University of Glasgow | School of Computing Science

Are matrix-based or node-linked graphs more readable when representing causal relationships for social and health data?

Kristina Lazarova

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

Level 4 Project — March, 2017

Abstract

We show how to produce a level 4 project report using latex and pdflatex using the style file l4proj.cls

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: _____ Signature: _____

Contents

Chapter 1

Literature Review

1.1 Data Visualisation

Data visualisation has been described as a technique that makes use of computer-supported, interactive illustrations to deepen human's understanding of a dataset [4]. Information visualisation is necessary when dealing with increasingly large and complex data.

Information visualisation systems are most helpful when a set of data is being explored [8]. Usually this occurs not when someone is looking for a specific answer to a question, but when a deeper understanding of that data set is sought. It was suggested that the value of information visualisation is not in understanding a specific question, it is about developing and deepening one's insights of a set of data [8]. Data visualisation is able to facilitate this process because there are a number of cognitive benefits associated with it. A large visual cue that illustrates data becomes a single point of reference for human cognitive processes. Visuals become external cognition helpers in facilitating human memory by providing a bigger working set for analysing data. Furthermore, visual architecture and design applied at company and department levels have been reported to be successful due to the low cognitive burden for visualization reading [12].

//examples, why is visualisation good and areas in which it helps

//readability

A well known issue with data visualisation is that sometimes it is challenging for people to understand it. A study aimed to examine familiarity of museum visitors with different visualisation techniques [3]. They included charts, maps, graphs, and networks to reveal how familiar people are with them. It was found that even though most participants were interested in science and art they have difficulties naming and reading the visualizations. It was concluded that people are interested in visualisation techniques, but have significant difficulties in naming and understanding them.

In order to solve this issue, another research area has focused on understanding how information visualization novices think and the approaches that can facilitate their learning. A study used sales data to find the barriers for novices when reading iterative visualisation construction and the way they think about visualisation specifications [10]. They found that the biggest barriers were interpreting questions into data factors, visual mappings, and understanding the visual representations. It was found that there is a need for instruments that suggest possible visualizations, facilitate help with learning, and are integrated with tool support for the whole analytic process. Furthermore, recent research acknowledged that individual differences between people will have influence on the way they interpret graphical representations [15]. They suggested that visualization performance can be improved by personalising visuals according to one's needs, abilities and preferences.

1.2 Causal Relationships

Causal relationships are of great interest for scientist who examine influence of different factors on each other.

A study looking to identify factors influencing blog design used the Decision Making Trial and Evaluation Laboratory method (DALMATEL) which is used to illustrate the relationships between factors and allows causal relationships to be shown [11]. Some of the causal relationships they found were that color arrangement directly impacts simplicity of layout, colour arrangement directly impacts font arrangement, and color arrangement impacts itself.

ReView is a tool for finding causal relationships in anomalies in network traffic. It has been suggested to facilitate better understanding of network representations [17]. One of its features is minimizing the detailed information while showing the causal relationship. ReView can also quickly navigate the user through networks with a large number of requests and levels of abstractions.

1.3 Node-Link Graphs

Node-link graphs consist of nodes which are connected by edges. Large amount of work has been dedicated to visualizing those structures in the most readable way. However, the larger the network that is being illustrated, the higher the possibility that the graph becomes dense and unreadable. This is especially true when the direction of the edges is important [7]. When representing causal relationships with node-link graphs if node n is causing node m, then the edge between those two nodes will point towards node m. In order to avoid the complexity of the large network some researchers have tried to add interaction with the graph [9], while others' intention was to create visuals that do not change the structure of the network and does not require interaction [7]. An innovative interaction technique was designed by Abello et al [1] where a user is able to navigate through a hierarchically clustered base of the graph. In the non interactive technique researchers decreased the complexity of large directed graphs by replacing single edges with edges which connect to groups of nodes [7].

Data driven journalism is also concerned with difficulties in showing directed relationships in large datasets [14].

1.4 Matrix-based Graphs

An Adjacency matrix is frequently used to represent a network [13]. If a network consists of n nodes, the matrix will consists n x n grid of cells. This is considered to be an unambiguous way of representing data. However, some of its disadvantages are that the area increases quadratically and as large networks are sparse there will mainly empty space on the matrix.

A new technique called Compressed Adjacency Matrices was introduced in 2012 for visualising gene regulatory networks [6]. As those directed networks have specific structural traits, standard representations such as adjacency matrices, and node-link diagrams are unable to depict all traits. Compressed Adjacency Matrices cut and rearrange adjacency matrix so that no space is wasted in case of sparse network. There are specific structures which represent sub networks. This is how scientists came up with a new data structure in order to fit the characteristics of the data they analyse.

Furthermore, PathwayMatrix is another visualisation tool that represents specific relations between proteins in a pathway [5]. The implementation of the tool consists of adjacent matrices that interact with each other. Additional features were added to facilitate the data analysis. This visualisation software received positive feedback

in the specific area of representing relations in proteins pathways. Consequently, there is no one best representations technique. Depending on the dataset specifications, the complexity and size of the data there might be many or only few sufficient ways to visualise it.

Chapter 2

Introduction

Introduction to the specific area of my research

Brain connectivity visualisations are in the form of weighted graphs which are node-link graphs in which each edge is given a numerical weight. Alper et al [2] compared augmented adjacency matrix with node-link visualization by conducting a controlled experiment. They found that matrix-based graphs outperform node-linked graphs.

Chapter 3

Implementation details

3.1 Software tools and technologies

Given the opportunity to chose any tools and technologies for the development of this web application was a very exciting task. However, I had to be certain that the right decisions are made. After a research period followed by a trial-error week it was decided that the backend of the application will built with Node.js and JavaScript combined with the web application framework Express. Node.js was chosen on the grounds of being event-driven, non-blocking I/O model which contributes to a very efficient and lightweight software. A Node.js JavaScript engine is also used in the Google Chrome browser. JavaScript servers have incredible performance due to their asynchronous I/O. Node.js appears to be single-threaded from a developer's point of view, as there is no thread management involved in the development process. However, behind the scenes Node.js handles threading, file system events, implements the event loop, feature thread pooling etc. Coming from a Java background, the Maven equivalent in Node.js is NPM. By using NPM commands the developer is able to install variety of different modules to help the implementation process. NPM executes the function of a package manager. Express is the standard server framework for Node.js. It is usually described as a minimal and flexible Node.js web application framework. Many popular frameworks such as KeystoneJS, Kraken and Sails, are built on Express.

AngularJS 1 was chosen for management of frontend functionality. Even though there is a newer version of the product, the lack of documentation and support online, was a sufficient reason for using the older AngularJS. It uses HTML as a template and enables the developers to extend it to express the application's components clearly. AngularJS supports features such as data binding and dependency injection which decreases the amount of code that a developer would usually write to implement them.

The database system chosen for the project is PostgreSQL. Considering the size of the project an object-relational database was chosen. In addition, it decided on PostgreSQL in particular because it is open source and has gained a reputation of a reliable database system. Also previous experience with PosgreSQL from developers point of view made the decision easier.

* maybe database schema will be added here *

3.2 Development Process

3.2.1 User Stories

The first step of the implementation was understanding the requirements and creating user stories. Some of the most important user stories are:

```
As a researcher, I want to be able to see participant's answers, so that I can anal
As a researcher, I want to keep participant's scores anonymous, so that my experime
As a participant, I want to be able to see graphs and associated questions, so that
As a participant, I want to be unable to go the next question, before completing th
```

3.2.2 Wire-frames

After the requirements gathering analysis, development of wire-frames followed. Balsamiq Mockups 3 is the software used for the creation of wire-frames. An example of the research question page can be found in figure 3.1.

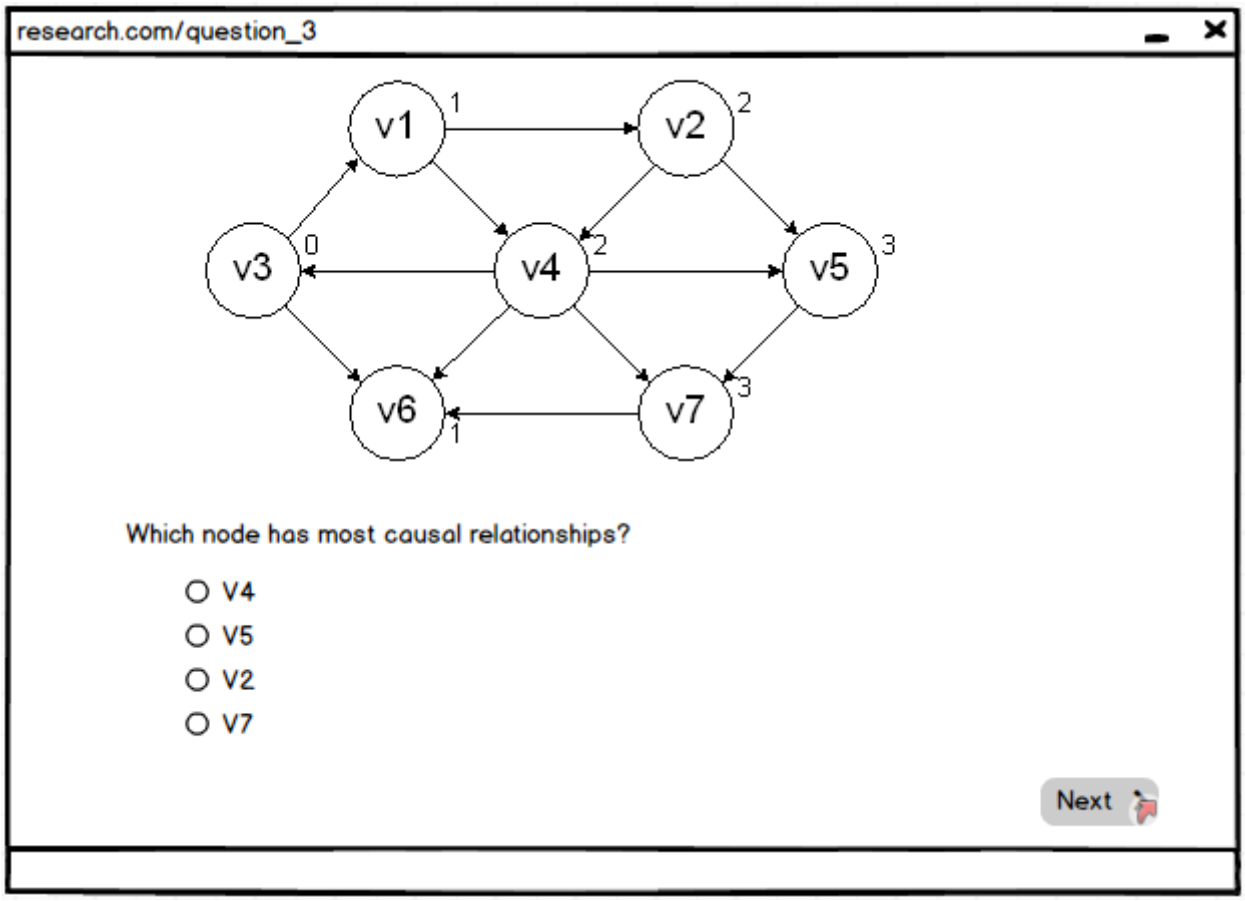

Figure 3.1: Research Question Wire-frame

After discussing the wire-frames it appeared that some important features are missing. One of those features was a participants training session. The aim of the experiment is to test which graph is more readable for people

who do not have regular exposure to such data visualisation. Therefore, it is important to make the participants aware of how to read each graph before the actual experiment. This way, the requirements specification became an iterative process during which a better understanding of the product evolved.

3.2.3 System Design

Designing the system was the next stage in the process.

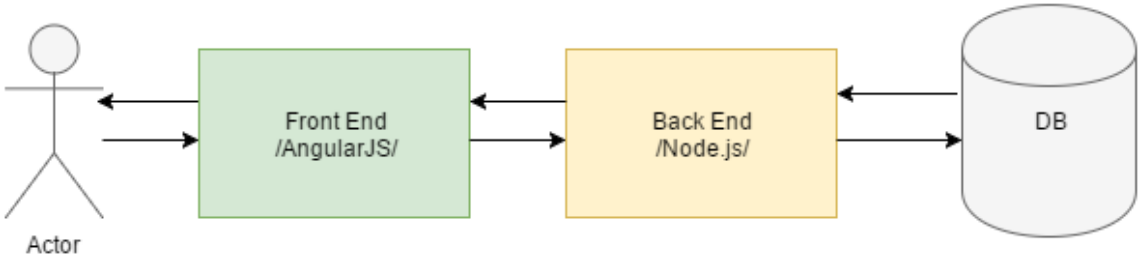

Figure 3.2: An abstract representation of the system design

Figure 3.2 shows an abstract view of the system design. There is an Actor who will either be a participant in the study or a researcher. They will interact with the front-end which will be in the form of a web application in a browser. The front end will communicate with the back-end which will be implemented in Node.js. The back-end will make requests to the database to retrieve and send information.

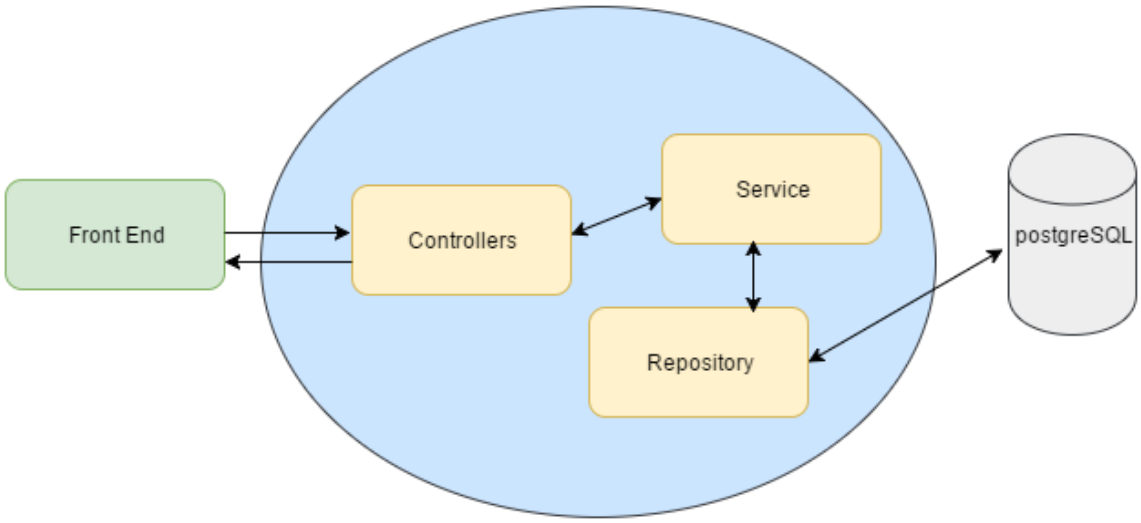

Figure 3.3: A more specific representation of the system design

Figure 3.3 displays a more detailed version of the system design. This particular design has been implemented to separate the different concerns in this specific system. When the Actor interacts with the application, the front-end will send information to the Controllers. There are many controllers because there is a controller for each page with front-end functionality. The Controller component decides what should the next action be according to the user input. It has control over the front-end logic and sending requests to the service if information from the database or the server is requested. For example, the answers to all questions are kept in the front-end until a ”Submit” button is clicked on. This action triggers a request to the Service. The Service component works with the back-end logic. It can send and retrieve data from the database and keep the information in the Repository.

The Service also deals with the requests for the different web-pages. Also it ensures that the project dependencies are loaded.

3.2.4 Implementation

The development process was split into front-end and back-end. Without using any frameworks, the front-end Hhtml pages were created following the wire-frames. Bootstrap was added to the html to improve the UI design and make it look more appealing for the participant.

By this time, it was clear that Node.js will be used to create a server so the next step was to implement it. The decision to use Express as a framework with Node.js followed and the html pages were mapped to a handlebars or hbs files.

The following couple of weeks were dedicate on work on the database system: creating database schema, and tables, and work on connecting it with the server.

3.3 Challenges

```
Spring idea failed
changed to Node.js
Angular compatibility with Node.js
```

In the beginning of this project the Java framework Spring was going to be used in the implementation as it is among the most widely used frameworks in industry [16]. This decision was supported by extensive previous experience with Java from developer's point of view and the applicability of the skills to be acquired. However, one of the reasons why Spring is used in industry is because of the large and complex systems that exist there. The Spring framework works on a very high level of abstraction where you can easily write configuration files to add dependencies from different project. Therefore, it is considered rather unfriendly for small independent projects and developers with limited Spring experience. The reasoning behind this conclusion was provoked after a couple of unsuccessful attempts to set relative paths to different CSS and JavaScript files. The issue was found to be in the web application configuration file. This is how the very simple task of reading a css file turned to be a long tedious debugging process after which the realisation that Spring is unnecessary abstract and complex for this project occurred.

A new research for web-application frameworks followed. Node.js backend was chosen because of its event-driven, non-blocking I/O model which creates an efficient and lightweight server-side of the application. Another challenge appeared when trying to incorporate AngularJS with Node.js. Usually in AngularJS one uses curly braces to reference data structure from the AngularJS controller. However, Node.js also uses curly brackets to reference information from the backend in the frontend. After a long research it was found that Node.js overrides the use of curly braces and the application is not displaying Angular data as it expects it come from the backend. Unfortunately, an appropriate error message does exist and it all had to be discovered during the development process. Instead of using curly brackets one can also use "ng-bind" and achieve the same result. This approached solved the issue until "ng-bind" information was need in "ng-src" to display the appropriate graph image. It is not possible to use "ng-bind" inside "ng-src" so the present solution at the time was no longer solving the problem. Therefore, the Angular configurations had be altered to use a different symbol. Implementing this solved the problem entirely.

3.4 Software reliability testing

Chapter 4

Evaluation

4.1 Design

4.2 Participants

4.3 Procedure

4.4 Results

4.5 Discussion

Chapter 5

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

5.1 The Lazy Dog

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog. The quick brown jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

Figure 5.1: An alternative hierarchy of the algorithms.

Appendices

Appendix A

Database Schema

```
participants_answers(question_id, participant_id, answer, time)
participants_info(participant_id,participant_name, email, uni_degree, age)
questions(question_id, question, one, two, three, four, correct, image)
```

Appendix B

Running the Programs

An example of running from the command line is as follows:

```
> java MaxClique BBMC1 brock200_1.clq 14400
```

This will apply $BBMC$ with $style = 1$ to the first brock200 DIMACS instance allowing 14400 seconds of cpu time.

Appendix C

Generating Random Graphs

We generate Erdós-Rënyi random graphs $G(n, p)$ where n is the number of vertices and each edge is included in the graph with probability p independent from every other edge. It produces a random graph in DIMACS format with vertices numbered 1 to n inclusive. It can be run from the command line as follows to produce a clq file

```
> java RandomGraph 100 0.9 > 100-90-00.clq
```

Bibliography

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

[2] Basak Alper, Benjamin Bach, Nathalie Henry Riche, Tobias Isenberg, and Jean-Daniel Fekete. Weighted graph comparison techniques for brain connectivity analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 483–492. ACM, 2013.

[3] Katy Börner, Adam Maltese, Russell Nelson Balliet, and Joe Heimlich. Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization*, page 1473871615594652, 2015.

[4] Stuart K Card, Jock D Mackinlay, and Ben Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[5] Tuan N Dang, Paul Murray, and Angus G Forbes. Pathwaymatrix: visualizing binary relationships between proteins in biological pathways. In *BMC proceedings*, volume 9, page S3. BioMed Central Ltd, 2015.

[6] Kasper Dinkla, Michel A Westenberg, and Jarke J van Wijk. Compressed adjacency matrices: untangling gene regulatory networks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2457–2466, 2012.

[7] Tim Dwyer, Nathalie Henry Riche, Kim Marriott, and Christopher Mears. Edge compression techniques for visualization of dense directed graphs. *IEEE transactions on visualization and computer graphics*, 19(12):2596–2605, 2013.

[8] Jean-Daniel Fekete, Jarke J Van Wijk, John T Stasko, and Chris North. The value of information visualization. In *Information visualization*, pages 1–18. Springer, 2008.

[9] Emden R Gansner, Yehuda Koren, and Stephen C North. Topological fisheye views for visualizing large graphs. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):457–468, 2005.

[10] Lars Grammel, Melanie Tory, and Margaret-Anne Storey. How information visualization novices construct visualizations. *IEEE transactions on visualization and computer graphics*, 16(6):943–952, 2010.

[11] Chun-Cheng Hsu. Evaluation criteria for blog design and analysis of causal relationships using factor analysis and dematel. *Expert Systems with Applications*, 39(1):187–193, 2012.

[12] John King, Kathy Sonderer, and Kevin Lynch. Cognitive benefits of a simple visual metrics architecture. In *International Conference on HCI in Business, Government and Organizations*, pages 319–329. Springer, 2016.

[13] William JR Longabaugh. Combing the hairball with biofabric: a new approach for visualization of large networks. *BMC bioinformatics*, 13(1):1, 2012.

[14] Christina Niederer, Wolfgang Aigner, and Alexander Rind. Survey on visualizing dynamic, weighted, and directed graphs in the context of data-driven journalism. *Proceedings of the International Summer School on Visual Computing*, pages 49–58, 2015.

[15] Ben Steichen, Giuseppe Carenini, and Cristina Conati. User-adaptive information visualization: Using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 317–328, New York, NY, USA, 2013. ACM.

[16] Kuikui Liuös Xiujin Shiás and Yue Liës. Integrated Architecture for Web Application Development Based on Spring Framework and Activiti Engine. *The International Conference on E-Technologies and Business on the Web (EBW2013)*, pages 52–56, 2013.

[17] Hao Zhang, Maoyuan Sun, Danfeng (Daphne) Yao, and Chris North. Visualizing traffic causality for analyzing network anomalies. In *Proceedings of the 2015 ACM International Workshop on International Workshop on Security and Privacy Analytics*, IWSPA '15, pages 37–42, New York, NY, USA, 2015. ACM.