People Data Labs

Exploring Employee Headcount Accuracy

A detailed look into where industry standard data sources fall short and how to improve them

Table of Contents

Introduction

Headcount information is frequently leveraged by investors, market analysts, and human resource professionals as a powerful signal for revenue growth, corporate strategy, and company health.

Yet, there is no universal source of truth for headcount data.

While publicly traded companies are required to disclose headcount information once per year, private market companies do not have the same disclosure requirements.

Over time, LinkedIn has come to be regarded as the perceived authority for company headcount data, supported by its deep repository of crowdsourced information on company and employee profiles. However, producing accurate and consistent headcount estimates is surprisingly challenging and nuanced.

When looking closely at LinkedIn's headcount data, significant biases become evident arising from the platform's reliance on user-generated profiles, algorithmic approximations and imprecise data reporting practices.

At People Data Labs, we are uniquely focused on capturing the relationships between people and the companies they work for. Generating precise and reliable headcounts is an area we have explored deeply since our company's founding. We've seen firsthand how even subtle decisions can have an outsized impact on the final headcount estimates.

In this report, we will explore 3 factors of LinkedIn's methodology that we have observed lead to inaccurate and misleading headcounts in their data:

01. **Quality of User Profiles**

02. **Person-to-Company Matching**

03. **Parent-Subsidiary Relationships**

For each of these topics, we will explore concrete examples that reveal the systematic biases present in LinkedIn's headcount estimates and share strategies to overcome these shortcomings.

But first,
how is headcount data generated?

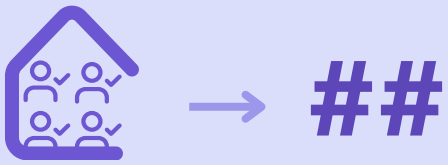
LinkedIn's headcount data is primarily estimated using the employment information contained in each LinkedIn user's profile. At a high level, there are 2 parts to this process:

Step 1. Matching user profiles to company profiles

The first part of this process involves matching user profiles to the company profiles using the information provided in the user's work experience fields. This process relies heavily on the company the user selected when inputting their work information. However, LinkedIn also applies text-based matching algorithms to generate additional matches between users and companies (e.g. by trying to identify company name references in the work experience fields).

Step 2. Aggregating the number of matched user profiles for each company

In the second step of this process, LinkedIn sums up the number of person profiles associated with each company to generate their headcount estimate. By accounting for start and end dates, they can also provide historical headcount information.



Note: This is just a high-level overview of the core process. LinkedIn also adds a variety of corrections to this estimate based on information like public records and direct, self-reported headcounts on the company profile.

The shortcomings in LinkedIn's methodology arise from the nuances in their implementation of this process.

Seemingly minor decisions, such as how lenient LinkedIn is in their text-matching process or how they choose to aggregate profiles, can have significant impacts on the generated headcount estimates at the end of the process.

PDL uses an analogous bottom-up approach to generating headcount estimates using our own independently sourced Person Dataset which contains current and historic employment information on the individual level.

Though we share the same high-level approach of first matching person records to company records and aggregating across each company record, we have also included a series of improvements around the edge cases we have observed firsthand as well as through reports from our customers.

The most impactful of these differences are our:

- Higher thresholds for **filtering out low-quality**, **incomplete**, and **inaccurate** profiles.
- Approach to **standardizing / canonicalizing companies** which leads to stricter matching between person records and company profiles.
- Separate accounting for employees at **parent organizations and their subsidiaries**.

Now, let's dig into these differences in detail.

Factor One:

Handling Low-Quality User Profiles

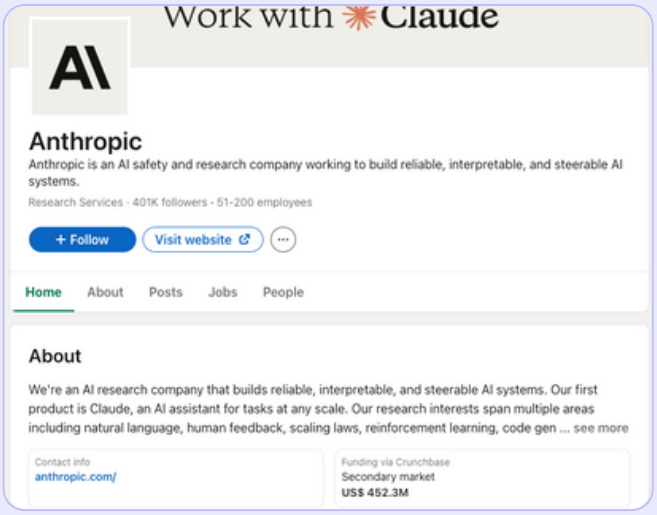
One of the most significant sources of error in LinkedIn's headcount data is the **inclusion of low-quality employee profiles in their headcount estimates.** These profiles come in a variety of forms; particularly as incomplete or inaccurate profiles.

PDL's own internal research on user profile quality indicates that **only around 10-20% of the 1B+ users on LinkedIn meaningfully represent real-world individuals** (with complete and up-to-date employment information).

~10-20%

of the 1B+ users on LinkedIn meaningfully represent real-world individuals

In many cases, these profiles are relatively easy to identify due to the lack of information they contain. However, these profiles are still quite pervasive throughout the site and are frequently included in LinkedIn's headcount calculations.



(Source: LinkedIn)

One clear example of this is the company profile for **Anthropic**, a leading US-based AI research organization known for building a popular alternative to OpenAI's ChatGPT. As a young and trending company with a rapid growth trajectory, this company is an ideal example to illustrate the prevalence of low-quality profiles that self-associate themselves with this organization on LinkedIn.

According to Anthropic's LinkedIn page, the company has a headcount of 839 employees and a steep 89% growth rate over the past 6 months (Figure 1).

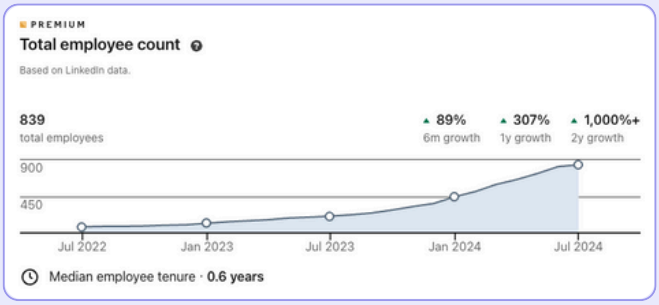


Figure 1
As of July 2024, LinkedIn reports total headcount of 839 and a 6 month growth rate of over 89%. (Source: LinkedIn)

Looking at the employee profiles (Figure 2), we see that the majority of them seem reasonable at first glance (e.g., located in the US/Bay area with educational backgrounds from leading CS programs).

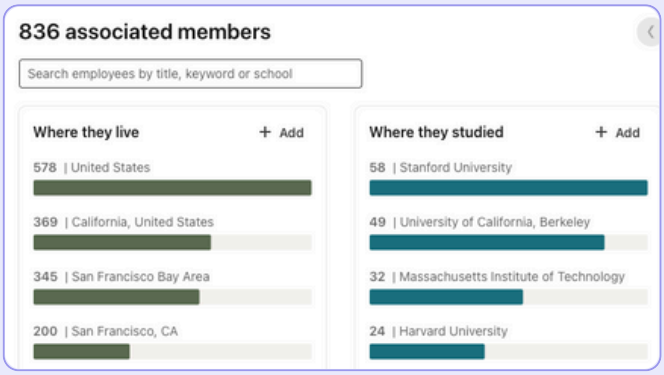


Figure 2
Employees associated with Anthropic's LinkedIn profile are largely based out of the US / Bay Area and studied at top Computer Science programs. (Source: LinkedIn)

However, looking at the long tail of LinkedIn employee profiles associated with this company is quite revealing.

Figure 3 illustrates the profiles that are representative of the bottom 15% of LinkedIn profiles associated with Anthropic, all of which contain missing or largely irrelevant information.

While this is only a first-order estimate of the number of low-quality profiles associated with the Anthropic profile, **the takeaway is that these profiles comprise a non-negligible percentage of the 839 employee profiles associated with the company's LinkedIn profile.**

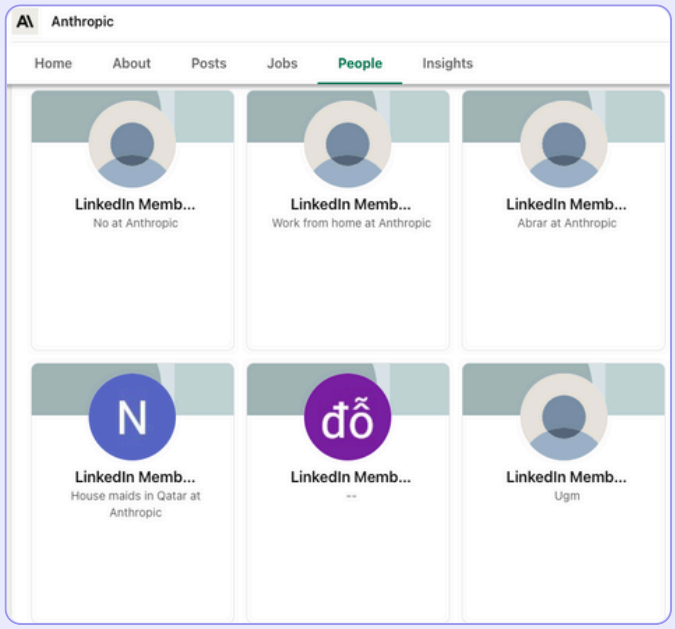


Figure 3
The low-quality profiles on Anthropic's LinkedIn account, included in their headcount estimates, represent the bottom ~120 employee profiles. (Source: LinkedIn)

This example also illustrates another important bias that is present in the LinkedIn data: well-known companies are more likely to find themselves associated with low-quality profiles.

Because LinkedIn does not filter out these low-quality profiles and includes them in their headcount calculations, popular companies will have inflated headcounts on LinkedIn.

PDL's Approach:

Filtering Low-Quality User Profiles

To correct for this type of bias in our own headcount estimation process, we flag and exclude person profiles, like those shown in the previous section, using a variety of unique filters.

The filters we apply are based on a variety of factors such as the level of profile completeness, as well as various "common sense" criteria (for example, excluding profiles reporting a start date before the company's founding date). Profiles that don't meet these quality thresholds are excluded from our person to company matching process.

As a result, these types of "low-quality" profiles are not included in our headcount estimates.

Looking at the Anthropic company record in the PDL dataset, instead of the 839 profiles reported in the LinkedIn headcount, our data shows a more modest headcount of 483 employees, but a similarly steep 6-month growth rate of 75%.


```
"name": "Anthropic"
"employee_count": 483
"employee_growth_rate":
  "3_month": 0.2282
  "6_month": 0.7546
  "12_month": 1.691
  "24_month": 7.1186
```


(Source: PDL Dataset)

Although the PDL headcount is noticeably lower than the LinkedIn headcount, our filtering process ensures that the number reflects the **true count of "high-quality" employee profiles for this company**. Despite the lower headcount, our data still reports a steep headcount growth.

Our lower headcount number compared to LinkedIn aligns with an adjustment for the popularity bias affecting well-known LinkedIn company profiles.

To summarize:

- LinkedIn's headcounts are commonly inflated by a large number of low-quality profiles attracted to popular companies.
- In contrast, PDL filters out profiles that have missing, incomplete, or inconsistent information.
- While the filtering process results in lower headcount numbers compared to LinkedIn, our data still captures the same workforce trends. **We believe this approach provides a more accurate indicator of the true workforce trends within a company**.

Factor Two:

Person-to-Company Matching

Another equally impactful decision is the choice of how individuals are matched to companies. While this may seem trivial, there are important nuances to the approach that we will explore in this section.

LinkedIn's approach to matching employees to their respective employers is largely driven by the input that each LinkedIn user provides. This means that if an individual selects a company as one of their past or current employers, then LinkedIn will match that individual to the selected company with a high likelihood. However, in cases where a user doesn't select their employer, LinkedIn will still attempt to generate person to company matches using text-based matching logic.

The issues here stem from a combination of user error when selecting employers and the lenient string matching logic that LinkedIn employs to identify companies.

To see this in action, we can look at the company **Railway**, a cloud-infrastructure provider based in San Francisco.

Despite being founded in 2020 and raising $24M, LinkedIn reports several thousand associated profiles and a headcount of nearly 4,000. Tellingly, the company still self-reports its size range as 11-50.

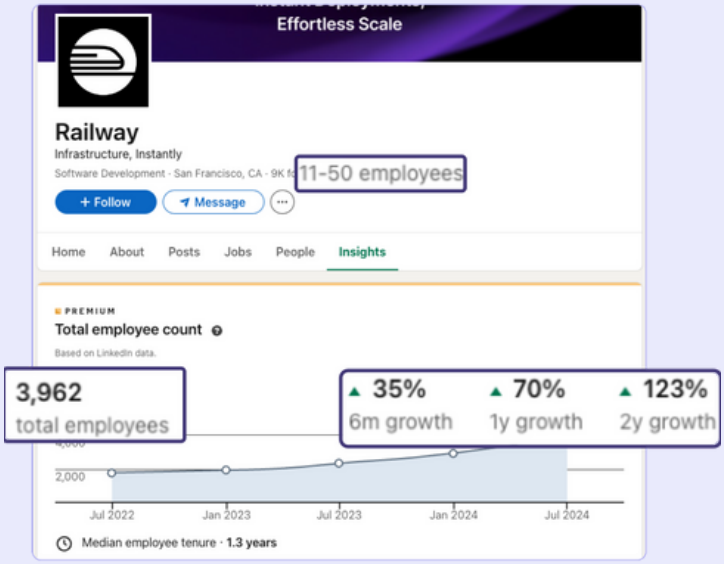


Figure 4
As of July 2024, LinkedIn reports total headcount of 3962 and a 6 month growth rate of over 35%. This contradicts the ~20 people on the About page and the "roughly 30 employees" number that they've shared with investors. (Source: LinkedIn)

To understand this better, we can look at the employee profiles associated with the company's LinkedIn page.

In Figure 5 below, LinkedIn shows thousands of employees based out of India with a variety of railroad and transportation related jobs.

As we've stated, LinkedIn's methodology for matching employees to companies is heavily influenced by which companies users self-selected when inputting their work information.

This is a clear demonstration of LinkedIn's sensitivity to **user-error**, which is especially likely when users from a non-English-speaking background interface with the platform's heavily-skewed distribution of English-speaking companies.

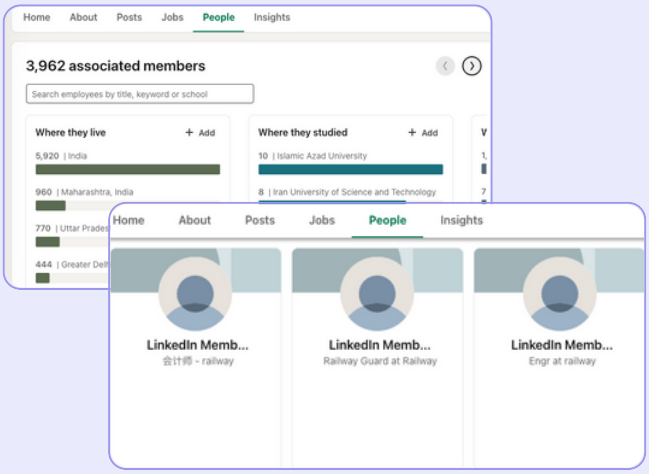


Figure 5
Employee profiles associated with Railway are largely based out of India, with a variety of transportation-related jobs that include the free text "railway." (Source: LinkedIn)

However, this issue is further compounded by LinkedIn's lenient text-based matching logic, which attempts to match user profiles to companies using the company name information provided in their work experiences.

As a result, even when users did not previously select a company profile as their employer, LinkedIn will still generate matches based on user-supplied text fields.

This can be seen in the example profile (Figure 6) which shows a user that did not select the Railway company profile in their work experience (as evidenced by the lack of a company logo in the work experience), but that LinkedIn still matched to the company using the "Railway" user-supplied text.

Figure 6
An example employee profile that was matched to the Railway company profile based on the user-supplied text, rather than selecting Railway when inputting their work experience. This is demonstrated by the lack of company logo next to their work experience. (Source: LinkedIn)

The result of these two factors is that thousands of international railroad workers were incorrectly mapped to the Railway company profile, dramatically inflating the headcount by several orders of magnitude.

While this may seem like an edge case, it is easy to find multiple examples of this in the LinkedIn dataset (see Appendix). **Overall, this example illustrates a clear, systematic bias in LinkedIn's headcount data and an important reason for caution when leveraging its data as a source of truth**.

PDL's Approach:

Strict Company Standardization

PDL's method of addressing this issue is our unique approach to company standardization (which we refer to as **canonicalization**). Our canonicalization process leverages information such as website, social profiles, and company name to deterministically identify companies from our company dataset.

We use canonicalization in our process of matching person records to company records, which helps us accurately determine if a person's employment information corresponds to a company in our dataset.

Just as importantly, this process also helps us identify person records with insufficient, inaccurate or contradictory employer information, which would not meet our threshold for canonicalization. Employer information that is unable to be canonicalized is excluded from our matching process and therefore excluded from our headcount calculations. This approach improves on LinkedIn's relatively simpler text-based matching logic, dramatically reducing our sensitivity to the failure mode demonstrated by the Railway example earlier.

Looking at the same record in the PDL dataset, we instead see a headcount of 35 and a growth rate of 9%.


```
"name": "Railway"
"employee_count": 35
"employee_growth_rate":
  "3_month": 0.0286
  "6_month": 0.0909
  "12_month": 0.2414
  "24_month": 0.7143
```


(Source: PDL Dataset)

This is the result of our stricter canonicalization thresholds which allows us to exclude the employees from our headcount calculation that were otherwise incorrectly mapped to the Railway company profile in LinkedIn's case. PDL will also still canonicalize employees with a display name that more clearly indicates their affiliation with Railway, (i.e. Railway.app) even if the employee does not directly tie themselves to Railway.

To summarize:

- LinkedIn relies on user input and basic string matching to match employees to companies.
- In contrast, PDL enhances this with probabilistic matching, resulting in more consistent and reliable matches at scale, especially for edge case profiles.

Factor Three:

Parent-Subsidiary Relationships

The third factor to consider when evaluating LinkedIn's headcount estimates is the case of companies with corporate subsidiaries.

When LinkedIn reports headcounts for companies with subsidiary entities, they provide a full aggregation of the employees at not only the parent organization but all of its subsidiaries as well. While this is standard practice among public companies reporting human capital disclosures to the SEC, this approach makes it difficult to isolate the headcount directly at the parent organization.

Importantly, this type of full aggregation dilutes important headcount signals within the parent organization.

For example, we can look at **LVMH**, a multinational luxury goods company with over 75 subsidiaries. LinkedIn reports a headcount of over 146,000 employees and a 6 month growth rate of around 4%.

In this aggregate view, trends across each of LVMH's subsidiaries are blended together. While this macro-level data is valuable, it loses the granularity of headcount trends within the parent organization itself.

Furthermore, for companies with many subsidiaries, LinkedIn's full aggregation across subsidiaries exacerbates the two inflationary biases presented in the previous sections.

Because LinkedIn includes the headcounts from each subsidiary, any biases present in the subsidiary headcounts will also be aggregated in the parent organization's headcount as well.

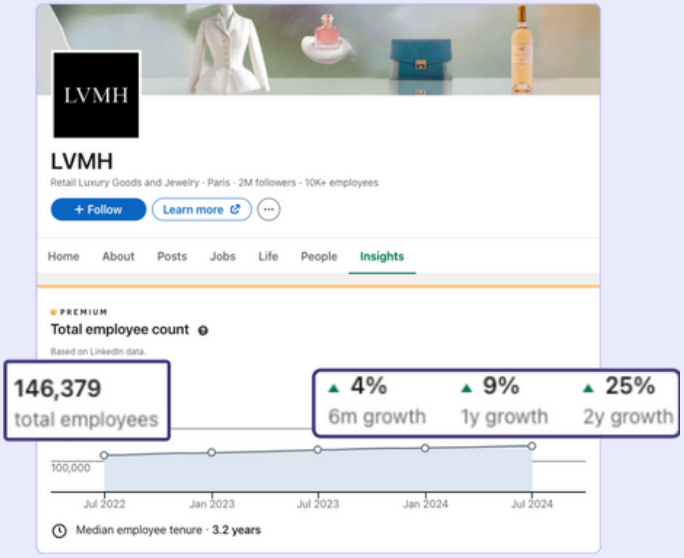


Figure 7
As of July 2024, LinkedIn reports total headcount of 146,379, and a 6 month growth rate of over 4%. (Source: LinkedIn)

PDL's Approach:

Accounting for Parent & Subsidiaries

In contrast, PDL's approach is to aggregate headcounts using only the direct employees within an organization and to **exclude employees from subsidiaries in the headcount calculation**. This approach reveals more nuanced signals around the headcount trends within an organization, while still allowing for a full aggregation across subsidiaries if desired.

Looking at the LVMH example in the PDL dataset, we see a dramatically smaller headcount of around 4,000 employees and a nearly flat 6 month growth rate.


```
"name": "LVMH"
"employee_count": 4596
"employee_growth_rate":
  "3_month": 0.00057
  "6_month": -0.0007
  "12_month": 0.0065
  "24_month": 0.0554
```


(Source: PDL Dataset)

The data here tells a much more granular story of a relatively stable and mature parent organization that is otherwise obscured when reporting the full headcount aggregation over the subsidiaries. In practice, we've seen this type of granularity be impactful across a variety of use cases from talent modeling to target account segmentation.

In addition, our data is designed to support reconstructing the full macro-scale headcount aggregation if desired.

This can be accomplished by using our **direct_subsidiaries** or **all_subsidiaries** fields, which can be iterated over to recover the full organizational headcount and growth trends.

To summarize:

- For companies with many subsidiaries, LinkedIn automatically aggregates headcounts across all subsidiaries, which is typical for public company reporting but obscures trends within the parent organization.
- In contrast, PDL reports only direct headcounts within an organization while still allowing for the reconstruction of total headcount across subsidiaries for more nuanced insights.

Let's wrap up.

Conclusion

While LinkedIn's headcount data is easy to rely on as a source of truth, we hope this report has highlighted several important limitations and biases present in LinkedIn's current headcount calculation methodology.

The key biases outlined in this report are:

- Limited filtering of low-quality employee profiles which impacts popular and well-known companies in particular.
- Lenient text-matching algorithms that magnify user error (particularly from non-English-speaking users).
- Combined parent and subsidiary headcounts that aggregate biases from the subsidiary organizations into the parent organization.

When combined, these biases lead to **inflated headcount estimates that dilute and obscure the true headcount trends** within an organization.

In contrast, the PDL headcount fields have been built with corrections that:

- Place stricter thresholds for matching person records to company records.
- Filter out low-quality profiles from our headcount estimates.
- Provide separate accounting for parent companies and their subsidiaries.

These efforts yield more representative headcounts and provide a better signal of the underlying growth trends occurring within each company.

If you are interested in learning more about the People Data Labs, our Company Dataset, and headcount data, visit our website and our documentation.

Appendix

In this section, we will take a look at some additional examples to supplement the discussions above.

Anyscale

LinkedIn Headcount

- **Employee Profiles**: 220
- **Self-Reported Size Range**: 51-200

Issues: Low-quality employee profiles associated with Anyscale

PDL Headcount

- **Employee Profiles:** 152
- **Self-Reported Size Range:** 51-200

How PDL corrects for this: Filtering out employee profiles that don't meet a sufficient completeness threshold

What's going on here?

The correct headcount as reported by one of Anyscale's investors (a PDL customer) is in-line with the PDL reported headcount ~150 rather than the 220+ headcount reported on LinkedIn. To understand the inflated LinkedIn headcount, we can look at the employee profiles on LinkedIn.

As shown in Figure 8, there are a non-negligible amount of low-quality user profiles, with only ~173 profiles appearing to be reasonably complete.

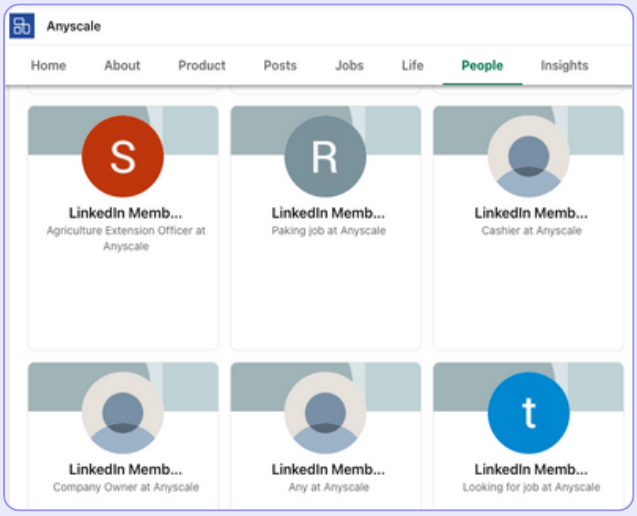


Figure 8
The low quality profiles associated with Anyscale's LinkedIn profile that are included in their headcount estimates. These profiles are representative of the bottom ~70 employee profiles. (Source: LinkedIn)

Privado

LinkedIn Headcount

- **Employee Profiles**: 230
- **Self-Reported Size Range**: 51-200

Issues: Low-quality employee profiles associated with Privado and user-error related to non-english speaking users

PDL Headcount

- **Employee Profiles:** 53
- **Self-Reported Size Range:** 51-200

How PDL corrects for this: Filtering out employee profiles that don't meet a sufficient completeness threshold

What's going on here?

Again, the correct headcount for this company as confirmed by a PDL customer is close to 60 (at the time of writing).

In this case, there are multiple biases leading to the grossly overestimated LinkedIn headcount. The first of these is the presence of low-quality profiles as before and seen in Figure 9. What's unusual is the dramatically higher proportion of low-quality profiles (nearly 77%) as compared to previous examples.

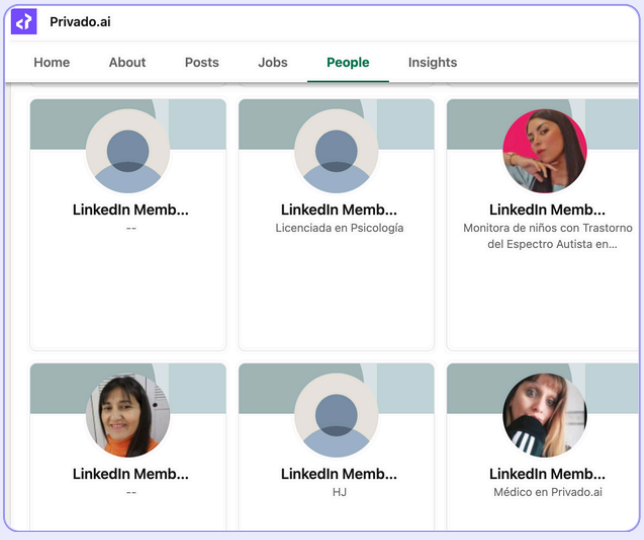


Figure 9
The low quality profiles associated with Privado's LinkedIn profile that are included in their headcount estimates. These profiles are representative of the bottom ~180 employee profiles. (Source: LinkedIn)

Looking at the geographic representation of these profiles reveals an overrepresentation of Spanish-speaking countries in the LinkedIn employee profiles.

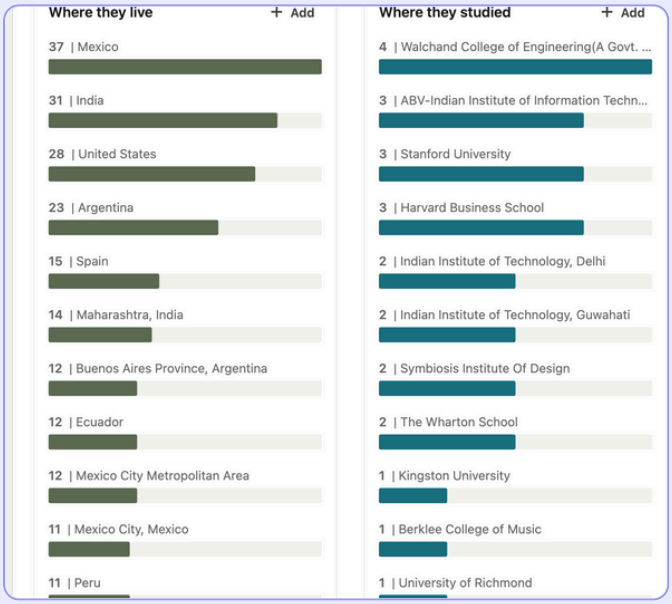


Figure 10
A large percentage of Privado employee profiles are primarily from spanish-speaking locations. (Source: LinkedIn)

Given that "privado" translates to "private" in Spanish, it's plausible that Spanish-speaking users entered "privado" as an attempt to hide the name of their employer, but as a result were attributed to the Privado company record by LinkedIn's fuzzy matching. This is similar to the Railway case noted earlier.