*Everything You Always Wanted to Know About AI**

(*But Were Afraid to Ask)

Nowcasting Digital Skills with Wikipedia

(Working Paper - Version 16/9/19)

Fabian Stephany¹²³

¹ Humboldt Institute for Internet and Society, Französische Straße 9, 10117 Berlin, Germany

² Wittgenstein Centre, Welthandelsplatz 2, 1020 Vienna, Austria

³ AtomLeap GmbH, Oranienstraße 183, 10999 Berlin, Germany

OrcidID: 0000-0002-0713-6010, mail@fabianstephany.com

1 Introduction

Digital technologies have a pervasive effect on our society. They augment or transform various previously analogue processes of value creation, capture, and exchange [1,2]. Hence, on the labour market, the skillful development and application of relevant digital technologies are in strong demand. However, early research findings indicate that the labour demand of certain tech industries is not met. The talent pool does not grow at the pace of industry demand and precise skill requirements related to growing technologies, such as Artificial Intelligence (AI), remain opaque [3].

This work proposes a network perspective in order to empirically identify the relevant ICT skills related to AI, how their composition changes over time, and how they could be predicted with online data. With the example of the US tech sector, two data sources are employed: The US' most popular online tech-job platform *dice.com* allows to relate ICT skills, in a network structure, from an industry perspective. Two skills are connected in an industry demand network, if they are jointly required by the same job advertisement. In addition, data from the online encyclopedia *Wikipedia* allows to create a network on the online knowledge side of ICT skills⁴. Here, skills are related in a network, if their respective articles are connected by a reference hyperlink (Figure 2 in the Appendix illustrates).

Similarly, past skill networks can be constructed: Information about previous job advertisements is stored in older versions of *dice.com* on the web-archive⁵ since 2004 and *Wikipedia's* reference history allows to reconstruct article networks at any point in time during the history of the encyclopedia. Over time, composition developments of the skill networks can be compared between and within the two network environments. The underlying hypothesis is that developments in the online knowledge network (*Wikipedia*) proceed changes in the industry demand network (*dice.com*). While it is one of the main challenges of this work to assess the validity of this hypothesis, previous studies have shown that the edit activities of the *Wikipedia* crowd enable early

⁴It is reasonable to assumed that *Wikipedia* article editors possess relevant topical knowledge on the digital skill the edited article is about.

⁵http://web.archive.org/web/*/http://www.dice.com/jobs

predictions on movie sales [4], electoral popularity [5], stock prices [6], knowledge hubs [7] or even global spreading of diseases [8].

2 Results

For the first exploratory part of the research, two ego-centred networks are regarded: All job postings advertised with the tag "Artificial Intelligence"⁶ on *dice.com* on September 14th, 2019, are considered. Skill tags are connected as nodes in a network, if they appear in the same advertisement. Similarly, all articles linked with the *Wikipedia* article "Artificial Intelligence"⁷ and the articles they refer to are connected in a network⁸. For both networks, a common set of most relevant overlapping skill tags are identified⁹. With the use of the *Wikipedia* edit history and the web-archive, the date, when these tags entered the respective ego-centred AI-networks on both *Wikipedia* and *dice.com* are registered¹⁰. On average, skill nodes joined the *Wikipedia* AI-network 15 weeks before they appeared in relation to AI on *dice.com*, as illustrated in Figure 1.

Fig. 1: In both ego-networks of AI (*Wikipedia* and *dice.com*), new nodes join over time. However, nodes entered significantly earlier on *Wikipedia*. For a set of selected skills, on average, their articles have been linked to "Artificial Intelligence" about 15 weeks before they had been announced on job advertisements about AI.

With this first indication of the predictive potential of *Wikipedia* data on digital skills, future extensions of this work focus on an identification of AI-cliques by clustering algorithms and under the consideration of all relevant skill tags that are currently

⁶<https://www.dice.com/jobs?q=Artificial+intelligence&l=>

⁷https://en.wikipedia.org/wiki/Artificial_intelligence

⁸Nodes can, at most, have a distance of one iteration from the original AI article.

⁹ApacheMX NeT, Big data, Caffe, Computer vision, Data mining, Data science, Deep learning, Keras, Machine learning, Natural language processing, Neural networks, Predictive analytics, Python, PyTorch, TensorFlow, Theano, R, RNN

¹⁰For five of the key skills it was possible to find the entry into the respective networks.

in use on *dice.com* and their respective *Wikipedia* articles. Moreover, future extension take centrality metrics, e.g., *Eigenvector centrality*, into account for comparing the state and development of skill nodes in both network environments. Lastly, the comparison of network similarities, e.g., *Jaccard similarity*, allows future investigations to evaluate the development of the AI-cliques within and across the two network environments. In addition, following investigations should take regional digital knowledge geographies [9] into account, too.

Overall, the insights of this project can support businesses in developing a data-driven strategy for the acquisition and the development of adequate skills needed to implement and leverage new technologies at best. Furthermore, the empirical relationship of digital skill sets will help to establish a common taxonomy to be used by policy makers, education providers, and recruiters, so that job market mismatches can be reduced. Lastly, a potential predictive power of the online knowledge network could help to develop farsighted programmes for the training of digital skills in the future.

Summary. With the use of online data from the tech job platform *dice.com* and the online encyclopedia *Wikipedia*, two networks of digital skills are created around the topic of Artificial Intelligence. Initial research indicates that new skill tags first join the *Wikipedia* network, before they appear in AI-related job announcements on *dice.com*. The findings of this work could be used in order to create a data-driven strategy for the acquisition and the development of adequate skills needed to implement and leverage new technologies at best.

References

1. Yoo, Youngjin, Ola Henfridsson, and Kalle Lyytinen. "Research commentarythe new organizing logic of digital innovation: an agenda for information systems research." *Information systems research* 21, no. 4 (2010): 724-735.
2. Nambisan, Satish, Kalle Lyytinen, Ann Majchrzak, and Michael Song. "Digital Innovation Management: Reinventing innovation management research in a digital world." *Mis Quarterly* 41, no. 1 (2017).
3. De Mauro, Andrea, Marco Greco, Michele Grimaldi, and Paavo Ritala. "Human resources for Big Data professions: A systematic classification of job roles and required skill sets." *Information Processing Management* 54, no. 5 (2018): 807-817.
4. Mestyn, Mrton, Taha Yasseri, and Jnos Kertsz. "Early prediction of movie box office success based on Wikipedia activity big data." *PloS one* 8, no. 8 (2013): e71226.
5. Yasseri, Taha, and Jonathan Bright. "Can electoral popularity be predicted using socially generated big data?." *it-Information Technology* 56, no. 5 (2014): 246-253.
6. Moat, Helen Susannah, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis. "Quantifying Wikipedia usage patterns before stock market moves." *Scientific reports* 3 (2013): 1801.
7. Stephany, Fabian, and Fabian Braesemann. "An exploration of wikipedia data as a measure of regional knowledge distribution." In *International Conference on Social Informatics*, pp. 31-40. Springer, Cham, 2017.
8. Generous, Nicholas, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. "Global disease monitoring and forecasting with Wikipedia." *PLoS computational biology* 10, no. 11 (2014): e1003892.
9. Stephany, Fabian, Fabian Braesemann, and Mark Graham. "Coding Together-Coding Alone: The Role of Trust in Collaborative Programming." (2019).

3 Appendix

Fig. 2: Two skill networks are constructed with the example of Wikipedia articles (lhs) and dice.com job advertisements (rhs). The stronger the edges between the nodes, the more references articles share with each other / the more jobs have jointly advertised the same skills. The example of the nodes, most strongly connected to the software TensorFlow, serves as an illustration. TensorFlow belongs to the field of machine learning skills and was written in the programming languages Python and C++. Considering the subgraph of these four tags, exclusively, one can see that C++ is more strongly connected in the online knowledge network than in the industry demand network. This could potentially be explained by the fact that C++ is less relevant for the application of TensorFlow, than it was for the development of the software.