GRAM: Global Research Activity Map

Randy Burd*
randy.burd@liu.edu

Kimberly Andrews Espy†
kespy@email.arizona.edu

Md Iqbal Hossain†
hossain@email.arizona.edu

Stephen Kobourov†
kobourov@cs.arizona.edu

Nirav Merchant†
nirav@email.arizona.edu

Helen Purchase‡
helen.purchase@glasgow.ac.uk

ABSTRACT

The Global Research Activity Map (GRAM) is an interactive web-based system for visualizing and analyzing worldwide scholarship activity as represented by research topics. The underlying data for GRAM is obtained from Google Scholar academic research profiles and is used to create a weighted topic graph. Nodes correspond to self-reported research topics and edges indicate co-occurring topics in the profiles. The GRAM system supports map-based interactive features, including semantic zooming, panning, and searching. Map overlays can be used to compare human resource investment, displayed as the relative number of active researchers in particular topic areas, as well scholarly output in terms of citations and normalized citation counts. Evaluation of the GRAM system, with the help of university research management stakeholders, reveals interesting patterns in research investment and output for universities across the world (USA, Europe, Asia) and for different types of universities. While some of these patterns are expected, others are surprising. Overall, GRAM can be a useful tool to visualize human resource investment and research productivity in comparison to peers at a local, regional and global scale. Such information is needed by university administrators to identify institutional strengths and weaknesses and to make strategic data-driven decisions.

CCS CONCEPTS

• **Information Visualization**; • **Visual Analytics**; • **Web Interfaces**;

KEYWORDS

Interactive visualization system, knowledge discovery, topics map

ACM Reference Format:
Randy Burd, Kimberly Andrews Espy, Md Iqbal Hossain, Stephen Kobourov, Nirav Merchant, and Helen Purchase. 2018. GRAM: Global Research Activity Map. In *AVI '18: 2018 International Conference on Advanced Visual Interfaces, AVI '18, May 29-June 1, 2018, Castiglione della Pescaia, Italy,* Massimo Mecella and Kent Norman (Eds.). ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3206505.3206531

*Long Island University, Brookville, NY
†University of Arizona, Tucson, Arizona
‡University of Glasgow, Scotland, UK

AVI '18, May 29-June 1, 2018, Castiglione della Pescaia, Italy

ACM ISBN 978-1-4503-5616-9/18/05…$15.00
https://doi.org/10.1145/3206505.3206531

1 INTRODUCTION

University research activity is diverse and distributed, and it is difficult for university managers to get a comprehensive overview of the research strengths and weaknesses in their own institution or to efficiently compare the overall research profile of a university with those of its peers and competitors. This is important for university strategy, since managers need to address the following questions:

- Q1 Where are the strengths and weaknesses in our institution? In which particular research areas do we employ individuals (or groups of individuals) who are recognized as making a valuable contribution to knowledge?
- Q2 How do these strengths and weaknesses compare with those of our competitor universities? In which areas are we comparatively strong/weak? Where should we invest human capital to improve our standing amongst our peers?

Citation information is, of course, important for addressing both questions. Current tools (e.g., Google Scholar) show published papers and citation counts for individual researchers, and tell us the most cited researcher in particular areas. They also tell us what topics individual researchers work on, and who their collaborators are. However, they do not provide an overall citation profile for an entire institution that is based on research topic areas, and do not permit comparison of the human resource investment between institutions.

Our prototype GRAM system is already used by senior managers at our university to visualize, explore, compare and contrast our human resource investment and citation output with that of other universities and benchmark-sets of universities, and to make decisions on upcoming faculty appointments. Key to the GRAM system is the classification and organization of human knowledge into topics, since it is these topics that the university managers relate to when making their humans resource investment decisions (for example, by advertising for faculty in particular research areas).

There have been many attempts at classifying and organizing topics of human knowledge, mostly using a top-down and hierarchical approach, by dividing known fields of study into sub-categories [1, 20]. For example, we know that "computer science" includes the sub-topics of "operating systems" and "algorithms." Taking this approach means that we assign known labels to fields of study, and make hierarchical connections between them based on what we know about them, as in the ACM classification; see Fig. 1(a). However, a more realistic view of knowledge is a non-hierarchical one which allows us to see, say in the form of a graph, different types of connections between topics; see Fig. 1(b). Such knowledge graphs can also be created in a top-down manner we know about them) but both of these approaches can be criticized as being biased by the views and extent of understanding of the knowledge graph creator.

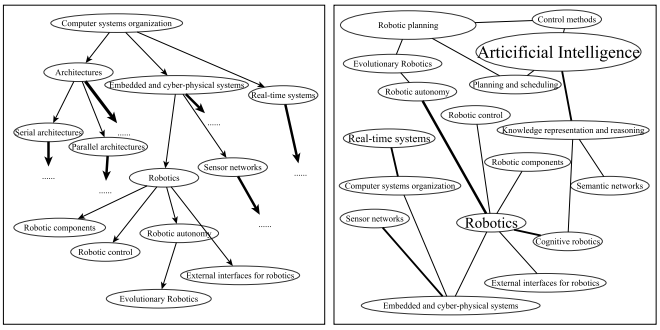


Figure 1: (a) Part of the ACM hierarchical topic classification; (b) Part of a non-hierarchical topic network.

A more justifiable approach to representing the categories of human knowledge and the relationships between them is to collect and organize "bottom-up" data, that is, information about what researchers actually study, and which topics are naturally linked together through their common research activity. For example, in creating a network of research articles, the list of references cited in one article can be represented as being linked to each other, since they are assumed have a common theme, even if only a loose one. Providing an overview of human knowledge and areas of active research is important not only in documenting and exploring the trajectory of global research endeavors - an important contribution to the history of knowledge - but also in benchmarking and comparing individual researchers and their institutions.

GRAM takes a novel approach to the representation of human research endeavor by (a) using bottom-up data provided by Google Scholar; (b) depicting the knowledge network with a map-like visualization; (c) supporting real-time in-the-browser semantic zooming; and (d) providing overlays that depict the relative number of researchers working on topics and the relative number of citations for topics, with respect to each institution (or set of institutions). GRAM makes it possible to see a high-level view research activity at universities worldwide and interact with the data in an intuitive and familiar way. Unlike expensive commercial tools, GRAM is open-source and free and has the potential to be useful at many universities interested in a high-level overview and comparisons with specific other institutions or with national and international aggregate data.

2 RELATED WORK

There is related work in different domains: from science classification and topic analysis, to visualizations for text and large graphs.

Knowledge classification: The most comprehensive bottom-up classification of science topics [15] uses data from ten years of Thomson Reuters' Web of Science [8] and eight years of Elsevier's Scopus [7] to group over 25,000 journals into 554 subdisciplines, each of which is associated with exactly one of 13 disciplines (e.g., Mathematics, Physics, Social Science). While the extensive graph can be presented with different views, it is not interactive, and its presentation as an "overview" makes it difficult to elicit details. The Microsoft Academic Graph is regularly updated, and is built from indexing research papers, each of which is classified into over 50,000 "fields of study" [40]. No visualization of the graph is provided, although it can be queried using different search methods, and an API is provided. The "fields of study" classifications have been found to be dynamic and too specific, and the hierarchies not always meaningful [30].

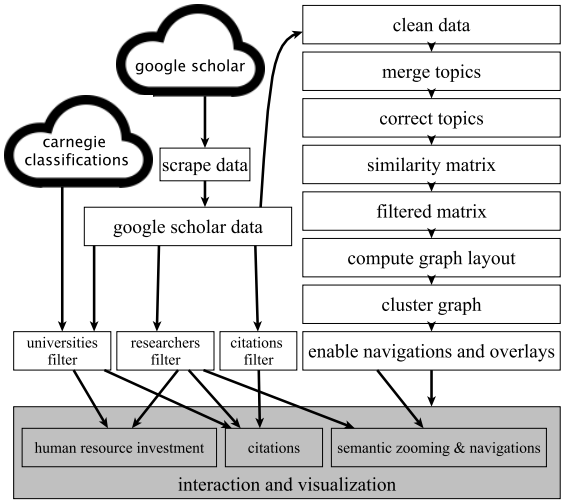


Figure 2: Overview of the GRAM system.

Topic extraction: The hierarchical latent tree method extracts a set of hierarchical topics to summarize the corpus at different levels of abstraction - where a "topic" is determined by words that appear in high frequency in the topic, and low frequency in others. While this method has been implemented in a visual analytics system [45] it has not, to our knowledge, been applied to extensive databases containing a large number of topics. Another analysis of a limited corpus of papers (Proceedings of the National Academy of Sciences from 1982-2001) uses a burst detection algorithm and co-word occurrence analysis to find salient topics and trends over time [32].

Aside from analyzing the full text of articles, topics have also simply been identified from the combination of paper titles and cited references (papers in the "Information Science" journal [43]), paper titles only (computer science papers [23]), and medical records [46]. In all these cases, the methodology proposed is demonstrated on only a small, well-defined corpus. Other examples of limited application of topic extraction include: computer science conferences and journals from the DLPB database [23], trends in computer science research [20], and publications in data visualization [28].

Graph visualization: Graph drawing libraries and toolkits make it easy for a graph to be visualized (e.g., GraphViz [5], OGDF [18], MSAGL [33], and VTK [38]); few of these support interaction or navigation – features essential for exploration of large graphs. Even visualization toolkits supporting graph manipulation (e.g., Prefuse [27], Tulip [10], Gephi [12], yEd [44]) have difficulty rendering large graphs in a manner that makes them easy to use. Multi-level interfaces for large graph exploration (e.g ASK-GraphView [9], topological fisheye views [25], and Grokker [35]) and domain-specific software (Pajek [19] for social networks, and Cytoscape [39] for biological data) rely on meta-graph information comprising meta-nodes and meta-edges: representations that make direct navigation of and interaction with large graphs counter-intuitive.

Mapping knowledge is often associated with trying to map the information, often in a form that relates to geographical maps, e.g., "Atlas of Science" [13] and "Atlas of Knowledge" [14]. Navigating and reading large networks represented as node-link diagrams are often difficult for non-experts while map-based visualizations of the same type of data have been shown to be both effective (in terms of time and error) [37] but also more memorable and engaging [36].

Our contributions: In the context of the prior research on knowledge classification, topic identification and large graph visualization, the contribution of our work include

(1) An approach to deriving a representation of knowledge is based on bottom-up analysis of publicly available data relating to active researchers (rather than research outputs).
(2) A system that allows us a glimpse in this large knowledge landscape using a geographical map metaphor and supporting map-exploration interactions: semantic zooming, panning, searching, and application of overlays.
(3) A system that provides real-time semantic zooming interactions with a large graph implemented in the browser.
(4) Overlays which highlight human resource investment and areas of research activity for individual universities or for aggregates (e.g, by type of university), functionality which to the best of our knowledge is not available in other systems.

3 NETWORK GENERATION

A knowledge network represents topics as nodes and uses edges to 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 [23, 45] - but these methods do not allow for easy identification of general topics (e.g., mathematics, physics) as sub-graphs, nor do they include very specific topics (e.g., symmetry detection algorithms, interactive graph visualization) as nodes.

Our approach rests 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 as defined by researchers stored in the Google Scholar (GS) database (note that other sources such as DBLP [3], index only a subset of science publications and do not provide research topics associated with publications). In GS, each researcher listed can modify their profile to list the research topics that they work on, and the co-occurrence of topics within a researcher's list indicates a relationship between them in our knowledge network.

Our network is generated using the following steps: scraping the data from GS, extracting information about researchers and their topics, cleaning the data to reduce ambiguity and duplication in the topic labels, splitting topic phrases into constituent topics, merging topics with common stems, and correcting anomalies. The result of this process is a diagonal similarity matrix with topics as rows and columns, with each cell representing the similarity between a pair of topics, calculated as the number of times the topics co-occur.

Data Scraping: While some prior research of GS data exists [11, 22, 31], these tend to focus on analysis and comparisons of index and citation data, rather than research topics. Data retrieval from GS is laborious due to the lack of an API and metadata scarcity [16]. The scope of our data is defined by including all GS entries associated with the world's top 1,000 universities (as listed by the Center for World University Rankings [2]). We extracted the institution IDs from GS (for example, MIT's ID is 6345133980181568013) and then scraped the URL associated with each institution to collect research profiles of all individuals associated with the institution. Using a regular expression to match relevant fields in the HTML, we collected name, affiliation, total number of citations, and list of research topics from each research profile. The total number of topics extracted 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.

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,"); we removed conjunctions (and, or) and other words with no semantic weight (for, of), thus splitting topic phrases into their constituents.

Topic correction: Recent changes to GS mean that researchers creating their topic list are prompted with auto-suggestions, and are limited to five topics. Previously there was no constraint on the number of topics, and they were all self-defined. Hence, there are naturally a large number of typing errors and acronyms in the dataset. We used Google's OpenRefine [6] to identify and resolve typing errors, and to find alternate representations of the same topic [17, 21, 29] (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.

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.

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 [34] 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 avoid this, we choose the main topic to be the one with the highest frequency amongst all topics with the same stem. This resulted in 35,028 topics.

Network reduction: We further reduced the size of the network by removing leaf nodes, i.e., nodes that have only one edge connecting them to other nodes. This brought the number of nodes to 34,774.

The final network contains 34,774 nodes and 646,582 edges. There are 17 components, including one giant connected component (34,741 nodes and 646,565 edges). The average shortest path length is 3.141, indicating that the topic network is highly connected. The graph has a low global clustering coefficient of 0.09 (defined as the ratio of the number of triangles over the total number of node triples) suggesting that topics are not typically tightly clustered into

Topics	Degree
machine learning	3314
artificial intelligence	2404
neuroscience	2033
modeling	1902
bioinformatics	1878
climate change	1846
optimization	1827
education	1808
nanotechnology	1788
statistics	1659

Topics	Researchers
machine learning	10726
artificial intelligence	5766
neuroscience	5655
computer vision	5372
bioinformatics	4943
robotics	3398
data mining	3334
ecology	3281
materials science	3193
genetics	2951

Figure 3: Top 10 topics by degree and number of researchers.

triples. The node "machine learning" has the highest degree and more researchers report working on this topic than any other. Figure 3 shows the top ten topics by degree and by number of researchers.

Interestingly, some universities seem to have more GS profiles than academic staff, (likely due to doctoral and postdoctoral students), although the majority of the universities are associated with fewer profiles than the size of their academic staff. On average most universities in our list are well represented by GS profiles; see Fig. 4.

Name of University	Acad. Staff	# Profiles
Stanford University	2118	8104
University of Washington	5803	5562
Harvard University	4671	5356
Massachusetts Institute of Technology	1021	3527
University of Michigan	6771	3413
University of Toronto	2547	3148
University of Cambridge	6645	2669
Texas A&M University	2700	2515
University of Minnesota	3804	2511
Pennsylvania State University	8864	2368

Figure 4: Academic staff size (Wikipedia) and GS profile count.

4 MAP GENERATION

Map-like representations provide a way to visualize relational data. Graphs are a standard way to visualize relational data, with the objects defining nodes and the relationships defining edges. It requires an additional step to get from graphs to maps: clusters of well-connected vertices form countries, and countries share borders when neighboring clusters are tightly interconnected. Maps are helpful in visually representing clusters by explicitly defining the boundary of the clusters and coloring the regions. While it often takes us considerable effort to understand graphs, a map representation is intuitive, as most people are familiar with the notions of searching, panning, and zooming. Finally, while edges in large graphs often end up creating a "hairball" effect, edges can be removed from maps as we rely on the Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things" [42].

We reduced the network a bit more with an edge filter: only those pairs of nodes corresponding to topics that co-occur at least 10 times were retained. We call this the BaseGraph-1000 network and it contains 6,052 nodes and 26,162 edges as it is based on GS profiles from researchers in the top 1,000 universities in the world [2].

Visualizing the BaseGraph. The network was first embedded on the plane using the Scalable Force-Directed graph layout algorithm provided by GraphViz [5]. K-means clustering was then used to group nodes into topic-clusters. To create the geographic map look, we use a modified Voronoi diagram based on the embedding and clustering, ensuring that the geographic regions are colored such that no two adjacent countries have similar colors, using the spectral vertex labeling method, following the GMap framework [26]. Each geographic region represents a topic cluster. This diagram is the BaseMap-1000.

Exploring the BaseMap. GMap produces a visual map from a given graph which is a static image that is not ideal for user interaction, such as zooming, panning, and searching. We enable exploration of the BaseMap with the help of the google maps API [4]. 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), and provide eight zoom levels, each showing subgraphs at varying levels of detail.

Labelling the Topics. Each node in the network represents a single topic and when viewing the BaseMap topic labels should not overlap. We use the GraphViz implementation of node-overlap removal provided by PRISM [24]. However, we wished to allow for seven further levels of semantic zooming, each providing more detail of the map. The Google Maps API handles the modifications required for nodes, edges and clusters (and heatmaps, to be discussed later), but does not cater for the introduction of node-label overlaps as more detail is revealed through zooming.

To ensure that neither nodes nor labels overlap at any zoom-level, we compute different node visibilities for different zoom-levels. For each zoom-level, we sort the nodes by their weight, where node weight is proportional to the number of researchers working on the topic associated with the node. We make *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 $1, 2, \cdots, (i-1)$. Figure 5 shows how the local neighborhood of the "computer vision" topic is changing in different zoom levels.

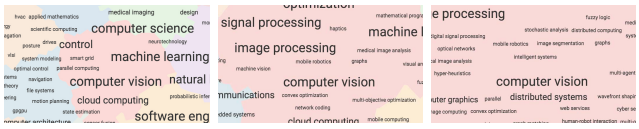


Figure 5: Three zoom-level views near "computer vision."

The size of the font label for topic t is directly proportional to the number of researchers working on that topic, denoted by the weight: $w(t)$. We assign font size from the range 100% to 300% of the default browser font size, as follows:

$$\mathscr{F}_t = \begin{cases} 100 & \text{if } w_t/10 \leq 100 \\ 300 & \text{if } w_t/10 \geq 300 \\ w_t/10 & \text{otherwise} \end{cases}$$

Searching the BaseMap. We provide basic search functionality, which locates topics in the map containing given query terms. Clicking on a node shows the number of people who work on that topic and highlights edges to adjacent nodes, that is, the other topics that are frequently co-listed with that topic.

Alternative BaseGraphs. The complete "whole-world" of topics is represented in the BaseGraph-1000 network, aggregating data over all 1,000 institutions. Different base graphs can be obtained by aggregating the data over specified sub-sets of institutions: this changes the weighting of the topic nodes (representing the number of researchers working on that topic) and the connections between the nodes. Indeed, some nodes may even disappear if no researcher in the set of specified institutions works on that topic. As well as the BaseGraph-1000, we have a base graph network which aggregates all 215 USA universities represented within the original 1,000 (BaseGraph-USA). Both these base maps can be used as reference points for comparing individual universities (or sets of universities).

5 STRENGTHS AND WEAKNESSES

Each topic node in the graph is associated with the name, institution, departmental affiliation and the number of citations for each researcher who works in that topic area. We use this information to facilitate further exploration of the data by departmental affiliation, via visual "overlays" representing knowledge strength and weaknesses (in terms of number of researchers and citations).

When visualizing these overlays we use the same visual map representation as BaseMap-1000 (thus preserving the reader's mental map by ensuring that topics are always located in the same position), but since the number of researchers working on each topic will vary, the node-weights (and hence the label font-sizes) are different.

Human resource investment: We use the data associating topics with researchers (and their institutions) to provide an overlay representation of human resource investment (HRI) in each topic for a given institution, relative to the HRI of a larger set of institutions.

For example, we can compare the HRI for the aggregated set of universities in Europe with reference to the "whole-world" topic graph (BaseGraph-1000), indicating where the HRI for each topic is higher or lower than the "whole-world" average. That is, to determine the HRI in topic t in European universities when compared with the reference network BaseGraph-1000, we calculate the difference between the percentage of researchers in Europe who work on topic t and the percentage of researchers in BaseGraph-1000 who work on topic t. If this difference is positive (negative) then we consider this a human resource strength (weakness) of the set of universities. This is illustrated on the BaseMap-1000 with circles of different color: green for strength and purple for weakness. The size of the circles is proportional to the magnitude of the difference and this is shown in a legend; see Fig. 6.

Citations: Each topic in a base graph has associated with it the total number of research article citations, calculated as the sum of the citation count (as recorded by GS) for all researchers working on the topic. In the absence of information as to how citations are distributed amongst the several topics that a researcher works on, we associate the citation count for one researcher with all the topics that the researcher works on. Thus, for each of our base graphs, we have a citation count for each topic, which can be overlaid on the BaseMap-1000 and visualized as with proportional circles or with a heatmap (as done in most examples in this paper); see Fig. 6.

These citation counts are raw aggregates, and do not take into account the fact that not all research fields cite at the same rate; e.g., "particle physics" is associated with more citations than average

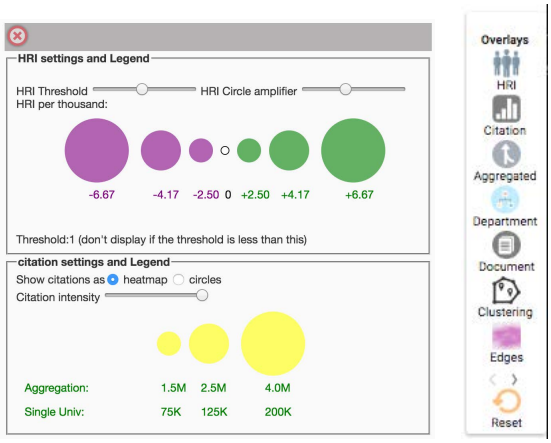


Figure 6: Legends (for HRI and citations) and settings of the system (e.g., for selecting individual or aggregate overlays).

(due to high number of co-authors and citations per paper). Figure 7 shows the topics with the highest number of citations per person.

Topics	Cite/Person
particle physics	15906
high energy physics	15768
cosmology	7037
...	
...	
(2563) information visualization	1642
(2799) artificial intelligence	1551
(3526) machine learning	1263

Figure 7: Topics ordered by number of citations per person in the Base-1000 graph.

With this in mind, we provide a normalized citation heatmap visualization. The normalized citations for topic t at a given university X with respect to BaseGraph B is $nc(t,X,B) = c_X(t) * c(t)/C$. Here $c(t) = \sum_{r \in B \& t \in r} cite(r)$ is the number of researchers r working on topic t in the BaseGraph universities, $c_X(t) = \sum_{r \in X \& t \in r} cite(r)$ is the number of citations for researchers r working on topic t at university X, and $C = \sum_{r \in B} cite(r)$ is the number of citations in the entire BaseGraph. For simplicity, $t \in r$ means a topic t from the list of topics for researcher r, and $r \in B$ means a researcher r from a university in B. When data is aggregated over several universities, this formula is extended so that X represents a set of institutions.

Table 1 shows aggregate normalized citation counts for universities in the USA, in Europe and in Asia. Figure 8 compares the raw citation count with the normalized version for a randomly selected university, with respect to BaseGraph-1000. In the remainder of this paper, we present only the normalized citation count heatmaps.

6 IMPLEMENTATION

For each researcher, our database stores name, GS id, university id, total citations, email address domain name, affiliation, listed research topics, research phrases, and stemmed phrases. We use a variety of tools to clean, store, and process our data: mongodb scripts, sqllite,

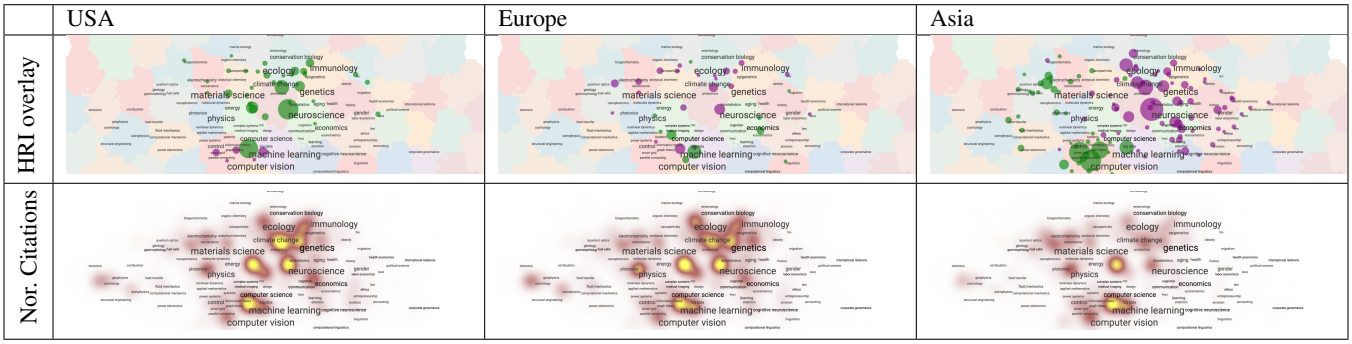


Table 1: Aggregate overlays for universities in the USA, Europe and Asia, each shown in relation to the reference graph BaseGraph-1000. There is a clear progression from left to right: significant HRI emphasis on medical sciences in the USA, contrasted by HRI emphasis on engineering and computer science in Asia.

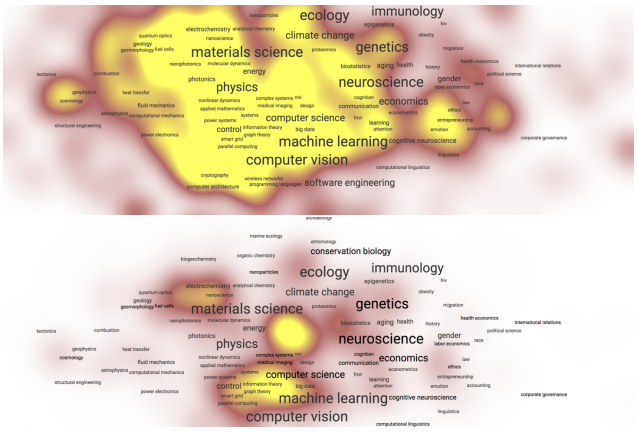


Figure 8: A heatmap showing raw citation and normalized citation counts (which factors in the relative frequency of citation of different research topics) for the same university.

python, R, Java-Lucene, openrefine. The Google Maps API and jquery are used for map drawing and to handle user interaction in the web application. We run python-django for the webserver and mongodb for database storage and query. Generating the topics map in svg format (layout, clustering, node-overlap removal) takes 14 seconds. Loading the initial base map takes 12,638ms, including 7,836ms for scripts, 2,444ms for rendering, and 275ms for painting (in Google Chrome v.58). Interaction with the BaseMap, map navigation, zooming with edges takes 1,515ms. Computing HRI overlays takes 3,129ms and citation heatmaps require 1,231ms. The system currently runs as a virtual machine on a Dell PowerEdge R430 server with 2 Intel(R) Xeon(R) CPU E5-2530 v4 @2.20GHz processors and 32GB of memory.

7 USE CASES

GRAM provides free access to aggregated global research data in a way that no other system does. Such information is of particular interest to university managers and strategists, for whom the ability to compare the performance, strength and weaknesses of their institution against others (or against the aggregates of others) and to explore their researcher profiles can drive their decision-making, planning and institutional reviews. There are two main use cases: external institutional comparison (comparing one institution against others) and internal institutional research profiling (identifying institutional strengths in research areas, and facilitating collaborations).

Commercial organizations provide access to (and often visualizations of) similar data for an institutional fee. For the comparison use case, SciVal (Elsevier) "offers quick, easy access to the research performance of 8,500 research institutions and 220 nations worldwide." Academic Analytics, which focuses on research universities in the USA and the UK, specifically supports "the strategic decision-making process as well as a method for benchmarking in comparison to other institutions." In profiling an institution, Pure (Elsevier) "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 In Cites (Thomson Reuters) 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.

We developed GRAM with input from managers in charge of research and development at the University of Arizona, seeking feedback on the usefulness of the system from the perspective of conducting institutional review, comparison and strategy. The primary visualizations they requested are those that show the strengths and weaknesses of one institution when compared with others. We demonstrate this use case by showing the HRI and normalized citations for two randomly selected universities, using the BaseGraph-1000 for comparison (Table 3).

A further request involved meta-analyses of different university classifications, in order to identify patterns (and confirm some "folklore knowledge"). In particular, we show comparisons between different types of universities included in the list of 115 "Highest Research Activity" universities in the USA, as defined by the Carnegie Classification of Institutions of Higher Education. We compare "public land grant" universities (i.e., those given federal land by the Morrill Acts of 1862 and 1890 specifically for agriculture and mechanical learning) versus "public non-land grant" universities (Table 4), universities with/without Medical Schools (Table 2), and private/public

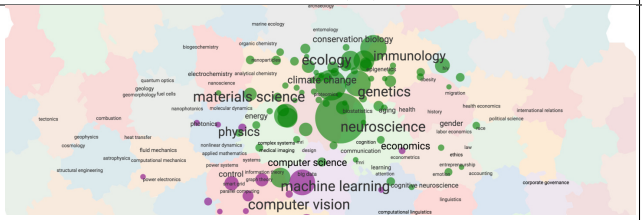
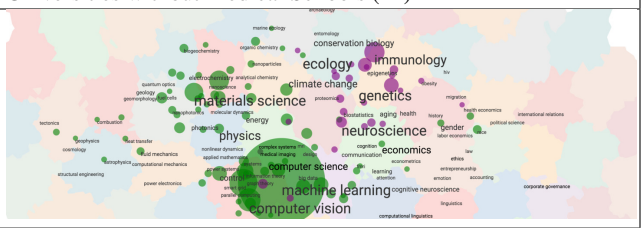
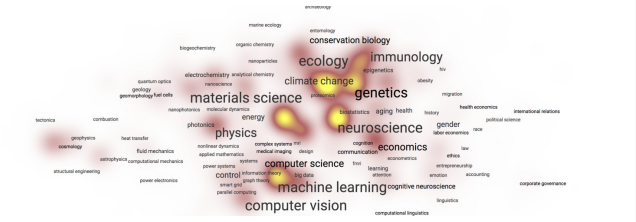
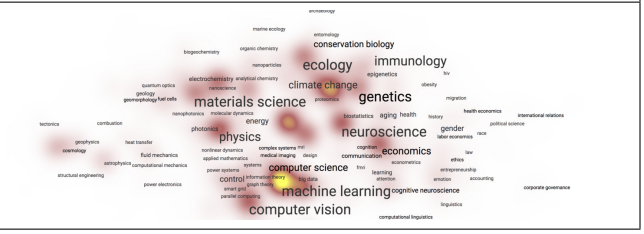
	Universities with Medical Schools (71)	Universities without Medical Schools (44)
HRI overlay		
Normalized Citations		

Table 2: Comparison of universities with/without Medical Schools, showing an expected division in HRI between medical sciences (e.g., genetics, neuroscience) and computer science (e.g., machine learning, computer vision). Note that despite lower-than-average HRI in machine learning for universities with Medical Schools, normalized citations for this topic are comparable to the other universities.

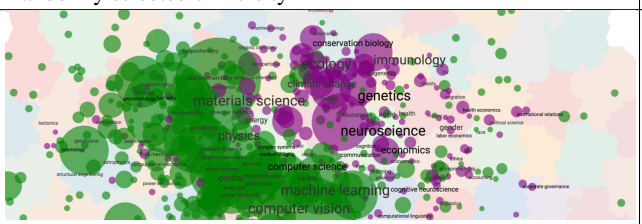
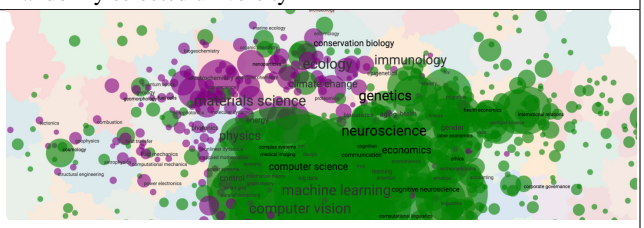
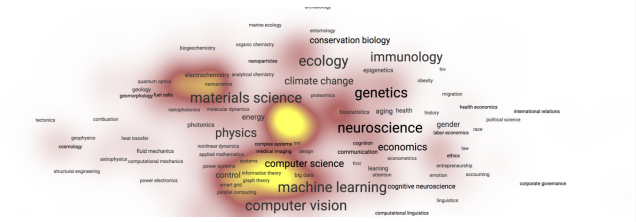
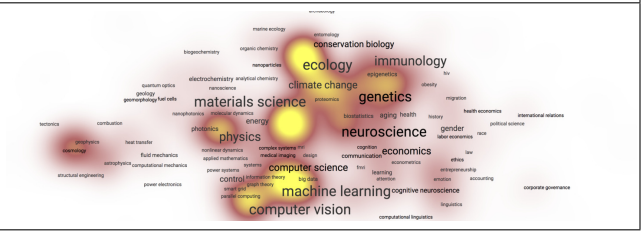
	Randomly selected university 1	Randomly selected university 2
HRI overlay		
Normalized Citations		

Table 3: The HRI, citations and normalized citations for two randomly selected universities when compared with the BaseGraph-1000. University 1 shows a clear strategy in investing in physical sciences and technology (at the expense of medical sciences). University 2 does not show any obvious HRI strategy, with lower HRI topics in the same topic clusters as higher HRI topics, and, despite its higher-than-average investment in most topics across the whole map, it has very low normalized citation counts.

members of the Association of American Universities (Table 5). In all cases, the images are shown with reference to BaseGraph-USA.

The response from the stakeholders to the GRAM system is overwhelmingly positive; they particularly welcome its flexibility. Despite the acknowledged limitations of the data source (discussed below), they see the system as being highly instrumental for informing senior management about research strengths and weaknesses of our institution, and in influencing future strategy.

8 DISCUSSION AND LIMITATIONS

We use Google Scholar as the source for our data, with all of its advantages (e.g., a large amount of information) and 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. Researchers from different universities also use Google Scholar profiles at different rates.

Once the data has been gathered, the choice of universities used to create base graphs has a non-trivial impact on the comparisons made. Focusing only on English language terms biases the results, and despite our attempts to clean, split and merge topics, several issues remain. For example, use of acronyms (e.g., NLP for natural language processing) requires further expansion and merging.

Our HRI-based strengths and weaknesses calculations are associated with other biases: the numbers used are not guaranteed to be accurate reflections of an institution's human investment in a

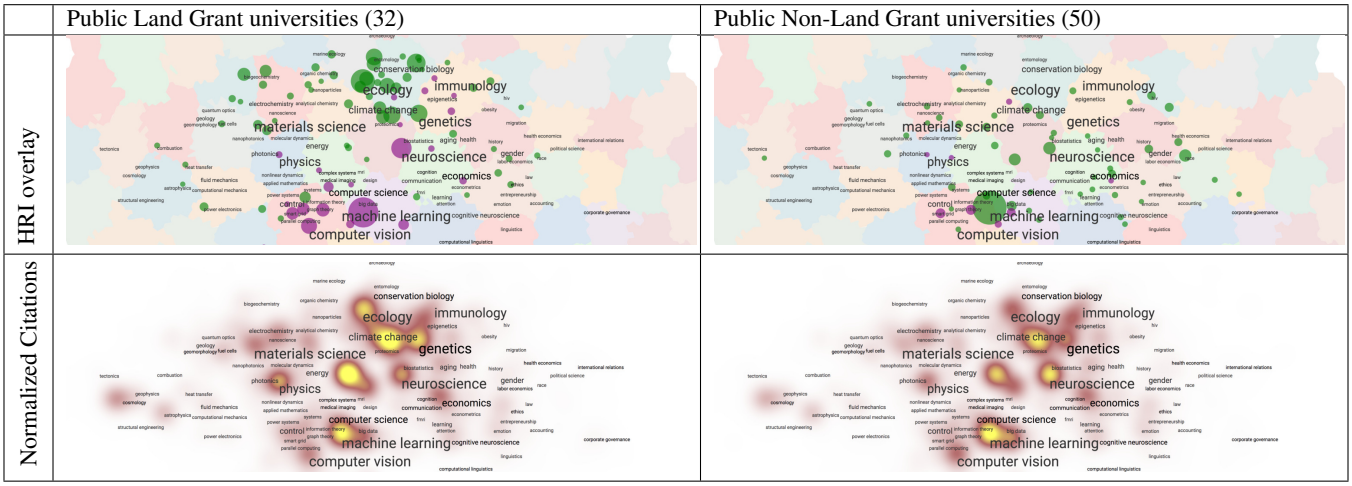


Table 4: Comparison of Public Land Grant and Public Non-Land Grant Universities in the USA, showing the expected dominance of agricultural topics (including ecology and conservation biology) on the left, and a "close-to-average" profile on the right.

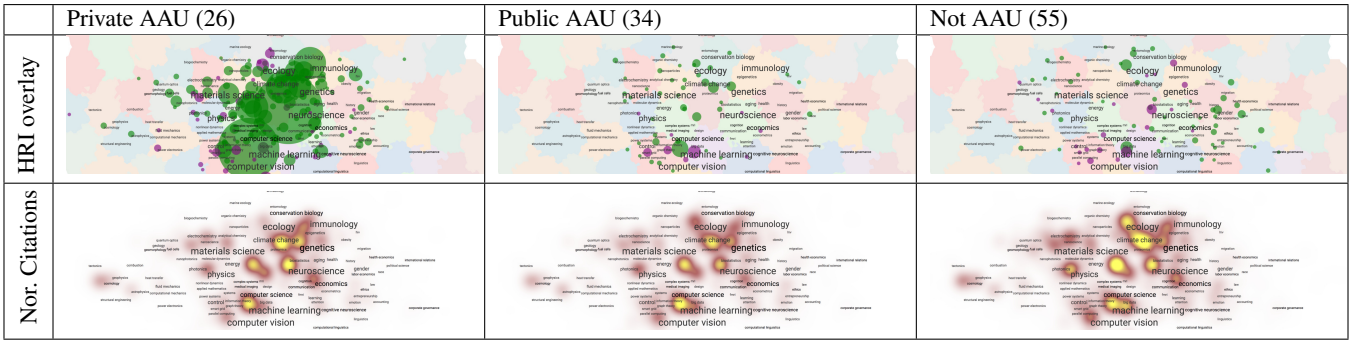


Table 5: Comparison of universities with respect to membership of the Association of American Universities. The private universities invest most heavily in the hottest current topics: machine learning and neuroscience, genetics, and immunology. The other two profiles show a comparatively more balanced research profile, both surprisingly with under-investment in machine learning.

research topic since we cannot distinguish tenure-track faculty from other type of staff (e.g., doctoral and postdoctoral students) in the GS profiles. Citation-based calculations are also biased, e.g., due to misattributed papers, the difficulty in perfectly matching specific citations to specific topics associated with a researcher, and distributing the citation contribution among its co-authors.

9 CONCLUSIONS AND FUTURE WORK

One of the impacts of big data is an unexpected one, where solutions to problems are being overlooked, as many tasks must cross several disciplines/domains that produce considerable amounts of data but interact only minimally. "Undiscovered public knowledge," named so by Swanson [41], is exactly such an example as knowledge can be public, yet undiscovered, if independently created fragments are logically related but never retrieved, brought together, and interpreted.

In the proposed GRAM system we attempt to retrieve publicly available data, bring it together via text processing and graph and map visualization techniques, in order to interpret and analyze worldwide research activity. Despite non-trivial limitations, the system is novel as it is based on large quantities of real, self-reported, bottom-up information, unlike traditional top-down hierarchical taxonomies and ontologies (which depend upon the creation of abstract category labels). The GRAM system implements in-the-browser, map-based interactive navigation of a large underlying network, supports panning, zooming and searching, and (with the help of map overlays) makes it possible to visualize human resource investments and scholarly output for different academic institutions. The GRAM system is open source and is available here: https://uamap-dev.arl.arizona.edu/ To the best of our knowledge, this is the only free, publicly available tool enabling global overview of research topic activity, researcher investment and researcher outputs.

Adding more data can augment the picture of a specific university, or enable more detailed comparisons between different universities. Discussions with university stakeholders indicate that there is a real demand for a tool such as GRAM that facilitates both comparison with competitor or benchmark institutions, while at the same time providing information about active institutional research that can help in directing university strategy.

REFERENCES

[1] [n. d.]. 2010 Mathematics Subject Classification - MSC2010 database. www.ams.org/msc/msc2010.html. Accessed: 05-04-2017.

[2] [n. d.]. CWUR | Center for World University Rankings. http://cwur.org/. Accessed: 09-13-2016.

[3] [n. d.]. dblp: computer science bibliography. http://dblp.uni-trier.de/. Accessed: 20-01-2018.

[4] [n. d.]. Google Maps APIs | Google Developers. https://developers.google.com/maps/. Accessed: 09-27-2017.

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

[6] [n. d.]. OpenRefine. http://openrefine.org/. Accessed: 05-04-2017.

[7] [n. d.]. Scopus | The largest database of peer-reviewed literature | Elsevier. https://www.elsevier.com/solutions/scopus. Accessed: 09-27-2017.

[8] [n. d.]. Web of Science - Clarivate Analytics. http://wokinfo.com/. Accessed: 09-27-2017.

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

[10] David Auber, Daniel Archambault, Romain Bourqui, Antoine Lambert, Morgan Mathiaut, Patrick Mary, Maylis Delest, Jonathan Dubois, and Guy Mélançon. 2012. *The Tulip 3 framework: A scalable software library for information visualization applications based on relational data*. Technical Report RR-7860. INRIA.

[11] Judit Bar-Ilan. 2007. Which h-index?-A comparison of WoS, Scopus and Google Scholar. *Scientometrics* 74, 2 (2007), 257–271.

[12] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. *ICWSM* 8 (2009), 361–362.

[13] Katy Börner. 2010. *Atlas of Science: Visualizing What We Know*. The MIT Press, Cambridge, MA.

[14] Katy Börner. 2015. *Atlas of Knowledge: Anyone Can Map*. The MIT Press, Cambridge, MA.

[15] Katy Börner, Richard Klavans, Michael Patek, Angela M. Zoss, Joseph R. Biberstine, Robert P. Light, Vincent Larivière, and Kevin W. Boyack. 2012. Design and Update of a Classification System: The UCSD Map of Science. *PLoS ONE* 7, 7 (07 2012), e39464.

[16] Lutz Bornmann, Andreas Thor, Werner Marx, and Hermann Schier. 2016. The application of bibliometrics to research evaluation in the humanities and social sciences: An exploratory study using normalized Google Scholar data for the publications of a research institute. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2778–2789.

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

[18] Markus Chimani, Carsten Gutwenger, Michael Jünger, Gunnar W Klau, Karsten Klein, and Petra Mutzel. 2011. The open graph drawing framework (OGDF). *Handbook of Graph Drawing and Visualization* (2011), 543–569.

[19] Wouter De Nooy, Andrej Mrvar, and Vladimir Batagelj. 2011. *Exploratory social network analysis with Pajek*. Vol. 27. Cambridge University Press.

[20] Suhendry Effendy and Roland H.C. Yap. 2017. 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)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1245–1250. https://doi.org/10.1145/3041021.3053064

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

[22] Matthew E Falagas, Eleni I Pitsouni, George A Malietzis, and Georgios Pappas. 2008. Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. *The FASEB journal* 22, 2 (2008), 338–342.

[23] Daniel Fried and Stephen G. Kobourov. 2014. Maps of Computer Science. *2014 IEEE Pacific Visualization Symposium (PacificVis)* 00 (2014), 113–120. https://doi.org/doi.ieeecomputersociety.org/10.1109/PacificVis.2014.47

[24] Emden Gansner and Yifan Hu. 2010. Efficient, Proximity-Preserving Node Overlap Removal. *Journal of Graph Algorithms and Applications* 14, 1 (2010), 53–74. https://doi.org/10.7155/jgaa.00198

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

[26] E. R. Gansner, Y. Hu, and S. Kobourov. 2010. GMap: Visualizing graphs and clusters as maps. In *2010 IEEE Pacific Visualization Symposium (PacificVis)*. 201–208. https://doi.org/10.1109/PACIFICVIS.2010.5429590

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

[28] F. Heimerl, Q. Han, S. Koch, and T. Ertl. 2016. CiteRivers: Visual Analytics of Citation Patterns. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan 2016), 190–199. https://doi.org/10.1109/TVCG.2015.2467621

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

[30] Sven E. Hug, Michael Ochsner, and Martin P. Brändle. 2016. Citation Analysis with Microsoft Academic. *CoRR* abs/1609.05354 (2016). http://arxiv.org/abs/1609.05354

[31] Péter Jacsó. 2005. Google Scholar: the pros and the cons. *Online information review* 29, 2 (2005), 208–214.

[32] Ketan K. Mane and Katy Börner. 2004. Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences* 101, suppl 1 (2004), 5287–5290. https://doi.org/10.1073/pnas.0307626100

[33] Lev Nachmanson, George Robertson, and Bongshin Lee. 2008. Drawing graphs with GLEE. In *Graph Drawing*. Springer, 389–394.

[34] Martin F Porter. 2001. Snowball: A language for stemming algorithms.

[35] Walky Rivadeneira and Benjamin B Bederson. 2003. A Study of Search Result Clustering Interfaces: Comparing Textual and Zoomable User Interfaces. *Studies* 21 (2003), 5.

[36] Bahador Saket, Carlos Scheidegger, Stephen G. Kobourov, and Katy Borner. 2015. Map-based Visualizations Increase Recall Accuracy of Data. *COMPUTER GRAPHICS FORUM* 34, 3 (2015), 441–450.

[37] Bahador Saket, Paolo Simonetto, Stephen Kobourov, and Katy Börner. 2014. Node, Node-Link, and Node-Link-Group Diagrams: An Evaluation. *IEEE Transactions on Visualization & Computer Graphics* 20, 12 (2014), 2231–2240.

[38] William J Schroeder, Lisa Sobierajski Avila, and William Hoffman. 2000. Visualizing with VTK: a tutorial. *Computer Graphics and Applications* 20, 5 (2000), 20–27.

[39] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 13, 11 (2003), 2498–2504.

[40] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june Paul Hsu, and Kuansan Wang. 2015. An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th international conference on world wide web*. ACM, 243–246.

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

[42] Waldo Tobler. 2004. On the first law of geography: A reply. *Annals of the Association of American Geographers* 94, 2 (2004), 304–310.

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

[44] Roland Wiese, Markus Eiglsperger, and Michael Kaufmann. 2001. yFiles: Visualization and Automatic Layout of Graphs. In *GD*. 453–454.

[45] Yi Yang, Quanming Yao, and Huamin Qu. 2017. VISTopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. *Visual Informatics* (2017), –. https://doi.org/10.1016/j.visinf.2017.01.005

[46] Naiyun Zhou, Joel Saltz, and Klaus Mueller. 2016. *Maps of Human Disease: A Web-Based Framework for the Visualization of Human Disease Comorbidity and Clinical Profile Overlay*. Springer International Publishing, Cham, 47–60. https://doi.org/10.1007/978-3-319-41576-5_4