Universiteit Leiden

Opleiding Informatica

ACE: Academic Collaboration Explorer

An author-centered view on academic collaboration

Name: Gilles B. Ottervanger

Date: August 23, 2018

1st supervisor: Holger H. Hoos
2nd supervisor: Marie Anastacio

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Introduction

An important part of academic research is working together with others [1]. This consists of reading and reviewing literature written by colleague researchers, collaborating on joint projects and attending conferences. Central to this, is staying up to date on work of colleague researchers. However, more than a million new academic articles are published every year, and this number is increasing exponentially [2]. As a result, getting insight and overview of academic output has become increasingly difficult.

Multiple approaches are possible for extracting information from research output. The first aspect to consider, is the kind of information that is desired. Information can be centered around different kinds of *objects of interest* like journals, fields of research, authors or individual papers. The extracted information can either provide details of one object of interest, or give an overview of many objects and their interrelations. The second aspect to consider, is how to present the extracted information. This depends on what information is extracted. Possibilities include tabular representation and different kinds of visualisations like charts, graphs or maps.

In this thesis, we propose an approach focused on co-authorship relations. Our goal is to give insight into collaboration between authors. Co-authorship relations are well-documented and easy to understand relations that describe collaboration between authors. The idea was to consider one author or a group of authors of interest and provide an overview of the collaborative relations with other researchers, thus creating a profile of the authors of interest and simultaneously showing their connections with the rest of the academic community. We chose to represent this relational information visually using an interactive graph. This graph uses different visual features to give more information at a glance and additionally uses interactive features to provide more detailed information. This provides both quick insight in the information and allows exploration of the underlying data.

We created a simple and accessible tool, called Academic Collaboration Explorer or ACE. ACE is a web-based application that enables the user to create author-centered graphs. In the produced graph, the authors of interest are placed at the center. The center node is surrounded by nodes representing research institutes that are related to the center node by co-authorship. This way, the graph provides an initial overview of the co-authorship relations. Interaction with the graph enables the extraction of more detailed information and provides links to external resources. Additionally, ACE enables filtering of the data represented in order to get more fine-grained insights.

The strength of ACE lies in three important aspects. Firstly, ACE is easily accessible. ACE is implemented as a purely web-based application, requiring no installation or browser extensions. Starting ACE is as simple as visiting a web page. ACE is designed for researchers of any field and does not require specific knowledge of data visualisation or bibliometric analysis. This is in contrast to most existing tools for data visualisation which commonly require both installation of software and collection and processing of data.

Secondly, ACE provides an author-centred view. This way of visualisation provides relevant information to the user in an understandable view. Many existing bibliometric visualisations focus on networks of authors or

journals. Although these large scale visualisations are interesting, they mostly appeal to people interested in bibliometrics and provide less insight on the level of individual authors. ACE, on the other hand, is created for a broader public, and provides information that is relevant for researchers of any field. The third strength of ACE lies in its intuitive design. By providing an simple and clean user interface, information is presented in a manageable way reducing information overload. Links to external resources provide the option to continue exploration, even when the limits of ACE itself are reached. This means that ACE can either be used to extract information directly or as a tool to navigate to other sources of information.

This thesis is structured as follows. Chapter 1 addresses related work in the fields of bibliometrics and interactive visualisation design and shows tools comparable to ACE. Chapter 2 provides a detailed description of the implementation of ACE as a web application. This includes data retrieval, graph generation and interactive design. In Chapter 3, the practical value of ACE is illustrated by discussing several use cases. Finally, in Chapter 4, we summarise our work, discuss the possible future development of ACE and the path that future research might follow in the field as a whole.

Chapter 1

Related Work

The work that is relevant for this project can be divided into three separate parts. We first consider research in the field of *bibiliometrics*, concerning the analysis of written publications. Next, we consider *data visualisation*, which is the effective communication of data by producing a visual representation. Since this field is broad, we focus mainly on visualisation of bibliometric data. Third, we consider *interactive design* of visualisations. Interaction enables the extraction of more detailed information from the data. Interactive design is an important part of the user experience. Finally, this chapter is concluded by a set of prominent applications within the area of bibliometric visualisation that illustrate how the theory is put into practice.

1.1 Bibliometrics

The field of bibiliometrics concerns the analysis of written publications. Although the main objects of study in bibliometrics are publications, in many cases it can also provide information about related objects of interest, including journals, fields of research and authors. In bibliometrics, multiple techniques are used, which can be separated into two main categories: evaluative techniques and relational techniques [3]. Evaluative techniques assess productivity or impact of the object of study. Paper citation count is used to measure impact of publications. More complex metrics have been devised to give measures of different objects as well. An example is *journal impact factor*, which is generally measured as an average number of citations per publication per year over a period of two years [4]. For authors, the h-index can be used. The value of the h-index is the highest value of h for which an author has published at least h papers with at least h citations [5]. In policy and management context, evaluative metrics are used to evaluate individual researchers for promotion and funding [6].

Relational techniques enable the quantification of the relations between objects. Different metrics are used to relate different kinds of objects. Co-citation relates papers that are likely covering similar topics. Two papers are said to be co-cited when they are both cited by a third publication. Co-word (or co-term) relations indicate the appearance of similar words in two texts and therefore, similar to co-citation, indicate similarity between the research topics of two papers. Relations between papers can be aggregated in order to relate journals, and they can be clustered to reveal different fields and sub-fields and their interrelations [7]. Co-authorship is a metric indicating relationships between authors, and reflects the number of papers two authors have both collaborated on. The main focus of this research is these co-authorship relations.

Analysis of co-authorship has some important benefits. Firstly, co-authorship is well-documented. In contrast to co-citation or co-word metrics, co-authorship does not require large scale analysis or complex computation: it can be directly retrieved from publication meta-data. Secondly, co-authorship is easily interpreted. Co-citation metrics are biased by the fact that researchers have varying motivations for citing literature aside from academic relevance. For example, researchers tend to cite papers by colleagues more frequently [8].

Co-word metrics are expressed as the inverse of the distance or the cosine of an angle between vectors in high-dimensional space, making them hard to interpret. Co-authorship, on the other hand, is simply the number of co-authored publications making it easy to understand. A third advantage of co-authorship is that studies about the topology of co-authorship networks have shown relations between network metrics (e.g. network efficiency and node centrality) on the one hand, and quality and impact of academic output on the other hand [9, 10, 11]. This suggests that providing an overview of co-authorship relations can potentially also provide valuable insight into the academic impact of one or more authors.

There are multiple ongoing projects with the goal of collecting and indexing bibliometric data. Examples are Microsoft Academic (MA) [12], Google Scholar [13], AMiner [14], Scopus [15] and Web of Science [16]. Many of these projects enable interaction with this data by searching the database. Additionally, they may provide bibliometric statistics and indices. For a comprehensive list of academic databases, we refer the reader to [17].

1.2   Data visualisation

Data visualisation refers to the process of mapping data onto visual objects (e.g. lines, point or shapes). The main goal of data visualisation is communicating data in an efficient way. It enables the interpretation of large amounts of data and enables the communication of hidden or emergent patterns in the data. Furthermore, data visualisation can help a broader, possibly non-technical, audience gain an overview of general patterns [7]. A straightforward method of representing data is the use of a table, where each row represents one object and lists attributes column-wise. Although this can be useful for looking up exact values, it is difficult to observe relations and get an overview of the data. Information hidden in data can be conveyed more effectively using graphical methods like plots and charts.

Data visualisation is an important tool for representing bibliometric data, especially in the case of relational data. An example is journal maps, where journals are mapped as nodes and the distance between the nodes represent similarity of covered topics. Similar mappings can be produced using academic fields specialities, conferences or authors. The production of these maps is referred to as *similarity mapping* and requires a mapping of multi-dimensional similarity data onto a 2d space [18]. These mappings are non-trivial and can be produced by using dimensionality reduction techniques, including principle component analysis (PCA), factor analysis (FA), multidimensional scaling (MDS), Pathfinder network scaling, and self-organizing maps. A detailed description of these methods can be found in [7].

Another common form of data visualisation is network visualisation. Network visualisation can be used to visualise co-authorship relations, where the nodes of the network represent authors, and nodes are connected by an edge if the two corresponding authors have co-authored at least one publication together. Mapping networks onto a 2d space can be done using a *force directed layout* [19]. The general idea of a force directed layouts is the simulation of forces among the nodes. The simulation reaches a mechanical equilibrium when the nodes are configured so that all forces cancel out. This equilibrium state is the force directed layout. Below, a more technical description of the production of force directed layouts is provided.

The simulation starts with initial positions for all nodes, this can be done randomly. However, random initialisation will result in inconsistent layouts for different visualisations of the same data. A solution is a pseudo-random initialisation that produces the same initial positions if the data is the same. This results in a more consistent layout. Other methods can be employed in order to assure convergence in case of dense networks. However, since the network we aim to visualize is sparse, we do not need these methods and they will not be discussed here.

There are two forces that are of primary importance to force directed layouts. First is a repulsive force between nodes comparable to electric charge based on Coulomb's law: the force is inversely proportional to the square of the distance between two nodes [19]. The second is a spring-like attractive force between nodes

connected by edges. The attractive force is based on Hooke's law: the force is proportional to the distance [19]. In the force simulation nodes are allowed to move as a result of the forces acting on them. The desired layout is achieved by reaching a configuration were the repulsive and attractive forces are in equilibrium. A useful addition to the primary forces is implementing friction. This damps harmonic oscillation and thereby enables settling of the layout.

There are several benefits to the use of force directed layout. Force directed layout is simple and easy to understand. It can be easily understood based on intuition and knowledge about common physical objects like springs. A more important benefit is that it produces good quality results with uniform node distributions which are both understandable and aesthetically pleasing. An additional benefit is the flexibility of force directed layout. Additional forces and constraints can be implemented in order to express additional information or to meet additional aesthetic criteria. For a more technical description of force directed layout we refer to [19].

In addition to the layout of the visual objects, more information can be encoded in visual features like shape, size, colour, opacity and labels of objects. When considering a network representation, the main layout and links represent the relations between nodes. Additional visual features can add information about individual nodes by adding colour and size to represent relative importance. Visual features added to edges can represent more detailed information about the relations between the nodes. In context of bibliometric visualisation, the layout and visual features of edges can represent relational information where visual features of nodes can encode evaluative metrics. This design has been applied in ACE.

1.3   Interactive visualisation design

In order to facilitate exploration of data visualisations, interactive components can be implemented. Interactiveness enables a more detailed view of part of the presented data. Interactiveness can include panning and zooming, filtering, querying and view distortion (e.g., a fish-eye effect) [7]. Interactive design can implement *detail on demand*: Users are able to view more detailed information about select parts of the data, allowing to *drill down* from a broad overview to detailed data.

Pirolli and Card introduced the theory of information foraging [20]. In this theory, a user interacting with a program is described as hungry for information. The user acts on imperfect proximal information in order to feed this hunger. Therefore, in order to keep the user involved there needs to be either information presented that satisfies the users information hunger or there should be promising heuristic cues that the user can follow in order to continue the search for information. These cues indicate either value of information or the possibility to access more information. In information foraging theory, these cues are called *information scent* [20]. Within the context of bibliometrics, scent can for instance be a paper title or the number of citations. Within a broader context, scent can be cues like size, shape, colour and location of objects. This idea has been incorporated in the design of ACE in order to facilitate exploration.

1.4   Real-world examples

The rest of this chapter provides a selection of applications of bibliometric data visualisation closely related to our work. Some of the examples are aimed to present similar information as ACE. Others demonstrate visualisation methods that are also applied in ACE. The order of addressing these applications roughly follows the same structure as the discussion of the theory.

1.4.1   Publish or Perish

Publish or Perish (PoP) is advertised as a tool to help individual academics highlight their most important work [21]. PoP provides a tabular representation of publication data. The primary aim of PoP is retrieving

evaluative metrics. PoP provides the ability to search through multiple resources and returns data as a list of papers with bibliometric data including citations, citations per year and rank as well as author list, title and year. PoP also implements interactiveness. The list of papers can for example be sorted by any attribute. The main table allows for further drilling down by linking directly to external resources for viewing the full publication.

In a sidebar, PoP shows aggregated metrics of a selection of papers including the range of publication years, citations per author, papers per author and h-index, among others. This enables to quickly evaluate the impact of an author. In theory, PoP can be used in various cases, for example literature research or for viewing most recent work of a specific author. PoP can be downloaded for free at [21].

1.4.2   VOSviewer

VOSviewer is a tool developed by Van Eck and Waltman [22] for the production of similarity maps. The idea behind VOSviewer is the implementation of the VOS mapping technique. This technique is a form of dimensionality reduction, mapping high-dimensional similarity data onto a 2d space. VOSviewer is capable of processing various kinds of data, including bibliometric specific data, and produces a mapping based on similarity. VOSviewer has shown to be useful for bibliometric visualisation for various types of data. Figure 1.1 shows a co-citation map of authors. In this map, authors working on similar topics are closer together and/or linked.
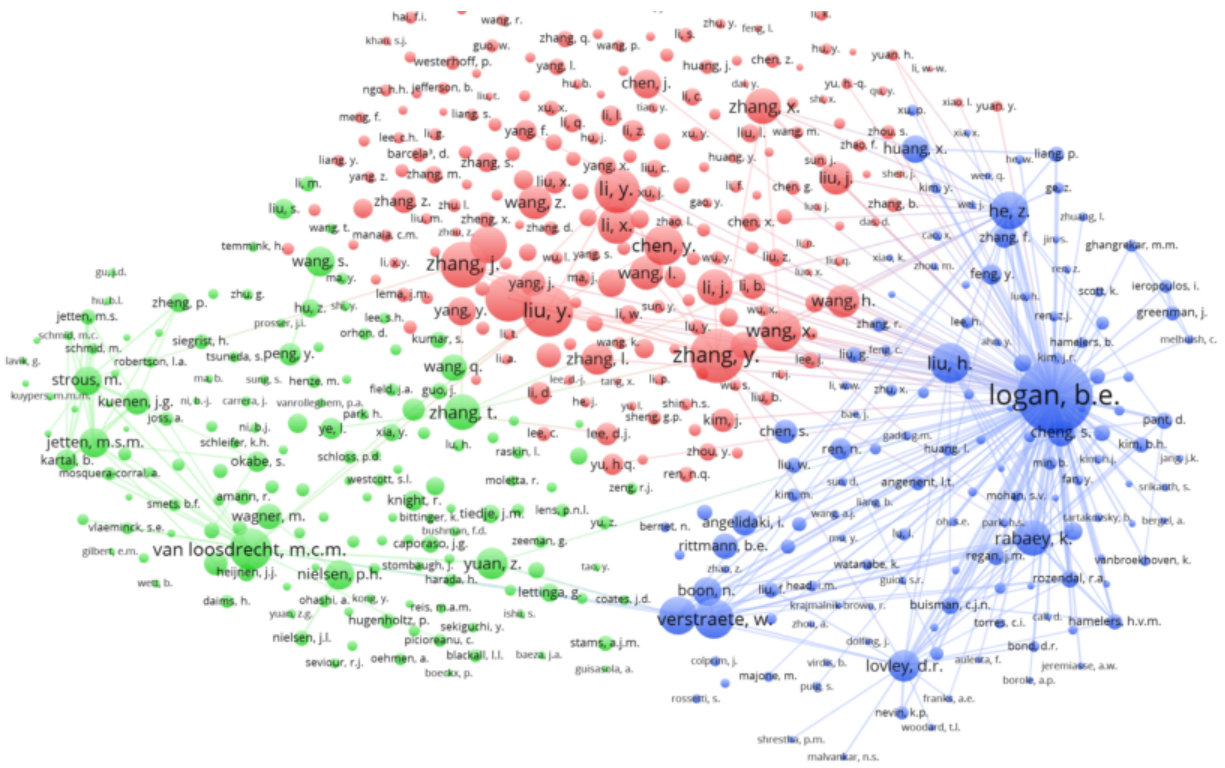

Figure 1.1: VOSviewer map based on co-citation [23].

VOSviewer allows a clear overview of complex maps and implements panning and zooming. VOSviewer implements detail on demand by providing data values on hovering over elements. In addition, filtering and

querying of elements is implemented. A limitation of VOSviewer is the fact that there is little room for drilling down. Users are restricted to the limited data that defines the mapping. It allows no insight into the actual underlying data, such as the publications themselves. A second limitation of VOSviewer is the difficulty of producing a custom map. It requires multiple steps of retrieving data from external sources and transforming them into the required format, before anything can be viewed. Although, there is a set of examples produced, which are available for viewing by simply downloading preprocessed data [24]. VOSviewer is available as either a desktop application, or launch-able from their website [25]. The latter option requires an installation of Java.

1.4.3   Heliocentric maps

In [26], mapping is done by what they refer to as *heliocentric maps*. These maps are similar to the maps that can be created using VOSviewer; however, one particular element to be analysed is placed in the centre of the map. Other elements "revolve" around this centre with their distance depending on the intensity of their relationship. This type of visualisation provides an individualized characteristic representation of the centre element. Heliocentric maps draw the focus of attention to the main object of interest. In addition, they remove clutter that can distract the user from the relevant pattern that is meant to be represented. This idea is used in the ACE visualisation.

1.4.4   AMiner

First introduced as Arnetminer in [14], AMiner is a project similar to Google Scholar and Mircosoft Academic. AMiner has a database of publication, author and affiliation records. Although not its main purpose, AMiner implements several small data visualisations on author profile pages. In Figure 1.2 these visualisations are shown. The "Ego network" (Fig. 1.2c) shows a similar layout as the earlier discussed "heliocentric map". In this implementation, there is detail on demand implemented by showing the related author information. However, this visualisation does not add much insight compared to an ordered list of related authors. ACE is similar to AMiner in the sense that it attempts to summarize author data by visualisation. AMiner can be accessed here [27].

1.4.5   Acemap

Acemap [28] is a project by researcher of the Shanghai Jiao Tong University. Acemap is based on the same data as AMiner, with a heavy focus on data visualisation. Several large bibliometric maps are linked on Acemap's home page. However, all of them are either difficult to navigate or hard to interpret. Based on the recent commits on GitHub [29], we assume it is still an ongoing project. However, little information is available in English. Like AMiner and MA, Acemap enables searching for different entities. For individual authors it is possible to view a full co-authorship ego network as shown in Figure 1.3. This is an application of network visualisation. It enables "stepping" through to other authors by double clicking, as well as detail on demand in the side bar. However, functionality is not intuitive and behaves unexpected. For example, the centre node is not shown when filtering by publication year.

1.4.6   Eigenfactor Project

The Eigenfactor project [31] aims to provide newly developed metrics for the evaluation of academic impact. In order to showcase their metrics, they have published a set of visualisations called "Well-formed" [32]. These visualisations are mostly interesting for their elegant design. They implement interactive components, like fish-eye distortion and detail on demand, in a way that is intuitive, useful and smooth. This results in graphs that are both visually pleasing and provide clear overview of the data. An example is shown in Figure 1.4.
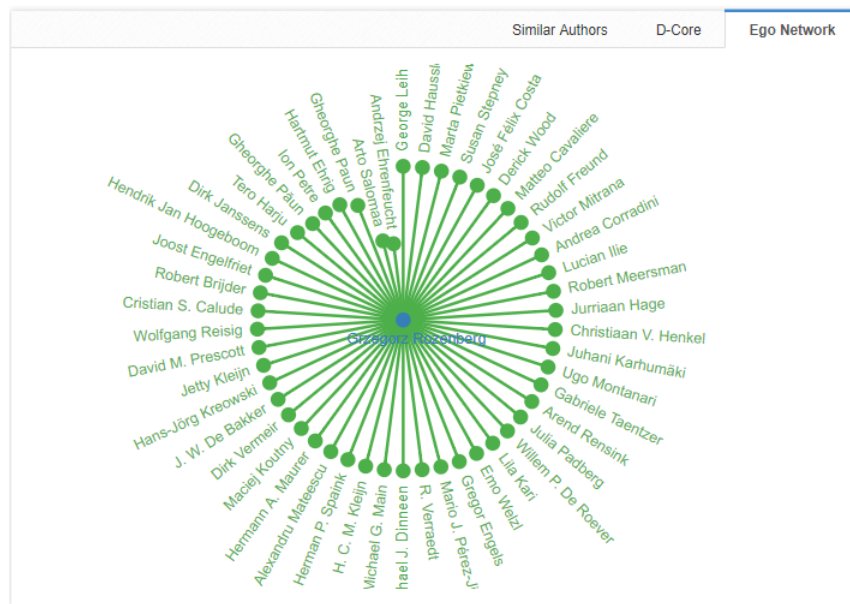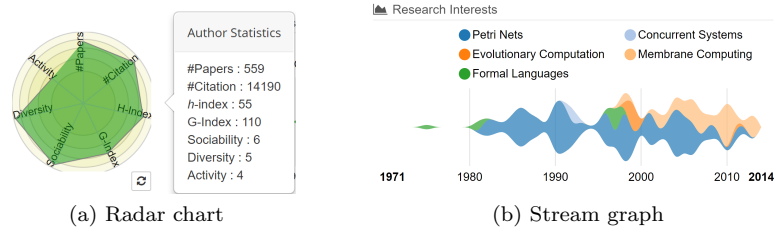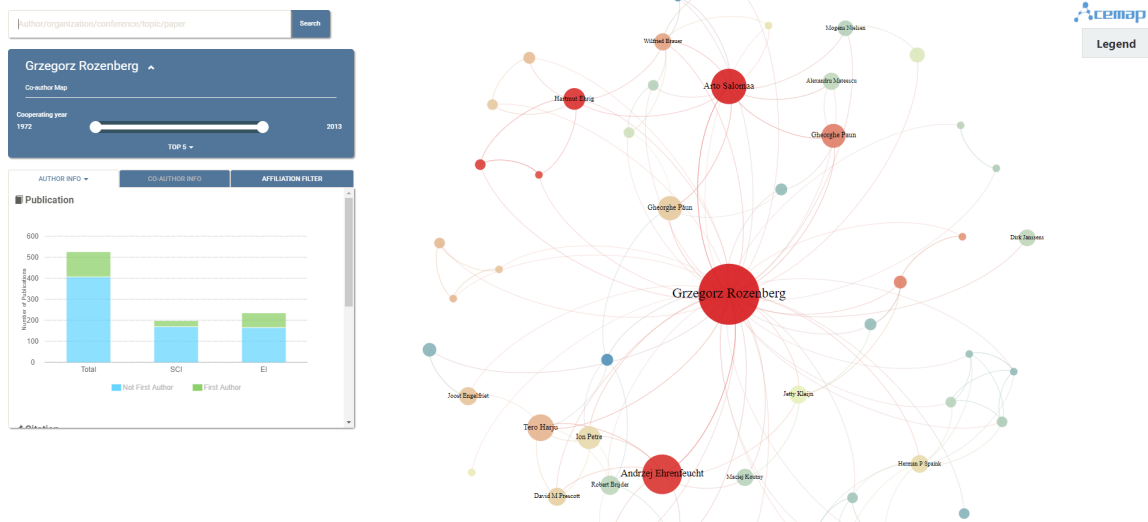
(a) Radar chart

(b) Stream graph

(c) Ego network

Figure 1.2: AMiner visualisations.

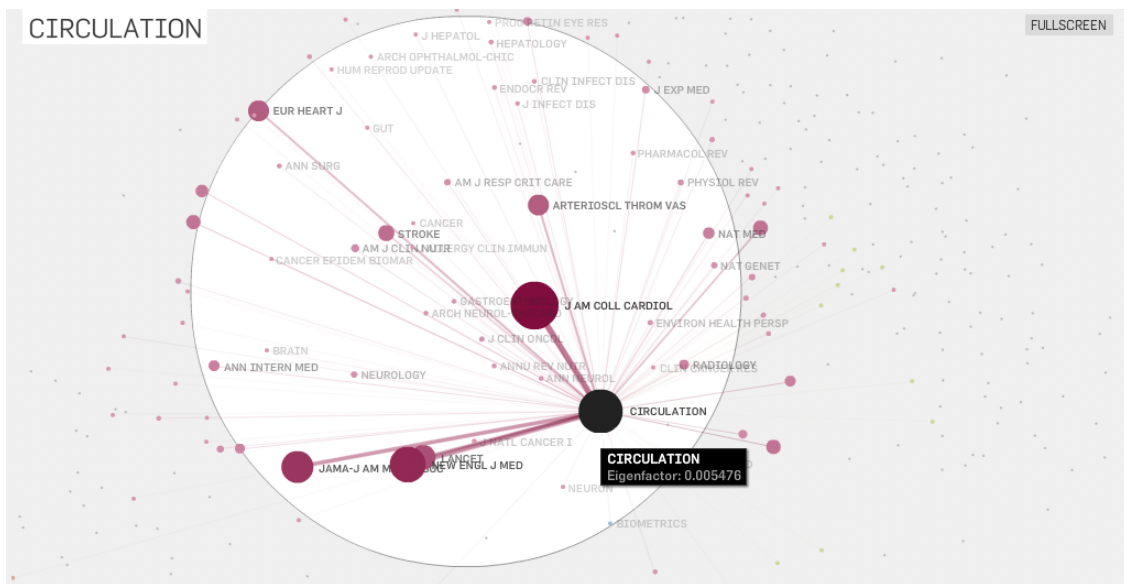Figure 1.3: Snapshot from Acemap's Co-author Map [30].

Figure 1.4: Snapshot from Eigenfactor's "Well-Formed" Map [32].

Chapter 2

Method

In this chapter, a detailed technical description of ACE is given. As mentioned in the introduction, ACE is a web-based visualisation tool that enables the creation of author centred co-authorship graphs. Additionally, ACE enables exploration of the underlying data, either within ACE or by linking to relevant external resources. First, a description is given of the data that ACE is based on and how this is retrieved. Next, the main data visualisation is addressed. This is followed by a description of the interactive features of ACE. The chapter concludes with a description of the efforts made in attempt to make ACE well accessible for users by optimizing for web search and social media.

2.1 Data

ACE is mainly based on the Microsoft Academic Graph (MAG) [12]. MAG is a weekly-updated heterogeneous graph consisting of publications, authors, institutions, journals, conferences and fields of study. MAG is part of the Microsoft Academic project and is freely accessible through Microsoft's "Cognitive Services Academic Knowledge API" [33].

As of December 2017, MAG contains records of around 160 million publications, 206 million authors and 19 000 institutions [34]. According to recent analysis, MAG is of similar quality as the Scopus and Web of Science databases [35]. MAG provides coverage of most research fields including economics and business, computer sciences and mathematics. However, similar to Scopus and Web of Science, MAG shows a slight bias against humanities, non-English publications and open-access publications [35]. During the development of ACE, we encountered two types of data inconsistencies that might influence results. Firstly, a significant portion of author entries lacks affiliation attributes. Secondly, for some authors there exist duplicate entries with the same or slightly different names and only one or a few linked publications.

The data ACE requires, is data about all combined publications by one or more authors and about all involved co-authors and their affiliations. In addition to the MAG affiliation entries, we matched affiliation names with entries in the Global Research Identifier Database (GRID) [36], adding country and institute homepage attributes as well as abbreviated names which are all not available in MAG.

2.2 Data retrieval

The process of Data retrieval is shown in Figure 2.1. Generally the process consists of two main sub-processes. First, the user searches for authors to include in the graph. During the search, author data is retrieved repeatedly for each character that is typed. Second, when the user has selected a set of authors, all data for the visualisation is retrieved. Since author search consists for a large part of user interaction, it is discussed

in Subsection 2.3.3. Below, the data retrieval for the visualisation is discussed further.
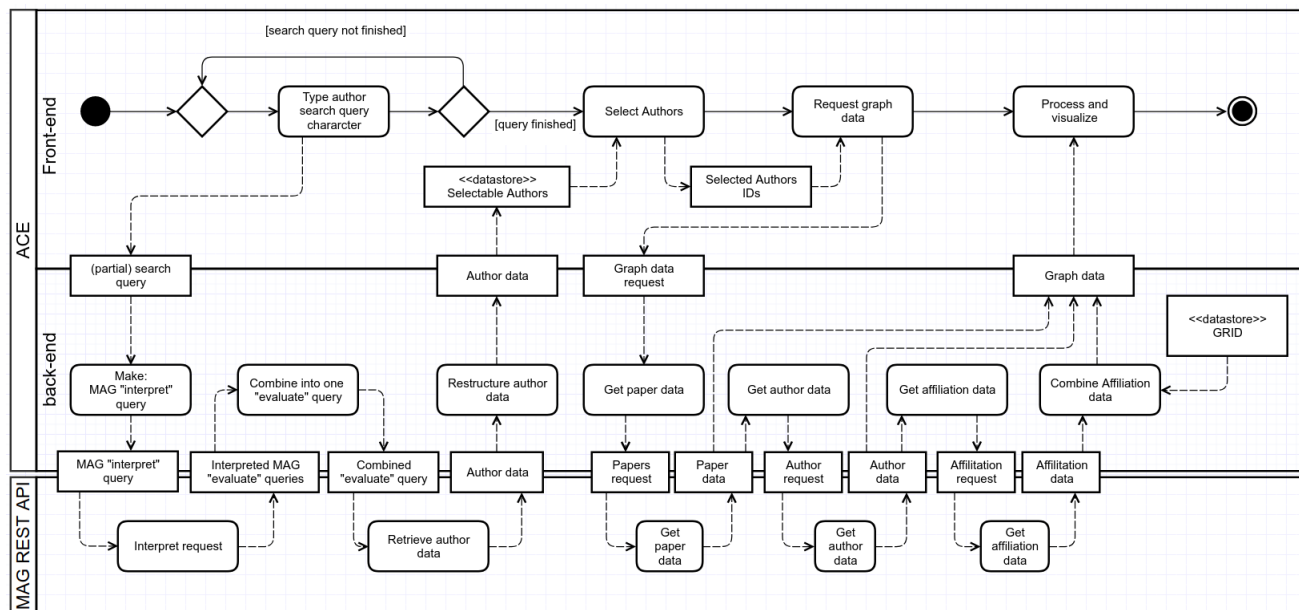

Figure 2.1: Process flow of data retrieval.

ACE uses the author, papers and affiliation entities which are connected in MAG following the schema shown in Figure 2.2. Retrieval of data is a multi-staged process. Data request is processed in three separate steps, each consisting of a request from the MAG API. First, based on a list of author ids, all paper entities are retrieved. Second, based on the set of papers, all co-author entities are retrieved. The third step is to extract affiliation data. This is done by grouping authors by affiliation and retrieving more affiliation specific data from MAG and adding the GRID data. Once all necessary data has been retrieved, it is ready to be visualised. This process is explained next.

2.3 Data Visualisation

ACE presents different kinds of data using a variety of methods. These can be split up in the relatively static graph and the interactive sidebar. First, we show how the graph layout is produced. Next, we show how data is encoded into colour, opacity and labels. This is followed by a description of the interactive components of ACE.

2.3.1 Force layout

The general layout of the graph places a single author or a group of authors at the centre. Placed around it, are all related research institutions. The positioning of nodes is the result of a force simulation, as described in Chapter 1, implemented in the D3 library for JavaScript [38]. D3's force simulation is able to implement a wide variety of forces, including the basic node repulsion and link attraction forces. D3 additionally allows for many more forces, including custom made forces in order to achieve the desired behaviour.

The initial node positions are with the centre node at the centre of the screen and all others in a circular pattern around the centre node. This initialisation ensures fast convergence of the simulation and ensures
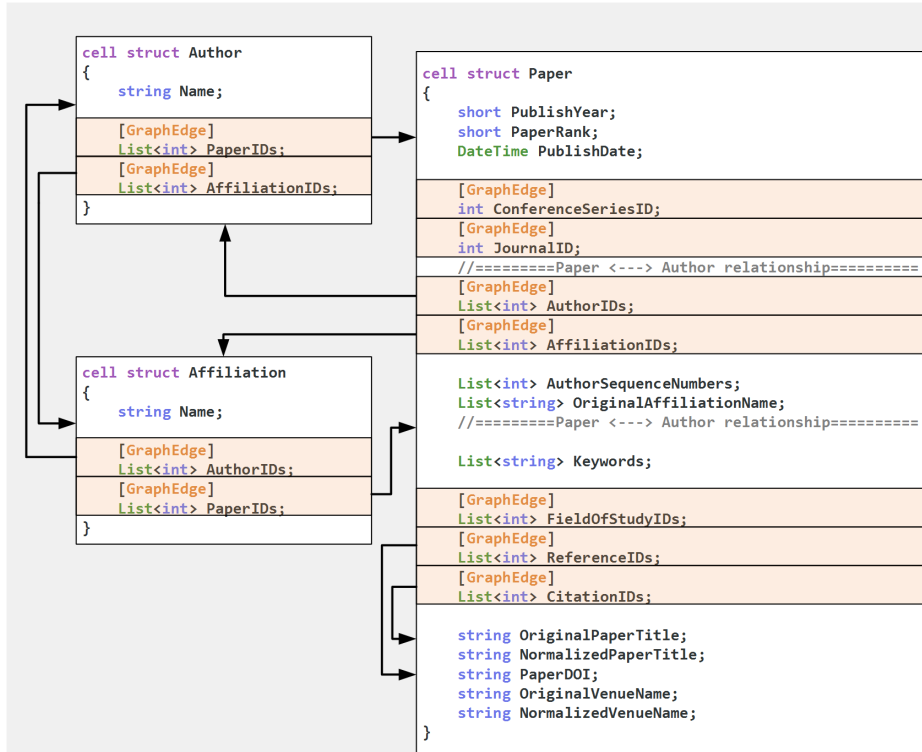
```
cell struct Author
{
    string Name;

    [GraphEdge]
    List<int> PaperIDs;
    [GraphEdge]
    List<int> AffiliationIDs;
}


cell struct Affiliation
{
    string Name;

    [GraphEdge]
    List<int> AuthorIDs;
    [GraphEdge]
    List<int> PaperIDs;
}
```

```
cell struct Paper
{
    short PublishYear;
    short PaperRank;
    DateTime PublishDate;

    [GraphEdge]
    int ConferenceSeriesID;
    [GraphEdge]
    int JournalID;
    //=========Paper <---> Author relationship==========
    [GraphEdge]
    List<int> AuthorIDs;
    [GraphEdge]
    List<int> AffiliationIDs;

    List<int> AuthorSequenceNumbers;
    List<string> OriginalAffiliationName;
    //=========Paper <---> Author relationship==========

    List<string> Keywords;

    [GraphEdge]
    List<int> FieldOfStudyIDs;
    [GraphEdge]
    List<int> ReferenceIDs;
    [GraphEdge]
    List<int> CitationIDs;

    string OriginalPaperTitle;
    string NormalizedPaperTitle;
    string PaperDOI;
    string OriginalVenueName;
    string NormalizedVenueName;
}
```

Figure 2.2: Part of the Microsoft Academic Graph Schema [37].

relatively consistent layout for the same data. All nodes repel each other equally in order to get an evenly distributed graph. The link attraction is dependent on the co-authorship relation strength. Since each node has a single link: only with the centre node. This results in a layout where node distance to the centre represents strength of the relationship. In addition to the node repulsion and link attraction, we implemented several additional forces to achieve an improved layout. A collision force was added to prevent nodes from overlapping. A rectangular box constraint was added so that nodes cannot escape from the view.

Achieving a good layout also requires tweaking of some variables which determine force strength. In the development of ACE, we tried to optimize these variables in order to always get a visually attractive and useful graph. However, depending on the data, the user might want to adjust these layout parameters. Therefore, these variables are made user controllable in the setting tab. Examples of the force layout at work can be seen in Figures A.1 through A.4 in Appendix A.

An option to group institutes by country was implemented to bring more structure in the layout of the graph. This introduces an additional force. This force adds an attractive force towards the largest of the nodes of the same country. As a result, nodes form the same country end up in clusters. The clustering of nodes facilitates exploration of the graph, especially for very dense graphs. An example can be seen in Figure 3.3.

2.3.2 Visual features

In the main graph, several visual features are meant to guide the user towards the most relevant information. These visual features are shown in Table 2.1. The relations between the visual features and the attributes they represent are chosen to improve the ability to distinguish important differences. For example, the distances to the centre are inversely proportional to the logarithm of the joint paper count. This results in a graph that is

in most cases evenly distributed and still clearly represents the relationship strengths. When the nodes are grouped by country, an additional colour accent is added to each node in order to facilitate the distinction of different countries.

Table 2.1: Visual features in the ACE graph.

Feature	Representing	relation
Node size	Affiliation total citation count	$y = x^{0.17}$
Node opacity	Joint paper count	$y = 1 - 1/\ln(x + 2)$
Distance to centre	Joint paper count	$y = C/\ln(x + 2)$

Node labels consist of either the abbreviated name of the institute or the shortest available name where the word "University" is replaced with the letter "U". This minimizes redundant label text, which might clutter the graph. The number of labelled nodes is restricted to a user adjustable number defaulting at 20 nodes in order to provide a clutter-free view. Node labelling is prioritized by the size of the node, labelling bigger nodes first which is equivalent to labelling nodes with the total citation count first. The label restriction is applied to all nodes in view. As a result, zooming in on a section of the graph causes smaller nodes to be labelled, enabling better inspection of the graph.

2.3.3 Interaction

Before any graph is shown, the first form of interaction the user will have with ACE is the search for authors. Author search is done in the search window which is shown in Figure 2.3. On the left, a user searches for authors. For each character that is typed, a request to MAG is done (indirecly through the ACE back-end) using the "interpret" method. This method enables auto-completion of search queries, resulting in a smooth search experience. The returned author entities are added to the multi-select list below the search box. Author names are shown with the author's affiliation and the number of times they were cited. This enables the user to distinguish the appropriate entry from possible duplicates. Upon selection, the author is added to the list on the right. This list represents the group that the graph will be about. At the top of the right hand side list, the group can be given a name that will be used in the visualisation. When the "show graph" button is clicked the required data will be retrieved and the graph will be produced.

In the discussion of node labels, we already hinted at interaction by panning and zooming. This allows to focus attention to a part of the data. Hovering over a node will show an extended label displaying the full name and country of the institute as well as the number of (joint) papers and (co-)authors. This can help the user to decide which nodes are interesting to explore further.

Selecting a node (by clicking) will show details regarding the selected node in the sidebar. Figure 2.4 shows an example of the sidebar. The details shown, include the number of authors, the number of publications and the number of times cited. The sidebar also shows the most common fields of study of the set of publications. This allows for a quick insight in the subjects that have been collaborated on. The rest of the sidebar shows the list of authors represented by the node. For each author, a list of co-authored publications can be displayed in the sidebar. Since ACE is fully centred on co-authorship relations, it is useful to get insight in the actual publications that were co-authored.

Wherever possible, links to external resources are added in the sidebar. These include links to institute websites and links to institute, author and paper entries in MA. This ensures that the exploration experience is not limited by the boundaries of ACE. Linked pages are themselves also highly linked, enabling users to dig as deep as they desire. In addition to external links, each author or group of authors displayed in ACE is accompanied by a red "ACE" link. The ACE-links will show the search window and automatically add the
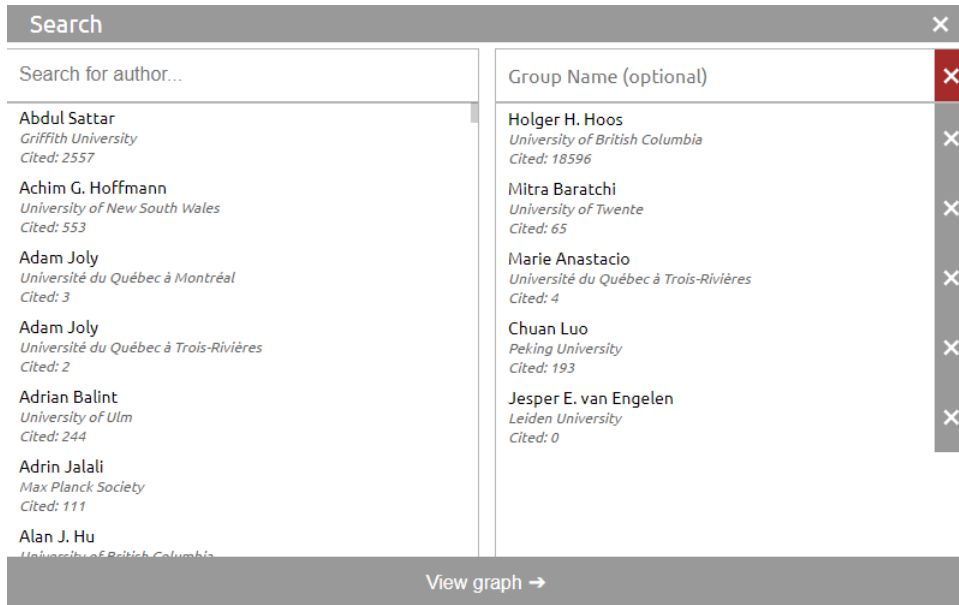
Figure 2.3: ACE search window.

Figure 2.4: ACE info bar with publication list.

desired authors, with the possibility to refine the search. This allows stepping though the collaboration network.

An important part of interaction is filtering the data that is presented. ACE implements filtering by year, enabling users to set a range of the years to include. Viewing only the most recent publications can for example show more relevant information. By comparing data from different time periods, temporal patterns can be observed. This can for example show shifts in research focus.

2.4 Ease of access

ACE is designed to be easily usable and also easily findable. ACE is equipped with the right meta data to act as a *rich object*. This meta data is, for instance, used by Facebook to represent links to ACE optimally. Figure 2.5 shows how links to ACE will appear on Facebook and other social media. The links include the recognisable colours and a brief description of what ACE does. Since social media is being adapted by academia [39], ACE implements easy sharing buttons and the ability to take a quick snapshot of your view. The data represented in ACE concerns people and relations between people; therefore, it is convenient to be able to share insights found using ACE with peers.
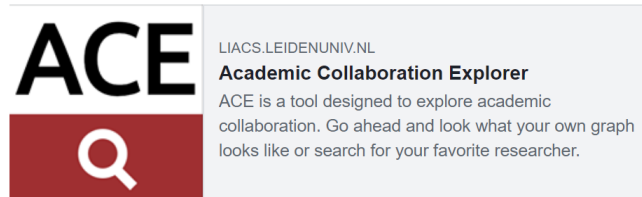
Figure 2.5: Link to ACE as it will appear on Facebook [40].

Chapter 3

ACE use cases

ACE can be used in a variety of situations. In this chapter, we discuss several use cases. First, looking up individual authors. Viewing their collaborative relations and a summary of research output. Second, we discuss comparing authors. ACE can for instance show differences in collaboration intensity or research focus. Third, we look how ACE can be used for promotion. Promoting a research group by showing ACE emphasises collaboration intensity. Forth, we show how ACE could be used for academic management.

3.1 Author lookup

The simplest use case is looking up a single author. On the first glance, one gets an overview of the most important institutes the author has collaborated with. The density of the graph gives an indication of the general intensity of collaboration with different institutes. In the sidebar, one can get a quick view of the research interests of the author, how many papers have been published and total citation counts giving indication of impact. All this information can be found in the initial view of the graph. Once one starts exploring, much more questions can be answered.

One can look whether the author of interest has ties with a specific university, by checking if the university is represented as a node in the graph. When the university is present, it can be selected. This will show the people involved in this relationship. Finding related authors can be useful for finding common acquainted colleagues, which in turn can be useful for making new contacts.

3.2 Author comparison

ACE can be used to compare authors based on publication output, impact and connectedness with the scientific community. Although metrics like h-index can be informative, they have their limitations. One could open two browser windows with ACE side-by-side and search for two researcher to be compared. One can select the same time range for both researchers, and compare research output. Number of papers and citation count provide useful metrics, especially when both researchers are active in the same fields. One can easily see if there are slight differences in research interest in the sidebar. The structure of the graph can indicate differences in intensity of collaboration. An example of side-by-side author comparison is shown in Figure 3.1.

3.3 Group promotion

An important feature of ACE, is the ability to show data based on a group of authors instead of only single authors. This way, it is possible to view the relations of a research group or a whole institute. This feature can
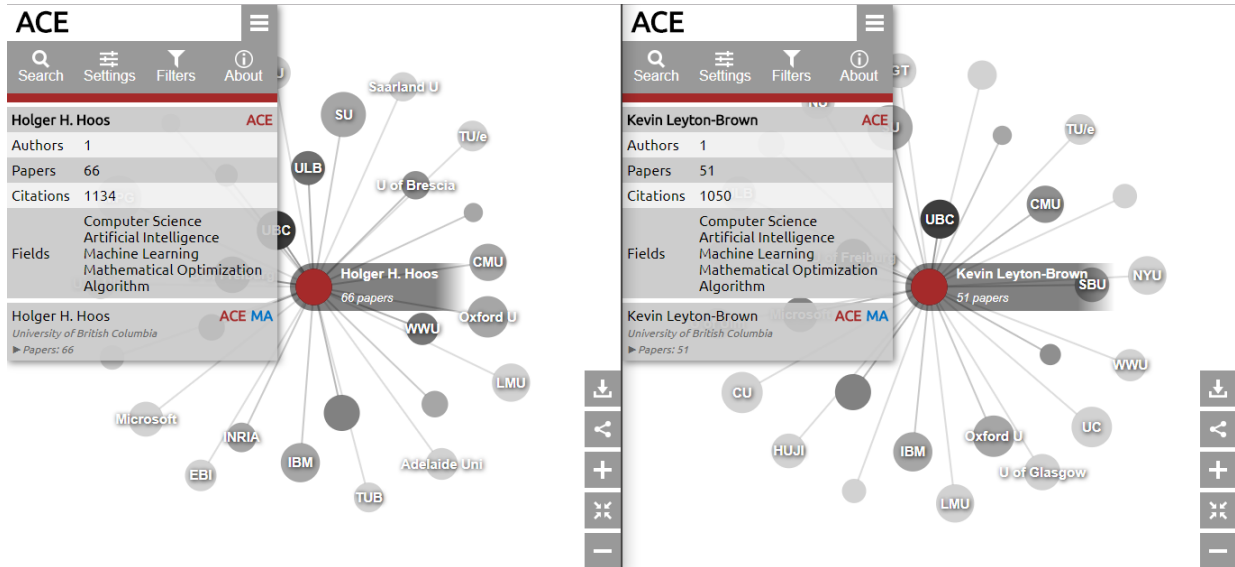
Figure 3.1: Side-by-side comparison of two authors.

be used to communicate the ties with the academic community to the outside world. Using the sharing options, one can link to ACE on the web page of a research group in order to display their collaboration with colleagues. Or one could take a snapshot of the ACE graph for the same purpose. Figure 3.2 shows the investigation of relations of a research group of five authors. Figure 3.3 shows relations of a full institute of almost 50 authors.

The advantage of sharing information this way, is to encourage exploration of the actual work instead of just listing several output statistics. Readers of the homepage of a research group might be interested in reading the most recently published papers. It gives a better understanding of the research being done by the group. For the research group, it provides a way to reach more people. An additional advantage of promotion using ACE requires no maintenance, since the data is kept up-to-date by Microsoft.

3.4 Academic management

As mentioned earlier, ACE can be used to quickly get an overview of the work of individual authors. Besides for academic purposes, this overview can be valuable for assessment of candidates for positions or funding. The author-group feature can be of value in management context as well. Management of larger groups of researchers can benefit from a clear and up-to-date view of external collaborations. Strategy of an institution might motivate the increase of collaboration with specific colleague institutes. For example, institutes that have expertise in a particular field. In these situations, it can be valuable to leverage already existing relations by strengthening the existing relations and possibly let them lead to increased collaboration. ACE could potentially help in finding these already existing relations.
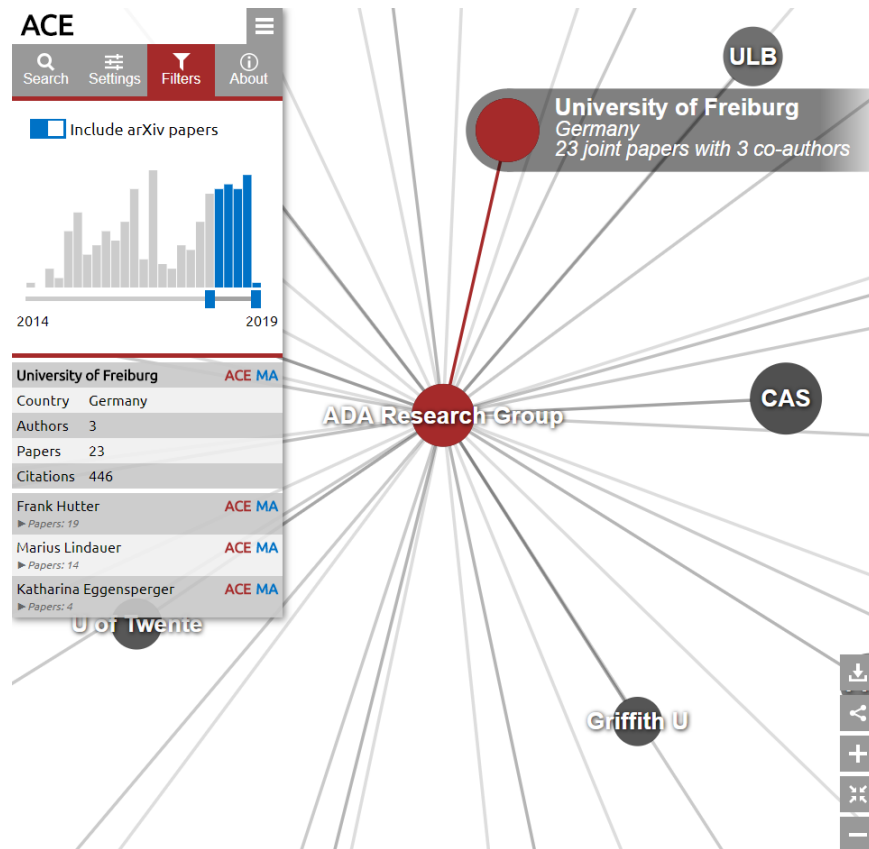
Figure 3.2: ACE showing details about the relation between the ADA research group and co-authors affiliated with the University of Freiburg over the last five years.
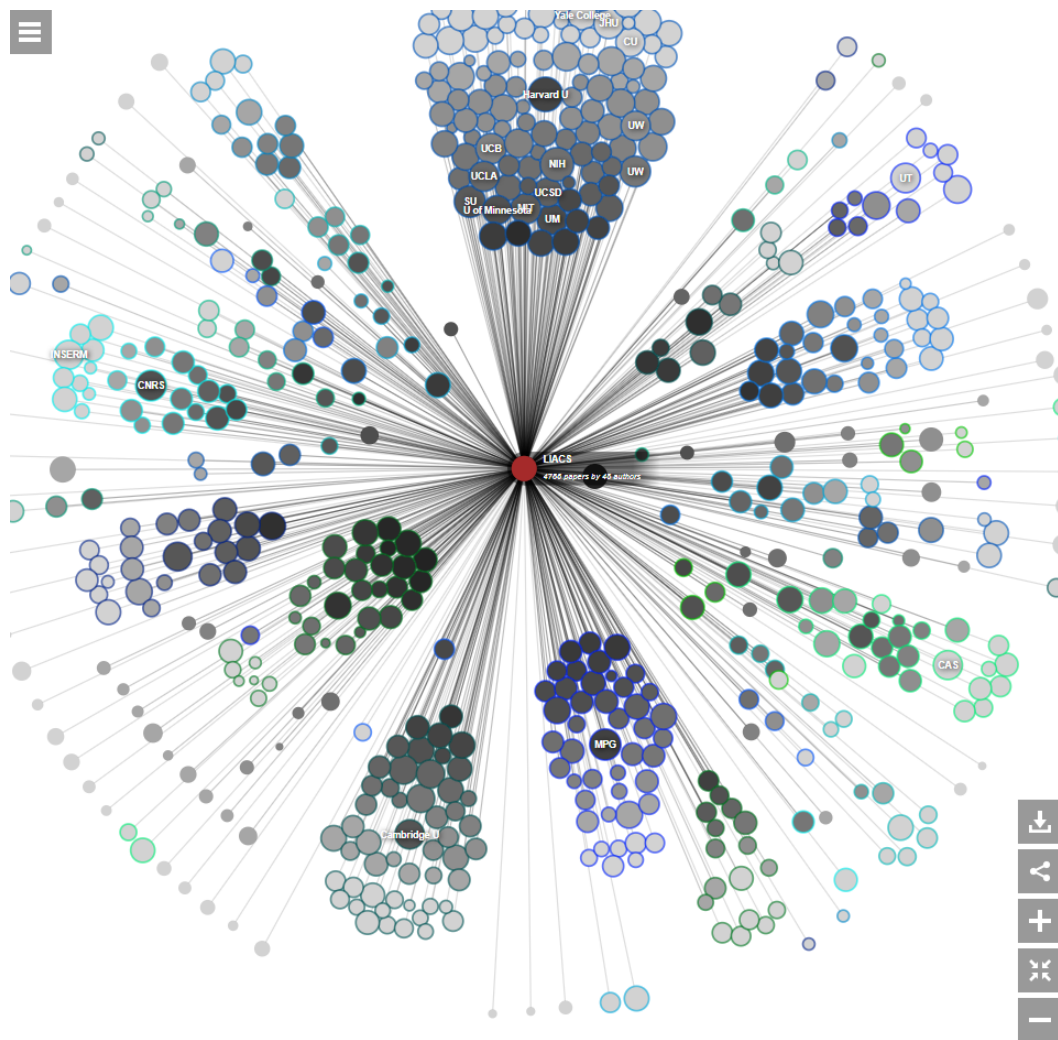
Figure 3.3: ACE overview of all institutes, grouped by country, related to researchers at the Leiden Institute for Advanced Computer Science.

Chapter 4

Conclusions and Future Work

In this thesis, we presented ACE: A web-based tool for visualising author-centred collaboration data. Bibliometric data is inherently complicated, since it consists of different entities that have complex relations that can change over time. In this project, we aimed to present data piecewise by focusing on a set of authors and showing their first order relations, preventing information overload. As a result, ACE offers information that is relevant and useful for researchers in any discipline. One advantage of ACE over alternative visualisation tools, is the fact that data retrieval is integrated: no separate data retrieval or processing is required. Compared to other tools for exploring academic output, ACE has the advantage of allowing to view a group of authors as a whole, instead of only individual authors. In order to make ACE accessible to a broad audience, much attention went into making ACE easy to use, easy to find and easy to share with others.

However, ACE is still an ongoing project: there are multiple opportunities for improvement. The ACE graph should provide as much meaningful data as possible without getting cluttered. This can be done by improving the ability to filter. Furthermore, the layout could be improved to better represent relations between related institutes, comparable to the Acemap's co-author map. The publication list can be improved by enabling sorting by citations in order to find more relevant work. This could be inspired by the tabular representation of Publish or Perish.

In the field of bibliometic visualisation as a whole, there are still challenges for the near future. At this point, there are multiple platforms being developed which help researchers navigate information that is published and shared with the academic community. However, the whole infrastructure of academic communication still depends on many different channels. Some of these channels are quite dated such as printed journals, mailing lists and RSS feeds. Others are just not optimized for academia such as social media. It seems like there is still room for a system that puts all relevant information in one place, and makes it accessible in a way that is intuitive and designed for day-to-day use.

Acknowledgements

I would like to thank Prof. Hoos for sharing his thoughts and his energetic involvement in this project. I especially want to thank Jesper van Engelen for the time and effort he has put into supporting my work and providing critical feedback.

Bibliography

[1] A. Perianes-Rodrguez, Z. Chinchilla-Rodrguez, B. Vargas-Quesada, C. O. Gmez, and F. Moya-Anegn, "Synthetic hybrid indicators based on scientific collaboration to quantify and evaluate individual research results," *Journal of Informetrics*, vol. 3, no. 2, pp. 91 – 101, 2009.

[2] B. C. Björk, A. Roos, and M. Lauri, "Scientific journal publishing: Yearly volume and open access availability.," *Information Research*, vol. 14, no. 1, 2009.

[3] M. A. Koseoglu, "Growth and structure of authorship and co-authorship network in the strategic management realm: Evidence from the strategic management journal," *BRQ Business Research Quarterly*, vol. 19, no. 3, pp. 153 – 170, 2016.

[4] E. Garfield and I. H. Sher, "New factors in the evaluation of scientific literature through citation indexing," *American Documentation*, vol. 14, no. 3, pp. 195–201, 1963.

[5] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences*, vol. 102, no. 46, pp. 16569–16572, 2005.

[6] L. Leydesdorff and S. Milojevic, "Scientometrics," *CoRR*, vol. abs/1208.4566, 2012.

[7] K. Börner, C. Chen, and K. W. Boyack, "Visualizing knowledge domains," *The Artist and Journal of Home Culture*, vol. 37, no. 1, pp. 179–255, 2005.

[8] J. R. Cole and S. Cole, *Social stratification in science*. Chicago: University of Chicago Press, 1973.

[9] A. Abbasi, K. S. K. Chung, and L. Hossain, "Egocentric analysis of co-authorship network structure, position and performance," *Information Processing & Management*, vol. 48, no. 4, pp. 671 – 679, 2012.

[10] E. Y. Li, C. H. Liao, and H. R. Yen, "Co-authorship networks and research impact: A social capital perspective," *Research Policy*, vol. 42, no. 9, pp. 1515 – 1530, 2013.

[11] A. Ebadi and A. Schiffauerova, "How to become an important player in scientific collaboration networks?," *Journal of Informetrics*, vol. 9, no. 4, pp. 809 – 825, 2015.

[12] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang, "An overview of Microsoft Academic Service (MAS) and applications," in *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, (Florence, Italy), pp. 243–246, ACM, 2015.

[13] "Google scholar." https://scholar.google.com/. Accessed: 19-7-2018.

[14] J. Li, J. Tang, J. Zhang, Q. Luo, Y. Liu, and M. Hong, "Arnetminer: expertise oriented search using social networks," *Frontiers of Computer Science in China*, vol. 2, pp. 94–105, Mar 2008.

[15] "Scopus." https://www.scopus.com/. Accessed: 19-7-2018.

[16] "Web of science." https://clarivate.com/products/web-of-science/. Accessed: 19-7-2018.

[17] "List of academic databases and search engines." https://en.wikipedia.org/wiki/List_of_academic_databases_and_search_engines. Accessed: 19-7-2018.

[18] M. Cobo, A. López-Herrera, E. Herrera-Viedma, and F. Herrera, "Science mapping software tools: Review, analysis, and cooperative study among tools," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 7, pp. 1382–1402, 2011.

[19] S. G. Kobourov, "Spring embedders and force directed graph drawing algorithms," *CoRR*, vol. abs/1201.3011, 2012.

[20] P. Pirolli and S. Card, "Information foraging," *Psychological review*, vol. 106, no. 4, p. 643, 1999.

[21] "Harzing - publish or perish." https://harzing.com/resources/publish-or-perish. Accessed: 28-6-2018.

[22] N. J. V. Eck and L. Waltman, "Software survey: Vosviewer, a computer program for bibliometric mapping," *Scientometrics*, vol. 84, no. 2, pp. 523–538, 2010.

[23] "Vos viewer author landscape." http://aida.tudelft.nl/cases/c14. Accessed: 29-6-2018.

[24] "Automatic identification of research trends - aida." http://aida.tudelft.nl/toolbox/aida-booklet. Accessed: 30-6-2018.

[25] "Vosviewer download." http://www.vosviewer.com/download. Accessed: 30-6-2018.

[26] B. Vargas-Quesadaa, F. de Moya-Anegóna, Z. Chinchilla-Rodrígueza, and A. Gonzalez-Molinab, "Domain analysis by means of the visualization of maps of vast scientific domains," in *I Internacional Conference on Multidisciplinary Information Sciences & Technologies*, (Mérida, Spain), 2006.

[27] "Aminer." https://aminer.org/. Accessed: 30-6-2018.

[28] "Acemap." http://acemap.sjtu.edu.cn/. Accessed: 30-6-2018.

[29] "Acemap github repository." https://github.com/cristianoBY/Acemap-Academic-Article-Analysis-Project. Accessed: 4-8-2018.

[30] "Acemap co-author map." http://acemap.sjtu.edu.cn/coauthormap?AuthorID=7D084C28. Accessed: 30-6-2018.

[31] "Eigenfactor project." http://eigenfactor.org/about.php. Accessed: 30-6-2018.

[32] "Eigenfactor project." http://eigenfactor.org/projects/well-formed/index.html. Accessed: 30-6-2018.

[33] "Microsoft cognitive services labs project academic knowledge." https://labs.cognitive.microsoft.com/en-us/project-academic-knowledge. Accessed: 28-6-2018.

[34] "Microsoft academic graph update." https://www.microsoft.com/en-us/research/project/academic/articles/december-2017-graph-update/. Accessed: 29-6-2018.

[35] S. E. Hug and M. P. Brändle, "The coverage of microsoft academic: analyzing the publication output of a university," *Scientometrics*, vol. 113, pp. 1551–1571, Dec 2017.

[36] "Global research identifier database." https://www.grid.ac/. Accessed: 28-6-2018.

[37] "Academic knowledge documentation." https://docs.microsoft.com/en-us/azure/cognitive-services/academic-knowledge/graphsearchmethod. Accessed: 28-6-2018.

[38] M. Bostock, "D3js - data-driven documents." https://d3js.org/, 2017. Accessed: 29-6-2018.

[39] A. Gruzd, K. Staves, and A. Wilk, "Connected scholars: Examining the role of social media in research practices of faculty using the utaut model," *Computers in Human Behavior*, vol. 28, no. 6, pp. 2340 – 2350, 2012.

[40] "Ace in facebook sharing debugger." `https://developers.facebook.com/tools/debug/sharing/?q=http%3A%2F%2Fliacs.leidenuniv.nl%2F~s1309773%2FAcademic%2F`. Accessed: 29-6-2018.
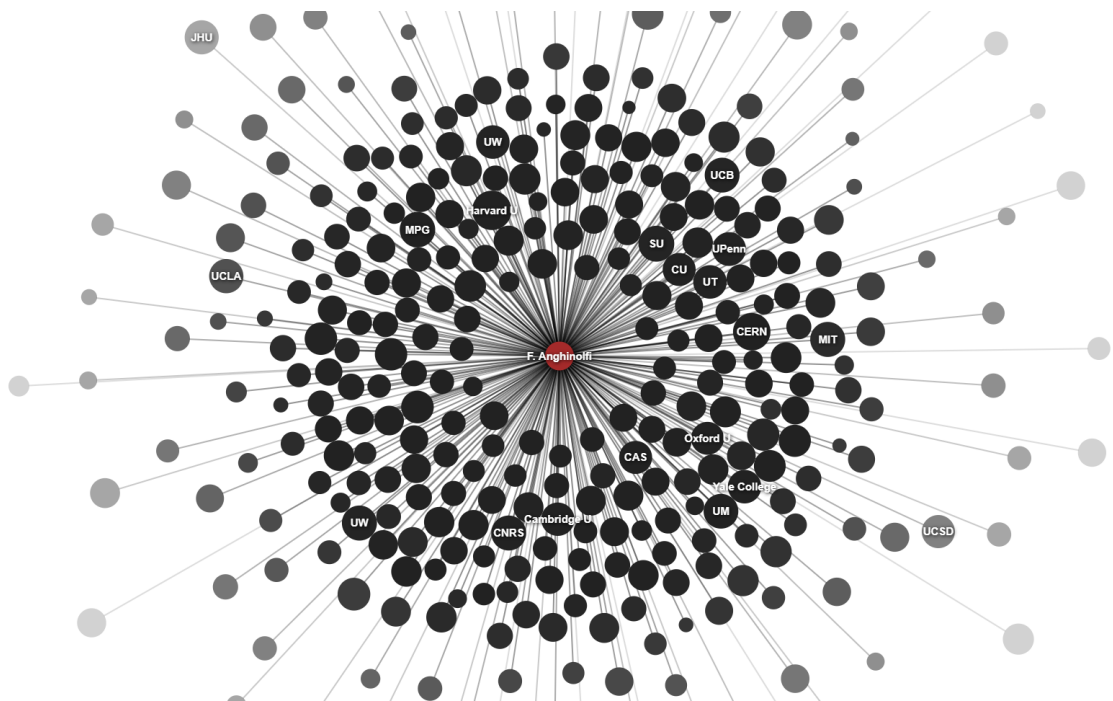
Appendix A

Demo Graphs

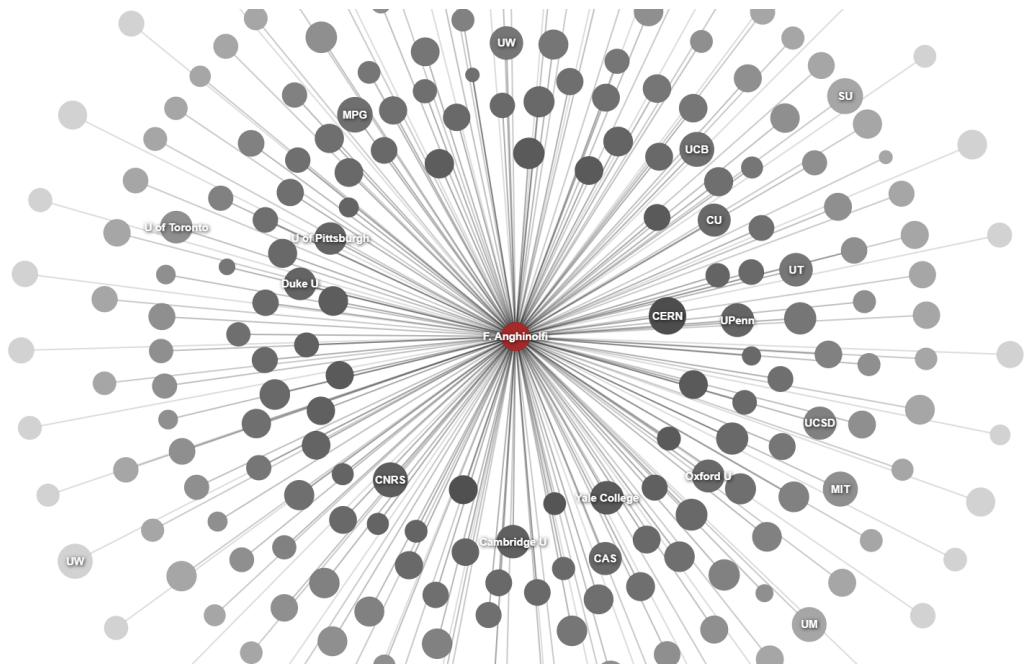Figure A.1: ACE graph: High collaboration in research in particle physics.

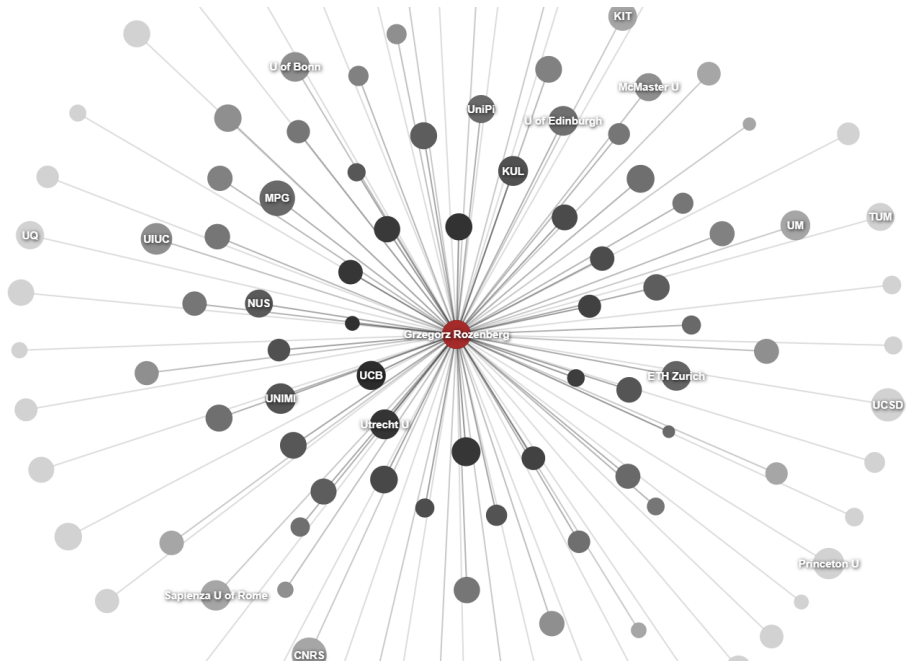Figure A.2: ACE graph: Data from Fig. A.1 between 2007 and 2010.
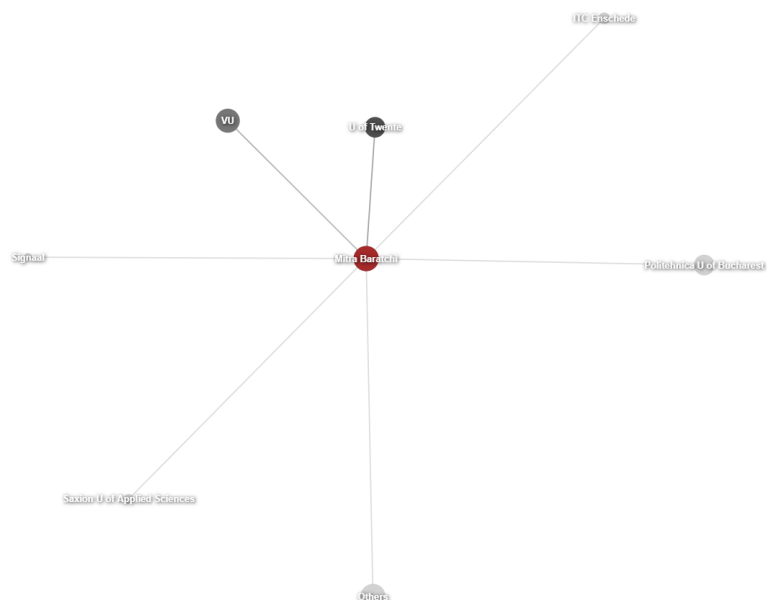

Figure A.3: ACE graph: A full career in computer science.

Figure A.4: ACE graph: A young researcher in computer science.