Technological University Dublin
ARROW@TU Dublin

Dissertations

School of Computing

2018

Visualization of Co-authorshipin DIT Arrow

Dan Xu
Technological University Dublin

Follow this and additional works at: https://arrow.tudublin.ie/scschcomdis

 Part of the Computer Sciences Commons

Recommended Citation

Xu, Dan (2018). *Visualization of co-authorship in DIT Arrow.* Masters dissertation, DIT, 2018.

This Dissertation is brought to you for free and open access by the School of Computing at ARROW@TU Dublin. It has been accepted for inclusion in Dissertations by an authorized administrator of ARROW@TU Dublin. For more information, please contact yvonne.desmond@tudublin.ie, arrow.admin@tudublin.ie, brian.widdis@tudublin.ie.

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License

Visualization of Co-authorship in DIT Arrow

DAN XU

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

January 2018

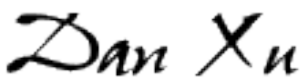
Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Stream), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institutes guidelines for ethics in research.

Signed:



Date: 03/01/2018

Abstract

With the popularization of information technology and the unprecedented development of online reading, the management and service of the library are facing severe challenges; the traditional library operation mode has been challenging to optimize the service. At the same time, there is also a fatal impact on library collection and systematic management, however, with the development of visualization techniques in management and service, the library can alleviate the effect of the current network information basically, which achieves the intellectual development of library field.

This study empirically provides the evidence to indicate that the force directed layout has the statistically significant performance than the radial layout for visualization of co-authorship in DIT Arrow repository based on the results of surveys.

Keywords: Co-authorship, Data Visualization, Layout Algorithms

Acknowledgments

I would like to express my special thanks of gratitude to my supervisor Dr.John McAuley, who with extraordinary patience and consistent encouragement, gave me great help by providing me with necessary materials, advice of great value and inspiration of new ideas. Without his strong support, this thesis could not been the present form.

My heartfelt thanks also go to Prof.Sarah Jane Delany, who gave me the opportunity and idea to do this wonderful project, and many constructive suggestions at the beginning.

Last but not least, my thanks would go to my beloved family for their loving supports and great confidence in me all through these years. I also owe my sincere gratitude to my boyfriend Kaiqiang Huang, who gave me the support and encouragement for helping me work out my problems during the difficult course of the thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Dublin Institute of Technology (DIT) has a digital collection of research publications produced by researchers in DIT called Arrow (Archiving Research Resources on the Web), which contains the institutes research materials such as articles, journals, and reports at arrow.dit.ie. The information in Arrow and other institutional repositories can be represented in a graph or network structure with the nodes representing the researchers and researchers who have co-authored work being connected. However, Arrow does not provide any graphics visual representation of this information. Visualization is the process of converting data, information, and knowledge into visual forms of representation. Visualization technology provides an interface between the two most potent information processing systems of human and computer. Moreover, academic papers not only provide scientists with a way to acquire professional achievements and knowledge but also provide efficient means for researchers to obtain scientific resources and establish the theoretical communication network between scientific circles and industrial elites. Therefore, the co-authorship is a critical standard that can evaluate the academic influence of the authors on the particular network.

Furthermore, the layout algorithm is the core and foundation of the visualization of the graph, which has an essential influence on the real-time processing of the graph data. Besides, the layout algorithm, which integrates the aesthetic standards, can also improve the ability and readability of the map to a certain extent. Force-Directed Layout algorithms are graph drawing algorithms based only on the information contained within the structure of the graph itself rather than relying on the contextual information. The most straightforward Force-Directed algorithm uses repulsive forces between nodes and attractive forces between adjacent nodes. Radial visualization is a wide range of fields, and applications are targeted to different subject areas. However, this does not mean that radial visualization is the correct solution for every visual need.

1.2 Research Project

Today, there are many research bibliography sites on the web such as IEEE Xplore and Institutional Repositories such as Arrow which provide detail information about specific authors, papers or journals, and it is useful for a research survey. However, when a researcher is getting into the unfamiliar field of research, grasping an overview of the research field, such as bibliographic network, is essential (Kurosawa & Takama, 2011). Because such a network structure is usually vast and complicated, to solve this problem, some visualization techniques of social network primarily for co-authorship networks are commonly used for the analysis.

The study aims to investigate the comparison of data visualization between the force directed layout and radial layout used in DIT Arrow repository. Then, the research question is defined as follows.

Can the co-authorship visualization using force-directed approach provide the more readable representation to the viewers in Arrow?

In order to answer the defined research question, the objectives of the research are to determine whether the force directed layout can yield the better visualization than radial layout in DIT Arrow repository. Few designed experiments will be conducted to generate the suitable and readable visualization, and then, the statistical test will be performed to prove the hypothesis as well.

1.3 Research Methodologies

The type of this research is primary research because the dataset will be generated by this study. And also, it is a combination of qualitative and quantitative research, which contains lab experiments, formal methods, mathematical modeling, and surveys. The form and reasoning methods are empirical research and deductive research.

1.4 Scope and Limitations

This research focuses on the area of web scraping and data visualization in the digital library. In the ideal case, web scraping would not be essential, and each site will provide an API to share data in a structured format. In fact, some websites do provide API, but they are usually limited by the availability of data and the frequency of access. Also, the central task for web development is to maintain the front end of the interface than the back end API. Furthermore, there have limited research and study for the comparison of force-directed layout and radial layout.

1.5 Document Outline

The structure of the document is outlined as follows.

- **Chapter 2** (Review of existing literature) reviews the existing research and study related to the area of data visualization in the digital library. And also, the different layouts of visualization are introduced. The evaluation of visualization is discussed and review as well. Furthermore, the technology of web scraping and the software of visualization are presented in general.
- **Chapter 3** (Experiment design and methodology) provides the process design of creating the dataset for experiments. Also, the methods of data preparation and selecting visualization tool are also presented and discussed.
- **Chapter 4** (Implementation and results) Gives the details of implementation of generating the various visualization for DIT Arrow. And the general analysis of the network is discussed.
- **Chapter 5** (Evaluation and analysis) provides the details of survey design and implementation, and also provides the relevant statistical test for giving the robust results.
- **Chapter 6** (Conclusion) makes the conclusions for this study, including the aspects of project overview, problem definition, experiment results, contributions, and future work.

Chapter 2

Review of Existing Literature

2.1 Introduction

This chapter will focus on the literature review about the research's underlying domain. The first section generally introduces the background of digital library evaluation which is the macro concept of library. Then, a briefly introduction and history of data visualization are presented. Exploration on visualization technology in the field of library will be introduced. Then, the literature review of social network analysis in co-authorship will be presented. Then, the most important part in this chapter is the layout algorithms, in this section, both force directed layout and radial layout will be described in details of history, development, research domain and existing research/paper work. Then, visualization tools and evaluation will be clearly presented. Finally, the scraping technology will be introduced briefly.

2.2 Digital Library Evaluation

Library evaluation is necessarily a process of value judgment. It refers to the process of making an objective analysis by specific evaluation criteria and evaluation methods. In the early twentieth Century, the bud of library evaluation appeared, but systematic and conscious library evaluation began in the middle and American countries in the middle of the Twentieth Century. Overall, after half a century of development, the performance evaluation of library experience from theoretical research to practical application, from individual academic point of view to the development of international standards, based on readers return value of library research and development to the development and implementation of the new evaluation model to test the library overall benefit, from simple service output expansion measurement to the library-wide social value analysis confirmed the transformation process. Performance evaluation has been widely used in various types of libraries in many countries and has accumulated rich research and practical experience. The primary assessment

aspects are the analysis of co-authorship, analysis of organization cooperation, analysis of keywords co-occurrence, citation analysis of literature.

2.3 Data Visualization

In ancient Greece, people used curves to represent the relationship between functions and variables, which was the beginning of human using visualization. In February 1987, the National Science Foundation held a conference on scientific computing visualization in Washington. The conference held that "applying graphics and image technology to scientific computing is a new field," and pointed out that "scientists not only need to analyze the calculated data obtained by computers but also need to know the data in the process of computation. Changes in the situation, and these need to be used in computer graphics and image processing technology. The development of visualization in scientific computing has promoted the research of data mining visualization. In recent years, with the explosion of the Internet, the popularization of computers and the development of data warehouse, visualization technology has made significant progress. Information visualization is to deal with the data types of this information and their related tasks in many fields, to find patterns, clustering, distinction and connection, trend and so on.

2.4 Exploration on Visualization Technology in the Field of Library

Visualization is the process of converting data, information, and knowledge into visual forms of representation. Visualization technology provides an interface between the two most potent information processing systems of human and computer. Using an efficient visual interface, you can quickly and efficiently interact with a large number of data to discover hidden features, relationships, patterns, and trends. The information pattern and data association or direction are intuitively presented to decision-makers. Data makers can interact with each other through visualization technology. Visualization has a broad and significant impact, which can lead to the new foresight and more efficient decision making. Data mining visualization is the process of finding and analyzing databases to find potentially useful information. More specifically defined data mining visualization is a process to find a specific subset of the database to assist in decision making. The application prospect of visualization technology in library service is shown in the following aspects:

- Using visualization technology to improve the interface of Online Public Access Catalogue (OPAC)

  OPAC is a software system designed to provide comprehensive and integrated bibliographic information services for librarians and users. However, OPAC is gradually disregarded by users

because there are many ways for users to obtain resources, it is no longer the best way for users to get the required bibliographic information. In order to improve OPAC, combined with a visual retrieval system called AquaBrowser Library (Kaizer & Hodge, 2005) is a good choice. With the support of this visualization technology, OPAC can not only help users explore, filter and discover knowledge systems and information features hidden inside bibliographic information, but also improve user search behavior efficiency in information retrieval process. By visualizing the multidimensional view, we can not only guide users' decisions at the semantic level but also reveal the internal relations of data, information and knowledge from the perspective of knowledge management, so that information recall and precision rate that users didn't notice were improved. The visualization technology of OPAC can reflect the retrieval process, make information retrieval process transparent, and provide friendly man-machine conversation and communication environment for users, enabling users' cognitive ability to integrate into information retrieval and information browsing naturally. Such an OPAC is no longer a simple bibliographic query system in the library. It should be a visual tool for users to do data mining, information processing, and knowledge management. The use of information visualization OPAC can improve the level of library resource information retrieval and enrich the experience of users. It is also one of the development directions of the next generation of OPAC.

- Using visualization technology for reference and consultation

  Reference service is to provide the information and help the user as many and quickly as possible by a librarian or information expert in a professional way. The essence of reference service is that the consultant participates in the interaction of the user, and seeks the solution that the user produces in the process of using the library.

- Using visualization technology to serve special readers

  Deaf-mute readers cannot accept traditional library services due to physical reasons, but the development of visualization technology provides opportunities and possibilities for deaf-mute readers to make use of libraries. Generally speaking, most deaf and dumb readers have good vision or better vision than ordinary people. The library can make full use of visualization intuitive and vivid features for their services, to expand the space of library services, broaden the service object, expand the scope of the library, improve the utilization of library.

2.5 Social Network Analysis in Co-authorship

Academic papers not only provide scientists with a way to acquire professional achievements and knowledge but also provide efficient means for researchers to obtain scientific resources and establish the academic communication network between scientific circles and industrial elites. Academic papers

often contain two or more authors, and monographs are declining. This phenomenon shows that with the development of science and technology changes faster and faster, theoretical and engineering disciplines are increasingly integrated, and new frontier disciplines are emerging as well. Scientific cooperation will help to promote the sharing of resources, the exchange of ideas and knowledge, and improve the efficiency of scientific research output. The traditional method of mathematical statistics is not suitable for the needs of the current management, so the co-authorship network is abstracted as complex social network, the author co-authored papers abstracted in social networks, expressed as the correlation between node and node, using the method of complex social network has become the primary technical means of the current.

Since 1994, since the advent of large digital libraries such as IEEE and ACM, some scholars have begun to pay attention to and study the coauthor network. The main reason is that the observation of the current state and structure of the co-authored network can provide a lot of valuable data. Co-authorship network is a typical social complex network. It has some structural characteristics, such as small world characteristics, cohesive tendencies and scale-free features in the network. Therefore, some tools and methods for analyzing complex networks can also be applied to the co-authored network, in order to dig into the useful information hidden in the co-authored network. These tools and methods have become the principal means to study the coauthor network.

Newman (2001) aimed at the fields of biology, mathematics and physics and constructed a separate co-authored network for the papers in four databases and made a quantitative analysis of the cooperative pattern of scientific papers by researchers. Newman (2001) analyzed the essential structural characteristics of the co-authored network, including the number of papers, the number of authors, and the changing features of cooperation patterns under different disciplines and time, it is observed that the co-author network has the following characteristics:

1. The average number of co-authors in the field of computer and theoretical physics is relatively small. The average amount of co-authors in experimental physics field is enormous, and there are significant differences between different disciplines and different directions.

2. The most substantial connected subset formed typically contains 80-90% of authors, indicating that academic groups developed by the same subject are interconnected.

3. In the same subject, the average distance between any two authors is 6, noting that the co-author network conforms to the characteristics of the small world.

2.6 Layout Algorithms For Co-authorship

The layout algorithm is the core and foundation of the visualization of the graph, which has an essential influence on the real-time processing of the graph data. Besides, the layout algorithm,

which integrates the aesthetic standards, can also improve the ability and readability of the map to a certain extent. Many scholars have studied the layout algorithm of the graph. S divides all graph visualization methods into nine categories, but there are overlapping problems in their classification, that is, one can belong to two categories at the same time. Cui divides it into five categories, but she mistakenly includes the layout algorithm of the hierarchical data into the layout algorithm of the graph data. The standard feature of these two classifications is that all of them involve the force directed layout algorithm. The force directed algorithm is the most widely used algorithm in literature. It can fully display the overall structure and Automorphism characteristics of the graph and has a substantial versatility. It plays a leading role in the layout algorithm of the graph.

2.6.1 Force Directed Layout

Force-Directed Layout algorithms are graph drawing algorithms based only on information contained within the structure of the graph itself rather than relying on contextual information. The most straightforward Force-Directed algorithm uses repulsive forces between nodes and attractive forces between adjacent nodes. Back to 1963, the graphical rendering algorithm of Tutte (1963) is a representation of the first force directed graph based on the center of gravity. More traditionally,the spring layout of Coleman and Parker (1996) and the algorithm of Fruchterman and Reingold (1991) both rely on the spring force algorithm, similar to Hooke's law. In these methods, all nodes have the repulsion force, but there is a mutual attraction between the adjacent nodes.

Alternatively, the forces between the nodes can be calculated according to their graph distance, determined by the length of the shortest path between them. Kamada and Kawai (1989) uses a spring force to proportion the graph theoretic distance. In general, force-directed method defines a target function, which maps the layout of each graph into a number, representing the energy of the graph. This function is defined in the way that low energy corresponds to the arrangement of adjacent nodes near each other at a certain distance, and the interval between non adjacent nodes spaced well. Then, the layout of a graph is calculated by finding the minimum value of a (usually local) function of the objective function.

The utility of the force directed approach is limited to the small graph, and the result is not good for the graphs of hundreds of vertices. There are many reasons for the poor performance of the traditional force directed algorithm on the large graph. One of the main obstacles to the scalability of these methods is that the physical model usually has a number of local minima. Even with the help of the complex mechanism of avoiding local minima, the basic force directed algorithm can not consistently generate a good layout for the large graph. The method of center of gravity is also not well done, mainly due to the problem of resolution: the minimum vertex separation of large graphs is often very small, which leads to unable to read drawings. In the late 90s of the last century, there were several techniques that extended the function of the force - directed method to graphs with

tens of thousands or even hundreds of thousands of vertices. In these methods, a common thread is multilevel layout technology, where the graph is represented by a series of simplified structures and arranged in opposite order: from the simplest to the most complex. These structures can be coarser graphs, as in the approach of (Hadany & Harel, 2001), (Harel & Koren, 2000), (Walshaw et al., 2000), or vertex filtrations as in the approach of (Gajer, Goodrich, & Kobourov, 2000). They have improved the basic algorithm of the force directed layout from the theoretical basis, the aesthetic standard and the ability to display so that the force directed layout algorithm is becoming more and more perfect.

The articles from the journal Bioinformatics, with 10 years of publications in the DBLP article database are used for the paperwork (Santamaría & Therón, 2008), the graphic uses a force-directed layout with two kinds of forces, and both forces are determined by the distance between the nodes. The overall result is that the nodes in the same group are often closer and separated from the nodes in the different groups. The layout is iterative, so after each cycle, the node will depend on the force applied, and the force recalculate the new location of the node. For each layout cycle, the node is bounded in the calculation position. In addition, each group is drawn with a round and transparent shape instead of drawing their edges. The outermost nodes are calculated by checking the location of each node in each group and determining which nodes are on the periphery of each moment. In order to improve the understanding of inter group relationships, the cross nodes are drawn into a pie chart, with the same number of groups in which the nodes belong. It is noticeable that the author's nationality is reflected in the form and issue of research groups. in figure 2.1, the central organization is the most influential author of bioinformatics and it also included the grouping of nationalities in Russian research field, and established contacts with German colleagues.

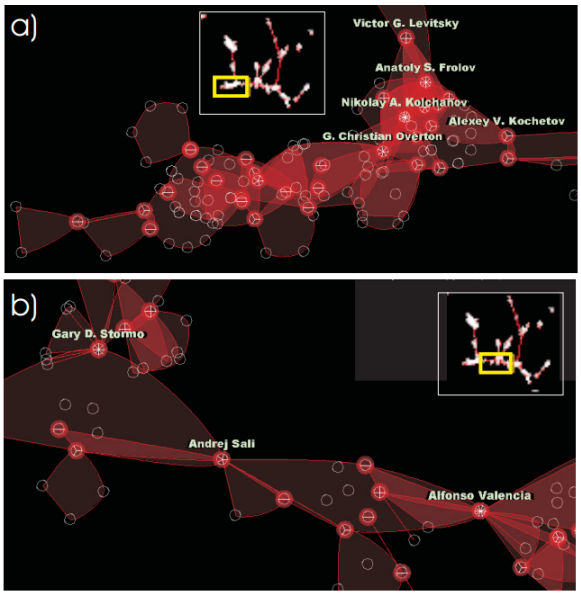


Figure 2.1: A force directed layout with two forces in co-authorships database (Santamaría & Therón, 2008)

The technique in the (Spritzer et al., n.d.) utilized builds upon the physics metaphor of traditional force-directed graph layouts to provide the user with interactive tools for the manipulation of the graph. The graph built with the database contains 474 nodes and 1252 edges. Each node represents an author while each edge represents all the publications between two authors. As attributes, each node contains the authors name, degree (number of edges connected to it), total number of publications, number of publications in conference proceedings, number of publications in journals, number of publications in books and their category (whether they are faculty members, students or external collaborators). Each edge contains the id of their two nodes, the years of their publications, the types of the publications (journals, conference proceedings or book) and the number of common publications.

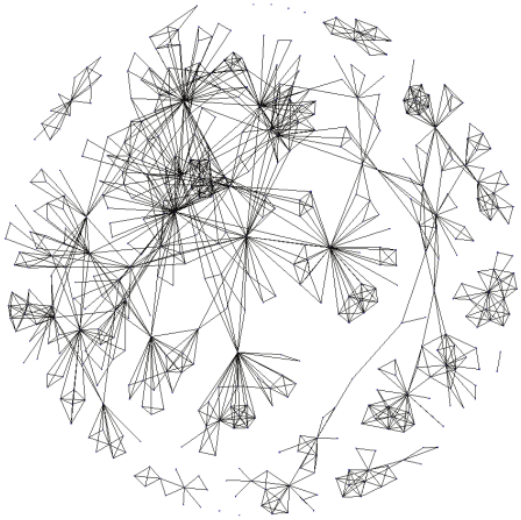


Figure 2.2: A force directed layout for co-authorships (Spritzer et al., n.d.)

An approach is proposed in paper (Collberg, Kobourov, Nagra, Pitts, & Wampler, 2003) that combines both readability and mental map preservation. The algorithms used to display a variety of program structure diagrams are based on GRIP. Through the stratified filtration calculation diagram, GRIP can draw a very large map in reasonable time. Figure 2.3 shows snapshots of the SandMark call-graph in a force directed layout which name is Frunchterman Reingold. Firstly, it shows that an early part of the system consisted of two main parts, the top on the left and the bottom on the right. Then nodes start out red. Then the node starts red. As time goes by, a node does not change, it turns purple and turns blue finally. When another change is affected, the node becomes red again.

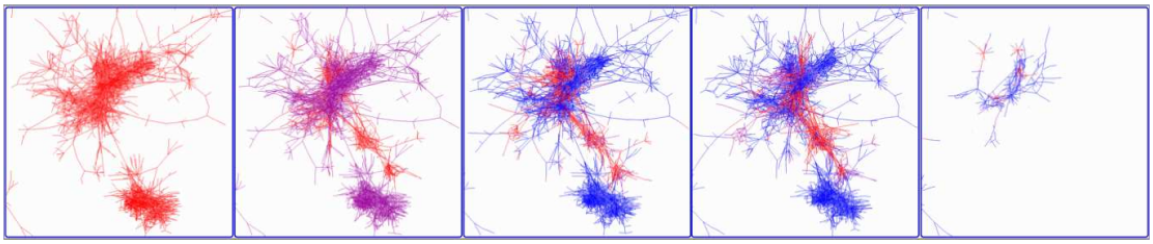


Figure 2.3: A force directed layout for co-authorships (Spritzer et al., n.d.)

2.6.2 Radial Layout

The term *radial visualization* seems to have been created by Hoffman, Grinstein, Marx, Grosse, and Stanley (1997) in 1990s, but its basic concepts are firmly rooted in the statistical graphic literature of nineteenth Century. Today, technologies such as pie charts, starplot, and radar plot are often used in the visual media of the business and communication of numerical data. These figures are the common ancestors of almost all radial visualization methods found in the most advanced studies.

- Pie charts

  A radial display is a visualization paradigm that arranges information in a circular or elliptical pattern. Perhaps, the earliest use of the radial display in the statistical graph was the pie chart. The first known occurrence of a pie chart was in William Playfair's 1801 treatise, a detailed list of the population and wealth of European countries in nineteenth Century (Playfair, 1801). Spoerri (2004) reviewed the historical background of Playfair's work and the influence of modern statistical graphics. Although often used in the mainstream media, the pie chart has some limitations. In particular, when the wedge in the pie chart is almost the same size, it is difficult to directly determine which one is the largest wedge.

- Star Plots

  The star plot is another form of the radial graph, which plays an important role in the statistical graph. This form of chart is alternately called a Kiviat graph or a spider web, which is specially designed to look at multiple systems in a compact form. A star plot is constructed by mapping each variable to one of several axes that are radiated from a common center point. The distance between the axes of each axis is proportional to the range of each variable, and the length of each axis is the same. Then draw the data points in the appropriate position on the axis and draw the straight lines that connect them. A star diagram is a radial equivalent of parallel coordinates (Inselberg & Dimsdale, 1987). In addition, multiple star plots can be superimposed to be compared between several different data sets with common field names. This produces a graph called radar plot (Wilkinson, 2006). When a region surrounded by an entity is completely

contained in another area or uses transparency, radar plot is the most effective choice, and it can compare the relative area that does not obscure.

- Sociograms

  A radial display of the first application is for sociometry, which is the study and measurement of interpersonal relationships in a group of people. The target sociogram, introduced by Northway (1952), describes the boundary between people and people in the small circle of human beings in figure 2.4. Radial visualization extends the concept of radial display, including the interactive operation of data. In general, this means that radial visualization has been implemented as part of a computer program, but this is not necessarily the case. For example, Northway (1952) also discusses Interactive simulation of a social graph called a the peg board sociogram. People are shown as pegs nailed to any place on the board, and the relationship is represented by a rubber band nailed to a nail. In this way, the location of the user and the relationship can be modified interactively.

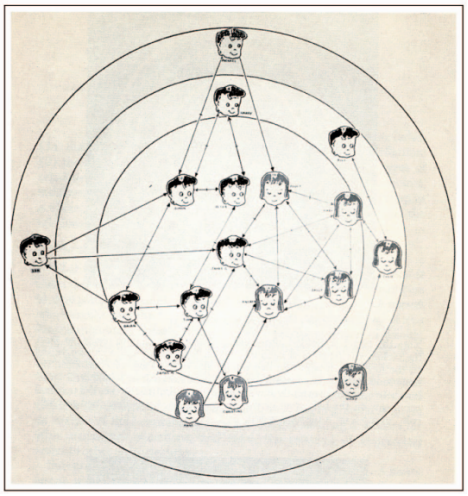


Figure 2.4: The target sociogram, an early example of radial display (Northway, 1952)

Radial visualization is a wide range of fields, and applications are targeted to different subject areas. However, this does not mean that radial visualization is the correct solution for every visual need. There are certain application domains that seem to be particularly suitable. Based on survey of the field, four basic application fields or kinds of data are identified, These techniques are successfully visualized by using radial technology. Which are:

- Hierarchical structure (Trees)

  Hierarchical visualization is still one of the most important problems in information visualization. File systems and organizational hierarchies are included in the commonly used datasets (Hong, DAndries, Richman, & Westfall, 2003; Teoh & Kwan-Liu, 2002; Wu & Takatsuka, 2006).

- The relationship between different entities

  In many multidimensional data, the relationship between several variables is often not obvious. For example, computer network traffic and alerts, population surveys,social networks (Livnat, Agutter, Moon, Erbacher, & Foresti, 2005; Van Berendonck & Jacobs, 2003; Keim, Mansmann, Schneidewind, & Schreck, 2006).

- The ranking of search results

  In modern computing, search engines are ubiquitous, making the application of this radial visualization particularly noticeable. Although the actual problem of search results ranking is a purely algorithmic problem, rather than a visualization problem, visualization is still an effective way to transfer relative rankings to users, so as to make final decisions (Institute, 2001; Spoerri, 2004; Torres, Silva, Medeiros, & Rocha, 2003).

- Serial periodic data

  It refers to continuous data, and shows a predictable repetitive structure. The most common example is time series data (Carlis & Konstan, 1998; Suntinger, Obweger, Schiefer, & Groller, 2008; Weber, Alexa, & Müller, 2001).

Biuk-Aghai (2006) using the modified radial layout which is called star layout in figure 2.5 to display the relationships between authors in the Wikipedia database, it shows co-authorships between a Wikipedia entity and all other related Wikipedia entities, regardless of type. Typically, it is used to show other Wikipedia entities related to a given Wikipedia article. To accommodate a larger number of nodes at each level, it has modified the radial layout slightly to arrange child nodes in a semi-spherical area around the parent node. When only one star is displayed, the positions of nodes around the central node are calculated so as to evenly distribute nodes in the star.

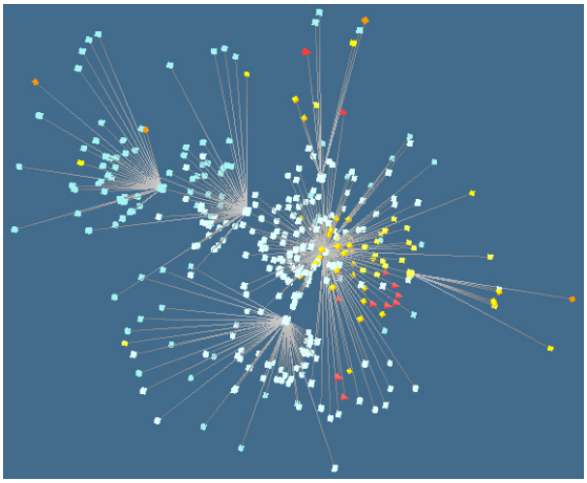


Figure 2.5: the modified radial layout for co-authorships in Wikipedia (Biuk-Aghai, 2006)

The multi-circular layout model is discussed and built for data visualization. The partitions of radial level layouts are placed on nested concentric circles (levels) and edges are drawn as curves between consecutive partitions. The positions of the vertices depict centrality measures. Additional information is reflected by the color, shape, size, and width of the vertices and edges, which are shown in figure 2.6. It is necessary that this prohibits intra-partition edges and edges connecting non-consecutive partitions.Edges are considered to be directed from lower to higher levels For technical reasons. In order to overcome the drawbacks of the radial layout algorithms described before, an extension of the sifting heuristic which computes a complete multi-circular layout is proposed by Baur et al. (2009) and edge crossings for optimizing both vertex order and edge winding should be considered, thus it is expected to generate better layouts.

Figure 2.6: (Baur et al., 2009)

In the Sparkler system, Hetzler et al. (1998) propose a Star-based visualization for performing queries on a collection of documents. Their visualization supports the depiction of multiple queries at once. The Starstruck system uses the Star pattern for showing relationships among themes within a document. The center of the visualization represents a given document and each theme is depicted as a small icon at the end of a line segment emanating from the center point. The single planar starbursts of the group is an invisible space along the pole. When the view is changed to look at the bar, all the starbursts are displayed the theme distribution given by the superposition presented in figure and the shape of the group represents the impression of the subject distribution in the group. As long as the rotation of the view is changed, the user can switch to view a single document as a whole with the theme of the whole group as a whole. Starstruck model allows to display several options for showing strength, which can be combined or used alone: the length and brightness of the ray can be fixed, and can also be changed according to the intensity of the subject, and it can be connected to the starbursts endpoint, so that each file forms a spider graph. and the subject ray label can be hidden or displayed as a vertex, and the color values can be specified for the rays and

the spider graph lines.

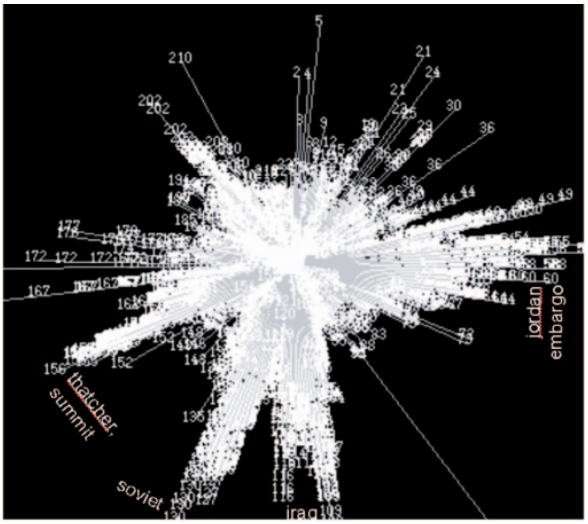


Figure 2.7: the Starstruck model (Hetzler et al., 1998)

2.7 Visualization Tools

2.7.1 Gephi

Gephi is an interactive visualization platform for various networks and complex systems, and is also a tool to facilitate people to explore and understand maps. Rich interaction means its characteristics and advantages. It can not only change the structure, shape, and color of every element of the graph, but also display information hidden behind the data, and it also possesses the animation ability that common visual tools do not possess. Gephi can increase the extension of the plug-in to support algorithm so that it can help the latest algorithm. At the same time, it also uses the three-dimensional engine to display the big picture network and accelerate data exploration process in real time. Its flexible multitask structure makes it possible to process elaborate data sets and produce valuable visual results.

2.7.2 UCINET

UCINET is a classic full-featured complex graph analysis tool. It is mainly used for the analysis of the social network. It is the most famous and most commonly used network analysis software package in the field of social network analysis. Compared with ease of use, UCINET pays more attention to speed, so its interaction interface and interaction means are not satisfactory. However, integration with other visualization tools, such as Pajek, Mage, and NetDraw, to some extent, has made up for this deficiency.

2.7.3 Tulip

Tulip is a visualization analysis framework for relational data. Its purpose is to provide developers with a complete library and Interactive design of information visualization for supporting relational data. The framework is developed in C + + language, which can further improve the algorithm, interactive technology, data model and unique domain visualization.

2.8 Visualization Evaluation

In the literature survey of about fifty user studies using the information visualization system, Plaisant (n.d.) found four subject areas of evaluation:

- Compare the control experiments of the design elements.

  This kind of research may compare unique components, such as comparison of alphaslider design (Ahlberg & Shneiderman, 1994). To compare different designs, an experiment was conducted to map 10000 items to a small portion of the pixels in an Alphaslider. The independent variable was the type of interface: Position interface, Acceleration interface, Micrometer interface, and Scrollbar interface. The dependent variables were: time to locate an item in the list and subjective satisfaction. For each interface, 25 tasks are randomly presented from 10000 MovieTitle lists with an average length of 19 characters. These tasks are generated when the start button is pressed at runtime. Five practice tasks are proposed for each interface topic. The slider pointer is returned to the middle of the slider before each task.

- Usability evaluation of tools.

  Some studies may provide feedback on the problems that the users will encounter with the tools and show how the designers improve the design (Sutcliffe, Ennis, & Hu, 2000; Byrd, 1999). (Byrd, 1999) designed a fully visualized experimental system compared with a control system without visualization, except to highlight the words in a single text. Byrd (1999) made two types of measurements: goals, including participants' judgments of "official" relevance, and the speed of their review of documents; For instance, how much they like to use visualizations. In order to minimize the discrepancy between the experimental system and the control system, the scroll bar code of the control system is the same as the experimental system code, except that the control system skips the drawing icon.

- Compare the control experiments of two or more tools.

  This is a common type of research. For example, three tree visualization tools are compared: SpaceTree, Hyperbolic and Window Explorer(Plaisant, Grosjean, & Bederson, 2002). These studies usually try to compare a new technology with state of the art. Topology tasks are used

in (Plaisant et al., 2002) to evaluate, such as listing all the ancestors of a node, find 3 nodes that have more than 10 direct descendants, which of the 3 branches of measurements contains a larger number of nodes, etc,. All tasks are used to compare all three tree visualization tools.

- The case study of tools in the realistic settings.

  This is the most unusual kind of research. The advantage of the case study is that they report users to do real tasks in a natural environment, and demonstrate the feasibility and contextual usefulness. The disadvantage is that when they are consuming, the results may not be replicated and promoted. The experiment in (Trafton et al., 2000) is a characteristic and exploratory study of a part of a field of research. the actual weather and the occasional computer problems can not be controlled, such as the web site of the WWW, the computer crashes, and so on. These are realistic problems normally encountered by METOC (ME Teorological and OCeanographic) forecasters.

2.9 Web Scraping Technology

In the ideal case, web scraping would not be necessary, and each site will provide an API to share data in a structured format. In fact, some websites do provide API, but they are usually limited by the availability of data and the frequency of access. In addition, the main task for web development is to maintain the front end of the interface than the back end API. Lawson (2015) said, "In short, we cannot rely on APIs to access the online data we may want and therefore, need to learn about web scraping techniques." DIT Arrow has the same API problem, so scraping technology is necessary. there are three different approaches to scraping data, regular expressions, popular BeautifulSoup module, and powerful lxml module.

2.9.1 Regular Expressions

Regular expressions (called REs, or regexes, or regex patterns) are a small, highly specialized programming language that embeds Python and passes through the re module. In this little language, "this set might contain English sentences, or e-mail addresses, or TeX commands, or anything you like. You can then ask questions such as Does this string match the pattern?, or Is there a match for the pattern anywhere in this string?. You can also use REs to modify a string or to split it apart in various ways. explained by Lawson (2015). Regular expressions are more forward-looking but challenging to build and become unreadable. Besides, there are a few other small layout changes that will destroy it, such as the title attribute is added to the label. It's clear that regular expressions provide a way to grab data quickly, but it's too fragile and easy to interrupt when updating a web page.

2.9.2 Beautiful Soup

The Beautiful Soup is a popular module, parsing a web page, and providing a convenient interface to navigate. Richardson (2013): " Beautiful Soup helps you pull particular content from a web page, remove the HTML markup, and save the information. It is a tool for web scraping that helps you clean up and parse the documents you have pulled down from the web." This approach is more verbose than regular expressions but is easier to construct and understand. At the same time, the change in the small layout is no need to worry, such as the extra space or the attribute of the tag.

2.9.3 lxml

lxml is a Python wrapper on top of the libxml2 XML parsing library written in C, which helps make it faster than Beautiful Soup but also harder to install on some computers. lxml is the most feature-rich and easy-to-use library for processing XML and HTML in the Python language, and it is the Python binding for two C libraries of libxml2 and libxslt. Its uniqueness lies in considering both the speed and functional integrity of these libraries, as well as the conciseness of Python API, mostly compatible but superior to the well-known ElementTree API.

2.10 Conclusion

In this chapter, the review of the existing literature is shown in three aspects, Firstly, the general introduction of the background and related domain are listed for a better understanding of the project. Secondly, the details of layout algorithms have been described, such as the history of the algorithm's development, the types contained, the involved domains about the algorithms, and the specific present paper about the algorithms' usage in co-authorship network. Then, some visualization tools, evaluation methods, and web scraping technology are reviewed for correct direction.

After the detailed literature review, the gap of this area is that there is no specific study about comparative layouts algorithms in co-authorships, usually are a design of particular visualization model or comparing two different layouts but not force directed nor radial layout, which makes this research meaningful.

Chapter 3

Experiment design and methodology

3.1 Introduction

This chapter will discuss the design methodology used in this research, the main sections are the dataset design and the scraping design, which need to confirm that what key data will be used in this research then how to extract it from DIT Arrow. Then, representing data in the force directed layout and the radial layout using visualization tools. Lastly, which layout can provide the more readable representation to the viewers in Arrow will be confirmed through evaluation.

An overview of the design is shown in figure[3.1], which outlines all the steps have been executed to perform the whole design, and the subsections of this chapter are as following:

- Database Design: the descriptions of the data source and the data details will be presented.
- Data Scraping Design: including the basic flow of the scraping design and the used scraper technology.
- Data Preparation
- Selections of Visualization Layouts and Tools: introduces the standards and the reasons for choosing layouts and tools.
- Evaluation Methodology: briefly introduces the overall design of evaluation.

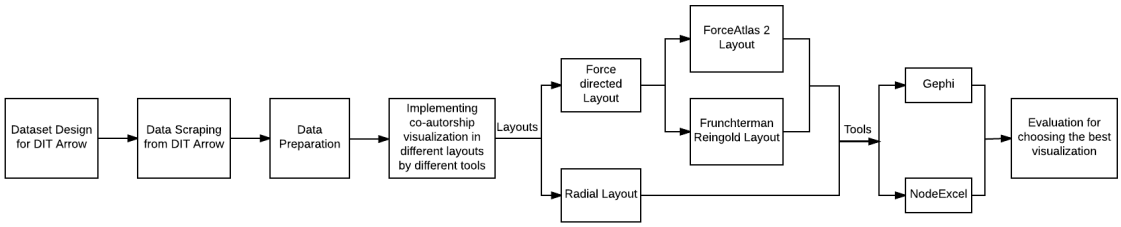


Figure 3.1: An overview of the design

3.2 Database Design

3.2.1 Data Source: DIT Arrow

Arrow (Archiving Research Resources on the Web) is a digital collection of research publications produced by researchers at Dublin Institute of Technology (https://arrow.dit.ie/). Arrow is an online institutional repository which brings together all of a University's research with the aim of preserving and providing access to that study. The information is provided in a textual format and is maintained by researchers adding new research works as they are published/developed. DIT Arrow currently contains an approximate of eleven thousand papers and grows steadily over time. However, there are only two visual graphs shown in the home page, one is about the readers' distribution which is a real-time display in the form of a map, and another one is for users who want to explore works in 765 disciplines in DIT Arrow. It is evident that Arrow still lacking in visualization especial in the area of relationships between authors or articles.

DIT Arrow provides many functions which allow users to search and seek the educational materials they are looking for. There are five different categories can be found on the home page, which are collections, journal collections, special collections, disciplines and DIT authors. In this research, I will use DIT authors category as the start, because the research focuses on the co-authorship. In this category, the list of DIT authors is represented in alphabetic order, and all work of each author can be found then, which offers key information such as the author's name, the articles title, document type, disciplines and the date it was added to the repository, and the papers are provided with the abstract, the citation and a link to the full document or DOI number.

3.2.2 Data Description

In this design, the detailed information of authors and articles will be needed, which is highlighted in the table 3.1, and the ideal database format is shown in the table 3.2, the method of scraping these data will be displayed in the next section.

document name	description
article title	the name of the article
author name	the name of the author which can be searched in the DIT authors category
co-author name	the name of the author who is coauthored in the same article
publication title	the type of the article
university	the university of the author
publication date	the publication date of the article

Table 3.1: Data description

article title	author name	co-author name	publication title	university	publication date
A Case-based Technique for Tracking Concept Drift in Spam Filtering	Padraig Cunningham	Alexey Tsymbal	Articles	Dublin Institute of Technology	Aug 05

Table 3.2: Database format

3.3 Data Scraping Design

3.3.1 The Basic Flow of the Design

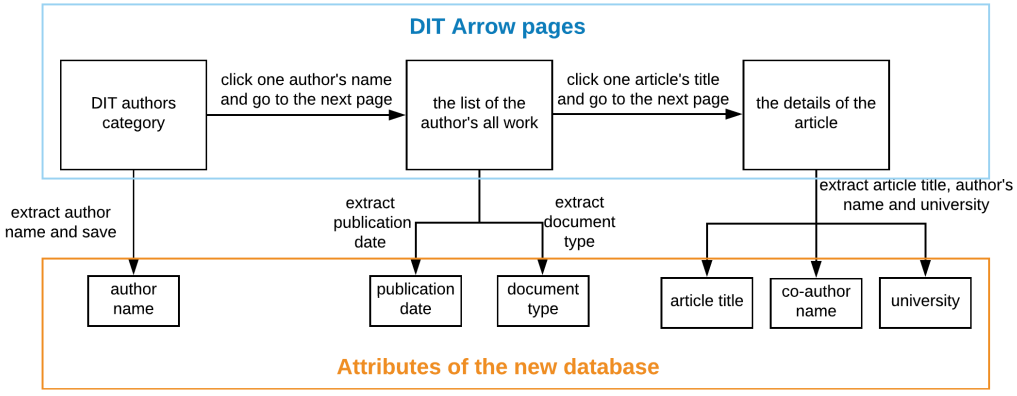


Figure 3.2: The overall flow chart about scraping design

In the part of scraping, because Arrow uses a third library API (Application Program Interface) to get information, but it is not allowed to access to public, which means the user data cannot be obtained

directly from API, the most feasible approach in this situation is using web scraping technology to extract the practical information from DIT Arrow for further research.

In this research, the overall scraping design for how to extract informative data from DIT Arrow and save them into the new dataset is shown in figure 3.2, and the detailed steps are as follows.

The ideal steps of scraping contain two main parts, which are data scraping and data storing. Firstly, DIT Authors category is selected as the start of scraping, extracting author's name in the listing of authors and saving them into the attribute of author name in the new database, but the order of the first name and last name is opposite which will be processed in the section of data preparation to guarantee the consistency of data. Secondly, extracting information of publication date and document type in the page of the list of each author's all work, the position is highlighted in figure 3.3. Lastly, extracting article title, author's name and university's name one by one, then store them into the attributes of the article title, co-author name and university separately in the database, the information which should be extracted is shown in figure 3.4. The most significant problem in this step is the duplicate; it is apparent that the same author name will be extracted twice and store in both attributes of author name and co-author name, which will be displayed in table 3.3. This problem will be solved in the section of data preparation as well. After macro scraping design in three necessary steps, an example of expected results after scraping is shown in table 3.3.

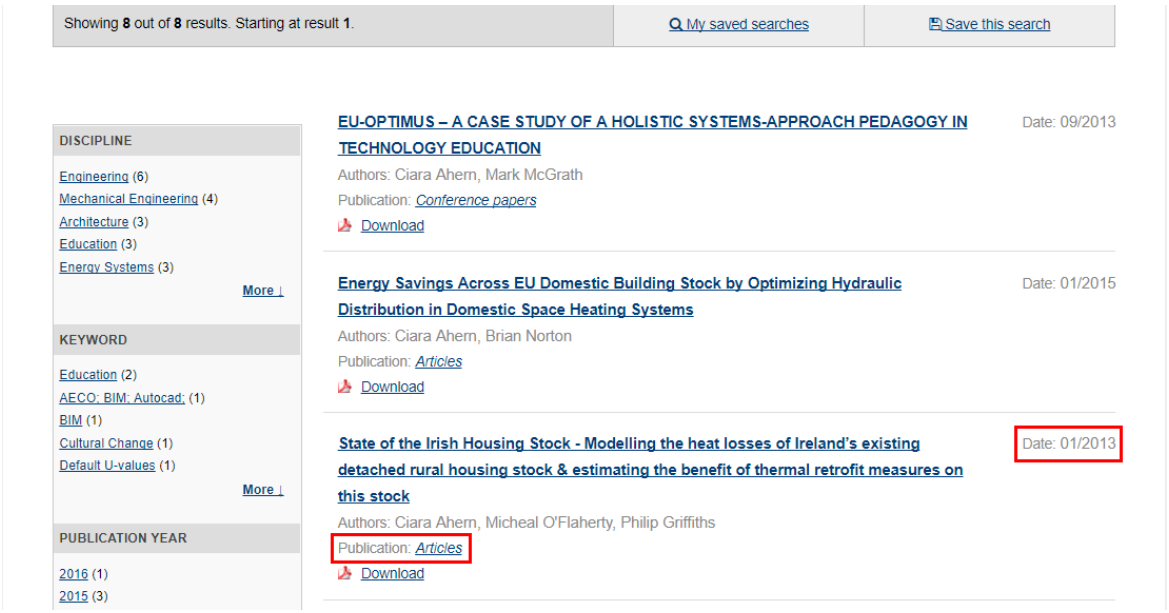


Figure 3.3: The scraping positions of date and type

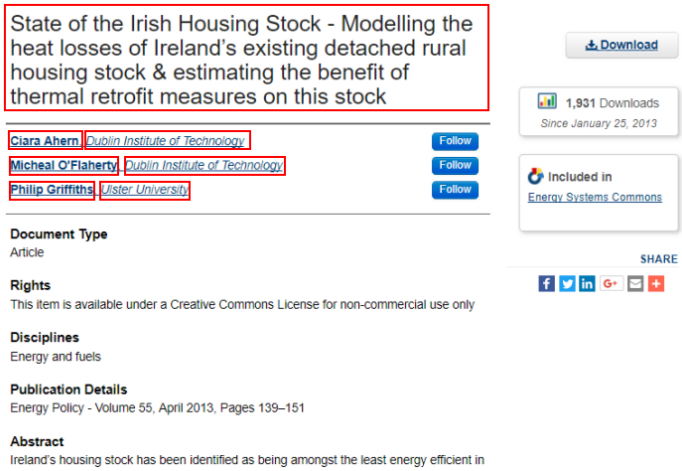


Figure 3.4: The scraping positions of author, article title and university

article title	author name	co-author name	publication title	university	publication date
State of the Irish Housing Stock - Modelling the heat loses of Ireland's existing detached rural housing stock & estimating the benefit of thermal retrofit measures on this stock	Ahern, Ciara	Ciara Ahern	Articles	Dublin Institute of Technology	Jan 13
State of the Irish Housing Stock - Modelling the heat loses of Ireland's existing detached rural housing stock & estimating the benefit of thermal retrofit measures on this stock	Ahern, Ciara	Micheal O'Flaherty	Articles	Dublin Institute of Technology	Jan 13
State of the Irish Housing Stock - Modelling the heat loses of Ireland's existing detached rural housing stock & estimating the benefit of thermal retrofit measures on this stock	Ahern, Ciara	Philip Griffiths	Articles	Ulster University	Jan 13

Table 3.3: An example of expected results after scraping

3.3.2 Web Scraper Tools

The selection of web scraper tools is also an essential process in this project because the accuracy and universality of the database should be guaranteed for further study. In the aspect of used tools, python is selected as the methodology for both scraping and storing in this research. The package lxml.etree is used for scraping, and the package pymysql is used for storing scraped data into the new database.

For the scraping part, one of the most critical parts of scraping the data in Python is to extract the required data from the obtained HTML page. There are three basic scraping methods in python, which are the regular expression, the lxml library such as etree, as well as BeautifulSoup. A conventional method of data extraction is to use regular expressions for matching extraction, which is a general way of string matching analysis. But for HTML pages, it does not make good use of its structural characteristics. lxml is the most feature-rich and easy-to-use library for processing XML and HTML in the Python language, and it is the Python binding for two C libraries of libxml2 and libxslt. Its uniqueness lies in considering both the speed and functional integrity of these libraries,

as well as the conciseness of Python API, mostly compatible but superior to the well-known ElementTree API. The package lxml.etree is a third-party library that combines the fast and powerful features of libxml2 and the ease of use in the Python language, which has a higher performance than the BeautifulSoup in parsing a web page, and BeautifulSoup also has slow speed, and the flexibility of parsing is not good, which is not suitable for DIT Arrow because of a large amount of data, that's the reason why use etree instead of BeautifulSoup in this project. Etree in the lxml package from python provides a better way to extract HTML page data more quickly and conveniently.

Element is a class of lxml, most of the XML is stored through this class. In lxml, the root can be created by the element method, and the tag attribute of the root is invoked, then add a child node to the root node. In order to facilitate access to child nodes, these child nodes are stored in a list, and the attribute format of XML Element is dictionary format to add or obtain. Then, start parsing a text, using Xpath to get the static text. Xpath is the XML path language, which is used to determine a location in the XML document. Xpath is similar to the function of the latitude and longitude network in geography, which is used to determine a particular position on the earth. The syntax of Xpath is similar to the regular expression. When Xpath is used to get the text in a web page, the required Xpath can be directly copied by examining the element function.

For the storing part, after finishing scraping useful data from DIT Arrow, the next step is stored them in a new database. pysql is used for storing in this research. pymsql is a module for manipulating MySQL in Python, and the way it handles is almost the same as MySQLdb.

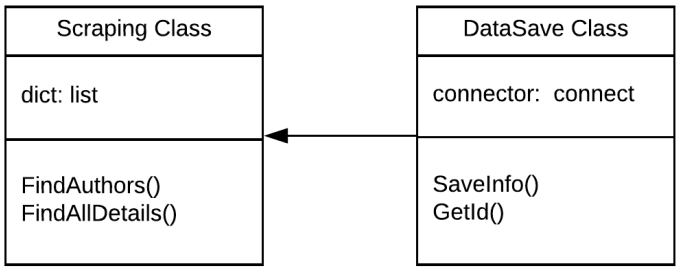


Figure 3.5: The class diagram

The class diagram is designed in figure 3.5. It contains scraping class and DataSave class. In scraping class, the variable of dict is defined as the list type in order to combine all details for each author, including author name, article title, university, publication date, publication title and download link. The method FindAuthors() is used to find the name of authors based on the URL of DIT authors list (https://arrow.dit.ie/authors.html), and the method FindAllDetails() is used to find

all attributes for each author, which is given by the previous method FindAuthors(). In DataSave Class, the variable of the connector is defined as the connected type in pymysql, in order to connect the python to MySQL database. The method of SaveInfo() is to save the scraped data into a new database, GetId() is to fetch the ID of each record to keep the integrity in the database.

3.4 Data Preparation

Not like Science Direct, IEEE Xplore Digital Library or ACM Digital Library have a unified format for every document, articles in DIT Arrow has many different forms which are hard to guarantee the consistency of the data, it is estimated that more time will be spent in the process of data preparation. As mentioned in the section of database design, firstly, in the DIT authors listing page, the order of the authors' name is surname name first and first name last, which is in the opposite order, should be changed. Secondly, the records of the same author's name are presented in both attributes of author name, and co-author name should be deleted. Thirdly, the scraping order is based on the DIT authors' list, and all authors' names will be extracted from the same article, so the duplication is unavoidable, for example, in the table 3.4, both two records represent the same relationship, because there is no priority of authors in DIT Arrow, one of them should be deleted. Lastly, there are still lots of inconsistent records should be cleaned and standardized, such as case sensitive, unexpected symbols, abbreviation/non-abbreviation, the appellation of the name, etc.

article title	**author name**	**co-author name**
A Case-based Technique for Tracking Concept Drift in Spam Filtering	Padraig Cunningham	Alexey Tsymbal
A Case-based Technique for Tracking Concept Drift in Spam Filtering	Alexey Tsymbal	Padraig Cunningham

Table 3.4: duplicate

3.5 Selections of Visualization Layouts and Tools

3.5.1 Visualization Layouts Selection

According to the aim in the introduction section and gaps in the literature review chapter mentioned before, force-directed layout and radial layout are selected in this experiment, in order to find out the best visualization for DIT Arrow. There are many different algorithms designed for force-directed layouts, such as Frunchternman Reingold algorithm, Force Atlas algorithm, ForceAtlas2 algorithm, OpenOrd algorithm, Yi Fan Hu algorithm, etc., Which are described in detail in the literature review chapter. In this research, most of these algorithms will experiment in order to find the most readable visualization for DIT Arrow. Unlike force-directed layout, the algorithms of the radial layout are not many. However, there are still many changes can be implemented in radial layout, such as the

selections of group type, order type, whether drawing spar/axis as the spiral, etc., All these possible approaches of bringing different visualizations specialized for DIT Arrow will be tested in the next chapter.

3.5.2 Visualization Tools Selection

After the selections of layouts are confirmed, the visualization tools should be considered, as described in the chapter of literature review, Gephi and NodeXL are chosen for further study. Gephi is a flexible and multi-task architecture which brings new possibilities to work with sophisticated database and produce valuable visual results. It can provide very high-quality visualizations, and it can also handle relatively large graphs, probably the most famous network visualization package can be found in Gephi, which is suitable and flexible for many layout approaches mentioned in the previous paragraph and also fit the large database created for DIT Arrow. Also, a few of the more common metrics such as degree, centrality, density, etc., can be calculated by Gephi. NodeXL is an Excel add-in which is easy to use based on Excel, and it can visualize and analyze complex social networks as well as it is also applicable to the vast database. However, NodeXL doesn't have all of the flexibility of Gephi regarding visualization but can produce some quality visualizations.

3.6 Evaluation Methodology

The Evaluation methodology of the readability of visualization layouts can only be carried out for a set of tasks and a set of graphs in this research. Two independent groups of users will be allowed to complete a series of the same tasks but different layouts through an online survey. Each group should contain at least 20 users to ensure the reliability and feasibility of the results for the further statistic analysis. The number of tasks should be at least 5, and the final score should consider both aspects of the accuracy of each question and the answer time because, in different visual graphs, the more readable a figure, the faster the user executes the task at hand and the less he makes mistakes. After collecting the results, the answers will be checked based on the statistical presentation; then the score will be obtained as a small data set for statistical testing, the final step is to analysis the results of the survey, to answer which layout is the most readable visualization for DIT Arrow. The design and implementation in detail will be shown in the chapter of evaluation because the elaborate taxonomy should be designed based on the specific visualization.

3.7 Conclusion

This section has shown the whole process of design, including database design, data scraping design, data preparation, selections of visualization layouts and tools, and the evaluation design. At first,

it has present what informative data should be extracted for visualization and how to scrape them correctly then store them in a new database, then using the database to implement different visualization using different tools. At last, an online survey has been chosen as the methodology to evaluate the readability of the representation.

Chapter 4

Implementation and Results

4.1 Introduction

This chapter presents the detailed process of implementing the proposed design, and the final visualizations will be shown. Based on the previous model, there are some changes performed during the implementation, the details of modifications will be explained in detail and the reasons why these changes have happened as well. The structure of this chapter is similar to the chapter 3; it is the process of implementation following the sequence of steps in chapter 3.

4.2 Data Generation

The process of data generation is implemented and discussed based on the designed methodologies in the previous chapter, and the high-level stages are following as creating the database schema, data scraping, and data preparation. Initially, the designed database is generated by SQL script based on the table 3.1 to save the information about co-authorship, created by scraping process.

After creating the desired database, the process of web scraping is rendered on Arrow DIT repository. As the discussion made in the previous chapter, the codes are divided into two files to process the steps of scraping and storing, respectively. In the class of scraping, it contains the functions of FindAuthors and FindAllDetails. Furthermore, in the function of FindAuthors, the method of HTML within the object of etree aims to extract and express the web page as JSON format based on the given URL that contains all authors' names at DIT. And also, the method of xpath, along with the object of etree, tends to extract all names of authors on this page, and call the other function, namely FindAllDeatils, within the for a loop. Also, in the function of FindAllDetails, the method of getting within the package of requests is to generate a JSON object based on the given URL, which displays all articles for each author. After that, in the for loop, the detailed information is temporarily saved to the list type of Dict. Then, it calls the function of saveInfo in the class of dataSave to transfer

the details from Dict to Mysql database in the loop as well. Besides, in the class of dataSave, the function of SaveInfo aims to connect python file to Mysql database using the method of attaching in the package of pymysql. In other words, it is similar to Open Database Connectivity (ODBC) driver. Also, it is required to execute the SQL statement to insert all details into the database with the error exceptions. The function of GetId aims to check the identification for each record so that it can make sure where has no condition on missing values.

Admittedly, one of the most unexpected severe mistakes occurs during the process of data scraping. Some authors could have many articles that exceed one page because one page only contains 25 pieces. Therefore, it can lead to the problem of losing data. Regarding the comparison after modifying codes, more than 10,000 new records are inserted into the database. This modification can robustly improve the reliability of visualization generation afterward.

4.2.1 Data Preparation

Selecting the necessary data from the original database to create a new dataset, which just contains author name, co-author name and article name, and then describe them. As the consideration about selected data, there still have some problems with data. At first, the order of name for each author is different between the column of author name and co-author name. In the pages of author list, the surname is in the front of the first name, but in the pages of the article, the surname is behind the first name. Therefore, for this issue, the order of author names should be consistently modified, that is the first name is in the front of the surname, using R script. Moreover, the duplicates cannot be avoided because of the complex structure of Arrow repository, so they should be removed. Besides, the issue occurs that the values are the same, treated as self-link node in visualization, which needs to be removed using R script as well. Furthermore, the issues of the case-sensitive and unexpected symbol should be handled as well to keep the datasets reliability and consistency. Moreover, the weight, defined as the values of cooperation frequencies between two authors, is calculated by R script. It is used to the undirected network to illustrate how the strong co-authorships perform among all authors. The higher values of weight mean, the stronger co-authorship between two nodes. This attribute is also an excellent measure for evaluation of network visualization.

4.3 Visualization Implementation

4.3.1 Visualization tools selection

As the discussion in the literature review, many data visualization softwares provide the functions and algorithms to implement the process of data visualization, such as Gephi and NodeXL. Therefore, first of all, this experiment aims to determine which the mentioned software for data visualization can deliver the better performance of co-authorship visualization for DIT Arrow repository. The data

generated and processed by the previous steps, such as scraping, storing and cleansing, is involved in this experiment to create the various network by various softwares. Also, the stage of software selection plays a significant role in the evaluation of network by surveys, because the best performance of visualization can provide the reliable and robust evidence to prove and answer the research question.

In Gephi, it provides many algorithms for network visualization and many measurements of network attributes. For example, the algorithms of Frunchterman Reingold, OpenOrd, Yifan Hu and radial axis layout are provided, and also the measurements of the network are offered as well, such as average degree, average weighted degree, network diameter and average path length. Another option is using NodeXL, which is a free and open-source network analysis and visualization software package for Microsoft. The NodeXL workbooks contain four worksheets that are Edges, Vertices, Groups, and Overall Metrics. The relevant data about entities in the graph and relationships between them are located in the appropriate worksheet in row format. Graph metrics and edge and vertex visual properties appear as additional columns in the individual sheets. It allows leveraging the Excel spreadsheet to edit existing node properties quickly and to generate new ones, for instance by applying Excel formulas to existing columns. Each designed layout is implemented and generated by both Gephi and NodeXL. However, the visualization presented in NodeXL is not very clear and not much useful information as well. The final representations are shown in Appendix A.6 and Appendix A.2. It can be seen that the nodes are too concentrated and not flexible to modified, because of the huge database, there exist too many overlaps in these two visualizations generated by NodeXL, which is not informative and hard for users to do the survey. After comparing these two data visualization softwares on the property of generated visualization, Gephi is considered to be selected for the further step that surveys evaluation.

4.3.2 Force directed layout

As mentioned in chapter 2 and 3, there are many kinds of force directed layouts algorithms can be chosen for the most suitable one, so most of them will be generated in Gephi for comparison, which are ForceAtlas, ForceAtlas2, OpenOrd, Yifan Hu, Frunchterman Reingold. After creating all different layouts, the layout algorithms of ForceAtlas2 and Frunchterman Reingold are kept for further research, because these two visualizations can represent the valid information and explicit relationships for such big database from DIT Arrow than other layouts. For different layouts, there are many overlaps exist in ForceAtlas layout, which is a severe problem because the database is too big, too many overlaps would affect the visual display of users. As for Yifan Hu algorithm, the distribution of nodes and links are also scattered, the nodes and relationships near the edge would be ignored, which influence the integrity of the database. For OpenOrd layout, the groups of authors are shown clear and easy to be recognized, but it also has the overlapping problem, and it is a little bit scatted as well. To summarize, in this project, the layout algorithms of Frunchterman Reingold

and ForceAtlas2 are the best two layouts among other force-directed layouts.

Frunchterman Reingold layout

The layout of Frunchterman Reingold is used to generate the visualization of co-authorship in Gephi. The relevant parameters of this layout are defined as follows. The settings of the area, gravity, and speed in Frunchterman Reingold are the corresponding to 10,000, 10.0 and 1.0, respectively. Then, the co-authorship visualization is shown in figure 4.1. As can be seen, the author Hugh Byrne has the most value of degree due to having the most prominent node in the network, and also some middle-size nodes are around Hugh Byrne. It indicates that Hugh Byrne is the most active author in DIT, and even this author has an academic networking in DIT as well. Furthermore, there have many clear blue lines in the network. Therefore, the relationship of co-authorship among all DIT's authors is slightly stable according to the amount of blue line in the network. Also, the depth of color for blue line illustrates what extent co-authorship between two authors. In a word, the more in-depth color means, the stronger cooperation and co-authorship. However, there also have many single nodes in this network, because some authors could be a student who only has an article for the thesis. Furthermore, the relationship between the students and their supervisors could just have one cooperation that reflects the visualization network as a single edge, and the value of weight is one.

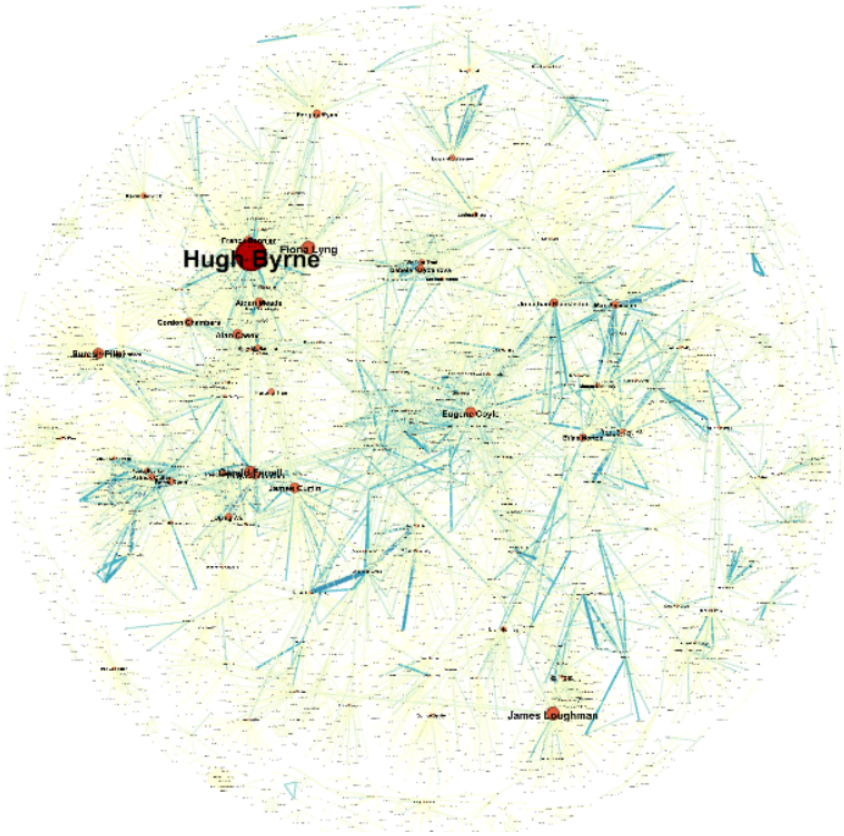
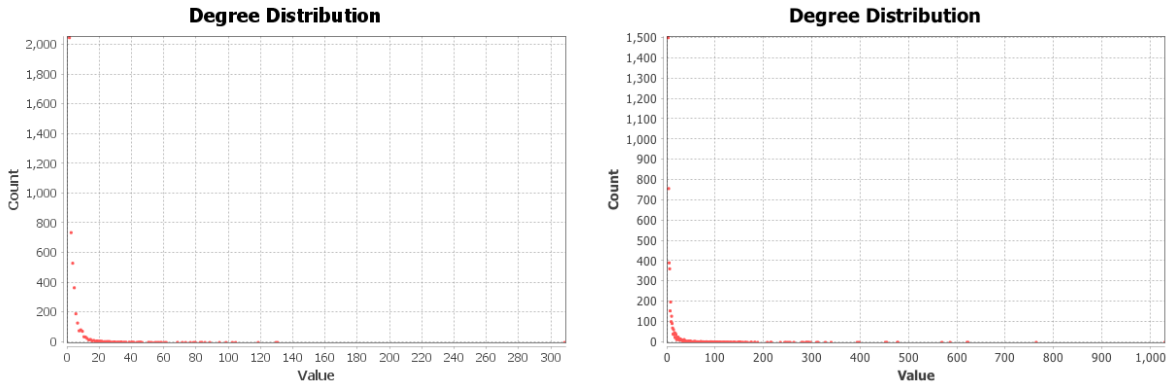


Figure 4.1: Frunchterman Reingold layout



(a) Degree Distribution

(b) Weighted Degree Distribution

Figure 4.2: The Degree Distribution

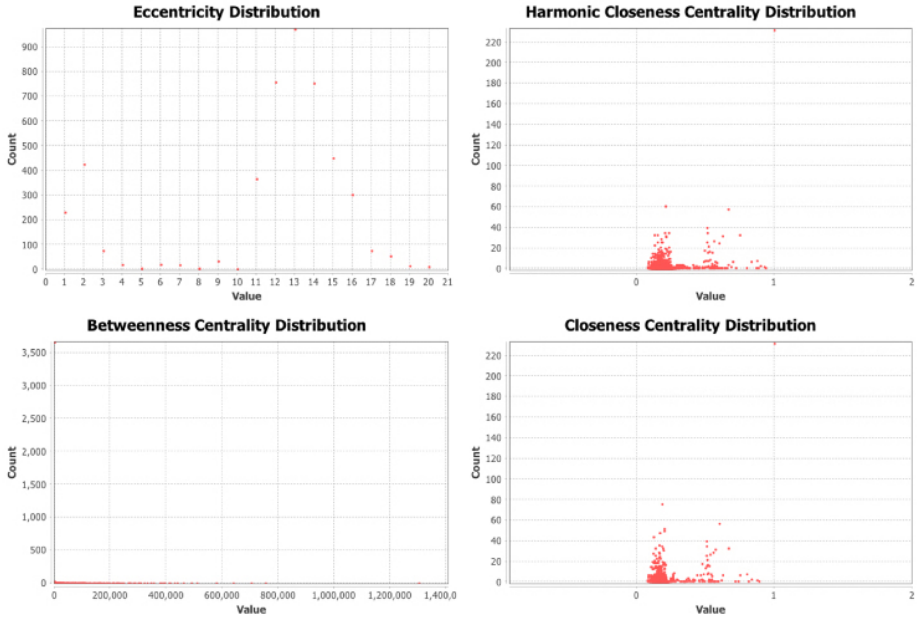


Figure 4.3: The results of measurements

Besides, the measurements are conducted to explore the insight property in this network. The results of the measure for this network are 4.083, 11.331, 20, and 6.288, corresponding to the methods of average degree, average weighted degree, network diameter and average path length, respectively. And also the distributions are provided in figure 4.3, including the distributions of eccentricity, harmonic closeness centrality, betweenness centrality, and closeness centrality. What is more, the average path length is a concept in the network topology that is defined as the average number of steps along the shortest paths for all possible pairs of network nodes. It is a measure of the efficiency of information or mass transport on a network. And also, the distribution of degrees, such

as average degree and weighted degree, can reveal the overview distribution of degree for each node on a network. The figure 4.2 illustrates both distributions of degree and weighted degree on the network of co-author. As can be seen, the most of authors with the degree of co-authorship are around 4. In another word, the authors of DIT nearly have 4-times article co-authorship based on the measure of the network.

ForceAtlas2 layout

With the same steps, the ForceAtlas2 layout is implemented in Gephi with the relevant parameter settings for this layout. The parameters of the number of thread, approximation, scaling, and gravity are set as 3.0, 1,2, 2.0, and 1.0. The parameter of estimate describes the theta of Branes Hut optimization, and gravity attracts nodes to the center, so that prevent islands from drifting away. The final visualization using ForceAtlas2 layout is shown in figure 4.4. And also, the relevant measures are generated by the algorithms as well. However, the results of each measure for this network are the entirely same as the previous results, such as the average degree, average weighted degree, average path length, and network diameter. Therefore, it is clear that the various layouts or algorithms do not affect the property and attribution of the network. The network only can be influenced by data itself only, and the different layouts can impact on the performance of application for end users.

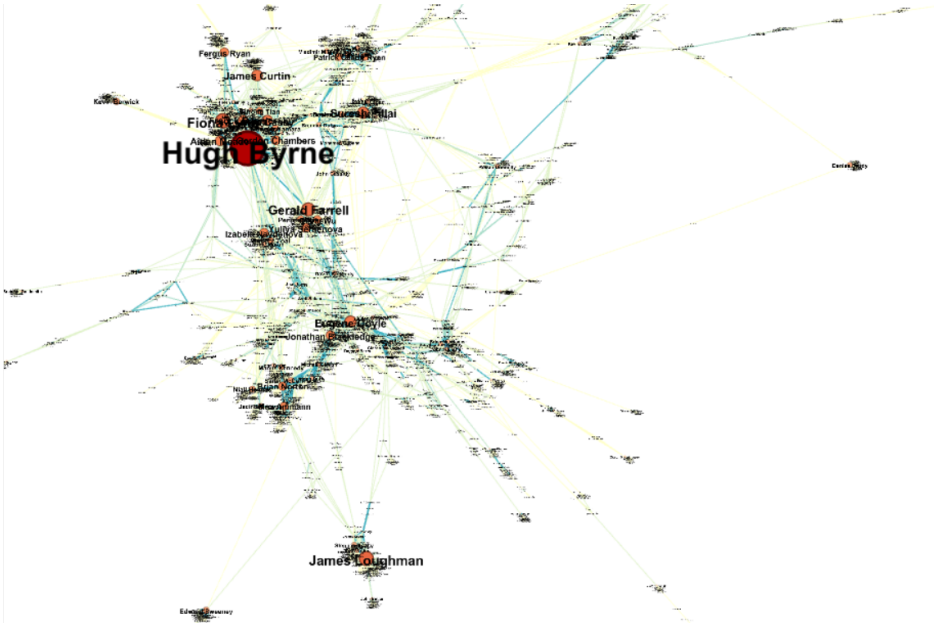


Figure 4.4: ForceAtlas2 layout

As can be seen above, Hugh Byrne is still easily and explicitly identified as the most significant node on this network. Unlike the circle-shape network using Frunchterman Reingold layout, the layout of ForceAtlas2 is similar to the shape of a spider web. Therefore, the different algorithms for

visualization layout can yield the different outputs with the various performances. Furthermore, the groups can be viewed and identified on this network using ForceAtlas2 layout, and even the small groups on the network. It is critical that the ForceAtlas2 layout and Frunchterman Reingold layout could be used in the different applications based on their purposes.

Radial layout

According to the literature review in chapter 2, another layout, called radial layout, is generated by Gephi, shown in figure 4.5, compared to the force directed layout. However, not as expected, the radial layout seems not fit DIT Arrow, all different settings are arranged and combined, the most evident radial visualization is shown in figure 4.5, it is apparent that figure 4.5 is not a good visualization, the distribution of nodes and links are unreasonable, and really hard to find the relationships even a single node or link, not to mention that if the labels are added, the whole picture will be more chaotic. The reason why this happened maybe because DIT Arrow is not a hierarchical co-author network, there is no priority in the authors for the articles, and the groups of authors also not defined, it can only be analyzed through modularity of authors' relationships, which is not correct because there are more than 26,000 co-authorships in this project, the modularity analysis cannot guarantee the accuracy. In this situation, for further research, the radial layout will not be involved, the evaluation of two force-directed layouts will be processed to find the most readable visualization for DIT Arrow.

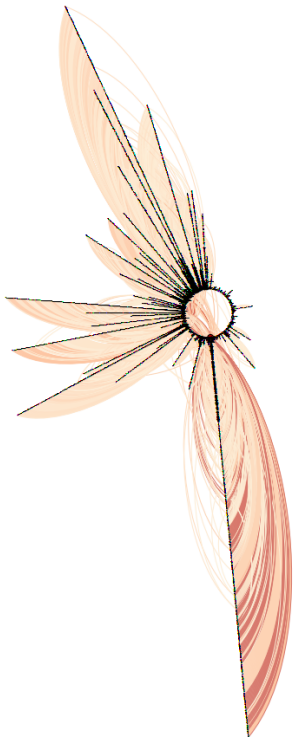


Figure 4.5: Radial layout

4.4 Conclusion

This chapter presented the steps and details to implement the various visualizations using different layout algorithms for DIT Arrow repository. First of all, the data was generated by Python script with the relevant data cleansing process. Then, the visualizations using the methods of force-directed layout and radial layout are generated with the network measures. The selection of data visualization software was discussed as well. Furthermore, the comparisons among all generated visualizations were considered and explained. In conclusion, the visualization obtained by the algorithm of the force directed layout will be evaluated and discussed in next chapter.

The next chapter, namely evaluation, and analysis will present the contents about the process of assessment and the relevant analysis for each visualization, including the survey design and implementation, the analysis of results with the statistical test.

Chapter 5

Evaluation and Analysis

5.1 Introduction

This section will evaluate the final visualization implemented in the previous chapter using the method of online survey, in order to find which visualization can perform best in DIT Arrow. There are two parts in this section which are evaluation implementation and analysis implementation. In the role of evaluation implementation, the experimental design will be shown in detail, including task taxonomy, users' description, features in the process of the survey. Then, the results of the survey are collected and collated, then creating a small dataset based on the accuracy of questions and the answer time. In the part of analysis implementation, using the dataset created in the previous section, doing some statistical testing and drawing detailed graphs to figure out which one is the most suitable visualization for DIT Arrow.

The structure of this chapter contains two main implementations:

- Evaluation Implementation presents the experimental design, the process, and the final results.
- Analysis Implementation: shows the process of data preparation, statistical testing, and the statistical results.

5.2 Evaluation Implementation

5.2.1 The Experimental Design

Fueled by the rapid growth of social networks and social media, the interest in more powerful network visual analysis tools and methods is growing as well. One of the most pressing challenges in facilitating network evolution analysis. The comparison of two visualization layouts can only be carried out for a set of tasks and a set of graphs through the survey in this research. Two independent groups of

users will be allowed to complete a series of the same tasks but different layouts through an online survey.

In this design, Typeform (https://admin.typeform.com) is used to build and design the survey sheet. Typeform has a concise user interface and powerful functional options and supports nearly 20 types of questions; each type has a wealth of opportunities can be set, such as add video or pictures, the color and font of the questionnaire can also be customized. Once the survey is created, the questionnaire links can be shared and sent to the users, which is convenient. Typeform also collected relevant information for the users who released the survey, including the number of answers, the number of responses, the completion time, the average answer time, the use of terminals (computers, flat panels, smartphones, etc.), which is helpful for the further study. Finally, Typeform's statistical results can be displayed and exported to xls format or online instant report, which saves a lot of time for statistical testing in the next chapter.

Task Taxonomy

The readability of a visual graph must be related to the ability of the user to answer some questions about the overview. A readable layout for a figure is one that shows the underlying relationships, so the problems should be tackled by considering the most generic tasks of visualization and attached with corresponding visual graphs. In this research, the same tasks will be shown to two independent groups, so the design of task taxonomy should be generic and not biased for either visualization and focusing only on general characteristics of graphs. In both visualizations, there are only presents the author with node and their relationships with links, so the type of task taxonomy is attribute-Based Task, which contains the nodes and links. There are seven tasks designed for the survey, three of them are about the nodes, and other four are about the links. Each question's score is 10, and the overall score is 70, and the correct answer will be given based on the results of data statistic. All tasks will be presented below.

Q1: Which author has the most co-authorship in Arrow DIT? (Which node is the biggest in this graph?)

In this question, the aim is to check whether the author who has the most co-authorship can be presented in both visualizations. This question is one of the most representative attribute-based tasks to test the readability of a visual graph. In the database, there is only one author has the most co-authorships and outclasses other authors according to data statistics, so there is only one right answer worthing 10 points.

Q2: How many authors have a visible number of co-authors? (How many individual nodes can you identifying)

This question is to explore how many authors often coauthored with other authors in DIT. There are above 26,000 records in the database, which is too many to be presented clearly in visualization,

so there is no specific answer in this question, but based on the statistics, the degree of nodes should be greater than or equal to 30 can be regarded as identifiable, which are 75 authors, so the answer between 65 and 85 can be considered as the correct answer. It seems not convincing, but for this vast database, the results obtained still have some statistic significance.

Q3: How many visible co-authors can you identify? (How many individual links can you identify that are associated with identifiable nodes))

This question aims to find the visual links which represent which authors often coauthored together in DIT. The more times they coauthored, the thicker links are. The same problem of the massive database as the previous task, after statistical analysis, the weight of links should be more prominent or equal to 30 can be treated as discernible, which are 56 links, the answer is the interval of [46, 66] can be checked as correct. It is predictable that both Q2 and Q3 will not get the high accuracy.

Q4: For author Hugh Byrne's co-authorship, which author collaborated with him the most? (Which node has the thickest link with Hugh?))

In this task, choosing author Hugh Byrne who has the most co-authorships in DIT authors as the target, then find who collaborated with him the most frequent. The graph should be zoomed in, and the node with label 'Hugh Byrne' should be set in the center of the graph for users' clear watch, in order to find out the thickest link connected with Hugh Byrne. It is am accessibility tasks of topology-based tasks. In the data statistic, there are 69 coauthored times between Hugh Byrne and Frank Bonnier which is the most co-authorships connected with Hugh Byrne, so in both surveys, answer 'Frank Bonnier' is right.

Q5: Can you find author Gary Henehan/Andreas Schwarzbacher??)

In this task, it is also an attribute-based task, a graph of partial visualization is shown, the users will be allowed to find a small but identifiable node, which degree is about 25 such as Gary Henehan and Andreas Schwarzbacher. It is a natural 'Yes/No' question.

Q6:Which author has co-authorship with Sarah Jane Delany and John Kelleher both? (Which node is a common neighbor between Sarah Jane Delany and John Kelleher?)

In this task, users need to find who has the relationship with both Sarah Jane Delany and John Kelleher. It is an accessibility testing tasks of topology-based tasks. The aim is to find a node that has the strong connection with two other nodes in many intricate relationships to test the readability of both visualizations.

Q7: How many groups which authors always work together can be found in this graph? (like a closed triangle or polygon)

This question is aimed to find groups in the visual graph. There are many closed triangles and polygons presented in the visualization, the more obvious they are, the more they collaborate. Through this task, we can find which authors in DIT always work together, which is significant for DIT Arrow. However, this one also has no specific answer same as Q2 and Q3; the estimated value

can be set between 20 to 30 based on statistical analysis.

The final score should consider both aspects of the accuracy of each question and the answer time, because in different visual graphs, the more readable a graph, the faster the user executes the task at hand and the less he/she makes mistakes. If the user answers quickly and correctly, the visualization is very readable for the task. On the contrary, if the user needs a lot of time or if the answer he provides is wrong, then the visualization is not well-suited for that task.

The Population

The population that participate this survey are students mostly, and they will be separated into two independent groups for different visualization but same questions to ensure the fairness and non-biased, because the readability also depends on the familiarity of visualizations to users and the answers for two separate surveys are equal, so one participant can only complete one survey. Each group should contain more than 20 users to ensure the reliability and feasibility of the results for the further statistic analysis.

5.2.2 Survey Result

Once the surveys are built, both two links are sent to 40 people separately, after two days' collection, survey 1 (Frunchterman Reingold layout) has received 33 responses, and survey 2 (ForceAtlas2 layout) has received 35 responses. The initial results of both surveys has been exported into xlm format as new datasets, which are shown in Appendix A.7 and Appendix A.8. In order to start the comparative analysis, the amounts of records in both datasets should be same, after careful review of both datasets, there are a few low-quality records exist in both datasets, such as the same answer for all questions, fill number or unexpected symbols but text required from questions, no answers but submitted, irrelevant answers or reasoning. After cautious deletion, 30 records with guaranteed quality have been kept in both datasets for further research in the next section.

5.3 Analysis Implementation

5.3.1 Data Preparation

Both datasets are small datasets so that they can be processed and scored manually. The details of scoring are shown in figure 5.1 and figure 5.2. The main factors in this dataset are scores and time. Based on the scoring criteria mentioned in the section of the experimental design, the scoring results of each question and overall final marks for every user are calculated manually. Then, based on the start date and the submit date, the duration of the answer time can be calculated by Excel. Also, the accuracy rate of each question is calculated and recorded in the second row in both datasets for

further research.

	Q1	mark	Q2	mark	Q3	mark	Q4	mark	Q5	mark	Q6	mark	Q7	mark	FinalScore	Duration
2	**Hugh Byrne**	**66.7**	**75 [65,85**	**26.7**	**56 [46, 66**	**10**	**Franck Bonnier**	**67**	**Yes**	**56.7**	**Brian MacNamee**	**97**	**[20, 30]**	**63.4**	**70**	
3	Hug Byrne	10	43	0	It's difficul	0	Franck Bonnier	10	Yes	10	Bran Mac Naree	10	9	0	**40**	07:44
4	Hugh Byrne	10	77	10	121	0	Franck Bonnier	10	Yes	10	Brian Macnamee	10	23	10	**60**	10:16
5	Hugh Byrne	10	65	10	125	0	Franck Bonnier	10	No	0	Brian MacNamee	10	11	0	**40**	05:51
6	Bryne	10	33	0	5	0	Fiona lyng	0	Yes	10	Brian MacNamee	10	4	0	**30**	03:04
7	Hugh-Byrne	10	13	0	66	10	Franck-Bonnier	10	Yes	10	Brain MacNamee	10	28	10	**60**	17:39
8	Hugh Byrne	10	70	10	80	0	Fiona Lyng	0	No	0	Brian MacNamee	10	8	0	**30**	07:32
9	Hugh-Byrne	10	21	0	9	0	Kevin Berwick	0	Yes	10	Gemma Kinsella	0	22	10	**30**	14:34
10		0	48	0	45	0	Fiona	0	No	0	Brain	10	18	0	**10**	02:48
11	Hugh Byrne	10	4	0	4	0	Franck Bonnier	10	No	0	Brian MacNamee	10	8	0	**30**	17:00
12	Hugh Byrne	10	12	0	2	0	Franck Bonnier	10	Yes	10	Brian Mac Namee	10	5	0	**40**	04:45
13	hugh byrne	10	48	0	30	0	franck bonnier	10	Yes	10	brian macnamee	10	8	0	**40**	03:02
14	Hugh Byrne	10	48	0	120	0	Frank Bonnier	10	Yes	10	Brian Mac Namee	10	10	0	**40**	08:56
15	Hugh Byrne	10	22	0	24	0	Frank Bonnier	10	No	0	Brian Macnamee	10	28	10	**40**	04:23
16	hugh	10	50	0	50	10	12	0	Yes	10	brian	10	20	10	**50**	02:42
17	Byrne	10	69	10	48	10	Franck Bonnier	10	Yes	10	Brian Mac Namee	10	30	10	**70**	41:35
18	hugh byrne	10	83	10	32	0	kevin berwick	0	Yes	10	brian macnamee	10	11	0	**40**	09:13
19	Hugh Byrne	10	65	10	20	0	franck bonnier	10	Yes	10	Brian MacNamee	10	26	10	**60**	25:16
20	Hugh Byrne	10	45	0	30	0	Franck Bonnier	10	No	0	Brian MacNamee	10	30	10	**40**	06:21
21	Hugh Byne	10	58	0	23	0	Flona Lyng	0	No	0	Brain MacNamee	10	12	0	**20**	05:16
22	Hugh Byrne	10	82	10	30	0	Kevin Bonnier	0	Yes	10	Brian Mca Namee	10	7	0	**40**	03:13
23	Hugh Byrne	10	35	0	Too many	0	Franck Bonnier	10	No	0	Brian MacNamee	10	1000	0	**30**	03:15
24	Hugh Byrne	10	76	10	17	0	Frank Bonnier	10	Yes	10	Brian MacNamee	10	7	0	**50**	08:24
25	hugh byrne	10	36	0	13	0	franck bonnier	10	No	0	Brian Mac	10	7	0	**30**	05:04
26	Hugh Byrne	10	6	0	10	0	Franck Bonnier	10	No	0	Brian MacNamee	10	4	0	**30**	03:15
27	hugh byrne	10	16	0	22	0	franck bonnier	10	No	0	brian macnamee	10	23	10	**40**	05:25
28	Hugh Byrne	10	50	0	5	0	Franck Bonnier	10	Yes	10	Brian MacNamee	10	6	0	**40**	07:25
29	Hugh Byrne	10	29	0	104	0	Franck Bonnier	10	Yes	10	Brian MacNamee	10	10	0	**40**	08:15
30	Hugh byrne	10	58	0	30	0	Flona lyng	0	No	0	Brian mac namee	10	8	0	**20**	03:28
31	Hugh Byrne	10	40	0	30	0	Franck Bonnier	10	No	0	Brian MacNamee	10	20	10	**40**	03:34
32	I	0	50	0	17	0	Fiona Lyng	0	Yes	10	Brian MacNamee	10	30	10	**30**	05:58

Figure 5.1: The results of scoring survey 1

	Q1	mark	Q2	mark	Q3	mark	Q4	mark	Q5	mark	Q6	mark	Q7	mark	Final Score	Duration
2	**Hugh Byrne**	**1**	**75 [65,85]**	**0.03**	**56 [46, 66]**	**0**	**Franck Bonnier**	**0.83**	**Yes**	**0.63**	**Brian MacNamee**	**0.867**	**[20, 30]**	**0.134**	**70**	
3	Hugh Byrne	10	8	0	22	0	Franck Bonnier	10	Yes	10	Brian Macnamme	10	26	10	50	04:02
4	Hugh Byrne	10	20	0	10	0	Franck Bonnier	10	Yes	10	Brian MacNamee	10	11	0	40	03:15
5	Hugh Byrne	10	41	0	13	0	Franck Bonnier	10	No	0	Brian MacNamme	10	7	0	30	03:19
6	Hugh Byrne	10	33	0	11	0	Franck Bonnier	10	No	0	Brian MacNamee	10	10	0	30	04:17
7	Hugh Byrne	10	29	0	15	0	Franck Bonnier	10	Yes	10	Brian MacNamee	10	3	0	40	02:51
8	Hugh Byrne	10	25	0	it is too mar	0	Aidan meade	0	No	0	Brian MacNamee	10	it is too n	0	20	04:26
9	hugh byrne	10	28	0	11	0	mary mc	0	No	0	brian macnamee	10	12	0	20	05:02
10	Hugh Byrne	10	20	0	15	0	Franck Bonnler	10	Yes	10	brian macnamee	10	10	0	40	02:13
11	Hugh Byrne	10	30	0	200	0	Franck Bonnier	10	No	0	Brian MacNamee	10	100	0	30	05:26
12	Hugh Byrne	10	25	0	21	0	Franck Bonnier	10	Yes	10	Brian MacNamee	10	10	0	40	02:52
13	Hugh Byrne	10	28	0	11	0	Franck Bonnier	10	Yes	10	Brian MacNamee	10	11	0	40	06:15
14	Hugh Byrne	10	25	0	30	0	Franck Bommier	10	Yes	10	Brian MacNamee	10	28	10	50	06:35
15	Hugh byrne	10	56	0	6	0	Franck bonnier	10	No	0	Susan MC keaver	0	7	0	20	02:37
16	hugh Byrne	10	23	0	6	0	frank bonnier	10	No	0	a tarasov	0	22	10	30	02:55
17	Hugh Byrne	10	44	0	33	0	Fiona lyng	0	Yes	10	Brian MacNamee	10	7	0	30	29:48
18	Hugh Byrne	10	40	0	30	0	Franck Bonnier	10	Yes	10	Brian MacNamee	10	4	0	40	04:09
19	Hugh Byrne	10	26	0	16	0	Mary	0	Yes	10	Brian	10	15	0	30	03:38
20	Hugh Byrne	10	47	0	22	0	furong tian	0	Yes	10	Brian MacNamee	10	14	0	30	03:43
21	Hugh Byrne	10	29	0	13	0	Franck Bonnier	10	Yes	10	Patrick Lindstorm	0	30	10	40	04:50
22	Hugh Bryne	10	28	0	8	0	Franck Bonnier	10	Yes	10	Bryan Duggan	0	7	0	30	05:44
23	Hugh Byrne	10	60-80	10	Too much!	0	Franck Bonnier or A	10	Yes	10	Brain MacNamee	10	7	0	50	08:42
24	Hugh Byrne	10	28	0	7	0	Franck Bonnier	10	No	0	Brian MacNamee	10		0	30	04:02
25	Hugh Byrne	10	53	0	24	0	Franck Bonnier	10	Yes	10	Brian MacNamee	10	9	0	40	05:59
26	Hugh byrne	10	38	0	41	0	Franck Bonnier	10	Yes	10	Brian MacNamee	10	96	0	40	08:20
27	Hugh Bryan	10	22	0	108	0	Frank bonnier	10	No	0	Brian MacNamee	10	8	0	30	05:01
28	Hugh Byrne	10	18	0	24	0	Franck Bonnier	10	No	0	Brian Macnamme	10	10	0	30	05:20
29	Hugh Byrne	10	8	0	0	0	French bonnier	10	No	0	Brian mcnamee	10	8	0	30	02:51
30	Hugh Byrne	10	32	0	25	0	Franck Bonnier	10	Yes	10	Brian MacNamee	10	7	0	40	03:08
31	Hugh Byrne	10	21	0	4	0	Frank Bonnier	10	Yes	10	Brian MacNamee	10	7	0	40	05:10
32	Hugh Byrne	10	36	0	10	0	Franck Bonnier	10	Yes	10	Brian MacNamee	10	9	0	40	05:46

Figure 5.2: The results of scoring survey 2

5.3.2 Statistical Analysis

In this section, analysis of surveys' results will be analyzed through Statistical Package for the Social Sciences (SPSS) software. The various features of the data are analyzed to facilitate the description of the distinct characteristics of the data in the database and the overall characteristics of the data. First of all, descriptive statistics will be shown in both datasets. Then, the correlations between score and time will be presented in two datasets separately, in order to check whether there is a statistical significance between them. Lastly, U-test will be used to figure out which visualization is more readable.

Descriptive Statistics

Descriptive statistics is an integral part of the statistical analysis, which can present useful features of the dataset. In this section, analyzing the comparison of accuracy rate in each question in two surveys, then statistics of survey 1 and survey 2 will be shown separately and comparatively.

The corresponding histograms shown in figure 5.3 is the comparison of accuracy rate in each question of both survey 1 and survey 2. In this figure, question 1, question 4 and question 6 have greater than sixty percent accuracy rate in both surveys, question 4 and question 6 are both accessibility testing tasks of topology-based tasks and also the only two of all, it can be estimated that both layouts can present accessibility of nodes and links well. On the other hand, question 2 and question 3 both have lower than 30% precision of accuracy, for question 3, even the accuracy is 0, which means no one got the right answer in the group of survey 2. Both questions are focusing on the number of nodes or links, the reasons of this problem probably because the database is too huge which contains more than 26,000 records, there are also many nodes and links in the visualization, it is difficult to count apparently. The overall distribution of score is presented in this figure.

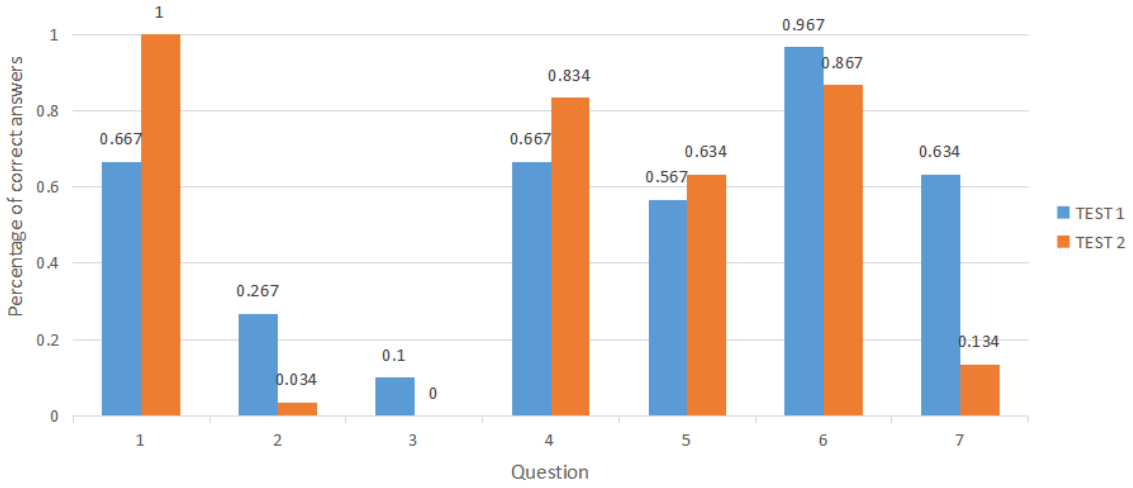


Figure 5.3: The comparison of accuracy in each question

The descriptive statistics of survey 1 is generated in figure 5.4,. From this table, it is obvious that the minimum score in survey 1 is 10 and the maximum score is 70, and the minimum time is 162 seconds and the maximum time is 2495 seconds. Because the format of the hour: minute: second cannot be used for analysis, changing it into seconds. The descriptive statistics of survey 2 is shown in figure 5.5, which presents basic information of results of survey 2, the minimum score and time are 20 and 133, the maximum score and time are 50 and 1788. In the comparison of the minimum/maximum of two datasets, it can be found that the gaps of the score range in survey 1 is much bigger than it in survey 2, which can be estimated that the ForceAtlas2 layout has higher and more stable performance than the Frunchterman Reingold layout.

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Score	30	60.00	10.00	70.00	38.6667	2.33580
Time	30	2333.00	162.00	2495.00	510.4333	88.99582
Valid N (listwise)	30					

Descriptive Statistics

	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Score	12.79368	163.678	.372	.427	.750	.833
Time	487.45020	237607.702	2.813	.427	9.300	.833
Valid N (listwise)						

Figure 5.4: descriptive statistics of survey 1

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Score	30	30.00	20.00	50.00	35.0000	1.49712
Time	30	1655.00	133.00	1788.00	324.5333	53.40392
Valid N (listwise)	30					

Descriptive Statistics

	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Score	8.20008	67.241	.000	.427	-.347	.833
Time	292.50532	85559.361	4.593	.427	23.290	.833
Valid N (listwise)						

Figure 5.5: descriptive statistics of survey 2

After completing descriptive statistics of survey 1 and survey 2 separately, do descriptive statistics for comparison of both. The tables of statistics will are shown in Appendix A.9, and box plots are used to analysis the difference between the results of survey 1 and survey 2 in both score and time variables, which is displayed in figure 5.6.

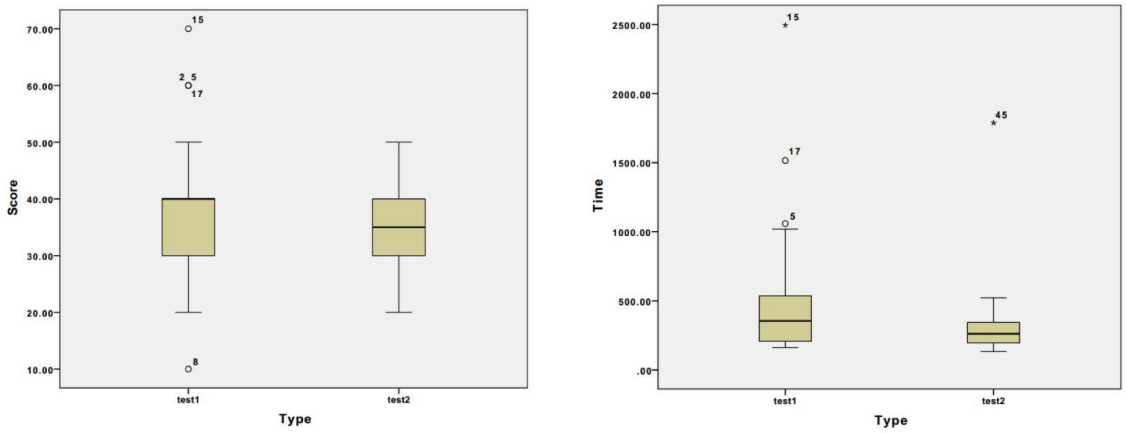


Figure 5.6: The box plots of time and score in the comparison of survey 1 and survey 2

In the variable of score, it can be seen that both boxes of survey 1 and survey 2 are the same size and position located between score 30 to 40, which means the overall data distribution of both surveys are equal, but there are three outliers exist in survey 1 and no outliers in survey 2, which says the results of survey 2 are more stable and concentrated, it also means the difficulty degree of readability of the ForceAtlas2 layout is more durable than the Frunchterman Reingold layout. On the other hand, the median of survey 1 is higher than survey 2, which means the accuracy rate of the results of survey 1 is higher than survey 2, it also says the accuracy of readability of the Frunchterman Reingold layout is higher than the ForceAtlas2 layout.

In the variable of time, it is evident that the time used in survey 2 is much less than survey 1, the median is lower as well, and there are three outliers exist in survey 1 but only one outlier in survey 2, in addition, the data distribution in survey 2 is more concentrated, which mean the time users spent in survey 2 is fundamentally similar. So in the variable of time, the ForceAtlas layout performs better than the Frunchterman layout.

Descriptive statistics is not enough for analyzing the final results, but some information can be seen for essential assessment. In conclusion of descriptive statistics, the accuracy of the score of survey 1 is higher than survey 2 but also spend more time, and the distribution of score and time in survey 2 is much more stable and concentrated than survey 1.

Correlation Analysis

After descriptive statistics, the correlations of score and time will be analyzed in test 1 and test 2 separately. The Pearson correlation coefficient is used to examines the relations of these two factors. For survey 1, the results are shown in figure 5.7, the Pearson correlation coefficient is 0.633, which is in the interval between 0.40 and 0.69, means there is a moderate correlation between score and time in survey 1, and also it is apparent that it is a positive correlation. The significant difference between two variable is 0, which means there is a significant linear correlation between them, so it can be indicated that the more time spending, the higher score obtained in survey 1.

Correlations

		Score	Time
Score	Pearson Correlation	1	.633**
	Sig. (2-tailed)		.000
	N	30	30
Time	Pearson Correlation	.633**	1
	Sig. (2-tailed)	.000	
	N	30	30

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 5.7: The correlations analysis of score and time in survey 1

The results of survey 2 are presented in figure 5.8, the Pearson correlation coefficient is 0.002, which means there is no correlation between score and time in survey 2, and also the significant difference between two variable is 0.992, which means there is no significant linear correlation between them.

Correlations

		Score	Time
Score	Pearson Correlation	1	.002
	Sig. (2-tailed)		.992
	N	30	30
Time	Pearson Correlation	.002	1
	Sig. (2-tailed)	.992	
	N	30	30

Figure 5.8: The correlations analysis of score and time in survey 2

Nonparametric Test – U Test

The last test is to check whether there is a statistically significant difference between these two layouts with two independent groups. In this research, the sample is two independent groups, so MannWhitney U test can be used to determine whether two separate samples were selected from populations having the same distribution. In other words, it means the results of which layout can provide the more readable representation for co-authorship in Arrow will be figured out.

Variable score is examined by U test first in figure 5.9. The null hypothesis is that the distribution of score is same in both surveys. The asymptotic significances of the score is 0.240, which is much bigger than the significance level 0.5, the null hypothesis is retained, so there is no significant difference between survey 1 and survey 2 in the variable of the score.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Score is the same across categories of Type.	Independent-Samples Mann-Whitney U Test	.240	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Figure 5.9: u-test for the distribution of score

Variable time is examined by U test as well in figure 5.10. The null hypothesis is that the distribution of time is same in both surveys. The asymptotic significances of the score is 0.038, which is smaller than the significance level 0.5, the null hypothesis is rejected, so there is a significant difference between survey 1 and survey 2 in the variable of the score.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Time is the same across categories of Type.	Independent-Samples Mann-Whitney U Test	.038	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Figure 5.10: u-test for the distribution of time

Based on the previous analysis, there is no significant difference between survey 1 and survey 2 in variable of score but there is significant difference between survey 1 and survey 2 in variable of

time, and according to the descriptive statistics mentioned before, survey 2 performs better than survey 1, which means the time used less in survey 2 than survey 1, which can conclude that the results of survey 2 are better than survey 1, in another word, ForceAtlas2 approach can provide the more readable representation than Frunchterman Reingold approach applied in the visualization for co-authorship in DIT Arrow.

5.3.3 Summary and discussion

In this chapter, an online survey has been designed and built for evaluation of two different visualizations, including seven questions focused on nodes tasks, links tasks, and accessibility tasks. Then this online survey has been sent to two independent groups, 40 people each, 68 results in total have been collected, after data cleaning, 30 results of each group have been saved into new two datasets. These two datasets have been analyzed by statistical testing, which contained three processes: descriptive statistics, Pearson correlation, and MannWhitney U test. Based on these statistical tests, the ForceAtlas2 layout can provide the more readable representation than Frunchterman Reingold layout to the viewers in DIT Arrow.

In the analysis of MannWhitney U test, the results of the variable of score means that there is no significant difference between two surveys, in other words, both layouts can express the correct information about authors and co-authorships between authors. However, there exists significant difference in the variable of time, which indicates that the time spent differently in two layouts, and according to previous statistics, time paid much less in survey 2 than survey 1, and the distribution of time in survey 2 is much more stable and concentrated, based on the readability contains both accuracies of answers as well as the time spent, so the precise results can be derived that ForceAtlas2 layout can provide the most readable graphs of co-authorships in DIT Arrow, it is probably suggested that ForceAtlas2 layout can be used in DIT Arrow for proving an excellent clear visualization to the users in the future.

Chapter 6

Conclusion

6.1 Research Overview

This research investigates the comparison of visualization between the layout of forces directed and the layout of radial. Begin with the research, the related work about data visualization are reviewed and studied, including the contents of applications, algorithms, and the relevant tools. The most complicated and confusing part of this research is data scraping, which contains some necessary processes, which are the analysis of Arrow structure, the implementation of coding, and the data cleansing.

The visualization for Arrow is implemented using Gephi, and the designed survey is conducted to obtain the results of answering from end users. Furthermore, the Mann-Whitney U test is performed to determine whether the significant difference appears. It reveals that the force directed layout is much more suitable to be used in the data visualization for Arrow. The discussion and analysis of key findings are presented as well.

6.2 Problem Definition

In fact, there has no signification research paper and applications about data visualization on DIT Arrow. The research is proposed and motivated by this point. What is more, both of the force directed layout, and the radial layout can produce the readable and explicit visualizations in some areas, whereas the radial layout could not have a positive effect on the Arrow visualization, even the other digital library systems. Therefore, the comparison of these two visualization layouts will be evaluated by surveys.

6.3 Experiment, Evaluation and Results

The designed experiment in this research was conducted to generate the various visualizations based on the different layout algorithms. After obtaining all visualizations, the Mann-Whitney U test is conducted to find out the significant differences among all visualizations. The result is that the visualization based on the force directed layout is better than radial layout's for DIT Arrow repository. In another word, the radial layout is not suitable to be used for the visualization in the field of digital library.

6.4 Contributions and Impact

One of the contribution in this research is that the primary co-authorship dataset is designed, created and prepared. The research indicated that the force directed layout can yield the high-quality visualization over radial layout's output. It suggest that the force directed layout can be used for the visualization of co-authorship rather than use radial layout.

6.5 Future Work & Recommendations

According to the consideration of the limitation in this research, few future work can be done to improve the quality of the study.

- The information of reference could be treated as a factor that probably influences in the insight of network and the structure of co-authorship visualization.
- The attribution of discipline for each article is an important factor that allows clients to view the relationships between authors based on the discipline. And the relationship of discipline can indicate the relationship of cooperation between authors.
- The more clients would be involved in the survey testing, providing the adequate records to guarantee the reliability of results.
- The API of scraping could be created to dynamically obtain and update the database on the back-end for DIT Arrow repository.
- The dynamic survey will be provided for viewing and testing by more clients because the dynamic network can provide the multiple types of visualizations.

References

Ahlberg, C., & Shneiderman, B. (1994). The alphaslider: a compact and rapid selector. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 365–371).

Baur, M., Brandes, U., Lerner, J., & Wagner, D. (2009). Group-level analysis and visualization of social networks. *Algorithmics of large and complex networks*, 330–358.

Biuk-Aghai, R. P. (2006). Visualizing co-authorship networks in online wikipedia. In *Communications and information technologies, 2006. iscit'06. international symposium on* (pp. 737–742).

Byrd, D. (1999). A scrollbar-based visualization for document navigation. In *Proceedings of the fourth acm conference on digital libraries* (pp. 122–129).

Carlis, J. V., & Konstan, J. A. (1998). Interactive visualization of serial periodic data. In *Proceedings of the 11th annual acm symposium on user interface software and technology* (pp. 29–38).

Coleman, M. K., & Parker, D. S. (1996). Aesthetics-based graph layout for human consumption. *Software: Practice and Experience*, *26*(12), 1415–1438.

Collberg, C., Kobourov, S., Nagra, J., Pitts, J., & Wampler, K. (2003). A system for graph-based visualization of the evolution of software. In *Proceedings of the 2003 acm symposium on software visualization* (pp. 77–ff).

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, *21*(11), 1129–1164.

Gajer, P., Goodrich, M. T., & Kobourov, S. G. (2000). A fast multi-dimensional algorithm for drawing large graphs. In *Graph drawing00 conference proceedings* (pp. 211–221).

Hadany, R., & Harel, D. (2001). A multi-scale algorithm for drawing graphs nicely. *Discrete Applied Mathematics*, *113*(1), 3–21.

Harel, D., & Koren, Y. (2000). A fast multi-scale method for drawing large graphs. In *Proceedings of the working conference on advanced visual interfaces* (pp. 282–285).

Hetzler, B., Whitney, P., Martucci, L., & Thomas, J. (1998). Multi-faceted insight through interoperable visual information analysis paradigms. In *Information visualization, 1998. proceedings. ieee symposium on* (pp. 137–144).

Hoffman, P., Grinstein, G., Marx, K., Grosse, I., & Stanley, E. (1997). Dna visual and analytic data mining. In *Visualization'97., proceedings* (pp. 437–441).

Hong, J. Y., DAndries, J., Richman, M., & Westfall, M. (2003). Zoomology: comparing two large hierarchical trees. *Posters Compendium of Information Visualization*, 120–121.

Inselberg, A., & Dimsdale, B. (1987). Parallel coordinates for visualizing multi-dimensional geometry. In *Computer graphics 1987* (pp. 25–44). Springer.

Institute, B. M. (2001). Interactive visualization of multiple query results. In *Presented at ieee symposium on information visualization.*

Kaizer, J., & Hodge, A. (2005). Aquabrowser library: Search, discover, refine. *Library Hi Tech News*, *22*(10), 9–12.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information processing letters*, *31*(1), 7–15.

Keim, D. A., Mansmann, F., Schneidewind, J., & Schreck, T. (2006). Monitoring network traffic with radial traffic analyzer. In *Visual analytics science and technology, 2006 ieee symposium on* (pp. 123–128).

Kurosawa, T., & Takama, Y. (2011). Predicting researchers' future activities using visualization system for co-authorship networks. In *Proceedings of the 2011 ieee/wic/acm international conferences on web intelligence and intelligent agent technology-volume 01* (pp. 332–339).

Lawson, R. (2015). Web scraping with python.

Livnat, Y., Agutter, J., Moon, S., Erbacher, R. F., & Foresti, S. (2005). A visualization paradigm for network intrusion detection. In *Information assurance workshop, 2005. iaw'05. proceedings from the sixth annual ieee smc* (pp. 92–99).

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

Northway, M. L. (1952). A primer of sociometry.

Plaisant, C. (n.d.). Information visualization evaluation.

Plaisant, C., Grosjean, J., & Bederson, B. B. (2002). Spacetree: Supporting exploration in large node link tree, design evolution and empirical evaluation. In *Information visualization, 2002. infovis 2002. ieee symposium on* (pp. 57–64).

Playfair, W. (1801). *The statistical breviary; shewing, on a principle entirely new, the resources fo every state and kingdom in europe; illustated with stained copper-plate charts, representing the physical powers of each distinct nation with ease and perspicuity. by william playair.* T. Bensley, J. Wallis [etc., etc.].

Richardson, L. (2013). Beautiful soup. *Crummy: The Site*.

Santamaría, R., & Therón, R. (2008). Overlapping clustered graphs: Co-authorship networks visualization. In *International symposium on smart graphics* (pp. 190–199).

Spoerri, A. (2004). Rankspiral: Toward enhancing search results visualizations. In *Information visualization, 2004. infovis 2004. ieee symposium on* (pp. p18–p18).

Spritzer, A. S., Volquind, F. P., & Freitas, C. M. (n.d.). Case study of co-authorship networks using a tool for graph visualization.

Suntinger, M., Obweger, H., Schiefer, J., & Groller, M. E. (2008). The event tunnel: Interactive visualization of complex event streams for business process pattern analysis. In *Visualization symposium, 2008. pacificvis'08. ieee pacific* (pp. 111–118).

Sutcliffe, A. G., Ennis, M., & Hu, J. (2000). Evaluating the effectiveness of visual user interfaces for information retrieval. *International Journal of Human-Computer Studies*, *53*(5), 741–763.

Teoh, S. T., & Kwan-Liu, M. (2002). Rings: A technique for visualizing large hierarchies. In *International symposium on graph drawing* (pp. 268–275).

Torres, R. S., Silva, C. G., Medeiros, C. B., & Rocha, H. V. (2003). Visual structures for image browsing. In *Proceedings of the twelfth international conference on information and knowledge management* (pp. 49–55).

Trafton, J. G., Kirschenbaum, S. S., Tsui, T. L., Miyamoto, R. T., Ballas, J. A., & Raymond, P. D. (2000). Turning pictures into numbers: extracting and generating information from complex visualizations. *International Journal of Human-Computer Studies*, *53*(5), 827–850.

Tutte, W. T. (1963). How to draw a graph. *Proceedings of the London Mathematical Society*, *3*(1), 743–767.

Van Berendonck, C., & Jacobs, T. (2003). Bubbleworld: a new visual information retrieval technique. In *Proceedings of the asia-pacific symposium on information visualisation-volume 24* (pp. 47–56).

Walshaw, C., et al. (2000). A multilevel algorithm for force-directed graph drawing. In *Graph drawing* (Vol. 1984, pp. 171–182).

Weber, M., Alexa, M., & Müller, W. (2001). Visualizing time-series on spirals. In *Infovis* (Vol. 1, pp. 7–14).

Wilkinson, L. (2006). *The grammar of graphics*. Springer Science & Business Media.

Wu, Y., & Takatsuka, M. (2006). Visualizing multivariate network on the surface of a sphere. In *Proceedings of the 2006 asia-pacific symposium on information visualisation-volume 60* (pp. 77–83).

Appendix A

Additional content

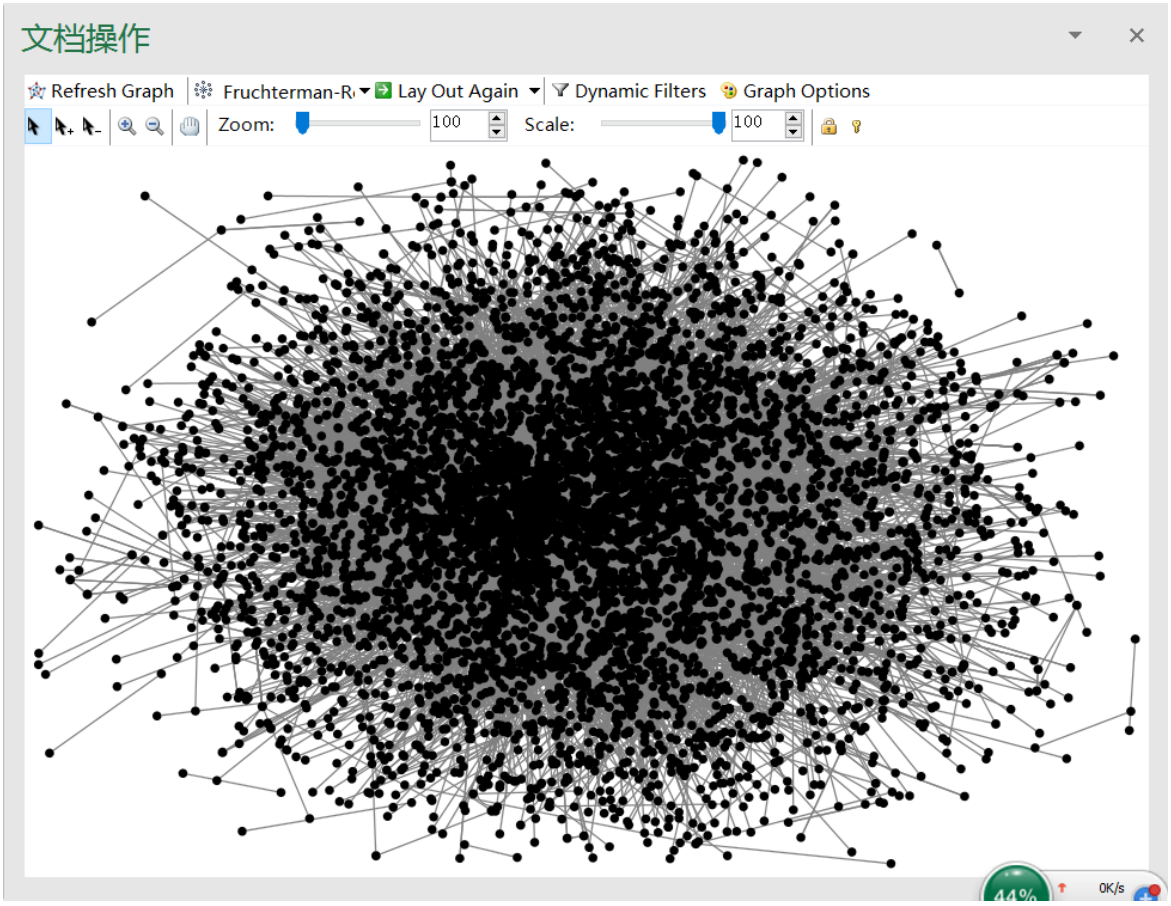


Figure A.1: The Frunchterman Reingold layout generated in NodeXL

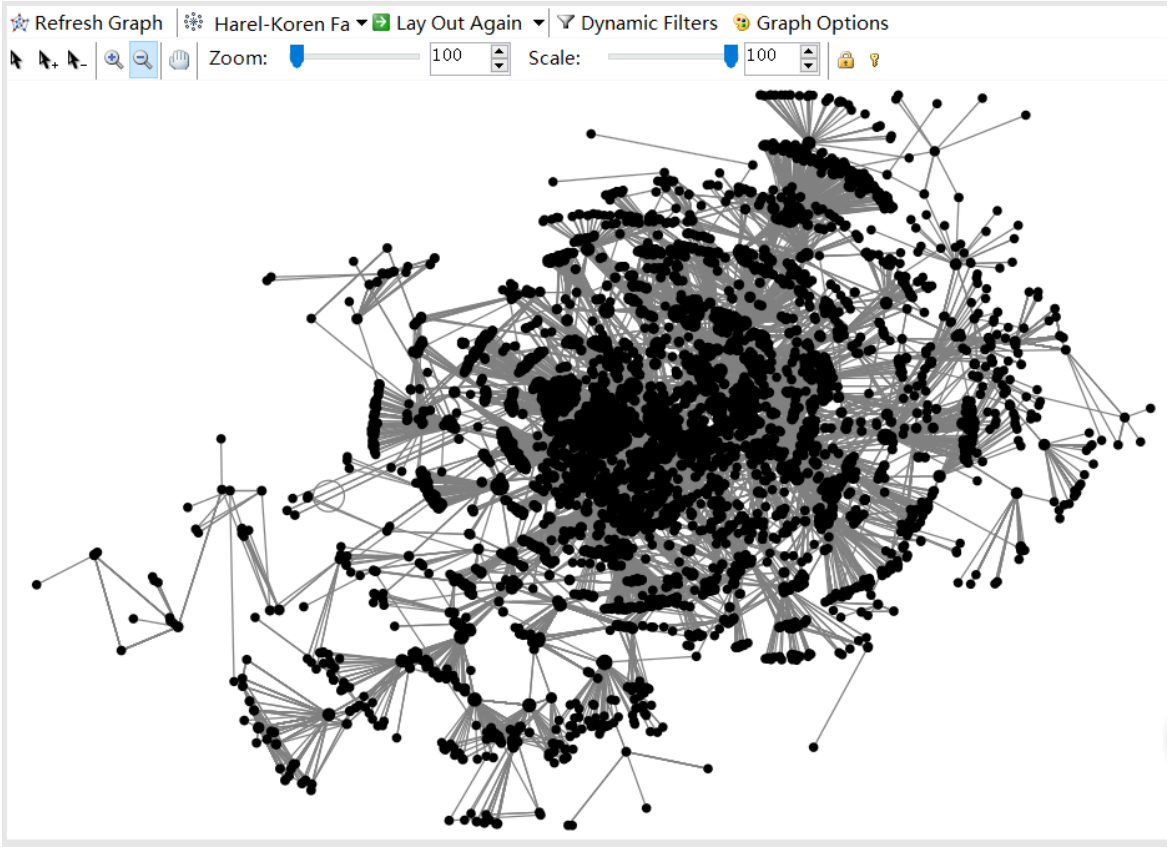


Figure A.2: The Harel-Koren layout generated in NodeXL

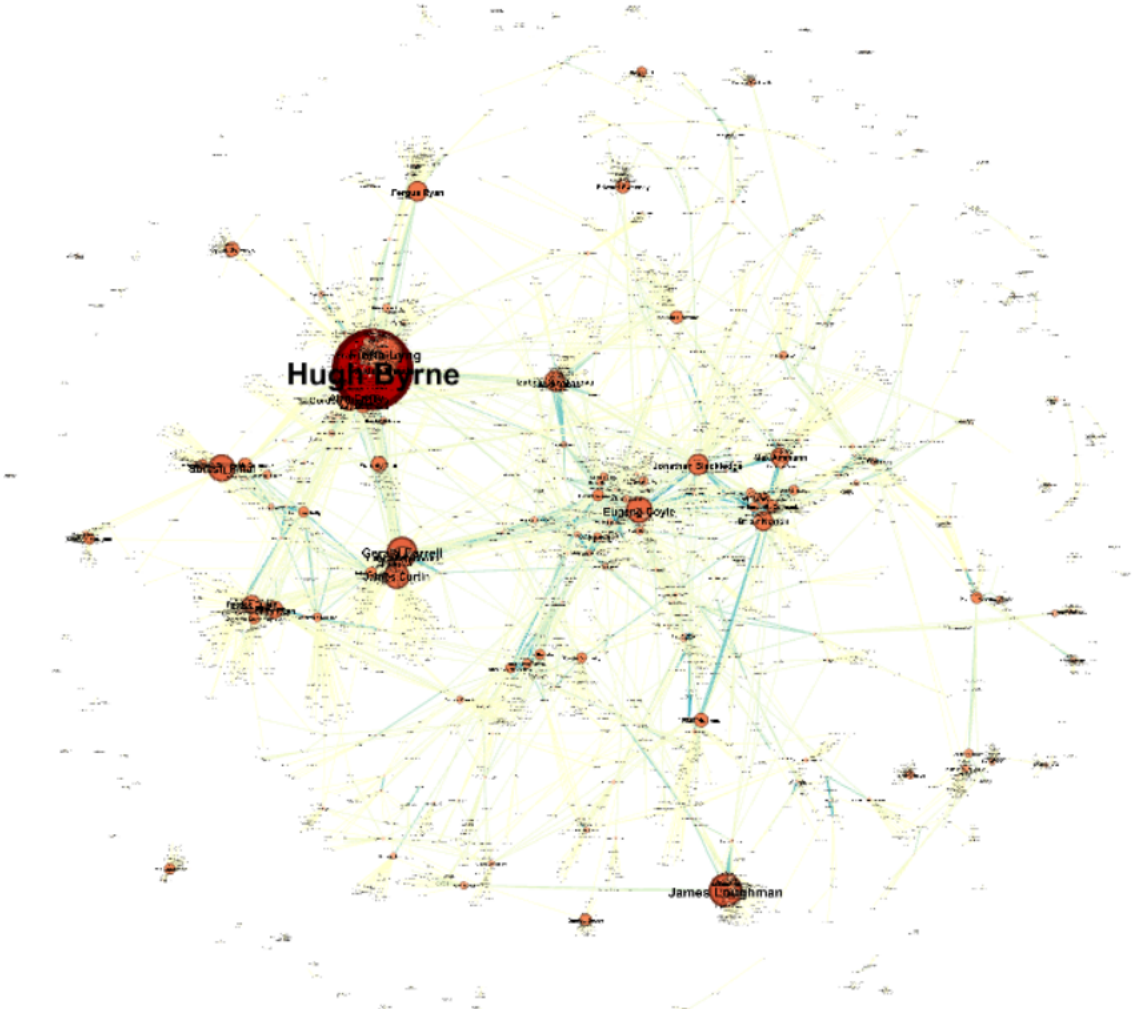


Figure A.3: The ForceAtlas layout

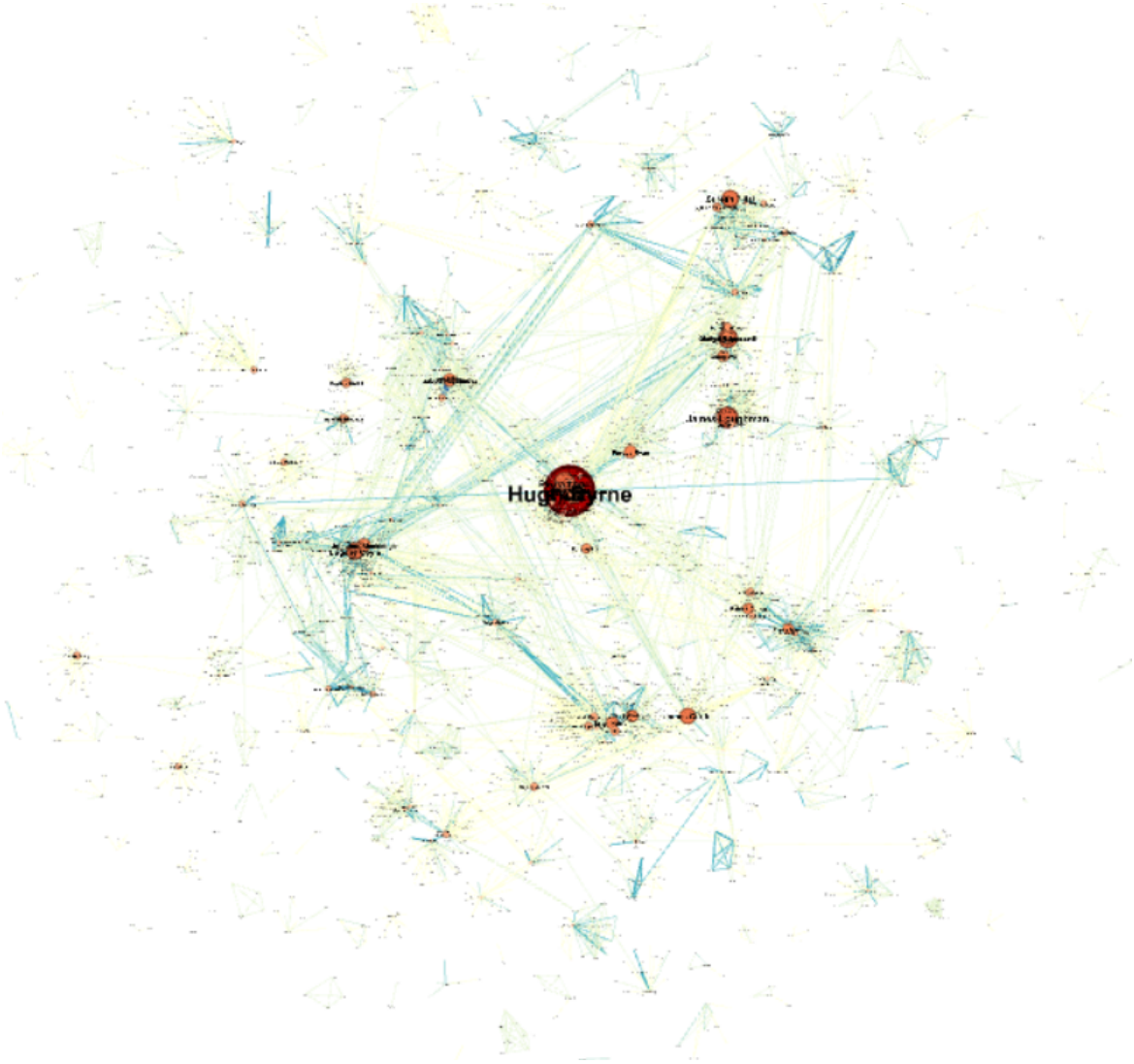


Figure A.4: The OpenOrd layout

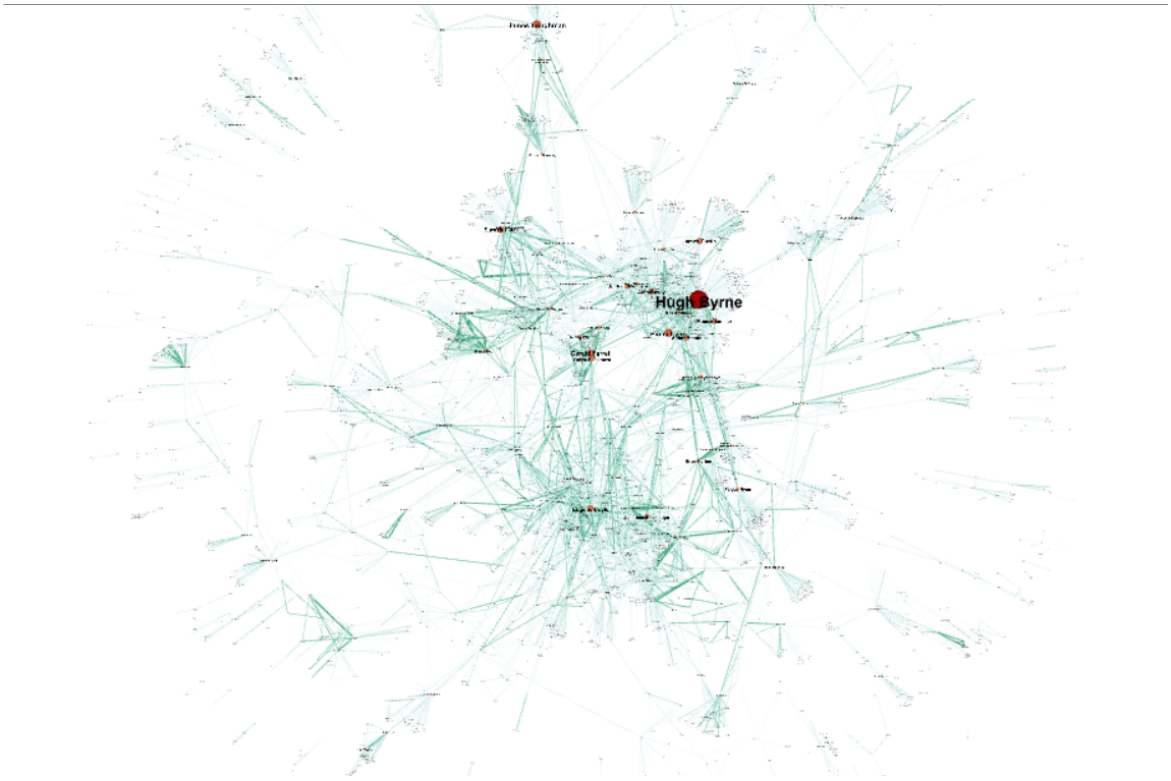


Figure A.5: The Yifan Hu layout

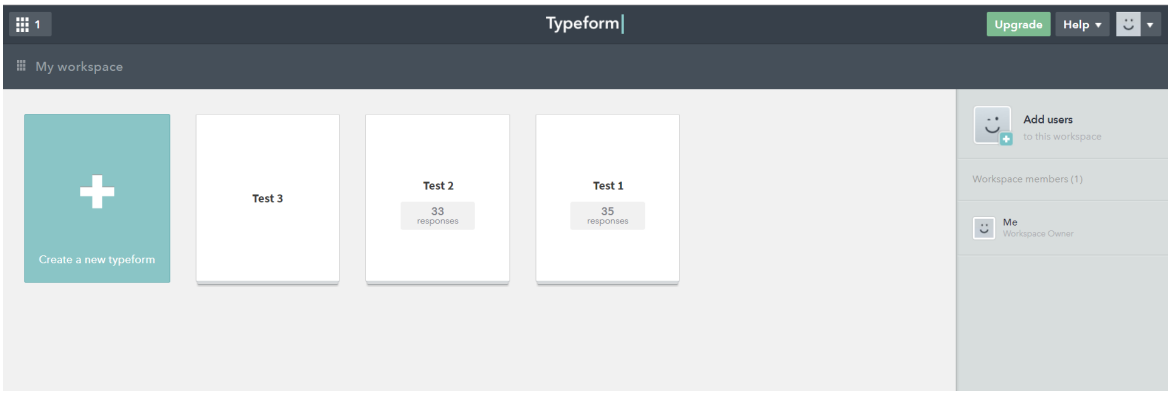
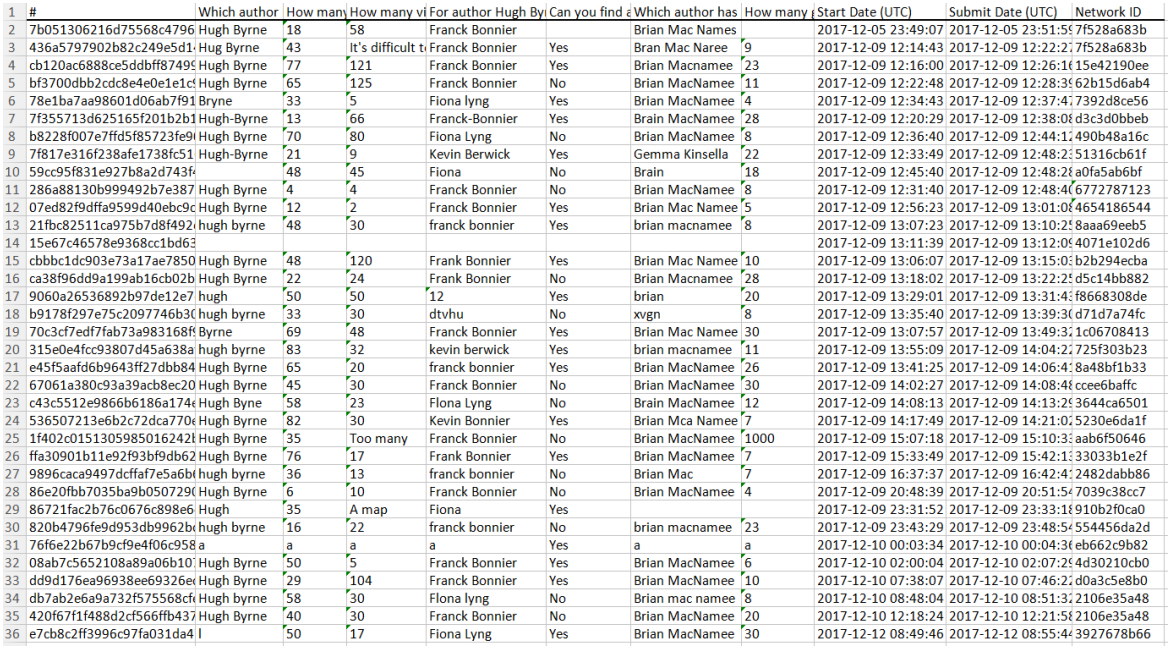
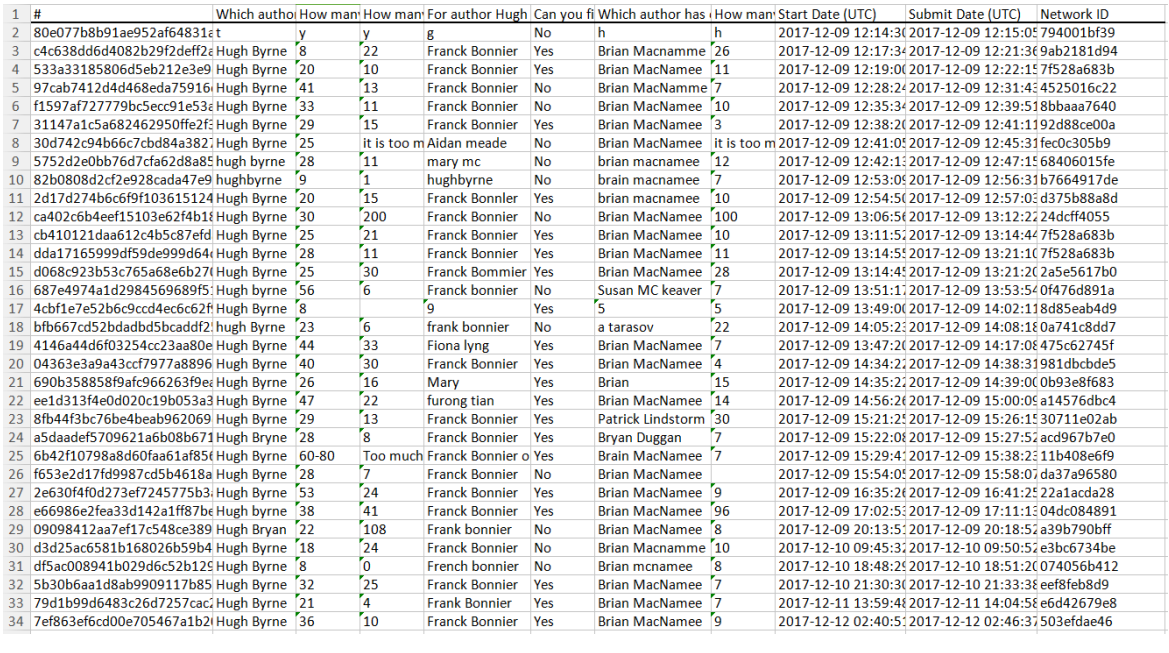


Figure A.6: The screen-shot of the final results of survey



#	Which author	How many	How many vi	For author Hugh By	Can you find a	Which author has	How many	Start Date (UTC)	Submit Date (UTC)	Network ID
7b051306216d75568c4796	Hugh Byrne	18	58	Franck Bonnier		Brian Mac Names		2017-12-05 23:49:07	2017-12-05 23:51:59	7f528a683b
436a5797902b82c249e5d1	Hug Byrne	43	It's difficult t	Franck Bonnier	Yes	Bran Mac Naree	9	2017-12-09 12:14:43	2017-12-09 12:22:27	7f528a683b
cb120ac6888ce5ddbff8749	Hugh Byrne	77	121	Franck Bonnier	Yes	Brian Macnamee	23	2017-12-09 12:16:00	2017-12-09 12:26:16	15e42190ee
bf3700dbb2cdc8e4e0e1e1c	Hugh Byrne	65	125	Franck Bonnier	No	Brian MacNamee	11	2017-12-09 12:22:48	2017-12-09 12:28:39	62b15d6ab4
78e1ba7aa98601d06ab7f91	Bryne	33	5	Fiona lyng	Yes	Brian MacNamee	4	2017-12-09 12:34:43	2017-12-09 12:37:47	7392d8ce56
7f355713d625165f201b2b1	Hugh-Byrne	13	66	Franck-Bonnier	Yes	Brain MacNamee	28	2017-12-09 12:20:29	2017-12-09 12:38:08	d3c3d0bbeb
b8228f007e7ffd5f85723fe9	Hugh Byrne	70	80	Fiona Lyng	No	Brian MacNamee	8	2017-12-09 12:36:40	2017-12-09 12:44:11	490b48a16c
7f817e316f238afe1738fc51	Hugh-Byrne	21	9	Kevin Berwick	Yes	Gemma Kinsella	22	2017-12-09 12:33:49	2017-12-09 12:48:23	51316cb61f
59cc95f831e927b8a2d743f		48	45	Fiona	No	Brain	18	2017-12-09 12:45:40	2017-12-09 12:48:28	a0fa5ab6bf
286a88130b999492b7e387	Hugh Byrne	4	4	Franck Bonnier	No	Brian MacNamee	8	2017-12-09 12:31:40	2017-12-09 12:48:40	6772787123
07ed82f9dffa9599d40ebc9c	Hugh Byrne	12	2	Franck Bonnier	Yes	Brian Mac Namee	5	2017-12-09 12:56:23	2017-12-09 13:01:08	4654186544
21fbc82511ca975b7d8f492	hugh byrne	48	30	franck bonnier	Yes	brian macnamee	8	2017-12-09 13:07:23	2017-12-09 13:10:25	8aaa69eeb5
15e67c46578e9368cc1bd63								2017-12-09 13:11:39	2017-12-09 13:12:09	4071e102d6
cbbbc1dc903e73a17ae7850	Hugh Byrne	48	120	Frank Bonnier	Yes	Brian Mac Namee	10	2017-12-09 13:06:07	2017-12-09 13:15:03	b2b294ecba
ca38f96dd9a199ab16cb02b	Hugh Byrne	22	24	Frank Bonnier	No	Brian Macnamee	28	2017-12-09 13:18:02	2017-12-09 13:22:25	d5c14bb882
9060a26536892b97de12e7	hugh	50	50	12	Yes	brian	20	2017-12-09 13:29:01	2017-12-09 13:31:43	f8668308de
b9178f297e75c2097746b3(	hugh byrne	33	30	dtvhu	No	xvgn	8	2017-12-09 13:35:40	2017-12-09 13:39:30	d71d7a74fc
70c3cf7edf7fab73a983168f	Byrne	69	48	Franck Bonnier	Yes	Brian Mac Namee	30	2017-12-09 13:07:57	2017-12-09 13:49:32	1c06708413
315e0e4fcc93807d45a638a	hugh byrne	83	32	kevin berwick	Yes	brian macnamee	11	2017-12-09 13:55:09	2017-12-09 14:04:22	725f303b23
e45f5aafd6b9643ff27dbb84	Hugh Byrne	65	20	franck bonnier	Yes	Brian MacNamee	26	2017-12-09 13:41:25	2017-12-09 14:06:43	8a48bf1b33
67061a380c93a39acb8ec20	Hugh Byrne	45	30	Franck Bonnier	No	Brian MacNamee	30	2017-12-09 14:02:27	2017-12-09 14:08:48	ccee6baffc
c43c5512e9866b6186a174	Hugh Byne	58	23	Flona Lyng	No	Brain MacNamee	12	2017-12-09 14:08:13	2017-12-09 14:13:29	3644ca6501
536507213e6b2c72dca770	Hugh Byrne	82	30	Kevin Bonnier	Yes	Brian Mca Namee	7	2017-12-09 14:17:49	2017-12-09 14:21:05	5230e6da1f
1f402c0151305985016242	Hugh Byrne	35	Too many	Franck Bonnier	No	Brian MacNamee	1000	2017-12-09 15:07:18	2017-12-09 15:10:33	aab6f50646
ffa30901b11e92f93bf9db62	Hugh Byrne	76	17	Frank Bonnier	Yes	Brian MacNamee	7	2017-12-09 15:33:49	2017-12-09 15:42:13	33033b1e2f
9896caca9497dcffaf7e5a6b	hugh byrne	36	13	franck bonnier	No	Brian Mac	7	2017-12-09 16:37:37	2017-12-09 16:42:43	2482dabb86
86e20fbb7035ba9b0507291	Hugh Byrne	6	10	Franck Bonnier	No	Brian MacNamee	4	2017-12-09 20:48:39	2017-12-09 20:51:54	7039c38cc7
86721fac2b76c0676c898e6	Hugh	35	A map	Fiona	Yes			2017-12-09 23:31:52	2017-12-09 23:33:18	910b2f0ca0
820b4796fe9d953db9962b	hugh byrne	16	22	franck bonnier	No	brian macnamee	23	2017-12-09 23:43:29	2017-12-09 23:48:54	554456da2d
76f6e22b67b9cf9e4f06c958	a	a	a	a	Yes	a	a	2017-12-10 00:03:34	2017-12-10 00:04:36	eb662c9b82
08ab7c5652108a89a06b10	Hugh Byrne	50	5	Franck Bonnier	Yes	Brian MacNamee	6	2017-12-10 02:00:04	2017-12-10 02:07:29	4d30210cb0
dd9d176ea96938ee69326e	Hugh Byrne	29	104	Franck Bonnier	Yes	Brian MacNamee	10	2017-12-10 07:38:07	2017-12-10 07:46:22	d0a3c5e8b0
db7ab2e6a9a732f575568cf	Hugh Byrne	58	30	Flona lyng	No	Brian mac namee	8	2017-12-10 08:48:04	2017-12-10 08:51:32	2106e35a48
420f67f1f488d2cf566ffb437	Hugh Byrne	40	30	Franck Bonnier	No	Brian MacNamee	20	2017-12-10 12:18:24	2017-12-10 12:21:58	2106e35a48
e7cb8c2ff3996c97fa031da4	I	50	17	Fiona Lyng	Yes	Brian MacNamee	30	2017-12-12 08:49:46	2017-12-12 08:55:44	3927678b66

Figure A.7: The initial results of survey 1



#	Which autho	How man	How man	For author Hugh	Can you fi	Which author has	How man	Start Date (UTC)	Submit Date (UTC)	Network ID
80e077b8b91ae952af64831	t	y	y	g	No	h	h	2017-12-09 12:14:30	2017-12-09 12:15:05	794001bf39
c4c638dd6d4082b29f2deff2	Hugh Byrne	8	22	Franck Bonnier	Yes	Brian Macnamme	26	2017-12-09 12:17:34	2017-12-09 12:21:36	9ab2181d94
533a33185806d5eb212e3e9	Hugh Byrne	20	10	Franck Bonnier	Yes	Brian MacNamee	11	2017-12-09 12:19:00	2017-12-09 12:22:15	7f528a683b
97cab7412d4d468eda75916	Hugh Byrne	41	13	Franck Bonnier	No	Brian MacNamme	7	2017-12-09 12:28:24	2017-12-09 12:31:43	4525016c22
f1597af727779bc5ecc91e53	Hugh Byrne	33	11	Franck Bonnier	No	Brian MacNamee	10	2017-12-09 12:35:34	2017-12-09 12:39:51	8bbaaa7640
31147a1c5a682462950ffe2f	Hugh Byrne	29	15	Franck Bonnier	Yes	Brian MacNamee	3	2017-12-09 12:38:20	2017-12-09 12:41:19	2d88ce00a
30d742c94b66c7cbd84a382	Hugh Byrne	25	it is too m	Aidan meade	No	Brian MacNamee	it is too m	2017-12-09 12:41:05	2017-12-09 12:45:31	ec0c305b9
5752d2e0bb76d7cfa62d8a85	hugh byrne	28	11	mary mc	No	brian macnamee	12	2017-12-09 12:42:13	2017-12-09 12:47:15	68406015fe
82b0808d2cf2e928cada47e9	hughbyrne	9	1	hughbyrne	No	brain macnamee	7	2017-12-09 12:53:09	2017-12-09 12:56:31	b7664917de
2d17d274b6c6f9f103615124	Hugh Byrne	20	15	Franck Bonnler	Yes	brian macnamee	10	2017-12-09 12:54:56	2017-12-09 12:57:03	d375b88a8d
ca402c6b4eef15103e62f4b1	Hugh Byrne	30	200	Franck Bonnier	No	Brian MacNamee	100	2017-12-09 13:06:56	2017-12-09 13:12:22	24dcff4055
cb410121daa612c4b5c87efd	Hugh Byrne	25	21	Franck Bonnier	Yes	Brian MacNamee	10	2017-12-09 13:11:52	2017-12-09 13:14:44	7f528a683b
dda17165999df59de999d64	Hugh Byrne	28	11	Franck Bonnier	Yes	Brian MacNamee	11	2017-12-09 13:14:55	2017-12-09 13:21:10	7f528a683b
d068c923b53c765a68e6b27	Hugh Byrne	25	30	Franck Bommier	Yes	Brian MacNamee	28	2017-12-09 13:14:45	2017-12-09 13:21:20	2a5e5617b0
687e4974a1d2984569689f5	Hugh Byrne	56	6	Franck bonnier	No	Susan MC keaver	7	2017-12-09 13:51:17	2017-12-09 13:53:54	0f476d891a
4cbf1e7e52b6c9ccd4ec6c62f	Hugh Byrne	8		9	Yes	5	5	2017-12-09 13:49:00	2017-12-09 14:02:11	8d85eab4d9
bfb667cd52bdadbd5bcaddf2	hugh Byrne	23	6	frank bonnier	No	a tarasov	22	2017-12-09 14:05:23	2017-12-09 14:08:18	0a741c8dd7
4146a44d6f03254cc23aa80e	Hugh Byrne	44	33	Fiona lyng	Yes	Brian MacNamee	7	2017-12-09 13:47:20	2017-12-09 14:17:08	475c62745f
04363e3a9a43ccf7977a8896	Hugh Byrne	40	30	Franck Bonnier	Yes	Brian MacNamee	4	2017-12-09 14:34:22	2017-12-09 14:38:31	981dbcbde5
690b358858f9afc966263f9e	Hugh Byrne	26	16	Mary	Yes	Brian	15	2017-12-09 14:35:22	2017-12-09 14:39:00	0b93e8f683
ee1d313f4e0d020c19b053a3	Hugh Byrne	47	22	furong tian	Yes	Brian MacNamee	14	2017-12-09 14:56:26	2017-12-09 15:00:09	a14576dbc4
8fb44f3bc76be4beab962069	Hugh Byrne	29	13	Franck Bonnier	Yes	Patrick Lindstorm	30	2017-12-09 15:21:25	2017-12-09 15:26:15	30711e02ab
a5daadef5709621a6b08b671	Hugh Bryne	28	8	Franck Bonnier	Yes	Bryan Duggan	7	2017-12-09 15:22:08	2017-12-09 15:27:52	acd967b7e0
6b42f10798a8d60faa61af85	Hugh Byrne	60-80	Too much	Franck Bonnier o	Yes	Brain MacNamee	7	2017-12-09 15:29:47	2017-12-09 15:38:23	11b408e6f9
f653e2d17fd9987cd5b4618a	Hugh Byrne	28	7	Franck Bonnier	No	Brian MacNamee		2017-12-09 15:54:05	2017-12-09 15:58:07	da37a96580
2e630f4f0d273ef7245775b3	Hugh Byrne	53	24	Franck Bonnier	Yes	Brian MacNamee	9	2017-12-09 16:35:26	2017-12-09 16:41:25	22a1acda28
e66986e2fea33d142a1ff87b	Hugh Byrne	38	41	Franck Bonnier	Yes	Brian MacNamee	96	2017-12-09 17:02:53	2017-12-09 17:11:13	04dc084891
09098412aa7ef17c548ce389	Hugh Bryan	22	108	Frank bonnier	No	Brian MacNamee	8	2017-12-09 20:13:53	2017-12-09 20:18:52	a39b790bff
d3d25ac6581b168026b59b4	Hugh Byrne	18	24	Franck Bonnier	No	Brian Macnamme	10	2017-12-10 09:45:32	2017-12-10 09:50:52	e3bc6734be
df5ac008941b029d6c52b125	Hugh Byrne	8	0	French bonnier	No	Brian mcnamee	8	2017-12-10 18:48:29	2017-12-10 18:51:20	074056b412
5b30b6aa1d8ab9909117b85	Hugh Byrne	32	25	Franck Bonnier	Yes	Brian MacNamee	7	2017-12-10 21:30:30	2017-12-10 21:33:38	eef8feb8d9
79d1b99d6483c26d7257cac	Hugh Byrne	21	4	Frank Bonnier	Yes	Brian MacNamee	7	2017-12-11 13:59:48	2017-12-11 14:04:58	e6d42679e8
7ef863ef6cd00e705467a1b2	Hugh Byrne	36	10	Franck Bonnier	Yes	Brian MacNamee	9	2017-12-12 02:40:53	2017-12-12 02:46:37	503efdae46

Figure A.8: The initial results of survey 2

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Score	60	60.00	10.00	70.00	36.8333	1.39595
Time	60	2362.00	133.00	2495.00	417.4833	52.85688
Valid N (listwise)	60					

Descriptive Statistics

	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Score	10.81300	116.921	.505	.309	1.220	.608
Time	409.42761	167630.966	3.362	.309	12.886	.608
Valid N (listwise)						

Figure A.9: descriptive statistics of survey1 and survey 2