Performing Social Data Analysis with Neo4j: Workforce Trends & Corporate Information Leakage

Călin CONSTANTINOV, Lucian IORDACHE, Adrian GEORGESCU, Paul-Ştefan POPESCU, Mihai MOCANU
Department of Computers and Information Technology
University of Craiova, Romania
constantinov.calin@ucv.ro, popescu.paulstefan@ucv.ro, mmocanu@software.ucv.ro

***Abstract*—Not until long ago, performing a study on how the graduates of a University programme have integrated in the labour market was a tedious task, requiring the sending and, hopefully, the receiving of (online) questionnaires. Fortunately, given the popularity of professional social networks, carrying out a survey is now much easier, as former students can simply export their online profiles while giving consent for processing their data. As a small-scale case-study, this paper uses professional information from a number of alumni of the Computer specialisations of the Faculty of Automation, Computers and Electronics, University of Craiova. Given its highly-interconnected structure, for modelling this snapshot of the local Information and Communication Technology (ICT) workforce, we chose to use Neo4j, the most popular graph database. By relying on a set of data visualisation techniques, the resulting system allowed for the uncovering of valuable feedback and insights from within this data. Several remarks are made regarding the observed trends among current employees. Moreover, a discussion is carried out on of how corporate information can unknowingly get leaked on social media by a collective of employees. While this case-study is limited for a very specific use-case, the proposed approaches can easily be extended for various scenarios.**

***Index Terms*—Data visualisation, Graph databases, Professional network analysis**

I. Introduction

During one of our previous research efforts, summarised in work [1], we detailed how educational institutions can use information available on professional social networks as feedback for improving their programmes. For instance, growing market demand as well as loss in interests for certain technologies might need to get reflected in what current students are studying. Similarly, monitoring the same data can be used as a direct and very relevant assessment on how successfully a graduate has managed to integrate in the job market and how his career path has evolved over time. For this study, we asked a number of our alumni to export their professional network profiles and allow us to anonymise and store their data for performing various analysis. For the experiment, only professionals which are based in Craiova were considered. While not in the initial scope of our experiments, as an interesting side-effect, not only did the obtained data allow us to find details regarding the skills of our graduates, but valuable information has also surfaced concerning the companies for which they work.

While our population is not very large, we chose to model and store our data using technologies that can easily be applied

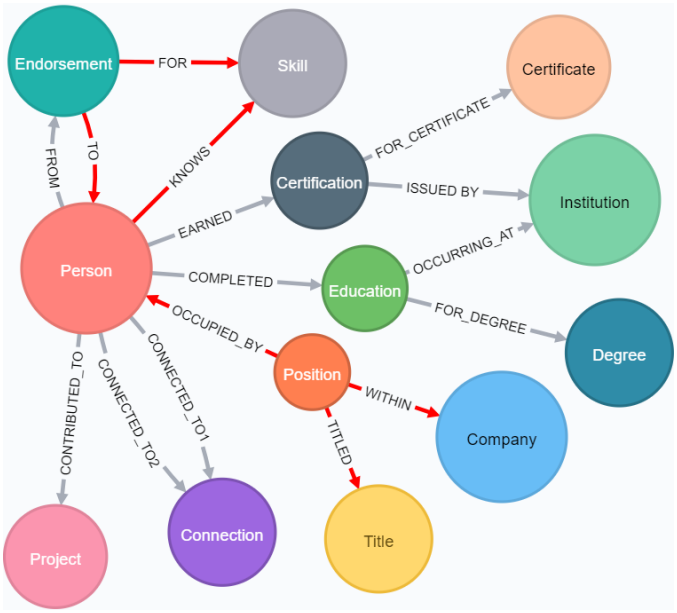

Figure 1. The Meta-Graph

to experiments of any sizes. Considering both its scalability and analytical power and given our previous experience with large-scale data analysis, as described by papers [2] and [3], we once again chose to use Neo4j. Another advantage of using this technology is that the persisted structure can easily be changed, enabling us to maintain a model that can fit data coming from any professional network, whenever needed. After analysing typical information available in these platforms, for storing the obtained snapshot of the local ICT workforce, we propose and use the graph database structure displayed in Figure 1. For this first analysis, we limited to using information contained by the the edges highlighted in red, along with the connecting nodes. As an observation, in some cases, an educational institution can also be an employing company. Along the same lines, a company might be the institution offering a certification. Some further adjustments would be required in order for our model to handle these scenarios. Future experiments that can leverage all available data to a grater extend are under development.

By the end of 2016, a number of 206 professionals answered our call and provided us with an export of their LinkedIn profiles, while also allowing us to use their data for this case-study. We estimate that this accounts for about 30%–35% of

978-1-5386-4444-7/18/$31.00 ©2018 IEEE

Table I
RELEVANT SNAPSHOT DETAILS

Entity	Count
Nodes	24497
Relationships	82163
Users Total	2044
↪ Authorising Users	206
Jobs Positions Total	775
↪ Active Jobs Positions	273
Job Titles	408
Companies	291
↪ Companies with local HQ	27
Skills	991
Endorsements	19421

our graduates from the last 10 years, which are currently based in Craiova. Table I provides details on the structure of the database, as resulting after processing every available profile.

Each person in the dataset typically showcases 14 skills on his profile for which he receives, on average, 94 endorsements, showing that our population is mostly formed of heavy users of the platform. Because endorsements are also coming from people outside our set of authorising users, some sort of data for a total of 2044 profiles was captured by our system. The database also contains a total of 775 job positions out of which about 35% were active, in Craiova (as in, the person is still occupying the job). After merging duplicate entries, we identified 27 companies having a local headquarter in Craiova for which these professionals work or have worked. Additionally, there were also some professionals that are currently located in Craiova but are freelancers or work remotely for companies based elsewhere. Many of our graduates are often tempted to move to other cities, especially Bucharest. At some point, a fraction of them chose to return home, reason for which the dataset also contains information about past job entries for "non-local" companies. Starting from the experiments detailed in work [4], in a future study, we plan a more in-depth analysis on the number of people that chose to relocate, along with their typical professional profile. The database also contains other types on entities, as per the previous meta-graph, but, as they were not part of the experiment, we chose not to include them in this table. Overall, from a graph database perspective, the whole dataset consisted of about 24k nodes, linked by around 82k relationships.

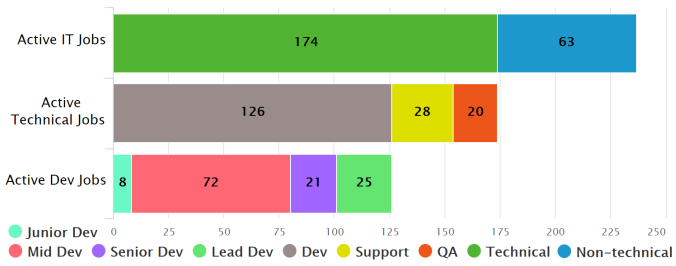

Figure 2. Job Types Distribution

Table II
COMMON JOB TITLES

Job Title	Count
Software Developer	50
Software Engineer	34
Programmer	20
Project Manager	15
Senior Software Engineer	14
QA Engineer	13
Java Developer	13
Owner	9
Developer	8
Junior Software Engineer	8

II. CASE STUDY: LOCAL ICT ALUMNI

A. Employee insights

We begin the analysis by looking over the structure of the currently active jobs, as seen in Figure 2. A first observation is that 86% of them were related to the ICT sector, showing that our graduates can largely find work in the area of their specialisation. Consistent with the findings of studies [5], [6] and [7], we also see that, at a certain point in their career, some professionals easily switch to jobs requiring business and managerial skills. More specifically, 26% of the active jobs are non-technical, including positions ranging from business consultants and analysts up to VPs, Directors or even C-level officers. While most of our graduates now work as mid-level developers, there are also a significant number of QA specialist roles as well as support jobs, including complementary roles such network or system administrators.

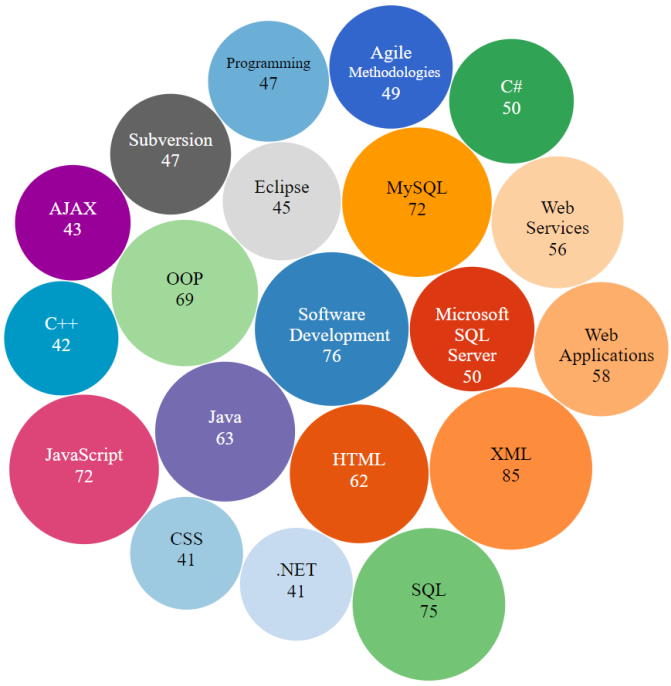

Figure 3. Top 20 showcased skills, by the number of occurrences

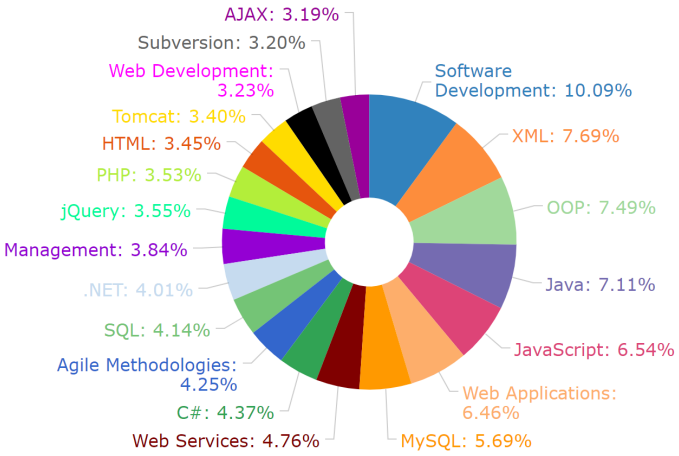

Figure 4. Percentages for top 20 endorsed skills

As a more in-depth view, Table II shows most common job titles, now including both active and inactive jobs. The table shows that there can be many names for the same job description, often even requiring the same level of seniority. More specifically, there are 163 job titles for the 273 active jobs in our dataset. Thus, a mechanism for normalising these positions would be needed for better insights. Possible solutions are mentioned in works [8] and [9]. Figure 3 shows which are the most commonly showcased skills by the professionals volunteered in our experiment. Interestingly, some of the skills are quite generic, such as *Programming*, while others are not very meaningful, such as *Eclipse*. Moreover, as some of these items are overlapping, such as *SQL* and *Microsoft SQL Server*, there is a clear need formalise and standardise these skills. Possible approaches are given by works [10] and [11]. For consistency, the same colours for each skill is kept throughout all the charts of this study. Figure 4 highlights which are the top endorsed skills. Some degree of correlation between the most showcased and most endorsed skills can be observed. Of course, there are also differences: while not being very showcased, *Management*, *jQuery*, *PHP* and *Tomcat* generate a more significant number of endorsements than *Microsoft SQL Server*, *Programming*, *Eclipse* and *CSS*, which no longer make it to the top. For a significant part of the skills, the percentages are somewhat uniformly distributed. As detailed in a following section, viewing the same data on a per-company basis will lead to much more insightful conclusions. The experiment

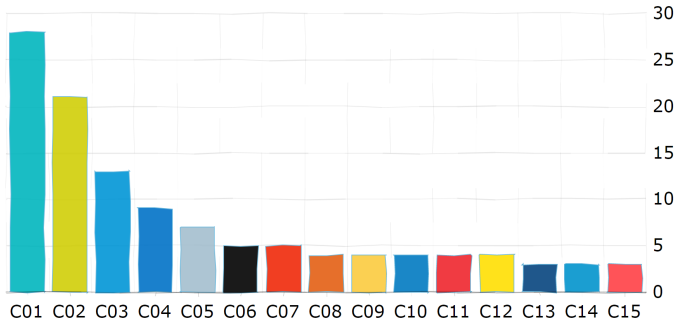

Figure 5. Top 15 companies by number of active jobs (employees)

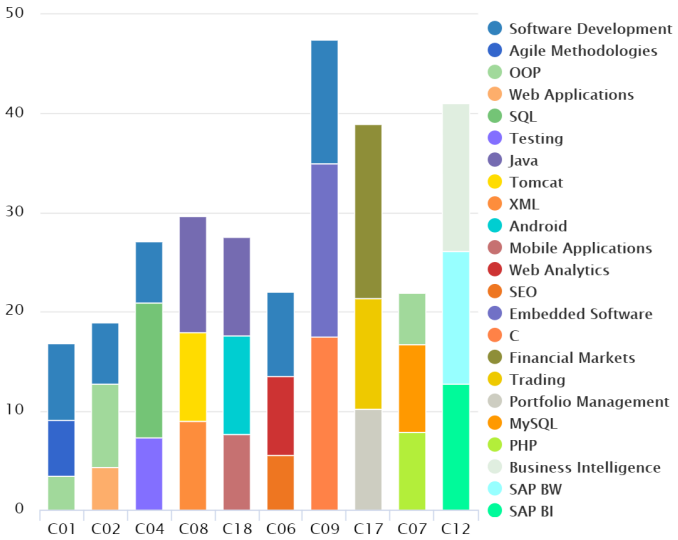

Figure 6. Percentages for top 3 endorsed skills for selected companies

leads to the conclusion that our graduates are successful in pursuing careers in the ICT sector. Moreover, information coming from what alumni chose to showcase and endorse can be used as initial input for any decision to modify the syllabi of the courses from our study programme. In a future analysis, we plan to enhance the data with temporal information, attempting to discover how interest in each skill evolved over time. In an era characterised by an abundance of emerging programming languages, paradigms and computational models, these types of studies have the potential of leading to a valuable prediction model which can help anticipate the next technology hype.

B. Company profiles

As our data also contained information regarding the organisations for which our graduates work, we also decided to obtain a more detailed view of the largest ICT companies that are either based in Craiova or have a local office here. Figure 5 shows the number of active jobs (or employees) for the top

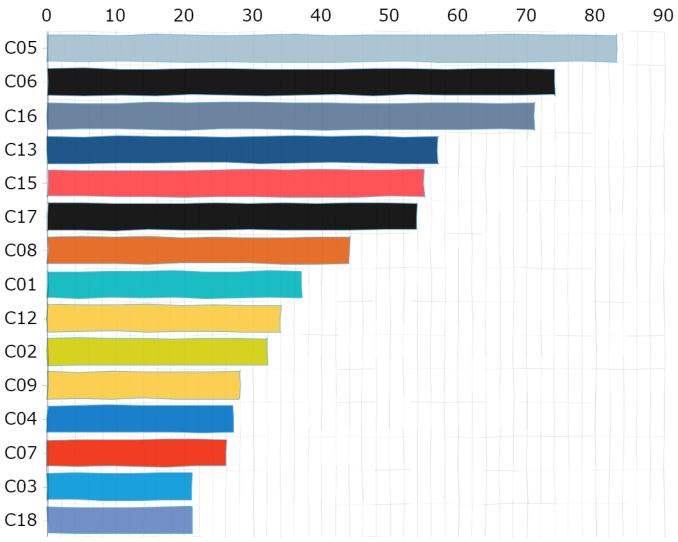

Figure 7. Average time (in months) an employee spends with the company

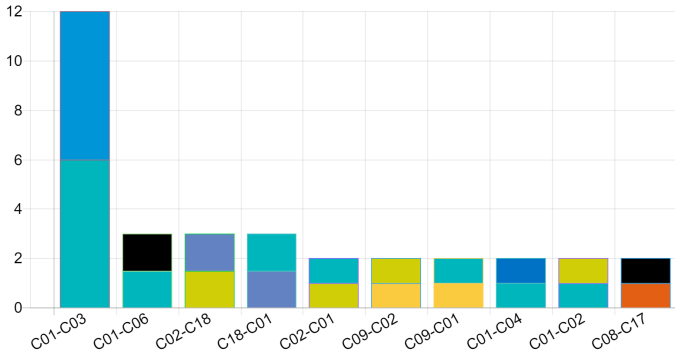

Figure 8. Top 10 leaves from one company to another

15 companies, as can be seen from our 206 users. While we willingly chose to anonymise the data, the colour selected for representing an entity is the one most predominant on that company's official logo. Thus, similar shades are used for multiple companies. The same colours are consistently used for the reminder of this study. As they have publicly embraced our experiment and encouraged their employees to participate, some companies are more strongly present in the dataset, thus leading to a certain bias. However, the sample is still representative for the proportions of most of the companies.

A interesting view of the data is obtained when correlating endorsements and companies, as per Figure 6. Here, all the endorsements of all the active employees of a company are counted, for each skill. This is a very basic example of how, by collectively analysing employee data, information which might not be publicly available can be obtained. The phenomenon, typically coined "light leakage" is discussed in greater depth by article [12], where details on how employees can unwittingly share sensitive information are given. From this data, companies which are highly specialised over an industry sector or are offering their own Intellectual Property products, such as *C09*, *C17* or *C12*, can be observed. For them, the top 3 most endorsed skills are responsible for generating around 40% of all the endorsements. Similarly, companies for which the business model relies on providing a wide palette of software outsourcing services, such as *C01* and *C02*, can be distinguished. In their case, the top 3 most endorsed skills rarely reach 20%. There is no information for when an endorsement was received, meaning that some were received when these people were working for different companies. Nevertheless, the data can still be successfully used for classifying the companies. An ideal candidate profile for each of these companies could also be built based on this data. The next two figures contain information which can be linked to company attractiveness and attrition. Figure 7 shows how much time an employee spends, on average, at a company, as can be deduced from data available on LinkedIn. In case of two or more successive positions within the same company, the entire period was considered as a whole. The top is dominated by companies which are well established locally, being active for more then 15 years, such as *C05*, *C06* and *C16*. However, *C02*, which has also been operating locally for a long time, is not as successful in keeping employees from leaving. A final observation is that not all companies which rely on selling their own products occupy high positions in this top. This is surprising as, losing talent is likely to be more problematic for them. Finally, Figure 8 shows the dynamics of the local ICT job market by highlighting common company switches. As the data contains some temporal information, a much more valuable analysis might be possible if modelling and interpreting this data as a time-series. For instance, an initial analysis shows that the general tendency to spend a large period with a company has slowly, but constantly, decreased.

III. Conclusion

In this case-study, a number of data modelling along with adequate visualisation techniques have been proposed for performing an analysis on professional social network profiles. The results show that graduates from a local University programme easily integrate in the job market. Details on how to extract up-to-date information regarding the most attractive skills are also outlined. Finally, a discussion is carried out on how the same dataset can be used for obtaining corporate details which might not be publicly available. The same approach can easily be applied for other industrial sectors and can be geographically extended to include regional information.

References

[1] C. Constantinov, P. Ş. Popescu, C. M. Poteraş, and M. L. Mocanu, "Preliminary results of a curriculum adjuster based on professional network analysis," in *System Theory, Control and Computing (ICSTCC), 2015 19th International Conference on*. IEEE, 2015, pp. 860–865.

[2] C. Constantinov, M. L. Mocanu, and C. M. Poteras, "Running complex queries on a graph database: A performance evaluation of neo4j," *Annals of the University of Craiova*, vol. 12, no. 1, pp. 38–44, 2015.

[3] C. Constantinov, C. M. Poteras, and M. L. Mocanu, "Performing real-time social recommendations on a highly-available graph database cluster," in *Carpathian Control Conference (ICCC), 2016 17th International*. IEEE, 2016, pp. 116–121.

[4] R. Y. Tantawy, Z. Farouk, S. Mohamed, and A. H. Yousef, "Using professional social networking as an innovative method for data extraction: The ict alumni index case study," *CoRR*, vol. abs, 2014.

[5] R. Colomo-Palacios, E. Tovar-Caro, Á. García-Crespo, and J. M. Gómez-Berbís, "Identifying technical competences of it professionals: The case of software engineers," *International Journal of Human Capital and Information Technology Professionals*, pp. 31–43, 2010.

[6] T. Case, A. Gardiner, P. Rutner, and J. Dyer, "A linkedin analysis of career paths of information systems alumni," *Journal of the Southern Association for Information Systems*, vol. 1, no. 1, pp. 1–13, 2013.

[7] L. Li, G. Zheng, S. Peltsverger, and C. Zhang, "Career trajectory analysis of information technology alumni: A linkedin perspective," in *Proceedings of the 17th Annual Conference on Information Technology Education*. ACM, 2016, pp. 2–6.

[8] P. Garg, R. Rani, and S. Miglani, "Mining professional's data from linkedin," in *Advances in Computing and Communications (ICACC), 2015 Fifth International Conference on*. IEEE, 2015, pp. 98–101.

[9] V. Ha-Thuc, Y. Xu, S. P. Kanduri, X. Wu, V. Dialani, Y. Yan, A. Gupta, and S. Sinha, "Search by ideal candidates: Next generation of talent search at linkedin," in *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 195–198.

[10] M. Diaby and E. Viennet, "Taxonomy-based job recommender systems on facebook and linkedin profiles," in *Research Challenges in Information Science (RCIS), 2014 IEEE Eighth International Conference on*. IEEE, 2014, pp. 1–6.

[11] V. Shankararaman and S. Gottipati, "Mapping information systems student skills to industry skills framework," in *Global Engineering Education Conference (EDUCON), 2016 IEEE*. IEEE, 2016.

[12] D. Bradbury, "Data mining with linkedin," *Computer Fraud & Security*, vol. 2011, no. 10, pp. 5–8, 2011.