LinkedIn Economic Graph

TECHNICAL NOTE

LinkedIn STEM Classification Methodology

Matthew Baird
LinkedIn Economic Graph Research & Insights

Nikhil Gahlawat
LinkedIn Economic Graph Research & Insights

Rosie Hood
LinkedIn Economic Graph Research & Insights

Paul Ko
LinkedIn Economic Graph Research & Insights

Silvia Lara
LinkedIn Economic Graph Research & Insights

economicgraph.linkedin.com

1. Introduction

Over the last few decades, STEM (Science, Technology, Engineering, and Mathematics) has gained significant importance as a categorization of educational programs and resulting occupations in science-related fields. Governments typically rely on expert panels to create lists of STEM fields of study and occupations, which are then used to inform explicit government policies such as work visas and funding decisions.

LinkedIn is an ideal data source for measuring trends in STEM and making comparisons across different groups, such as by gender, by time in their career, and internationally. This report outlines the methodology used to create international taxonomies that classify each representative occupational group as STEM or non-STEM. We take a skills-based approach, as STEM is primarily a conceptualization about skills. The methodology involves three basic steps:

1. Identify STEM degree holders
2. Identify STEM skills based on STEM degree holders
3. Identify STEM occupations based on STEM skills

We intentionally developed a data-driven methodology that is consistent across countries, while still allowing for variations between them. For instance, certain occupations may require different skills in different countries. In the following section, we will describe our approach in detail.

2. Sample Inclusion

In order to ensure sufficient data quality, we make several restrictions on whether a country is included in our methodology. First, we require that countries have a minimum of 100,000 STEM degree holders, which is enough to estimate STEM skills at more refined levels without introducing significant noise. Second, we only evaluate countries that are considered key markets according to our Data for Impact program. Third, we restrict which skills we evaluate at a country-level, and only include them if

the skill has 100 degree holders in the country who have added the skill. For gender-based analysis, we only consider countries where our gender-inference coverage is estimated to be greater than 67%. In cases where members have not explicitly self-identified their gender, we have inferred it based on their first name and other relevant information.

3. STEM degree holders

Our approach relies on first identifying STEM degree holders. To accomplish this, we used the U.S. Department of Homeland Security's classification of STEM degrees using Classification of Instructional Program (CIP) codes. We decided to use the US STEM CIP classification for several reasons. First, it aligns with our data, as our standardization team at LinkedIn has created a taxonomy of fields of study in college degrees with CIP codes as the output. Second, STEM definitions are generally similar across countries (see below for a note on international comparisons). Third, the US has the largest user base of LinkedIn members of any country.

We used the DHS list of STEM fields of study to create a sample of both STEM and non-STEM degree holders, including graduates of any post-secondary level, from associate degrees through doctoral and professional degrees. In some cases, we were unable to classify degree holders as STEM or non-STEM. This occurred when the CIP code was missing (i.e., when the member did not provide any information about their degree's field of study on their profile), or when the CIP code was too general (i.e., when the standardization team was unable to classify them with a full 4-digit ##.## CIP code and only provided two or three digits, and the children codes under that parent level were not wholly STEM or non-STEM). If the major groups included both STEM and non-STEM degrees, then we were unable to classify these individuals as either STEM or non-STEM degree holders, and they were excluded from the calculations.

4. STEM skills

4.1. STEM skills methodology

After classifying degree holders as STEM or non-STEM, we proceeded to classify skills as either STEM or non-STEM. We defined a skill as STEM if it was significantly more likely to be held by a STEM degree holder than by a non-STEM degree holder.

To determine whether a given skill was classified as STEM, we utilized the LinkedIn profile feature that allows users to add skills. For each skill, we calculated the probability of a STEM degree holder adding that skill by country. To do this, we used the empirical ratio of the proportion of STEM degree holders who added the skill to the proportion of non-STEM degree holders who added the skill. This yielded the probability of skill j being added by STEM degree holders in country c (p_{jc}^{STEM}) using the empirical ratio:

$$p_{jc}^{STEM} = \frac{Number\ of\ STEM\ degree\ holders\ with\ skill\ j\ in\ country\ c}{Number\ of\ STEM\ degree\ holders\ in\ country\ c}$$

Similarly, we calculated the empirical ratio for non-STEM degree holders, which gave us the probability of skill j being added by non-STEM degree holders in country c ($p_{jc}^{NONSTEM}$).:

$$p_{jc}^{NONSTEM} = \frac{Number\ of\ Non-STEM\ degree\ holders\ with\ skill\ j\ in\ country\ c}{Number\ of\ Non-STEM\ degree\ holders\ in\ country\ c}$$

We then calculated the odds ratio (r_{jc}) for each skill j in each country c, which measures how much more likely it is for a STEM degree holder to add the skill to their LinkedIn profile compared to a non-STEM degree holder. This ratio is obtained by dividing the probability of a STEM degree holder adding the skill (p_{jc}^{STEM}) by the probability of a non-STEM degree holder adding the skill ($p_{jc}^{NONSTEM}$). The odds ratio indicates the extent to which a skill is associated with STEM degrees in a given country

$$r_{jc} = \frac{p_{jc}^{STEM}}{p_{jc}^{NONSTEM}}$$

A skill was classified as STEM if r_{jc} was greater than or equal to five in a given country, or if $p_{jc}^{NONSTEM}$ was zero and p_{jc}^{STEM} was non-zero, provided that at least 100 STEM degree holders in the country had added that skill. We chose the threshold of 5 as it provides a common benchmark for all countries, is an intuitive round number, and is around the 75th percentile of the distribution of r_{jc} (76th percentile). Additionally, this approach yields estimates of the STEM workforce size and STEM gender composition in the US that are consistent with other estimates from the literature.

Figure 1 displays the kernel density of the log skill ratio for the 50 countries included in our analysis. We chose to use the natural log due to a long right tail in the data, where some skills have many STEM holders but only a few non-STEM holders. The vertical green line indicates the threshold used for classifying skills as STEM or non-STEM. Skills to the left of the green line are not classified as STEM, while those to the right are classified as STEM. A value of zero indicates a skill ratio of 1 (ln(1)=0), where STEM degree holders are equally likely to add the skill as non-STEM degree holders. We observe a large concentration of points just below zero, and the threshold of five approximately separates this group from a smaller group of STEM-dominated points.

Figure 1: Kernel density of log skill ratio
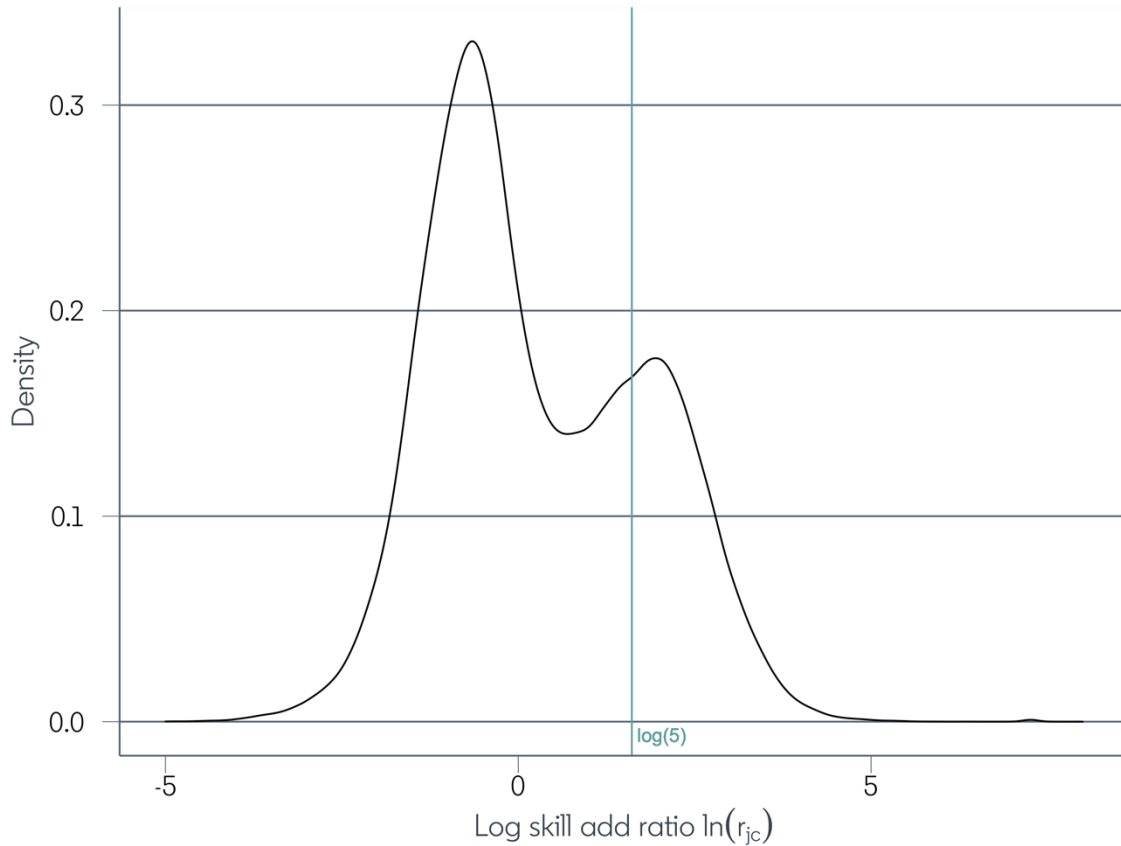
We can gain further insights into the distribution of skill ratios and thresholds by examining the empirical cumulative density function (CDF). Figure 2 displays the CDF estimates by country, with each line representing a different country among the 50 included. Although the trends are very similar, some variation exists between countries. The vertical line at log(5) indicates the percentile on the y-axis at which the threshold separates the skills. Globally, the threshold of 5 times as likely corresponds to the 76th percentile across all skill ratios. The percentiles at the odds ratio of 5 range from 69.8 percentile to 83.2 percentile across countries. Therefore, the threshold yields a reasonable separation in all cases, with none falling below the 50th percentile. The similarity across countries provides additional reassurance that the methodological choices are appropriate for a common approach.

Figure 2: Empirical cumulative density function of log skill ratio by country (each line is a different country)
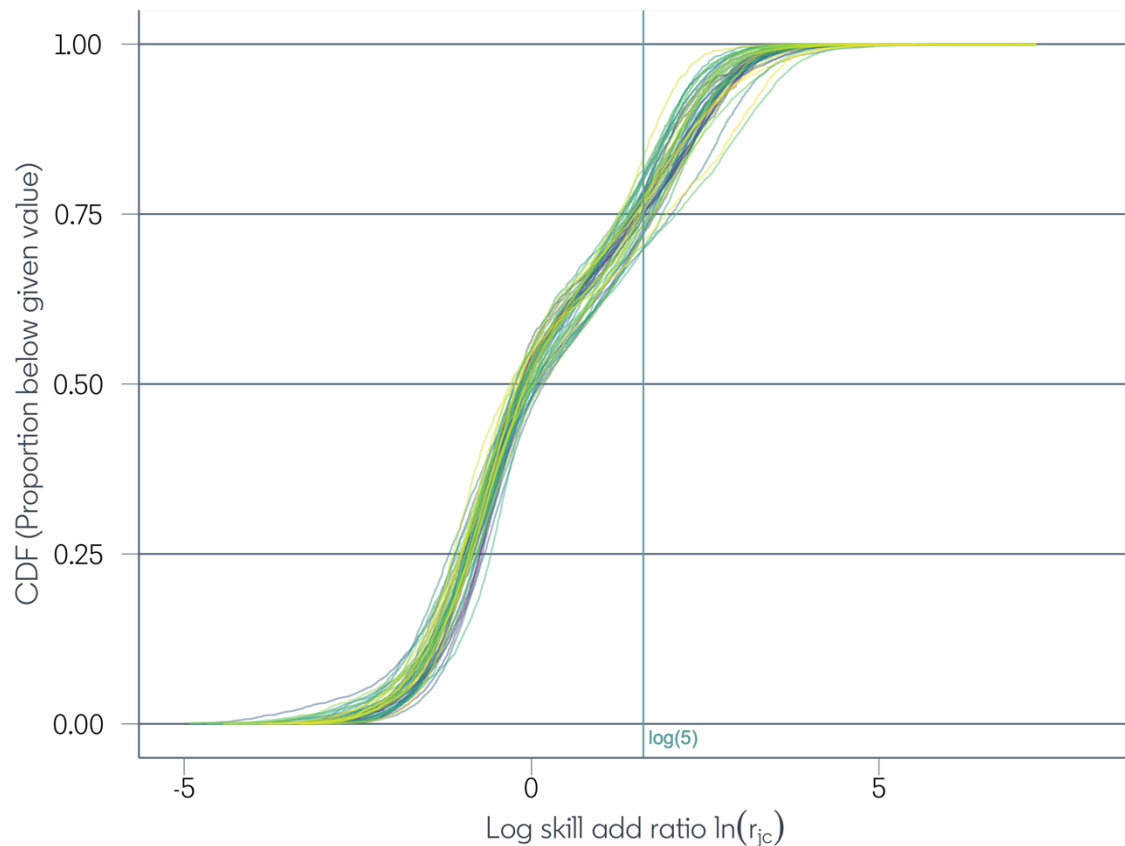
4.2. STEM skills statistics

Table 1 shows the top 10 STEM skills in terms of numbers of members who have added it out of the 50 countries included, as well as for the five countries with the highest LinkedIn membership. The lists are dominated by computer programming and engineering skills. For global, each of these skills has over 2 million members (cumulatively across the countries) who have added the skill across the 50 countries.

Table 1: Most common STEM skills (by number of members with skill added)

Global	Engineering, Python (Programming Language), SQL, Java, JavaScript, C++, C (Programming Language), MySQL, Software Development, Linux
United States	Python (Programming Language), SQL, Engineering, Java, JavaScript, C++, MATLAB, Software Development, Linux, C (Programming Language)
India	Engineering, SQL, C (Programming Language), Python (Programming Language), Java, HTML, C++, JavaScript, Cascading Style Sheets (CSS), MySQL
Brazil	Engineering, SQL, JavaScript,), HTML, Python (Programming Language), Java, MySQL, Linux, Technical Support
China	Java, C++, Linux, JavaScript, Software Development, MATLAB, C (Programming Language), MySQL, Machine Learning, Cascading Style Sheets (CSS)
United Kingdom	Engineering, Python (Programming Language), SQL, JavaScript, Java, Software Development, MATLAB, C++, Mathematics, C#

We can analyze the ranking of STEM skills by the odds ratio r_{jc} to understand how much more likely STEM degree holders are to add the skill compared to non-STEM degree holders. When we examine the top 10 skills for each country, we find that the ratios are in the hundreds, indicating that STEM degree holders are significantly more likely to add the skill compared to non-STEM degree holders. For instance, in the United States, the top five skills with the highest odds ratios are Geomodelling, GAMESS, Hit-to-Lead, SQL*NET, and Biocatalysis. Interestingly, the first three skills have no non-STEM degree holders who have added them, while the latter two have a relatively small number of STEM degree holders who have added them (each with hundreds, but not more than 1,000). To provide an alternative perspective, we also examine the highest ranked STEM skills among common skills (over 500,000 members who have added them globally) in Table 2. While the lists differ somewhat from those presented in Table 1, computer programming skills still dominate the highest ranked STEM skills in all five countries.

Table 2: Highest STEM-ranked skills (by ratio of probability added)

United States	Core Java (34.5), Spring Framework (32.2), Algorithms (27.9), Spring Boot (27.9), C (Programming Language) (25.1)
India	Spring Boot (18.6), Spring Framework (17.7), AngularJS (16.3), JSON (15.8), Node.js (15.6)
Brazil	Algorithms (24.2), LaTeX (23.4), Computer Science (21.7), MATLAB (18.2), C++ (17.5)
China	Spring Boot (18.5), REST APIs (15.7), Object-Oriented Programming (OOP) (12.7), Angular JS (12.6), JSON (12.3)
United Kingdom	Core Java (24.1), Spring Boot (21.1), Spring Framework (18.9), LaTeX (18.3), Algorithms (18.0)

Note: number in parentheses is r_{jc}, the ratio of probability of STEM degree holders adding the skill to non-STEM degree holders' probability

5. STEM occupations

5.1. STEM occupations methodology

To classify STEM occupations, we use the LinkedIn Skills Genome[1], which creates a ranked list of skills attached to each occupation based on term frequency-inverse document frequency (TF-IDF) ratings. Essentially, a skill has a higher TF-IDF score for an occupation if members in that occupation are more likely to add that skill than members in all other occupations. We define an occupation as STEM if one of its most important skills is a STEM skill, as defined by the TF-IDF score ranking. For example, if an occupation's most important skills are all non-STEM skills, it will not be classified as a STEM occupation. It's worth noting that while a common skill like Microsoft Office may be important for an engineer, it would not rank high based on TF-IDF because it is also listed as a skill by many other occupations. In contrast, some of the programming languages listed in Tables 1 and 2 may be rated higher based on their TF-IDF scores.

We chose to use occupational representative groups to classify STEM. LinkedIn has different levels of specificity for its occupational taxonomies. The occupational representative groups contains around 3,194 unique occupation groups, making it

---

[1] https://engineering.linkedin.com/blog/2019/how-we-mapped-the-skills-genome-of-emerging-jobs

already quite a bit more specific than for example the Standard Occupational Classification (SOC) used by the U.S. Bureau of Labor Statistics (867 occupations).

5.1. STEM occupations statistics

Table 3 presents the most common STEM occupations. Engineering and analysts make a strong appearance throughout, as expected. Note too that given the methodological decisions of the level of occupational taxonomy to use as well as the skills-based approach, such occupations as professor and project manager make the top of these lists.

Table 3: Most common STEM occupations

Global	Software Engineer, Project Manager, Professor, System Engineer, Data Analyst, Design Engineer, Manufacturing Engineer, Business Analyst, Civil Engineer, Mechanical Engineer
United States	Software Engineer, Professor, Manufacturing Engineer, System Engineer, Data Analyst, Quality Assurance Manager, Engineering Manager, Mechanical Engineer, Director of Information Technology, Design Engineer
India	Software Engineer, System Engineer, Founder, General Manager, Design Engineer, Business Analyst, Programming Analyst, Data Analyst, Information Technology Analyst, Quality Assurance Engineer
Brazil	Software Engineer, Civil Engineer, System Analyst, Technical Support Analyst, Health Safety Environment Officer, Pharmacist, Founder, Business Analyst, Electrician, Information Technology Analyst
China	Software Engineer, Professor, Design Engineer, Frontend Developer, Data Analyst, Algorithm Engineer, Engineering Team Lead, Director Information Technology Operations, Chief Technology Officer, Business Analyst
United Kingdom	Software Engineer, Lecturer, Data Analyst, Teaching Assistant, Design Engineer, Professor, Technical Support Analyst, Solutions Architect, Support Team Lead, Technical Support Engineer

Table 3 shows the most common STEM occupations, while Table 4 provides a rough approximation of how STEM-intensive an occupation is by summing the skills ratios of its top 10 skills. By summing the skills ratio, this is in essence a weighted sum of total top 10 skills that are STEM, with the weights given by the degree of how STEM-

intensive the skill is. This is only used for descriptive purposes in ranking here. We assign a value of zero to skills that are not STEM-related. This sum is larger if an occupation has more STEM skills in its top 10 or if the included skills have a higher skills ratio, indicating that they are more STEM-centered. The top 10 occupations in Table 4 are dominated by STEM skills, with an average of over nine STEM skills in their top 10. This list is heavily weighted towards engineering, programming, and hard science occupations.

Table 4: Most STEM-intensive occupations

United States	Geology Specialist, Exploration Manager, Java Consultant, Thermal Engineer, Geophysicist, Computational Biologist, Analytical Chemist, Professor of Chemistry, Research Instructor, Bioinformatician
India	Dotnet Developer, Mobile Application Developer, Microbiologist, Full Stack Engineer, Computer Scientist, Interactive Developer, Back End Developer, Information Technology Associate
Brazil	Geophysicist, Geologist, Bioinformatician, Exploration Manager, Hardware Developer, Analytical Chemist, Embedded Software Engineer, Electronic Designer, Hardware Engineer, Machine Learning Researcher
China	Full Stack Engineer, Bioinformatician, Research Instructor, Chemist, Web Developer, Site Reliability Engineer, Artificial Intelligence Specialist, Java Consultant, DevOps Consultant, Mobile Application Developer
United Kingdom	Bioinformatician, Research Instructor, Biologist, Research Technologist, Chemist, Chemistry Supervisor, Lecturer in Chemistry, Physicist, Geologist, Chemical Engineer

Note: STEM occupations are ordered by the sum of the STEM/non-STEM skills ratio r_{jc}

6. Benchmark Comparisons

The purpose of this section is to compare our methodology for defining STEM to other available benchmarks. It should be noted from the outset that there is no single, universal classification of STEM, and different approaches can lead to different estimates of, for example, the proportion of the workforce that is STEM or the proportion of STEM workers that are women. Common approaches to defining STEM often use a top-down methodology, in which an expert panel subjectively classifies occupations as STEM or non-STEM. This is exemplified by methods used by the U.S. Census Bureau that rely on the Standard Occupational Classification (SOC). An alternative approach, used by Rothwell (2013), is a skills-based methodology that leverages O*NET to list common skills

for each occupation. Anderson et al. (2021) used a bottom-up approach by surveying individuals and asking them directly whether their job was in STEM. Our approach is primarily top-down, as it is built on the subjective classification of STEM fields of study by the United States using CIP codes. However, our methodology is also skills-based and data-driven, which may help alleviate some of the human biases that can arise from a purely subjective classification approach. In this section, we will compare our results to those obtained using these alternative approaches.

Table 5 presents some key estimates based on our data (as of January 2023) and compares them to external estimates derived from other classifications. As there is no universally accepted definition of STEM, different methods can yield different estimates of the size and composition of the STEM workforce. For example, the SOC, which uses a top-down approach based on expert ratings, tends to produce narrower estimates of STEM than a skills-based approach such as ours (Rothwell, 2013; Anderson et al., 2021).

Despite these differences, our estimates fall broadly within the range of other benchmarks, which lends additional credibility to our approach. For instance, our estimate of 19.4% for the fraction of the U.S. workforce in STEM is higher than some benchmarks, such as the Bureau of Labor Statistics, but lower than others, such as the estimates derived from O*NET data (Rothwell, 2013) and Okrent and Burke (2021). It is worth noting that a skills-based approach like ours may be expected to yield a larger estimate than a top-down approach, as it takes into account a wider range of skills and occupations that are relevant to STEM.

Another possible explanation for our higher estimate is that our sample of LinkedIn members may be more representative of workers in tech and engineering, which are highly represented in STEM, than the overall U.S. population. However, more research would be needed to confirm this hypothesis. However, our estimates are still well within the range of other estimates.

Table 5: Comparison of our STEM estimates against external benchmarks

Measure	Our estimate	Benchmarks
Fraction of U.S. workforce in STEM	19.4%	6.2%: BLS[1] 5.7%: Anderson et al. (2021) using Census classification[2] 9%: Rothwell for 2011, strict classification[3] 11.8%: Anderson et al. (2021) using Rothwell strict classification[2] 19.8%: Anderson et al. (2021) using self-reports[2] 20%: Rothwell for 2011, broad classification[3] 23%: Okrent and Burke (2021)[4] 26.5%: Anderson et al. (2021) using Rothwell broad classification[2]
Fraction of U.S. bachelor's or higher STEM graduates in STEM	42.7%	40.5%: Baird et al. (2017)[5]
Fraction of U.S. bachelor's or higher STEM male graduates in STEM	47.4%	49.1% Baird et al. (2017)[5]
Fraction of U.S. bachelor's or higher STEM female graduates in STEM	34.2%	29.7% Baird et al. (2017)[5]
Fraction of U.S. STEM workers that are women	30.1%	32.4%: NCES[6] 26%: Anderson et al. (2021) using Census definition[2] 23%: Anderson et al. (2021) using Rothwell strict definition[2] 32%: Anderson et al. (2021) using Rothwell broader definition[2] 34%: Okrent and Burke (2021)[4] 43%: Anderson et al. (2021) using self-reports[2]
Fraction of U.S. STEM workers with STEM degrees that are women	27.5%	34%: Anderson et al. (2021) using census classification[2] 26%: Anderson et al. (2021) using Rothwell strict classification[2] 35%: Anderson et al. (2021) using Rothwell broad classification[2] 50%: Anderson et al. (2021) using self-reports[2]

1. https://www.bls.gov/emp/tables/stem-employment.htm
2. Anderson, D., M. Baird, and R. Bozick (2021). "Who gets counted as STEM? A new approach for measuring the STEM workforce and its implications for identifying gender disparities in the labor market." *International Journal of Gender, Science and Technology,* 13(3), 254-279.
3. Rothwell, J. (2013). *The hidden STEM economy*. Washington, DC: Metropolitan Policy Program at Brookings.
4. Okrent, A., & Burke, A. (2021). The STEM labor force of today: Scientists, engineers, and skilled technical workers. Alexandria, VA: National Science Foundation.

5. Baird, M., R. Bozick and M. Harris (2017). "Postsecondary education and STEM employment in the United States: An analysis of national trends with a focus on the oil and natural gas industry." RR-2115-AMPI.

6. https://usafacts.org/articles/women-stem-degrees/?utm_source=google&utm_medium=cpc&utm_campaign=ND-Education-Childcare&gclid=Cj0KCQiA-oqdBhDfARIsAO0TrGH7dgFQfRL9vJp2X8tQq6rldvP17-INpBbNiBV715COxGR0tsAC7FAaAnZKEALw_wcB

7. Conclusion

This technical note explains the methodology underlying LinkedIn Economic Graph's definitions of STEM skills and STEM occupations. This data-driven, skills-based approach offers a taxonomy rooted in the definition of STEM (skills developed and used), and may thus be more inclusive against traditional hierarchal, subjective classifications. It also allows for an internationally consistent approach which can be leveraged across several countries. Additionally, this approach can help identify emerging STEM skills and occupations that may not have been traditionally classified as such. A skills-based approach can also help identify areas where workers may need additional training and development to meet the demands of the evolving job market. Finally, by producing first a list of STEM skills, it offers an opportunity for deeper investigation into the heart of what drives the STEM workforce and the potential to understand patterns in and out of STEM work based on skill development patterns.