resulting feature importances (for k=20 and an XGB regression tree) are summarized in the figure. When looking at the individual feature importances, we found that the dominant ones at small Horizon were the time-related features and the betweenness centrality. These features decreased in importance as the Horizon got larger. The number of publications and number of assignments to different organisational units were quite important too, and became dominant at large Horizon values. Other social network related features (e.g. various degree and centrality metrics) were relevant throughout the entire range of Horizon values. The access to research managers profiles in the network showed a minor but statistically significant effect at short Horizons. An interesting observation was related to the GraphWave features from which we used (the similarity to) 20 different network roles in the model. It was found that for the XGB trained model, these features - when taken jointly - were in fact the dominant feature category for all Horizons suggesting that the role of a researcher in its collaborative network adds significant predictive power to the model, besides the other feature categories.

This work significantly improves upon the state of the art in at least two ways: first of all, the combination of data sources used is very rare since financial data of this nature are typically not available to the broader research community, especially when combined with an extensive set of bibliometric data as well as HR related data. Secondly, the work went beyond the traditional HR and bibliometric methods used to approximate the human capital (such as degree, betweenness, ...) and also included information on the role of academics in their collaborative environment. As such, it demonstrates the applicability of network role mining to extract meaningful information for policy making in research intense universities. Future work can look more in detail into specific financial metrics (e.g. instead of the total yearly revenue one can focus on specific income sources by category). Alternatively one can also treat the financial data as time series and look into the effect of exogenous 'shocks' such as changes in legislative environment. Additionally, one can also investigate more in detail the GraphWave features and learn which type of network roles are more predictive of financial success.

References

1. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794 (2016)
2. CSIRO's Data61: Node representations with graphwave. URL: https://stellargraph.readthedocs.io/en/v1.2.1/demos/embeddings/graphwave-embeddings.html, last checked on 2021-10-03
3. Donnat, C., Zitnik, M., Hallac, D., Leskovec, J.: Learning structural node embeddings via diffusion wavelets. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1320–1329 (2018)
4. Epps, T.W.: Characteristic functions and their empirical counterparts: Geometrical interpretations and applications to statistical inference. The American Statistician **47**(1), 33–38 (1993)
5. Fyen, W.: A study of the effects of human capital on financial performance of academics in a research intense university (2021)
6. KU Leuven: Corporate booklet. URL: https://www.kuleuven.be/over-kuleuven/pdf/corporate-booklet, last checked on 2021-06-03
7. Newman, M.E.J.: Scientific collaboration networks. i. network construction and fundamental results. Physical Review E **64**(1), 16131 (2001)