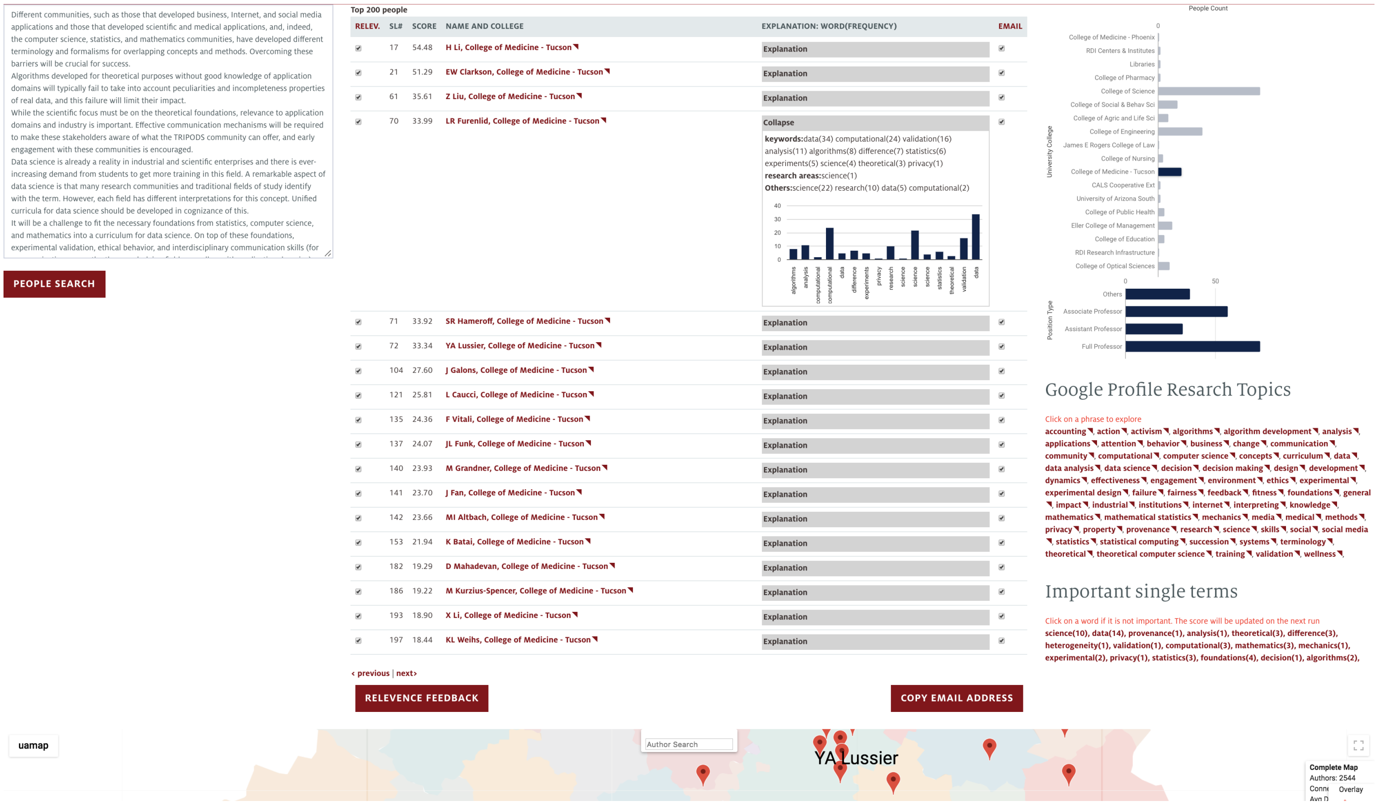
REMatch: Research Expert Matching System

Md Iqbal Hossain[1], Stephen Kobourov[1], Helen Purchase[2], Mihai Surdeanu[1]
{hossain,kobourov,msurdeanu}@cs.arizona.edu, helen.purchase@glasgow.ac.uk
[1]Department of Computer Science, University of Arizona, USA
[2]School of Computing Science, University of Glasgow, Scotland



Different communities, such as those that developed business, Internet, and social media applications and those that developed scientific and medical applications, and, indeed, the computer science, statistics, and mathematics communities, have developed different terminology and formalisms for overlapping concepts and methods. Overcoming these barriers will be crucial for success.
Algorithms developed for theoretical purposes without good knowledge of application domains will typically fail to take into account peculiarities and incompleteness properties of real data, and this failure will limit their impact.
While the scientific focus must be on the theoretical foundations, relevance to application domains and industry is important. Effective communication mechanisms will be required to make these stakeholders aware of what the TRIPODS community can offer, and early engagement with these communities is encouraged.
Data science is already a reality in industrial and scientific enterprises and there is ever-increasing demand from students to get more training in this field. A remarkable aspect of data science is that many research communities and traditional fields of study identify with the term. However, each field has different interpretations for this concept. Unified curricula for data science should be developed in cognizance of this.
It will be a challenge to fit the necessary foundations from statistics, computer science, and mathematics into a curriculum for data science. On top of these foundations, experimental validation, ethical behavior, and interdisciplinary communication skills (for

Top 200 people

RELEV.	SL#	SCORE	NAME AND COLLEGE	EXPLANATION: WORD(FREQUENCY)	EMAIL
☑	17	54.48	H Li, College of Medicine - Tucson ◥	Explanation	☑
☑	21	51.29	EW Clarkson, College of Medicine - Tucson ◥	Explanation	☑
☑	61	35.61	Z Liu, College of Medicine - Tucson ◥	Explanation	☑
☑	70	33.99	LR Furenlid, College of Medicine - Tucson ◥	Collapse **keywords:**data(34) computational(24) validation(16) analysis(11) algorithms(8) difference(7) statistics(6) experiments(5) science(4) theoretical(3) privacy(1) **research areas:**science(1) **Others:**science(22) research(10) data(5) computational(2)	☑
☑	71	33.92	SR Hameroff, College of Medicine - Tucson ◥	Explanation	☑
☑	72	33.34	YA Lussier, College of Medicine - Tucson ◥	Explanation	☑
☑	104	27.60	J Galons, College of Medicine - Tucson ◥	Explanation	☑
☑	121	25.81	L Caucci, College of Medicine - Tucson ◥	Explanation	☑
☑	135	24.36	F Vitali, College of Medicine - Tucson ◥	Explanation	☑
☑	137	24.07	JL Funk, College of Medicine - Tucson ◥	Explanation	☑
☑	140	23.93	M Grandner, College of Medicine - Tucson ◥	Explanation	☑
☑	141	23.70	J Fan, College of Medicine - Tucson ◥	Explanation	☑
☑	142	23.66	MI Altbach, College of Medicine - Tucson ◥	Explanation	☑
☑	153	21.94	K Batai, College of Medicine - Tucson ◥	Explanation	☑
☑	182	19.29	D Mahadevan, College of Medicine - Tucson ◥	Explanation	☑
☑	186	19.22	M Kurzius-Spencer, College of Medicine - Tucson ◥	Explanation	☑
☑	193	18.90	X Li, College of Medicine - Tucson ◥	Explanation	☑
☑	197	18.44	KL Weihs, College of Medicine - Tucson ◥	Explanation	☑

Google Profile Resarch Topics

Click on a phrase to explore
accounting ◥, action ◥, activism ◥, algorithms ◥, algorithm development ◥, analysis ◥, applications ◥, attention ◥, behavior ◥, business ◥, change ◥, communication ◥, community ◥, computational ◥, computer science ◥, concepts ◥, curriculum ◥, data ◥, data analysis ◥, data science ◥, decision ◥, decision making ◥, design ◥, development ◥, dynamics ◥, effectiveness ◥, engagement ◥, environment ◥, ethics ◥, experimental ◥, experimental design ◥, failure ◥, fairness ◥, feedback ◥, fitness ◥, foundations ◥, general ◥, impact ◥, industrial ◥, institutions ◥, internet ◥, interpreting ◥, knowledge ◥, mathematics ◥, mathematical statistics ◥, mechanics ◥, media ◥, medical ◥, methods ◥, privacy ◥, property ◥, provenance ◥, research ◥, science ◥, skills ◥, social ◥, social media ◥, statistics ◥, statistical computing ◥, succession ◥, systems ◥, terminology ◥, theoretical ◥, theoretical computer science ◥, training ◥, validation ◥, wellness ◥,

Important single terms

Click on a word if it is not important. The score will be updated on the next run
science(10), data(14), provenance(1), analysis(1), theoretical(3), difference(3), heterogeneity(1), validation(1), computational(3), mathematics(3), mechanics(1), experimental(2), privacy(1), statistics(3), foundations(4), decision(1), algorithms(2),

Figure 1: A screenshot of the REMatch system: input text query from a call for proposals (top-left), list of matching research experts along with explanations (top-middle), college filter bar chart, position filter bar chart, research topics and matching terms (top-right), map with matched researchers from the selected college.

***Abstract*—We describe a system designed to process, analyze and visualize academic data, from research papers and research proposals to list of courses taught, consulting, internal and external service. This can be helpful in identifying experts in a given field for future collaborations, as well as in putting together strong multi-disciplinary teams to apply for future research funding. Our REMatch system aims to support such tasks by leveraging natural language processing, machine learning, and interactive visualization. Specifically, REMatch provides a functional system that implements in-the-browser, map-based interactive navigation of a large underlying network, supporting semantic zooming, panning, searching, and map overlays. A prototype of the system is evaluated with a small-scale case study.**

I. Introduction

The explosion of data leads to unexpected effects, where solutions to critical problems are overlooked, since many tasks need to cross several disciplines or domains that produce considerable amounts of data but interact only minimally. Swanson predicted this problem more than three decades ago, and described it as *undiscovered public knowledge*:

> "Knowledge can be public, yet undiscovered, if independently created fragments are logically related but never retrieved, brought together, and interpreted." [46].

In this work we focus on the complex task of matching academics to calls for proposals. This task impacts universities and other research-oriented organizations, which must constantly answer difficult questions such as: How can we identify experts in a given field? How can we identify gaps in our strategic areas of expertise? How can we match calls for proposals with the right set of experts? How can we put together a strong multi-disciplinary team to apply for big research proposals?

Importantly, although these are complex tasks, there is great deal of data from sources such as research papers, research proposals, and courses taught that can be used to address them. This information can also be used to identify current strengths and weaknesses, as well as finding patterns and trends over time. The main challenge and our major contributions are in gathering and processing the needed data, putting together a collection of new and existing tools, designing an intuitive interactive interface, and packaging all of this in a functional system that non-experts can use.

A. Context

Encouraging research experts (tenure-track research faculty and research scientists) to apply for external funding is an important task for university research offices. This task is simultaneously very difficult (given thousands of research experts across dozens of colleges and hundreds of departments), while also of critical importance (given the major role of research funding in the face of steady cuts in state and nationwide education funding). In the rest of this paper we focus on the University of Arizona (UA), although the design principles can be applied to other universities.

The Research Office (RO) staff collect information about calls for funding from a wide range of sources (e.g., federal funds, industry proposals, and named foundations), with the aim of alerting research experts for whom the call is relevant. In particular, they would like to encourage people who might not have considered the call as being one for which they should apply, or who might have missed the call altogether.

The continuous stream of information (calls for funding proposals, information requests, white papers) is monitored and filtered by the RO staff, and a weekly digest is emailed to all university research experts. Small specialist teams (e.g., physical sciences, arts and humanities, clinical and biomedical) consider the calls in their particular area, and, after filtering, do one or both of two things: (a) they forward the calls to the relevant College Dean of Research (e.g., Engineering, Arts and Humanities, Science) asking them to in turn forward to appropriate faculty in their colleges, and/or (b) they directly contact particular individuals who are known to have specialist knowledge in the area relating to the call. When contacting individuals directly, RO staff rely on internal information (faculty profiles) and commercial tools (e.g., the PIVOT database, https://pivot.cos.com/). Despite these resources, RO staff readily acknowledge that a major source of knowledge about the research expertise of the institution's researchers is "in their heads" – personal knowledge that they have built up over time through networking events and individual contacts, and which is lost when RO staff members change jobs.

It is the process of "finding the perfect fit" that the RO staff welcome support with – knowing exactly who would be the best people to target and encourage to respond to a particular call for funding. They would like to reduce the amount of targeting emails that they send (so as to cut down on emails considered "spam" by the university research experts), while being confident that the sources of information that they use in this targeting are robust enough to ensure both high precision (the people targeted are appropriate) and high recall (appropriate people are not omitted).

B. The REMatch System

The Research Expert Matching (REMatch) System relies on data gathered from university databases (e.g., current staff, research proposals), as well as external sources (e.g., research publications, research funding awards). This data is analyzed using machine learning (ML) and natural language processing (NLP) components and visualized with a interactive map-based network and overlays. In particular, given a specific call for proposal (CFP) or a research paper (or even a bit of relevant text from a CFP or paper), we can find the research experts who best match the query and locate them on the map. The query can be refined by specifying one or more colleges (e.g., show only experts in the College of Science), and further refined by marking some of the phrases and words from the input as "stop-words" (i.e., contentless words). Each person matched is associated with an explanation, showing which topics and terms were matched, and therefore explaining why that person has been included in the list. Exploring the map can show additional information, such as past collaborations, past funding data, and citation counts.

We gather data from different sources (several university databases, as well as external ones such as Google Scholar) to build the collaboration network. We then use multi-level force-directed placement, node overlap removal, and clustering algorithms to represent the network as a map. Nodes, node labels, polygon colors, and edges are transformed into Google Map objects, which are then drawn in the browser using the Google Maps API. Seven different level-of-detail (semantic zoom levels) are precomputed, which entails determining which nodes are present on a given level, computing label font sizes, and ensuring no label overlaps. To process a text query and find research experts that match it, we extract research topics and terms associated with the query. Then we perform multiple information retrieval queries on a precomputed Lucene index, which includes information about each expert's research topics, publication details, and grant proposals. Figure. 2 shows an overview of our system.

The REMatch system provides several novel features: (a) leveraging knowledge and information from different domains to find relevant research expertise and research experts; (b) combining machine learning, natural language processing and visualization techniques in one functional system that (unlike other stand-alone graph-drawing tools) is implemented in the browser, and is therefore readily accessible; (c) providing interactive map-based visualization of the network of research experts and supporting common map-exploration interactions such as multiple zooming levels, semantic zooming, panning, and searching; and (e) overlaying additional information such as collaborations, citations, and funding levels. The current prototype of REMatch is already used by senior managers at the university to identify potential new research collaborations and to create multi-disciplinary research teams. As an open-

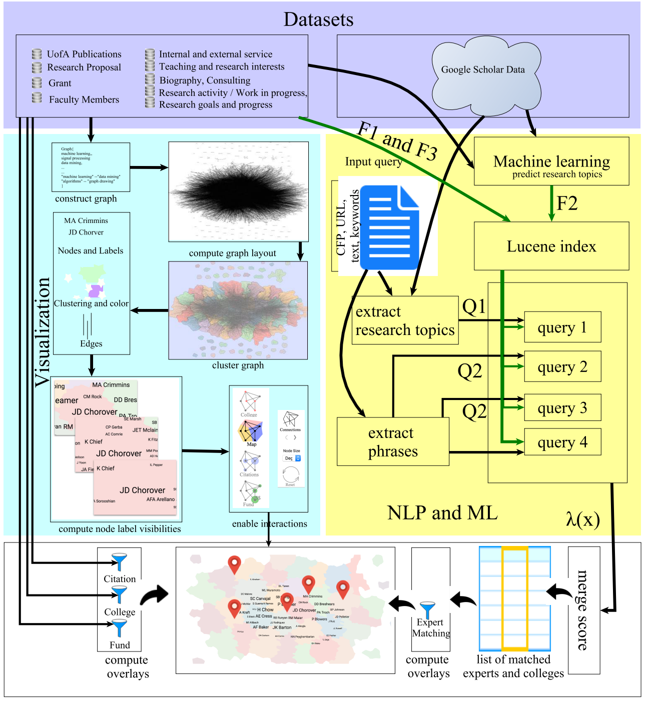


Figure 2: Overview of the REMatch system.

source system, we expect that REMatch can be customized to different institutions and that additional features will make it useful beyond the scope of the current prototype.

II. Related work

There is a great deal of related work in different domains: from collaboration networks and topic analysis, to visualization techniques for text data and large graphs.

Research collaboration networks are widely studied [24], [33]. Words from paper titles have been used as indicators for the content of a research topic, which in turn are used to create topic visualizations [17], [47], [56]. Some earlier approaches focus on analyzing specific journals, conferences, or research areas, e.g., analyzing computer science conferences and journals [17], trends in computer science research [15], the International Conference on Data Mining (ICDM) [30], publications in data visualization [21]. Domenico *et al.* [11] quantify attractive topics (i.e., topics that attract researchers from different areas). Sun *et al.* [44] build a network, with computer science conferences as nodes and edges between two conferences with common authors. Map-based visualization has been used for document visualization [22], [42], [53]. Dunne *et al.* [13] built a tool to combine reference management, citation text, automatic summarization, ranking and filtering, and network visualization for documents.

Citations are considered an important contribution measurement [55] and are used in visualizations of scholar profiles [35] and paper recommendation systems [51]. Citation data from the Web of Science [34] and from Microsoft's Academic Graph [25] have been analyzed. CiteRivers [21] and Cite-VIS [43] analyze and visualize IEEE VIS conference citations, as do Ke *et al.* [26].

Also related to our work are many graph visualization techniques and tools; see surveys by Von Landesberger *et al.* [49] and Vehlow *et al.* [48]. Graph layout algorithms are provided in several libraries, such as GraphViz [3], OGDF [10], MSAGL [32], and VTK [40], which however, do not support interaction, navigation, and data manipulation. Visualization toolkits such as Prefuse [20], Tulip [7], Gephi [8], and yEd [52] support visual graph manipulation, and while they can handle large graphs, their rendering does not: even for graphs with a few thousand vertices, the amount of information rendered statically on the screen makes the visualization difficult to use. Further, there are research papers that describe interactive multi-level interfaces for exploring large graphs such as ASK-GraphView [6], topological fisheye views [18], and Grokker [37]. Software applications such as Pajek [12] for social networks, and Cytoscape [41] for biological data provide limited support for multi-level network visualization. These approaches rely on meta-graphs, made out of meta-vertices and meta-edges, which make interactions such as semantic zooming, searching, and navigation counter-intuitive. Not many of the tools and systems above provide browser-level navigation and interaction for large graphs.

Liu *et al.* collect publications and projects of potential supervisors from internal data sources of University of Leeds so that new PhD applicants can find potential supervisors [27]. As this approach relies on manually defined rules, it does not easily extend to larger scale datasets such as ours. Automated recommender systems such as blended recommending [28], expertise recommender [29], ruled-based mapping [31], and group recommender systems [50] deal with large-scale problems such as ours, but are not clearly applicable to academic publications, research grant proposals, and course descriptions.

Commercial organizations provide access to (and, in most cases, visualizations of) data about research activity for an institutional fee. SciVal (Elsevier, www.elsevier.com/solutions/scival) "offers quick, easy access to the research performance of 8,500 research institutions and 220 nations worldwide," and Academic Analytics (www.academicanalytics.com) focusses on research universities in the United States and the United Kingdom, specifically supporting "the strategic decision-making process as well as a method for benchmarking in comparison to other institutions." In profiling an institution, Pure (Elsevier, www.elsevier.com/solutions/pure) "aggregates your organization's research information ... enables your organization to build reports, carry out performance assessments, manage researcher profiles, enable research networking and expertise," while InCites (Thomson Reuters, http://clarivate.libguides.com/incites) allows you to "analyze institutional productivity, monitor collaboration activity, identify influential researchers, showcase strengths, and discover areas of opportunity." Universities pay hundreds of thousands of dollars for these services, typically in the form of multi-year contracts, yet these services lack some critical pieces of information that we do have access to (e.g., past research proposals) and do not match research experts to calls for proposals.

Our work builds on this previous knowledge, and proposes an integrated search and visualization platform for the important task of matching research experts with calls for proposals, a challenging task that is currently handled manually by university staff.

III. REMatch System Overview

The REMatch system relies on a wide range of data collected from the following internal and external sources:

1) UAVitae: an internal online system for university faculty and staff, which includes research publications;
2) UAIR: university internal analytics and institutional research, which keeps track of staff arrivals and departures;
3) UAR: UA's research activity system, which keeps track of research proposals and awards information internally;
4) Google Scholar Profiles: we scraped UA Google Scholar profiles for publications data (e.g., title, abstract, citations, research areas);
5) Google Scholar Topics: we scraped half a million research profiles from the top one thousand universities (according to the Center for World University Rankings [2]), in order to build a topic network connecting co-occurring research topics.

These data sources are integrated so as to associate a rich individual research profile with each member of the institution.

The system interaction entails the entry of a URL address for a call for proposals (CFP) or plain text, typically the content-rich passages from a CFP. The output is a ranked list of research experts who match the CFP, i.e., a list of people who RO staff might reasonably contact and encourage to respond to the CFP. Clicking beside each name reveals the relevant matching research topics and terms. A bar chart shows the number of matches in each college and clicking on it filters the results. Another bar chart shows the number of matches by type of position (e.g., Assistant Professor) and clicking on it further filters the results. The set of the keywords extracted from the CFP is shown - clicking on any of them makes them stop-words for this query (removing them from the set) and the list of experts is updated with this reduced set. The collaboration map highlights the research experts in the current matched list - this map can be navigated and explored to find more information about the highlighted experts, including their collaborators, the nature of these collaborations (titles of joint papers or research proposals), level of current funding, and number of citations.

The steps in this process are therefore (1) building the university-wide collaboration map, (2) building a database of research expertise, (3) extracting research topics and terms from the CFP, (4) identifying the relevant faculty members, (5) displaying them on the map, (6) filtering with respect to college, (7) removing irrelevant keywords. We discuss each of these steps below.

IV. Research Collaboration Map

We chose to use a map-based visualization to show the results in the context of cross-university collaboration. Maps have been shown to be effective and memorable representations of networks [19], [38], [39]. In our case, the underlying data is a collaboration network, in which each node is a person, and an edge exists between two nodes if the corresponding pair have past collaborations (e.g., joint research paper, joint research proposal). Collaborations are based on publication data in the databases scraped (described in the previous section). Named-entity recognition is one of the challenging issues when working with multiple datasets. In all internal university systems, researchers are identified by an unique ID. Each Google Scholar profile also assigns unique IDs. We match IDs from different databases using a semi-automated system based on OpenRefine [5].

We use the GMap framework [19] to generate a map-like visualization of the collaboration network and extend it to support semantic zooming. The process can be summarized as follows: (1) embed the collaboration network in the plane, (2) group nodes into clusters, (3) create the geographic map representation, (4) compute multiple level-of-detail views, and (5) provide support for interactions (pan, zoom, search) and overlays.

We embed the network using a scalable force-directed algorithm (`sfdp` from graphviz) and then group the nodes using *k*-means clustering. To create the geographic-map look, we use a modified Voronoi diagram based on the obtained embedding and clustering. The geographic regions are colored such that no two adjacent countries have colors that are too similar, using the spectral vertex labeling method [19]. We use the GraphViz implementation of node-overlap removal provided by `prism`. Note that `prism` provides non-overlapping labels only for the complete basemap (showing all nodes), and not for the other 6 level-of-detail views, needed for semantic zooming.

The semantic zoom in REMatch requires modifications to nodes, edges, clusters, and heatmap overlays; see Fig. 3. We use the Google Maps API which handles most of these issues, with the notable exception of node-overlap (and hence node-label overlap), which is a natural side effect of zooming. To ensure that neither nodes nor labels overlap on any level of detail, we compute different *node visibilities* for different zoom-levels. For each level, we sort the nodes by their weight (node-weight is determined by the degree of the node, or amount of funding, or number of citations). We make the i-th node visible on the j-th level if the bounding box of the i-th node does not overlap with the bounding boxes of nodes $1, 2, \cdots, (i-1)$. This algorithm requires $O(n^2)$ time but it could be improved using different techniques as in [14].

V. Matching Research Experts

In this section we describe how we find the research experts who best match a given text query, (e.g., a CFP). We treat this as an information retrieval (IR) problem, where the text query forms the (large) input query, and each individual researcher

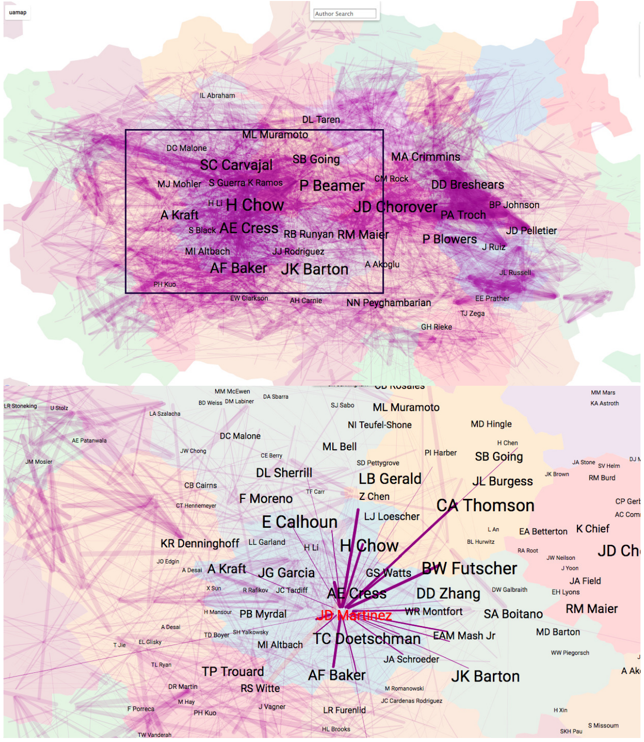


Figure 3: Top: a high level view of the collaboration map; Bottom: zooming in one level deeper provides more details; a mouse-over event highlights a node and its strongest connections.

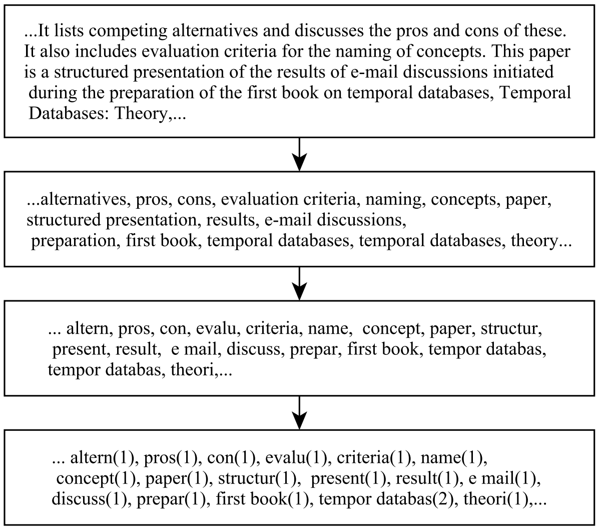
is represented as a single "document" containing all associated research materials (e.g., research topics, publications, submitted proposals). We further extend this process with a ML component, as described below.

A. Building a Database of Research Expertise

We first assign research topics to all researcher experts. In general, Google Scholar (GS) profiles provide self-reported description of a researcher's interests or research topics. However, many researchers do not have a GS profile and many have a profile that is very sparse (containing only one or two research topics), which complicates the IR task addressed in this section. To mitigate this problem, we implement an ML component that predicts five research topics for such researchers.

Out of the current total 2,187 researchers at UA, 1,127 have GS profiles. Of these, 314 provide the maximum possible five research topics and no prediction is necessary. For all others we use the method described below. When a researcher has no GS profile we use the top 5 predicted topics. When a researcher has a GS profile with fewer than 5 topics, we augment the given topics with our prediction to again obtain exactly 5 topics. The prediction process relies on a multi-class, multi-label classification with regularized logistic regression which is implemented in the LiblineaR package.

Task: Determine Top-5 research topics from an expert's data.



...It lists competing alternatives and discusses the pros and cons of these. It also includes evaluation criteria for the naming of concepts. This paper is a structured presentation of the results of e-mail discussions initiated during the preparation of the first book on temporal databases, Temporal Databases: Theory,...

...alternatives, pros, cons, evaluation criteria, naming, concepts, paper, structured presentation, results, e-mail discussions, preparation, first book, temporal databases, temporal databases, theory...

... altern, pros, con, evalu, criteria, name, concept, paper, structur, present, result, e mail, discuss, prepar, first book, tempor databas, tempor databas, theori,...

... altern(1), pros(1), con(1), evalu(1), criteria(1), name(1), concept(1), paper(1), structur(1), present(1), result(1), e mail(1), discuss(1), prepar(1), first book(1), tempor databas(2), theori(1),...

Figure 4: Preprocessing text for the machine learning step.

Data: We scraped GS profiles, gathering the self-reported areas of expertise and publication details. Note that multiple research topics are typically mentioned in each profile. We consider all publications in a given profile as belonging to each of the research areas provided in the profile. We use this data as training data. To predict the Top-5 research topics associated with each researcher, we use all associated data (research publications, grant proposals, etc.).

Data preprocessing: To prepare the dataset for the machine learning model we first extract important phrases (i.e., noun phrases) from the documents. We rely on the Natural Language Toolkit (NLTK) [4] for this task, using the grammar $NP : \{(< JJR > | < JJS > | < JJ > | < NN >| < NNS > | < NNP > | < NNPS >)*\}$. We then stem the phrases with using a series of operations from Lucene [1]: StandardFilter, LowerCaseFilter, StopFilter, and SnowballFilter. Finally we compute term frequency as shown in Fig 4.

Training: Formally, we consider the self-reported research topics in GS profiles for the 1,127 academics as labels (Y). For all researchers, we build a matrix X from their publications: each row corresponds to an individual publication; each column represents a noun phrase found in the corresponding text, with a value set to its term frequency (TF) in the corresponding document. The label(s) assigned to rows in X are taken from all research topics reported by the corresponding researchers in GS. The matrix X used for training contains 38,886 rows and 8,462 columns. This matrix is used to train the multi-class logistic regression classifier.

Prediction: At prediction time, the model returns classification probabilities as a $n \times k$ matrix, where k is the number of research topics, and n is the number of researchers to be classified. We take only the top five columns for each person, i.e., the top predicted research topics of a person. Where needed (e.g., for researchers with fewer than 5 topics in their GS profile) we augment the list to obtain the Top-5, making sure there are no repeats.

B. Extracting Google Research Topics

In a topic network, topics are the nodes and edges indicate that topics are related to each other. Extracting topics from research articles (with topic co-occurrence within an article indicating topic relationship) is a popular approach to creating a knowledge network [17], [54]), but these methods do not allow for easy identification of general topics (e.g., mathematics, physics), nor do they include very specific topics (e.g., symmetry detection algorithms, interactive graph visualization) as nodes.

Our approach for creating a topic network relies on the assumption that people know the topics that they work on: nobody is better placed to categorize researchers' topic areas than the researchers themselves, and, while document analysis might automatically identify and extract topic labels from an article, only the researchers who wrote the article know precisely the key topics of the paper. We therefore use the self-reported areas of study in the Google Scholar (GS) database, as defined by researchers themselves and the co-occurrence of topics within a researcher's list indicates a relationship between them in our knowledge network. Specifically, the topics network is generated using the following steps:

1) Data Scraping: Our data gathering is limited to GS entries associated with the world's top 1,000 universities (according to the Center for World University Rankings [2]). We extracted the institution IDs from GS and then scraped the URL associated with each institution to collect research profiles of all individuals associated with the institution, focusing on the list of research topics from each research profile. The total number of topics extracted from this raw data was 190,137, but after standardizing the topic separators within the topic list, and using beautifulsoup [2] to tidy up html tags for consistency, the number of distinct topics rose to 222,459.
2) Data Cleaning: We removed leading or trailing spaces, inconsistent use of upper and lower case letters, unnecessary punctuation and control characters, and duplicate topics. Many topics were phrases or composite terms (e.g., "statistics for neuroscience," "data and model management," "group theory and combinatorics," "symmetries of graphs"); we removed conjunctions (and, or) and other words with no semantic weight (for, of), thus splitting topic phrases into their constituents.
3) Topic Correction: Typos and acronyms frequently occur. We used Google's OpenRefine [5] to identify and resolve typing errors, and to find alternate representations of the same topic [9], [16], [23] (e.g "Computer Human-Interaction" is equivalent to "Human-Computer Interaction"; "Primary education" is the same as "Elementary education"). This process reduced the number of unique topics to 210,588.
4) Topic Removal: We dropped topics that were associated with four or fewer people (aware that these topics might be topic labels in which there were typing errors that were not captured by OpenRefine), and topics that we identified as not being in English. This reduced the number of topics to 39,067.

Table I: Data summary.

Field	Data Description	#Records
F1	Research paper titles, abstracts	165501
	Grant proposals	44970
	Abstracts of grant proposals	1355
F2	Top 5 research predicted areas	2544
F3	Courses taught	240823
	Internal and external service	18256
	Teaching and research interests	1706
	Consulting	1690
	Biography	1222
	Research activity / Work in progress	2032
	Research goals and progress	6844

5) Topic Merging: Merging was required for topics that are similar, but are listed slightly differently; for example, "algorithm," "algorithms," "algorithmics" are all the same topic, as are "organization," "organizational" and "organizing." We used snowball [36] to find the root word by applying stemming processes (removing endings such as -s, -ed, -ing). "Algorithm," "algorithms", and "algorithmics" thus all become "algorithm;" however "applied" and "applications" become the meaningless term "appli." To mitigate against nonsensical resolution, we choose the main topic to be the one with the highest frequency amongst all topics with the same stem.

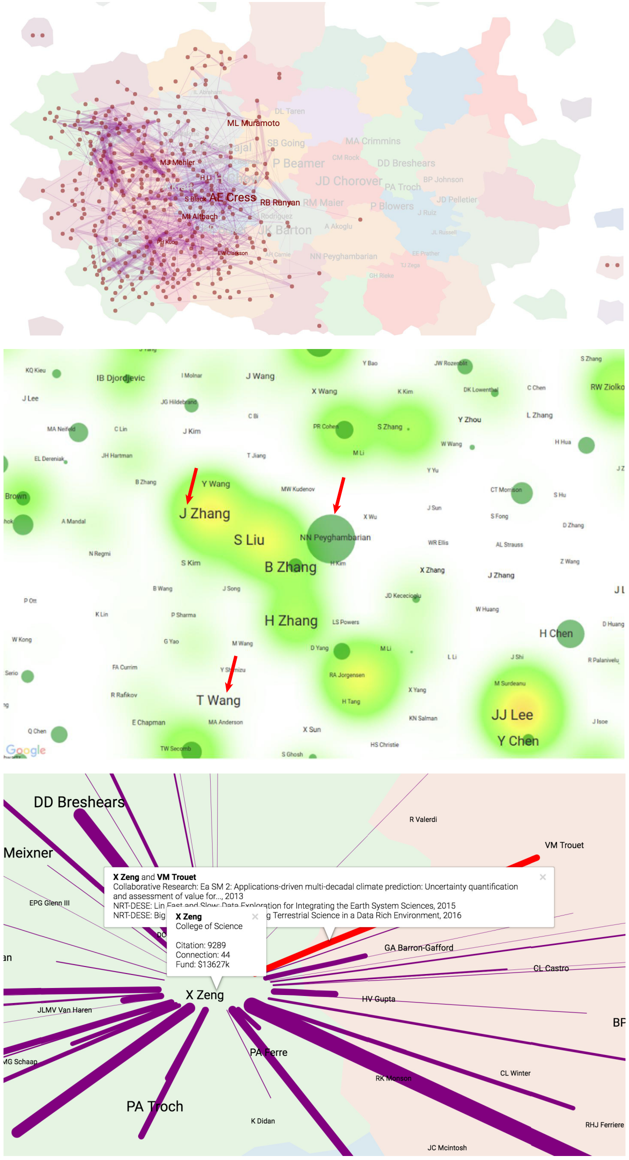
C. Lucene Indexing

As mentioned earlier, we treat the problem of matching academics with a given research topic as an IR task, which we implemented using Apache Lucene [1]. In particular, we construct a Lucene document for each individual researcher, with two data fields. The first field ($F1$) contains titles and abstracts of research papers and grant proposals. The second field ($F2$) contains the research topics of the corresponding person (from their GS profile, or predicted as described above). The third field ($F3$) contains information from biographies/CVs, list of courses taught, internal and external service, consulting, etc.; see Table I.

This approach allows us to combine three sources of information in a single query: $F1$ contains data that is (more or less) objective since it comes from academic publications, but it may be too verbose, $F2$ contains precise information, but which may be biased since it is self-reported, and $F3$ contains high level research areas that are not covered by $F1$ and $F2$.

D. Query Pre-Processing

Given an input query (e.g., a URL for a CFP, plain text, or a collection of keywords), we convert it to an actual Lucene query as follows: The text is cleaned, tokenized, and stemmed. We next compute the term frequency of unigrams, bigrams, and trigrams, which become our *candidate research topics*. Then we match the candidate research topics with the research topics extracted from GS, by taking the intersection of the two sets of topics (in our network we only include topics that



X Zeng and **VM Trouet**
Collaborative Research: Ea SM 2: Applications-driven multi-decadal climate prediction: Uncertainty quantification and assessment of value for..., 2013
NRT-DESE: Lin Fast and Slow: Data Exploration for Integrating the Earth System Sciences, 2015
NRT-DESE: Big g Terrestrial Science in a Data Rich Environment, 2016

X Zeng
College of Science

Citation: 9289
Connection: 44
Fund: $13627k

Figure 5: Top: collaborations within the College of Medicine; Middle: collaboration (indicated by font size), citations (heatmap), and funding (green circles); Bottom: information about an individual researcher.

occur more than once in the 500,000 analyzed GS profiles). The result is a collection of known research topics associated with the input query. This set of research topics is denoted by $Q1$, while the list of unigrams in the input query is denoted by $Q2$.

E. Lucene Queries and Score Merge

Using this information, we construct three separate queries, and compute a combined score through linear interpolation (with hyper parameters λ_i), as follows:

$Score_1$: Normalized score for phrase query on index field $F2$, using query $Q1$;

$Score_2$: Normalized score for query on index field $F2$, with query $Q2$;

$Score_3$: Normalized score for query on index field $F1$, with query $Q2$.

$Score_4$: Normalized score for phrase query on index field $F3$, using query $Q1$;

The overall score S for a given research expert p is then computed as: $S(p) = \sum_1^4 \{\lambda_i \times Score_i(p)\}$ where $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$. Since $Score1$ captures a match between the research areas of the input query and the expert's research areas λ_1 is larger than λ_2 and λ_3. In our implementation we use $\lambda = [0.3, 0.1, 0.1, 0.5]$, that is, $F1$ and $F2$ contribute a combined 50% and $F3$ the other 50% of the total score.

VI. Map Overlays

Currently the REMatch system provides the following overlays:

Citation Heatmap: highlights highly-cited experts, based on citation counts or normalized citation counts

Funds Overlay: shows funding levels (overall or from specific funding agencies) using a green circle with radius proportional to current funds associated with each research expert

Edges: shows all connections between research experts, as by default these are not shown, relying on Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things"

College Overlay: highlights all research expert (nodes) from a given college and/or all collaborations between pairs of research experts from the same college

Individual Details: We provide basic search functionality which locates people on the map. Clicking on a node shows the name of the person, along with the number of connections, citations count, and college affiliation; edges to collaborators are also shown; see Fig. 5. By clicking on an edge we can see information about the collaborative work.

Label Size: Depending on what we want to show, we can change node importance (which is reflected in the font sizes of the labels) by considering the degree of the node, the research funding amount, or number of citations associated with this node; see Fig. 6.

Several of these overlays can be shown concurrently, which helps visualize different data dimensions, e.g., identifying research experts who are producing highly cited work, research experts who are well-connected in the university, and research experts who are good at securing funding; see Fig. 5.

VII. Implementation

Given a CFP, a research paper, or even a bit of relevant text from a CFP or paper, our system matches the best experts (as described in Section V) and shows the results on the map. We can filter by college and explore the explanations for the matches. Moreover, the system shows the important single terms identified from the query, any of which can be marked as stop-words. Clicking on a word removes it from the query

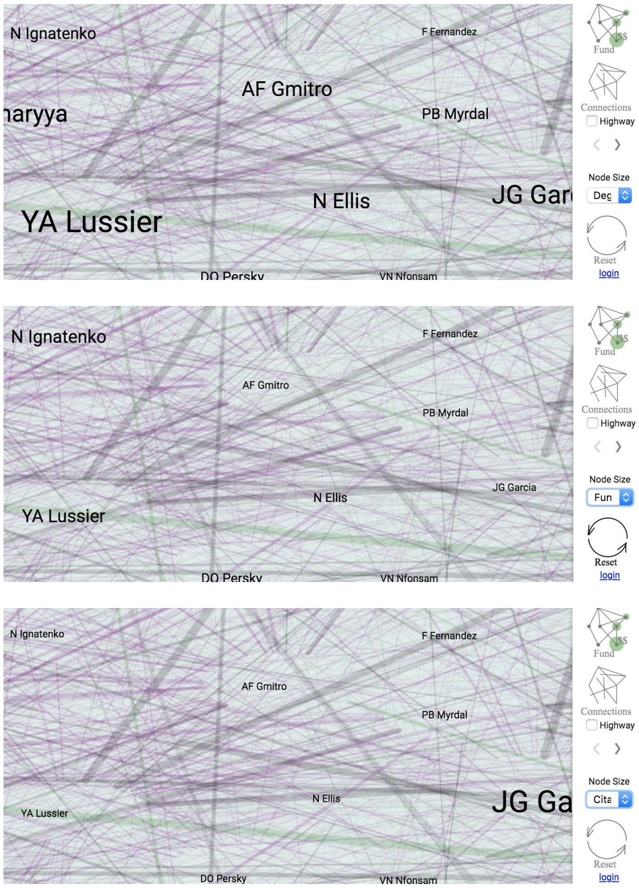


Figure 6: Node and label sizes are determined by one of three variables: number of collaborations (node degree), funding amounts, and citations. This is illustrated for the same portion on the map above. Note that some nodes disappear (when they drop below the threshold size for the current level of detail) and that their font sizes change.

and can improve the results. Figure 1 shows a screenshot of the system.

We use a variety of tools to clean, store, and process our data: mongodb scripts, sqllite, python, R, Java-Lucene, openrefine. Google maps API and jquery are used for map drawing and to handle interactions in the web application. We run python-django for the webserver and mongodb for database storage and queries.

GMap produces a "basemap" from the given graph, which is a *static image* that is not well-suited for user interaction, such as zooming, panning, and searching. We enable interactions with the help of the google maps API [45]. Specifically, we take the output from GMap and convert it into google map objects, i.e., *google.maps.SymbolPath*, *google.maps.Polygon*, *google.maps.Polyline*, etc. For the web interface we provide 7 levels of details, showing different subgraphs, depending on the zoom level.

VIII. Evaluation

We performed a case-study evaluation with two Research Office (RO) staff members, both of whom perform the task of matching funding calls to faculty on a regular basis – one in the Medical and Biological Sciences, and one in Environmental Science. They each brought two relevant funding calls to the meeting, together with a list of people they would target for these calls, having followed their normal processes in devising these lists. We gave a live demonstration of the system – for each call, we entered relevant content-rich paragraphs into the system, showed the list of names, filtered by colleges as appropriate, and discussed the results with them. There were three responses to each of the four lists:

1) Why is X there? This happened in the cases where the RO staff members know a specific person well, and are surprised to see them matched for the particular call.
2) Why is X not there? This happened for people included on the RO staff member's list who are ranked low on the REMatch list or do not appear at all.
3) I don't know X. These are people that the RO staff member does not know at all, and would be interested in finding out why they were included in the list.

(1) and (2) were responses that are a result of personal knowledge of the RO staff – they rely on the information about faculty that they keep in their heads – information that can be lost with staff turnover and which is hard to collect given thousands of university researchers. (3) is a positive response, since this encourages the RO staff to find out more about university researchers who they do not already know. It also is useful for new RO staff members.

In discussion, we discovered that anomalies were closely related to context. For example, for one of the calls, an extensive multi-disciplinary team is needed, spanning several colleges – requiring separate filtering for each relevant college. In another, people were missed because they have joint cross-college appointments (and did not show up in the first college selected). For one proposal, a senior team of established researchers was needed, and so names of junior university researchers are not good matches. We discovered that several people were included on the REMatch list because of very generic terms (e.g., "environment") and the facility to "turn off" such all-encompassing stop-words was welcomed.

Overall, the prototype REMatch system received a very positive response from the RO staff, despite the few anomalies in these four specific calls. For those names that they had not thought of (category (3) above), they were intrigued ("[...] is an interesting fit. Someone like that would be good."; "none of these are totally wacky") and recognized the usefulness of being directed to people they had not previously considered. One approach that they use that is currently not supported is the ability to identify people who have previously failed in a funding bid. This information would enable the RO staff to focus on people who might need support in further applications, or to identify other appropriate opportunities that a failed bid might be recycled for. However, they acknowledge that this is sensitive information. From a process perspective, the two RO staff members were impressed with the fact that the text entered can be customized – other research management systems (for example, Academic Analytics) require the entry of specified keywords, making them "not as intuitive and easy to use as this system." In effect, our system serves as a pre-processing step by extracting the relevant keywords.

The RO staff members were confident that the REMatch would be a very useful complement to their current processes, and would save them a great deal of time. The ability to "easily refine and hone down" results means that they have control over the search, while taking advantage of the explanations, collaboration map and faculty profile information that is easily accessible.

IX. Discussion and Limitations

Anecdotally, our REMatch system seems to be fulfilling its purpose well. For example, it placed all 8 of the actual principal investigators and senior personnel of a recent proposal sent to the National Science Foundation TRIPODS program among the top 15 matches (as shown in the accompanying video, available at https://uamap-dev.arl.arizona.edu). Note that this particular CFP is recent and was not included in the training data for REMatch. Similarly, out of four recently submitted proposals (that were also not part of the training data), REMatch identified three of the Principal Investigators who submitted proposals to these programs. The University's Research, Discovery & Innovation Office is using the REMatch system to put together research teams for large-scale, multi-disciplinary projects, even though it is still under development. We plan to add new features based on suggestions we have already received.

We rely on several internal and external sources for gathering the data that makes REMatch work. For example, relying on Google Scholar has some advantages (e.g., a large amount of information) but also many disadvantages (e.g., the data is not curated). Further, different research areas differ in the extent of their representation in Google Scholar. For example, there seem to be many more computer science and physics profiles than history and psychology ones. Data from both internal and external sources needs to be updated regularly in order to provide current and relevant results.

Before arriving at our four-level scoring function (described in Section V-E) we tried several simpler approaches, all of which resulted in poor matches. We considered treating both research experts and CFPs as documents represented by high dimensional vectors of unigrams weighted by frequency or TF/IDF (term frequency/inverse document frequency) and finding the best matches via cosine similarity between the angles of a given CFP vector and the research expert vectors. This resulted in bad matches due to the high importance of adjectives and articles. Limiting to noun-phrases did not help as common CFP terms such as grants, proposers, projects carried little signal. Treating such phrases as stop-words and even using a different English language corpus as a baseline did not yield the desired results. Several other attempts failed before we arrived at the four-level scoring function we currently use. Still, we hope to further improve our scoring function.

We are aware that our focus on the underlying computational processes and the map visualization means that our interface could do with some improvements with respect to human-computer interaction design, in particular when considering the nature of the users and their specific visual analytics tasks.

X. Conclusions

The REMatch system helps identify areas of expertise, as well as experts in a given field, which is useful in the context of putting together strong multi-disciplinary research teams. We consider the main challenge and our major contributions to be in gathering and processing the needed data, putting together a collection of new and existing tools, designing an intuitive interactive interface, and packaging all of this in a functional system that non-experts can use. In addition to the case studies above, we plan to formally validate several of its components, starting with the quality of the predicted research topics and the quality of the proposal-person matches.

Despite non-trivial limitations, REMatch is novel as it provides a functional system that implements in-the-browser, map-based interactive navigation of a large underlying network, supporting panning, zooming, and searching, as well as map overlays. The REMatch system is open source and a video showing the system in action is available at https://uamap-dev.arl.arizona.edu.

References

[1] Apache lucene - welcome to apache lucene. https://lucene.apache.org. Accessed: 06-18-2018.

[2] Cwur | center for world university rankings. http://cwur.org/. Accessed: 09-13-2016.

[3] Graphviz | Graphviz - Graph Visualization Software. http://www.graphviz.org/. Accessed: 05-25-2017.

[4] Natural language toolkit – nltk 3.0 documentation. http://www.nltk.org/. Accessed: 05-14-2016.

[5] Openrefine. http://openrefine.org/. Accessed: 05-04-2017.

[6] J. Abello, F. Van Ham, and N. Krishnan. Ask-GraphView: A large scale graph visualization system. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):669–676, 2006.

[7] D. Auber, D. Archambault, R. Bourqui, A. Lambert, M. Mathiaut, P. Mary, M. Delest, J. Dubois, and G. Mélançon. The Tulip 3 framework: A scalable software library for information visualization applications based on relational data. Technical Report RR-7860, INRIA, 2012.

[8] M. Bastian, S. Heymann, and M. Jacomy. Gephi: an open source software for exploring and manipulating networks. *ICWSM*, 8:361–362, 2009.

[9] W. B. Cavnar, J. M. Trenkle, et al. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175, 1994.

[10] M. Chimani, C. Gutwenger, M. Jünger, G. W. Klau, K. Klein, and P. Mutzel. The open graph drawing framework (OGDF). *Handbook of Graph Drawing and Visualization*, pages 543–569, 2011.

[11] M. De Domenico, E. Omodei, and A. Arenas. Quantifying the diaspora of knowledge in the last century. *Applied Network Science*, 1(1):15, 2016.

[12] W. De Nooy, A. Mrvar, and V. Batagelj. *Exploratory social network analysis with Pajek*, volume 27. Cambridge University Press, 2011.

[13] C. Dunne, B. Shneiderman, R. Gove, J. Klavans, and B. Dorr. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, 63(12):2351–2369.

[14] T. Dwyer, K. Marriott, and P. J. Stuckey. Fast node overlap removal. In *International Symposium on Graph Drawing*, pages 153–164. Springer, 2005.

[15] S. Effendy and R. H. Yap. Analysing trends in computer science research: A preliminary study using the microsoft academic graph. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1245–1250, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[16] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on knowledge and data engineering*, 19(1), 2007.

[17] D. Fried and S. G. Kobourov. Maps of computer science. *2014 IEEE Pacific Visualization Symposium (PacificVis)*, 00:113–120, 2014.

[18] E. Gansner, Y. Koren, and S. North. Topological fisheye views for visualizing large graphs. *TVCG*, 11(4):457–468, July 2005.

[19] E. R. Gansner, Y. Hu, and S. Kobourov. Gmap: Visualizing graphs and clusters as maps. In *2010 IEEE Pacific Visualization Symposium (PacificVis)*, pages 201–208, March 2010.

[20] J. Heer, S. K. Card, and J. A. Landay. Prefuse: a toolkit for interactive information visualization. In *Proc. SIGCHI conference on Human factors in computing systems*, pages 421–430. ACM, 2005.

[21] F. Heimerl, Q. Han, S. Koch, and T. Ertl. Citerivers: Visual analytics of citation patterns. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):190–199, Jan 2016.

[22] F. Heimerl, M. John, Q. Han, S. Koch, and T. Ertl. Docucompass: Effective exploration of document landscapes. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 11–20, Oct 2016.

[23] G. R. Hjaltason and H. Samet. Index-driven similarity search in metric spaces (survey article). *ACM Transactions on Database Systems (TODS)*, 28(4):517–580, 2003.

[24] J. Huang, Z. Zhuang, J. Li, and C. L. Giles. Collaboration over time: Characterizing and modeling network evolution. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 107–116, New York, NY, USA, 2008. ACM.

[25] S. E. Hug, M. Ochsner, and M. P. Brändle. Citation analysis with microsoft academic. *CoRR*, abs/1609.05354, 2016.

[26] W. Ke, K. Borner, and L. Viswanath. Major information visualization authors, papers and topics in the acm library. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages r1–r1. IEEE, 2004.

[27] P. Liu, J. Curson, and P. Dew. Use of rdf for expertise matching within academia. *Knowl. Inf. Syst.*, 8(1):103–130, July 2005.

[28] B. Loepp, K. Herrmanny, and J. Ziegler. Blended recommending: Integrating interactive information filtering and algorithmic recommender techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 975–984, New York, NY, USA, 2015. ACM.

[29] D. W. McDonald and M. S. Ackerman. Expertise recommender: A flexible recommendation system and architecture. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, pages 231–240, New York, NY, USA, 2000. ACM.

[30] K. Misue. Visual exploration of a series of academic conferences. In *2014 IEEE International Conference on Data Mining Workshop*, pages 314–320, Dec 2014.

[31] B. Mutlu, E. Veas, C. Trattner, and V. Sabol. Vizrec: A two-stage recommender system for personalized visualizations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, IUI Companion '15, pages 49–52, New York, NY, USA, 2015. ACM.

[32] L. Nachmanson, G. Robertson, and B. Lee. Drawing graphs with GLEE. In *Graph Drawing*, pages 389–394. Springer, 2008.

[33] M. E. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132, 2001.

[34] A. Perianes-Rodriguez and J. Ruiz-Castillo. University citation distributions. *Journal of the Association for Information Science and Technology*, 2015.

[35] J. Portenoy and J. D. West. Visualizing scholarly publications and citations to enhance author profiles. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1279–1282, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[36] M. F. Porter. Snowball: A language for stemming algorithms, 2001.

[37] W. Rivadeneira and B. B. Bederson. A study of search result clustering interfaces: Comparing textual and zoomable user interfaces. *Studies*, 21:5, 2003.

[38] B. Saket, C. Scheidegger, S. G. Kobourov, and K. Borner. Map-based visualizations increase recall accuracy of data. *Computer Graphics Forum*, 34(3):441–450, 2015.

[39] B. Saket, P. Simonetto, S. Kobourov, and K. Börner. Node, node-link, and node-link-group diagrams: An evaluation. *IEEE Transactions on Visualization & Computer Graphics*, 20(12):2231–2240, 2014.

[40] W. J. Schroeder, L. S. Avila, and W. Hoffman. Visualizing with VTK: a tutorial. *Computer Graphics and Applications*, 20(5):20–27, 2000.

[41] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[42] A. Skupin. A cartographic approach to visualizing conference abstracts. *IEEE Computer Graphics and Applications*, 22(1):50–58, 2002.

[43] J. Stasko, J. Choo, Y. Han, M. Hu, H. Pileggi, R. Sadana, and C. D. Stolper. Citevis: Exploring conference paper citation data visually. *Posters of IEEE InfoVis*, 2, 2013.

[44] X. Sun, K. Ding, and Y. Lin. Mapping the evolution of scientific fields based on cross-field authors. *Journal of Informetrics*, 10(3):750 – 761, 2016.

[45] G. Svennerberg. *Beginning Google Maps API 3*. Apress, 2010.

[46] D. R. Swanson. Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118, 1986.

[47] P. Van den Besselaar and G. Heimeriks. Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, 68(3):377–393, 2006.

[48] C. Vehlow, F. Beck, and D. Weiskopf. Visualizing group structures in graphs: A survey. *Comput. Graph. Forum*, 36(6):201–225, Sept. 2017.

[49] T. Von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. W. Fellner. Visual analysis of large graphs: state-of-the-art and future research challenges. In *Computer graphics forum*, volume 30, pages 1719–1749. Wiley Online Library, 2011.

[50] W. Wang, G. Zhang, and J. Lu. Hierarchy visualization for group recommender systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, PP(99):1–12, 2017.

[51] J. D. West, I. Wesley-Smith, and C. T. Bergstrom. A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE Transactions on Big Data*, 2(2):113–123, June 2016.

[52] R. Wiese, M. Eiglsperger, and M. Kaufmann. yFiles: Visualization and automatic layout of graphs. In *GD*, pages 453–454, 2001.

[53] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Proceedings of Visualization 1995 Conference*, pages 51–58, Oct 1995.

[54] Y. Yang, Q. Yao, and H. Qu. Vistopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. *Visual Informatics*, pages –, 2017.

[55] D. Zhao and A. Strotmann. Analysis and visualization of citation networks. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 7(1):1–207, 2015.

[56] N. Zhou, J. Saltz, and K. Mueller. *Maps of Human Disease: A Web-Based Framework for the Visualization of Human Disease Comorbidity and Clinical Profile Overlay*, pages 47–60. Springer International Publishing, Cham, 2016.