A proposal for characterizing funded companies using Dimensions and CrunchBase

Overview: This document outlines a proposed approach to characterize the companies supported by a U.S. government funder of scientific research. Relevant company characteristics include those that highlight the general composition of the funder's portfolio, the relevance of the funder's support for the company's success, adherence of the support to principles of equity, and their scientific and technical outputs. To this end, this proposal leverages the unique and powerful bibliographic database ***Dimensions***, and enriches it with company-level information sourced from the business information platform ***CrunchBase***.

Data Sources

Dimensions: Dimensions is one of the most comprehensive bibliographic databases in existence, containing, at the time of writing, information on over 127 million publications, 6 million grants, 11 million datasets, 200 million online mentions, 743 thousand policy documents, 695 thousand clinical trials, and 145 million patents. The authors of this proposal have full rights to access and use the full Dimensions database through the use of their official API and via Google's BigQuery. While Dimensions is powerful, its purpose is indexing scientific products, and thus it may not include information on companies with little public scientific activity.

Global Research Identifier Database The Global Research Identifier Database (GRID) uniquely identifies over 100 thousand research organizations around the world, including among them approximately 29,500 companies[1]. The overwhelming majority of indexed organizations contain metadata on their physical address and primary URL. Dimensions already incorporates GRID to disambiguate researcher affiliations and funding organizations, and so the procedure outlined in this proposal will focus on linking CrunchBase to GRID.

CrunchBase: CrunchBase is a business intelligence company that aggregates company level information from public data sources, data partners, and crowdsourced from users. Indexed data includes basic company metadata, information about total funding, investments made, employee information (particularly board of directors and Executive-level positions), technologies used, website traffic, and links to news sources. Basic data is free for any purpose with sufficient attribution [2], but only provides limited usage of the CrunchBase API and metadata. Complete enterprise-level API access can be purchased, but the costs this will incur are unknown pending contact of the CrunchBase sales department[3]. While full access will be necessary to derive final company-level features, basic access allows for bulk downloading of metadata and is sufficient for linking data sources. Once full access purchased, additional company information can be queried for each matched company directly using the CrunchBase API.

CrunchBase indexes far more information about many more entities than comparable data sources (e.g., OpenCorporates). However, it comes with known limitations. Much of the CrunchBase data is crowdsourced from their community, meaning that some records will be incomplete or inaccurate. Additionally, lacking ground-truth data on non-private company information (e.g., investments, technology), it is difficult, if not impossible, to appropriately validate the data's integrity. Still, for conducting high-level portfolio analysis, the advantages of CrunchBase easily make up for these shortcomings.

Data Linkage

Pre-processing: Before linking data, all relevant string fields in CrunchBase and GRID will be converted to lowercase and replacing Unicode characters with ASCII alternatives. Each data source will also be filtered to remove non-companies and companies based in countries not relevant to the funder.

URL matching: URLs are unique identifiers, and so identical URLs between CrunchBase and GRID will offer the most unambiguous match. To aid matching, URLs in both data sources will be normalized by removing leading and trailing characters, leaving only the base name, e.g., `http://www.digital-science.com/` would become `digital-science.com`. Successfully matched companies are excluded from the following steps.

Dimensions API matching: Next, potential matches will be determined based on company name and geographic metadata. First, a match is attempted using the `extract_affiliations` function implemented as part of the Dimensions API[4], which identifies possible matches between unstructured text and GRID identifiers. If an unambiguous match is made, then the pair is is considered successfully matched. If the match is ambiguous—returning more than one organization—then we consider steps to resolve the ambiguity.

Name matching: Multiple name matching strategies will be attempted on remaining unmatched publications. First, an exact name match will be attempted between the GRID and CrunchBase. Modified exact matching will then be attempted on the unmatched companies, which repeats the exact matching procedure

[1] Regularly updated statistics of GRID coverage can be found at `https://www.grid.ac/stats`
[2] Attribution must state "Powered by Crunchbase" and must contain a link to the CrunchBase domain.
[3] `https://about.crunchbase.com/products/crunchbase-enterprise/`
[4] `https://api-lab.dimensions.ai/cookbooks/8-organizations/1-GRID-preview.html`

but after removing common suffixes (e.g., "ltd.", "inc.", "co."). Next, the remaining companies will undergo fuzzy string matching based on the weighted Jaccard similarity between names; this approach measures similarity by comparing words in each string and weighting matching words by their occurence across all strings (e.g., "Corporation" is a common word, and thus will receive a low weight). The top five matches for each company in terms of Jaccard similarity will be returned and the final match resolved in the following steps.

Resolving ambiguous matches: Because name-based matching is imprecise, or even because multiple companies may share the same names, an additional procedure is needed to resolve these to one-to-one matches. Specifically, geographic metadata is used to identify a final match. For all potential record pairs, exact string matching is attempted based first on city name, and it that fails, then on state or region. If necessary, selection can also be made by comparing the founding dates of the companies across the two data sources. Pairs that remain ambiguous will be manually resolved.

Validation: This matching procedure will be validated by randomly sampling at least 20 matched company pairs assigned by each strategy outlined above and manually annotating them as correct or incorrect. Doing so will provide an overall sense of the integrity of linkage, as well as how different linking strategies compare.

Product: The entire data linkage workflow outlined here will be implemented in python and incorporated into an automated `snakemake` workflow that will allow easy edits and expansion of the data into the future. The product will be a crosswalk table linking GRID Ids with CrunchBase records. This crosswalk will be stored on Google's BigQuery to allow easy linkages with Dimensions.

Caveats: Inevitably, there will be unmatched records or records which cannot be adequately resolved; these records will be marked and reported on in future analyses. Some records may simply not appear in GRID or CrunhBase. In particular, GRID's focus on research organizations is likely to underrepresent companies which do not engage in public scientific activity. When no matches can be made for a funder's company, additional information will need to be provided in an attempt to identify corresponding data records.

Derived features

Basic features: Basic company metadata will be derived from CrunchBase, including the company's location, their number of employees (a categorical variable, e.g., "251-500"), the founding year of the company, and its CrunchBase industrial classification as multiple selections of 47 coarse and 744 granular categories.

Leadership demographics: The names the leadership in each company will be identified from CrunchBase by extracting the first names of employees with terms like "director", "VP", or "Chief" in their job titles. A gender assignment will be made for each employee based on their first name using `genderize.io`, which infers genders based on the gender distribution of names in the U.S. census.

Financial features: CrunchBase provides information on total funding received by a company. When paired with information from the funder and Dimensions, the size of the funder support relative to the company's total funding can be derived.

S&T Outputs: Data will be sourced from Dimensions to quantify each company's scientific and technical outputs, including their publication count, scientific impact (measured by citations), social impact (measured by Altmetrics score), their published patents, datasets, and clinical trials.

Deliverables

Raw data: The raw data files for matched companies with all relevant metadata will be provided to the funder in a standard text format for use and analysis as they see fit, so long as they abide by the policies of CrunchBase[5] and the contract with Dimensions.

Interactive dashboard: An interactive dashboard will be created based on the raw matched data which will allow the funder to explore the data and derived features. Additionally, by drawing on the full content of Dimensions and CrunchBase, this dashboard will provide quantitative comparisons between the matched companies and other organizations, allowing the funder to explore how their portfolio compares to science and business writ large. This dashboard will be developed using the library `Shiny` written in R and deployed to `shinyapps.io` and made accessible online with a password; hosting costs will depend on the size of the final dashboard and the funder's expected hours of usage per month.

Summary report: A brief summary report will be provided alongside the interactive dashboard, outlining highlights of the analysis and discussing their significance in the context of the domain and methodological expertise of the team at Digital Science & Research Solutions ltd.

Risks and caveats: The production of these deliverable depends on the success of the data linkage procedure, which in turn depends on CrunchBase's data quality and the presence of the funder's portfolio companies in each data source. In the case that a company can only be matched to one data source, then features derived only from that resource will still be reported. Due to the messiness of CrunchBase's crowdsourced data, final results must be considered with caution, and interpretations should be made using aggregates of data, rather than at the level of individual companies.

[5] https://about.crunchbase.com/terms-of-service/