# MACRE

## Medical and Academic Collaboration Recommendation Engine

Authors: Santiago Buenahora, Garrett Darley, Kunal Garg, Nicholas Keenan
Advisors: Ani Nenkova

### Abstract

The goal of MACRE is to connect academic researchers to foster more effective collaboration, find experts in a given field, and characterize the research being done at large research organizations. MACRE examines a large set of published research from the University of Pennsylvania and the Children's Hospital of Philadelphia (CHOP). We use machine learning algorithms to determine the similarity of publications relative to each other. Using these weights and the network of past collaborations, we recommend future collaborators for each author. Using our tool, users can easily find authors and publications by their name, subject of research, and publication keywords.

### Motivation

Over the course of years in quality research and development, the University of Pennsylvania Children's Hospital of Philadelphia (CHOP) has established a large research network via natural cross-collaboration. Medical researchers rely largely on social circles to find suitable colleagues for collaboration. This becomes an issue at large research organizations such as Penn and CHOP, which have 3000+ researchers combined.

Our team found this problem by being connected to the SCOSY team at CHOP, which authored a paper about a better biomedical collaboration system. The SCOSY project was proof to us that CHOP was invested in the idea of tech-enabled research recommendations. This gave us motivation for MACRE because it proved to us the real need for a software tool to further medical research collaboration.

### Technical Approach

**Recommendation Engine**

1. Publication Similarity: The recommendation engine searches for the papers that are most similar to the target author. How papers are assigned a similarity weight is described below. Authors of these papers are assigned a weight proportional to this similarity. If an author has several papers that are very similar to the target author's, then this author is assigned a weight proportional to the most similar paper.

2. Past Collaboration Network: The engine then uses the network of past collaborators to recommend future ones. Authors with smaller degrees of separation from the target author are given higher weights.

3. Combine Weights: The different weights are combined for each recommendation. The engine then sorts and returns the highest ranked recommended collaborators. The recommendation engine does not recommend past collaborators because these are trivial recommendations.

**Project Pipeline**

1. Pubmed Database: We use the Entrez Programming Utilities, the public API to the NCBI Entrez system, to query the PubMed database for author and publication data. We filter the data for only authors in our target organizations, in this case those are UPenn and CHOP, and store their data and their publications.

2. Data Cleaning: The data from PubMed has several issues. Authors are not associated with any unique identifiers, so there are many duplicate authors. We attempt to combine these duplicates based on their

name and organization. There may be some false positives, where we may combine two authors who are actually distinct, but it is not common nor particularly debilitating towards the success of the product. During this stage, we assign authors unique identifiers to be used later in the pipeline. Many of the publications are missing MeSH terms. We search the Other Terms and the publication's abstract for any matching MeSH terms. Finally, we remove any abstracts that are 'noise' and not actually descriptive of the paper.

3. Preprocessing: We calculate the similarity of each document relative to the other. We use two separate measurements for similarity. The first uses Term Frequency-Inverse Document Frequency (TF-IDF) to measure the similarity of two documents based on their abstracts. Because not every publication we get from PubMed has an abstract associated with it, we use a second measurement based on the subject list of the publication. This measurement is found using the same TF-IDF approach.

4. MACRE API: The backend of MACRE responds to search queries with a list of matching authors and publications. Searches can be made for the author name, publication MeSH term, or publication keywords. The MACRE API also hosts the recommendation engine. Recommendation requests are made for a given author. The response contains a list of recommended collaborators mapped to the recommendation weight the engine assigns to them.

5. Web Application: The web application gives users an easy way to search for authors and publications and find recommendations. We also included a visualization element in order to see the the tree of mesh terms that exist within the database.

## Design & UI

1. Opening Screen: To create a minimal and understandable user experience for our users our landing page has two key components:
   a. Search: The first visual component of the screen is a simple search bar. Next to the search bar is a toggle button to navigate the search type and "search" button that sends the final query with the keyword and the search type. These three components utilize React's Boostrap library to make a clean user interface.
   b. Zoomable Burst: This robust visualization is used to help the user navigate the MeSH term tree hierarchy. This component utilizes D3 libraries[1]. This component features a circular set of rings. Each ring represents a level in the hierarchy tree. Each ring is broken into several slices, where each slice is a node for that specific level's tree. The rings closer to the center indicate a higher level than those closer to the outside of the center. Clicking on a specific slice will recalculate the rings such that the ring closest to the center is the selected slice's direct children and the successive rings moving outward are the levels in the hierarchy below. Clicking on the center of this component allows the user to jump up to a level such that the current slice will now exist as a part of the ring in the innermost circle.
2. Results: After a query is sent and a response has received the results are organized into two main categories, resulting in the authors and publications:
   a. Authors: The authors are organized in a clean and organized table that has table fields of "Name", "Roles", "More Info." Each row in the table lists serves a single author. The Name column simply lists the author's name in "LAST_NAME, FIRST_NAME" format. The roles column consists of bootstrap pills that are and labeled to represent the number of times the author served in a specific role for a unique paper. For example, if the author was the "Chief Author" in five papers and the "Ordinary Author" in two other papers, there will be two pills in the Roles column: one green with the text "Chief Author (5)" and another blue with the text "Ordinary Author

[1] https://observablehq.com/@d3/zoomable-sunburst

(2)”. Finally, there is a “More Info” column which has a button that has the label: “View Recs.” By clicking on this button a user will open a modal that displays the results of the recommendation engine for the specific author that corresponds to the row. In the modal the recommended collaborator authors are listed in decreasing order from authors with the most weight highest recommendation at the top. A horizontal slider or  “progress bar” helps visualize the weight as a percentage.

b. Publications: The publications tab has 5 fields in the following order: “Title”, “PMIDs”, “Author List”, “MeSH Terms”, and “PubMed Link”. The results of the search list the publications that have some connection to the query, where each row is a different publication. The MeSH terms are listed if there are any associated values and the column is left blank if the query does not have any associated values. The PubMed Link is hyperlinked to the actual PubMed article’s webpage,

## Evaluation

### Methodology

In order to test the accuracy of our model, it wasn’t feasible in the time we had to record the actual effectiveness of the recommendation engine in real cases. So we used the assumption that past collaborations were successful collaborations. To evaluate our model, we would select a target author for the test. For each of the past collaborators of this author, we removed the target author’s connection to any overlapping publications. In this altered data set, the two authors would no longer be collaborators. We then ran the recommendation engine for this author using this altered data set. The test was successful if the original collaborator was highly recommended.

### Limitations

Not every author can be tested using our evaluation method. Each author that we test is required to have more than one publication. More specifically, a test is only possible if the author has publications remaining after the overlapping ones are removed. Otherwise, we have no data to use in our recommendation.

## Findings

### Results From Method 1 Evaluation

There are two ways to interpret the evaluation results. The first is evaluated per target author. The test is successful if at least one of their past collaborators was returned as a recommended collaborator in our test. The percentage of test that passed with this evaluation is approximately 90%. This is expected as considering most research isn’t authored by a single person, connections are more likely.

### Results From Method 2 Evaluation

The other interpretation is evaluated per past collaborator, rather than the target author. So we measure how many collaborators are successfully recommended in our tests per target author. Our recommendation engine is successful about 30% of the time with this evaluation method.

## Ethical Considerations

### Open Sourced Data

One of the primary concerns with our platform is sourcing data from private databases. If our platform used papers that were not publicly available, then our platform could potentially allow users to gain insights on some researcher’s private work. However, MACRE utilizes the PubMed API to collect publicly available information, mitigating this major risk.

### Bias Towards Established Authors

Our recommendation model gives better recommendations when the target author has more publications. Similarly, an author is more likely to be recommended if they have more publications. This can put new researchers at a disadvantage.

### Ethical Considerations of Recommendations

One of the major risks we run with making recommendations is the assumption that any individual is willing to collaborate with any other researcher in the network. Our platform does not take into consideration whether the recommended individuals have left the industry, have personal conflicts with individual making the recommendation search, or disagree with the hypothesis made by the individual researcher.

## Product Vision

In our meeting with CHOP, one of the questions they asked us was “if a venture capitalist gave you one million dollars for this project, what would you turn this into”. In the world of product management, the term North Star is often used to describe exactly this case. If the project had infinite resources, time, and money, what would it look like that is aligned with the project’s original mission? Understanding the North Star is important because it allows us to have a clear vision of what we are working towards and what the project may look like in the best case scenario.

Our team thought of our North Star as follows. If given significant resources, we would build a product that is useful not only when grants come and researchers need collaborators, but also on a daily basis. At its core, our project built the MVP for inputting a single researcher and outputting a community of peer researchers. This is useful for more purposes than just setting up researcher collaboration. This portal could be a resource for a community of researchers in one area, to be updated on each other’s progress and collaborate on a digital forum on a daily basis. By centralizing researchers in an area in one place, updating as new research pieces become public, and providing a forum to chat, our platform would provide huge value by enabling collaboration and by keeping researchers up to date. Another value of this project is that it can be used by administrators to keep a pulse on the researchers in the organization and all the research topics they are covering. This may be useful for tracking the quality of work and research performance.

All in all, this project is an example of how software can transform coordination in research organizations. While our project dwarfs in comparison to the north star, by understanding the vision, we were motivated and made better decisions.

## Business Plan

### Stakeholders

Currently, our major stakeholders include Jorge Marin, Dr. Winston, CHOP, University of Philadelphia Hospital for whom we are building this platform. By creating a versatile recommendation platform, we can increase impactful interactions between PI’s and other faculty.

### Market Opportunity

#### Customer segment(s)

We have identified a few key customer segments. Currently, we are viewing our customer segments as the parties that connect grants with research faculty, individuals in the medical field looking to expand their network, and researchers looking for mentorship opportunities. Potentially, individuals looking to do a literature review on a certain topic could use our platform as a springboard for further research.

#### Estimated of Size and Growth of Market Segment

The market segment that we have identified has an estimated size and growth that is a large untapped potential. Millions of research dollars and faculty hours are invested in improving the health of our planet. However, getting the right people for specific jobs can be difficult. The NIH alone spends ~39.2 a year on research [2]. According to the American Hospital Association, there are 6210 hospitals in the U.S. alone[3].

### Competition

#### who are they?

Our biggest competitor is the simple word of mouth action that takes place when a peer or colleague asks for a collaborator recommendation or reference. Even one of our major stakeholders relies heavily on his personal connections and knowledge of the intricate network of researches. ScholarSearch is a platform that targets health informatics and is referenced as an exception to the traditional recommendation systems currently existing as competition. Finally, one of the most popular open-source platforms, Profiles Research Network Software (RNS), creates “career snapshots”, combining directory information, user-contributed content and publications content extracted from PubMed (Marin).

[2] https://www.nih.gov/about-nih/what-we-do/budget

[3] https://www.aha.org/statistics/fast-facts-us-hospitals

How are they addressing the problem?

As per Jorge Marin’s paper on Scosy A Biomedical Recommendation System[4], there are three main algorithms backing traditional recommendation systems: 1. Collaborative Filtering, 2. Content-Based Filtering, and 3. Hybrid Filtering.

What is it that differentiates you?

Various factors help us differentiate ourselves from key market players. The two major limitations of the existing platforms are the following and have been listed in *SCOSY: A Biomedical Collaboration Recommendation System*: 1. The algorithms use publication data and set a PI or faculty’s expertise based on a few set keywords 2. “It requires administrative privileges to extend the functionality of the system to include ontology-based semantic web features”

SWOT Analysis

| | **Helpful** | **Harmful** |
|---|---|---|
| **Internal** | Strengths<br>● Jorge & Asif’s domain expertise<br>● Existing Qlikview platform & ML algorithm<br>● Access to CHOP proprietary data<br>● **Collaborator** recommendation rather than **research paper** recommendation.<br>● Santiago’s webapp building experience | Weaknesses<br>● Data restrictions by the department<br>● Novice skills in LDA/ML algorithms for all Senior Design teammates and Python for select members of the team |
| **External** | Opportunities<br>● Access to various data sources | Threats<br>● Emerging competitors |

| | | |
|---|---|---|
| | PubMed, Penn/CHOP collaborations<br>● MeSH labeling | ● ScholarSearch (targets health informatics) - does collaborator recommendation system.<br>● Open-source platforms such as Profiles Research Network Software (RNS) |

Cost

Overall our main cost will be website maintenance and hosting. In initial conversations with CIS400 and CHOP faculty, the university will provide us initial AWS credits to host our platform on Amazon Web Services. Moving forward, Jorge Marin, has mentioned that we could make a pitch to CHOP to see if they will be willing to allocate funds for this maintenance.

Revenue model

A few revenue models can be considered for this comprehensive platform.

1. Subscription-based model for individual users: In the case that there are individuals with a single sign-on to our platform, subscription access can be provided to people who need to frequently make searches.
2. Enterprise/Institution subscription: For research institutions like universities or hospitals, a larger access fee can be charged for multiple accounts.
3. Banner Ads - Google AdSense

## Conclusions

Our team built a web application that recommends biomedical researchers for collaboration. Our project is a minimum viable product for CHOP to explore the utility of tech-enabled collaboration matchings

[4] Guerra, Jorge & Quan, Wei & Li, Kai & Ahumada, Luis & Winston, Flaura & Desai, Bimal. (2018). SCOSY: A Biomedical Collaboration Recommendation System. Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference. 2018. 10.1109/EMBC.2018.8513268.

One can think of our project as one of the many that will shape how large research organizations will use software to dramatically improve their processes in allocating resources, both funds and researchers. One can imagine that in ten years, an organization like CHOP will use software to keep a pulse on the productivity of its researchers, to make sure ongoing research is up to date and not duplicating existing work, to communicate its research efforts and updates with similar institutions, and to match diverse research talents to further the biomedical frontier.

The ultimate goal of our project is to make CHOP more efficient and to have better workflows. As of this report, our team has gotten positive responses from a CHOP associate. We've also scheduled a handoff meeting with CHOP so that they can start using our work. We're hoping that this tool will make it easier for CHOP to coordinate research funds and collaboration.