World Patent Information 73 (2023) 102193

Contents lists available at ScienceDirect

World Patent Information

journal homepage: www.elsevier.com/locate/worpatin

Forecasting labor needs for digitalization: A bi-partite graph machine learning approach ☆

Dimitri Percia David [a,b,c,*], Santiago Anton Moreno [d], Loïc Maréchal [e], Thomas Maillart [a], Alain Mermoud [b]

[a] Information Science Institute, GSEM, University of Geneva, 40 Boulevard du Pont-d'Arve, Geneva 1211, Switzerland
[b] Cyber-Defence Campus, armasuisse Science and Technology, Feuerwerkstrasse 39, Thun 3602, Switzerland
[c] University of Applied Sciences of Western Switzerland (HES-SO Valais-Wallis), Institute of Entrepreneurship & Management, Techno-Pôle 1, Sierre 3960, Switzerland
[d] EPFL, Section of Mathematics, EPFL FSB SMA, Station 8, Lausanne 1015, Switzerland
[e] University of Lausanne, HEC Lausanne, Department of Information Systems, Internef, Lausanne 1015, Switzerland

ARTICLE INFO

Keywords:
High skilled labor
Intellectual property
Technology mining
Bi-partite networks
Graph machine learning
Link prediction

ABSTRACT

We use a unique database of digital, and cybersecurity hires from Swiss organizations and develop a method based on a temporal bi-partite network, which combines local and global indices through a Support Vector Machine. We predict the appearance and disappearance of job openings from one to six months horizons. We show that global indices yield the highest predictive power, although the local network does contribute to long-term forecasts. At the one-month horizon, the "area under the curve" and the "average precision" are 0.984 and 0.905, respectively. At the six-month horizon, they reach 0.864 and 0.543, respectively. Our study highlights the link between the skilled workforce and the digital revolution and the policy implications regarding intellectual property and technology forecasting.

1. Introduction

In 2005, the Swiss Economic Institute at ETH Zurich produced a study on the importance of computerization on workplace organization, skilled labor, and firm productivity [1]. This study followed another on the importance of skilled labor in information technologies to the adaptation of organizations and their competitiveness [2]. Twenty years onward, the digitalization of business processes has become a fierce global battle between organizations [3] with great challenges of business adaptation [4,5]. This battle involves the production of intellectual property (IP) such as industry secrets, patents, and open source software, as well as specific IP arrangements for the semiconductor industry [6], and for the highly skilled labor required to produce it [7]. In the international context, developing human capital for producing IP is a high priority for national competitiveness [8]. Similarly, at an age of high cybersecurity concerns by companies and governments [9] regarding the preservation of data privacy or the reliability [10,11] and the resilience of critical infrastructures to cyber attacks, the shortage of skills is a priority in public policy [12].

Labor skills in digitalization remain a key topic globally. In this context, Swiss companies seek to remain competitive while operating in one of the most expensive countries in the world [13]. Yet, human expertise in cybersecurity is in particularly short supply. This scarcity leads to important opportunity costs for organizations [14–16], as they evaluate which labor force to acquire and when. In cybersecurity, a technological advantage can be obtained through the acquisition of (i) material, (ii) human, and (iii) knowledge resources. Hiring human expertise is one way to acquire knowledge and IP, which then turns into a defense capability [17] to overcome digital threats [18]. While knowledge production and IP closely relate to digitalization and cybersecurity to skills, we question how job openings in digital technologies can be predicted, as they signal capacity building in the deployment of a technology [19]. We question the predictability of job openings as an input to the production of IP. We use a unique dataset of job openings

Abbreviations: IP, Intellectual Property; IPR, Intellectual Property Regime; PA, Preferential Attachment; HS, Hyperbolic Sine; SVM, Support Vector Machine; TPR, True Positive Rate; FPR, False Positive Rate; ARIMA, Autoregressive Integrated Moving Average; ROC, Receiver Operating Characteristic; AUC, Area Under the Curve; PR, Precision-Recall; AP, Average Precision; TMM, Technology & Market Monitoring

☆ Our code is available at: https://github.com/technometrics-lab/6-Link_Prediction.

* Corresponding author at: University of Applied Sciences of Western Switzerland (HES-SO Valais-Wallis), Institute of Entrepreneurship & Management, Techno-Pôle 1, Sierre 3960, Switzerland.
E-mail address: dimitri.perciadavid@hevs.ch (D. Percia David).

https://doi.org/10.1016/j.wpi.2023.102193
Received 12 February 2022; Received in revised form 13 February 2023; Accepted 14 March 2023
Available online 28 March 2023

highly related to the production of IP by organizations in Switzerland. We use a graph machine-learning approach to predict job openings as a bi-partite network of (i) Swiss organizations (*e.g.*, UBS, Nestle, or Roche) and (ii) IP-intensive software technologies (*e.g.*, blockchain or artificial intelligence).

The remainder of this paper proceeds as follows. Section 2 presents the research background. Section 3 introduces our data and methods, Section 4 details our results, Section 5 discusses their importance and Section 6 concludes.

2. Related work

2.1. Intellectual property, labor, digitalization, and cybersecurity

Intellectual property regimes (IPRs) for software and hardware include trade secrets, patents, and open source. Organizations aim to keep their knowledge secret [20], protected [21], while ensuring the production of non-rival, non-exclusive collective goods [22]. Regardless of the IPRs, the production of knowledge proceeds from high-skilled labor [2]. The relation between labor and IP production can be traced back to John Locke (1690), in which he argues that *person who labors that are either unowned or held as a commons has a natural property right to the fruits of her efforts*. This idea is especially applicable to IP, for which the raw material (knowledge) is held in common and where labor contributes to the value of finished products [23,24].

Although technological advancements in all fields have evolved, digitalization and cybersecurity technologies play a special role in knowledge production and for organizations at the strategic level [25], for their business models [4,5], and business processes [26] to compete on global markets [3,27]. Some organizations face specific challenges when they are located in countries with expensive labor [13]. For instance, the semiconductor industry exhibits an increasing division of labor, which foster a separation between knowledge and hardware production [6]. More broadly, the digital industry is subject to high specialization, and modularization [28,29]. This applies to technological development and criminal activities [30]. As a result, skill shortage impedes innovation for digitalization [12,31], and other fields [1,2,32]. Skilled workers contribute to the production of IP regardless of IPRs [24]. Finally, IP production preempts competitiveness, warrants trade secrets, and ensures that organizations compete in the knowledge-intensive industry of digitalization [13].

Given the high cybersecurity concerns for companies [33], and governments [9] regarding data privacy or the reliability of critical infrastructures to cyber attacks, the shortage of skills is a major concern in public policy [12,16]. Altogether, the importance of high-skilled labor as the main input for producing IP, especially for digitalization and cybersecurity, has received little attention. More broadly, the understanding of labor needs by organizations to produce IP remains limited in the literature, especially in that using technology mining methods in science, technology, and innovation [34].

2.2. Bi-partite networks and graph learning

To understand the labor needs as input for the organizations' IP in digitalization, one may consider a dynamic network with two types of nodes (organizations and digital technology). Literature on two-node networks or bi-partite networks with machine learning includes Benchettara et al. (2010), who adapt link-prediction metrics, to enhance forecasting performance over traditional metrics [35]. Taking the bi-partite nature of the graph into account enhances the performance of prediction models. Silva et al. (2012) and Tylenda et al. (2009) explore time-dependent metrics using time-series within link-prediction analysis and show a significant improvement over time-independent methods [36,37]. Link-prediction methods also include supervised learning. Mohammad et al. (2006) apply supervised learning to a co-authoring network using nodal features for several classification algorithms [38]. Deep-learning algorithms also help to predict network links and exhibit improved performance [39,40]. The last strand of the literature combines Markov processes and random walkers to link node types [41–43]. However, these approaches converge slowly and sometimes ambiguously [43,44].

Graph machine learning is powerful for analyzing large online social networks, and their entities [45]. This approach permits prediction, in a general context, of the appearance and disappearance of links in networks [46]. Link prediction addresses challenges in several research fields, such as healthcare and gene expression [47], business partner search [48], and social network recommendations [49]. Accounting for network structures and exogenous variables, link-prediction methods extract metrics for the likelihood of links (dis)appearances through time [50]. For instance, Kim et al. (2019) use link prediction to forecast technology convergence using Wikipedia hyperlinks and obtain statistically significant results for the 3D printing industry [51]. Lee et al. (2021) use F-terms (a patent classification code) to build a technology network and identify technology opportunities [52]. Kim and Geum (2021) develop a data-driven technology roadmap using patent data and market-trend publications to create a "keyword co-occurrence network". They then use link predictions to detect new opportunities in the road map [53]. Although these methods are well suited for large networks, their performance is similar to classic supervised methods in most cases [54].

3. Data and methods

3.1. Data

To predict the labor needs of organizations in digitalization and cybersecurity, we collect job-openings data from Indeed.com from March 2018 until December 2020 (*i.e.*, 34 months) in Switzerland (see Appendix for a detailed description) [55].[1] These job opening advertisements are sifted to identify predefined keywords related to 124 cybersecurity technologies. We generate a bi-partite network of 1805 organizations and 46 technologies. The link weights represent active job postings by one organization for a given technology. Weights vary over time. The average lifespan of a link is 2.8 months, meaning that the average job opening gets filled or deleted after slightly less than three months. Fig. 2 shows the network of job postings as of December 2020.

Cloud computing is the most linked technology throughout the study period. In second place are generally technologies related to data analysis, with a few exceptions in March and April 2018, where machine learning came second. The third place is taken by one of three technologies at each time step: machine learning, the Internet of Things, or artificial intelligence. Fig. 2 shows the diversity of technologies in job postings by organizations throughout the study period.

Besides staffing companies, the most influential organizations in our network are Roche, Novartis, and UBS, showing that large corporations have turned into technology companies not only for developing their core business but also for efficiently protecting it in cyberspace. We also collect patent information for seven technologies. Fig. 1B shows the relationship between job postings and patents per technology, which is best captured by a scaling relationship, $\text{Patents} \sim \text{Jobs}^{\mu}$ with $\mu = 0.7$ ($p < 0.01$). At the aggregate level for Switzerland and despite the very small sample ($n = 7$), this result suggests a relation between technology jobs and patent production, hence motivating further the investigation of highly skilled labor in the context of the production of IP for digitalization.

[1] https://ch.indeed.com/.

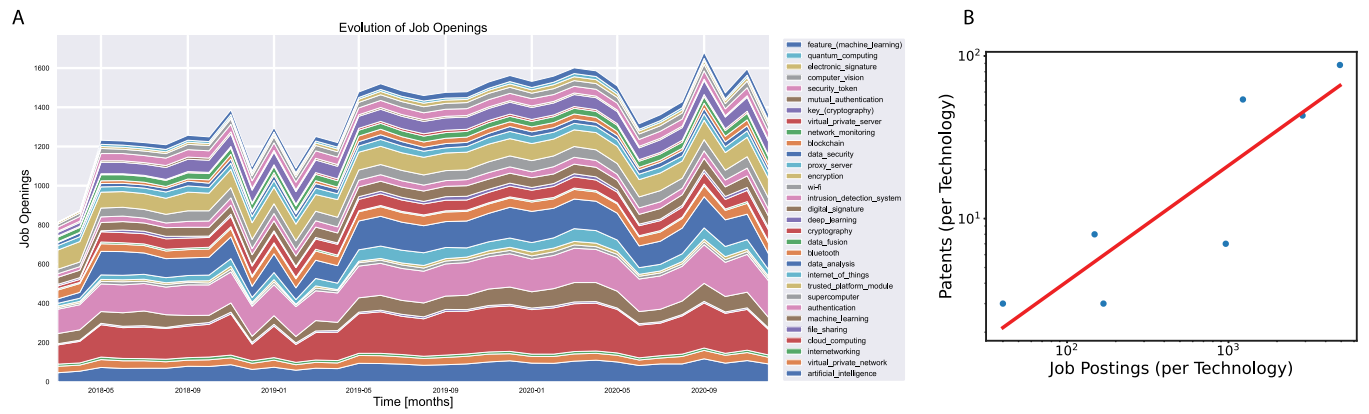

Fig. 1. A. Evolution of job openings by technology. Only technologies with more than 200 job openings are represented. B. Scaling relation between job openings and patents. Based on the evidence provided by our dataset, the Spearman rank correlation between job postings and patents in technologies is $\rho = 0.83$ ($p < 0.05$). Given the limited sample size ($n = 7$), the best relationship with could find is a scaling with $\text{Patents} \sim \text{Jobs}^{\mu}$ with $\mu = 0.7$ ($p < 0.01$).

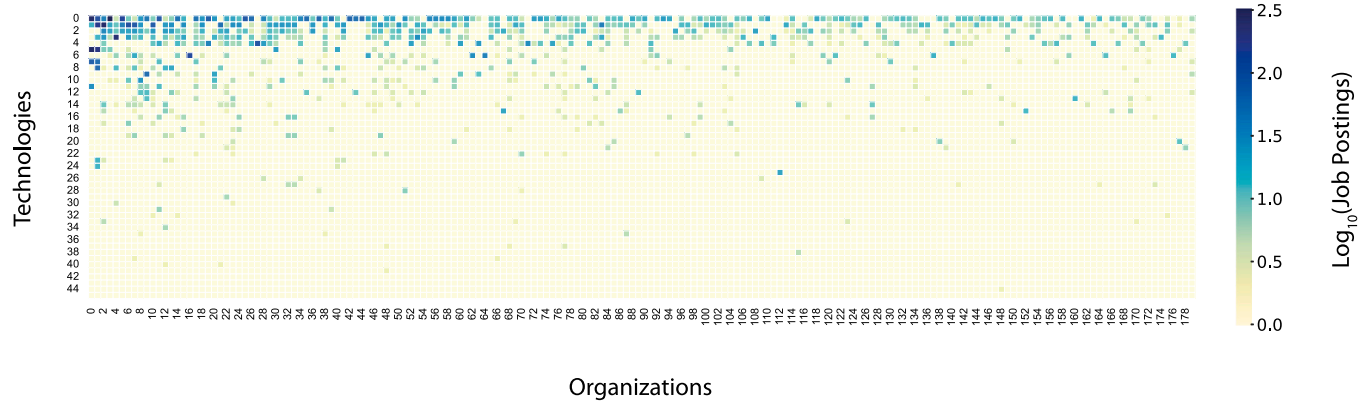

Fig. 2. Job postings by technology (in decreasing order of overall job postings) and by organizations (in decreasing order of overall job postings). Only organizations with more than 15 job openings are represented. Overall, organizations tend to post jobs in various technology fields, regardless of their size.

3.2. Method

Labor is a key economic input for the production of IP [24], and even more so for digitalization and cybersecurity [4,26]. We thus develop a graph machine learning method to forecast fine-grained labor needs in Switzerland, a small yet IP-intensive country [1,13].

We consider a prediction based on the network topology rather than by including additional features on nodes. Two reasons underly this choice. First, it remains unclear which features we could reliably gather on organizations in the context of labor and digitalization. Second, many organizations in our dataset are staffing companies (*e.g.*, Manpower, or Adecco Group), acting as proxies for other organizations and, as such, display labor needs at an aggregate level. Our contribution lies in the robustness and wide range of applications of our method that can be used in any technology mining context.

We rely on the similarity between entities in the graph and assign a score mapping each organization with technologies (proximity). We consider proximity indices because our network nodes do not carry additional variables. Proximity scores can be classified into three categories: (i) local, (ii) global, and (iii) quasi-global [46,50]. Quasi-global indices are simplified versions of global indices which do not significantly decrease computation time and that we do not consider. Instead, we adapt and combine local and global similarity indices to match the constraints of the bi-partite network [35], which is defined as $G = (V, E)$. V is any finite set called the node set and $E \subseteq V \times V$, corresponds to links between elements of V called the link set. Let $x, y \in V$, such as:

- the neighborhood of x is $\Gamma(x) = \{y \in V \, s.t. \, (x, y) \in E\}$;
- the degree of x is $\delta_x = |\Gamma(x)|$;
- there is a path between x and y if there exists $(x_0, x_1, \ldots, x_n)$ such that $x_0 = x$, $x_n = y$ and $(x_i, x_{i+1}) \in E \; \forall \, 0 \leq i \leq n-1$;
- a graph G is said to be bi-partite if there exists $A, B \subset V$ such that if $(x, y) \in E$ then x and y are not in the same subset A, B;
- Let $|V| = n$, $A \in \mathbf{R}^{n \times n}$ is an adjacency matrix of G if and only if $\forall x, y \in V \, A_{x,y} = 1$ implies $(x, y) \in E$ and $A_{x,y} = 0$ otherwise.

V corresponds to organizations and technologies and E to links between them. As our model focuses on the relationships between organizations and technologies, the links can only link one organization to a technology. We define $\mathbf{G}$ as the set containing all the graphs modeled on the time dimension: $\mathbf{G}_0, \mathbf{G}_{33} \in \mathbf{G}$ are graphs representative of the network in March 2018, and December 2020, respectively. We define $\mathbf{G_{i-j}}$ with $i < j \in 0, 1, \ldots, 32$ as the subset of $\mathbf{G}$ that contains all graphs starting at $\mathbf{G}_i$ and ending $\mathbf{G}_j$ (both ends included). In the following sections, $\mathbf{G}_{i-j}$ will also be referred to as a set of training graphs.

3.2.1. Local indices

We build local indices on the immediate neighborhood of two nodes. Local indices do not carry information on the global structure of the graph to make them more tractable. We consider the following local indices: *Common Neighbors, Jaccard index, Sorensen index, Adamic–Adar coefficient, Preferential Attachment index, Resources Allocation index, and Salton index*. Unfortunately, most of these indices depend on the intersection of the neighborhood of two network nodes. the bi-partite network formalism imposes that the neighborhood between two entities

(an organization and a technology) is always empty. To avoid a significant loss of graph structural information, we depart from Benchettera et al. (2010) [35] who project the bi-partite network on two mono-partite networks of (i) organizations and (ii) technologies. Thus, the local index, which does not depend on the neighborhood intersection, is the Preferential Attachment (PA) index and depends on the node degree (*i.e.*, the number of neighbors).

3.2.2. Global indices

Global indices rely on the entire network structure, *i.e.*, the ensemble of paths between two entities. Two entities connected through several paths will likely connect directly in the future. However, long paths (many internodes between two nodes of interest) are less influential than shorter paths. The computation time for these indices is an exponential function of network size. We consider the following global indices: Katz, Leicht–Holme–Newman, Average Commute, Hyperbolic Sine, and Simrank. Besides Katz and Hyperbolic Sine (HS), these indices rely on randomization algorithms to reduce computation time. In the literature [46,50], Katz performs better than the other indices on several networks. We also test these indices on our data set and find results on par with the literature. Although Katz is theoretically computationally intensive, we reduce the computing cost with linear algebra algorithms optimized for sparse matrices implemented in Phython [56]. The Hyperbolic Sine index is specifically designed for considering only paths of an odd size, which are the only possible paths between two nodes of the same type on a bi-partite network.

3.2.3. Combining information from local and global indices

We consider that information from local and global indices matters for job openings prediction by organizations in digital technologies. We, therefore, explore how to combine these indices as features in several supervised learning models. We consider the following algorithms: Logistic Regression, Decision Trees, and Support Vector Machine (SVM). Logistic regression does not allow non-linear features and was discarded. Decision trees and SVM are more appropriate for handling non-linear features. However, the "kernel trick" offered by SVM brings more flexibility and allows more nuanced results because similarity indices are easily separable in a vector space [57].[2] They also provide better diagnostic capabilities on the relative importance of features in the prediction. Moreover, decision trees are often used for categorical features, which is not the case here. We test our design assumptions with these two models. SVM with Radial Basis Function brings significantly better results than a decision tree. We subsequently test a random forest to match SVM predictions. However, the size of the resulting random forest grows too large and introduces a significant over-fitting risk.

We thus select PA as the local index. We also select the Katz and HS as global indices. We define these three indices as,

(1) ***Preferential Attachment Index*** [58] is a local index that assumes that the higher the degree of the two nodes, the higher the likelihood of them connecting. It is defined as,

$$S_{xy}^{PA} = \delta_x \cdot \delta_y, \tag{1}$$

where δ_x and δ_y are the degrees of respectively x and y, respectively, organization and technology nodes (see Appendix).

(2) ***Katz Index*** [59]: Given $x, y \in V$, this global index counts the number of paths between x and y. The Katz Index is defined as,

$$S_{xy}^{Katz} = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{xy}^{\langle l \rangle}| = \beta A_{xy} + \beta^2 (A^2)_{xy} + \beta^3 (A^3)_{xy} + \cdots, \tag{2}$$

where β is the damping coefficient in $[0, 1]$, $paths_{xy}^{\langle l \rangle}$ is the set of all paths with length l connecting x and y, and A is the adjacency matrix of the graph (see Fig. 2). It should be noted that β should be smaller than $\frac{1}{\lambda_{max}}$ for the series to converge where λ_{max} is the largest eigenvalue of A. The Katz index is a weighted sum of the number of paths of different lengths, with the weights decreasing exponentially for longer paths.

(3) ***Hyperbolic Sine Index*** [60] is similar to the Katz Index, but with its weights are related to the exponential weights. In matrix form, it is defined as,

$$S^{HS} = \sinh(\alpha A) = \sum_{l=0}^{\infty} \frac{\alpha^{1+2i}}{(1+2i)!} A^{1+2i}, \tag{3}$$

with α being a parameter weighting the influence of long-range paths versus short-range paths. We set α to 0.01, which does not evolve significantly through optimization.

3.2.4. Graph learning pipeline

Link prediction is a highly unbalanced classification problem, with non-existing links far outnumbering existing ones in the overall network structure. We implement a graph learning pipeline, which is robust to this major constraint. Fig. 3 shows the supervised learning pipeline, with the training set, which first embeds the network measures: (i) PA index, (ii) Katz Index (Katz), and (iii) HS index. The probabilities of link appearance through each measure can either be (i) validated separately on the graph or be combined and used as features in the SVM to generate an aggregate prediction, which can be validated in turn on the graph.

Given a set of training graphs $\mathbf{G}_{i-j}$ and a forecast range t (in our case $1 < t < 6$ for t months), the algorithm predicts the existence or non-existence of a link between each organization and each technology in $\mathbf{G}_{j+t}$. We use the graphs in a certain time range to predict the state of the graph t months later. One constraint is that $j + t$ must be smaller than 33, as our dataset is limited to a time series of 33 observations (see Section 3.1). Each pair of nodes x and y is given a score s_{xy}, which is defined as the similarity between the nodes. The greater s_{xy} is, the higher the likelihood of an existing link between x and y. Given scores s_{xy} for $x, y \in V$ and a threshold θ, we predict the existence of a link between the two nodes if $s_{xy} \geq \theta$. If $s_{xy} < \theta$, this link does not exist between two nodes. We obtain the threshold θ by maximizing a simple function of the True Positive Rate (TPR) and False Positive Rate (FPR). We optimize the difference between TPR and FPR, using the standard Youden's J statistic [61], which is similar to other threshold-moving techniques [62], such as the geometric mean of sensitivity and specificity, as well as the F-score of precision and recall [62]. We compute the similarities for each graph present within $\mathbf{G}_{i-j}$ and use this data to forecast the similarities for the graph $\mathbf{G}_{j+t}$ using several forecasting models, such as PA, Katz, HS on the one hand (three models) and on the other hand PA, Katz and HS together as input features for SVM (one model) see Fig. 3. In the latter case, we use these three indices as additional features for our supervised classification algorithm. Each link is represented as a feature vector consisting of four values: the three indices presented above and the number of job openings linking each pair of nodes consisting of an organization and a technology. We use all possible links in the graphs present in $\mathbf{G}_{i-j}$ as training sets for the SVM using the computed scores and job-openings data. We have a series of score matrices S_t^{ind} $i < t < j$ and $ind \in \{Katz, PA, HS\}$ for each similarity measure presented above, as well as a trained SVM model. With the series of scores for each link S_{xy}^t, we can train a time-series model to forecast the score τ steps (*i.e.*, months) ahead [63].[3] This approach has proven more robust for graphs than linear regression and Autoregressive Integrated Moving Average (ARIMA) models [36]. With the forecast measures and a baseline obtained by the final $\mathbf{G}_j$ in $\mathbf{G}_{i-j}$, we establish the confusion matrix for the prediction of $\mathbf{G}_{j+t}$ for the four graph machine learning models.

[2] The kernel trick allows to project data in a different space, in which they exhibit statistical behaviors closer to linearity.

[3] We omit the *ind* for clarity, but we do this for each similarity measure.

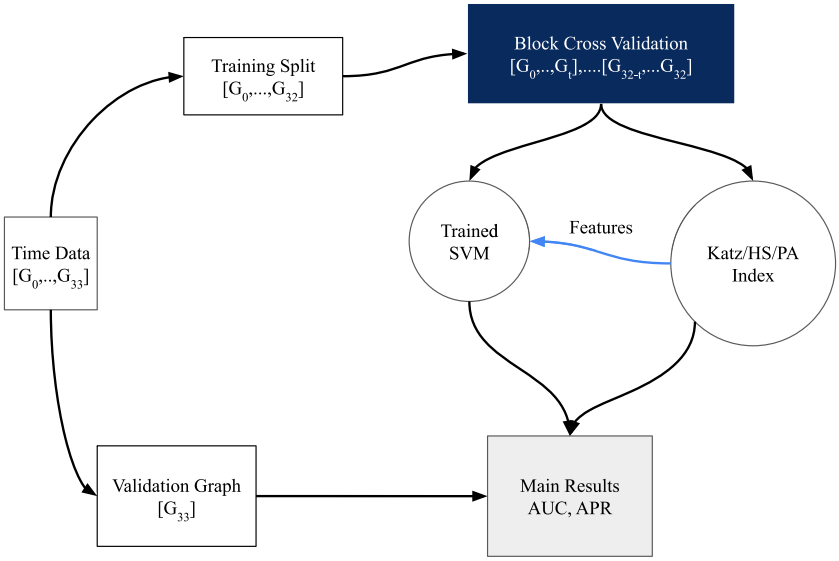

Fig. 3. This figure represents the algorithm from training to validation. Data is split between the training and validation sets. Using block cross validation, the probabilities of link appearance through each measure (*i.e.*, PA, Katz, HS) can then either be (i) validated separately on the validation graph or be combined and used as features in the SVM to generate an aggregate prediction, which can be validated in turn on the validation graph.

3.2.5. Block cross-validation

To overcome the problem of unbalanced data, we use an under-sampling method to pick the negative class (no job opening) compared to the positive class (job opening or closing) to ensure a balanced dataset. We implement blocked cross-validation with different training set sizes and forecast ranges among our different models [64]. Given a block size i, *i.e.*, the number of months used for training, and a forecast range j. We partition our dataset in $trunc(34/(i-j),0)$ distinct sets. The first block contains the i graphs used for training and the validation graph $i+j$. The indices (Katz, PA, HS) are computed for each graph in the training set, and the parameters are optimized for this block (Young statistics). We obtain scores associated with the graph $i+j$ by linear regression using the scores obtained during training. We then optimize parameters and the scores to compute performance on graph $i+j$, namely Receiver Operating Characteristic (ROC), Area Under the Curve (AUC), and Precision-Recall (PR), which can be measured as a scalar through the Average Precision (AP) (see Appendix, and [65,66]). For SVM, all indices are used as parameters, and the model is trained on all graphs of the training block. We subsequently use the indices to predict graph $i+j$ and compute the precision. We iterate this process on all blocks and average the results computed over the $trunc(34/(i-j),0)$ blocks. We repeat the procedure for values $i = \{2,3,4,5,6\}$ and $j = \{1,2,3,4,5,6\}$. The main tools used were Phython and its associated packages such as NetworkX [67], and scikit-learn.

4. Results

Over a 34 month period, we predict job openings for one-month to six-month horizons. We perform prediction using blocked cross-validation for training sizes of 2, 3, 4 and 6 months. Additionally, we predict using the maximum available training set, given the forecasting range. Fig. 4A shows the mean area under the ROC Curve (AUC) for PA, Katz, and HS indices as well as for the SVM approach combining the local (PA) and global (Katz, HS) indices for the prediction of the last month of our dataset ($\mathbf{G}_3 3$) from $\mathbf{G}_{0-32}$. AUC is high >0.95 for all models, albeit slightly worse for PA. Fig. 4B shows the PR curve of the same models. The AP of the SVM (= 0.905) is superior to that of HS ($AP = 0.846$), and Katz ($Katz = 0.835$) models. However, in terms of PR, PA performs badly ($AP = 0.364$). Fig. 4C shows the results for predictions of the last month in the dataset ($\mathbf{G}_3 3$) from $\mathbf{G}_{0-27}$ (*i.e.*, prediction of job-openings six months ahead). AUC is almost equal for HS ($AUC = 0.863$), Katz ($AUC = 0.863$) and SVM ($AUC = 0.864$) and slightly worse for PA ($AUC = 0.848$).

Fig. 4D shows the PR curve. In this case, the SVM performs much better than the other models ($AP = 0.543$ for SVM versus $AP = 0.356$ for Katz and HS and $AP = 237$ for PA), suggesting that combining local and global network indices is relevant for long forecast ranges. Fig. 5 shows the evolution of mean AUC as a function of forecast ranges. The AUC remains consistent across models as a function of the forecast range. We provide a detailed account of AUC for all forecasting ranges in Appendix, Table 3.

We find that global network indices (Katz and HS) capture remarkably well job-opening predictions for short-term horizons. However, the local index (PA) does not perform well and improves only marginally the model for short-term horizons. For long-term horizons, the SVM model that combines local (PA) and global (Katz, HS) indices perform much better than models with individual indices, in particular in terms of AP (PR curve).

Lastly, we want to test if the results (*i.e.*, the median AUC) of our different machine-learning classifiers display statistically significant differences. To reach this purpose, we can use a non-parametric statistical test, such as the Kruskal–Wallis test, to compare the mean AUC of multiple classifiers [68]. This test is a non-parametric version of one-way ANOVA, and it can be used to test the Null hypothesis (H0) that the population medians (in our case, AUCs) of all models are equal. The Kruskal–Wallis test is appropriate for comparing more than two groups when the dependent variable is continuous and the distributions are not normal. Hence,

Null hypothesis, H0: *The population medians of the AUC values for all models are equal.*

Alternative hypothesis, Ha: *At least two population medians of the AUC values are different.*

The respective median AUCs are obtained by extracting the results of each cross-validation iteration. Then, we perform the Kruskal–Wallis test for each training size and forecast range. Table 5 shows the results ($p-values$) for each combination of training size and forecast range. Overall, we notice that the more we increase the training size and forecast range, the more the Null Hypothesis (H0) is rejected. As a conclusion, we can say that the differences in results of our machine-learning classifiers are indeed statistically significant.

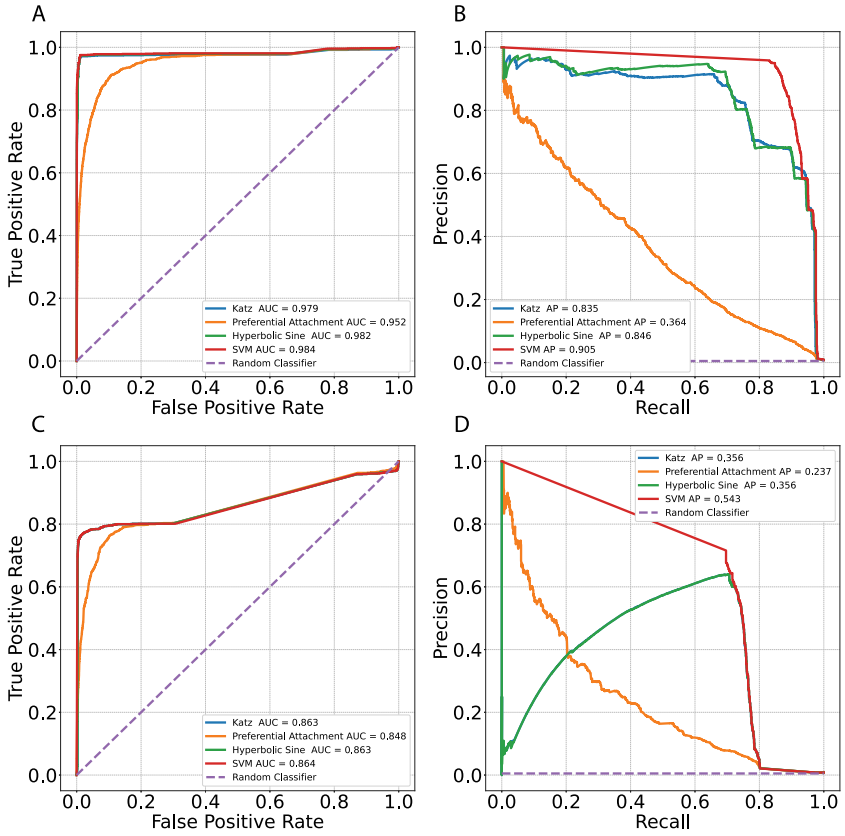

Fig. 4. A ROC curve for prediction the last month $\mathbf{G}_{33}$ with data from $\mathbf{G}_{0-32}$. The AUC is high for all individual indices (PA, Katz, and HS) and for the combined SVM model. B Corresponding PR curves for one-month forecast range. SVM performs much better than the models based on individual indices. C ROC curve for predicting the last month $\mathbf{G}_{33}$ with data from $\mathbf{G}_{0-27}$. The AUC is high for all individual indices (PA, Katz, and HS) and for the combined SVM model. D Corresponding PR curves for six months forecast range. SVM performs much better than the models based on individual indices.

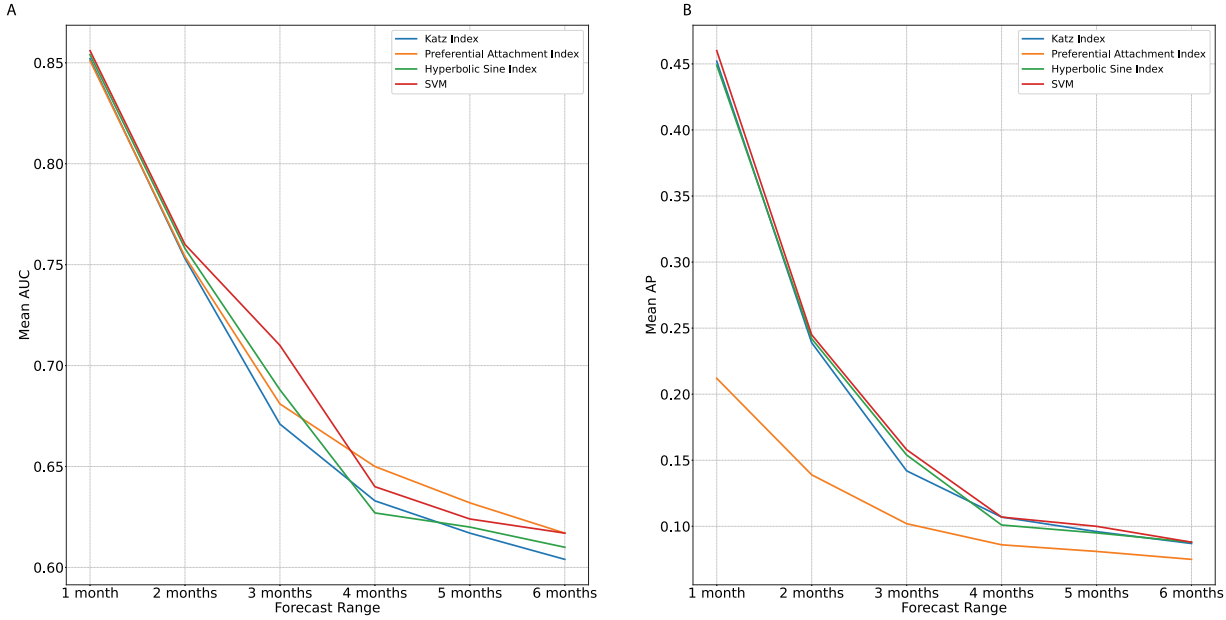

Fig. 5. A. Mean AUC as a function of forecast range, obtained through block cross-validation using training sets of size 6. The standard deviation is between 0.03 and 0.07. B. Mean AP as a function of forecast range, obtained through block cross-validation using training sets of size 6. The standard deviation is between 0.02 and 0.08 and the threshold of significance for AP is 0.01.

5. Discussion

High-skilled workers are the main source of IP. For digital technologies and cybersecurity, labor forces have been in scarce supply for decades, impeding the sustainable progress of digital technologies globally. This scarcity of talent calls for better management. To address this issue, we propose a graph machine-learning approach to predict labor needs in digital technologies. The robustness of our approach lies in the bi-partite representation of the network [35], which has yielded significant results for investigating productive capabilities at the country level [41,42] as well as for the production of open source knowledge [43]. To capture best the evolution of the bi-partite network of 1805 organizations opening (resp. closing) job-openings in 124 digital technologies (second node type) over a 2.5 years period, we select local (PA) and global (Katz and HS) network metrics. We use these metrics to predict job openings, individually or in combination, using SVM. Our results show that global indices capture job openings the best, while the local index significantly under-performs. This supports the view that the network's global properties are essential for the prediction. Our model learns primarily from the job openings of all organizations to predict those of individual ones. The network structure supports the evidence that organizations face similar labor needs in terms of digitalization (see Fig. 2). However, when considering long-term horizons (*i.e.*, six months), combining global indices with a local index (*i.e.*, PA, Katz, HS indices combined with SVM) yields the best predictions. This result suggests that the organizations' peculiarities must be considered when envisioning the longer-term evolution of digital skills.

These results are important for management and policy-making as high-skilled labor has become crucial for the IP production [12], and as organizations and countries heavily compete in digitalization [1,6,8,12]. Although our data do not allow us to infer any causal link between job openings and IP, we do contribute with a preliminary identification of their commonalities. Moreover, our approach permits prediction at a resolution including each organization and technology, and at a monthly frequency. In the business context, predicting labor needs in a specific technology at the aggregate level or anticipating the competitors' next move provides invaluable information to adapt business strategies. Furthermore, in countries with an endemic labor shortage of highly skilled workers, such as Switzerland, knowing in advance job openings by competitors may provide organizations with a first-mover advantage and facilitate talent recruitment. Conversely, talents could use prediction of job openings to adapt their job market strategies, *e.g.*, by adopting "option to wait for" strategy [69].

Finally, the competitiveness of nearly all countries relies on the mastery of digital technologies. It is the responsibility of policymakers to ensure that the labor is adequately supplied. Thus, our approach would help them to set higher incentives for people to up-skill their profile and adapt to the demand [70]. Along with the monetary policy, labor is one of the most important macroeconomic variables for developing economies. With more complete data across industrial sectors, our research contribution could be extended to develop labor market forecasts enabling targeted interventions, which predicted effects could, in turn, be measured.

6. Conclusion

Intellectual property is mostly produced through the labor of highly skilled people. This labor force has notoriously been a scarce resource for decades [1,2] and has been the subject of an increasing rivalry between organizations [6] and even Nation states [3–5]. Using a unique dataset of job openings in Switzerland between 2018 and 2020 and by developing a link prediction method in a bi-partite network of technologies and organizations, we find that job openings by organizations in digital technologies are predictable up to the six-month horizon. Although a horizon of a few months is surely not enough to adapt to the highly skilled labor market of digitalization (*e.g.*, through education), gaining labor market predictability helps to monitor the allocation of human capital in countries such as Switzerland, which increasingly rely on the production of intellectual property, including patents, to ensure their global competitiveness.

CRediT authorship contribution statement

Dimitri Percia David: Conceptualization, Methodology, Investigation, Visualization, Validation, Software, Writing – original draft, Writing – review & editing. **Santiago Anton Moreno:** Data curation, Methodology, Visualization, Investigation, Software, Writing – original draft, Writing – review & editing. **Loïc Maréchal:** Investigation, Writing – review & editing. **Thomas Maillart:** Supervision, Conceptualization, Writing – review & editing. **Alain Mermoud:** Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This document is the result of a research project funded by the Cyber-Defence (CYD) Campus, armasuisse S+T. Dimitri Percia David was funded through the Swiss Confederation DDPS Research Contract on Aramis: CYD-C-2020015.

Appendix

Detailed description of the dataset

We use a dataset that was collected to build a technology-mining project entitled *Technology & Market Monitoring* (TMM). TMM is developed by armasuisse Science and Technology (S + T), the Swiss Federal Office for Defence Procurement of Switzerland. TMM is designed to gather automatically, process, and exploit open-source information for intelligence purposes. The TMM system crawls and aggregates information from different online resources such as commercial registers (*Zefix*[4]), websites (*Wikipedia*, and *Indeed*[5]) to obtain a list of organizations and job openings based in Switzerland.

TMM compiles job openings monthly from March 2018 to December 2020. These job openings are sifted to identify keywords associated with certain technologies. A list of predefined keywords related to cybersecurity technologies is used to compute word similarity with the technologies coded in the TMM database. Such a list is defined and exploited by the TMM platform and uses a *key-BERT* approach. For additional information, please get in touch with the authors. We then use the difflib[6] library in Phython to obtain good matches with the TMM data. We verify the obtained list afterward to delete any irrelevant matches, and thus we obtain 124 keywords[7] from TMM. These 124 keywords correspond to 46 technologies (see Appendix, Table 1). By using the organization list and job openings data, we link

[4] https://www.zefix.ch.
[5] https://indeed.com.
[6] https://tinyurl.com/8wvkfha2.
[7] Keywords link https://tinyurl.com/jswtsmmn.

Table 1
Table with Job openings and patents by technology between March 2018 and December 2020.

Technology	Job openings	Patents
Cloud computing	4994	88
Data analysis	2921	43
Artificial intelligence	1283	54
Machine learning	1279	–
Internet of things	969	7
Trusted platform module	634	–
Blockchain	548	–
Virtual private network	313	–
Supercomputer	204	–
Proxy server	195	–
Intrusion detection system	179	–
Authentication	168	3
Cryptography	156	8
Wi-fi	118	–
Deep learning	116	–
Computer vision	114	–
Penetration test	85	–
Predictive analytics	63	–
Encryption	46	–
Quantum computing	43	–
Virtual private server	43	–
Bluetooth	43	3
Data security	39	–
Reinforcement learning	31	–
Feature (machine learning)	30	–
Padding (cryptography)	27	–
5g	27	–
Electronic signature	24	–
Digital signature	20	–
Strong authentication	15	–
Deep packet inspection	14	–
Internetworking	11	–
https	11	–
ntfs	11	–
Data fusion	8	–
File sharing	8	–
Network monitoring	8	–
mqtt	7	–
Adversarial machine learning	4	–
Distributed ledger	3	–
Distributed algorithm	2	–
Key (cryptography)	1	–
Federated search	1	–
Security token	1	–
Deep linking	1	–
Closed-loop authentication	1	–

Table 2
Job openings and patents by the organization between March 2018 and December 2020. All organizations with at least one patent are reported, whereas only organizations with 65 job openings are reported.

Organization	Job openings	Patents
Universal-Job AG	627	–
Manpower SA	613	–
Roche AG	400	–
Adecco Group AG	378	–
Quaker United Nations Office Geneva Association	264	–
Careerplus SA	188	–
Novartis AG	186	–
Universitätsstiftung Basel	168	–
UBS AG	156	9
Arobase SA	151	–
Philip Morris International Management SA	151	–
yellowshark AG	145	–
ETH Zürich SEC AG	118	–
EPFL-WISH FOUNDATION	107	–
Marco R. Fuhrer Unternehmensberatung	105	–
Deloitte AG	105	–
BDO AG	105	–
myitjob GmbH	104	–
ABB Ltd	104	–
Honeywell AG	101	17
...	...	...
Accenture AG	65	4
...	...	...
Ericsson AG	–	12
INGRAM MICRO GmbH	–	10
Bayer Consumer Care AG	–	9
Infosys Limited, Bangalore, Zweigniederlassung ...	–	7
Riverbed Technology AG	–	6
salesforce.com Sàrl	–	6
Bayer GmbH	–	10
JPMorgan Chase Bank, National Association, Columbus, Zurich Branch	–	5
Wipro Limited, Bangalore, succursale de Genève	–	4
Garrett Motion Sàrl	–	3
Kudelski S.A.	–	3
Dell SA	–	3
Caterpillar SARL	–	2
TOPREX AG	–	2
Cyberlink AG	–	2
Pharma Development AG	–	1
Thales Suisse SA	–	1
Box GmbH	–	1

organizations to certain technologies thanks to the keywords present in the job postings. This results in a bi-partite network with 1712 nodes of *organization* type and 46 nodes for technologies (see Fig. 2 for a partial representation of the bi-partite network as a heatmap) consisting of 14,819 links over the study period. For each organization, we also sum up the number of technologies mentioned in their posted job openings and use this metric as an additional link attribute in the created graph. The average lifespan of a job opening is 2.8 months, meaning that the average job opening gets filled or deleted after slightly less than 3 months. Cloud computing is the most linked technology throughout the study period. In second place are generally technologies related to data analysis, with a few exceptions in March and April 2018, where machine learning came second. The third place is taken by one of three technologies at each time step: machine learning, the Internet of Things, or artificial intelligence. The most influential organizations in our network besides staffing companies (*e.g.*, Universal-Job AG, Manpower SA, Adecco Group AG) are Roche, Novartis, and UBS, showing that large corporations have turned into technology companies, not only for developing digitalization for their core business through but also possibly for efficiently protecting their business in the cyberspace.

Similarly, we have used 206 patents released by Swiss organizations in the same technology fields and compiled by TMM over the same study period.

Job openings and patents by technology

See Table 1.

Job openings and patents by company

See Table 2.

Detailed results of AUC and PR

We use the ROC curve as a graphical representation of performance [65,66]. The ROC curve plots the TPR against the FPR. The metric we seek to optimize is the AUC. The AUC takes values between 0 and 1 and indicates how well a model can separate the two output classes. The closer the AUC is to 1, the better the model predicts both existing and non-existing links. Because the AUC measures the performance of the model for both possible outputs, it is less affected by unbalanced datasets. We also compute PR curves, where $Precision =$

Table 3
AUC. Mean AUC was obtained through blocked cross-validation, according to the training set size and forecasting ranges. The thresholds used to compute accuracy are obtained by optimizing Younden's $Jstatistic$. The standard deviation of those means is between 0.03 and 0.07. The best AUC for a given forecast range is highlighted in red.

Method	Training size	Forecast range					
		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
PA index	2 months	0.818	0.709	0.642	0.610	0.587	0.576
	3 months	0.838	0.734	0.652	0.617	0.599	0.588
	4 months	0.843	0.736	0.657	0.620	0.601	0.594
	6 months	0.851	0.754	0.681	0.650	0.632	0.617
Katz index	2 months	0.832	0.731	0.638	0.618	0.595	0.596
	3 months	0.837	0.732	0.655	0.617	0.601	0.598
	4 months	0.844	0.728	0.649	0.615	0.601	0.587
	6 months	0.852	0.753	0.671	0.633	0.617	0.604
HS index	2 months	0.832	0.738	0.669	0.644	0.629	0.627
	3 months	0.837	0.734	0.663	0.633	0.625	0.618
	4 months	0.844	0.727	0.65	0.623	0.615	0.606
	6 months	0.854	0.758	0.688	0.627	0.62	0.61
SVM	2 months	0.836	0.740	0.674	0.650	0.635	0.629
	3 months	0.838	0.740	0.667	0.638	0.629	0.621
	4 months	0.849	0.733	0.657	0.628	0.618	0.609
	6 months	0.856	0.760	0.710	0.640	0.624	0.617

Table 4
AP obtained through blocked cross-validation, according to the training set size and forecasting ranges. The thresholds used to compute accuracy are obtained by optimizing Younden's $Jstatistic$. The standard deviation of those means is between 0.02 and 0.08. The threshold of significance for AP is 0.01. The best AP for a given forecast range is highlighted in red.

Method	Training size	Forecast range					
		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
PA index	2 months	0.214	0.135	0.078	0.063	0.052	0.047
	3 months	0.230	0.140	0.088	0.070	0.060	0.055
	4 months	0.220	0.137	0.090	0.073	0.065	0.061
	6 months	0.212	0.139	0.102	0.086	0.081	0.075
Katz index	2 months	0.530	0.312	0.136	0.099	0.082	0.076
	3 months	0.560	0.271	0.136	0.100	0.087	0.080
	4 months	0.502	0.240	0.127	0.097	0.084	0.076
	6 months	0.452	0.239	0.142	0.107	0.096	0.087
HS index	2 months	0.530	0.317	0.146	0.107	0.092	0.085
	3 months	0.560	0.273	0.140	0.106	0.094	0.087
	4 months	0.501	0.241	0.129	0.100	0.090	0.081
	6 months	0.449	0.242	0.154	0.101	0.095	0.088
SVM	2 months	0.513	0.295	0.146	0.110	0.094	0.089
	3 months	0.514	0.283	0.141	0.107	0.093	0.086
	4 months	0.479	0.256	0.127	0.102	0.091	0.082
	6 months	0.460	0.245	0.158	0.107	0.100	0.088

$\frac{tp}{tp+fp}$ and $Recall = tpr$, as well as AP. The AP of a random classifier should, on average, be equal to $\frac{P}{P+F}$ and the AUC of such a classifier should be around 0.5. The problem with AP as a performance measure is that it is incompatible with cross-validation since the scale of AP varies greatly depending on the distribution of the test set and, thus, is sometimes hard to interpret. We compute the ROC Curve, TPR, FPR, AUC, PR, and AP to the full test set, as well as to a randomized, balanced test set. In all cases, we obtain a numerical result equal to up to the third decimal place (see Table 4).

Table 5
The $p-value$ represents the probability of observing a test statistic as extreme or more extreme than the observed test statistic, assuming the Null hypothesis (H0) is true. If the $p-value$ is less than a significance level (alpha; typically * set at 0.1, ** set at 0.05, and *** set at 0.01), we reject the Null hypothesis and conclude that there is a significant difference in performance between the models.

Training size	Forecast range	p-value
2	1	0.25
2	2	0.39
2	3	0.13
2	4	0.12
2	5	0.15
2	6	0.04**
3	1	0.13
3	2	0.20
3	3	0.20
3	4	0.20
3	5	0.07*
3	6	0.05**
4	1	0.15
4	2	0.13
4	3	0.02**
4	4	0.02**
4	5	0.06*
4	6	0.06*
6	1	0.23
6	2	0.11
6	3	<0.01***
6	4	<0.01***
6	5	<0.01***
6	6	0.013**

References

[1] Spyros Arvanitis, Computerization, workplace organization, skilled labour and firm productivity: Evidence for the Swiss business sector, Econ. Innov. New Technol. 14 (4) (2005) 225–249.

[2] Timothy F. Bresnahan, Erik Brynjolfsson, Lorin M. Hitt, Information technology, workplace organization, and the demand for skilled labor: Firm-Level evidence*, Q. J. Econ. 117 (1) (2002) 339–376.

[3] George Emm. Halkos, Nickolaos G. Tzeremes, International competitiveness in the ict industry: Evaluating the performance of the top 50 companies, Glob. Econ. Rev. 36 (2) (2007) 167–182.

[4] Dominik Paulus-Rohmer, Heike Schatton, Thomas Bauernhansl, Ecosystems, strategy and business models in the age of digitization - How the manufacturing industry is going to change its logic, Proc. CIRP 57 (2016) 8–13.

[5] Marie-Sophie Denner, Louis Christian Püschel, Maximilian Röglinger, How to exploit the digitalization potential of business processes, Bus. Inf. Syst. Eng. 60 (4) (2018) 331–349.

[6] Rajà Attia, Isabelle Davy, Roland Rizoulières, Innovative labor and intellectual property market in the semiconductor industry, in: Bernard Guilhon (Ed.), Technology and Markets for Knowledge: Knowledge Creation, Diffusion and Exchange Within a Growing Economy, in: Economics of Science, Technology and Innovation, Springer US, Boston, MA, 2001, pp. 137–180.

[7] C.I. Orji, Digital business transformation: Towards an integrated capability framework for digitization and business value generation, J. Glob. Bus. Technol. 15 (1) (2019) 47–57.

[8] Briana Boland, Kevin Dong, Jude Blanchette, Ryan Hass, How China's Human Capital Impacts Its National Competitiveness, Technical report, Center for Strategic and International Studies, 2022, p. 15.

[9] T. Casey Fleming, Eric L. Qualkenbush, Anthony M. Chapa, The secret war against the united states:The top threat to national security and the American dream cyber and asymmetrical hybrid warfare an urgent call to action, Cyber Def. Rev. 2 (3) (2017).

[10] Alex Wilner, Cyber deterrence and critical-infrastructure protection: Expectation, application, and limitation, Comp. Strategy 36 (4) (2017) 309–318.

[11] Nabin Chowdhury, Vasileios Gkioulos, Cyber security training for critical infrastructure protection: A literature review, Comp. Sci. Rev. 40 (2021) 100361.

[12] Tommaso De Zan, Mind the Gap: The Cyber Security Skills Shortage and Public Policy Interventions, Policy Report, Global Cyber Security Center Report - University of Oxford, 2019.

[13] Jean-Pierre Jeannet, Thierry Volery, Heiko Bergmann, Cornelia Amstutz, Leveraging local competitiveness, in: Jean-Pierre Jeannet, Thierry Volery, Heiko Bergmann, Cornelia Amstutz (Eds.), Masterpieces of Swiss Entrepreneurship: Swiss SMEs Competing in Global Markets, Springer International Publishing, Cham, 2021, pp. 235–255.

[14] Lance Hoffman, Diana Burley, Costis Toregas, Holistically building the cybersecurity workforce, IEEE Secur. Priv. 10 (2) (2011) 33–39, Publisher: IEEE.

[15] Harry Scarbrough, Knowledge management, HRM and the innovation process, Int. J. Manpow. (2003) Publisher: MCB UP Ltd.

[16] Dimitri Percia David, Marcus Matthias Keupp, Ricardo Marino, Patrick Hofstetter, The persistent deficit of militia officers in the Swiss Armed Forces: An opportunity cost explanation, Def. Peace Econ. 30 (1) (2019) 111–127, Publisher: Taylor & Francis.

[17] Dimitri Percia David, Three Articles on the Economics of Information-Systems Defense Capability. Material-, Human-, and Knowledge-Resources Acquisition for Critical Infrastructures (Ph.D. thesis), Université de Lausanne, Faculté des hautes études commerciales, 2020.

[18] Julian Jang-Jaccard, Surya Nepal, A survey of emerging threats in cybersecurity, J. Comput. System Sci. 80 (5) (2014).

[19] John J. Horton, The effects of algorithmic labor market recommendations: Evidence from a field experiment, J. Labor Econ. 35 (2) (2017) 345–385, Publisher: University of Chicago Press Chicago, IL.

[20] D.S. Levine, T. Sichelman, Why do startups use trade secrets, Notre Dame Law Rev. 94 (2) (2018) 751–820.

[21] B.H. Hall, M. MacGarvie, The private value of software patents, Res. Policy 39 (7) (2010) 994–1009.

[22] J.P. Johnson, Open source software: Private provision of a public good, J. Econ. Manag. Strategy 11 (4) (2002) 637–662.

[23] William Fisher, Theories of intellectual property, 2001, p. 30, Available At: https://Cyber.Harvard.Edu/People/Tfisher/Iptheory.Pdf.

[24] Bryan Cwik, Labor as the basis for intellectual property rights, Ethical Theory Moral Pract. 17 (4) (2014) 681–695.

[25] Max Bankewitz, Carl Aberg, Christine Teuchert, Digitalization and boards of directors: A new era of corporate governance? Bus. Manag. Res. 5 (2) (2016) p58.

[26] Shahar Markovitch, Paul Willmott, Accelerating the digitization of business processes, Technical report, Mc Kinsey & Company, 2014, p. 4.

[27] Michel Ferrary, Market for competences: When attractiveness drives competitiveness, in: Managing Competences, Taylor & Francis, 2020.

[28] Ron Sanchez, Joseph Mahoney, Modularity, flexibility, and knowledge management in product and organization design, Strateg. Manag. J. 17 (1996) 63–76, URL http://dx.doi.org/10.2307/2486991.

[29] T. Maillart, D. Sornette, S. Spaeth, G. von Krogh, Empirical tests of Zipf's law mechanism in open source linux distribution, Phys. Rev. Lett. 101 (21) (2008) 218701+, URL http://scitation.aip.org/getabs/servlet/GetabsServlet{p}rog=normal&id=PRLTAO000101000021218701000001&idtype=cvips&gifs=yes.

[30] Lillian Ablon, Martin C. Libicki, Andrea A. Golay, Markets for Cybercrime Tools and Stolen Data: Hackers' Bazaar, Rand Corporation, 2014.

[31] John Forth, Do ICT Skill Shortages Hamper Firms Performance?, Technical Report 281, National Institute of Economic and Social Research, 2006.

[32] Angus C. Chu, Shiyuan Pan, Minjuan Sun, When does elastic labor supply cause an inverted-U effect of patents on innovation? Econom. Lett. 117 (1) (2012) 211–213.

[33] M. Lundstrom, Applied physics: Enhanced: Moore's law forever? Science 299 (5604) (2003) 210–211.

[34] Yi Zhang, Ying Huang, Denise Chiavetta, Alan L. Porter, An introduction of advanced tech mining: Technical emergence indicators and measurements, Technol. Forecast. Soc. Change 182 (2022) 121855.

[35] Nesserine Benchettara, Rushed Kanawati, Celine Rouveirol, Supervised machine learning applied to link prediction in bipartite social networks, in: 2010 International Conference on Advances in Social Networks Analysis and Mining, IEEE, 2010, pp. 326–330.

[36] Paulo Ricardo da Silva Soares, Ricardo Bastos Cavalcante Prudêncio, Time series based link prediction, in: The 2012 International Joint Conference on Neural Networks (IJCNN), IEEE, 2012, pp. 1–7.

[37] Tomasz Tylenda, Ralitsa Angelova, Srikanta Bedathur, Towards time-aware link prediction in evolving social networks, in: Proceedings of the 3rd Workshop on Social Network Mining and Analysis, 2009, pp. 1–10.

[38] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, Mohammed Zaki, Link prediction using supervised learning, in: SDM06: Workshop on Link Analysis, Counter-Terrorism and Security, Vol. 30, 2006, pp. 798–805.

[39] Lu Huang, Xiang Chen, Xingxing Ni, Jiarun Liu, Xiaoli Cao, Changtian Wang, Tracking the dynamics of co-word networks for emerging topic identification, Technol. Forecast. Soc. Change 170 (2021) 120944.

[40] Muhan Zhang, Yixin Chen, Link Prediction Based on Graph Neural Networks, 2018, ArXiv:1802.09691 [Cs, Stat].

[41] César A Hidalgo, Bailey Klinger, A-L Barabási, Ricardo Hausmann, The product space conditions the development of nations, Science 317 (5837) (2007) 482–487.

[42] César A. Hidalgo, Ricardo Hausmann, The building blocks of economic complexity, Proc. Natl. Acad. Sci. 106 (26) (2009) 10570–10575.

[43] Maximilian Klein, Thomas Maillart, John Chuang, The virtuous circle of Wikipedia: recursive measures of collaboration structures, in: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, 2015, pp. 1106–1115.

[44] Andrea Tacchella, Matthieu Cristelli, Guido Caldarelli, Andrea Gabrielli, Luciano Pietronero, A new metrics for countries' fitness and products' complexity, Sci. Rep. 2 (1) (2012) 723.

[45] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, Jure Leskovec, Open graph benchmark: Datasets for machine learning on graphs, Adv. Neural Inf. Process. Syst. 33 (2020) 22118–22133.

[46] Peng Wang, BaoWen Xu, YuRong Wu, XiaoYu Zhou, Link prediction in social networks: the state-of-the-art, Sci. China Inf. Sci. 58 (1) (2015) 1–38.

[47] Wadhah Almansoori, Shang Gao, Tamer N. Jarada, Abdallah M. Elsheikh, Ayman N. Murshed, Jamal Jida, Reda Alhajj, Jon Rokne, Link prediction and classification in social networks and its application in healthcare and systems biology, Netw. Model. Anal. Health Inform. Bioinform. 1 (1) (2012) 27–36.

[48] Junichiro Mori, Yuya Kajikawa, Hisashi Kashima, Ichiro Sakata, Machine learning approach for finding business partners and building reciprocal relationships, Expert Syst. Appl. 39 (12) (2012) 10402–10407.

[49] Cuneyt Gurcan Akcora, Barbara Carminati, Elena Ferrari, Network and profile based measures for user similarities on social networks, in: 2011 IEEE International Conference on Information Reuse Integration, 2011, pp. 292–298.

[50] Linyuan Lü, Tao Zhou, Link prediction in complex networks: A survey, Phys. A 390 (6) (2011) 1150–1170.

[51] Juram Kim, Seungho Kim, Changyong Lee, Anticipating technological convergence: Link prediction using Wikipedia hyperlinks, Technovation 79 (2019).

[52] Jiho Lee, Namuk Ko, Janghyeok Yoon, Changho Son, An approach for discovering firm-specific technology opportunities: Application of link prediction to F-term networks, Technol. Forecast. Soc. Change 168 (2021).

[53] Junhan Kim, Youngjung Geum, How to develop data-driven technology roadmaps:The integration of topic modeling and link prediction, Technol. Forecast. Soc. Change 171 (2021) 120972.

[54] Shaohuai Shi, Qiang Wang, Pengfei Xu, Xiaowen Chu, Benchmarking state-of-the-art deep learning software tools, in: 2016 7th International Conference on Cloud Computing and Big Data (CCBD), IEEE, 2016, pp. 99–104.

[55] 'Indeed.com', Indeed job openings, 2020, Indeed.com, data crawled on January 2021, from https://ch.indeed.com.

[56] Timothy A. Davis, John R. Gilbert, Stefan I. Larimore, Esmond G. Ng, Algorithm 836: COLAMD, a column approximate minimum degree ordering algorithm, ACM Trans. Math. Software 30 (3) (2004) 377–380.

[57] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Appl. 13 (4) (1998) 18–28.

[58] Albert-László Barabási, Réka Albert, Emergence of scaling in random networks, Science 286 (5439) (1999).

[59] Leo Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43.

[60] Jérôme Kunegis, Ernesto W. De Luca, Sahin Albayrak, The link prediction problem in bipartite networks, in: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, 2010, pp. 380–389.

[61] W.J. Youden, Index for rating diagnostic tests, Cancer 3 (1) (1950) 32–35.

[62] David Powers, Evaluation: From precision, recall and F-Factor to ROC, informedness, markedness & correlation, Mach. Learn. Technol. 2 (2008).

[63] Douglas C. Montgomery, Introduction to Time Series Analysis and Forecasting, in: Wiley Series in Probability and Statistics, 526, 2008, p. 5.

[64] Christoph Bergmeir, José M. Benítez, On the use of cross-validation for time series predictor evaluation, Inform. Sci. 191 (2012) 192–213.

[65] Tom Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.

[66] Yang Yang, Ryan N. Lichtenwalter, Nitesh V. Chawla, Evaluating link prediction methods, Knowl. Inf. Syst. 45 (3) (2015).

[67] Aric Hagberg, Pieter Swart, Daniel S. Chult, Exploring network structure, dynamics, and function using NetworkX, Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[68] William H. Kruskal, W. Allen Wallis, Use of ranks in one-criterion variance analysis, J. Amer. Statist. Assoc. 47 (260) (1952) 583–621.

[69] M.A. Brach, Real Options in Practice, John Wiley & Sons, 2003, Google-Books-ID: ttrWdu3uEgkC.

[70] A. Meijer, M. Wessels, Predictive Policing: Review of Benefits and Drawbacks, Int. J. Public Adm. 42 (12) (2019) 1031–1039.